



BE COMP MCQ PDF

ML

Jordan PDF

APPROVED

Our Telegram Channel

https://t.me/SPPU_BE_COMP_BOOKS_EXAMS

Team Members:
Tatyा Vinchu
Sergio Marquina

This sheet is for 3 Mark questions

S.r No	Question	Image	a	b	c	d	Correct Answer
e.g 1	Write down question	img.jpg	Option a	Option b	Option c	Option d	a/b/c/d
1	Which of the following is characteristic of best machine learning method ?		fast	accuracy	scalable	All above	D
2	What are the different Algorithm techniques in Machine Learning?		Supervised Learning and Semi-	Unsupervised Learning and Transduction	Both A & B	None of the Mentioned	C
3	_____ can be adopted when it's necessary to categorize a large amount of data with a few complete examples or when there's the need to impose some _____.		Supervised	Semi-supervised	Reinforcement	Clusters	B
4	In reinforcement learning, this feedback is usually called as _____.		Overfitting	Overlearning	Reward	None of above	C
5	In the last decade, many researchers started training bigger and bigger models, built with several different layers that's why this approach is called _____.		Deep learning	Machine learning	Reinforcement learning	Unsupervised learning	A
6	What does learning exactly mean?		Robots are programed so that they can	A set of data is used to discover the potentially predictive relationship.	Learning is the ability to change	It is a set of data is used to discover the	C
7	When it is necessary to allow the model to develop a generalization ability and avoid a common problem called _____.		Overfitting	Overlearning	Classification	Regression	A
8	Techniques involve the usage of both labeled and unlabeled data is called _____.		Supervised	Semi-supervised	Unsupervised	None of the above	B
9	there's a growing interest in pattern recognition and associative memories whose structure and functioning are similar to what happens in the neocortex. Such an _____ showed better performance than other approaches, even without a context-based model		Regression	Accuracy	Modelfree	Scalable	C
10			Machine learning	Deep learning	Reinforcement learning	Supervised learning	B
11	Which of the following sentence is correct?	--	Machine learning relates with the study.	Data mining can be defined as the process in which the unstructured data tries to extract knowledge or unknown interesting patterns.	Both A & B	None of the above	C
12	What is 'Overfitting' in Machine learning?	--	when a statistical model describes random error or noise instead of	Robots are programed so that they can perform the task based on data they gather from sensors.	While involving the process of learning 'overfitting' occurs.	a set of data is used to discover the potentially predictive relationship	A

13	What is ‘Test set’?	--	Test set is used to test the accuracy of the hypotheses generated by the learner.	It is a set of data is used to discover the potentially predictive relationship.	Both A & B	None of above	A
14	what is the function of ‘Supervised Learning’?	--	Classifications, Predict time series, Annotate strings	Speech recognition, Regression	Both A & B	None of above	C
15	Commons unsupervised applications include	--	Object segmentation	Similarity detection	Automatic labeling	All above	D
16	Reinforcement learning is particularly efficient when_____.	--	the environment is not completely deterministic	it's often very dynamic	it's impossible to have a precise error measure	All above	D
17	During the last few years, many _____ algorithms have been applied to deep neural networks to learn the best policy for playing Atari video games and to teach an agent how to associate the right action with an input representing the state.	--	Logical	Classical	Classification	None of above	D
18	Common deep learning applications include_____	--	Image classification, Real-time visual tracking	Autonomous car driving, Logistic optimization	Bioinformatics, Speech recognition	All above	D
19	if there is only a discrete number of possible outcomes (called categories), the process becomes a _____.	--	Regression	Classification.	Modelfree	Categories	B
20	Which of the following are supervised learning applications	--	Spam detection, Pattern detection, Natural Language Processing	Image classification, Real-time visual tracking	Autonomous car driving, Logistic optimization	Bioinformatics, Speech recognition	A

21	<p>Let's say, you are working with categorical feature(s) and you have not looked at the distribution of the categorical variable in the test data.</p> <p>You want to apply one hot encoding (OHE) on the categorical feature(s). What challenges you may face if you have applied OHE on a categorical variable of train dataset?</p>	--	All categories of categorical variable are not present in the test dataset.	Frequency distribution of categories is different in train as compared to the test dataset.	Train and Test always have same distribution.	Both A and B	D
22	Which of the following sentence is FALSE regarding regression?	--	It relates inputs to outputs.	It is used for prediction.	It may be used for interpretation.	It discovers causal relationships.	D
23	Which of the following method is used to find the optimal features for cluster analysis	--	k-Means	Density-Based Spatial Clustering	Spectral Clustering Find clusters	All above	D
24	scikit-learn also provides functions for creating dummy datasets from scratch:	--	make_classification()	make_regression()	make_blobs()	All above	D
25	_____ which can accept a NumPy RandomState generator or an integer seed.	--	make_blobs	random_state	test_size	training_size	B
26	In many classification problems, the target dataset is made up of categorical labels which cannot immediately be processed by any algorithm. An encoding is needed and scikit-learn offers at least _____ valid options	--	1	2	3	4	B
27	In which of the following each categorical label is first turned into a positive integer and then transformed into a vector where only one feature is 1 while all the others are 0.	--	LabelEncoder class	DictVectorizer	LabelBinarizer class	FeatureHasher	C
28	_____ is the most drastic one and should be considered only when the dataset is quite large, the number of missing features is high, and any prediction could be risky.	--	Removing the whole line	Creating sub-model to predict those features	Using an automatic strategy to input them according to the other known values	All above	A
29	It's possible to specify if the scaling process must include both mean and standard deviation using the parameters _____.	--	with_mean=True/False	with_std=True/False	Both A & B	None of the Mentioned	C
30	Which of the following selects the best K high-score features.	--	SelectPercentile	FeatureHasher	SelectKBest	All above	C

31	How does number of observations influence overfitting? Choose the correct answer(s). Note: Rest all parameters are same 1. In case of fewer observations, it is easy to overfit the data. 2. In case of fewer observations, it is hard to overfit the data. 3. In case of more observations, it is easy to overfit the data. 4. In case of more observations, it is hard to overfit the data.	--	1 and 4	2 and 3	1 and 3	None of theses	A
32	Suppose you have fitted a complex regression model on a dataset. Now, you are using Ridge regression with tuning parameter lambda to reduce its complexity. Choose the option(s) below which describes relationship of bias and variance with lambda.	--	In case of very large lambda; bias is low, variance is low	In case of very large lambda; bias is low, variance is high	In case of very large lambda; bias is high, variance is low	In case of very large lambda; bias is high, variance is high	C
33	What is/are true about ridge regression? 1. When lambda is 0, model works like linear regression model 2. When lambda is 0, model doesn't work like linear regression model 3. When lambda goes to infinity, we get very, very small coefficients approaching 0 4. When lambda goes to infinity, we get very, very large coefficients approaching infinity	--	1 and 3	1 and 4	2 and 3	2 and 4	A
34	Which of the following method(s) does not have closed form solution for its coefficients?	--	Ridge regression	Lasso	Both Ridge and Lasso	None of both	B
35	Function used for linear regression in R is _____	--	lm(formula, data)	lr(formula, data)	lrm(formula, data)	regression.linear(formula, data)	A
36	In the mathematical Equation of Linear Regression $Y = \beta_1 + \beta_2 X + \epsilon$, (β_1, β_2) refers to	--	(X-intercept, Slope)	(Slope, X-Intercept)	(Y-Intercept, Slope)	(slope, Y-Intercept)	C
37	Suppose that we have N independent variables (X_1, X_2, \dots, X_n) and dependent variable is Y. Now Imagine that you are applying linear regression by fitting the best fit line using least square error on this data. You found that correlation coefficient for one of it's variable(Say X_1) with Y is -0.95. Which of the following is true for X_1 ?	--	Relation between the X_1 and Y is weak	Relation between the X_1 and Y is strong	Relation between the X_1 and Y is neutral	Correlation can't judge the relationship	B

38	We have been given a dataset with n records in which we have input attribute as x and output attribute as y. Suppose we use a linear regression method to model this data. To test our linear regressor, we split the data in training set and test set randomly. Now we increase the training set size gradually. As the training set size increases, what do you expect will happen with the mean training error?	--	Increase	Decrease	Remain constant	Can't Say	D
39	We have been given a dataset with n records in which we have input attribute as x and output attribute as y. Suppose we use a linear regression method to model this data. To test our linear regressor, we split the data in training set and test set randomly. What do you expect will happen with bias and variance as you increase the size of training data?	--	Bias increases and Variance increases	Bias decreases and Variance increases	Bias decreases and Variance decreases	Bias increases and Variance decreases	D
40	Suppose, you got a situation where you find that your linear regression model is under fitting the data. In such situation which of the following options would you consider?1. I will add more variables2. I will start introducing polynomial degree variables3. I will remove some variables	--	1 and 2	2 and 3	1 and 3	1, 2 and 3	A
41	Problem: Players will play if weather is sunny. Is this statement is correct?	weather data.jpg	TRUE	FALSE			A
42	Multinomial Naïve Bayes Classifier is distribution		Continuous	Discrete	Binary		B
43	For the given weather data, Calculate probability of not playing	weather data.jpg	0.4	0.64	0.36	0.5	C
44	Suppose you have trained an SVM with linear decision boundary after training SVM, you correctly infer that your SVM model is under fitting. Which of the following option would you more likely to consider iterating SVM next time?	--	You want to increase your data points	You want to decrease your data points	You will try to calculate more variables	You will try to reduce the features	C
45	The minimum time complexity for training an SVM is O(n ²). According to this fact, what sizes of datasets are not best suited for SVM's?	--	Large datasets	Small datasets	Medium sized datasets	Size does not matter	A

46	The effectiveness of an SVM depends upon:	--	Selection of Kernel	Kernel Parameters	Soft Margin Parameter C	All of the above	D
47	What do you mean by generalization error in terms of the SVM?	--	How far the hyperplane is from the support vectors	How accurately the SVM can predict outcomes for unseen data	The threshold amount of error in an SVM		B
48	What do you mean by a hard margin?	--	The SVM allows very low error in classification	The SVM allows high amount of error in classification	None of the above		A
49	We usually use feature normalization before using the Gaussian kernel in SVM. What is true about feature normalization? 1. We do feature normalization so that new feature will dominate other 2. Some times, feature normalization is not feasible in case of categorical variables 3. Feature normalization always helps when we use Gaussian kernel in SVM	--	1	1 and 2	1 and 3	2 and 3	B
50	Support vectors are the data points that lie closest to the decision surface.	--	TRUE	FALSE			A
51	Which of the following is not supervised learning?	--	PCA	Decision Tree	Naive Bayesian	Linerar regression	A
52	Suppose you are using RBF kernel in SVM with high Gamma value. What does this signify?	--	The model would consider even far away points from hyperplane for modeling	The model would consider only the points close to the hyperplane for modeling	The model would not be affected by distance of points from hyperplane for modeling	None of the above	B
53	Gaussian Naïve Bayes Classifier is distribution	--	Continuous	Discrete	Binary		A

54	If I am using all features of my dataset and I achieve 100% accuracy on my training set, but ~70% on validation set, what should I look out for?	--	Underfitting	Nothing, the model is perfect	Overfitting		C
55	What is the purpose of performing cross-validation?	--	a. To assess the predictive performance of the models b. To judge how the trained model performs outside the sample on test data		c. Both A and B		C
56	Which of the following is true about Naive Bayes ?	--	a. Assumes that all the features in a dataset are equally important b. Assumes that all the features in a dataset are independent		c. Both A and B d. None of the above option		C
57	Suppose you are using a Linear SVM classifier with 2 class classification problem. Now you have been given the following data in which some points are circled red that are representing support vectors.If you remove the following any one red points from the data. Does the decision boundary will change?	svm.jpg	yes	no			A
58	Linear SVMs have no hyperparameters that need to be set by cross-validation	--	TRUE	FALSE			B
59	For the given weather data, what is the probability that players will play if weather is sunny	weather data.jpg	0.5	0.26	0.73	0.6	D
60	100 people are at party. Given data gives information about how many wear pink or not, and if a man or not. Imagine a pink wearing guest leaves, what is the probability of being a man	man.jpg	0.4	0.2	0.6	0.45	B
61	Problem: Players will play if weather is sunny. Is the statement TRUE or FALSE	weather data.jpg	TRUE	FALSE			a
62	For the given weather data, Calculate probability	weather data.jpg	0.4	0.64	0.29	0.75	b
63	For the given weather data, Calculate probability	weather data.jpg	0.4	0.64	0.36	0.5	c
64	For the given weather data, what is the probability	weather data.jpg	0.5	0.26	0.73	0.6	d
65	100 people are at party. Given data gives information about how many wear pink or not, and if a man or not. Imagine a pink wearing guest leaves, what is the probability of being a man	man.jpg	0.4	0.2	0.6	0.45	b

66	100 people are at party. Given data gives information about them. What is the best way to represent this data?	man.jpg	TRUE	FALSE			a
67	What do you mean by generalization error in terms of the SVM?	How far the hypothesis is from the decision boundary.	How accurately the SVM can predict outcome.	The threshold amount of error.			b
68	What do you mean by a hard margin?	The SVM allows high amount of error in classification.	The SVM allows high amount of error in classification.	None of the above			a
69	The minimum time complexity for training an SVM is O(n^2). According to you, which type of datasets will take more time to train?	Large datasets	Small datasets	Medium sized datasets	Size does not matter		a
70	The effectiveness of an SVM depends upon:	Selection of Kernel Parameters	Kernel Parameters	Soft Margin Parameter	All of the above		d
71	Support vectors are the data points that lie closest to the decision boundary.	TRUE	FALSE				a
72	The SVM's are less effective when:	The data is linearly separable.	The data is clean and ready to use.	The data is noisy and contains outliers.			c
73	Suppose you are using RBF kernel in SVM with high Gamma value. What will happen?	The model would consider only the points close to the center.	The model would consider only the points close to the center.	The model would consider all the points.	None of the above		b
74	The cost parameter in the SVM means:	The number of cross-validations to be made	The kernel to be used	The tradeoff between misclassification and simplicity of the model	None of the above		c
75	If I am using all features of my dataset and I achieve 100% accuracy, what is the problem?	Underfitting	Nothing, the model is perfect	Overfitting			c
76	Which of the following are real world applications of the SVM?	Text and Hypothesis	Image Classification	Clustering of NLP	All of the above		d
77	Suppose you have trained an SVM with linear decision boundary after training SVM, you correctly infer that your SVM model is under fitting. Which of the following option would you more likely to consider iterating SVM next time?	You want to increase the margin.	You want to decrease your data points.	You will try to change the kernel.	You will try to reduce the cost parameter.		
78	We usually use feature normalization before using the Gaussian kernel.	1	1 and 2	1 and 3	2 and 3		b
79	Linear SVMs have no hyperparameters that need to be set by cross-validation.	TRUE	FALSE				b
80	In a real problem, you should check to see if the SVM is separable and then choose the right kernel.	TRUE	FALSE				b

This sheet is for 2 Mark questions

S.r No	Question	Image	a	b	c	d	Correct Answer
e.g 1	Write down question	img.jpg	Option a Programmer	Option b Teacher	Option c Author	Option d Farmer	a/b/c/d B
1	A supervised scenario is characterized by the concept of a _____.						B
2	overlearning causes due to an excessive _____.		Capacity	Regression	Reinforcement	Accuracy	A
3	If there is only a discrete number of possible outcomes called _____.		Modelfree	Categories	Prediction	None of above	B
4	What is the standard approach to supervised learning?		split the set of example into the training set and the test	group the set of example into the training set and the test	a set of observed instances tries to induce a general	learns programs from data	A
5	Some people are using the term _____ instead of prediction only to avoid the weird idea that machine learning is a sort of modern magic.		Inference	Interference	Accuracy	None of above	A
6	The term _____ can be freely used, but with the same meaning adopted in physics or system theory.		Accuracy	Cluster	Regression	Prediction	D
7	Which are two techniques of Machine Learning ?		Genetic Programming and Inductive Learning	Speech recognition and Regression	Both A & B	None of the Mentioned	A
8	Even if there are no actual supervisors _____ learning is also based on feedback provided by the environment		Supervised	Reinforcement	Unsupervised	None of the above	B
9	Common deep learning applications / problems can also be solved using_____		Real-time visual object identification	Classic approaches	Automatic labeling	Bio-inspired adaptive systems	B
10	Identify the various approaches for machine learning.		Concept Vs Classification Learning	Symbolic Vs Statistical Learning	Inductive Vs Analytical	All above	D
11	what is the function of "Unsupervised Learning"?		Find clusters of the data and find low-dimensional	Find interesting directions in data and find novel coordinates and associations / database cleaning	All		D
12	What are the two methods used for the calibration in Supervised Learning?		Platt Calibration and Isotonic Regression	Statistics and Informal Retrieval			A
13	What is the standard approach to supervised learning?		split the set of example into the training set and the test	group the set of example into the training set and the test	a set of observed instances tries to induce a general	learns programs from data	A
14	Which of the following is not Machine Learning?		Artificial Intelligence	Rule based inference	Both A & B	None of the Mentioned	B
15	What is Model Selection in Machine Learning?		The process of selecting models among different	when a statistical model describes random error or noise instead of underlying	Find interesting directions in data and find novel	All above	A
16	_____ provides some built-in datasets that can be used for testing purposes.		scikit-learn	classification	regression	None of the above	A
17	While using _____ all labels are turned into sequential numbers.		LabelEncoder class	LabelBinarizer class	DictVectorizer	FeatureHasher	A
18	_____ produce sparse matrices of real numbers that can be fed into any machine learning model.		DictVectorizer	FeatureHasher	Both A & B	None of the Mentioned	C
19	scikit-learn offers the class _____, which is responsible for filling the holes using a strategy based on the mean, median, or frequency		LabelEncoder	LabelBinarizer	DictVectorizer	Imputer	D
20	Which of the following scale data by removing elements that don't belong to a given range or by considering a maximum absolute value.		MinMaxScaler	MaxAbsScaler	Both A & B	None of the Mentioned	C
21	Which of the following model include a backwards elimination feature selection routine?		MCV	MARS	MCRS	All above	B
22	Can we extract knowledge without apply feature selection		YES	NO			A

23	While using feature selection on the data, is the number of features decreases.		NO	YES			B
24	Which of the following are several models for feature extraction		regression	classification	None of the above		C
25	scikit-learn also provides a class for per-sample normalization, _____		Normalizer	Imputer	Classifier	All above	A
26	_____ dataset with many features contains information proportional to the independence of all features and their variance.		normalized	unnormalized	Both A & B	None of the Mentioned	B
27	In order to assess how much information is brought by each component, and the correlation among them, a useful tool is the _____.		Concurrent matrix	Convergence matrix	Supportive matrix	Covariance matrix	D
28	The _____ parameter can assume different values which determine how the data matrix is initially processed.		run	start	init	stop	C
29	allows exploiting the natural sparsity of data while extracting principal components.		SparsePCA	KernelPCA	SVD	init parameter	A
30	Which of the following is an example of a deterministic algorithm?		PCA	K-Means	None of the above		A
31	Let's say, a "Linear regression" model perfectly fits the training data (train error is zero). Now, Which of the following statement is true?		A. You will always have test error zero	B. You can not have test error zero	C. None of the above		c
32	In a linear regression problem, we are using "R-squared" to measure goodness-of-fit. We add a feature in linear regression model and retrain the same model.Which of the following option is true?		A. If R Squared increases, this variable is significant. B. If R Squared decreases, this variable is not significant.		C. Individually R squared cannot tell about variable importance. We can't say anything about it right now. D. None of these.		c
33	Which of the one is true about Heteroskedasticity?		A. Linear Regression with varying error terms B. Linear Regression with constant error terms	C. Linear Regression with zero error terms	D. None of these		a
34	Which of the following assumptions do we make while deriving linear regression parameters?1. The true relationship between dependent y and predictor x is linear2. The model errors are statistically independent3. The errors are normally distributed with a 0 mean and constant standard deviation4. The predictor x is non-stochastic and is measured error-free		A. 1,2 and 3.	B. 1,3 and 4.	C. 1 and 3.	D. All of above.	d
35	To test linear relationship of y(dependent) and x(independent) continuous variables, which of the following plot best suited?		A. Scatter plot B. Barchart	C. Histograms	D. None of these		a
36	Generally, which of the following method(s) is used for predicting continuous dependent variable?1. Linear Regression2. Logistic Regression		A. 1 and 2	B. only 1	C. only 2	D. None of these.	b
37	Suppose you are training a linear regression model. Now consider these points.1. Overfitting is more likely if we have less data2. Overfitting is more likely when the hypothesis space is small.Which of the above statement(s) are correct?		A. Both are False B. 1 is False and 2 is True	C. 1 is True and 2 is False D. Both are True			c
38	Suppose we fit "Lasso Regression" to a data set, which has 100 features (X1,X2...X100). Now, we rescale one of these feature by multiplying with 10 (say that feature is X1), and then refit Lasso regression with the same regularization parameter.Now, which of the following option will be correct?		A. It is more likely for X1 to be excluded from the model B. It is more likely for X1 to be included in the model	C. Can't say D. None of these			b
39	Which of the following is true about "Ridge" or "Lasso" regression methods in case of feature selection?		A. Ridge regression uses subset selection of features B. Lasso regression uses subset selection of features	C. Both use subset selection of features D. None of above			b
40	Which of the following statement(s) can be true post adding a variable in a linear regression model?1. R-Squared and Adjusted R-squared both increase2. R-Squared increases and Adjusted R-squared decreases3. R-Squared decreases and Adjusted R-squared decreases4. R-Squared decreases and Adjusted R-squared increases		A. 1 and 2 B. 1 and 3	C. 2 and 4 D. None of the above			a
41	We can also compute the coefficient of linear regression with the help of an analytical method called "Normal Equation". Which of the following is/are true about "Normal Equation"?1. We don't have to choose the learning rate2. It becomes slow when number of features is very large3. No need to iterate		A. 1 and 2 B. 1 and 3.	C. 2 and 3. D. 1,2 and 3.			d

42	How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?		A. 1	B. 2	C. Can't Say		b
43	If two variables are correlated, is it necessary that they have a linear relationship?		A. Yes	B. No			b
44	Correlated variables can have zero correlation coefficient. True or False?		A. True	B. False			a
45	Which of the following option is true regarding "Regression" and "Correlation" ?Note: y is dependent variable and x is independent variable.		A. The relationship is symmetric between x and y in both.	B. The relationship is not symmetric between x and y in both.	C. The relationship is not symmetric between x and y in case of correlation but in case of regression it is symmetric.	D. The relationship is symmetric between x and y in case of correlation but in case of regression it is not symmetric.	d
46	What is/are true about kernel in SVM?1. Kernel function map low dimensional data to high dimensional space2. It's a similarity function		1	2	1 and 2	None of these	c
47	Suppose you are building a SVM model on data X. The data X can be error prone which means that you should not trust an	Misclassification would	Data will be correctly classified	Can't say	None of these	a	
48	Suppose you are using a Linear SVM classifier with 2 class classification problem. Now you have b	svm.jpg	yes	no			a
49	If you remove the non-red circled points from the data, the decision boundary will change?	svm.jpg	TRUE	FALSE			b
50	When the C parameter is set to infinite, which of the following holds true?		The optimal hyperplane if exists, will be the one that completely separates the data	The soft-margin classifier will separate the data	None of the above		a
51	Suppose you are building a SVM model on data X. The data X can be error prone which means that you should not trust an	We can still classify data	We can not classify data correctly	Can't Say	None of these	a	
52	SVM can solve linear and non-linear problems		TRUE	FALSE			a
53	The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number		TRUE	FALSE			a
54	Hyperplanes are _____ boundaries that help classify the data points.	usual	decision	parallel			b
55	The _____ of the hyperplane depends upon the number of features.	dimension	classification	reduction			a
56	Hyperplanes are decision boundaries that help classify the data points.	TRUE	FALSE				a
57	SVM algorithms use a set of mathematical functions that are defined as the kernel.	TRUE	FALSE				a
58	In SVM, Kernel function is used to map a lower dimensional data into a higher dimensional data.	TRUE	FALSE				a
59	In SVR we try to fit the error within a certain threshold.	TRUE	FALSE				a
60	When the C parameter is set to infinite, which of the following holds true?		The optimal hyperplane if exists, will be the one that completely separates the data	The soft-margin classifier will separate the data	None of the above		a
61	How do you handle missing or corrupted data in a dataset?	a. Drop missing rows or columns b. Replace missing values with mean/median/mode	c. Assign a unique category to missing values d. All of the above				d

62	What is the purpose of performing cross-validation?		a. To assess the predictive performance of the models	b. To judge how the trained model performs outside the sample on test data	c. Both A and B		c
63	Which of the following is true about Naive Bayes ?		a. Assumes that all the features in a dataset are equally important	b. Assumes that all the features in a dataset are independent	c. Both A and B	d. None of the above option	c
64	Which of the following statements about Naive Bayes is incorrect?		A. Attributes are equally important.	B. Attributes are statistically dependent of one another given the class value.	C. Attributes are statistically independent of one another given the class value.	D. Attributes can be nominal or numeric	b
65	Which of the following is not supervised learning?		PCA	Decision Tree	Naive Bayesian	Linear regression	a
66	How can you avoid overfitting ?	--	By using a lot of data	By using inductive machine learning	By using validation only	None of above	A
67	What are the popular algorithms of Machine Learning?	--	Decision Trees and Neural Networks (back propagation)	Probabilistic networks and Nearest Neighbor	Support vector machines	All	D
68	What is 'Training set'?	--	Training set is used to test the accuracy of the hypotheses generated by the learner.	A set of data is used to discover the potentially predictive relationship.	Both A & B	None of above	B
69	Identify the various approaches for machine learning.	--	Concept Vs Classification Learning	Symbolic Vs Statistical Learning	Inductive Vs Analytical Learning	All above	D
70	what is the function of 'Unsupervised Learning'?	--	Find clusters of the data and find low-dimensional representations of the data	Find interesting directions in data and find novel observations/ database cleaning	Interesting coordinates and correlations	All	D
71	What are the two methods used for the calibration in Supervised Learning?	--	Platt Calibration and Isotonic Regression	Statistics and Informal Retrieval			A
72	_____ can be adopted when it's necessary to categorize a large amount of data with a few complete examples or when there's the need to impose some constraints to a clustering algorithm.	--	Supervised	Semi-supervised	Reinforcement	Clusters	B
73	In reinforcement learning, this feedback is usually called as _____.	--	Overfitting	Overlearning	Reward	None of above	C
74	In the last decade, many researchers started training bigger and bigger models, built with several different layers that's why this approach is called _____.	--	Deep learning	Machine learning	Reinforcement learning	Unsupervised learning	A
75	there's a growing interest in pattern recognition and associative memories whose structure and functioning are similar to what happens in the neocortex. Such an approach also allows simpler algorithms called _____.	--	Regression	Accuracy	Modelfree	Scalable	C
76	_____ showed better performance than other approaches, even without a context-based model	--	Machine learning	Deep learning	Reinforcement learning	Supervised learning	B
77	Common deep learning applications / problems can also be solved using _____	--	Real-time visual object identification	Classic approaches	Automatic labeling	Bio-inspired adaptive systems	B
78	Some people are using the term _____ instead of prediction only to avoid the weird idea that machine learning is a sort of modern magic.	--	Inference	Interference	Accuracy	None of above	A
79	The term _____ can be freely used, but with the same meaning adopted in physics or system theory.	--	Accuracy	Cluster	Regression	Prediction	D
80	If there is only a discrete number of possible outcomes called _____.	--	Modelfree	Categories	Prediction	None of above	B

81	A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college. Which of the following statement is true in following case?	--	Feature F1 is an example of nominal variable.	Feature F1 is an example of ordinal variable.	It doesn't belong to any of the above category.	Both of these	B
82	What would you do in PCA to get the same projection as SVD?	--	Transform data to zero mean	Transform data to zero median	Not possible	None of these	A
83	What is PCA, KPCA and ICA used for?	--	Principal Components Analysis	Kernel based Principal Component Analysis	Independent Component Analysis	All above	D
84	Can a model trained for item based similarity also choose from a given set of items?	--	YES	NO			A
85	What are common feature selection methods in regression task?	--	correlation coefficient	Greedy algorithms	All above	None of these	C
86	The parameter _____ allows specifying the percentage of elements to put into the test/training set	--	test_size	training_size	All above	None of these	C
87	In many classification problems, the target _____ is made up of categorical labels which cannot immediately be processed by any algorithm.	--	random_state	dataset	test_size	All above	B
88	_____ adopts a dictionary-oriented approach, associating to each category label a progressive integer number.	--	LabelEncoder class	LabelBinarizer class	DictVectorizer	FeatureHasher	A
89	_____ is much more difficult because it's necessary to determine a supervised strategy to train a model for each feature and, finally, to predict their value	--	Removing the whole line	Creating sub-model to predict those features	Using an automatic strategy to input them according to the other known values	All above	B
90	How it's possible to use a different placeholder through the parameter _____.	--	regression	classification	random_state	missing_values	D
91	If you need a more powerful scaling feature, with a superior control on outliers and the possibility to select a quantile range, there's also the class _____.	--	RobustScaler	DictVectorizer	LabelBinarizer	FeatureHasher	A
92	scikit-learn also provides a class for per-sample normalization, Normalizer. It can apply _____ to each element of a dataset	--	max, l0 and l1 norms	max, l1 and l2 norms	max, l2 and l3 norms	max, l3 and l4 norms	B
93	There are also many univariate methods that can be used in order to select the best features according to specific criteria based on _____.	--	F-tests and p-values	chi-square	ANOVA	All above	A
94	Which of the following selects only a subset of features belonging to a certain percentile	--	SelectPercentile	FeatureHasher	SelectKBest	All above	A
95	_____ performs a PCA with non-linearly separable data sets.	--	SparsePCA	KernelPCA	SVD	None of the Mentioned	B
96	If two variables are correlated, is it necessary that they have a linear relationship?	--	Yes	No			B
97	Correlated variables can have zero correlation coefficient. True or False?	--	TRUE	FALSE			A
98	Suppose we fit "Lasso Regression" to a data set, which has 100 features (X1,X2...X100). Now, we rescale one of these feature by multiplying with 10 (say that feature is X1), and then refit Lasso regression with the same regularization parameter.Now, which of the following option will be correct?	--	It is more likely for X1 to be excluded from the model	It is more likely for X1 to be included in the model	Can't say	None of these	B
99	If Linear regression model perfectly first i.e., train error is zero, then _____	--	Test error is also always zero	Test error is non zero	Couldn't comment on Test error	Test error is equal to Train error	C
100	Which of the following metrics can be used for evaluating regression models?i) R Squaredii) Adjusted R Squarediii) F Statisticsiv) RMSE / MSE / MAE	--	ii and iv	i and ii	ii, iii and iv	i, ii, iii and iv	D
101	In syntax of linear model lm(formula,data,...), data refers to _____	--	Matrix	Vector	Array	List	B
102	Linear Regression is a supervised machine learning algorithm.	--	TRUE	FALSE			A
103	It is possible to design a Linear regression algorithm using a neural network?	--	TRUE	FALSE			A
104	Which of the following methods do we use to find the best fit line for data in Linear Regression?	--	Least Square Error	Maximum Likelihood	Logarithmic Loss	Both A and B	A
105	Suppose you are training a linear regression model. Now consider these points.1. Overfitting is more likely if we have less data2. Overfitting is more likely when the hypothesis space is small.Which of the above statement(s) are correct?	--	Both are False	1 is False and 2 is True	1 is True and 2 is False	Both are True	C

106	We can also compute the coefficient of linear regression with the help of an analytical method called “Normal Equation”. Which of the following is/are true about “Normal Equation”?1. We don’t have to choose the learning rate2. It becomes slow when number of features is very large3. No need to iterate	--	1 and 2	1 and 3.	2 and 3.	1,2 and 3.	D
107	Which of the following option is true regarding “Regression” and “Correlation”? Note: y is dependent variable and x is independent variable.	--	The relationship is symmetric between x and y in both.	The relationship is not symmetric between x and y in both.	The relationship is not symmetric between x and y in both.	The relationship is symmetric between x and y in case of correlation but in case of regression it is symmetric.	D
108	In a simple linear regression model (One independent variable), If we change the input variable by 1 unit. How much output variable will change?	--	by 1	no change	by intercept	by its slope	D
109	Generally, which of the following method(s) is used for predicting continuous dependent variable?1. Linear Regression2. Logistic Regression	--	1 and 2	only 1	only 2	None of these.	B
110	How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?	--	1	2	3	4	B
111	Suppose you are building a SVM model on data X. The data X can be error prone which means that you should not trust any specific data point too much. Now think that you want to build a SVM model which has quadratic kernel function of polynomial degree 2 that uses Slack variable C as one of its hyper parameter. What would happen when you use very large value of C(C->infinity)?	--	We can still classify data correctly for given setting of hyper parameter C	We can not classify data correctly for given setting of hyper parameter C	Can't Say	None of these	A
112	SVM can solve linear and non-linear problems	--	TRUE	FALSE			A
113	The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points.	--	TRUE	FALSE			A
114	Hyperplanes are boundaries that help classify the data points.	--	usual	decision	parallel		B
115	When the C parameter is set to infinite, which of the following holds true?	--	The optimal hyperplane if exists, will be the one that completely separates the data	The soft-margin classifier will separate the data	None of the above		A
116	SVM is a ----- learning	--	Supervised	Unsupervised	Both	None	A
117	The linear SVM classifier works by drawing a straight line between two classes	--	True	FALSE			A
118	In a real problem, you should check to see if the SVM is separable and then include slack variables if it is not separable.	--	TRUE	FALSE			B
119	Which of the following are real world applications of the SVM?	--	Text and Hypertext Categorization	Image Classification	Clustering of News Articles	All of the above	D
120	The ----- of the hyperplane depends upon the number of features.	--	dimension	classification	reduction		A
121	Hyperplanes are decision boundaries that help classify the data points.	--	TRUE	FALSE			A
122	SVM algorithms use a set of mathematical functions that are defined as the kernel.	--	TRUE	FALSE			A
123	Naive Bayes classifiers are a collection ----- of algorithms	--	Classification	Clustering	Regression	All	A
124	In given image, P(H E) is ----- probability.	bayes.jpg	Posterior	Prior			A
125	Solving a non linear separation problem with a hard margin Kernelized SVM (Gaussian RBF Kernel) might lead to overfitting		True	FALSE			A
126	100 people are at party. Given data gives information about how many wear pink or not, and if a man or not. Imagine a pink wearing guest leaves, was it a man?	man.jpg	TRUE	FALSE			A
127	For the given weather data, Calculate probability of playing	weather data.jpg	0.4	0.64	0.29	0.75	B

128	In SVM, Kernel function is used to map a lower dimensional data into a higher dimensional data.	--	TRUE	FALSE			A
129	In SVR we try to fit the error within a certain threshold.	--	TRUE	FALSE			A
130	When the C parameter is set to infinite, which of the following holds true?	--	The optimal hyperplane if exists, will be the one that completely separates the data	The soft-margin classifier will separate the data	None of the above		A

This sheet is for 1 Mark questions								
S.r No	Question	Image	a	b	c	d	Correct Answer	
	Write down question	img.jpg	Option a	Option b	Option c	Option d	a/b/c/d	
1	In reinforcement learning if feedback is negative one it is defined as ____.		Penalty	Overlearning	Reward	None of above	A	
2	According to ___, it's a key success factor for the survival and evolution of all species.		Claude Shannon's theory	Gini Index	Darwin's theory	None of above	C	
3	How can you avoid overfitting ?		By using a lot of data	By using inductive machine learning	By using validation only	None of above	A	
4	What are the popular algorithms of Machine Learning?		Decision Trees and Neural Networks (back propagation)	Probabilistic networks and Nearest Neighbor	Support vector machines	All	D	
5	What is 'Training set'?		Training set is used to test the accuracy of the hypotheses generated by the learner	A set or data is used to discover the potentially predictive relationship	Both A & B	None of above	B	
6	Common deep learning applications include ____		Image classification, Real-time visual tracking	Autonomous car driving, Logistic	Bioinformatics,	All above	D	
7	what is the function of 'Supervised Learning'?		Classifications, Predict time series, Annotate strings	Speech recognition, Regression	Both A & B	None of above	C	
8	Commons unsupervised applications include		Object segmentation	Similarity detection	Automatic labeling	All above	D	
9	Reinforcement learning is particularly efficient when _____.		the environment is not completely deterministic	it's often very dynamic	it's impossible to have a	All above	D	
10	if there is only a discrete number of possible outcomes (called categories), the process becomes a .		Regression	Classification.	Modelfree	Categories	B	
11	Which of the following are supervised learning applications		Spam detection, Pattern detection, Natural Language Processing	Image classification, Real-time visual tracking	Autonomous car driving, Logistic optimization	Bioinformatics, Speech recognition	A	
12	During the last few years, many _____ algorithms have been applied to deep neural networks to learn the best policy for playing Atari video games and to teach an agent how to associate the right action with an input representing the state.		Logical	Classical	Classification	None of above	D	
13	Which of the following sentence is correct?		Machine learning relates with the study, design and development of the when a statistical model describes random error or noise instead of	Data mining can be defined as the process in which the unstructured	Both A & B	None of the above	C	
14	What is 'Overfitting' in Machine learning?		when a statistical model describes random error or noise instead of	Robots are programed so that they can perform the task	While involving the process of learning	a set of data is used to discover the potentially	A	

15	What is 'Test set'?		Test set is used to test the accuracy of the hypotheses	It is a set of data is used to discover the	Both A & B	None of above	A	
16	_____ is much more difficult because it's necessary to determine a supervised strategy to train a model for each feature and, finally, to predict their value		Removing the whole line	Creating sub-model to predict those features	Using an automatic strategy to input them according to the other known values	All above	B	
17	How it's possible to use a different placeholder through the parameter _____.		regression	classification	random_state	missing_values	D	
18	If you need a more powerful scaling feature, with a superior control on outliers and the possibility to select a quantile range, there's also the class _____.		RobustScaler	DictVectorizer	LabelBinarizer	FeatureHasher	A	
19	scikit-learn also provides a class for per-sample normalization, Normalizer. It can apply _____ to each element of a dataset		max, l0 and l1 norms	max, l1 and l2 norms	max, l2 and l3 norms	max, l3 and l4 norms	B	
20	There are also many univariate methods that can be used in order to select the best features according to specific criteria based on _____.		F-tests and p-values	chi-square	ANOVA	All above	A	
21	Which of the following selects only a subset of features belonging to a certain percentile		SelectPercentile	FeatureHasher	SelectKBest	All above	A	
22	_____ performs a PCA with non-linearly separable data sets.		SparsePCA	KernelPCA	SVD	None of the Mentioned	B	
23	A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college. Which of the following statement is true in following case?		Feature F1 is an example of nominal variable.	Feature F1 is an example of ordinal variable.	It doesn't belong to any of the above category.	Both of these	B	
24	What would you do in PCA to get the same projection as SVD?		Transform data to zero mean	Transform data to zero median	Not possible	None of these	A	
25	What is PCA, KPCA and ICA used for?		Principal Components Analysis	Kernel based Principal Component Analysis	Independent Component Analysis	All above	D	
26	Can a model trained for item based similarity also choose from a given set of items?		YES	NO			A	
27	What are common feature selection methods in regression task?		correlation coefficient	Greedy algorithms	All above	None of these	C	
28	The parameter _____ allows specifying the percentage of elements to put into the test/training set		test_size	training_size	All above	None of these	C	
29	In many classification problems, the target _____ is made up of categorical labels which cannot immediately be processed by any algorithm.		random_state	dataset	test_size	All above	B	
30	_____ adopts a dictionary-oriented approach, associating to each category label a progressive integer number.		LabelEncoder class	LabelBinarizer class	DictVectorizer	FeatureHasher	A	
31	If Linear regression model perfectly first i.e., train error is zero, then _____		a) Test error is also a) Test error is no c) Couldn't co d) Test error is c	a) Test error is also a) Test error is no c) Couldn't co d) Test error is c				
32	Which of the following metrics can be used for evaluating regression models? i) R Squared ii) Adjusted R Squared iii) F Statistics		a) ii and iv	b) i and ii	c) ii, iii and iv	d) i, ii, iii and iv		
33	How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?		a) 1	b) 2	c) 3	d) 4	b	

34	In a simple linear regression model (One independent variable), If we change the input variable by 1 unit. How much output will change?	a) by 1 b) no change c) by intercept d) by its slope				
35	Function used for linear regression in R is _____	a) lm(formula, data)	b) lr(formula, data)	c) lrm(formula, data)	d) regression.a	
36	In syntax of linear model lm(formula,data,...), data refers to _____	a) Matrix b) Vector c) Array d) List		b		
37	In the mathematical Equation of Linear Regression $Y = \beta_1 + \beta_2 X + \epsilon$, (β_1, β_2) refers to _____	a) (X-intercept, Slope) b) (Slope, X-Intercept) c) (Y-Intercept, Slope) d) (slope, Y-Intercept)		c		
38	Linear Regression is a supervised machine learning algorithm.	A) TRUE B) FALSE			a	
39	It is possible to design a Linear regression algorithm using a neural network?	A) TRUE B) FALSE			a	
40	Which of the following methods do we use to find the best fit line for data in Linear Regression?	A) Least Square Error B) Maximum Likelihood	Logarithmic Loss	D) Both A and B	a	
41	Which of the following evaluation metrics can be used to evaluate a model while modeling a continuous output variable?	A) AUC-ROC B) Accuracy	C) Logloss D) Mean-Squared-Error		d	
42	Which of the following is true about Residuals ?	A) Lower is better B) Higher is better	depend on the situation	D) None of these	a	
43	Overfitting is more likely when you have huge amount of data to train?	A) TRUE B) FALSE			b	
44	Which of the following statement is true about outliers in Linear regression?	regression is sensitive to outliers	regression is not sensitive to	D) None of these	a	
45	Suppose you plotted a scatter plot between the residuals and predicted values in linear regression and you found that there is a relationship between them. Which of the following conclusion do you make about this situation?	A) Since there is a relationship means our model is not good B) Since there is a relationship means our model is good	C) Can't say D) None of these		a	
46	Naive Bayes classifiers are a collection ----- of algorithms	Classification	Clustering	Regression	All	a
47	Naive Bayes classifiers is _____ Learning	Supervised	Unsupervised	Both	None	a
48	Features being classified is independent of each other in Naïve Bayes Classifier	False	TRUE			b
49	Features being classified is _____ of each other in Naïve Bayes Classifier	Independent	Dependent	Partial Dependent	None	a
50	Bayes Theorem is given by where 1. $P(H)$ is the probability of hypothesis H being true. 2. $P(E)$ is the probability of the evidence (regardless of the hypothesis). 3. $P(E H)$ is the probability of the evidence given that hypothesis is true. 4. $P(H E)$ is the probability of the hypothesis given that the evidence is there.	bayes.jpg	True FALSE			a
51	In given image, $P(H E)$ is _____ probability.	bayes.jpg	Posterior Prior			a
52	In given image, $P(H)$ is _____ probability.	bayes.jpg	Posterior Prior			b
53	Conditional probability is a measure of the probability of an event given that another event has already occurred.	True	FALSE			a
54	Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event.	True	FALSE			a
55	Bernoulli Naïve Bayes Classifier is _____ distribution	Continuous	Discrete	Binary		c
56	Multinomial Naïve Bayes Classifier is _____ distribution	Continuous	Discrete	Binary		b
57	Gaussian Naïve Bayes Classifier is _____ distribution	Continuous	Discrete	Binary		a
58	Binarize parameter in BernoulliNB scikit sets threshold for binarizing of sample features.	True	FALSE			a
59	Gaussian distribution when plotted, gives a bell shaped curve which is symmetric about the _____ of the feature values.	Mean	Variance	Discrete	Random	a
60	SVMs directly give us the posterior probabilities $P(y = 1 x)$ and $P(y = -1 x)$	True	FALSE			b
61	Any linear combination of the components of a multivariate Gaussian is a univariate Gaussian.	True	FALSE			a
62	Solving a non linear separation problem with a hard margin Kernelized SVM (Gaussian RBF Kernel) might lead to overfitting	True	FALSE			a
63	SVM is a ----- algorithm	Classification	Clustering	Regression	All	a

64	SVM is a ----- learning		Supervised	Unsupervised	Both	None	a	
65	The linear SVM classifier works by drawing a straight line between two classes		True	FALSE			a	
66	Which of the following function provides unsupervised prediction ?	--	cl_forecastB	cl_nowcastC	cl_precastD	None of the Mentioned	D	
67	Which of the following is characteristic of best machine learning method ?	--	fast	accuracy	scalable	All above	D	
68	What are the different Algorithm techniques in Machine Learning?	--	Supervised Learning and Semi-supervised Learning	Unsupervised Learning and Transduction	Both A & B	None of the Mentioned	C	
69	What is the standard approach to supervised learning?	--	split the set of example into the training set and the test	group the set of example into the training set and the test	a set of observed instances tries to induce a general rule	learns programs from data	A	
70	Which of the following is not Machine Learning?	--	Artificial Intelligence	Rule based inference	Both A & B	None of the Mentioned	B	
71	What is Model Selection in Machine Learning?	--	The process of selecting models among different mathematical models, which are used to describe the same data set	when a statistical model describes random error or noise instead of underlying relationship	Find interesting directions in data and find novel observations / database cleaning	All above	A	
72	Which are two techniques of Machine Learning ?	--	Genetic Programming and Inductive Learning	Speech recognition and Regression	Both A & B	None of the Mentioned	A	
73	Even if there are no actual supervisors _____ learning is also based on feedback provided by the environment	--	Supervised	Reinforcement	Unsupervised	None of the above	B	
74	What does learning exactly mean?	--	Robots are programed so that they can perform the task based on data they gather from sensors.	A set of data is used to discover the potentially predictive relationship.	Learning is the ability to change according to external stimuli and rememberin	It is a set of data is used to discover the potentially predictive relationship	C	
75	When it is necessary to allow the model to develop a generalization ability and avoid a common problem called _____ .	--	Overfitting	Overlearning	Classification	Regression	A	
76	Techniques involve the usage of both labeled and unlabeled data is called _____.	--	Supervised	Semi-supervised	Unsupervised	None of the above	B	

77	In reinforcement learning if feedback is negative one it is defined as ____.	--	Penalty	Overlearning	Reward	None of above	A	
78	According to____, it's a key success factor for the survival and evolution of all species.	--	Claude Shannon's theory	Gini Index	Darwin's theory	None of above	C	
79	A supervised scenario is characterized by the concept of a ____.	--	Programmer	Teacher	Author	Farmer	B	
80	overlearning causes due to an excessive ____.	--	Capacity	Regression	Reinforcement	Accuracy	A	
81	Which of the following is an example of a deterministic algorithm?	--	PCA	K-Means	None of the above		A	
82	Which of the following model include a backwards elimination feature selection routine?	--	MCV	MARS	MCRS	All above	B	
83	Can we extract knowledge without apply feature selection	--	YES	NO			A	
84	While using feature selection on the data, is the number of features decreases.	--	NO	YES			B	
85	Which of the following are several models for feature extraction	--	regression	classification	None of the above		C	
86	____ provides some built-in datasets that can be used for testing purposes.	--	scikit-learn	classification	regression	None of the above	A	
87	While using ____ all labels are turned into sequential numbers.	--	LabelEncoder class	LabelBinarizer class	DictVectorizer	FeatureHasher	A	
88	____ produce sparse matrices of real numbers that can be fed into any machine learning model.	--	DictVectorizer	FeatureHasher	Both A & B	None of the Mentioned	C	
89	scikit-learn offers the class ____ , which is responsible for filling the holes using a strategy based on the mean, median, or frequency	--	LabelEncoder	LabelBinarizer	DictVectorizer	Imputer	D	
90	Which of the following scale data by removing elements that don't belong to a given range or by considering a maximum absolute value.	--	MinMaxScaler	MaxAbsScaler	Both A & B	None of the Mentioned	C	
91	scikit-learn also provides a class for per-sample normalization,	--	Normalizer	Imputer	Classifier	All above	A	
92	____ dataset with many features contains information proportional to the independence of all features and their variance.	--	normalized	unnormalized	Both A & B	None of the Mentioned	B	
93	In order to assess how much information is brought by each component, and the correlation among them, a useful tool is the ____.	--	Concuttent matrix	Convergance matrix	Supportive matrix	Covariance matrix	D	
94	The ____ parameter can assume different values which determine how the data matrix is initially processed.	--	run	start	init	stop	C	
95	____ allows exploiting the natural sparsity of data while extracting principal components.	--	SparsePCA	KernelPCA	SVD	init parameter	A	
96	Which of the following evaluation metrics can be used to evaluate a model while modeling a continuous output variable?	--	AUC-ROC	Accuracy	Logloss	Mean-Squared-Error	D	
97	Which of the following is true about Residuals ?	--	Lower is better	Higher is better	A or B depend on the situation	None of these	A	
98	Overfitting is more likely when you have huge amount of data to train?	--	TRUE	FALSE			B	

99	Which of the following statement is true about outliers in Linear regression?	--	Linear regression is sensitive to outliers	Linear regression is not sensitive to outliers	Can't say	None of these	A	
100	Suppose you plotted a scatter plot between the residuals and predicted values in linear regression and you found that there is a relationship between them. Which of the following conclusion do you make about this situation?	--	Since the there is a relationship means our model is not good	Since the there is a relationship means our model is good	Can't say	None of these	A	
101	Let's say, a "Linear regression" model perfectly fits the training data (train error is zero). Now, Which of the following statement is true?	--	You will always have test error zero	You can not have test error zero	None of the above		C	
102	In a linear regression problem, we are using "R-squared" to measure goodness-of-fit. We add a feature in linear regression model and retrain the same model. Which of the following option is true?	--	If R Squared increases, this variable is significant.	If R Squared decreases, this variable is not significant.	Individually R squared cannot tell about variable importance. We can't say anything about it right now.	None of these.	C	
103	Which of the one is true about Heteroskedasticity?	--	Linear Regression with varying error terms	Linear Regression with constant error terms	Linear Regression with zero error terms	None of these	A	
104	Which of the following assumptions do we make while deriving linear regression parameters?1. The true relationship between dependent y and predictor x is linear2. The model errors are statistically independent3. The errors are normally distributed with a 0 mean and constant standard deviation4. The predictor x is non-stochastic and is measured error-free	--	1,2 and 3.	1,3 and 4.	1 and 3.	All of above.	D	
105	To test linear relationship of y(dependent) and x(independent) continuous variables, which of the following plot best suited?	--	Scatter plot	Barchart	Histograms	None of these	A	
106	which of the following step / assumption in regression modeling impacts the trade-off between under-fitting and over-fitting the most.	--	The polynomial degree	Whether we learn the weights by matrix inversion or gradient descent	The use of a constant-term		A	
107	Can we calculate the skewness of variables based on mean and median?	--	TRUE	FALSE			B	
108	Which of the following is true about "Ridge" or "Lasso" regression methods in case of feature selection?	--	Ridge regression uses subset selection of features	Lasso regression uses subset selection of features	Both use subset selection of features	None of above	B	
109	Which of the following statement(s) can be true post adding a variable in a linear regression model?1. R-Squared and Adjusted R-squared both increase2. R-Squared increases and Adjusted R-squared decreases3. R-Squared decreases and Adjusted R-squared decreases4. R-Squared decreases and Adjusted R-squared increases	--	1 and 2	1 and 3	2 and 4	None of the above	A	

110	How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?	--	1	2	Can't Say		B	
111	In given image, P(H) is _____ probability.	bayes.jpg	Posterior	Prior			B	
112	Conditional probability is a measure of the probability of an event given that another event has already occurred.	--	True	FALSE			A	
113	Gaussian distribution when plotted, gives a bell shaped curve which is symmetric about the _____ of the feature values.	--	Mean	Variance	Discrete	Random	A	
114	SVMs directly give us the posterior probabilities $P(y = 1 x)$ and $P(y = -1 x)$	--	True	FALSE			B	
115	SVM is a ----- algorithm	--	Classification	Clustering	Regression	All	A	
116	What is/are true about kernel in SVM?1. Kernel function map low dimensional data to high dimensional space2. It's a similarity function	--	1	2	1 and 2	None of these	C	
117	Suppose you are building a SVM model on data X. The data X can be error prone which means that you should not trust any specific data point too much. Now think that you want to build a SVM model which has quadratic kernel function of polynomial degree 2 that uses Slack variable C as one of it's hyper parameter.What would happen when you use very small C ($C \sim 0$)?	--	Misclassification would happen	Data will be correctly classified	Can't say	None of these	A	
118	The cost parameter in the SVM means:	--	The number of cross-validations to be made	The kernel to be used	The tradeoff between misclassification and simplicity of the model	None of the above	C	
119	Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event.	--	True	FALSE			A	
120	Bernoulli Naïve Bayes Classifier is _____ distribution	--	Continuous	Discrete	Binary		C	
121	If you remove the non-red circled points from the data, the decision boundary will change?	svm.jpg	TRUE	FALSE			B	
122	How do you handle missing or corrupted data in a dataset?	--	a. Drop missing rows or columns	b. Replace missing values with mean/median/mode	c. Assign a unique category to missing values	d. All of the above	D	
123	Binarize parameter in BernoulliNB scikit sets threshold for binarizing of sample features.	--	True	FALSE			A	
124	Which of the following statements about Naive Bayes is incorrect?	--	A. Attributes are equally important.	B. Attributes are statistically dependent of one another given the class value.	C. Attributes are statistically independent of one another given the class value.	D. Attributes can be nominal or numeric	B	

	The SVM's are less effective when:	--	The data is linearly separable	The data is clean and ready to use	The data is noisy and contains overlapping points		C	
125		--						
126	Naive Bayes classifiers is _____ Learning	--	Supervised	Unsupervised	Both	None	A	
127	Features being classified is independent of each other in Naive Bayes Classifier	--	False	TRUE			B	
128	Features being classified is _____ of each other in Naïve Bayes Classifier	--	Independent	Dependent	Partial Dependent	None	A	
129	Bayes Theorem is given by where 1. P(H) is the probability of hypothesis H being true. 2. P(E) is the probability of the evidence(regardless of the hypothesis). 3. P(E H) is the probability of the evidence given that hypothesis is true. 4. P(H E) is the probability of the hypothesis given that the evidence is there.	bayes.jpg	True	FALSE			A	
130	Any linear combination of the components of a multivariate Gaussian is a univariate Gaussian.	--	True	FALSE			A	

Machine Learning MCQs UNIT I

1. What is classification?

- a) when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”.
- b) when the output variable is a real value, such as “dollars” or “weight”.

Ans: A

2. What is regression?

- a) When the output variable is a category, such as “red” or “blue” or “disease” and “no disease”.
- b) When the output variable is a real value, such as “dollars” or “weight”.

Ans: B

3. What is supervised learning?

- a) All data is unlabelled and the algorithms learn to inherent structure from the input data
- b) All data is labelled and the algorithms learn to predict the output from the input data
- c) It is a framework for learning where an agent interacts with an environment and receives a reward for each interaction
- d) Some data is labelled but most of it is unlabelled and a mixture of supervised and unsupervised techniques can be used.

Ans: Solution B

4. What is Unsupervised learning?

- a) All data is unlabelled and the algorithms learn to inherent structure from the input data
- b) All data is labelled and the algorithms learn to predict the output from the input data
- c) It is a framework for learning where an agent interacts with an environment and receives a reward for each interaction
- d) Some data is labelled but most of it is unlabelled and a mixture of supervised and unsupervised techniques can be used.

Ans: Solution A

5. What is Semi-Supervised learning?

- a) All data is unlabelled and the algorithms learn to inherent structure from the input data
- b) All data is labelled and the algorithms learn to predict the output from the input data
- c) It is a framework for learning where an agent interacts with an environment and receives a reward for each interaction
- d) Some data is labelled but most of it is unlabelled and a mixture of supervised and unsupervised techniques can be used.

Ans: Solution D

6. What is Reinforcement learning?

- a) All data is unlabelled and the algorithms learn to inherent structure from the input data
- b) All data is labelled and the algorithms learn to predict the output from the input data
- c) It is a framework for learning where an agent interacts with an environment and receives a reward for each interaction
- d) Some data is labelled but most of it is unlabelled and a mixture of supervised and unsupervised techniques can be used.

Ans: Solution C

7. Sentiment Analysis is an example of:

- a)Regression,
- b)Classification
- c)Clustering
- d)Reinforcement Learning

Options:

- A. 1 Only

- B. 1 and 2
- C. 1 and 3
- D. 1, 2 and 4

Ans : Solution D

8. The process of forming general concept definitions from examples of concepts to be learned.

- a) Deduction
- b) abduction
- c) induction
- d) conjunction

Ans : Solution C

9. Computers are best at learning

- a) facts.
- b) concepts.
- c) procedures.
- d) principles.

Ans : Solution A

10. Data used to build a data mining model.

- a) validation data
- b) training data
- c) test data
- d) hidden data

Ans : Solution B

11. Supervised learning and unsupervised clustering both require at least one

- a) hidden attribute.
- b) output attribute.
- c) input attribute.
- d) categorical attribute.

Ans : Solution A

12. Supervised learning differs from unsupervised clustering in that supervised learning requires

- a) at least one input attribute.
- b) input attributes to be categorical.
- c) at least one output attribute.
- d) output attributes to be categorical.

Ans : Solution B

13. A regression model in which more than one independent variable is used to predict the

dependent variable is called

- a) a simple linear regression model
- b) a multiple regression models
- c) an independent model
- d) none of the above

Ans : Solution C

14. A term used to describe the case when the independent variables in a multiple regression model

are correlated is

- a) Regression
- b) correlation
- c) multicollinearity
- d) none of the above

Ans : Solution C

15. A multiple regression model has the form: $y = 2 + 3x_1 + 4x_2$. As x_1 increases by 1 unit (holding x_2 constant), y will

- a) increase by 3 units
- b) decrease by 3 units
- c) increase by 4 units
- d) decrease by 4 units

Ans : Solution C

16. A multiple regression model has

- a) only one independent variable
- b) more than one dependent variable
- c) more than one independent variable
- d) none of the above

Ans : Solution B

17. A measure of goodness of fit for the estimated regression equation is the

- a) multiple coefficient of determination
- b) mean square due to error
- c) mean square due to regression
- d) none of the above

Ans : Solution C

18. The adjusted multiple coefficient of determination accounts for

- a) the number of dependent variables in the model
- b) the number of independent variables in the model
- c) unusually large predictors
- d) none of the above

Ans : Solution D

19. The multiple coefficient of determination is computed by

- a) dividing SSR by SST
- b) dividing SST by SSR
- c) dividing SST by SSE
- d) none of the above

Ans : Solution C

20. For a multiple regression model, $SST = 200$ and $SSE = 50$. The multiple coefficient of

determination is

- a) 0.25
- b) 4.00
- c) 0.75
- d) none of the above

Ans : Solution B

21. A nearest neighbor approach is best used

- a) with large-sized datasets.
- b) when irrelevant attributes have been removed from the data.
- c) when a generalized model of the data is desirable.
- d) when an explanation of what has been found is of primary importance.

Ans : Solution B

22. Another name for an output attribute.

- a) predictive variable
- b) independent variable
- c) estimated variable
- d) dependent variable

Ans : Solution B

23. Classification problems are distinguished from estimation problems in that

- a) classification problems require the output attribute to be numeric.
- b) classification problems require the output attribute to be categorical.
- c) classification problems do not allow an output attribute.
- d) classification problems are designed to predict future outcome.

Ans : Solution C

24. Which statement is true about prediction problems?

- a) The output attribute must be categorical.
- b) The output attribute must be numeric.
- c) The resultant model is designed to determine future outcomes.
- d) The resultant model is designed to classify current behavior.

Ans : Solution D

25. Which statement about outliers is true?

- a) Outliers should be identified and removed from a dataset.
- b) Outliers should be part of the training dataset but should not be present in the test data.
- c) Outliers should be part of the test dataset but should not be present in the training data.
- d) The nature of the problem determines how outliers are used.

Ans : Solution D

26. Which statement is true about neural network and linear regression models?

- a) Both models require input attributes to be numeric.
- b) Both models require numeric attributes to range between 0 and 1.
- c) The output of both models is a categorical attribute value.
- d) Both techniques build models whose output is determined by a linear sum of weighted input attribute values.

Ans : Solution A

27. Which of the following is a common use of unsupervised clustering?

- a) detect outliers
- b) determine a best set of input attributes for supervised learning
- c) evaluate the likely performance of a supervised learner model
- d) determine if meaningful relationships can be found in a dataset

Ans : Solution A

28. The average positive difference between computed and desired outcome values.

- a) root mean squared error
- b) mean squared error
- c) mean absolute error
- d) mean positive error

Ans : Solution D

29. Selecting data so as to assure that each class is properly represented in both the training and test set.

- a) cross validation
- b) stratification
- c) verification
- d) bootstrapping

Ans : Solution B

30. The standard error is defined as the square root of this computation.

- a) The sample variance divided by the total number of sample instances.
- b) The population variance divided by the total number of sample instances.
- c) The sample variance divided by the sample mean.
- d) The population variance divided by the sample mean.

Ans : Solution A

31. Data used to optimize the parameter settings of a supervised learner model.

- a) Training
- b) Test
- c) Verification
- d) Validation

Ans : Solution D

32. Bootstrapping allows us to

- a) choose the same training instance several times.
- b) choose the same test set instance several times.
- c) build models with alternative subsets of the training data several times.
- d) test a model with alternative subsets of the test data several times.

Ans : Solution A

33. The correlation between the number of years an employee has worked for a company and the salary of the employee is 0.75. What can be said about employee salary and years worked?

- a) There is no relationship between salary and years worked.
- b) Individuals that have worked for the company the longest have higher salaries.
- c) Individuals that have worked for the company the longest have lower salaries.
- d) The majority of employees have been with the company a long time.
- e) The majority of employees have been with the company a short period of time.

Ans : Solution B

34. The correlation coefficient for two real-valued attributes is -0.85 . What does this value tell you?

- a) The attributes are not linearly related.
- b) As the value of one attribute increases the value of the second attribute also increases.
- c) As the value of one attribute decreases the value of the second attribute increases.
- d) The attributes show a curvilinear relationship.

Ans : Solution C

35. The average squared difference between classifier predicted output and actual output.

- a) mean squared error
- b) root mean squared error
- c) mean absolute error
- d) mean relative error

Ans : Solution A

36. Simple regression assumes a _____ relationship between the input attribute and output attribute.

- a) Linear
- b) Quadratic
- c) reciprocal
- d) inverse

Ans : Solution A

37. Regression trees are often used to model _____ data.

- a) Linear
- b) Nonlinear

- c) Categorical
- d) Symmetrical

Ans : Solution B

38. The leaf nodes of a model tree are
- a) averages of numeric output attribute values.
 - b) nonlinear regression equations.
 - c) linear regression equations.
 - d) sums of numeric output attribute values.

Ans : Solution C

39. Logistic regression is a _____ regression technique that is used to model data having a
_____ outcome.

- a) linear, numeric
- b) linear, binary
- c) nonlinear, numeric
- d) nonlinear, binary

Ans : Solution D

40. This technique associates a conditional probability value with each data instance.
- a) linear regression
 - b) logistic regression
 - c) simple regression
 - d) multiple linear regression

Ans : Solution B

41. This supervised learning technique can process both numeric and categorical input attributes.
- a) linear regression
 - b) Bayes classifier
 - c) logistic regression
 - d) backpropagation learning

Ans : Solution A

42. With Bayes classifier, missing data items are
- a) treated as equal compares.
 - b) treated as unequal compares.

- c) replaced with a default value.
- d) ignored.

Ans : Solution B

43. This clustering algorithm merges and splits nodes to help modify nonoptimal partitions.

- a) agglomerative clustering
- b) expectation maximization
- c) conceptual clustering
- d) K-Means clustering

Ans : Solution D

44. This clustering algorithm initially assumes that each data instance represents a single cluster.

- a) agglomerative clustering
- b) conceptual clustering
- c) K-Means clustering
- d) expectation maximization

Ans : Solution C

45. This unsupervised clustering algorithm terminates when mean values computed for the current iteration of the algorithm are identical to the computed mean values for the previous iteration.

- a) agglomerative clustering
- b) conceptual clustering
- c) K-Means clustering
- d) expectation maximization

Ans : Solution C

46. Machine learning techniques differ from statistical techniques in that machine learning methods

- a) typically assume an underlying distribution for the data.
- b) are better able to deal with missing and noisy data.
- c) are not able to explain their behavior.

d) have trouble with large-sized datasets.

Ans : Solution B

Machine Learning MCQs UNIT -II

1. True- False: Over fitting is more likely when you have huge amount of data to train?

A) TRUE

B) FALSE

Ans Solution: (B)

With a small training dataset, it's easier to find a hypothesis to fit the training data exactly i.e. over fitting.

2. What is pca.components_ in Sklearn?

A) Set of all eigen vectors for the projection space

B) Matrix of principal components

C) Result of the multiplication matrix

D) None of the above options

Ans A

3. Which of the following techniques would perform better for reducing dimensions of a data set?

A. Removing columns which have too many missing values

B. Removing columns which have high variance in data

C. Removing columns with dissimilar data trends

D. None of these

Ans Solution: (A) If a columns have too many missing values, (say 99%) then we can remove such columns.

4. It is not necessary to have a target variable for applying dimensionality reduction algorithms.

A. TRUE

B. FALSE

Ans Solution: (A)

LDA is an example of supervised dimensionality reduction algorithm

5. PCA can be used for projecting and visualizing data in lower dimensions.

- A. TRUE
- B. FALSE

Ans Solution: (A)

Sometimes it is very useful to plot the data in lower dimensions. We can take the first 2 principal components and then visualize the data using scatter plot.

6. The most popularly used dimensionality reduction algorithm is Principal Component Analysis (PCA). Which of the following is/are true about PCA?

- 1. PCA is an unsupervised method
 - 2. It searches for the directions that data have the largest variance
 - 3. Maximum number of principal components \leq number of features
 - 4. All principal components are orthogonal to each other
- A. 1 and 2
 - B. 1 and 3
 - C. 2 and 3
 - D. All of the above

Ans D

7. PCA works better if there is?

- 1. A linear structure in the data
 - 2. If the data lies on a curved surface and not on a flat surface
 - 3. If variables are scaled in the same unit
- A. 1 and 2
 - B. 2 and 3
 - C. 1 and 3
 - D. 1,2 and 3

Ans Solution: (C)

8. What happens when you get features in lower dimensions using PCA?

- 1. The features will still have interpretability
 - 2. The features will lose interpretability
 - 3. The features must carry all information present in data
 - 4. The features may not carry all information present in data
- A. 1 and 3
 - B. 1 and 4
 - C. 2 and 3
 - D. 2 and 4

Ans Solution: (D)

When you get the features in lower dimensions then you will lose some information of data most of the times and you won't be able to interpret the lower dimension data.

9. Which of the following option(s) is / are true?
1. You need to initialize parameters in PCA
 2. You don't need to initialize parameters in PCA
 3. PCA can be trapped into local minima problem
 4. PCA can't be trapped into local minima problem
- A. 1 and 3
B. 1 and 4
C. 2 and 3
D. 2 and 4

Ans Solution: (D)

PCA is a deterministic algorithm which doesn't have parameters to initialize and it doesn't have local minima problem like most of the machine learning algorithms has.

10. What is of the following statement is true about t-SNE in comparison to PCA?
- A. When the data is huge (in size), t-SNE may fail to produce better results.
 - B. T-SNE always produces better result regardless of the size of the data
 - C. PCA always performs better than t-SNE for smaller size data.
 - D. None of these

Ans Solution: (A)

Option A is correct

11. [True or False] PCA can be used for projecting and visualizing data in lower dimensions.

- A. TRUE
B. FALSE

Solution: (A)

Sometimes it is very useful to plot the data in lower dimensions. We can take the first 2 principal components and then visualize the data using scatter plot.

12. A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college.

- 1) Which of the following statement is true in following case?
A) Feature F1 is an example of nominal variable.
B) Feature F1 is an example of ordinal variable.

C) It doesn't belong to any of the above category.

D) Both of these

Solution: (B)

Ordinal variables are the variables which has some order in their categories. For example, grade A should be consider as high grade than grade B.

13. Which of the following is an example of a deterministic algorithm?

A) PCA

B) K-Means

C) None of the above

Solution: (A)

A deterministic algorithm is that in which output does not change on different runs. PCA would give the same result if we run again, but not k-means

Machine Learning MCQs UNIT –III

1. Which of the following methods do we use to best fit the data in Logistic Regression?

A) Least Square Error

B) Maximum Likelihood

C) Jaccard distance

D) Both A and B

Ans Solution: B

2. Choose which of the following options is true regarding One-Vs-All method in Logistic

Regression.

A) We need to fit n models in n-class classification problem

B) We need to fit n-1 models to classify into n classes

C) We need to fit only 1 model to classify into n classes

D) None of these

Ans Solution: A

3. Suppose, You applied a Logistic Regression model on a given data and got a training accuracy X and testing accuracy Y. Now, you want to add a few new features in the same data. Select the option(s) which is/are correct in such a case.

Note: Consider remaining parameters are same.

- A) Training accuracy increases
- B) Training accuracy increases or remains the same
- C) Testing accuracy decreases
- D) Testing accuracy increases or remains the same

Ans Solution: A and D

Adding more features to model will increase the training accuracy because model has to consider more data to fit the logistic regression. But testing accuracy increases if feature is found to be significant

4. Which of the following algorithms do we use for Variable Selection?

- A) LASSO
- B) Ridge
- C) Both
- D) None of these

Ans Solution: A

In case of lasso we apply a absolute penalty, after increasing the penalty in lassosome of the coefficient of variables may become zero

5. Which of the following statement is true about outliers in Linear regression?

- A) Linear regression is sensitive to outliers
- B) Linear regression is not sensitive to outliers
- C) Can't say
- D) None of these

Ans Solution: (A)

The slope of the regression line will change due to outliers in most of the cases. So Linear Regression is sensitive to outliers.

6. Which of the following methods do we use to find the best fit line for data in Linear Regression?

- A) Least Square Error
- B) Maximum Likelihood
- C) Logarithmic Loss
- D) Both A and B

Ans Solution: (A)

In linear regression, we try to minimize the least square errors of the model to identify the line of best fit.

7. Which of the following is true about Residuals?

- A) Lower is better

- B) Higher is better
- C) A or B depend on the situation
- D) None of these

Ans Solution: (A)

Residuals refer to the error values of the model. Therefore lower residuals are desired.

8. Suppose you plotted a scatter plot between the residuals and predicted values in linear regression and you found that there is a relationship between them. Which of the following conclusion do you make about this situation?

- A) Since the there is a relationship means our model is not good
- B) Since the there is a relationship means our model is good
- C) Can't say
- D) None of these

Ans Solution: (A)

There should not be any relationship between predicted values and residuals. If there exists any relationship between them, it means that the model has not perfectly captured the information in the data.

9. Suppose you have fitted a complex regression model on a dataset. Now, you are using Ridge regression with penalty x .

Choose the option which describes bias in best manner.

- A) In case of very large x ; bias is low
- B) In case of very large x ; bias is high
- C) We can't say about bias
- D) None of these

Ans Solution: (B)

If the penalty is very large it means model is less complex, therefore the bias would be high.

10. Which of the following option is true?

- A) Linear Regression errors values has to be normally distributed but in case of Logistic Regression it is not the case
- B) Logistic Regression errors values has to be normally distributed but in case of Linear Regression it is not the case
- C) Both Linear Regression and Logistic Regression error values have to be normally distributed
- D) Both Linear Regression and Logistic Regression error values have not to be normally distributed

Ans Solution: A

11. Suppose you have trained a logistic regression classifier and it outputs a new example x with

- a prediction $h_0(x) = 0.2$. This means Our estimate for $P(y=1 | x)$
- Our estimate for $P(y=0 | x)$
- Our estimate for $P(y=1 | x)$
- Our estimate for $P(y=0 | x)$

Ans Solution: B

12. True-False: Linear Regression is a supervised machine learning algorithm.

- A) TRUE
- B) FALSE

Solution: (A)

Yes, Linear regression is a supervised learning algorithm because it uses true labels for training. Supervised learning algorithm should have input variable (x) and an output variable (Y) for each example

13. True-False: Linear Regression is mainly used for Regression.

- A) TRUE
- B) FALSE

Solution: (A)

Linear Regression has dependent variables that have continuous values.

14. True-False: It is possible to design a Linear regression algorithm using a neural network?

- A) TRUE

B) FALSE

Solution: (A)

True. A Neural network can be used as a universal approximator, so it can definitely implement a linear regression algorithm.

15. Which of the following methods do we use to find the best fit line for data in Linear Regression?

- A) Least Square Error
- B) Maximum Likelihood
- C) Logarithmic Loss
- D) Both A and B

Solution: (A)

In linear regression, we try to minimize the least square errors of the model to identify the line of best fit.

16. Which of the following evaluation metrics can be used to evaluate a model while modeling a continuous output variable?

- A) AUC-ROC
- B) Accuracy
- C) Logloss
- D) Mean-Squared-Error

Solution: (D)

Since linear regression gives output as continuous values, so in such case we use mean squared error metric to evaluate the model performance. Remaining options are used in case of a classification problem.

True-False: Lasso Regularization can be used for variable selection in Linear Regression.

- A) TRUE
- B) FALSE

Solution: (A)

True, In case of lasso regression we apply absolute penalty which makes some of the coefficients zero.

17. Which of the following is true about Residuals ?

- A) Lower is better
- B) Higher is better
- C) A or B depend on the situation
- D) None of these

Solution: (A)

Residuals refer to the error values of the model. Therefore lower residuals are desired.

18. Suppose that we have N independent variables (X_1, X_2, \dots, X_n) and dependent variable is Y. Now Imagine that you are applying linear regression by fitting the best fit line using least square error on this data. You found that correlation coefficient for one of it's variable (Say X_1) with Y is -0.95.

Which of the following is true for X_1 ?

- A) Relation between the X_1 and Y is weak
- B) Relation between the X_1 and Y is strong
- C) Relation between the X_1 and Y is neutral
- D) Correlation can't judge the relationship

Solution: (B)

The absolute value of the correlation coefficient denotes the strength of the relationship. Since absolute correlation is very high it means that the relationship is strong between X_1 and Y.

19. Looking at above two characteristics, which of the following option is the correct for Pearson correlation between V_1 and V_2 ? If you are given the two variables V_1 and V_2 and they are following below two characteristics.

- 1. If V_1 increases then V_2 also increases
- 2. If V_1 decreases then V_2 behavior is unknown

- A) Pearson correlation will be close to 1
- B) Pearson correlation will be close to -1
- C) Pearson correlation will be close to 0
- D) None of these

Solution: (D)

We cannot comment on the correlation coefficient by using only statement 1. We need to consider the both of these two statements. Consider V1 as x and V2 as $|x|$. The correlation coefficient would not be close to 1 in such a case.

21. Suppose Pearson correlation between V1 and V2 is zero. In such case, is it right to conclude that V1 and V2 do not have any relation between them?

- A) TRUE
- B) FALSE

Solution: (B)

Pearson correlation coefficient between 2 variables might be zero even when they have a relationship between them. If the correlation coefficient is zero, it just means that they don't move together. We can take examples like $y=|x|$ or $y=x^2$.

22. True- False: Overfitting is more likely when you have huge amount of data to train?

- A) TRUE
- B) FALSE

Solution: (B)

With a small training dataset, it's easier to find a hypothesis to fit the training data exactly i.e. overfitting.

23. We can also compute the coefficient of linear regression with the help of an analytical method called "Normal Equation". Which of the following is/are true about Normal Equation?

- 1. We don't have to choose the learning rate
 - 2. It becomes slow when number of features is very large
 - 3. There is no need to iterate
- A) 1 and 2
 - B) 1 and 3

C) 2 and 3

D) 1,2 and 3

Solution: (D)

Instead of gradient descent, Normal Equation can also be used to find coefficients.

Question Context 24-26:

Suppose you have fitted a complex regression model on a dataset. Now, you are using Ridge regression with penalty x.

24. Choose the option which describes bias in best manner.

- A) In case of very large x; bias is low
- B) In case of very large x; bias is high
- C) We can't say about bias
- D) None of these

Solution: (B)

If the penalty is very large it means model is less complex, therefore the bias would be high.

25. What will happen when you apply very large penalty?

- A) Some of the coefficient will become absolute zero
- B) Some of the coefficient will approach zero but not absolute zero
- C) Both A and B depending on the situation
- D) None of these

Solution: (B)

In lasso some of the coefficient value become zero, but in case of Ridge, the coefficients become close to zero but not zero.

26. What will happen when you apply very large penalty in case of Lasso?

- A) Some of the coefficient will become zero
- B) Some of the coefficient will be approaching to zero but not absolute zero
- C) Both A and B depending on the situation
- D) None of these

Solution: (A)

As already discussed, lasso applies absolute penalty, so some of the coefficients will become zero.

27. Which of the following statement is true about outliers in Linear regression?

- A) Linear regression is sensitive to outliers
- B) Linear regression is not sensitive to outliers
- C) Can't say
- D) None of these

Solution: (A)

The slope of the regression line will change due to outliers in most of the cases. So Linear Regression is sensitive to outliers.

28. Suppose you plotted a scatter plot between the residuals and predicted values in linear regression and you found that there is a relationship between them. Which of the following conclusion do you make about this situation?

- A) Since the there is a relationship means our model is not good
- B) Since the there is a relationship means our model is good
- C) Can't say
- D) None of these

Solution: (A)

There should not be any relationship between predicted values and residuals. If there exists any relationship between them, it means that the model has not perfectly captured the information in the data.

Question Context 29-31:

Suppose that you have a dataset D1 and you design a linear regression model of degree 3

polynomial and you found that the training and testing error is "0" or in another terms it perfectly fits the data.

29. What will happen when you fit degree 4 polynomial in linear regression?

- A) There are high chances that degree 4 polynomial will over fit the data
- B) There are high chances that degree 4 polynomial will under fit the data
- C) Can't say
- D) None of these

Solution: (A)

Since is more degree 4 will be more complex(overfit the data) than the degree 3 model so it will again perfectly fit the data. In such case training error will be zero but test error may not be zero.

30. What will happen when you fit degree 2 polynomial in linear regression?

- A) It is high chances that degree 2 polynomial will over fit the data
- B) It is high chances that degree 2 polynomial will under fit the data
- C) Can't say
- D) None of these

Solution: (B)

If a degree 3 polynomial fits the data perfectly, it's highly likely that a simpler model(degree 2 polynomial) might under fit the data.

31. In terms of bias and variance. Which of the following is true when you fit degree 2 polynomial?

- A) Bias will be high, variance will be high
- B) Bias will be low, variance will be high
- C) Bias will be high, variance will be low
- D) Bias will be low, variance will be low

Solution: (C)

Since a degree 2 polynomial will be less complex as compared to degree 3, the bias will be high and variance will be low

Question Context 32-33:

We have been given a dataset with n records in which we have input attribute as x and output attribute as y. Suppose we use a linear regression method to model this data. To test our linear regressor, we split the data in training set and test set randomly.

32. Now we increase the training set size gradually. As the training set size increases, what do you expect will happen with the mean training error?

- A) Increase

- B) Decrease
- C) Remain constant
- D) Can't Say

Solution: (D)

Training error may increase or decrease depending on the values that are used to fit the model. If the values used to train contain more outliers gradually, then the error might just increase.

33. What do you expect will happen with bias and variance as you increase the size of training data?

- A) Bias increases and Variance increases
- B) Bias decreases and Variance increases
- C) Bias decreases and Variance decreases
- D) Bias increases and Variance decreases
- E) Can't Say False

Solution: (D)

As we increase the size of the training data, the bias would increase while the variance would decrease.

Question Context 34:

Consider the following data where one input(X) and one output(Y) is given

34. What would be the root mean square training error for this data if you run a Linear Regression model of the form ($Y = A_0 + A_1X$)?

- A) Less than 0
- B) Greater than zero
- C) Equal to 0
- D) None of these

Solution: (C)

We can perfectly fit the line on the following data so mean error will be zero.

Question Context 35-36:

Suppose you have been given the following scenario for training and validation error for Linear Regression.

35. Which of the following scenario would give you the right hyper parameter?

- A) 1
- B) 2
- C) 3
- D) 4

Solution: (B)

Option B would be the better option because it leads to less training as well as validation error.

36. Suppose you got the tuned hyper parameters from the previous question. Now, Imagine you want to add a variable in variable space such that this added feature is important. Which of the following thing would you observe in such case?

- A) Training Error will decrease and Validation error will increase
- B) Training Error will increase and Validation error will increase
- C) Training Error will increase and Validation error will decrease
- D) Training Error will decrease and Validation error will decrease
- E) None of the above

Solution: (D)

If the added feature is important, the training and validation error would decrease.

Question Context 37-38:

Suppose, you got a situation where you find that your linear regression model is under fitting
the data.

37. In such situation which of the following options would you consider?

- 1. I will add more variables
 - 2. I will start introducing polynomial degree variables
 - 3. I will remove some variables
- A) 1 and 2
 - B) 2 and 3
 - C) 1 and 3
 - D) 1, 2 and 3

Solution: (A)

In case of under fitting, you need to induce more variables in variable space or you can add

some polynomial degree variables to make the model more complex to be able to fit the data better.

38. Now situation is same as written in previous question (under fitting). Which of

following regularization algorithm would you prefer?

- A) L1
- B) L2
- C) Any
- D) None of these

Solution: (D)

I won't use any regularization methods because regularization is used in case of overfitting.

39. True-False: Is Logistic regression a supervised machine learning algorithm?

- A) TRUE
- B) FALSE

Solution: A

True, Logistic regression is a supervised learning algorithm because it uses true labels for training. Supervised learning algorithm should have input variables (x) and a target variable (Y) when you train the model .

40. True-False: Is Logistic regression mainly used for Regression?

- A) TRUE
- B) FALSE

Solution: B

Logistic regression is a classification algorithm, don't confuse with the name regression.

41. True-False: Is it possible to design a logistic regression algorithm using a Neural Network Algorithm?

- A) TRUE
- B) FALSE

Solution: A

True, Neural network is a universal approximator so it can implement linear regression algorithm.

42. True-False: Is it possible to apply a logistic regression algorithm on a 3-class Classification

problem?

- A) TRUE
- B) FALSE

Solution: A

Yes, we can apply logistic regression on 3 classification problem, We can use One Vsall method for 3 class classification in logistic regression.

43. Which of the following methods do we use to best fit the data in Logistic Regression?

- A) Least Square Error
- B) Maximum Likelihood
- C) Jaccard distance
- D) Both A and B

Solution: B

Logistic regression uses maximum likely hood estimate for training a logistic regression.

44. Which of the following evaluation metrics can not be applied in case of logistic regression output to compare with target?

- A) AUC-ROC
- B) Accuracy
- C) Logloss
- D) Mean-Squared-Error

Solution: D

Since, Logistic Regression is a classification algorithm so it's output can not be realtime value so mean squared error can not use for evaluating it

45. One of the very good methods to analyze the performance of Logistic Regression is AIC,

which is similar to R-Squared in Linear Regression. Which of the following is true about AIC?

- A) We prefer a model with minimum AIC value
- B) We prefer a model with maximum AIC value
- C) Both but depend on the situation
- D) None of these

Solution: A

We select the best model in logistic regression which can least AIC.

46. [True-False] Standardisation of features is required before training a Logistic Regression.

- A) TRUE

B) FALSE

Solution: B

Standardization isn't required for logistic regression. The main goal of standardizing features is to help convergence of the technique used for optimization.

47. Which of the following algorithms do we use for Variable Selection?

- A) LASSO
- B) Ridge
- C) Both
- D) None of these

Solution: A

In case of lasso we apply a absolute penalty, after increasing the penalty in lassosome of the coefficient of variables may become zero

Context: 48-49

Consider a following model for logistic regression: $P(y=1|x, w) = g(w_0 + w_1x)$ where $g(z)$ is the logistic function. In the above equation the $P(y=1|x; w)$, viewed as a function of x , that we can get by changing the parameters w .

48 What would be the range of p in such case?

- A) $(0, \infty)$
- B) $(-\infty, 0)$
- C) $(0, 1)$
- D) $(-\infty, \infty)$

Solution: C

For values of x in the range of real number from $-\infty$ to $+\infty$ Logistic function will give the output between $(0,1)$

49 In above question what do you think which function would make p between $(0,1)$?

- A) logistic function
- B) Log likelihood function
- C) Mixture of both
- D) None of them

Solution: A

Explanation is same as question number 10

50. Suppose you have been given a fair coin and you want to find out the odds of getting heads. Which of the following option is true for such a case?

- A) odds will be 0
- B) odds will be 0.5

- C) odds will be 1
- D) None of these

Solution: C

Odds are defined as the ratio of the probability of success and the probability of failure. So in case of fair coin probability of success is $1/2$ and the probability of failure is $1/2$ so odd would be 1

51. The logit function(given as $l(x)$) is the log of odds function. What could be the range of logit function in the domain $x=[0,1]$?

- A) $(-\infty, \infty)$
- B) $(0,1)$
- C) $(0, \infty)$
- D) $(-\infty, 0)$

Solution: A

For our purposes, the odds function has the advantage of transforming the probability function, which has values from 0 to 1, into an equivalent function with values between 0 and ∞ . When we take the natural log of the odds function, we get a range of values from $-\infty$ to ∞ .

52. Which of the following option is true?

- A) Linear Regression errors values has to be normally distributed but in case of Logistic Regression it is
not the case
- B) Logistic Regression errors values has to be normally distributed but in case of Linear Regression it is
not the case
- C) Both Linear Regression and Logistic Regression error values have to be normally distributed
- D) Both Linear Regression and Logistic Regression error values have not to be normally distributed

Solution:A

53. Which of the following is true regarding the logistic function for any value "x"?

Note:

Logistic(x): is a logistic function of any number " x "

Logit(x): is a logit function of any number " x " Logit_inv(x):

is a inverse logit function of any number " x "

- A) $\text{Logistic}(x) = \text{Logit}(x)$
- B) $\text{Logistic}(x) = \text{Logit_inv}(x)$
- C) $\text{Logit_inv}(x) = \text{Logit}(x)$
- D) None of these

Solution: B

54. How will the bias change on using high(infinite) regularisation?

Suppose you have given the two scatter plot "a" and "b" for two classes(blue for positive and red for negative class). In scatter plot "a", you correctly classified all data points using logistic regression (black line is a decision boundary).

- A) Bias will be high
- B) Bias will be low
- C) Can't say
- D) None of these

Solution: A

Model will become very simple so bias will be very high.

55. Suppose, You applied a Logistic Regression model on a given data and got a training accuracy X and testing accuracy Y. Now, you want to add a few new features in the same data. Select the option(s) which is/are correct in such a case.

Note: Consider remaining parameters are same.

- A) Training accuracy increases
- B) Training accuracy increases or remains the same
- C) Testing accuracy decreases
- D) Testing accuracy increases or remains the same

Solution: A and D

Adding more features to model will increase the training accuracy because model has to consider more data to fit the logistic regression. But testing accuracy increases if feature is found to be significant

56. Choose which of the following options is true regarding One-Vs-All method in

Logistic Regression.

- A) We need to fit n models in n-class classification problem
- B) We need to fit n-1 models to classify into n classes
- C) We need to fit only 1 model to classify into n classes
- D) None of these

Solution: A

If there are n classes, then n separate logistic regression has to fit, where the probability of each category is predicted over the rest of the categories combined.

57. Below are two different logistic models with different values for β_0 and β_1

Which of the following statement(s) is true about β_0 and β_1 values of two logistics models (Green, Black)?

Note: consider $Y = \beta_0 + \beta_1 * X$. Here, β_0 is intercept and β_1 is coefficient.

- A) β_1 for Green is greater than Black
- B) β_1 for Green is lower than Black
- C) β_1 for both models is same
- D) Can't Say

Solution: B

β_0 and β_1 : $\beta_0 = 0, \beta_1 = 1$ is in X1 color(black) and $\beta_0 = 0, \beta_1 = -1$ is in X4 color (green)

Context 58-60 Below are the three scatter plot(A,B,C left to right) and hand drawn decision boundaries for logistic regression.

58. Which of the following above figure shows that the decision boundary is overfitting the training data?

- A) A
- B) B
- C) C
- D) None of these

Solution: C

Since in figure 3, Decision boundary is not smooth that means it will over-fitting the data.

59. What do you conclude after seeing this visualization?

- 1. The training error in first plot is maximum as compare to second and third plot.
 - 2. The best model for this regression problem is the last (third) plot because it has minimum training error (zero).
 - 3. The second model is more robust than first and third because it will perform best on unseen data.
 - 4. The third model is overfitting more as compare to first and second.
 - 5. All will perform same because we have not seen the testing data.
- A) 1 and 3
 - B) 1 and 3
 - C) 1, 3 and 4
 - D) 5

Solution: C

The trend in the graphs looks like a quadratic trend over independent variable X. A higher degree(Right graph) polynomial might have a very high accuracy on the train population but is expected to fail badly on test dataset. But if you see in left graph we will have training error maximum because it underfits the training data

60. Suppose, above decision boundaries were generated for the different value of regularization. Which of the above decision boundary shows the maximum regularization?

- A) A
- B) B
- C) C
- D) All have equal regularization

Solution: A

Since, more regularization means more penalty means less complex decision boundary that shows in first figure A.

61. What would do if you want to train logistic regression on same data that will take less time as well as give the comparatively similar accuracy(may not be same)?

Suppose you are using a Logistic Regression model on a huge dataset. One of the problem you may face on such huge data is that Logistic regression will take very long time to train.

- A) Decrease the learning rate and decrease the number of iteration
- B) Decrease the learning rate and increase the number of iteration
- C) Increase the learning rate and increase the number of iteration
- D) Increase the learning rate and decrease the number of iteration

Solution: D

If you decrease the number of iteration while training it will take less time for surely but will not give the same accuracy for getting the similar accuracy but not exact you need to increase the learning rate.

62. Which of the following image is showing the cost function for $y = 1$.

Following is the loss function in logistic regression(Y-axis loss function and x axis log probability) for two class classification problem.

Note: Y is the target class

- A) A
- B) B
- C) Both
- D) None of these

Solution: A

A is the true answer as loss function decreases as the log probability increases

63. Suppose, Following graph is a cost function for logistic regression.

Now, How many local minimas are present in the graph?

- A) 1
- B) 2
- C) 3
- D) 4

Solution: C

There are three local minima present in the graph

64. Can a Logistic Regression classifier do a perfect classification on the below data?

Note: You can use only X1 and X2 variables where X1 and X2 can take only two binary values(0,1).

- A) TRUE
- B) FALSE
- C) Can't say
- D) None of these

Solution: B

No, logistic regression only forms linear decision surface, but the examples in the figure are not linearly separable

Machine Learning MCQs UNIT IV

1. The SVM's are less effective when:

- A) The data is linearly separable
- B) The data is clean and ready to use
- C) The data is noisy and contains overlapping points

Ans Solution: C

When the data has noise and overlapping points, there is a problem in drawing a clear

hyperplane without misclassifying.

2. The cost parameter in the SVM means:

- A) The number of cross-validations to be made
- B) The kernel to be used
- C) The tradeoff between misclassification and simplicity of the model
- D) None of the above

Ans Solution: C

The cost parameter decides how much an SVM should be allowed to “bend” with the data. For a low cost, you aim for a smooth decision surface and for a higher cost, you aim to classify more points correctly. It is also simply referred to as the cost of misclassification.

3. Which of the following are real world applications of the SVM?

- A) Text and Hypertext Categorization
- B) Image Classification
- C) Clustering of News Articles
- D) All of the above

Ans Solution: D

SVM's are highly versatile models that can be used for practically all real world problems ranging from regression to clustering and handwriting recognitions.

4. Which of the following is true about Naive Bayes ?

- A) Assumes that all the features in a dataset are equally important
- B) Assumes that all the features in a dataset are independent
- C) Both A and B – answer
- D) None of the above options

Ans Solution: C

5 What do you mean by generalization error in terms of the SVM?

- A) How far the hyperplane is from the support vectors
- B) How accurately the SVM can predict outcomes for unseen data
- C) The threshold amount of error in an SVM

Ans Solution: B

Generalisation error in statistics is generally the out-of-sample error which is the measure of how accurately a model can predict values for previously unseen data.

6 The SVM's are less effective when:

- A) The data is linearly separable
- B) The data is clean and ready to use
- C) The data is noisy and contains overlapping points

Ans Solution: C

When the data has noise and overlapping points, there is a problem in drawing a clear hyperplane without misclassifying.

7 What is/are true about kernel in SVM?

- 1. Kernel function map low dimensional data to high dimensional space
 - 2. It's a similarity function
- A) 1
 - B) 2
 - C) 1 and 2
 - D) None of these

Ans Solution: C

Both the given statements are correct

Question Context:8– 9

Suppose you are using a Linear SVM classifier with 2 class classification problem. Now you have been given the following data in which some points are circled red that are representing support vectors.

8. If you remove the following any one red points from the data. Does the decision boundary will change?

- A) Yes
- B) No

Solution: A

These three examples are positioned such that removing any one of them introduces slack in the constraints. So the decision boundary would completely change.

9. [True or False] If you remove the non-red circled points from the data, the decision boundary will change?

- A) True
- B) False

Solution: B

On the other hand, rest of the points in the data won't affect the decision boundary much.

10. What do you mean by generalization error in terms of the SVM?

- A) How far the hyperplane is from the support vectors
- B) How accurately the SVM can predict outcomes for unseen data
- C) The threshold amount of error in an SVM

Solution: B

Generalization error in statistics is generally the out-of-sample error which is the measure of how accurately a model can predict values for previously unseen data.

11. When the C parameter is set to infinite, which of the following holds true?

- A) The optimal hyperplane if exists, will be the one that completely separates the data
- B) The soft-margin classifier will separate the data
- C) None of the above

Solution: A

At such a high level of misclassification penalty, soft margin will not hold existence as there will be no room for error.

12. What do you mean by a hard margin?

- A) The SVM allows very low error in classification
- B) The SVM allows high amount of error in classification
- C) None of the above

Solution: A

A hard margin means that an SVM is very rigid in classification and tries to work extremely well in the training set, causing overfitting.

13. The minimum time complexity for training an SVM is $O(n^2)$. According to this fact, what sizes of datasets are not best suited for SVM's?

- A) Large datasets
- B) Small datasets
- C) Medium sized datasets
- D) Size does not matter

Solution: A

Datasets which have a clear classification boundary will function best with SVM's.

14. The effectiveness of an SVM depends upon:

- A) Selection of Kernel

- B) Kernel Parameters
- C) Soft Margin Parameter C
- D) All of the above

Solution: D

The SVM effectiveness depends upon how you choose the basic 3 requirements mentioned above in such a way that it maximises your efficiency, reduces error and overfitting.

15. Support vectors are the data points that lie closest to the decision surface.

- A) TRUE
- B) FALSE

Solution: A

They are the points closest to the hyperplane and the hardest ones to classify. They also have a direct bearing on the location of the decision surface.

16. The SVM's are less effective when:

- A) The data is linearly separable
- B) The data is clean and ready to use
- C) The data is noisy and contains overlapping points

Solution: C

When the data has noise and overlapping points, there is a problem in drawing a clear hyperplane without misclassifying.

17. Suppose you are using RBF kernel in SVM with high Gamma value. What does this signify?

- A) The model would consider even far away points from hyperplane for modeling
- B) The model would consider only the points close to the hyperplane for modeling
- C) The model would not be affected by distance of points from hyperplane for modeling
- D) None of the above

Solution: B

The gamma parameter in SVM tuning signifies the influence of points either near or far away from the hyperplane. For a low gamma, the model will be too constrained and include all points of the training dataset, without really capturing the shape. For a higher gamma, the model will capture the shape of the dataset well.

18. The cost parameter in the SVM means:

- A) The number of cross-validations to be made
- B) The kernel to be used
- C) The tradeoff between misclassification and simplicity of the model
- D) None of the above

Solution: C

The cost parameter decides how much an SVM should be allowed to “bend” with the data. For a low cost, you aim for a smooth decision surface and for a higher cost, you aim to classify more points correctly. It is also simply referred to as the cost of misclassification.

19. Suppose you are building a SVM model on data X. The data X can be error prone which means that you should not trust any specific data point too much. Now think that you want to build a SVM model which has quadratic kernel function of polynomial degree 2 that uses Slack variable C as one of its hyper parameter. Based upon that give the answer for following question.

What would happen when you use very large value of C(C->infinity)?Note:

For small C was also classifying all data points correctly

- A) We can still classify data correctly for given setting of hyper parameter C
- B) We can not classify data correctly for given setting of hyper parameter C
- C) Can't Say
- D) None of these

Solution: A

For large values of C, the penalty for misclassifying points is very high, so the decision boundary will perfectly separate the data if possible.

20. What would happen when you use very small C (C~0)?

- A) Misclassification would happen
- B) Data will be correctly classified
- C) Can't say
- D) None of these

Solution: A

The classifier can maximize the margin between most of the points, while misclassifying a few points, because the penalty is so low.

21. If I am using all features of my dataset and I achieve 100% accuracy on my training set, but ~70% on validation set, what should I look out for?

- A) We can still classify data correctly for given setting of hyper parameter C
- B) We can not classify data correctly for given setting of hyper parameter C
- C) Can't Say
- D) None of these

Solution: A

For large values of C, the penalty for misclassifying points is very high, so the decision boundary will perfectly separate the data if possible.

20. What would happen when you use very small C ($C \sim 0$)?

- A) Misclassification would happen
- B) Data will be correctly classified
- C) Can't say
- D) None of these

Solution: A

The classifier can maximize the margin between most of the points, while misclassifying a few points, because the penalty is so low.

21. If I am using all features of my dataset and I achieve 100% accuracy on my training set, but ~70% on validation set, what should I look out for?

- A) Underfitting
- B) Nothing, the model is perfect
- C) Overfitting

Solution: C

If we're achieving 100% training accuracy very easily, we need to check to verify if we're overfitting our data.

22. Which of the following are real world applications of the SVM?

- A) Text and Hypertext Categorization

- B) Image Classification
- C) Clustering of News Articles
- D) All of the above

Solution: D

SVM's are highly versatile models that can be used for practically all real world problems ranging from regression to clustering and handwriting recognitions.

Question Context: 23 – 25

Suppose you have trained an SVM with linear decision boundary after training SVM, you correctly infer that your SVM model is under fitting.

23. Which of the following option would you more likely to consider iterating SVM next time?

- A) You want to increase your data points
- B) You want to decrease your data points
- C) You will try to calculate more variables
- D) You will try to reduce the features

Solution: C

The best option here would be to create more features for the model.

24. Suppose you gave the correct answer in previous question. What do you think that is actually happening?

- 1. We are lowering the bias
 - 2. We are lowering the variance
 - 3. We are increasing the bias
 - 4. We are increasing the variance
- A) 1 and 2
 - B) 2 and 3
 - C) 1 and 4
 - D) 2 and 4

Solution: C

Better model will lower the bias and increase the variance

25. In above question suppose you want to change one of it's(SVM) hyperparameter so that effect would be same as previous questions i.e model will not under fit?

- A) We will increase the parameter C
- B) We will decrease the parameter C
- C) Changing in C don't effect

D) None of these

Solution: A

Increasing C parameter would be the right thing to do here, as it will ensure regularized model

26. We usually use feature normalization before using the Gaussian kernel in SVM.

What is true about feature normalization?

1. We do feature normalization so that new feature will dominate other
2. Some times, feature normalization is not feasible in case of categorical variables
3. Feature normalization always helps when we use Gaussian kernel in SVM

A) 1

B) 1 and 2

C) 1 and 3

D) 2 and 3

Solution: B

Statements one and two are correct.

Question Context: 27-29

Suppose you are dealing with 4 class classification problem and you want to train aSVM model on the data for that you are using One-vs-all method. Now answer the below questions?

27. How many times we need to train our SVM model in such case?

A) 1

B) 2

C) 3

D) 4

Solution: D

For a 4 class problem, you would have to train the SVM at least 4 times if you are using a one-vs-all method.

28. Suppose you have same distribution of classes in the data. Now, say for training 1 time in one vs all setting the SVM is taking 10 second. How many seconds would it require to train one-vs-all method end

to end?

A) 20

B) 40

C) 60

D) 80

Solution: B

It would take $10 \times 4 = 40$ seconds

29 Suppose your problem has changed now. Now, data has only 2 classes. What would you think how many times we need to train SVM in such case?

- A) 1
- B) 2
- C) 3
- D) 4

Solution: A

Training the SVM only one time would give you appropriate results

Question context: 30 –31

Suppose you are using SVM with linear kernel of polynomial degree 2. Now think that you have applied this on data and found that it perfectly fit the data that means, Training and testing accuracy is 100%.

30. Now, think that you increase the complexity (or degree of polynomial of this kernel). What would you think will happen?

- A) Increasing the complexity will over fit the data
- B) Increasing the complexity will under fit the data
- C) Nothing will happen since your model was already 100% accurate
- D) None of these

Solution: A

Increasing the complexity of the data would make the algorithm overfit the data.

31. In the previous question after increasing the complexity you found that training accuracy was still 100%. According to you what is the reason behind that?

- 1. Since data is fixed and we are fitting more polynomial term or parameters so the algorithm starts memorizing everything in the data
- 2. Since data is fixed and SVM doesn't need to search in big hypothesis space

- A) 1
- B) 2
- C) 1 and 2
- D) None of these

Solution: C

Both the given statements are correct.

32. What is/are true about kernel in SVM?

1. Kernel function map low dimensional data to high dimensional space
 2. It's a similarity function
- A) 1
B) 2
C) 1 and 2
D) None of these

Solution: C

Both the given statements are correct.

Machine Learning MCQs UNIT V

1. Which of the following is a widely used and effective machine learning algorithm based on the idea of bagging?

- a) Decision Tree
- b) Regression
- c) Classification
- d) Random Forest

Ans D

2. Which of the following is a disadvantage of decision trees?

- a) Factor analysis
- b) Decision trees are robust to outliers
- c) Decision trees are prone to be overfit
- d) None of the above

Ans C

3. Can decision trees be used for performing clustering?

- a. True
- b. False

Ans Solution: (A)

Decision trees can also be used to find clusters in the data but clustering often generates natural clusters and is not dependent on any objective function.

4. Which of the following algorithm is most sensitive to outliers?

- a. K-means clustering algorithm

- b. K-medians clustering algorithm
- c. K-modes clustering algorithm
- d. K-medoids clustering algorithm

Ans Solution: (A)

5 Sentiment Analysis is an example of:

- a. Regression
- b. Classification
- c. Clustering
- d. Reinforcement Learning

Options:

- a. 1 Only
- b. 1 and 2
- c. 1 and 3
- d. 1, 2 and 4

Ans D

6 Which of the following is the most appropriate strategy for data cleaning before performing clustering analysis, given less than desirable number of data points:

Capping and flooring of variables Removal of outliers

Options:

- a. 1 only
- b. 2 only
- c. 1 and 2
- d. None of the above

Ans A

7 Which of the following is/are true about bagging trees?

- 1. In bagging trees, individual trees are independent of each other
 - 2. Bagging is the method for improving the performance by aggregating the results of weak learners
- A) 1
B) 2
C) 1 and 2
D) None of these

Ans Solution: C

Both options are true. In Bagging, each individual trees are independent of each other because they consider different subset of features and samples.

8. Which of the following is/are true about boosting trees?

1. In boosting trees, individual weak learners are independent of each other
 2. It is the method for improving the performance by aggregating the results of weak learners
- A) 1
B) 2
C) 1 and 2
D) None of these

Ans Solution: B

In boosting tree individual weak learners are not independent of each other because each tree correct the results of previous tree. Bagging and boosting both can be consider as improving the base learners results.

9. In Random forest you can generate hundreds of trees (say T₁, T₂T_n) and then aggregate the results of these tree. Which of the following is true about individual (T_k) tree in Random Forest?

1. Individual tree is built on a subset of the features
 2. Individual tree is built on all the features
 3. Individual tree is built on a subset of observations
 4. Individual tree is built on full set of observations
- A) 1 and 3
B) 1 and 4
C) 2 and 3
D) 2 and 4

Ans Solution: A

Random forest is based on bagging concept, that consider fraction of sample and fraction of feature for building the individual trees.

10. Suppose you are using a bagging based algorithm say a RandomForest in model building.Which of the following can be true?

1. Number of tree should be as large as possible
 2. You will have interpretability after using Random Forest
- A) 1
B) 2

- C) 1 and 2
- D) None of these

Ans Solution: A

Since Random Forest aggregate the result of different weak learners, If It is possible we would want more number of trees in model building. Random Forest is a black box model you will lose interpretability after using it.

11. Which of the following is/are true about Random Forest and Gradient Boosting ensemble methods?

- 1. Both methods can be used for classification task
 - 2. Random Forest is use for classification whereas Gradient Boosting is use for regression task
 - 3. Random Forest is use for regression whereas Gradient Boosting is use for Classification task
 - 4. Both methods can be used for regression task
- A) 1
 - B) 2
 - C) 3
 - D) 4
 - E) 1 and 4

Solution: E

Both algorithms are design for classification as well as regression task

12. In Random forest you can generate hundreds of trees (say T₁, T₂T_n) and then aggregate the results of these tree. Which of the following is true about individual(T_k) tree in Random Forest?

- 1. Individual tree is built on a subset of the features
 - 2. Individual tree is built on all the features
 - 3. Individual tree is built on a subset of observations
 - 4. Individual tree is built on full set of observations
- A) 1 and 3
 - B) 1 and 4
 - C) 2 and 3
 - D) 2 and 4

Solution: A

Random forest is based on bagging concept, that consider fraction of sample and fraction of feature for building the individual trees.

13. Which of the following algorithm doesn't uses learning Rate as one of its hyperparameter?

1. Gradient Boosting

2. Extra Trees

3. AdaBoost

4. Random Forest

A) 1 and 3

B) 1 and 4

C) 2 and 3

D) 2 and 4

Solution: D

Random Forest and Extra Trees don't have learning rate as a hyperparameter.

14. Which of the following algorithm are not an example of ensemble learning algorithm?

A) Random Forest

B) Adaboost

C) Extra Trees

D) Gradient Boosting

E) Decision Trees

Solution: E

Decision trees doesn't aggregate the results of multiple trees so it is not an ensemble algorithm.

15. Suppose you are using a bagging based algorithm say a RandomForest in model building. Which of the following can be true?

1. Number of tree should be as large as possible

2. You will have interpretability after using RandomForest

A) 1

B) 2

C) 1 and 2

D) None of these

Solution: A

Since Random Forest aggregate the result of different weak learners, If It is possible we would want more number of trees in model building. Random Forest is a black box model you will lose interpretability after using it.

16. True-False: The bagging is suitable for high variance low bias models?

- A) TRUE
- B) FALSE

Solution: A

The bagging is suitable for high variance low bias models or you can say for complex models.

17. To apply bagging to regression trees which of the following is/are true in such case?

- 1. We build the N regression with N bootstrap sample
 - 2. We take the average the of N regression tree
 - 3. Each tree has a high variance with low bias
- A) 1 and 2
 - B) 2 and 3
 - C) 1 and 3
 - D) 1,2 and 3

Solution: D

All of the options are correct and self-explanatory

18. How to select best hyper parameters in tree based models?

- A) Measure performance over training data
- B) Measure performance over validation data
- C) Both of these
- D) None of these

Solution: B

We always consider the validation results to compare with the test result.

19. In which of the following scenario a gain ratio is preferred over Information Gain?

- A) When a categorical variable has very large number of category
- B) When a categorical variable has very small number of category
- C) Number of categories is the not the reason
- D) None of these

Solution: A

When high cardinality problems, gain ratio is preferred over Information Gain technique.

20. Suppose you have given the following scenario for training and validation error for Gradient Boosting. Which of the following hyper parameter would you choose in such case?

- A) 1
- B) 2
- C) 3
- D) 4

Solution: B

Scenario 2 and 4 has same validation accuracies but we would select 2 because depth is lower is better hyper parameter.

21. Which of the following is/are not true about DBSCAN clustering algorithm:

- 1. For data points to be in a cluster, they must be in a distance threshold to a core point
- 2. It has strong assumptions for the distribution of data points in dataspace
- 3. It has substantially high time complexity of order $O(n^3)$
- 4. It does not require prior knowledge of the no. of desired clusters
- 5. It is robust to outliers

Options:

- A. 1 only
- B. 2 only
- C. 4 only
- D. 2 and 3

Solution: D

DBSCAN can form a cluster of any arbitrary shape and does not have strong assumptions for the distribution of data points in the data space.

DBSCAN has a low time complexity of order $O(n \log n)$ only

22. Point out the correct statement.

- a) The choice of an appropriate metric will influence the shape of the clusters
- b) Hierarchical clustering is also called HCA
- c) In general, the merges and splits are determined in a greedy manner
- d) All of the mentioned

Answer: d

Explanation: Some elements may be close to one another according to one distance and farther away according to another.

23. Which of the following is required by K-means clustering?

- a) defined distance metric
- b) number of clusters
- c) initial guess as to cluster centroids
- d) all of the mentioned

Answer: d

Explanation: K-means clustering follows partitioning approach.

24. Point out the wrong statement.

- a) k-means clustering is a method of vector quantization
- b) k-means clustering aims to partition n observations into k clusters
- c) k-nearest neighbor is same as k-means
- d) none of the mentioned

Answer: c

Explanation: k-nearest neighbour has nothing to do with k-means.

25. Which of the following function is used for k-means clustering?

- a) k-means
- b) k-mean
- c) heat map
- d) none of the mentioned

Answer: a

Explanation: K-means requires a number of clusters.

Machine Learning MCQ

Faculty: Dr Madhavi Pradhan

1) Supervised learning is _____

1. learning through previously seen given examples
2. learning technique which uses rewards and punishment as a feedback
3. 1 & 2
4. None of the above

Answer 1

2) Unsupervised learning is _____

1. learning technique which uses rewards and punishment as a feedback
2. learning technique where the system discovers the patterns from the given example
- 3 . 1 and 2
4. None of the above

Answer 2

3) Which of the following is a supervised learning?

- 1 clustering
- 2 Associations
- 3 Classification
4. None of the above

Answer 3

4) Principle Component analysis technique is used for

1. feature selection

- 2. handling missing values
- 3. Both 1 & 2
- 4. None of above

5) Naïve Bayes is _____ algorithm

- 1 clustering
- 2 associations
- 3. classification
- 4. None of the above

Answer 3

6) Naïve Bayes algorithm makes an assumption known as

- 1. conditional independence
- 2. conditional dependence
- 3. makes no assumption
- 4. None of above

Answer 1

7) Support vector machine classify _____

- 1. only linear data
- 2. only non linear data
- 3. Both linear as well as non linear data
- 4. None of above

Answer 3

8) Decision tree is _____ algorithm

- 1 Clustering
- 2 Associations
- 3 Classification
- 4. None of the above

Answer 3

9) k-means algorithm is used for

- 1 Clustering
- 2 Associations
- 3 Classification
- 4 None of the above

Answer 1

10) Confusion matrix is

- 1 performance metric used for evaluation of classification algorithm
- 2 performance metric used for evaluation of association algorithm
- 3 Both 1&2
- 4 None of the above

Answer 1

11) ROC is used to assess a _____ model

- 1 Clustering
- 2 Associations
- 3 Classification
- 4 None of the above

Answer 3

- 12) ROC is abbreviation of
- 1 receiver operating characteristics
 - 2 receiver opinion characteristics
 - 3 receiver open characteristics
 - 4 None of the above

Answer 1

13) The ability to detect true positive samples among all the potential positives can be assessed using a measure called

- 1 Recall
- 2 Precision
- 3 Accuracy
- 4 None of the above

Answer 1

14) In classification, false positive rate is known as _____

- 1 specificity
- 2 sensitivity
- 3 accuracy
- 4 None of the above

Answer 1

15) Random forests are an example of _____

- 1 boosted trees
- 2 simple trees
- 3 bagged tree ensemble
- 4 None of the above

Answer 3

16 PCA stand for

- 1 Principle Constant Analysis
- 2 Principle Component Analysis
- 3 Principle Corelation Analysis
- 4 None of the above

[REDACTED]

17. LDA stand for

- 1 Linear Direct Analysis
- 2 Linear discriminant Analysis
- 3 Linear Different Analysis
- 4 None of the above

[REDACTED]

_____ regression model illustrate the relationship between two variables or factors

- 1 Linear
- 2 Ridge
- 3 Lasso
- 4 None of the above

Answer 1

18 RANSAC stand for

- 1 Random Sequence Consensus
- 2 Random Sample Consensus

3 Random Simple Consensus

4 None of the above

Answer 2

19 If there are two probabilistic events A and B with the conditional probabilities $P(A|B)$ and $P(B|A)$ then according to the baye's theorem $P(A|B)$ will be

- a. $P(B|A) P(A)/ P(B)$
- b. $P(A|B) P(A)/ P(B)$
- c. $P(B|A) P(B)/ P(A)$
- d. $P(B|A)/ P(B)$

Ans :- a

20 Which probabilistic distributions are used with the naive Bayes variants ?

- a. Bernoulli
- b. Multinomial
- c. Gaussian
- d. All the above

Ans :- d

21 Which of the following is true about Naive Bayes ?

- a. Assumes that all the features in a dataset are equally important
- b. Assumes that all the features in a dataset are independent
- c. Both A and B
- d. None of the above options

Ans :- c

23 Three companies A, B and C supply 25%, 35% and 40% of the notebooks to a school. Past experience shows that 5%, 4% and 2% of the notebooks produced by these companies are

defective. If a notebook was found to be defective, what is the probability that the notebook was supplied by A?

- a. 44/69
- b. 25/69
- c. 13/24
- d. 11/24

Ans :- b

Explanation: Let A, B and C be the events that notebooks are provided by A, B and C respectively.

Let D be the event that notebooks are defective

Then,

$$P(A) = 0.25, P(B) = 0.35, P(C) = 0.4$$

$$P(D|A) = 0.05, P(D|B) = 0.04, P(D|C) = 0.02$$

$$\begin{aligned}P(A | D) &= (P(D | A) * P(A)) / (P(D | A) * P(A) + P(D | B) * P(B) + P(D | C) * P(C)) \\&= (0.05 * 0.25) / ((0.05 * 0.25) + (0.04 * 0.35) + (0.02 * 0.4)) = 2000 / (80 * 69) \\&= 25/69.\end{aligned}$$

23 How many terms are required for building a bayes model?

- a. 1
- b. 2
- c. 3
- d. 4

Answer: c

Explanation: The three required terms are a conditional probability and two unconditional probability.

24 Where does the bayes rule can be used?

- a. Solving queries
- b. Increasing complexity
- c. Decreasing complexity
- d. Answering probabilistic query

Answer: d

Explanation: Bayes rule can be used to answer the probabilistic queries conditioned on one piece of evidence.

25 What does the Bayesian network provide?

- a. Complete description of the domain
- b. Partial description of the domain
- c. Complete description of the problem
- d. None of the mentioned

Answer: a

Explanation: A Bayesian network provides a complete description of the domain.

26 What is the consequence between a node and its predecessors while creating a Bayesian network?

- a. Functionally dependent
- b. Dependant
- c. Conditionally independent
- d. Both Conditionally dependant & Dependant

Answer: c

Explanation: The semantics to derive a method for constructing bayesian networks were led to the consequence that a node can be conditionally independent of its predecessors.

27 What is needed to make probabilistic systems feasible in the world?

- a. Reliability
- b. Crucial robustness
- c. Feasibility
- d. None of the mentioned

Answer: b

Explanation: On a model-based knowledge provides the crucial robustness needed to make a probabilistic system feasible in the real world.

28. What does the Bayesian network provide?

- a. Complete description of the domain
- b. Partial description of the domain
- c. Complete description of the problem
- d. None of the mentioned

Answer: a

Explanation: A Bayesian network provides a complete description of the domain.

29 How the entries in the full joint probability distribution can be calculated?

- a. Using variables
- b. Using information
- c. Both Using variables & information
- d. None of the mentioned

Answer: b

Explanation: Every entry in the full joint probability distribution can be calculated from the information in the network.

30 How the bayesian network can be used to answer any query?

- a. Full distribution
- b. Joint distribution
- c. Partial distribution
- d. All of the mentioned

Answer: b

Explanation: If a bayesian network is a representation of the joint distribution, then it can solve any query, by summing all the relevant joint entries.

31. How can the compactness of the bayesian network be described?

- a. Locally structured
- b. Fully structured
- c. Partial structure
- d. All of the mentioned

Answer: a

Explanation: The compactness of the bayesian network is an example of a very general property of a locally structured system.

32.What is the consequence between a node and its predecessors while creating bayesian network?

- a. Functionally dependent
- b. Dependant
- c. Conditionally independent

- d. Both Conditionally dependant & Dependant

Answer: c

Explanation: The semantics to derive a method for constructing bayesian networks were led to the consequence that a node can be conditionally independent of its predecessors.

33.Bayesian classifiers is

- a. A class of learning algorithms that tries to find an optimum classification of a set of examples using probabilistic theory.
- b. Any mechanism employed by a learning system to constrain the search space of a hypothesis
- c. An approach to the design of learning algorithms that is inspired by the fact that when people encounter new situations, they often explain them by reference to familiar experiences, adapting the explanations to fit the new situation.
- d. None of these

Ans :- Option: A

34.Naive prediction is

- a. A class of learning algorithms that try to derive a Prolog program from examples
- b. A table with n independent attributes can be seen as an n- dimensional space.
- c. A prediction made using an extremely simple method, such as always predicting the same output.
- d. None of these

Ans :- C

35.what is true about gaussian naive bayes?

- a. Gaussian Naive Bayes is one classifier model
- b. most popular one
- c. it is the simplest
- d. All

Ans :- d

36. Naive bayes is used for

- a. Recommendation System
- b. Text classification/ Spam Filtering/ Sentiment Analysis
- c. Intrusion detection
- d. Both a and b

Ans :- d

37. When the assumption of independence holds, a Naive Bayes classifier performs better compared to other models like logistic regression.

- a. True
- b. false

Ans :- a

38. Why is the naive bayes classifier used in machine learning?

- a. Technique based on Bayes' theorem with an assumption of independence between predictors
- b. Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature
- c. Naive Bayes classifier performs better compare to other models
- d. all

Ans :- d

39 Support Vector Machine works well with,

- a) Linear Scenarios
- b) Non-linear Scenarios
- c) Both of these
- d) None of these

Answer: c) Both of these

40. Which of the following is best for MNIST dataset classification,

- a) Naïve Bayes
- b) Support Vector Machines
- c) Random forest
- d) Decision tree

Answer: b) Support Vector Machines

41 Which method measure the usefulness of a subset of feature by actually training a model on it?

- a filter
- b wrapper
- c PCA
- d. None of the Above

Answer: b) wrapper

42 Which method measure the importance of feature by their corelation with dependent variable?

- a filter
- b wrapper
- c PCA
- d. None of the Above

Answer: a) filter

43 Two classes separated by a margin with two boundaries are called as,

- a) Linear Vectors
- b) Support Vectors
- c) Test Vectors
- d) None of these

Answer: b) Support Vectors

44 Scikit-learn supports which kernels,

- a) Polynomial kernels
- b) Sigmoid kernels

- c) Custom kernels
- d) All of these

Answer: d) All of these

- 45 Which of the following is the default kernel used in SVM,
- a) Polynomial kernel
 - b) Sigmoid kernel
 - c) Custom kernel
 - d) Radial Basis Function

Answer: d) Radial Basis Function

- 46 The gamma parameter in RBF determines,
- a) Amplitude of the function
 - b) Altitude of the function
 - c) Complexity of the function
 - d) None of these

Answer: a) Amplitude of the function

- 47 Scikit-learn allows us to create which kernel as a normal python function,
- a) Polynomial kernel
 - b) Custom kernel
 - c) Sigmoid kernel
 - d) All of these

Answer: b) Custom kernel

- 48 To find out a trade-off between precision and number of support vectors, scikit-learn provides an implementation called as,
- a) NuSVC
 - b) BuSVC
 - c) MuSVC
 - d) AuSVC

Answer: a) NuSVC

- 49 The RBF kernel is based on the function:

- a) $K(\bar{x}_i, \bar{x}_j) = e^{-\gamma|\bar{x}_i - \bar{x}_j|^2}$
- b) $K(\bar{x}_i, \bar{x}_j) = (\gamma \bar{x}_i^T \cdot \bar{x}_j + r)^c$
- c) $K(\bar{x}_i, \bar{x}_j) = \frac{1 - e^{-2(\gamma \bar{x}_i^T \cdot \bar{x}_j + r)}}{1 + e^{-2(\gamma \bar{x}_i^T \cdot \bar{x}_j + r)}}$

d) None of these

$$K(\bar{x}_i, \bar{x}_j) = e^{-\gamma |\bar{x}_i - \bar{x}_j|^2}$$

Answer: a)

50 The polynomial kernel is based on the function:

- a) $K(\bar{x}_i, \bar{x}_j) = e^{-\gamma |\bar{x}_i - \bar{x}_j|^2}$
- b) $K(\bar{x}_i, \bar{x}_j) = (\gamma \bar{x}_i^T \cdot \bar{x}_j + r)^c$
- c) $K(\bar{x}_i, \bar{x}_j) = \frac{1 - e^{-2(\gamma \bar{x}_i^T \cdot \bar{x}_j + r)}}{1 + e^{-2(\gamma \bar{x}_i^T \cdot \bar{x}_j + r)}}$
- d) None of these

$$K(\bar{x}_i, \bar{x}_j) = (\gamma \bar{x}_i^T \cdot \bar{x}_j + r)^c$$

Answer: b)

51 The sigmoid kernel is based on this function:

- a) $K(\bar{x}_i, \bar{x}_j) = e^{-\gamma |\bar{x}_i - \bar{x}_j|^2}$
- b) $K(\bar{x}_i, \bar{x}_j) = (\gamma \bar{x}_i^T \cdot \bar{x}_j + r)^c$
- c) $K(\bar{x}_i, \bar{x}_j) = \frac{1 - e^{-2(\gamma \bar{x}_i^T \cdot \bar{x}_j + r)}}{1 + e^{-2(\gamma \bar{x}_i^T \cdot \bar{x}_j + r)}}$
- d) None of these

$$K(\bar{x}_i, \bar{x}_j) = \frac{1 - e^{-2(\gamma \bar{x}_i^T \cdot \bar{x}_j + r)}}{1 + e^{-2(\gamma \bar{x}_i^T \cdot \bar{x}_j + r)}}$$

Answer: c)

52 What is/are true about kernel in SVM,

1. It maps low dimensional data to high dimensional data.
2. It is a similarity function.

- a) 1
- b) 2
- c) Both 1 and 2

d) None of these

Answer: c) Both 1 and 2

53 Which type of classifier is SVM,

- a) Discriminative
- b) Generative
- c) Both
- d) None of these

Answer: a) Discriminative

54 SVM is used to solve which type of problems,

- a) Classification
- b) Regression
- c) Clustering
- d) Both Classification and Regression

Answer: d) Both Classification and Regression

55 SVM is which type of learning algorithm,

- a) Supervised
- b) Unsupervised
- c) Both
- d) None of these

Answer: a) Supervised

56 The goal of SVM is to,

- a) Find the optimal separating hyperplane which minimizes the margin of training data.
- b) Find the optimal separating hyperplane which maximizes the margin of training data.
- c) Both
- d) None of these

Answer: b) Find the optimal separating hyperplane which maximizes the margin of training data.

57 The equation for hyperplane is,

$$w^T \bar{x} + b = 0 \text{ where } w = \begin{pmatrix} w_1 \\ \vdots \\ w_m \end{pmatrix} \text{ and } \bar{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}$$

a)

$$w^T \bar{x} - b = 0 \text{ where } w = \begin{pmatrix} w_1 \\ \vdots \\ w_m \end{pmatrix} \text{ and } \bar{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}$$

b)

$$w^T \bar{x} * b = 0 \text{ where } w = \begin{pmatrix} w_1 \\ \vdots \\ w_m \end{pmatrix} \text{ and } \bar{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}$$

c)

- d) None of these

$$\bar{w}^T \bar{x} + b = 0 \text{ where } \bar{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_m \end{pmatrix} \text{ and } \bar{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}$$

Answer: a)

- 58 What is a kernel in SVM?

- a) SVM algorithms use a set of mathematical functions that are defined as the kernel
- b) SVM algorithms use a set of logarithmic functions that are defined as the kernel
- c) SVM algorithms use a set of exponential functions that are defined as the kernel
- d) SVM algorithms use a set of algebraic functions that are defined as the kernel

Answer: a) SVM algorithms use a set of mathematical functions that are defined as the kernel

- 59 Which of the following is false,

- a) SVM's are very good when we have no idea on the data.
- b) It works well with unstructured and semi structured data.
- c) The kernel trick is real strength of SVM.
- d) It scales relatively well to low dimensional data.

Answer: d) It scales relatively well to low dimensional data.

- 60 Which of the following is false,

- a) SVM algorithm is suitable for large data sets.
- b) It does not perform well when the data has more noise.
- c) SVM algorithm is not suitable for large data sets.
- d) None of these

Answer: a) SVM algorithm is suitable for large data sets.

Name of UNIT: Regression

01	The number of correct predictions made by the model is the	
	Option A	Precision
	Option B	Accuracy
	Option C	Recall
	Option D	Sensitivity
	Correct Answer	B
02	_____ regression model illustrate the relationship between two variables or factors	
	Option A	Linear
	Option B	Ridge
	Option C	Lasso
	Option D	Elastic Net
	Correct Answer	A
03	In which type of regression uses squared penalty	
	Option A	Linear
	Option B	Ridge
	Option C	Lasso
	Option D	Elastic Net
	Correct Answer	B
04	RANSAC stand for	
	Option A	Random Sequence Consensus
	Option B	Random Sample Consensus
	Option C	Random Simple Consensus
	Option D	Random Square Consensus
	Correct Answer	B
05	Logistic Regression is type of	
	Option A	Reinforcement Algorithm
	Option B	Clustering Algorithm
	Option C	Classification Algorithm
	Option D	Unsupervised Algorithm
	Correct Answer	C
06	Isotonic regression is used in statistical inference	
	Option A	True
	Option B	False

Name of UNIT: Regression

	Correct Answer	A
07	In which type of regression the relationship between the independent variable x and the dependent variable y is modelled as an nth degree polynomial in x.	
	Option A	Linear Regression
	Option B	Ridge Regression
	Option C	Lasso Regression
	Option D	Polynomial Regression
	Correct Answer	D
08	_____ as an iterative method which finds the values od the parameters of a function that helps to minimize the cost function as much as possible	
	Option A	Logit function
	Option B	Gradient Decent
	Option C	Sigmoid function
	Option D	All of the above
	Correct Answer	B
09	A linear classifier is achieved by making a classification decision depending on the value of a linear combination of the characteristics.	
	Option A	True
	Option B	False
	Correct Answer	A
10	Out of these is example of non linear regression model	
	Option A	Linear Regression
	Option B	Ridge Regression
	Option C	Polynomial Regression
	Option D	Lasso Regression
	Correct Answer	C
11	The parameters those cannot be directly tuned from normal training process are known as	
	Option A	Non linear parameter
	Option B	Hyper parameter
	Option C	Non parameter
	Option D	Linear parameter
	Correct Answer	B
	Predicting whether a given mass of tissue is benign or malignant is example of	

Name of UNIT: Regression

	Option A	Regression
	Option B	Binary Classification
	Option C	Clustering
	Option D	Unsupervised Learning
	Correct Answer	B
13	Grid search is used to optimize	
	Option A	Non linear parameter
	Option B	Hyper parameter
	Option C	Non parameter
	Option D	Linear parameter
	Correct Answer	B
14	False Positive Rate is defined by,	
	Option A	$FPR = FN / (FP + TN)$
	Option B	$FPR = FP / (FP + TN)$
	Option C	$FPR = FP / (FP + TP)$
	Option D	$FPR = FP / (TP + TN)$
	Correct Answer	B
15	_____ is a tvalues always knowable test that is being used to define the performance of the classification model on the test data where true	
	Option A	
	Option B	Multiplication Matrix
	Option C	Confusion Matrix
	Option D	
	Correct Answer	C
16	True positive rate is defined by	
	Option A	$TPR = TP / (TP + FN)$
	Option B	$TPR = TF / (TP + FN)$
	Option C	$TPR = TP / (TF + FN)$
	Option D	$TPR = TP / (TP + FP)$
	Correct Answer	B
17	ROC stand for	
	Option A	Receiver Open Characteristic curve
	Option B	Receiver Operating Characteristic curve
	Option C	Receiver Operational Characteristic curve

Name of UNIT: Regression

	Option D	Receiver Operand Characteristic curve
	Correct Answer	B
18	Sigmoid function is used it exists between	
	Option A	2 to 6
	Option B	0 and 1
	Option C	0 to 3
	Option D	1 and 2
	Correct Answer	B
19	Roc curve plots	
	Option A	FNR vs FPR at different classification threshold
	Option B	TPR vs FPR at different classification threshold
	Option C	FPR vs FPR at different classification threshold
	Option D	TPR vs TPR at different classification threshold
	Correct Answer	B
20	The relationship between more than one explanatory variables and response variable is modeled by	
	Option A	Simple linear Regression
	Option B	Lasso Regression
	Option C	Ridge Regression
	Option D	Multiple linear Regression
	Correct Answer	D

Name of UNIT: Introduction to Machine learning

01	In which type of machine learning , a computer program interacts with a dynamic environment in which it perform a certain goal and is provided with feedback in terms of rewards and punishments as it navigates its problem space?	
	Option A	Supervised Learning
	Option B	Unsupervised Learning
	Option C	Reinforcement Learning
	Option D	Semi Supervised Learning
	Correct Answer	C
02	In which type of learning, the computer is presented with example inputs and their desired outputs, given by a “teacher”, and the goal is to learn a general rule that maps inputs?	
	Option A	Supervised Learning
	Option B	Unsupervised Learning
	Option C	Reinforcement Learning
	Option D	Semi Supervised Learning
	Correct Answer	A
03	Which statement is not correct statement for Machine Learning?	
	Option A	is the field of study that gives computers the capability to learn without being explicitly programmed.
	Option B	It gives the computer that makes it more similar to humans
	Option C	Replace man with machine
	Option D	can be explained as automating and improving the learning process of computers based on their experiences without being actually programmed
	Correct Answer	C
04	Which is not a type of Machine Learning algorithm?	
	Option A	Supervised
	Option B	Unsupervised
	Option C	Restoration Learning

Name of UNIT: Introduction to Machine learning

	Option D	Semi Supervised Learning
	Correct Answer	C
05	No labels are given to the learning algorithm, leaving it on its own to find structure in its input. It is used for to divide population in different groups.	
	Option A	Supervised
	Option B	Unsupervised
	Option C	Restoration Learning
	Option D	Semi Supervised Learning
	Correct Answer	B
06	After a model captures the probability distribution of your input data, it will be able to generate more data. This can be very useful to make your classifier more robust used in	
	Option A	Supervised
	Option B	Unsupervised
	Option C	Clustering
	Option D	Generative Models
	Correct Answer	D
07	Application/ Applications of reinforcement learning is/are	
	Option A	Video Gaming
	Option B	Industrial Simulation
	Option C	Resource management
	Option D	All of the above
	Correct Answer	D
08	In which method of machine learning Use the computer to help us visualize high dimension data.	
	Option A	High Dimension Visualization
	Option B	Generative model
	Option C	Clustering
	Option D	Deep Learning
	Correct Answer	

Name of UNIT: Introduction to Machine learning

09	_____ separates similar data into groups depending on various features.
Option A	High Dimension Visualization
Option B	Generative model
Option C	Clustering
Option D	Deep Learning
Correct Answer	C
10	Feed in DATA(Input) + Output, run it on machine during training and the machine creates its own program(logic), which can be evaluated while testing.
Option A	Deep Learning
Option B	Traditional Learning
Option C	Machine Learning
Option D	Neural Network
Correct Answer	C
11	Spam Filtering is the example of,
Option A	Classification
Option B	Regression
Option C	Clustering
Option D	Random Forest
Correct Answer	A
12	A set of inputs is to be divided into groups. The groups are not known beforehand, making this typically an unsupervised task is known as _____
Option A	Classification
Option B	Regression
Option C	Clustering
Option D	Random Forest
Correct Answer	C

01	A computer program interacts with a dynamic environment in which it must perform a certain goal .The program is provided feedback in terms of rewards and punishments as it navigates its problem
----	---

Name of UNIT: Introduction to Machine learning

	space.
	Option A Supervised Learning
	Option B Unsupervised Learning
	Option C Reinforcement Learning
	Option D Semi Supervised Learning
	Correct Answer
02	The computer is presented with example inputs and their desired outputs, given by a “teacher”, and the goal is to learn a general rule that maps inputs to output is used in which type of learning?
	Option A Supervised Learning
	Option B Unsupervised Learning
	Option C Reinforcement Learning
	Option D Semi Supervised Learning
	Correct Answer
03	Which statement is not correct statement for Machine Learning
	Option A is the field of study that gives computers the capability to learn without being explicitly programmed.
	Option B It gives the computer that makes it more similar to humans
	Option C Replace man with machine
	Option D can be explained as automating and improving the learning process of computers based on their experiences without being actually programmed
	Correct Answer
04	Which is not a type of Machine Learning algorithm?
	Option A Supervised
	Option B Unsupervised
	Option C Restoration Learning
	Option D Semi Supervised Learning

Name of UNIT: Introduction to Machine learning

	Correct Answer	
05	No labels are given to the learning algorithm, leaving it on its own to find structure in its input. It is used for to divide population in different groups.	
	Option A	Supervised
	Option B	Unsupervised
	Option C	Restoration Learning
	Option D	Semi Supervised Learning
	Correct Answer	
06	After a model captures the probability distribution of your input data, it will be able to generate more data. This can be very useful to make your classifier more robust used in	
	Option A	Supervised
	Option B	Unsupervised
	Option C	Clustering
	Option D	Generative Models
	Correct Answer	
07	Applications of reinforcement learning	
	Option A	Video Gaming
	Option B	Industrial Simulation
	Option C	Resource management
	Option D	All of the above
	Correct Answer	
08	In which method of machine learning Use the computer to help us visualize high dimension data.	
	Option A	High Dimension Visualization
	Option B	Generative model
	Option C	Clustering
	Option D	Deep Learning
	Correct Answer	

Name of UNIT: Introduction to Machine learning

09	In which method of machine learning computer to separate similar data into groups depend on various features
	Option A High Dimension Visualization
	Option B Generative model
	Option C Clustering
	Option D Deep Learning
	Correct Answer
10	Feed in DATA(Input) + Output, run it on machine during training and the machine creates its own program(logic), which can be evaluated while testing.
	Option A Deep Learning
	Option B Traditional Learning
	Option C Machine Learning
	Option D Neural Network
	Correct Answer
11	Spam Filtering is the example of,
	Option A Classification
	Option B Regression
	Option C Clustering
	Option D Random Forest
	Correct Answer
12	A set of inputs is to be divided into groups. The groups are not known beforehand, making this typically an unsupervised task is known as
	Option A Classification
	Option B Regression
	Option C Clustering
	Option D Random Forest
	Correct Answer

Name of UNIT: Naïve Bayes and Support Vector Machine

01	Naïve Bayes algorithm used for which type of classification problems	
	Option A	Binary
	Option B	Multiclass
	Option C	Binary and multiclassification problem
	Option D	None of above
	Correct Answer	C
02	When the predictor take up a continuous value and are not discrete these values are sampled from	
	Option A	Multinomial Naïve Bayes
	Option B	Gaussian Naïve Bayes
	Option C	Bernoulli Naïve Bayes
	Option D	Complement Naïve Bayes
	Correct Answer	B
03	Which of the following is correct statement	
	Option A	Naïve Bayes is known as decent classifier it is known to be bad estimator
	Option B	Naïve Bayes is known as bad classifier it is known to be bad estimator
	Option C	Naïve Bayes is known as decent classifier it is known to be good estimator
	Option D	Naïve Bayes is known as bad classifier it is known to be good estimator
	Correct Answer	A
04	What is the main objective for the selection of the hyperplane	
	Option A	Minimum possible margin between support vectors in the given datasets
	Option B	Average possible margin between support vectors in the given datasets
	Option C	Maximum possible margin between support vectors in the given datasets
	Option D	All of the above
	Correct Answer	C
05	In the SVM the distance between the either nearest points known as	
	Option A	Hyperplane
	Option B	Kernel
	Option C	Class
	Option D	Margin
	Correct Answer	D
06	Naïve Bayes classifier is a probabilistic machine learning model that's used for classification tasks.	
	Option A	True
	Option B	False

Name of UNIT: Naïve Bayes and Support Vector Machine

	Correct Answer	A
07	The SVM algorithm implemented using kernel takes	
	Option A	Low Dimensional input space and transforms it into a higher dimension space
	Option B	Non separable problem to separable problems by adding more dimensions
	Option C	High dimensional input space and transform it into a lower dimension space
	Option D	A and B both
	Correct Answer	D
08	Linear kernel equation for prediction for new input between input(x) and each support vector(x_i) is calculated as	
	Option A	$F(x)=B(0)+\sum(ai*(x,x_i))$
	Option B	$K(x,x_i)=1+\sum(x*x_i)^d$
	Option C	$K(x,x_i)=\exp(-gamma*\sum((x-x_i)^2))$
	Option D	Both A and B
	Correct Answer	A
09	SVM is a linear classifier that learns an (n-1) dimensional classifier for classification of data into two classes	
	Option A	True
	Option B	False
	Correct Answer	A
10	Polynomial kernel can be written as,	
	Option A	$F(x)=B(0)+\sum(ai*(x,x_i))$
	Option B	$K(x,x_i)=1+\sum(x*x_i)^d$
	Option C	$K(x,x_i)=\exp(-gamma*\sum((x-x_i)^2))$
	Option D	Both A and B
	Correct Answer	B
11	Polynomial and exponential kernels calculate the separation line in a higher dimension is called as	
	Option A	Hyperplane
	Option B	Kernel Trick
	Option C	Margin
	Option D	Class
	Correct Answer	B
	Radial Basis Kernel function is	

Name of UNIT: Naïve Bayes and Support Vector Machine

	Option A	$F(x)=B(0)+\sum(ai*(x,x_i))$
	Option B	$K(x,x_i)=1+\sum(x*x_i)^d$
	Option C	$K(x,x_i)=\exp(-\gamma*\sum((x-x_i)^2))$
	Option D	Both A and B
	Correct Answer	C
13	Which is the correct command to import SVMs Linear Classification in sklearn	
	Option A	from sklearn.datasets import make_classification
	Option B	import sklearn.datasets from make_classification
	Option C	import sklearn.svm from LinearSVC
	Option D	from sklearn.svm import LinearSVC
	Correct Answer	D
14	What is the meaning of C and Gamma parameters used to train the SVM with the Radial Basis Function(RBF) kernel	
	Option A	Cache size and Epsilon
	Option B	Verbose and Tolerance for stopping criterion
	Option C	Regularization parameter and kernel coefficient
	Option D	Kernel and Degree
	Correct Answer	C
15	In which Naïve Bayes predictors are Boolean variable	
	Option A	Multinomial Naïve Bayes
	Option B	Gaussian Naïve Bayes
	Option C	Bernoulli Naïve Bayes
	Option D	Multi Naïve Bayes
	Correct Answer	C
16	_____ parameter defines how far the influence of a single training example reaches with low values meaning ‘far’ and high values meaning ‘close’	
	Option A	Kernel
	Option B	Gamma
	Option C	Margin
	Option D	Class
	Correct Answer	B

Name of UNIT: Feature Selection

01	Transforming Nominal Attributes is a method to deal with (
	Option A	Sensor data
	Option B	Numerical data
	Option C	Categorical data
	Option D	Clinical data
	Correct Answer	C
02	When the limited data is available to build an accurate model and also possibly when a linear model is tried to build using non linear model this situation is called as,	
	Option A	Underfitting
	Option B	overfitting
	Option C	Reinforcement Learning
	Option D	Semi Supervised Learning
	Correct Answer	A
03	Datasets supported by the scikit -learn are	
	Option A	load_iris() Write only iris similar for other optins
	Option B	load_digits()
	Option C	load_heart()
	Option D	A and B both
	Correct Answer	D
04	Pruning is the _____	
	Option A	Deep Learning
	Option B	Methodologies for avoidance of underfitting
	Option C	Methodologies for avoidance of overfitting
	Option D	Semi Supervised Learning
	Correct Answer	C
05	Every machine learning algorithm has three components Representation,Evaluation and Optimization	
	Option A	True
	Option B	False
	Correct Answer	A
06	The range of data is changed in _____	
	Option A	Missing Values
	Option B	Scaling
	Option C	Normalization

Name of UNIT: Feature Selection

	Option D	Transforming
	Correct Answer	B
07	Following steps are used in which type of feature selection algorithm	
	1.Select all Features	
	2.Selecting the best subset	
	3.Learning Algorithm	
	4.Performances	
	Option A	Filter method
	Option B	Wrapper Method
	Option C	Embedded method
	Option D	Normalization
	Correct Answer	A
08	LDA stand for	
	Option A	Linear Domain Analysis
	Option B	Linear Direct Analysis
	Option C	Linear discriminant Analysis
	Option D	Linear Different Analysis
	Correct Answer	C
09	Feature selection Technique (Not understood)	
	Option A	Feature Importance
	Option B	Correlation matrix with Heatmap
	Option C	Univariate Selection
	Option D	All of the above
	Correct Answer	D
10	A linear dimensionality reduction technique is _____ is a linear dimensionality reduction technique.	
	Option A	LASSO
	Option B	PCA
	Option C	RIDGE
	Option D	ANOVA
	Correct Answer	B
11	Dictionary Learning technique used in (Pls check)	
	Option A	Reinforcing Learning

Name of UNIT: Feature Selection

	Option B	Supervised Learning
	Option C	Unsupervised Clustering
	Option D	Semi Supervised Learning
	Correct Answer	C
12	-----is broadly used as a tool for analysis of high dimensional data as it automatically allows to extract sparse and meaningful features from a set of nonnegative data vector.	
	Option A	NMF
	Option B	ANOVA
	Option C	PCA
	Option D	LASSO
	Correct Answer	A

Unit-1 MCQs **ML**

Q.1 Systems that can learn from their experiences and modify their behavior in order to maximize the possibility of reaching a specific goal are referred as _____

1. Classic systems
2. Coherent systems
- 3. Adaptive systems**
4. Automatic systems

Q.2. What is Machine learning?

1 set of techniques that allow implementing classic algorithms to make predictions and to auto-organize input data according to their common features

2. set of techniques that allow implementing classic algorithms to make predictions and to auto-organize input data according to their uncommon features

3. set of techniques that allow implementing classic algorithms to make predictions and to auto-organize input data according to their common features

4. set of techniques that allow implementing adaptive algorithms to make predictions and to auto-organize input data according to their uncommon features

Q.3 If there are n output classes, n classifiers will be trained in parallel considering there is always a separation between an actual class and the remaining ones. This Multiclass strategy is called _____

- 1. One-vs-all**
2. One-vs-one
3. One-vs-none
4. All-vs-none

Unit II

Q.1 Which of the following is option for managing missing features in a dataset?

1. Removing the whole line
2. Creating sub-model to predict those features
3. Using an automatic strategy to input them according to the other known values
- 4. All of the above**

Q.2 While splitting the dataset into the training and test sets the rules to be followed are _____

1. Both datasets must reflect the original distribution and randomly shuffled
2. Both datasets must not reflect the original distribution and are not randomly shuffled
3. Both datasets must of same size and randomly shuffled
4. Both datasets must of different size and are not randomly shuffled

Q.3 In scikit-learn which class is available for handling missing values?

- 1. LabelEncoder
- 2. Imputer**
- 3. RobustScaler
- 4. VarianceThreshold

Q.4 In scikit-learn which class is available for managing categorical data ?

- 1. LabelEncoder**

- 2. Imputer
- 3. RobustScaler
- 4. VarianceThreshold

Q.5 In scikit-learn which class is available for feature selection ?

- 1. LabelEncoder
- 2. Imputer
- 3. RobustScaler
- 4. VarianceThreshold**

Unit III

Q.1 Which regression method is use for a dataset of non-decreasing points which can present low-level oscillations?

- 1. Ridge Regression
- 2. Polynomial regression
- 3. Isotonic regression**
- 4. Lasso Regression

Q.2 In ROC curve _____

- 1. The x axis represents specificity and the y axis represents sensitivity**
- 2. The x axis represents specificity and the y axis represents sensitivity
- 3. The x axis represents precision and the y axis represents recall
- 4. The x axis represents recall and the y axis represents precision

Q.3. The logical operator XOR is an example of

- 1. Linearly separable problem
- 2. Linearly separable and simple problem
- 3. Linearly separable but complex problem
- 4 Non-linearly separable problem**

1.Previous probabilities in Bayes Theorem that are changed with help of new available information are classified as _____ [1 M]

- a) independent probabilities
- b) posterior probabilities**
- c) interior probabilities
- d) dependent probabilities

2. The method in which the previously calculated probabilities are revised with new probabilities is classified as [1 M]

- a] updating theorem
- b] revised theorem
- c] Bayes theorem**
- d]dependency theorem

3. The previous probabilities in Bayes Theorem that are changed with the help of new available information are classified as [1 M]

- a) independent probabilities
- b) posterior probabilities
- c) interior probabilities**
- d) dependent probabilities

4. The model which assumes that all our features are binary such that they take only two values is [1 M]

- a) Multinomial Naïve Bayes
- b) Gaussian Naïve Bayes
- c) Bernoulli Naïve Bayes**
- d) none

5. The effectiveness of an SVM depends upon: [1 M]

- a) Selection of Kernel**
- b) Kernel Parameters
- c) Soft Margin Parameter C
- d) All of the above

6. In Classification Model, Which Technique can help you to choose a threshold that balance sensitivity and specificity [1 M]

- a) Confusion Matrix**
- b) ROC curve**
- c) MAPE
- d) None of the Above

7. In Decision Tree, by comparing the impurity across all possible splits in all possible Predictors, the next split is chosen. How we can measure the Impurity ? [1 M]

- a) UC
- b) Entropy, Ginni-Index**
- c) ROC
- d) MAPE

8. How we can avoid the overfitting in Decision Tree [1 M]

- c) Both of above
- a) CHAID(Stopping the Tree Growth)
- b) Pruning the Full Grown Tree
- d) None of the Above

9. Predictive Errors are due to [1 M]

- a) Bias
- b) Variance
- c) Both of above
- d) None of the Above

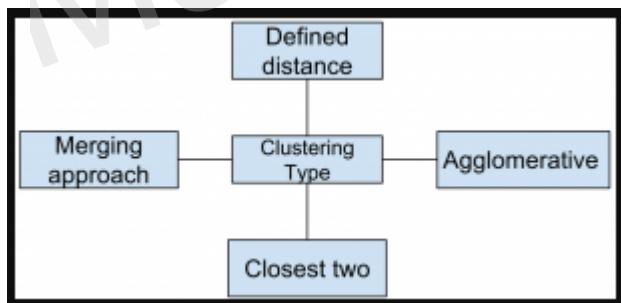
10. Any linear model can be turned into a non-linear model by applying the kernel trick to the model. [1 M]

- a) true
- b) false

11. Random Forest Modeling (Ensemble Modeling) uses . [1 M]

- a) Bagging(BootStrap Samples)
- b) Boosting
- c) Both of above
- d) None of the Above

12. Which of the following clustering type has characteristic shown in the below Figure? [1 M]



- a) Partitional
- b) Hierarchical
- c) Naive bayes
- d) None of the mentioned

Explanation: Hierarchical clustering groups data over a variety of scales by creating a cluster tree or dendrogram.

13. Point out the correct statement. [1 M]

- a) The choice of an appropriate metric will influence the shape of the clusters

- b) Hierarchical clustering is also called HCA
- c) In general, the merges and splits are determined in a greedy manner
- d) All of the mentioned

Explanation: Some elements may be close to one another according to one distance and farther away according to another.

14. Which of the following is finally produced by Hierarchical Clustering? [1 M]

- a) final estimate of cluster centroids
- b) tree showing how close things are to each other
- c) assignment of each point to clusters
- d) all of the mentioned.

Explanation: Hierarchical clustering is an agglomerative approach.

15. Which of the following is required by K-means clustering? [1 M]

- a) defined distance metric
- b) number of clusters
- c) initial guess as to cluster centroids
- d) all of the mentioned

Explanation: K-means clustering follows partitioning approach.

16. Point out the wrong statement. [1 M]

- a) k-means clustering is a method of vector quantization
- b) k-means clustering aims to partition n observations into k clusters
- c) k-nearest neighbor is same as k-means
- d) none of the mentioned

Explanation: k-nearest neighbor has nothing to do with k-means.

17. Which of the following clustering requires merging approach? [2 M]

- a) Partitional
- b) Hierarchical
- c) Naive Bayes
- d) None of the mentioned

Explanation: Hierarchical clustering requires a defined distance as well.

18. a system that is capable of predicting the future preference of a set of items for a user, and recommend the top items. [2 M]

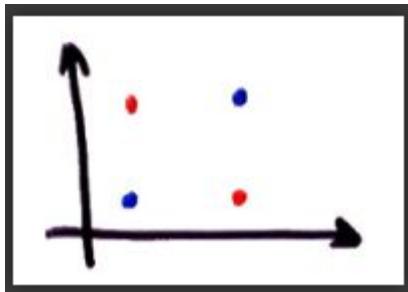
- a) Recommendation Systems
- b) collaborative filtering
- C) Content based Systems
- d) all of the above

19. A content based recommender works with data that the user provides, either explicitly (rating) or implicitly. [2 M]

a) true

b) false

20. Is the data linearly separable? [2M]



a) Yes

b) No

Explanation :If you can draw a line or plane between the data points, it is said to be linearly separable.

21. Which of the following are universal approximators? [2M]

a) Kernel SVM

b) Neural Networks

c) Boosted Decision Trees

d) All of the above

Explanation :All of the above methods can approximate any function.

22. Decision Tree is a display of an algorithm. [2M]

a) True

b) False

23. A _____ is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. [2M]

a) Decision tree

b) Graphs

c) Trees

d) Neural Networks

24. Choose from the following that are Decision Tree nodes? [2M]

a) Decision Nodes

b) End Nodes

c) Chance Nodes

d) All of the mentioned

25. . Which of the following is/are true about bagging trees? [2M]

1. In bagging trees, individual trees are independent of each other
 2. Bagging is the method for improving the performance by aggregating the results of weak learners
- a) 1
b) 2
c) 1 and 2
d) None of these

Explanation: Both options are true. In Bagging, each individual trees are independent of each other because they consider different subset of features and samples.

26 . Which of the following is/are true about Random Forest and Gradient Boosting ensemble methods? [2M]

1. Both methods can be used for classification task
 2. Random Forest is use for classification whereas Gradient Boosting is use for regression task
 3. Random Forest is use for regression whereas Gradient Boosting is use for Classification task
 4. Both methods can be used for regression task
- a) 1
b) 2
c) 3
d) 1&4

Explanation: Both algorithms are design for classification as well as regression task

27. Which of the following algorithm doesn't uses learning Rate as of one of its hyperparameter? [2M]

1. Gradient Boosting
 2. Extra Trees
 3. AdaBoost
 4. Random Forest
- a) 1 and 3
b) 1 and 4
c) 2 and 3
d) 2 and 4

Explanation: Random Forest and Extra Trees don't have learning rate as a hyperparameter

28. Which of the following algorithm are not an example of ensemble learning algorithm?

- a) Random Forest
b) Adaboost
c) Extra Trees
d) Decision Trees

Explanation: Decision trees doesn't aggregate the results of multiple trees so it is not an ensemble algorithm

29. Which of the following splitting point on feature x1 will classify the data correctly? [4]

- a) Greater than x11
- b) Less than x11
- c) Equal to x11
- d) None of above

Explanation: If you search any point on X1 you won't find any point that gives 100% accuracy

30. What will be the minimum accuracy you can get? [4]

- a) Always greater than 70%
- b) Always greater than and equal to 70%
- c) It can be less than 70%
- d) None of these

Explanation: Refer below table for models M1, M2 and M3.

Actual predictions	M1	M2	M3	Output
1	1	0	0	0
1	1	1	1	1
1	1	0	0	0
1	0	1	0	0
1	0	1	1	1
1	0	0	1	0
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1

Question	a	b	c	d	Answer	
In reinforcement learning if feedback is negative one it is defined as	a. Penalty	b. Overlearning	c. Reward	d. None of the above	A	
According to, it's a key success factor for the survival and evolution of all species.	Claude Shannon's theory	Gini Index	Darwin's theory	None of the above	C	
How can you avoid overfitting ?	By using a lot of data	By using inductive machine learning	By using validation only	None of the above	A	
What are the popular algorithms of Machine Learning?	Decision Trees and Neural Networks (back propagation)	Probabilistic networks and Nearest Neighbor	Support vector machines	All	D	
What is 't Training set'?	Training set is used to test the accuracy of the hypotheses generated by the learner.	A set of data is used to discover the potentially predictive relationship.	Both A & B	None of the above	B	
Common deep learning applications include	Image classification,	Autonomous car driving,	Bioinformatics,	All above	D	
what is the function of 't Supervised Learning'?	Classifications, Predict time series, Annotate strings	Speech recognition, Regression	Both A & B	None of the above	C	
Common unsupervised applications include	Object segmentation	Similarity detection	Automatic labeling	All above	D	
Reinforcement learning is particularly efficient when	the environment is not completely deterministic	it's often very dynamic	it's impossible to have a precise error measure	All above	D	
If there are many categories of outcomes (called categories),	Regression	Classification.	Model selection	Categories	B	
Which of the following are supervised learning applications?	Sparse detection,	image classification,	Autonomous car driving,	Bioinformatics,	A	
During the last few years, many	Logical	Classical	Classification	None of the above	D	
Which of the following sentence is correct?	Machine learning relates with the study, design and development of the algorithms that	Machine learning can be defined as the process in which the unstructured data tries	Robots are programmed so that they can perform the task based on data they get	While involving the process of learning 't overfitting' occurs	C	
What is 't Overfitting' in Machine learning?	Test set is used to test the accuracy of the hypotheses generated by the learner.	It is a set of data is used to discover the potentially predictive relationship.	Both A & B	A set of data is used to discover the potentially predictive relationship	A	
What is 't Test set'?	is much more difficult because it's necessary to determine a supervised strategy to train a model for each feature and, finally, to predict	Creating sub-model to predict those features	Using an automatic strategy to input them according to the other known values	None of the above	B	
How it's possible to use a different placeholder through the parameter	removing the whole line regression	classification	random_state	missing values	D	
If you need a more powerful scaling feature, with a superior control on outliers and the possibility to select a quantile range, there's also the class RobustScaler	RobustScaler	LabelBinarizer	FeatureFisher	Independent Component Analysis	A	
scikit-learn also provides a class for per-sample normalization, Normalizer. It can apply	max, I1 and I2 norms	max, I1 and I2 norms	max, I2 and I3 norms	max, I3 and I4 norms	B	
There are also many univariate methods that can be used in order to select the best features according to specific criteria based on	F-tests and p-values	chi-square	ANOVA	All above	A	
Which of the following selects only a subset of features belonging to a certain percentile	SelectPercentile	FeatureHasher	SelectKBest	All above	A	
..... performs a PCA with non-linearly separable data sets	SparsePCA	KernelPCA	SVD	None of the Mentioned	B	
A feature F1 can take certain value A, B, C, D, & F and represents grade of students from a college.	Feature F1 is an example of nominal variable.	Feature F1 is an example of ordinal variable.	It doesn't belong to any of the above category.	Both of these	B	
What would you do in PCA to get the same projection as SVD?	Transform data to zero mean	Transform data to zero median	Not possible	None of these	A	
What is PCA, KPCA and LDA used for?	Principal Components Analysis	Kernel based Principal Component Analysis	Independent Component Analysis	All above	D	
Can we combine linear based similarity also choose from a given set of items?	YES	NO	YES	NO	A	
What are common feature selection methods in regression task?	correlation coefficient	Greedy algorithms	All above	None of these	C	
The parameter	test_size	training_size	All above	None of these	C	
In many classification problems, the target	dataset	test_size	All above	None of these	B	
adopts a dictionary-oriented approach, associating to each category label a progressive integer number.	LabelEncoder class	DictVectorizer	FeatureFisher	Independent Component Analysis	A	
If Linear regression model perfectly fit i.e., train error is zero, then	a) Test error is also always zero	b) Test error is not zero	c) Couldn't comment on Test error	d) Test error is equal to Train error	C	
Which of the following metrics can be used for evaluating regression models? (R Squared) Adjusted R Squared(FI) F Statistic(FI) RMSE / MSE / MAE	a) i and iv	b) i and ii	c) ii, iii and iv	d) i, ii, iii and iv	d	
How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?	a) 1	b) 2	c) 3	d) 4	b	
In a simple linear regression model (One independent variable), if we change the input variable by 1 unit. How much output variable will change?	a) by 1	b) no change	c) by intercept	d) by its slope	d	
Function used for linear regression in R is	a) lm(formula, data)	b) lr(formula, data)	c) lm(formula, data)	d) regression.linear(formula, data)	a	
In syntax of linear model lm(formula, data), data refers to	a) Matrix	b) Vector	c) Array	d) List	b	
In the mathematical Equation of Linear Regression $Y_{\hat{A}} = \hat{A}_0 + \hat{A}_1 X_1 + \hat{A}_2 X_2 + \dots$ refers to _____	(X-Intercept, Slope)	(Slope, X-Intercept)	(Y-Intercept, Slope)	(slope, Y-Intercept)	c	
Linear Regression is a supervised machine learning algorithm.	a) TRUE	b) FALSE	c) UNKNOWN	d) FALSE	a	
It is possible to design a Linear regression algorithm using a neural network?	a) TRUE	b) FALSE	c) UNKNOWN	d) FALSE	a	
Which of the following methods do we use to find the best fit line for Data in Linear Regression?	a) Least Square Error	b) Maximum Likelihood	c) Logarithmic Loss	d) Both A and B	a	
Which of the following information must be used to evaluate a model while modeling a continuous output variable?	a) AUC-ROC	b) Accuracy	c) LogLoss	d) Mean-Squared-Error	d	
Which of the following is true about Residuals ?	a) Lower is better	b) Higher is better	c) AA or B depend on the situation	d) None of these	b	
Overfitting is more likely when you have huge amount of data to train?	a) TRUE	b) FALSE	c) UNKNOWN	d) UNKNOWN	b	
Which of the following statement is true about outliers in Linear regression?	a) Linear regression is sensitive to outliers	b) Linear regression is not sensitive to outliers	c) Can't say	d) None of these	a	
Suppose you plotted a scatter plot between the residuals and predicted values in linear regression and you found that there is a relationship bet	a) AA Since there is a relationship means our model is good	b) BA Since there is a relationship means our model is good	c) CA Can't say	d) DA None of these	a	
Naive Bayes classifiers are a collection of algorithms	Classification	Clustering	Regression	All	a	
Naive Bayes classifiers is	Learning	Supervised	Both	None	a	
Features being classified is independent of each other in Na've Bayes Classifier	FALSE	TRUE	1	0	b	
Features being classified is	of each other in Na've Bayes Classifier	Independent	Dependent	Partial Dependent	None	a
Bayes Theorem is given by where $P(H E)$ is the probability of hypothesis H being true.	TRUE	TRUE	0	0	a	
In given image, $P(H E)$ is probability.	Posterior	Prior	Posterior	Prior	a	
In given image, $P(H E)$ is probability.	Posterior	Prior	Posterior	Prior	b	
Conditional probability is a measure of probability of an event given that another event has already occurred.	TRUE	TRUE	0	0	a	
Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event.	TRUE	TRUE	0	0	a	
Bernoulli Na've Bayes Classifier is distribution	Continuous	Discrete	Binary	Both A and B	c	
Multinomial Na've Bayes Classifier is distribution	Continuous	Discrete	Binary	AA or B	b	
Gaussian Na've Bayes Classifier is distribution	Continuous	Discrete	Binary	BA	b	
Binarize parameter in Bernoulli's scikit sets threshold for binarizing of sample features.	TRUE	TRUE	0	0	a	
Gaussian distribution when plotted, elvls a bell shaped curve which is symmetric about the	of the feature values.	Mean	Variance	Discrete	Random	a
SVMS directly give us the posterior probabilities $P(Y=1 x)$ and $P(Y=0 x)$	TRUE	TRUE	0	0	b	
Any linear combination of the components of a multivariate Gaussian is a univariate Gaussian.	TRUE	TRUE	0	0	a	
Solving a non linear separation problem with a hard margin Kernelized SVM (Gaussian RBF Kernel) might lead to overfitting	TRUE	TRUE	0	0	a	
SVM is algorithm	Classification	Clustering	Regression	All	a	
SVM is learning	Supervised	Unsupervised	Both	None	a	
The linear' SVM classifier works by drawing a straight line between two classes	TRUE	TRUE	0	0	a	
Which of the following function provides unsupervised prediction ?	cl_forecastB	cl_forecastC	cl_forecastD	None of the Mentioned	D	
Which of the following is characteristic of best machine learning method ?	fast	accuracy	scalable	All above	D	
What are the different Algorithm techniques in Machine Learning?	Supervised Learning and Semi-supervised Learning	Unsupervised Learning and Transduction	Both A & B	None of the Mentioned	C	
What is the standard term for supervised learning?	split the set of example into the training set and the test	group the set of example into the training set and the test	a set of observed instances tries to induce a general rule	learns programs from data	A	
What is the difference between AI and Machine Learning?	Artificial Intelligence	Machine Learning	Both A & B	None of the Mentioned	B	
What is Model Selection in Machine Learning?	The process of selecting models among different mathematical models, which are used when a statistical model describes random error or noise instead of underlying	when a statistical model describes random error or noise instead of underlying	find interesting directions in data and find novel observations/ database cleaning	All above	A	
Which are two techniques of Machine Learning?	Genetic Programming and Speech recognition and Regression	Both A & B	None of the Mentioned	A		
Even if there are no actual supervisors learning is also based on feedback provided by the environment	Supervised	Reinforcement	Unsupervised	None of the above	B	
What does learning exactly mean?	Robots are programmed so that they can perform the task based on data they gather from a set of data is used to discover the potentially predictive relationship.	Learning is the ability to change according to external stimuli and remembering most of a set of data is used to discover the potentially predictive relationship.	Learn is the ability to change according to external stimuli and remembering most of a set of data is used to discover the potentially predictive relationship.	It is a set of data is used to discover the potentially predictive relationship.	C	
When it is necessary to allow the model to develop a generalization ability and avoid a common problem called	Overfitting	Overfitting	Classification	Regression	A	
Techniques involve the usage of both labeled and unlabeled data is called	Supervised	Semi-supervised	Unsupervised	None of the above	B	
A supervised scenario is characterized by the concept of a	Programmer	Teacher	Author	Farmer	B	
overlearning causes due to an excessive	Capacity	Regression	Reinforcement	Accuracy	A	
Which of the following is an example of a deterministic algorithm?	PCA	K-Means	None of the above	None of the above	A	
Which of the following model include a backwards elimination feature selection routine?	MCV	MARS	MCRS	All above	B	
Can we extract knowledge without apply feature selection?	YES	NO	YES	NO	A	
While using feature selection on the data, is the number of features decreases.	NO	YES	YES	NO	B	
Which of the following are several models for feature extraction	regression	classification	None of the above	None of the above	C	
..... provides some built-in datasets that can be used for testing purposes.	scikit-learn	classification	regression	None of the above	A	
While using feature selection models are	LabelEncoder class	LabelBinarizer class	DictVectorizer	FeatureFisher	A	
..... produce sparse matrix of real numbers that can be fed into any machine learning model.	run	start	init	stop	C	
scikit-learn offers the class, which is responsible for filling the holes using a strategy based on the mean, median, or frequency	DictVectorizer	LabelBinarizer	DictVectorizer	None of the Mentioned	C	
Which of the following scale data by removing elements that don't belong to a given range or by considering a maximum absolute value.	MinMaxScaler	MaxAbsScaler	Both A & B	None of the Mentioned	C	
scikit-learn also provides a class for per-sample normalization.	Normalizer	Imputer	Classifier	All above	A	
..... dataset with many features contains information proportional to the independence of all features and their variance.	normalized	unnormalized	Both A & B	None of the Mentioned	B	
In order to assess how much information is brought by each component, and the correlation among them, a useful tool is the	Concurrent matrix	Convergence matrix	Supportive matrix	Covariance matrix	D	
The parameter can assume different values which determine how the data matrix is initially processed.	run	start	init	stop	C	
allows exploiting the natural sparsity of data while extracting principal components.	SparsePCA	KernelPCA	SVD	Init parameter	A	
Which of the following evaluation metrics can be used to evaluate a model while modeling a continuous output variable?	AUC-ROC	Accuracy	LogLoss	Mean-Squared-Error	D	
Which of the following is true about residuals ?	Lower is better	Higher is better	A or B depend on the situation	None of these	A	
Oversampling is more likely when you have huge amount of data to train?	1	0	0	0	B	
Which of the following statement is true about outliers in Linear regression?	Linear regression is sensitive to outliers	Linear regression is not sensitive to outliers	Can't say	None of these	A	
Suppose you plotted a scatter plot between the residuals and predicted values in linear regression and you found that there is a relationship bet	Since the there is a relationship means our model is good	Since the there is a relationship means our model is good	Can't say	None of these	A	
Let't say, a linear regression model perfectly fits the training data (train error is zero). Now, Which of the following statement is true?	You will always have test error zero	You can not have test error zero	None of the above	None of these	C	
In a linear regression problem, we are using adj-R-squared to measure goodness-of-fit. We add a feature in linear regression model and retr	adj-R-squared increases, then it means the new feature is significant	adj-R-squared decreases, this variable is not significant.	Individually R squared cannot tell about variable importance. We can't say anything ab	None of these.	C	

Question	a	b	c	d	Answer
SVM algorithms use a set of mathematical functions that are defined as the kernel.		1	0		A
Naive Bayes classifiers are a collection of algorithms	Classification	Clustering	Regression	All	A
In given image, $P(H E)$ is probability.	Posterior	Prior			A
Solving a non linear separation problem with a hard margin Kernelized SVM (Gaussian RBF Kernel) might lead to overfitting.	TRUE		0		A
100 people are at party. Given data gives information about how many wear pink or not, and if a man or not. Imagine		1	0		A
For the given weather data, Calculate probability of playing		0.4	0.64	0.29	0.75 B
In SVM, Kernel function is used to map a lower dimensional data into a higher dimensional data.		1	0		A
In SVR we try to fit the error within a certain threshold.		1	0		A
Let's say, you are working with categorical feature(s) and you have not looked at the distribution of the	All categories of categorical variable are not present in the	Frequency distribution of categories is different in train	Train and Test always have same distribution.	Both A and B	D
Which of the following sentence is FALSE regarding regression?	It relates inputs to outputs.	It is used for prediction.	It may be used for interpretation.	It discovers causal relationships.	D
Which of the following method is used to find the optimal features for cluster analysis	k-Means	Density-Based Spatial Clustering	Spectral Clustering Find clusters	All above	D
scikit-learn also provides functions for creating	make_classification()	make_regression()	make_blobs()	All above	D
which can accept a NumPy RandomState generator or an integer seed.	make_blobs	random_state	test_size	training_size	B
In many classification problems, the target dataset is made up of categorical labels which cannot immediately be processed.		1	2	3	4 B
In which of the following each categorical label is first turned into a positive integer and then transformed into a vector	LabelEncoder class	DictVectorizer	LabelBinarizer class	FeatureHasher	C
is the most drastic one and should be considered only when the dataset is quite large, the number of missing features	Removing the whole line	Creating sub-model to predict those features	Using an automatic strategy to input them according to	All above	A
It's possible to specify if the scaling process must include both mean and standard deviation using the parameters	with_mean=True/False	with_std=True/False	Both A & B	None of the Mentioned	C
Which of the following selects the best K high-score features.	SelectPercentile	FeatureHasher	SelectKBest	All above	C
How does number of observations influence overfitting? Choose the correct answer(s). Note: Rest all parameters are same.	1 and 4	2 and 3	1 and 3	None of theses	A
Suppose you have fitted a complex regression model on a dataset. Now, you are using Ridge regression with tuning parameter lambda.	In case of very large lambda; bias is low, variance is high	In case of very large lambda; bias is low, variance is high	In case of very large lambda; bias is high, variance is low	In case of very large lambda; bias is high, variance is high	C
What is/are true about ridge regression? 1. When lambda is 0, model works like linear regression model2. When lambda is 1 and 3	1 and 4	2 and 3	2 and 4	2 and 4	A
Which of the following method(s) does not have closed form solution for its coefficients?	Ridge regression	Lasso	Both Ridge and Lasso	None of both	B
Function used for linear regression in R is	lm(formula, data)	lmr(formula, data)	regression.linear(formula, data)		A
In the mathematical Equation of Linear Regression $\hat{Y}_{i,j} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$, (β_1, β_2) refers to	(X-intercept, Slope)	(Slope, X-Intercept)	(Y-Intercept, Slope)	(slope, Y-Intercept)	C
Suppose that we have N independent variables (X_1, X_2, \dots, X_n) and dependent variable is Y. Now Imagine that you are	Relation between the X_1 and Y is weak	Relation between the X_1 and Y is strong	Relation between the X_1 and Y is neutral	Correlation can't judge the relationship	B
We have been given a dataset with n records in which we have input attribute as x and output attribute as y. Suppose	Increase	Decrease	Remain constant	Can't Say	D
We have been given a dataset with n records in which we have input attribute as x and output attribute as y. Suppose	Bias increases and Variance increases	Bias decreases and Variance increases	Bias decreases and Variance decreases	Bias increases and Variance decreases	D
Suppose, you got a situation where you find that your linear regression model is under fitting the data. In such situation	1 and 2	2 and 3	1 and 3	1, 2 and 3	A
Problem: A Players will play if weather is sunny. Is this statement is correct?		1	0		A
Multinomial Naïve Bayes Classifier is	distribution	Continuous	Discrete	Binary	B
For the given weather data, Calculate probability of not playing		0.4	0.64	0.36	0.5 C
Suppose you have trained an SVM with linear decision boundary after training SVM, you correctly infer that your SVM	You want to increase your data points	You want to decrease your data points	You will try to calculate more variables	You will try to reduce the features	C
The minimum time complexity for training an SVM is O(n^2). According to this fact, what sizes of datasets are not best:	Large datasets	Small datasets	Medium sized datasets	Size does not matter	A
The effectiveness of an SVM depends upon:	Selection of Kernel	Kernel Parameters	Soft Margin Parameter C	All of the above	D
What do you mean by generalization error in terms of the SVM?	How far the hyperplane is from the support vectors	How accurately the SVM can predict outcomes for unseen	The threshold amount of error in an SVM		B
What do you mean by a hard margin?	The SVM allows very low error in classification	The SVM allows high amount of error in classification	None of the above		A
We usually use feature normalization before using the Gaussian kernel in SVM. What is true about feature normalization?		1 and 2	1 and 3	2 and 3	B
Support vectors are the data points that lie closest to the decision surface.		1	0		A
Which of the following is not a supervised learning?	PCA	Decision Tree	Naive Bayesian	Linear regression	A
Suppose you are using RBF kernel in SVM with high Gamma value. What does this signify?	The model would consider even far away points from hyper	The model would consider only the points close to the	The model would not be affected by distance of points	None of the above	B
Gaussian Naïve Bayes Classifier is	distribution	Continuous	Discrete	Binary	A
If I am using all features of my dataset and I achieve 100% accuracy on my training set, but ~70% on validation set, what is the purpose of performing cross-validation?	Underfitting	Nothing, the model is perfect	Overfitting		C
What is the purpose of performing cross-validation?	a. To assess the predictive performance of the models	b. To judge how the trained model performs outside the	c. Both A and B		C
Which of the following is true about Naive Bayes ?	a. Assumes that all the features in a dataset are equally important	b. Assumes that all the features in a dataset are independent	c. Both A and B	d. None of the above option	C
Suppose you are using a Linear SVM classifier with 2 class classification problem. Now you have been given the following	yes	no			A
Linear SVMs have no hyperparameters that need to be set by cross-validation		1	0		B
For the given weather data, what is the probability that players will play if weather is sunny		0.5	0.26	0.73	0.6 D
100 people are at party. Given data gives information about how many wear pink or not, and if a man or not. Imagine		0.4	0.2	0.6	0.45 B
Problem: A Players will play if weather is sunny. Is this statement is correct?		1	0		A
For the given weather data, Calculate probability of playing		0.4	0.64	0.29	0.75 b
For the given weather data, Calculate probability of not playing		0.4	0.64	0.36	0.5 c
For the given weather data, what is the probability that players will play if weather is sunny		0.5	0.26	0.73	0.6 d
100 people are at party. Given data gives information about how many wear pink or not, and if a man or not. Imagine		0.4	0.2	0.6	0.45 b
100 people are at party. Given data gives information about how many wear pink or not, and if a man or not. Imagine		1	0		a
What do you mean by generalization error in terms of the SVM?	How far the hyperplane is from the support vectors	How accurately the SVM can predict outcomes for unseen	The threshold amount of error in an SVM		b
What do you mean by a hard margin?	The SVM allows very low error in classification	The SVM allows high amount of error in classification	None of the above		a
The minimum time complexity for training an SVM is O(n^2). According to this fact, what sizes of datasets are not best:	Large datasets	Small datasets	Medium sized datasets	Size does not matter	a
The effectiveness of an SVM depends upon:	Selection of Kernel	Kernel Parameters	Soft Margin Parameter C	All of the above	d
Support vectors are the data points that lie closest to the decision surface.		1	0		a
The SVMs are less effective when:	The data is linearly separable	The data is clean and ready to use	The data is noisy and contains overlapping points		c
Suppose you are using RBF kernel in SVM with high Gamma value. What does this signify?	The model would consider even far away points from hyper	The model would consider only the points close to the	The model would not be affected by distance of points	None of the above	b
The cost parameter in the SVM means:	The number of cross-validations to be made	The kernel to be used	The tradeoff between misclassification and simplicity of	None of the above	c
If I am using all features of my dataset and I achieve 100% accuracy on my training set, but ~70% on validation set, what is the purpose of performing cross-validation?	Underfitting	Nothing, the model is perfect	Overfitting		c
Which of the following are real world applications of the SVM?	Text and Hypertext Categorization	Image Classification	Clustering of News Articles	All of the above	d
Suppose you have trained an SVM with linear decision boundary after training SVM, you correctly infer that your SVM	You want to increase your data points	You want to decrease your data points	You will try to calculate more variables	You will try to reduce the features	c
We usually use feature normalization before using the Gaussian kernel in SVM. What is true about feature normalization?		1 and 2	1 and 3	2 and 3	b
Linear SVMs have no hyperparameters that need to be set by cross-validation		1	0		b
In a real problem, you should check to see if the SVM is separable and then include slack variables if it is not separable		1	0		b
What is Machine learning?	The autonomous acquisition of knowledge through the use	The autonomous acquisition of knowledge through the	The selective acquisition of knowledge through the use	The selective acquisition of knowledge through the use of	a
Movie Recommendation systems are an example of:	2 and 3	1 and 3	all of the mentioned		b
Sentiment Analysis is an example of: 1. Regression 2. Classification 3. Clustering	1,2,4	1,2	1,2,3		a
Which of the factors affect the performance of learner system does not include?	Representation scheme used	Training scenario	Type of feedback	Good data structures	d
Point out the correct statement.	Machine learning focuses on prediction, based on known prior	Data Cleaning focuses on prediction, based on known prior	Representing data in a form which both mere mortals	None of the mentioned	d

The problem of finding hidden structure in unlabeled data is called	Supervised learning	Unsupervised learning	Reinforcement learning	None of these	b
which of the following is not involve in data mining?	Knowledge extraction	Data archaeology	Data exploration	Data transformation	d
Which of the following is one of the key data science skills?	Statistics	Machine Learning	Data visualization	All of the above	d
Raw data should be processed only one time.		1	0		b
Which of the following is characteristic of best machine learning method ?	Fast	Accuracy	Scalable	All of the Mentioned	d
True-False: Linear Regression is a supervised machine learning algorithm.	TRUE		0		a
Linear Regression is mainly used for Regression.		1	0		b
Overfitting is more likely when you have huge amount of data to train?		1	0		
Supervised learning and unsupervised clustering both require which is correct according to the statement.	output attribute.	hidden attribute.	input attribute.	categorical attribute	b
Methodologies used for avoiding overfitting problems are	1.Cross- Validation 2.Pruning 3. Early stopping 4. Regularization	1and 2	2and 4	1,2and 4	d
Some telecommunication company wants to segment their customers into distinct groups in order to send appropriate offers.	Supervised learning	Data extraction	Serration	Unsupervised learning	d
Self-organizing maps are an example of	Supervised learning	Unsupervised learning	Reinforcement learning	Missing data imputation	b
You are given data about seismic activity in Japan, and you want to predict a magnitude of the next earthquake, this is	Supervised learning	Unsupervised learning	Serration	Dimensionality reduction	a
Adaptive system management is	It uses machine-learning techniques. Here program can learn from experience.	Computational procedure that takes some values as input	Science of making machines performs tasks that would	none of these	a
Background knowledge referred to	Additional acquaintance used by a learning algorithm to facilitate learning	A neural network that makes use of a hidden layer	It is a form of automatic learning	None of these	a
Data independence means	Data is defined separately and not included in programs	Programs are not dependent on the physical attributes	Programs are not dependent on the logical attributes of	Both (B) and (C)	d
Classification is	A subdivision of a set of examples into a number of classes	A measure of the accuracy, of the classification of a collection of examples	The task of assigning a classification to a set of examples	None of these	a
It is possible to design a Linear regression algorithm using a neural network?		1	0		a
Methodologies used for avoiding overfitting problems are	1.Cross- Validation 2.Evaluation 3. Early stopping 4. Regularization	1and 2	2and 3	1,4	d
Methodologies used for avoiding overfitting problems are	1.Cross- Validation 2.Evaluation 3. Early stopping 4. Regularization	1and 2	1 and 3	1,4	b
Which of the following Methodologies is not used for avoiding overfitting problems	1.Cross- Validation 2.Pruning		1	2	d
Which of the following Methodologies is not used for avoiding overfitting problems	1.Cross- Validation 2.Pruning		1	2	c
CNN and RNN stands for 1. content Neural Network and Robust Neural Network 2. circular Neural Network And reverse		1	2	3	c
which of the following problem(s) is/are not solved by supervised learning 1.regression 2.classification 3. association 4. clustering	1and 2	3and 4	1and 4	2and 3	b
Reinforcement learning is	1.The dataset is properly labeled, meaning, a set of data is provided to train the algorithm.	1	2	3	c
In the example of predicting number of babies based on storksâ€™ population size, number of babies is	outcome	feature	attribute	observation	a
Assume you want to perform supervised learning and to predict number of newborns according to size of storksâ€™ population	Classification	Regression	Clustering	Structural equation modeling	b

UNIT I

1. What is classification?
 - a) when the output variable is a **category**, such as “red” or “blue” or “disease” and “no disease”.
 - b) when the output variable is a **real value**, such as “dollars” or “weight”.

Ans: Solution A

2. What is regression?
 - a) When the output variable is a category, such as “red” or “blue” or “disease” and “no disease”.
 - b) When the output variable is a real value, such as “dollars” or “weight”.

Ans: Solution B

3. What is supervised learning?
 - a) All data is unlabelled and the algorithms learn to inherent structure from the input data
 - b) All data is labelled and the algorithms learn to predict the output from the input data
 - c) It is a framework for learning where an agent interacts with an environment and receives a reward for each interaction
 - d) Some data is labelled but most of it is unlabelled and a mixture of supervised and unsupervised techniques can be used.

Ans: Solution B

4. What is Unsupervised learning?
 - a) All data is unlabelled and the algorithms learn to inherent structure from the input data
 - b) All data is labelled and the algorithms learn to predict the output from the input data
 - c) It is a framework for learning where an agent interacts with an environment and receives a reward for each interaction
 - d) Some data is labelled but most of it is unlabelled and a mixture of supervised and unsupervised techniques can be used.

Ans: Solution A

5. What is Semi-Supervised learning?
 - a) All data is unlabelled and the algorithms learn to inherent structure from the input data
 - b) All data is labelled and the algorithms learn to predict the output from the input data
 - c) It is a framework for learning where an agent interacts with an environment and receives a reward for each interaction
 - d) Some data is labelled but most of it is unlabelled and a mixture of supervised and unsupervised techniques can be used.

Ans: Solution D

6. What is Reinforcement learning?
 - a) All data is unlabelled and the algorithms learn to inherent structure from the input data
 - b) All data is labelled and the algorithms learn to predict the output from the input data
 - c) It is a framework for learning where an agent interacts with an environment and receives a reward for each interaction
 - d) Some data is labelled but most of it is unlabelled and a mixture of supervised and unsupervised techniques can be used.

Ans: Solution C

7. Sentiment Analysis is an example of:

Regression,

Classification

Clustering

Reinforcement Learning

Options:

- A. 1 Only
- B. 1 and 2
- C. 1 and 3
- D. 1, 2 and 4

Ans : Solution D

8. The process of forming general concept definitions from examples of concepts to be learned.
 - a) Deduction
 - b) abduction
 - c) induction
 - d) conjunction

Ans : Solution C

9. Computers are best at learning
 - a) facts.
 - b) concepts.
 - c) procedures.
 - d) principles.

Ans : Solution A

10. Data used to build a data mining model.

- a) validation data
- b) training data
- c) test data
- d) hidden data

Ans : Solution B

11. Supervised learning and unsupervised clustering both require at least one

- a) hidden attribute.
- b) output attribute.
- c) input attribute.
- d) categorical attribute.

Ans : Solution A

12. Supervised learning differs from unsupervised clustering in that supervised learning requires

- a) at least one input attribute.
- b) input attributes to be categorical.
- c) at least one output attribute.
- d) output attributes to be categorical.

Ans : Solution B

13. A regression model in which more than one independent variable is used to predict the dependent variable is called

- a) a simple linear regression model
- b) a multiple regression models
- c) an independent model
- d) none of the above

Ans : Solution C

14. A term used to describe the case when the independent variables in a multiple regression model are correlated is

- a) Regression
- b) correlation
- c) multicollinearity
- d) none of the above

Ans : Solution C

15. A multiple regression model has the form: $y = 2 + 3x_1 + 4x_2$. As x_1 increases by 1 unit (holding x_2 constant), y will

- a) increase by 3 units
- b) decrease by 3 units
- c) increase by 4 units
- d) decrease by 4 units

Ans : Solution C

16. A multiple regression model has

- a) only one independent variable
- b) more than one dependent variable
- c) more than one independent variable
- d) none of the above

Ans : Solution B

17. A measure of goodness of fit for the estimated regression equation is the

- a) multiple coefficient of determination
- b) mean square due to error
- c) mean square due to regression
- d) none of the above

Ans : Solution C

18. The adjusted multiple coefficient of determination accounts for

- a) the number of dependent variables in the model
- b) the number of independent variables in the model
- c) unusually large predictors
- d) none of the above

Ans : Solution D

19. The multiple coefficient of determination is computed by

- a) dividing SSR by SST
- b) dividing SST by SSR
- c) dividing SST by SSE
- d) none of the above

Ans : Solution C

20. For a multiple regression model, $SST = 200$ and $SSE = 50$. The multiple coefficient of determination is

- a) 0.25

- b) 4.00
- c) 0.75
- d) none of the above

Ans : Solution B

21. A nearest neighbor approach is best used
- a) with large-sized datasets.
 - b) when irrelevant attributes have been removed from the data.
 - c) when a generalized model of the data is desirable.
 - d) when an explanation of what has been found is of primary importance.

Ans : Solution B

22. Another name for an output attribute.
- a) predictive variable
 - b) independent variable
 - c) estimated variable
 - d) dependent variable

Ans : Solution B

23. Classification problems are distinguished from estimation problems in that
- a) classification problems require the output attribute to be numeric.
 - b) classification problems require the output attribute to be categorical.
 - c) classification problems do not allow an output attribute.
 - d) classification problems are designed to predict future outcome.

Ans : Solution C

24. Which statement is true about prediction problems?
- a) The output attribute must be categorical.
 - b) The output attribute must be numeric.
 - c) The resultant model is designed to determine future outcomes.
 - d) The resultant model is designed to classify current behavior.

Ans : Solution D

25. Which statement about outliers is true?
- a) Outliers should be identified and removed from a dataset.
 - b) Outliers should be part of the training dataset but should not be present in the test data.
 - c) Outliers should be part of the test dataset but should not be present in the training data.
 - d) The nature of the problem determines how outliers are used.

Ans : Solution D

26. Which statement is true about neural network and linear regression models?
- a) Both models require input attributes to be numeric.
 - b) Both models require numeric attributes to range between 0 and 1.
 - c) The output of both models is a categorical attribute value.
 - d) Both techniques build models whose output is determined by a linear sum of weighted input attribute values.

Ans : Solution A

27. Which of the following is a common use of unsupervised clustering?
- a) detect outliers
 - b) determine a best set of input attributes for supervised learning
 - c) evaluate the likely performance of a supervised learner model
 - d) determine if meaningful relationships can be found in a dataset

Ans : Solution A

28. The average positive difference between computed and desired outcome values.
- a) root mean squared error
 - b) mean squared error
 - c) mean absolute error
 - d) mean positive error

Ans : Solution D

29. Selecting data so as to assure that each class is properly represented in both the training and test set.
- a) cross validation
 - b) stratification
 - c) verification
 - d) bootstrapping

Ans : Solution B

30. The standard error is defined as the square root of this computation.
- a) The sample variance divided by the total number of sample instances.
 - b) The population variance divided by the total number of sample instances.
 - c) The sample variance divided by the sample mean.
 - d) The population variance divided by the sample mean.

Ans : Solution A

31. Data used to optimize the parameter settings of a supervised learner model.

- a) Training
- b) Test
- c) Verification
- d) Validation

Ans : Solution D

32. Bootstrapping allows us to

- a) choose the same training instance several times.
- b) choose the same test set instance several times.
- c) build models with alternative subsets of the training data several times.
- d) test a model with alternative subsets of the test data several times.

Ans : Solution A

33. The correlation between the number of years an employee has worked for a company and the salary of the employee is 0.75. What can be said about employee salary and years worked?

- a) There is no relationship between salary and years worked.
- b) Individuals that have worked for the company the longest have higher salaries.
- c) Individuals that have worked for the company the longest have lower salaries.
- d) The majority of employees have been with the company a long time.
- e) The majority of employees have been with the company a short period of time.

Ans : Solution B

34. The correlation coefficient for two real-valued attributes is -0.85 . What does this value tell you?

- a) The attributes are not linearly related.
- b) As the value of one attribute increases the value of the second attribute also increases.
- c) As the value of one attribute decreases the value of the second attribute increases.
- d) The attributes show a curvilinear relationship.

Ans : Solution C

35. The average squared difference between classifier predicted output and actual output.

- a) mean squared error
- b) root mean squared error
- c) mean absolute error
- d) mean relative error

Ans : Solution A

36. Simple regression assumes a _____ relationship between the input attribute and output attribute.

- a) Linear

- b) Quadratic
- c) reciprocal
- d) inverse

Ans : Solution A

37. Regression trees are often used to model _____ data.

- a) Linear
- b) Nonlinear
- c) Categorical
- d) Symmetrical

Ans : Solution B

38. The leaf nodes of a model tree are

- a) averages of numeric output attribute values.
- b) nonlinear regression equations.
- c) linear regression equations.
- d) sums of numeric output attribute values.

Ans : Solution C

39. Logistic regression is a _____ regression technique that is used to model data having a _____ outcome.

- a) linear, numeric
- b) linear, binary
- c) nonlinear, numeric
- d) nonlinear, binary

Ans : Solution D

40. This technique associates a conditional probability value with each data instance.

- a) linear regression
- b) logistic regression
- c) simple regression
- d) multiple linear regression

Ans : Solution B

41. This supervised learning technique can process both numeric and categorical input attributes.

- a) linear regression
- b) Bayes classifier
- c) logistic regression
- d) backpropagation learning

Ans : Solution A

42. With Bayes classifier, missing data items are
- a) treated as equal compares.
 - b) treated as unequal compares.
 - c) replaced with a default value.
 - d) ignored.

Ans : Solution B

43. This clustering algorithm merges and splits nodes to help modify nonoptimal partitions.
- a) agglomerative clustering
 - b) expectation maximization
 - c) conceptual clustering
 - d) K-Means clustering

Ans : Solution D

44. This clustering algorithm initially assumes that each data instance represents a single cluster.
- a) agglomerative clustering
 - b) conceptual clustering
 - c) K-Means clustering
 - d) expectation maximization

Ans : Solution C

45. This unsupervised clustering algorithm terminates when mean values computed for the current iteration of the algorithm are identical to the computed mean values for the previous iteration.
- a) agglomerative clustering
 - b) conceptual clustering
 - c) K-Means clustering
 - d) expectation maximization

Ans : Solution C

46. Machine learning techniques differ from statistical techniques in that machine learning methods
- a) typically assume an underlying distribution for the data.
 - b) are better able to deal with missing and noisy data.
 - c) are not able to explain their behavior.
 - d) have trouble with large-sized datasets.

Ans : Solution B

UNIT -II

1. True- False: Over fitting is more likely when you have huge amount of data to train?

- A) TRUE
- B) FALSE

Ans Solution: (B)

With a small training dataset, it's easier to find a hypothesis to fit the training data exactly i.e. over fitting.

2. What is pca.components_ in Sklearn?

Set of all eigen vectors for the projection space

Matrix of principal components

Result of the multiplication matrix

None of the above options

Ans A

3. Which of the following techniques would perform better for reducing dimensions of a data set?

- A. Removing columns which have too many missing values
- B. Removing columns which have high variance in data
- C. Removing columns with dissimilar data trends
- D. None of these

Ans Solution: (A)

If a columns have too many missing values, (say 99%) then we can remove such columns.

4. It is not necessary to have a target variable for applying dimensionality reduction algorithms.

- A. TRUE
- B. FALSE

Ans Solution: (A)

LDA is an example of supervised dimensionality reduction algorithm.

5. PCA can be used for projecting and visualizing data in lower dimensions.

- A. TRUE
- B. FALSE

Ans Solution: (A)

Sometimes it is very useful to plot the data in lower dimensions. We can take the first 2 principal components and then visualize the data using scatter plot.

6. The most popularly used dimensionality reduction algorithm is Principal Component Analysis (PCA). Which of the following is/are true about PCA?

PCA is an unsupervised method

- It searches for the directions that data have the largest variance
 - Maximum number of principal components <= number of features
 - All principal components are orthogonal to each other
- A. 1 and 2
 - B. 1 and 3
 - C. 2 and 3
 - D. All of the above

Ans D

- 7. PCA works better if there is?
- A linear structure in the data
 - If the data lies on a curved surface and not on a flat surface
 - If variables are scaled in the same unit
- A. 1 and 2
 - B. 2 and 3
 - C. 1 and 3
 - D. 1,2 and 3

Ans Solution: (C)

- 8. What happens when you get features in lower dimensions using PCA?
- The features will still have interpretability
 - The features will lose interpretability
 - The features must carry all information present in data
 - The features may not carry all information present in data
- A. 1 and 3
 - B. 1 and 4
 - C. 2 and 3
 - D. 2 and 4

Ans Solution: (D)

When you get the features in lower dimensions then you will lose some information of data most of the times and you won't be able to interpret the lower dimension data.

- 9. Which of the following option(s) is / are true?
- You need to initialize parameters in PCA
 - You don't need to initialize parameters in PCA
 - PCA can be trapped into local minima problem
 - PCA can't be trapped into local minima problem
- A. 1 and 3
 - B. 1 and 4
 - C. 2 and 3
 - D. 2 and 4

Ans Solution: (D)

PCA is a deterministic algorithm which doesn't have parameters to initialize and it doesn't have local minima problem like most of the machine learning algorithms has.

10. What is of the following statement is true about t-SNE in comparison to PCA?

- A. When the data is huge (in size), t-SNE may fail to produce better results.
- B. T-SNE always produces better result regardless of the size of the data
- C. PCA always performs better than t-SNE for smaller size data.
- D. None of these

Ans Solution: (A)

Option A is correct

11. [True or False] PCA can be used for projecting and visualizing data in lower dimensions.

- A. TRUE
- B. FALSE

Solution: (A)

Sometimes it is very useful to plot the data in lower dimensions. We can take the first 2 principal components and then visualize the data using scatter plot.

12. A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college.

1) Which of the following statement is true in following case?

- A) Feature F1 is an example of nominal variable.
- B) Feature F1 is an example of ordinal variable.
- C) It doesn't belong to any of the above category.
- D) Both of these

Solution: (B)

Ordinal variables are the variables which has some order in their categories. For example, grade A should be consider as high grade than grade B.

13. Which of the following is an example of a deterministic algorithm?

- A) PCA
- B) K-Means
- C) None of the above

Solution: (A)

A deterministic algorithm is that in which output does not change on different runs. PCA would give the same result if we run again, but not k-means.

UNIT -III

1. Which of the following methods do we use to best fit the data in Logistic Regression?

- A) Least Square Error
- B) Maximum Likelihood
- C) Jaccard distance
- D) Both A and B

Ans Solution: B

2. Choose which of the following options is true regarding One-Vs-All method in Logistic Regression.

- A) We need to fit n models in n-class classification problem
- B) We need to fit n-1 models to classify into n classes
- C) We need to fit only 1 model to classify into n classes
- D) None of these

Ans Solution: A

3. Suppose, You applied a Logistic Regression model on a given data and got a training accuracy X and testing accuracy Y. Now, you want to add a few new features in the same data. Select the option(s) which is/are correct in such a case.

Note: Consider remaining parameters are same.

- A) Training accuracy increases
- B) Training accuracy increases or remains the same
- C) Testing accuracy decreases
- D) Testing accuracy increases or remains the same

Ans Solution: A and D

Adding more features to model will increase the training accuracy because model has to consider more data to fit the logistic regression. But testing accuracy increases if feature is found to be significant

4. Which of the following algorithms do we use for Variable Selection?

- A) LASSO
- B) Ridge
- C) Both
- D) None of these

Ans Solution: A

In case of lasso we apply a absolute penalty, after increasing the penalty in lasso some of the coefficient of variables may become zero

5. Which of the following statement is true about outliers in Linear regression?

- A) Linear regression is sensitive to outliers
- B) Linear regression is not sensitive to outliers
- C) Can't say
- D) None of these

Ans Solution: (A)

The slope of the regression line will change due to outliers in most of the cases. So Linear Regression is sensitive to outliers.

6. Which of the following methods do we use to find the best fit line for data in Linear Regression?

- A) Least Square Error
- B) Maximum Likelihood
- C) Logarithmic Loss
- D) Both A and B

Ans Solution: (A)

In linear regression, we try to minimize the least square errors of the model to identify the line of best fit.

7. Which of the following is true about Residuals?

- A) Lower is better
- B) Higher is better
- C) A or B depend on the situation
- D) None of these

Ans Solution: (A)

Residuals refer to the error values of the model. Therefore lower residuals are desired.

8. Suppose you plotted a scatter plot between the residuals and predicted values in linear regression and you found that there is a relationship between them. Which of the following conclusion do you make about this situation?

- A) Since the there is a relationship means our model is not good
- B) Since the there is a relationship means our model is good
- C) Can't say
- D) None of these

Ans Solution: (A)

There should not be any relationship between predicted values and residuals. If there exists any relationship between them, it means that the model has not perfectly captured the information in the data.

9. Suppose you have fitted a complex regression model on a dataset. Now, you are using Ridge regression with penalty λ .

Choose the option which describes bias in best manner.

- A) In case of very large x; bias is low
- B) In case of very large x; bias is high
- C) We can't say about bias
- D) None of these

Ans Solution: (B)

If the penalty is very large it means model is less complex, therefore the bias would be high.

10. Which of the following option is true?

- A) Linear Regression errors values has to be normally distributed but in case of Logistic Regression it is not the case
- B) Logistic Regression errors values has to be normally distributed but in case of Linear Regression it is not the case
- C) Both Linear Regression and Logistic Regression error values have to be normally distributed
- D) Both Linear Regression and Logistic Regression error values have not to be normally distributed

Ans Solution: A

11. Suppose you have trained a logistic regression classifier and it outputs a new example x with a prediction $h_0(x) = 0.2$. This means

- Our estimate for $P(y=1 | x)$
- Our estimate for $P(y=0 | x)$
- Our estimate for $P(y=1 | x)$
- Our estimate for $P(y=0 | x)$

Ans Solution: B

12. **True-False: Linear Regression is a supervised machine learning algorithm.**

- A) TRUE
- B) FALSE

Solution: (A)

Yes, Linear regression is a supervised learning algorithm because it uses true labels for training. Supervised learning algorithm should have input variable (x) and an output variable (y) for each example.

13. **True-False: Linear Regression is mainly used for Regression.**

- A) TRUE
- B) FALSE

Solution: (A)

Linear Regression has dependent variables that have continuous values.

14. True-False: It is possible to design a Linear regression algorithm using a neural network?

- A) TRUE
- B) FALSE

Solution: (A)

True. A Neural network can be used as a universal approximator, so it can definitely implement a linear regression algorithm.

15. Which of the following methods do we use to find the best fit line for data in Linear Regression?

- A) Least Square Error
- B) Maximum Likelihood
- C) Logarithmic Loss
- D) Both A and B

Solution: (A)

In linear regression, we try to minimize the least square errors of the model to identify the line of best fit.

16. Which of the following evaluation metrics can be used to evaluate a model while modeling a continuous output variable?

- A) AUC-ROC
- B) Accuracy
- C) Logloss
- D) Mean-Squared-Error

Solution: (D)

Since linear regression gives output as continuous values, so in such case we use mean squared error metric to evaluate the model performance. Remaining options are use in case of a classification problem.

17. True-False: Lasso Regularization can be used for variable selection in Linear Regression.

- A) TRUE
- B) FALSE

Solution: (A)

True, In case of lasso regression we apply absolute penalty which makes some of the coefficients zero.

18. Which of the following is true about Residuals ?

- A) Lower is better
- B) Higher is better

- C) A or B depend on the situation
- D) None of these

Solution: (A)

Residuals refer to the error values of the model. Therefore lower residuals are desired.

19. Suppose that we have N independent variables (X_1, X_2, \dots, X_n) and dependent variable is Y. Now Imagine that you are applying linear regression by fitting the best fit line using least square error on this data.

You found that correlation coefficient for one of its variable (Say X_1) with Y is -0.95.

Which of the following is true for X_1 ?

- A) Relation between the X_1 and Y is weak
- B) Relation between the X_1 and Y is strong
- C) Relation between the X_1 and Y is neutral
- D) Correlation can't judge the relationship

Solution: (B)

The absolute value of the correlation coefficient denotes the strength of the relationship.

Since absolute correlation is very high it means that the relationship is strong between X_1 and Y.

20. Looking at above two characteristics, which of the following option is the correct for Pearson correlation between V_1 and V_2 ?

If you are given the two variables V_1 and V_2 and they are following below two characteristics.

1. If V_1 increases then V_2 also increases
 2. If V_1 decreases then V_2 behavior is unknown
- A) Pearson correlation will be close to 1
 - B) Pearson correlation will be close to -1
 - C) Pearson correlation will be close to 0
 - D) None of these

Solution: (D)

We cannot comment on the correlation coefficient by using only statement 1. We need to consider the both of these two statements. Consider V_1 as x and V_2 as $|x|$. The correlation coefficient would not be close to 1 in such a case.

21. Suppose Pearson correlation between V_1 and V_2 is zero. In such case, is it right to conclude that V_1 and V_2 do not have any relation between them?

- A) TRUE
- B) FALSE

Solution: (B)

Pearson correlation coefficient between 2 variables might be zero even when they have a relationship between them. If the correlation coefficient is zero, it just means that they don't move together. We can take examples like $y=|x|$ or $y=x^2$.

22. True- False: Overfitting is more likely when you have huge amount of data to train?

- A) TRUE
- B) FALSE

Solution: (B)

With a small training dataset, it's easier to find a hypothesis to fit the training data exactly i.e. overfitting.

23. We can also compute the coefficient of linear regression with the help of an analytical method called “Normal Equation”. Which of the following is/are true about Normal Equation?

- 1. We don't have to choose the learning rate
- 2. It becomes slow when number of features is very large
- 3. There is no need to iterate

- A) 1 and 2
- B) 1 and 3
- C) 2 and 3
- D) 1,2 and 3

Solution: (D)

Instead of gradient descent, Normal Equation can also be used to find coefficients.

Question Context 24-26:

Suppose you have fitted a complex regression model on a dataset. Now, you are using Ridge regression with penalty λ .

24. Choose the option which describes bias in best manner.

- A) In case of very large λ ; bias is low
- B) In case of very large λ ; bias is high
- C) We can't say about bias
- D) None of these

Solution: (B)

If the penalty is very large it means model is less complex, therefore the bias would be high.

25. What will happen when you apply very large penalty?

- A) Some of the coefficient will become absolute zero
- B) Some of the coefficient will approach zero but not absolute zero
- C) Both A and B depending on the situation
- D) None of these

Solution: (B)

In Lasso some of the coefficient value become zero, but in case of Ridge, the coefficients become close to zero but not zero.

26. What will happen when you apply very large penalty in case of Lasso?

- A) Some of the coefficient will become zero

- B) Some of the coefficient will be approaching to zero but not absolute zero
- C) Both A and B depending on the situation
- D) None of these

Solution: (A)

As already discussed, lasso applies absolute penalty, so some of the coefficients will become zero.

27. Which of the following statement is true about outliers in Linear regression?

- A) Linear regression is sensitive to outliers
- B) Linear regression is not sensitive to outliers
- C) Can't say
- D) None of these

Solution: (A)

The slope of the regression line will change due to outliers in most of the cases. So Linear Regression is sensitive to outliers.

28. Suppose you plotted a scatter plot between the residuals and predicted values in linear regression and you found that there is a relationship between them. Which of the following conclusion do you make about this situation?

- A) Since the there is a relationship means our model is not good
- B) Since the there is a relationship means our model is good
- C) Can't say
- D) None of these

Solution: (A)

There should not be any relationship between predicted values and residuals. If there exists any relationship between them, it means that the model has not perfectly captured the information in the data.

Question Context 29-31:

Suppose that you have a dataset D1 and you design a linear regression model of degree 3 polynomial and you found that the training and testing error is "0" or in another terms it perfectly fits the data.

29. What will happen when you fit degree 4 polynomial in linear regression?

- A) There are high chances that degree 4 polynomial will over fit the data
- B) There are high chances that degree 4 polynomial will under fit the data
- C) Can't say
- D) None of these

Solution: (A)

Since is more degree 4 will be more complex(overfit the data) than the degree 3 model so it will again perfectly fit the data. In such case training error will be zero but test error may not be zero.

30. What will happen when you fit degree 2 polynomial in linear regression?

- A) It is high chances that degree 2 polynomial will over fit the data
- B) It is high chances that degree 2 polynomial will under fit the data
- C) Can't say
- D) None of these

Solution: (B)

If a degree 3 polynomial fits the data perfectly, it's highly likely that a simpler model(degree 2 polynomial) might under fit the data.

31. In terms of bias and variance. Which of the following is true when you fit degree 2 polynomial?

- A) Bias will be high, variance will be high
- B) Bias will be low, variance will be high
- C) Bias will be high, variance will be low
- D) Bias will be low, variance will be low

Solution: (C)

Since a degree 2 polynomial will be less complex as compared to degree 3, the bias will be high and variance will be low.

Question Context 32-33:

We have been given a dataset with n records in which we have input attribute as x and output attribute as y. Suppose we use a linear regression method to model this data. To test our linear regressor, we split the data in training set and test set randomly.

32. Now we increase the training set size gradually. As the training set size increases, what do you expect will happen with the mean training error?

- A) Increase
- B) Decrease
- C) Remain constant
- D) Can't Say

Solution: (D)

Training error may increase or decrease depending on the values that are used to fit the model. If the values used to train contain more outliers gradually, then the error might just increase.

33. What do you expect will happen with bias and variance as you increase the size of training data?

- A) Bias increases and Variance increases
- B) Bias decreases and Variance increases
- C) Bias decreases and Variance decreases

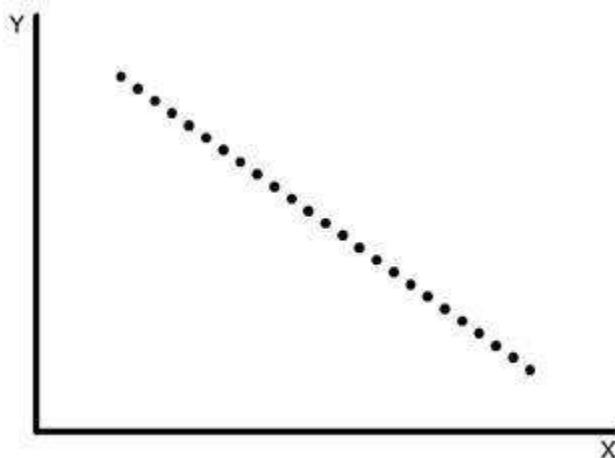
- D) Bias increases and Variance decreases
- E) Can't Say False

Solution: (D)

As we increase the size of the training data, the bias would increase while the variance would decrease.

Question Context 34:

Consider the following data where one input(X) and one output(Y) is given.



34. What would be the root mean square training error for this data if you run a Linear Regression model of the form ($Y = A_0 + A_1X$)?

- A) Less than 0
- B) Greater than zero
- C) Equal to 0
- D) None of these

Solution: (C)

We can perfectly fit the line on the following data so mean error will be zero.

Question Context 35-36:

Suppose you have been given the following scenario for training and validation error for Linear Regression.

Scenario	Learning Rate	Number of iterations	Training Error	Validation Error
1	0.1	1000	100	110
2	0.2	600	90	105

3	0.3	400	110	110
4	0.4	300	120	130
5	0.4	250	130	150

35. Which of the following scenario would give you the right hyper parameter?

- A) 1
- B) 2
- C) 3
- D) 4

Solution: (B)

Option B would be the better option because it leads to less training as well as validation error.

36. Suppose you got the tuned hyper parameters from the previous question. Now, Imagine you want to add a variable in variable space such that this added feature is important. Which of the following thing would you observe in such case?

- A) Training Error will decrease and Validation error will increase
- B) Training Error will increase and Validation error will increase
- C) Training Error will increase and Validation error will decrease
- D) Training Error will decrease and Validation error will decrease
- E) None of the above

Solution: (D)

If the added feature is important, the training and validation error would decrease.

Question Context 37-38:

Suppose, you got a situation where you find that your linear regression model is under fitting the data.

37. In such situation which of the following options would you consider?

1. I will add more variables
 2. I will start introducing polynomial degree variables
 3. I will remove some variables
- A) 1 and 2
 - B) 2 and 3
 - C) 1 and 3
 - D) 1, 2 and 3

Solution: (A)

In case of under fitting, you need to induce more variables in variable space or you can add some polynomial degree variables to make the model more complex to be able to fit the data better.

38. Now situation is same as written in previous question(under fitting). Which of following regularization algorithm would you prefer?

- A) L1
- B) L2
- C) Any
- D) None of these

Solution: (D)

I won't use any regularization methods because regularization is used in case of overfitting.

39. True-False: Is Logistic regression a supervised machine learning algorithm?

- A) TRUE
- B) FALSE

Solution: A

True, Logistic regression is a supervised learning algorithm because it uses true labels for training. Supervised learning algorithm should have input variables (x) and a target variable (Y) when you train the model .

40. True-False: Is Logistic regression mainly used for Regression?

- A) TRUE
- B) FALSE

Solution: B

Logistic regression is a classification algorithm, don't confuse with the name regression.

41. True-False: Is it possible to design a logistic regression algorithm using a Neural Network Algorithm?

- A) TRUE
- B) FALSE

Solution: A

True, Neural network is a universal approximator so it can implement linear regression algorithm.

42. True-False: Is it possible to apply a logistic regression algorithm on a 3-class Classification problem?

- A) TRUE
- B) FALSE

Solution: A

Yes, we can apply logistic regression on 3 classification problem, We can use One Vs all method for 3 class classification in logistic regression.

43. Which of the following methods do we use to best fit the data in Logistic Regression?

- A) Least Square Error
- B) Maximum Likelihood
- C) Jaccard distance
- D) Both A and B

Solution: B

Logistic regression uses maximum likely hood estimate for training a logistic regression.

44. Which of the following evaluation metrics can not be applied in case of logistic regression output to compare with target?

- A) AUC-ROC
- B) Accuracy
- C) Logloss
- D) Mean-Squared-Error

Solution: D

Since, Logistic Regression is a classification algorithm so it's output can not be real time value so mean squared error can not use for evaluating it

45. One of the very good methods to analyze the performance of Logistic Regression is AIC, which is similar to R-Squared in Linear Regression. Which of the following is true about AIC?

- A) We prefer a model with minimum AIC value
- B) We prefer a model with maximum AIC value
- C) Both but depend on the situation
- D) None of these

Solution: A

We select the best model in logistic regression which can least AIC.

46. [True-False] Standardisation of features is required before training a Logistic Regression.

- A) TRUE
- B) FALSE

Solution: B

Standardization isn't required for logistic regression. The main goal of standardizing features is to help convergence of the technique used for optimization.

47. Which of the following algorithms do we use for Variable Selection?

- A) LASSO
- B) Ridge
- C) Both
- D) None of these

Solution: A

In case of lasso we apply a absolute penalty, after increasing the penalty in lasso some of the coefficient of variables may become zero.

Context: 48-49

Consider a following model for logistic regression: $P(y=1|x, w) = g(w_0 + w_1x)$ where $g(z)$ is the logistic function.

In the above equation the $P(y=1|x; w)$, viewed as a function of x , that we can get by changing the parameters w .

48 What would be the range of p in such case?

- A) (0, inf)
- B) (-inf, 0)
- C) (0, 1)
- D) (-inf, inf)

Solution: C

For values of x in the range of real number from $-\infty$ to $+\infty$ Logistic function will give the output between (0,1)

49 In above question what do you think which function would make p between (0,1)?

- A) logistic function
- B) Log likelihood function
- C) Mixture of both
- D) None of them

Solution: A

Explanation is same as question number 10

50. Suppose you have been given a fair coin and you want to find out the odds of getting heads. Which of the following option is true for such a case?

- A) odds will be 0
- B) odds will be 0.5
- C) odds will be 1
- D) None of these

Solution: C

Odds are defined as the ratio of the probability of success and the probability of failure. So in case of fair coin probability of success is 1/2 and the probability of failure is 1/2 so odd would be 1

51. The logit function(given as l(x)) is the log of odds function. What could be the range of logit function in the domain x=[0,1]?

- A) $(-\infty, \infty)$
- B) $(0,1)$
- C) $(0, \infty)$
- D) $(-\infty, 0)$

Solution: A

For our purposes, the odds function has the advantage of transforming the probability function, which has values from 0 to 1, into an equivalent function with values between 0 and ∞ . When we take the natural log of the odds function, we get a range of values from $-\infty$ to ∞ .

52. Which of the following option is true?

- A) Linear Regression errors values has to be normally distributed but in case of Logistic Regression it is not the case
- B) Logistic Regression errors values has to be normally distributed but in case of Linear Regression it is not the case
- C) Both Linear Regression and Logistic Regression error values have to be normally distributed
- D) Both Linear Regression and Logistic Regression error values have not to be normally distributed

Solution:A

53. Which of the following is true regarding the logistic function for any value “x”?

Note:

Logistic(x): is a logistic function of any number “ x ”

Logit(x): is a logit function of any number “ x ”

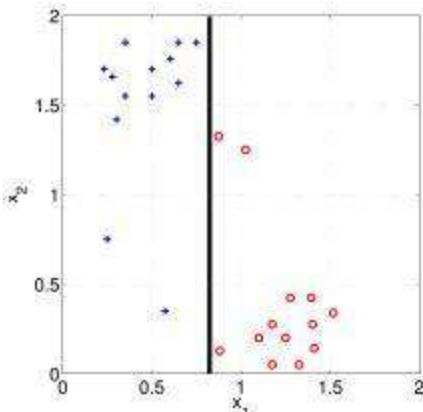
Logit_inv(x): is a inverse logit function of any number “ x ”

- A) $\text{Logistic}(x) = \text{Logit}(x)$
- B) $\text{Logistic}(x) = \text{Logit_inv}(x)$
- C) $\text{Logit_inv}(x) = \text{Logit}(x)$
- D) None of these

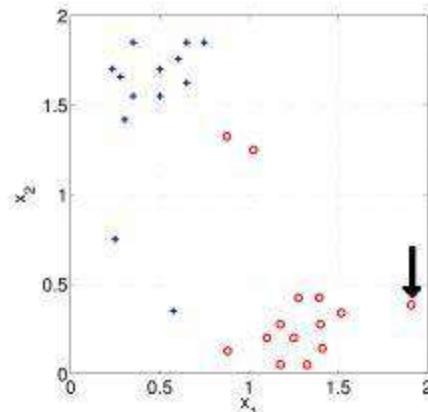
Solution: B

54. How will the bias change on using high(infinite) regularisation?

Suppose you have given the two scatter plot “a” and “b” for two classes(blue for positive and red for negative class). In scatter plot “a”, you correctly classified all data points using logistic regression (black line is a decision boundary).



(a)



(b)

- A) Bias will be high
- B) Bias will be low
- C) Can't say
- D) None of these

Solution: A

Model will become very simple so bias will be very high.

55. Suppose, You applied a Logistic Regression model on a given data and got a training accuracy X and testing accuracy Y. Now, you want to add a few new features in the same data. Select the option(s) which is/are correct in such a case.

Note: Consider remaining parameters are same.

- A) Training accuracy increases
- B) Training accuracy increases or remains the same
- C) Testing accuracy decreases
- D) Testing accuracy increases or remains the same

Solution: A and D

Adding more features to model will increase the training accuracy because model has to consider more data to fit the logistic regression. But testing accuracy increases if feature is found to be significant

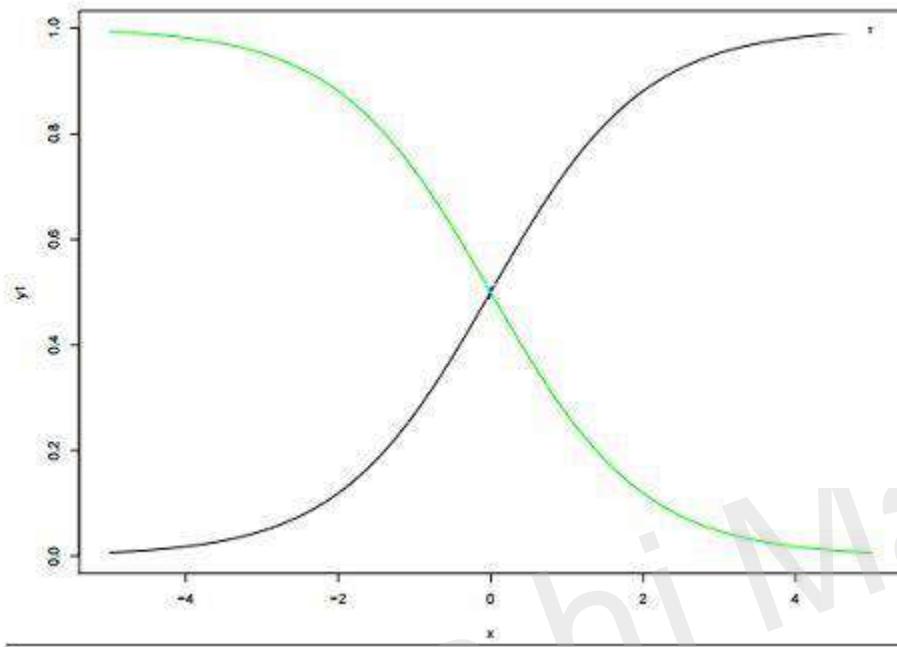
56. Choose which of the following options is true regarding One-Vs-All method in Logistic Regression.

- A) We need to fit n models in n-class classification problem
- B) We need to fit n-1 models to classify into n classes
- C) We need to fit only 1 model to classify into n classes
- D) None of these

Solution: A

If there are n classes, then n separate logistic regression has to fit, where the probability of each category is predicted over the rest of the categories combined.

57. Below are two different logistic models with different values for β_0 and β_1 .



Which of the

following statement(s) is true about β_0 and β_1 values of two logistics models (Green, Black)?

Note: consider $Y = \beta_0 + \beta_1 * X$. Here, β_0 is intercept and β_1 is coefficient.

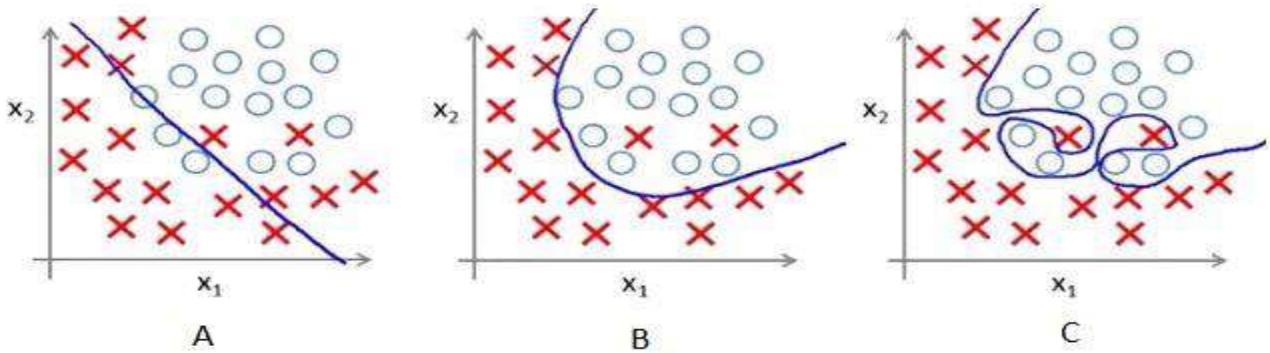
- A) β_1 for Green is greater than Black
- B) β_1 for Green is lower than Black
- C) β_1 for both models is same
- D) Can't Say

Solution: B

β_0 and β_1 : $\beta_0 = 0$, $\beta_1 = 1$ is in X1 color(black) and $\beta_0 = 0$, $\beta_1 = -1$ is in X4 color (green)

Context 58-60

Below are the three scatter plot(A,B,C left to right) and hand drawn decision boundaries for logistic regression.



58. Which of the following above figure shows that the decision boundary is overfitting the training data?

- A) A
- B) B
- C) C
- D) None of these

Solution: C

Since in figure 3, Decision boundary is not smooth that means it will over-fitting the data.

59. What do you conclude after seeing this visualization?

1. The training error in first plot is maximum as compare to second and third plot.
2. The best model for this regression problem is the last (third) plot because it has minimum training error (zero).
3. The second model is more robust than first and third because it will perform best on unseen data.
4. The third model is overfitting more as compare to first and second.
5. All will perform same because we have not seen the testing data.

- A) 1 and 3
- B) 1 and 3
- C) 1, 3 and 4
- D) 5

Solution: C

The trend in the graphs looks like a quadratic trend over independent variable X. A higher degree(Right graph) polynomial might have a very high accuracy on the train population but is expected to fail badly

on test dataset. But if you see in left graph we will have training error maximum because it underfits the training data

60. Suppose, above decision boundaries were generated for the different value of regularization.

Which of the above decision boundary shows the maximum regularization?

- A) A
- B) B
- C) C
- D) All have equal regularization

Solution: A

Since, more regularization means more penalty means less complex decision boundary that shows in first figure A.

61. What would do if you want to train logistic regression on same data that will take less time as well as give the comparatively similar accuracy(may not be same)?

Suppose you are using a Logistic Regression model on a huge dataset. One of the problem you may face on such huge data is that Logistic regression will take very long time to train.

- A) Decrease the learning rate and decrease the number of iteration
- B) Decrease the learning rate and increase the number of iteration
- C) Increase the learning rate and increase the number of iteration
- D) Increase the learning rate and decrease the number of iteration

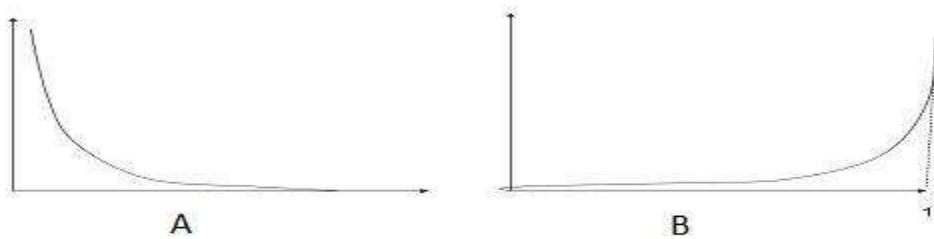
Solution: D

If you decrease the number of iteration while training it will take less time for surely but will not give the same accuracy for getting the similar accuracy but not exact you need to increase the learning rate.

62. Which of the following image is showing the cost function for $y=1$.

Following is the loss function in logistic regression(Y-axis loss function and x axis log probability) for two class classification problem.

Note: Y is the target class

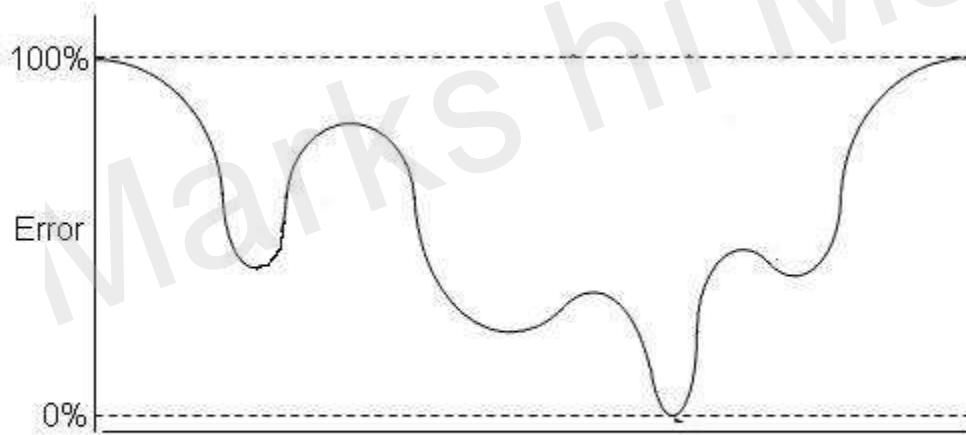


- A) A
 B) B
 C) Both
 D) None of these

Solution: A

A is the true answer as loss function decreases as the log probability increases

63. Suppose, Following graph is a cost function for logistic regression.



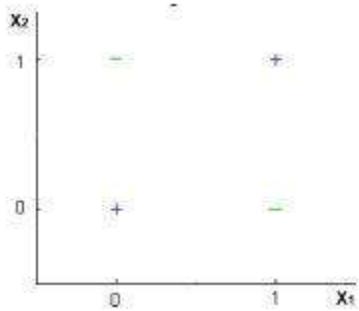
Now, How many local minimas are present in the graph?

- A) 1
 B) 2
 C) 3
 D) 4

Solution: C

There are three local minima present in the graph

64. Can a Logistic Regression classifier do a perfect classification on the below data?



Note: You can use only X1 and X2 variables where X1 and X2 can take only two binary values(0,1).

- A) TRUE
- B) FALSE
- C) Can't say
- D) None of these

Solution: B

No, logistic regression only forms linear decision surface, but the examples in the figure are not linearly separable.

UNIT IV

1. The SVM's are less effective when:

- A) The data is linearly separable
- B) The data is clean and ready to use
- C) The data is noisy and contains overlapping points

Ans Solution: C

When the data has noise and overlapping points, there is a problem in drawing a clear hyperplane without misclassifying.

2. The cost parameter in the SVM means:

- A) The number of cross-validations to be made
- B) The kernel to be used
- C) The tradeoff between misclassification and simplicity of the model
- D) None of the above

Ans Solution: C

The cost parameter decides how much an SVM should be allowed to "bend" with the data. For a low cost, you aim for a smooth decision surface and for a higher cost, you aim to classify more points correctly. It is also simply referred to as the cost of misclassification.

3. Which of the following are real world applications of the SVM?

- A) Text and Hypertext Categorization
- B) Image Classification
- C) Clustering of News Articles
- D) All of the above

Ans Solution: D

SVM's are highly versatile models that can be used for practically all real world problems ranging from regression to clustering and handwriting recognitions.

4. Which of the following is true about Naive Bayes ?

Assumes that all the features in a dataset are equally important

Assumes that all the features in a dataset are independent

Both A and B - answer

None of the above options

Ans Solution: C

5 What do you mean by generalization error in terms of the SVM?

- A) How far the hyperplane is from the support vectors
- B) How accurately the SVM can predict outcomes for unseen data
- C) The threshold amount of error in an SVM

Ans Solution: B

Generalisation error in statistics is generally the out-of-sample error which is the measure of how accurately a model can predict values for previously unseen data.

6 The SVM's are less effective when:

- A) The data is linearly separable
- B) The data is clean and ready to use
- C) The data is noisy and contains overlapping points

Ans Solution: C

When the data has noise and overlapping points, there is a problem in drawing a clear hyperplane without misclassifying.

7 What is/are true about kernel in SVM?

1. Kernel function map low dimensional data to high dimensional space
2. It's a similarity function

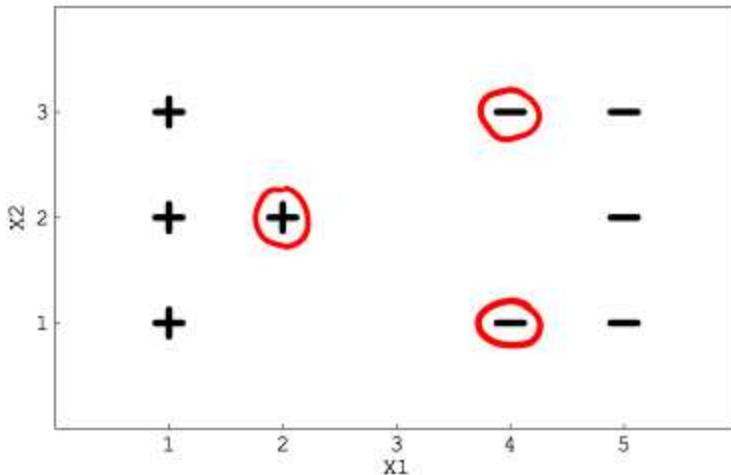
- A) 1
- B) 2
- C) 1 and 2
- D) None of these

Ans Solution: C

Both the given statements are correct.

Question Context:8–9

Suppose you are using a Linear SVM classifier with 2 class classification problem. Now you have been given the following data in which some points are circled red that are representing support vectors.



8. If you remove the following any one red points from the data. Does the decision boundary will change?

- A) Yes
- B) No

Solution: A

These three examples are positioned such that removing any one of them introduces slack in the constraints. So the decision boundary would completely change.

9. [True or False] If you remove the non-red circled points from the data, the decision boundary will change?

- A) True
- B) False

Solution: B

On the other hand, rest of the points in the data won't affect the decision boundary much.

10. What do you mean by generalization error in terms of the SVM?

- A) How far the hyperplane is from the support vectors
- B) How accurately the SVM can predict outcomes for unseen data
- C) The threshold amount of error in an SVM

Solution: B

Generalization error in statistics is generally the out-of-sample error which is the measure of how accurately a model can predict values for previously unseen data.

11. When the C parameter is set to infinite, which of the following holds true?

- A) The optimal hyperplane if exists, will be the one that completely separates the data
- B) The soft-margin classifier will separate the data
- C) None of the above

Solution: A

At such a high level of misclassification penalty, soft margin will not hold existence as there will be no room for error.

12. What do you mean by a hard margin?

- A) The SVM allows very low error in classification
- B) The SVM allows high amount of error in classification
- C) None of the above

Solution: A

A hard margin means that an SVM is very rigid in classification and tries to work extremely well in the training set, causing overfitting.

13. The minimum time complexity for training an SVM is $O(n^2)$. According to this fact, what sizes of datasets are not best suited for SVM's?

- A) Large datasets
- B) Small datasets
- C) Medium sized datasets
- D) Size does not matter

Solution: A

Datasets which have a clear classification boundary will function best with SVM's.

14. The effectiveness of an SVM depends upon:

- A) Selection of Kernel
- B) Kernel Parameters
- C) Soft Margin Parameter C
- D) All of the above

Solution: D

The SVM effectiveness depends upon how you choose the basic 3 requirements mentioned above in such a way that it maximises your efficiency, reduces error and overfitting.

15. Support vectors are the data points that lie closest to the decision surface.

- A) TRUE
- B) FALSE

Solution: A

They are the points closest to the hyperplane and the hardest ones to classify. They also have a direct bearing on the location of the decision surface.

16. The SVM's are less effective when:

- A) The data is linearly separable
- B) The data is clean and ready to use
- C) The data is noisy and contains overlapping points

Solution: C

When the data has noise and overlapping points, there is a problem in drawing a clear hyperplane without misclassifying.

17. Suppose you are using RBF kernel in SVM with high Gamma value. What does this signify?

- A) The model would consider even far away points from hyperplane for modeling
- B) The model would consider only the points close to the hyperplane for modeling
- C) The model would not be affected by distance of points from hyperplane for modeling
- D) None of the above

Solution: B

The gamma parameter in SVM tuning signifies the influence of points either near or far away from the hyperplane.

For a low gamma, the model will be too constrained and include all points of the training dataset, without really capturing the shape.

For a higher gamma, the model will capture the shape of the dataset well.

18. The cost parameter in the SVM means:

- A) The number of cross-validations to be made
- B) The kernel to be used
- C) The tradeoff between misclassification and simplicity of the model
- D) None of the above

Solution: C

The cost parameter decides how much an SVM should be allowed to “bend” with the data. For a low cost, you aim for a smooth decision surface and for a higher cost, you aim to classify more points correctly. It is also simply referred to as the cost of misclassification.

19. Suppose you are building a SVM model on data X. The data X can be error prone which means that you should not trust any specific data point too much. Now think that you want to build a SVM model which has quadratic kernel function of polynomial degree 2 that uses Slack variable C as one of its hyper parameter. Based upon that give the answer for following question.

What would happen when you use very large value of C($C \rightarrow \infty$)?

Note: For small C was also classifying all data points correctly

- A) We can still classify data correctly for given setting of hyper parameter C
- B) We can not classify data correctly for given setting of hyper parameter C
- C) Can't Say
- D) None of these

Solution: A

For large values of C, the penalty for misclassifying points is very high, so the decision boundary will perfectly separate the data if possible.

20. What would happen when you use very small C ($C \sim 0$)?

- A) Misclassification would happen
- B) Data will be correctly classified
- C) Can't say
- D) None of these

Solution: A

The classifier can maximize the margin between most of the points, while misclassifying a few points, because the penalty is so low.

21. If I am using all features of my dataset and I achieve 100% accuracy on my training set, but ~70% on validation set, what should I look out for?

- A) Underfitting
- B) Nothing, the model is perfect
- C) Overfitting

Solution: C

If we're achieving 100% training accuracy very easily, we need to check to verify if we're overfitting our data.

22. Which of the following are real world applications of the SVM?

- A) Text and Hypertext Categorization
- B) Image Classification
- C) Clustering of News Articles
- D) All of the above

Solution: D

SVM's are highly versatile models that can be used for practically all real world problems ranging from regression to clustering and handwriting recognitions.

Question Context: 23 – 25

Suppose you have trained an SVM with linear decision boundary after training SVM, you correctly infer that your SVM model is under fitting.

23. Which of the following option would you more likely to consider iterating SVM next time?

- A) You want to increase your data points
- B) You want to decrease your data points
- C) You will try to calculate more variables
- D) You will try to reduce the features

Solution: C

The best option here would be to create more features for the model.

24. Suppose you gave the correct answer in previous question. What do you think that is actually happening?

- 1. We are lowering the bias
- 2. We are lowering the variance
- 3. We are increasing the bias
- 4. We are increasing the variance

- A) 1 and 2
- B) 2 and 3
- C) 1 and 4
- D) 2 and 4

Solution: C

Better model will lower the bias and increase the variance

25. In above question suppose you want to change one of it's(SVM) hyperparameter so that effect would be same as previous questions i.e model will not under fit?

- A) We will increase the parameter C
- B) We will decrease the parameter C
- C) Changing in C don't effect
- D) None of these

Solution: A

Increasing C parameter would be the right thing to do here, as it will ensure regularized model

26. We usually use feature normalization before using the Gaussian kernel in SVM. What is true about feature normalization?

- 1. We do feature normalization so that new feature will dominate other
- 2. Some times, feature normalization is not feasible in case of categorical variables
- 3. Feature normalization always helps when we use Gaussian kernel in SVM

- A) 1
- B) 1 and 2
- C) 1 and 3
- D) 2 and 3

Solution: B

Statements one and two are correct.

Question Context: 27-29

Suppose you are dealing with 4 class classification problem and you want to train a SVM model on the data for that you are using One-vs-all method. Now answer the below questions?

27. How many times we need to train our SVM model in such case?

- A) 1
- B) 2
- C) 3
- D) 4

Solution: D

For a 4 class problem, you would have to train the SVM at least 4 times if you are using a one-vs-all method.

28. Suppose you have same distribution of classes in the data. Now, say for training 1 time in one vs all setting the SVM is taking 10 second. How many seconds would it require to train one-vs-all method end to end?

- A) 20
- B) 40
- C) 60
- D) 80

Solution: B

It would take $10 \times 4 = 40$ seconds

29 Suppose your problem has changed now. Now, data has only 2 classes. What would you think how many times we need to train SVM in such case?

- A) 1
- B) 2
- C) 3
- D) 4

Solution: A

Training the SVM only one time would give you appropriate results

Question context: 30 –31

Suppose you are using SVM with linear kernel of polynomial degree 2, Now think that you have applied this on data and found that it perfectly fit the data that means, Training and testing accuracy is 100%.

30. Now, think that you increase the complexity (or degree of polynomial of this kernel). What would you think will happen?

- A) Increasing the complexity will over fit the data
- B) Increasing the complexity will under fit the data
- C) Nothing will happen since your model was already 100% accurate
- D) None of these

Solution: A

Increasing the complexity of the data would make the algorithm overfit the data.

31. In the previous question after increasing the complexity you found that training accuracy was still 100%. According to you what is the reason behind that?

1. Since data is fixed and we are fitting more polynomial term or parameters so the algorithm starts memorizing everything in the data
 2. Since data is fixed and SVM doesn't need to search in big hypothesis space
- A) 1
B) 2
C) 1 and 2
D) None of these

Solution: C

Both the given statements are correct.

32. What is/are true about kernel in SVM?

1. Kernel function map low dimensional data to high dimensional space
 2. It's a similarity function
- A) 1
B) 2
C) 1 and 2
D) None of these

Solution: C

Both the given statements are correct.

UNIT V

1. Which of the following is a widely used and effective machine learning algorithm based on the idea of bagging?

- a) Decision Tree
- b) Regression
- c) Classification
- d) Random Forest

Ans D

2. Which of the following is a disadvantage of decision trees?

- a) Factor analysis
- b) Decision trees are robust to outliers
- c) Decision trees are prone to be overfit

- d) None of the above

Ans C

3. Can decision trees be used for performing clustering?

- a. True
- b. False

Ans Solution: (A)

Decision trees can also be used to find clusters in the data but clustering often generates natural clusters and is not dependent on any objective function.

4. Which of the following algorithm is most sensitive to outliers?

- a. K-means clustering algorithm
- b. K-medians clustering algorithm
- c. K-modes clustering algorithm
- d. K-medoids clustering algorithm

Ans Solution: (A)

5 Sentiment Analysis is an example of:

Regression

Classification

Clustering

Reinforcement Learning

Options:

- a. 1 Only
- b. 1 and 2
- c. 1 and 3
- d. 1, 2 and 4

Ans D

6 Which of the following is the most appropriate strategy for data cleaning before performing clustering analysis, given less than desirable number of data points:

Capping and flooring of variables

Removal of outliers

Options:

- a. 1 only
- b. 2 only
- c. 1 and 2
- d. None of the above

Ans A

7 Which of the following is/are true about bagging trees?

- 1. In bagging trees, individual trees are independent of each other
- 2. Bagging is the method for improving the performance by aggregating the results of weak learners

- A) 1
- B) 2
- C) 1 and 2
- D) None of these

Ans Solution: C

Both options are true. In Bagging, each individual trees are independent of each other because they consider different subset of features and samples.

8. Which of the following is/are true about boosting trees?

- 1. In boosting trees, individual weak learners are independent of each other
- 2. It is the method for improving the performance by aggregating the results of weak learners

- A) 1
- B) 2
- C) 1 and 2
- D) None of these

Ans Solution: B

In boosting tree individual weak learners are not independent of each other because each tree correct the results of previous tree. Bagging and boosting both can be consider as improving the base learners results.

9. In Random forest you can generate hundreds of trees (say T1, T2Tn) and then aggregate the results of these tree. Which of the following is true about individual (Tk) tree in Random Forest?

1. Individual tree is built on a subset of the features
 2. Individual tree is built on all the features
 3. Individual tree is built on a subset of observations
 4. Individual tree is built on full set of observations
- A) 1 and 3
B) 1 and 4
C) 2 and 3
D) 2 and 4

Ans Solution: A

Random forest is based on bagging concept, that consider fraction of sample and fraction of feature for building the individual trees.

10. Suppose you are using a bagging based algorithm say a RandomForest in model building. Which of the following can be true?

1. Number of tree should be as large as possible
 2. You will have interpretability after using Random Forest
- A) 1
B) 2
C) 1 and 2
D) None of these

Ans Solution: A

Since Random Forest aggregate the result of different weak learners, If it is possible we would want more number of trees in model building. Random Forest is a black box model you will lose interpretability after using it.

11. Which of the following is/are true about Random Forest and Gradient Boosting ensemble methods?

1. Both methods can be used for classification task
2. Random Forest is used for classification whereas Gradient Boosting is used for regression task
3. Random Forest is used for regression whereas Gradient Boosting is used for Classification task
4. Both methods can be used for regression task

- A) 1
- B) 2
- C) 3
- D) 4
- E) 1 and 4

Solution: E

Both algorithms are design for classification as well as regression task.

12. In Random forest you can generate hundreds of trees (say T₁, T₂T_n) and then aggregate the results of these tree. Which of the following is true about individual(T_k) tree in Random Forest?

- 1. Individual tree is built on a subset of the features
 - 2. Individual tree is built on all the features
 - 3. Individual tree is built on a subset of observations
 - 4. Individual tree is built on full set of observations
-
- A) 1 and 3
 - B) 1 and 4
 - C) 2 and 3
 - D) 2 and 4

Solution: A

Random forest is based on bagging concept, that consider fraction of sample and fraction of feature for building the individual trees.

13. Which of the following algorithm doesn't uses learning Rate as of one of its hyperparameter?

- 1. Gradient Boosting
 - 2. Extra Trees
 - 3. AdaBoost
 - 4. Random Forest
-
- A) 1 and 3
 - B) 1 and 4
 - C) 2 and 3
 - D) 2 and 4

Solution: D

Random Forest and Extra Trees don't have learning rate as a hyperparameter.

14. Which of the following algorithm are not an example of ensemble learning algorithm?

- A) Random Forest
- B) Adaboost
- C) Extra Trees
- D) Gradient Boosting
- E) Decision Trees

Solution: E

Decision trees doesn't aggregate the results of multiple trees so it is not an ensemble algorithm.

15. Suppose you are using a bagging based algorithm say a RandomForest in model building. Which of the following can be true?

1. Number of tree should be as large as possible
2. You will have interpretability after using RandomForest

- A) 1
- B) 2
- C) 1 and 2
- D) None of these

Solution: A

Since Random Forest aggregate the result of different weak learners, If It is possible we would want more number of trees in model building. Random Forest is a black box model you will lose interpretability after using it.

16. True-False: The bagging is suitable for high variance low bias models?

- A) TRUE
- B) FALSE

Solution: A

The bagging is suitable for high variance low bias models or you can say for complex models.

17. To apply bagging to regression trees which of the following is/are true in such case?

1. We build the N regression with N bootstrap sample
2. We take the average the of N regression tree
3. Each tree has a high variance with low bias

- A) 1 and 2
- B) 2 and 3
- C) 1 and 3
- D) 1,2 and 3

Solution: D

All of the options are correct and self-explanatory

18. How to select best hyper parameters in tree based models?

- A) Measure performance over training data
- B) Measure performance over validation data
- C) Both of these
- D) None of these

Solution: B

We always consider the validation results to compare with the test result.

19. In which of the following scenario a gain ratio is preferred over Information Gain?

- A) When a categorical variable has very large number of category
- B) When a categorical variable has very small number of category
- C) Number of categories is the not the reason
- D) None of these

Solution: A

When high cardinality problems, gain ratio is preferred over Information Gain technique.

20. Suppose you have given the following scenario for training and validation error for Gradient Boosting. Which of the following hyper parameter would you choose in such case?

Scenario	Depth	Training Error	Validation Error
1	2	100	110
2	4	90	105
3	6	50	100
4	8	45	105

5	10	30	150
---	----	----	-----

- A) 1
 B) 2
 C) 3
 D) 4

Solution: B

Scenario 2 and 4 has same validation accuracies but we would select 2 because depth is lower is better hyper parameter.

21. Which of the following is/are not true about DBSCAN clustering algorithm:

1. **For data points to be in a cluster, they must be in a distance threshold to a core point**
2. **It has strong assumptions for the distribution of data points in dataspace**
3. **It has substantially high time complexity of order $O(n^3)$**
4. **It does not require prior knowledge of the no. of desired clusters**
5. **It is robust to outliers**

Options:

- A. 1 only
 B. 2 only
 C. 4 only
 D. 2 and 3

Solution: D

- DBSCAN can form a cluster of any arbitrary shape and does not have strong assumptions for the distribution of data points in the data space.
- DBSCAN has a low time complexity of order $O(n \log n)$ only.

22. Point out the correct statement.

- a) The choice of an appropriate metric will influence the shape of the clusters
 b) Hierarchical clustering is also called HCA
 c) In general, the merges and splits are determined in a greedy manner
 d) All of the mentioned

Answer: d

Explanation: Some elements may be close to one another according to one distance and farther away according to another.

23. Which of the following is required by K-means clustering?

- a) defined distance metric
- b) number of clusters
- c) initial guess as to cluster centroids
- d) all of the mentioned

Answer: d

Explanation: K-means clustering follows partitioning approach.

24. Point out the wrong statement.

- a) k-means clustering is a method of vector quantization
- b) k-means clustering aims to partition n observations into k clusters
- c) k-nearest neighbor is same as k-means
- d) none of the mentioned

Answer: c

Explanation: k-nearest neighbour has nothing to do with k-means.

25. Which of the following function is used for k-means clustering?

- a) k-means
- b) k-mean
- c) heat map
- d) none of the mentioned

Answer: a

Explanation: K-means requires a number of clusters.

26. K-means is not deterministic and it also consists of number of iterations.

- a) True
- b) False

Answer: a

Explanation: K-means clustering produces the final estimate of cluster centroids.

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Choose the options that is incorrect regarding machine learning (ML) and artificial intelligence (AI)
((OPTION_A)) THIS IS MANDATORY OPTION	ML is an alternate way of programming intelligent machines.
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	ML and AI have very different goals
((OPTION_C)) This is optional	ML is a set of techniques that turns a dataset into a software.
((OPTION_D)) This is optional	AI is a software that can emulate the human mind
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	B
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Which of the following sentence is FALSE regarding regression
((OPTION_A)) THIS IS MANDATORY OPTION	It is used for prediction
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	It may be used for interpretation
((OPTION_C)) This is optional	It relates inputs to outputs.
((OPTION_D)) This is optional	It discovers causal relationships
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	D
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Grid search is
((OPTION_A)) THIS IS MANDATORY OPTION	Linear in D
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Exponential in D
((OPTION_C)) This is optional	Linear in N
((OPTION_D)) This is optional	Both B&C
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	D
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Find incorrect regarding Gradient of a continuous and differentiable function
((OPTION_A)) THIS IS MANDATORY OPTION	is zero at a minimum
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	is non-zero at a maximum
((OPTION_C)) This is optional	is zero at a saddle point
((OPTION_D)) This is optional	decreases as you get closer to the minimum
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	B
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Consider a linear-regression model with $N = 3$ and $D = 1$ with input-output pairs as follows: $y_1 = 22$, $x_1 = 1$, $y_2 = 3$, $x_2 = 1$, $y_3 = 3$, $x_3 = 2$. What is the gradient of mean-square error (MSE) with respect to β_1 when $\beta_0 = 0$ and $\beta_1 = 1$? Give your answer correct to two decimal digits.
((OPTION_A)) THIS IS MANDATORY OPTION	-1.66
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	2
((OPTION_C)) This is optional	3
((OPTION_D)) This is optional	4
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Let us say that we have computed the gradient of our cost function and stored it in a vector \mathbf{g} . What is the cost of one gradient descent update given the gradient?
((OPTION_A)) THIS IS MANDATORY OPTION	$O(D)$
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	$O(N)$
((OPTION_C)) This is optional	$O(ND)$
((OPTION_D)) This is optional	$O(ND_2)$
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	You observe the following while fitting a linear regression to the data: As you increase the amount of training data, the test error decreases and the training error increases. The train error is quite low (almost what you expect it to), while the test error is much higher than the train error. What do you think is the main reason behind this behavior. Choose the most probable option
((OPTION_A)) THIS IS MANDATORY OPTION	High variance
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	High model bias
((OPTION_C)) This is optional	High estimation bias
((OPTION_D)) This is optional	None of the above
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Adding more basis functions in a linear model... (pick the most probably option)
((OPTION_A)) THIS IS MANDATORY OPTION	Decreases model bias
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Decreases estimation bias
((OPTION_C)) This is optional	Decreases variance
((OPTION_D)) This is optional	Doesn't affect bias and variance
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	The problem of finding hidden structure in unlabeled data is called
((OPTION_A)) THIS IS MANDATORY OPTION	Supervised learning
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	UnSupervised learning
((OPTION_C)) This is optional	Reinforcement learning
((OPTION_D)) This is optional	None of the above
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	B
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Task of inferring a model from labeled training data is called
((OPTION_A)) THIS IS MANDATORY OPTION	Unsupervised learning
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	supervised learning
((OPTION_C)) This is optional	Reinforcement learning
((OPTION_D)) This is optional	None of the above
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	B
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Some telecommunication company wants to segment their customers into distinct groups in order to send appropriate subscription offers, this is an example of
((OPTION_A)) THIS IS MANDATORY OPTION	Supervised learning
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Data extraction
((OPTION_C)) This is optional	Serration
((OPTION_D)) This is optional	Unsupervised learning
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	D
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Self-organizing maps are an example of
((OPTION_A)) THIS IS MANDATORY OPTION	Unsupervised learning
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Supervised learning
((OPTION_C)) This is optional	Reinforcement learning
((OPTION_D)) This is optional	Missing data imputation
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	You are given data about seismic activity in Japan, and you want to predict a magnitude of the next earthquake, this is in an example of
((OPTION_A)) THIS IS MANDATORY OPTION	Supervised learning
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Unsupervised learning
((OPTION_C)) This is optional	Serration
((OPTION_D)) This is optional	None of the above
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Assume you want to perform supervised learning and to predict number of newborns according to size of storks' population (http://www.brixtonhealth.com/storksBabies.pdf), it is an example of
((OPTION_A)) THIS IS MANDATORY OPTION	Classification
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Regression
((OPTION_C)) This is optional	Clustering
((OPTION_D)) This is optional	None of the above
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	B
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Discriminating between spam and ham e-mails is a classification task, true or false?
((OPTION_A)) THIS IS MANDATORY OPTION	True
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	False
((OPTION_C)) This is optional	
((OPTION_D)) This is optional	
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	In the example of predicting number of babies based on storks' population size, number of babies is
((OPTION_A)) THIS IS MANDATORY OPTION	Outcome
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Feature
((OPTION_C)) This is optional	Attribute
((OPTION_D)) This is optional	None of the above
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	It may be better to avoid the metric of ROC curve as it can suffer from accuracy paradox.
((OPTION_A)) THIS IS MANDATORY OPTION	True
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	False
((OPTION_C)) This is optional	
((OPTION_D)) This is optional	
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	B
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	which of the following is not involve in data mining
((OPTION_A)) THIS IS MANDATORY OPTION	Knowledge extraction
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Data archaeology
((OPTION_C)) This is optional	Data exploration
((OPTION_D)) This is optional	Data transformation
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	D
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	The expected value or _____ of a random variable is the center of its distribution.
((OPTION_A)) THIS IS MANDATORY OPTION	Mode
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	median
((OPTION_C)) This is optional	mean
((OPTION_D)) This is optional	None of the above
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	C
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Point out the correct statement.
((OPTION_A)) THIS IS MANDATORY OPTION	Some cumulative distribution function F is non-decreasing and right-continuous
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Every cumulative distribution function F is decreasing and right-continuous
((OPTION_C)) This is optional	Every cumulative distribution function F is increasing and left-continuous
((OPTION_D)) This is optional	None of the above
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	D
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Which of the following of a random variable is a measure of spread
((OPTION_A)) THIS IS MANDATORY OPTION	variance
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	standard deviation
((OPTION_C)) This is optional	empirical mean
((OPTION_D)) This is optional	All above
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	The square root of the variance is called the _____ deviation
((OPTION_A)) THIS IS MANDATORY OPTION	empirical
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	mean
((OPTION_C)) This is optional	continuous
((OPTION_D)) This is optional	standard
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	D
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	For continuous random variables, the CDF is the derivative of the PDF.
((OPTION_A)) THIS IS MANDATORY OPTION	True
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	False
((OPTION_C)) This is optional	
((OPTION_D)) This is optional	
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	B
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Cumulative distribution functions are used to specify the distribution of multivariate random variables.
((OPTION_A)) THIS IS MANDATORY OPTION	True
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	False
((OPTION_C)) This is optional	
((OPTION_D)) This is optional	
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Consider the results of a medical experiment that aims to predict whether someone is going to develop myopia based on some physical measurements and heredity. In this case, the input dataset consists of the person's medical characteristics and the target variable is binary: 1 for those who are likely to develop myopia and 0 for those who aren't. This can be best classified as
((OPTION_A)) THIS IS MANDATORY OPTION	Regression
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Desicion Tree
((OPTION_C)) This is optional	Clustering
((OPTION_D)) This is optional	Association Rule
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	B
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	The purpose of a machine learning model is to approximate an unknown function that associates input elements to output ones
((OPTION_A)) THIS IS MANDATORY OPTION	True
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	False
((OPTION_C)) This is optional	
((OPTION_D)) This is optional	
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Training set is normally a representation of a global distribution
((OPTION_A)) THIS IS MANDATORY OPTION	True
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	False
((OPTION_C)) This is optional	
((OPTION_D)) This is optional	
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	2
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	The model has an excessive capacity and it's not more able to generalize considering the original dynamics provided by the training set. This problem is called as
((OPTION_A)) THIS IS MANDATORY OPTION	Underfitting
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Overfitting
((OPTION_C)) This is optional	Both
((OPTION_D)) This is optional	None
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	It can associate almost perfectly all the known samples to the corresponding output values, but when an unknown input is presented, the corresponding prediction error can be very high, This problem is called as
((OPTION_A)) THIS IS MANDATORY OPTION	Underfitting
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Overfitting
((OPTION_C)) This is optional	Both
((OPTION_D)) This is optional	None
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	----- may prove to be more difficult to discover as it could be initially considered the result of a perfect fitting
((OPTION_A)) THIS IS MANDATORY OPTION	Underfitting
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Overfitting
((OPTION_C)) This is optional	Both
((OPTION_D)) This is optional	None
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	when working with a supervised scenario, we define a non-negative error measure e_m which takes two arguments and allows us to compute a total error value over the whole dataset. Those two arguments are.
((OPTION_A)) THIS IS MANDATORY OPTION	expected and predicted output
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	calculated and predicted output
((OPTION_C)) This is optional	calculated and measured output
((OPTION_D)) This is optional	none
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Initial value represents a starting point over the surface of a n-variables function. A generic training algorithm has to find the global minimum or a point quite close to it (there's always a tolerance to avoid an excessive number of iterations and a consequent risk of overfitting). This measure is also called
((OPTION_A)) THIS IS MANDATORY OPTION	loss function
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	predicted output
((OPTION_C)) This is optional	measured output
((OPTION_D)) This is optional	mean square error
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	In 1984, the computer scientist L. Valiant proposed a mathematical approach to determine whether a problem is learnable by a computer. The name of this technique is
((OPTION_A)) THIS IS MANDATORY OPTION	Max likelihood
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Zero one loss error
((OPTION_C)) This is optional	Probably approximately correct
((OPTION_D)) This is optional	none
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	C
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	In particular, a concept is a subset of input patterns X which determine the same output element
((OPTION_A)) THIS IS MANDATORY OPTION	True
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	False
((OPTION_C)) This is optional	
((OPTION_D)) This is optional	
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Therefore, learning a concept (parametrically) means minimizing the corresponding loss function restricted to a specific class, while learning all possible concepts (belonging to the same universe), means finding the minimum of a global loss function
((OPTION_A)) THIS IS MANDATORY OPTION	True
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	False
((OPTION_C)) This is optional	
((OPTION_D)) This is optional	
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	An exponential time could lead to computational explosions when the datasets are too large or the optimization starting point is very far from an acceptable minimum. Moreover, it's important to remember the so-called
((OPTION_A)) THIS IS MANDATORY OPTION	curse of dimensionality
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Hughes phenomenon
((OPTION_C)) This is optional	Probably approximately correct
((OPTION_D)) This is optional	
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	In many cases, in order to capture the full expressivity, it's necessary to have a very large dataset and without enough training data, the approximation can become problematic. This is called...
((OPTION_A)) THIS IS MANDATORY OPTION	curse of dimensionality
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Hughes phenomenon
((OPTION_C)) This is optional	Probably approximately correct
((OPTION_D)) This is optional	
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	B
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	$P(h_{pi} X) \propto P(X h_{pi})P(h_{pi})$ First term is called as
((OPTION_A)) THIS IS MANDATORY OPTION	posteriori
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Apriori
((OPTION_C)) This is optional	likelihood.
((OPTION_D)) This is optional	
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	$P(h_{pi} X) \propto P(X h_{pi})P(h_{pi})$ second term is called as
((OPTION_A)) THIS IS MANDATORY OPTION	posteriori
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Apriori
((OPTION_C)) This is optional	likelihood
((OPTION_D)) This is optional	
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	B
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	$P(h_{pi} X) \propto P(X h_{pi})P(h_{pi})$ Third term is called as
((OPTION_A)) THIS IS MANDATORY OPTION	posteriori
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Apriori
((OPTION_C)) This is optional	likelihood
((OPTION_D)) This is optional	
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	C
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Which of the following step / assumption in regression modeling impacts the trade-off between under-fitting and over-fitting the most
((OPTION_A)) THIS IS MANDATORY OPTION	The polynomial degree
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Whether we learn the weights by matrix inversion or gradient descent
((OPTION_C)) This is optional	The use of a constant-term
((OPTION_D)) This is optional	
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1								
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Suppose you have the following data with one real-value input variable & one real-value output variable. What is leave-one out cross validation mean square error in case of linear regression ($Y = bX+c$)? <table border="1"> <thead> <tr> <th>X(independent variable)</th> <th>Y(dependent variable)</th> </tr> </thead> <tbody> <tr> <td>0</td> <td>2</td> </tr> <tr> <td>2</td> <td>2</td> </tr> <tr> <td>3</td> <td>1</td> </tr> </tbody> </table>	X(independent variable)	Y(dependent variable)	0	2	2	2	3	1
X(independent variable)	Y(dependent variable)								
0	2								
2	2								
3	1								
((OPTION_A)) THIS IS MANDATORY OPTION	10/27								
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	20/27								
((OPTION_C)) This is optional	50/27								
((OPTION_D)) This is optional	49/27								
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option									
((CORRECT_CHOICE)) Either A or B or C or D or E	D								
((EXPLANATION)) This is also optional									

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	<p>Which of the following is/ are true about “Maximum Likelihood estimate (MLE)”?</p> <ol style="list-style-type: none"> 1. MLE may not always exist 2. MLE always exists 3. If MLE exist, it (they) may not be unique 4. If MLE exist, it (they) must be unique
((OPTION_A)) THIS IS MANDATORY OPTION	1 and4
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	2 and3
((OPTION_C)) This is optional	1 and3
((OPTION_D)) This is optional	2 and4
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	C
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Let's say, a "Linear regression" model perfectly fits the training data (train error is zero). Now, Which of the following statement is true?
((OPTION_A)) THIS IS MANDATORY OPTION	You will always have test error zero
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	. You can not have test error zero
((OPTION_C)) This is optional	None of the above
((OPTION_D)) This is optional	
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	C
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Which one of the statement is true regarding residuals in regression analysis?
((OPTION_A)) THIS IS MANDATORY OPTION	A. Mean of residuals is always zero
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Mean of residuals is always less than zero
((OPTION_C)) This is optional	Mean of residuals is always greater than zero
((OPTION_D)) This is optional	There is no such rule for residuals.
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Which of the one is true about Heteroskedasticity?
((OPTION_A)) THIS IS MANDATORY OPTION	Linear Regression with varying error terms
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Linear Regression with constant error terms
((OPTION_C)) This is optional	Linear Regression with zero error terms
((OPTION_D)) This is optional	None of the above
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

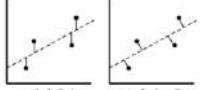
((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Which of the following indicates a fairly strong relationship between X and Y?
((OPTION_A)) THIS IS MANDATORY OPTION	A. Correlation coefficient = 0.9
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	. The p-value for the null hypothesis Beta coefficient =0 is 0.0001
((OPTION_C)) This is optional	The t-statistic for the null hypothesis Beta coefficient=0 is 30
((OPTION_D)) This is optional	None of these
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Which of the following assumptions do we make while deriving linear regression param 1. The true relationship between dependent y and predictor x is linear 2. The model errors are statistically independent 3. The errors are normally distributed with a 0 mean and constant standard deviation.
((OPTION_A)) THIS IS MANDATORY OPTION	1,2&3
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	1&3
((OPTION_C)) This is optional	All of above
((OPTION_D)) This is optional	
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	C
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	To test linear relationship of y(dependent) and x(independent) continuous variables, which of the following plot best suited?
((OPTION_A)) THIS IS MANDATORY OPTION	Scatter plot
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Barchart
((OPTION_C)) This is optional	Histograms
((OPTION_D)) This is optional	None of these
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Generally, which of the following method(s) is used for predicting continuous dependent variable? 1. Linear Regression 2. Logistic Regression
((OPTION_A)) THIS IS MANDATORY OPTION	1&2
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Only 1
((OPTION_C)) This is optional	Only 2
((OPTION_D)) This is optional	None f the above
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	B
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	<ul style="list-style-type: none"> • A correlation between age and health of a person found to be -1.09. On the basis of this you would tell the doctors that:
((OPTION_A)) THIS IS MANDATORY OPTION	<ul style="list-style-type: none"> . The age is good predictor of health
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	<ul style="list-style-type: none"> . The age is poor predictor of health
((OPTION_C)) This is optional	<ul style="list-style-type: none"> None of these
((OPTION_D)) This is optional	<ul style="list-style-type: none"> All of the above
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	C
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Which of the following offsets, do we use in case of least square line fit? Suppose horizontal axis is independent variable and vertical axis is dependent variable 
((OPTION_A)) THIS IS MANDATORY OPTION	Vertical offset
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Perpendicular offset
((OPTION_C)) This is optional	Both but depend on situation
((OPTION_D)) This is optional	Both a&b
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	<p>Suppose we have generated the data with help of polynomial regression of degree 3 (degree 3 will perfectly fit this data). Now consider below points and choose the option based on these points.</p> <p>1. Simple Linear regression will have high bias and low variance 2. Simple Linear regression will have low bias and high variance 3. polynomial of degree 3 will have low bias and high variance</p> <p>Polynomial of degree 3 will have low bias and Low variance</p>
((OPTION_A)) THIS IS MANDATORY OPTION	. Only 1
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	1&3
((OPTION_C)) This is optional	1&4
((OPTION_D)) This is optional	None of the above
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	C
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	<p>. Suppose you are training a linear regression model. Now consider these points.</p> <ol style="list-style-type: none"> 1. Overfitting is more likely if we have less data 2. Overfitting is more likely when the hypothesis space is small <p>Which of the above statement(s) are correct?</p>
((OPTION_A)) THIS IS MANDATORY OPTION	Both are False
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	1 is False and 2 is True
((OPTION_C)) This is optional	1 is True and 2 is False
((OPTION_D)) This is optional	None of the above
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	c
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	<p>Suppose we fit “Lasso Regression” to a data set, which has 100 features ($X_1, X_2 \dots X_{100}$). Now, we rescale one of these feature by multiplying with 10 (say that feature is X_1), and then refit Lasso regression with the same regularization parameter.</p> <p>Now, which of the following option will be correct?</p>
((OPTION_A)) THIS IS MANDATORY OPTION	It is more likely for X_1 to be excluded from the model
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	It is more likely for X_1 to be included in the model
((OPTION_C)) This is optional	. Can't say
((OPTION_D)) This is optional	None of the above
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	B
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Which of the following is true about “Ridge” or “Lasso” regression methods in case of feature selection?
((OPTION_A)) THIS IS MANDATORY OPTION	Ridge regression uses subset selection of features
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	. Lasso regression uses subset selection of features
((OPTION_C)) This is optional	Both use subset selection of features
((OPTION_D)) This is optional	All of the above
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	B
((EXPLANATION)) This is also optional	

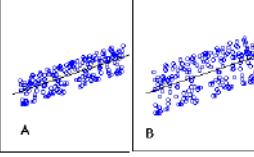
((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	<p>• Which of the following statement(s) can be true post adding a variable in a linear regression model?</p> <ol style="list-style-type: none"> 1. R-Squared and Adjusted R-squared both increase 2. R-Squared increases and Adjusted R-squared decreases 3. R-Squared decreases and Adjusted R-squared decreases 4. R-Squared decreases and Adjusted R-squared increases
((OPTION_A)) THIS IS MANDATORY OPTION	• 1 and 2
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	1 and 3
((OPTION_C)) This is optional	2 and 4
((OPTION_D)) This is optional	none of these
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	<ul style="list-style-type: none"> • Which of the following metrics can be used for evaluating regression models? <ol style="list-style-type: none"> 1. R Squared 2. Adjusted R Squared 3. F Statistics 4. RMSE / MSE / MAE
((OPTION_A)) THIS IS MANDATORY OPTION	2 and 4
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	1 and 2.
((OPTION_C)) This is optional	. 2, 3 and 4.
((OPTION_D)) This is optional	All of the above
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	D
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	<p>We can also compute the coefficient of linear regression with the help of an analytical method called “Normal Equation”. Which of the following is/are true about “Normal Equation”?</p> <ol style="list-style-type: none"> 1. We don't have to choose the learning rate 2. It becomes slow when number of features is very large 3. No need to iterate
((OPTION_A)) THIS IS MANDATORY OPTION	1 and 2
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	1&3
((OPTION_C)) This is optional	2&3
((OPTION_D)) This is optional	1,2&3
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	D
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	<p>. The expected value of Y is a linear function of the X(X1,X2....Xn) variables and regression line is defined as: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$</p> <p>Which of the following statement(s) are true?</p> <ol style="list-style-type: none"> 1. If X_i changes by an amount ΔX_i, holding other variables constant, then the expected value of Y changes by a proportional amount $\beta_i \Delta X_i$, for some constant β_i (which in general could be a positive or negative number). 2. The value of β_i is always the same, regardless of values of the other X's. 3. The total effect of the X's on the expected value of Y is the sum of their separate effects.
((OPTION_A)) THIS IS MANDATORY OPTION	. 1 and 2
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	1 and 3
((OPTION_C)) This is optional	2 and 3
((OPTION_D)) This is optional	1,2 and 3
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	D
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	<ul style="list-style-type: none"> • How many coefficients do you need to estimate in a simple linear regression model (One independent variable)
((OPTION_A)) THIS IS MANDATORY OPTION	1
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	2
((OPTION_C)) This is optional	CAN'T SAY
((OPTION_D)) This is optional	
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	B
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	2
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	. Below graphs show two fitted regression lines (A & B) on randomly generated data. Now, I want to find the sum of residuals in both cases A and B.  <p>Which of the following statement is true about sum of residuals of A and B</p>
((OPTION_A)) THIS IS MANDATORY OPTION	A has higher than B
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	A has lower than B
((OPTION_C)) This is optional	Both have same
((OPTION_D)) This is optional	None of these
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	C
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	If two variables are correlated, is it necessary that they have a linear relationsh
((OPTION_A)) THIS IS MANDATORY OPTION	YES
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	NO
((OPTION_C)) This is optional	Both a&b
((OPTION_D)) This is optional	None of the above
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	B
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Correlated variables can have zero correlation coefficient. True or False?
((OPTION_A)) THIS IS MANDATORY OPTION	TRUE
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	FALSE
((OPTION_C)) This is optional	
((OPTION_D)) This is optional	
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	<p>Suppose I applied a logistic regression model on data and got training accuracy X and testing accuracy Y. Now I want to add few new features in data. Select option(s) which are correct in such case.</p> <p>Note: Consider remaining parameters are same.</p> <ul style="list-style-type: none"> 1. Training accuracy always decreases. 2. Training accuracy always increases or remain same. 3. Testing accuracy always decreases <p>Testing accuracy always increases or remain same</p>
((OPTION_A)) THIS IS MANDATORY OPTION	Only 2
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Only 1
((OPTION_C)) This is optional	Only3
((OPTION_D)) This is optional	All of the above
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	The graph below represents a regression line predicting Y from X. The values on the graph shows the residuals for each predictions value. Use this information to compute the SSE.
((OPTION_A)) THIS IS MANDATORY OPTION	3.02
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	0.75
((OPTION_C)) This is optional	1.01
((OPTION_D)) This is optional	None of these
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CH OICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	<p>Suppose the distribution of salaries in a company X has median \$35,000, and 25th and 75th percentiles are \$21,000 and \$53,000 respectively.</p> <p>Would a person with Salary \$1 be considered an Outlier?</p>
((OPTION_A)) THIS IS MANDATORY OPTION	YES
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	NO
((OPTION_C)) This is optional	. More information is required
((OPTION_D)) This is optional	None of these
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	C
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Which of the following option is true regarding “Regression” and “Correlation” ? Note: y is dependent variable and x is independent variable.
((OPTION_A)) THIS IS MANDATORY OPTION	The relationship is symmetric between x and y in both.
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	The relationship is not symmetric between x and y in both.
((OPTION_C)) This is optional	The relationship is not symmetric between x and y in case of correlation but in case of regression it is symmetric.
((OPTION_D)) This is optional	The relationship is symmetric between x and y in case of correlation but in case of regression it is not symmetric.
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	B
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	True-False: Is Logistic regression a supervised machine learning algorithm?
((OPTION_A)) THIS IS MANDATORY OPTION	TRUE
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	FALSE
((OPTION_C)) This is optional	-
((OPTION_D)) This is optional	
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	True-False: Is Logistic regression mainly used for Regression?
((OPTION_A)) THIS IS MANDATORY OPTION	TRUE
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	FALSE
((OPTION_C)) This is optional	
((OPTION_D)) This is optional	
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	B
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	True-False: Is it possible to design a logistic regression algorithm using a Neural Network Algorithm?
((OPTION_A)) THIS IS MANDATORY OPTION	TRUE
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	FALSE
((OPTION_C)) This is optional	
((OPTION_D)) This is optional	
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	True-False: Is it possible to apply a logistic regression algorithm on a 3-class Classification problem?
((OPTION_A)) THIS IS MANDATORY OPTION	TRUE
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	FALSE
((OPTION_C)) This is optional	
((OPTION_D)) This is optional	
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Which of the following methods do we use to best fit the data in Logistic Regression?
((OPTION_A)) THIS IS MANDATORY OPTION	Least Square Error
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Maximum Likelihood
((OPTION_C)) This is optional	Jaccard distance
((OPTION_D)) This is optional	Both a&B
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	B
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	One of the very good methods to analyze the performance of Logistic Regression is AIC, which is similar to R-Squared in Linear Regression. Which of the following is true about AIC
((OPTION_A)) THIS IS MANDATORY OPTION	We prefer a model with minimum AIC value
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	We prefer a model with maximum AIC value
((OPTION_C)) This is optional	Both but depend on the situation
((OPTION_D)) This is optional	None of the above
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	True-False] Standardisation of features is required before training a Logistic Regression
((OPTION_A)) THIS IS MANDATORY OPTION	TRUE
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	FALSE
((OPTION_C)) This is optional	
((OPTION_D)) This is optional	
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	B
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Which of the following algorithms do we use for Variable Selection?
((OPTION_A)) THIS IS MANDATORY OPTION) LASSO
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Ridge
((OPTION_C)) This is optional	Both
((OPTION_D)) This is optional	All of these
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Suppose you have been given a fair coin and you want to find out the odds of getting heads. Which of the following option is true for such a case?
((OPTION_A)) THIS IS MANDATORY OPTION	odds will be 0
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	odds will be 0.5
((OPTION_C)) This is optional	odds will be 1
((OPTION_D)) This is optional	None of the above
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	C
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO) The logit function(given as $l(x)$) is the log of odds function. What could be the range of logit function in the domain $x=[0,1]$?
((OPTION_A)) THIS IS MANDATORY OPTION	($-\infty, \infty$)
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	(0,1)
((OPTION_C)) This is optional	(0, ∞)
((OPTION_D)) This is optional	($-\infty, 0$)
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Which of the following option is true?
((OPTION_A)) THIS IS MANDATORY OPTION	Linear Regression errors values has to be normally distributed but in case of Logistic Regression it is not the case
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Linear Regression errors values has to be normally distributed but in case of Logistic Regression it is not the case
((OPTION_C)) This is optional	Both Linear Regression and Logistic Regression error values have to be normally distributed
((OPTION_D)) This is optional	Both Linear Regression and Logistic Regression error values have not to be normally distributed
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	17) Which of the following is true regarding the logistic function for any value "x" Note: Logistic(x): is a logistic function of any number "x" Logit(x): is a logit function of any number "x" Logit_inv(x): is a inverse logit function of any number "x"?
((OPTION_A)) THIS IS MANDATORY OPTION	C) A) Logistic(x) = Logit(x)
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Logistic(x) = Logit_inv(x)
((OPTION_C)) This is optional	A) Logistic(x) = Logit(x)
((OPTION_D)) This is optional	None of these
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	B
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	2
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	<p>Suppose, You applied a Logistic Regression model on a given data and got a training accuracy X and testing accuracy Y. Now, you want to add a few new features in the same data. Select the option(s) which is/are correct in such a case.</p> <p>Note: Consider remaining parameters are same.</p>
((OPTION_A)) THIS IS MANDATORY OPTION	Training accuracy increases
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Training accuracy increases or remains the same
((OPTION_C)) This is optional	Testing accuracy decreases
((OPTION_D)) This is optional	Testing accuracy increases or remains the same
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	A&D
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Choose which of the following options is true regarding One-Vs-All method in Logistic Regression.
((OPTION_A)) THIS IS MANDATORY OPTION	We need to fit n models in n-class classification problem
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	We need to fit n-1 models to classify into n classes
((OPTION_C)) This is optional	We need to fit only 1 model to classify into n classes
((OPTION_D)) This is optional	
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	<p>What would do if you want to train logistic regression on same data that will take less time as well as give the comparatively similar accuracy(may not be same)?</p> <p>Suppose you are using a Logistic Regression model on a huge dataset. One of the problem you may face on such huge data is that Logistic regression will take very long time to train</p>
((OPTION_A)) THIS IS MANDATORY OPTION	Decrease the learning rate and decrease the number of iteration
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Decrease the learning rate and increase the number of iteration
((OPTION_C)) This is optional	Increase the learning rate and increase the number of iteration
((OPTION_D)) This is optional	Increase the learning rate and decrease the number of iteration
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	D
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	2
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	<p>Which of the following image is showing the cost function for $y=1$. Following is the loss function in logistic regression(Y-axis loss function and x axis log probability) for two class classification problem. Note: Y is the target class</p> 
((OPTION_A)) THIS IS MANDATORY OPTION	A
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	B
((OPTION_C)) This is optional	BOTH
((OPTION_D)) This is optional	NON OF THESE
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Logistic regression is used when you want to:
((OPTION_A)) THIS IS MANDATORY OPTION	Predict a dichotomous variable from continuous or dichotomous variables.
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Predict a continuous variable from dichotomous variables.
((OPTION_C)) This is optional	Predict any categorical variable from several other categorical variables.
((OPTION_D)) This is optional	Predict a continuous variable from dichotomous or continuous variables
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	A
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	The odds ratio is
((OPTION_A)) THIS IS MANDATORY OPTION	The ratio of the probability of an event not happening to the probability of the event happening.
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	The probability of an event occurring.
((OPTION_C)) This is optional	The ratio of the odds after a unit change in the predictor to the original odds.
((OPTION_D)) This is optional	The ratio of the probability of an event happening to the probability of the event not happening.
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	C
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Large values of the log-likelihood statistic indicate:
((OPTION_A)) THIS IS MANDATORY OPTION	That there are a greater number of explained vs. unexplained observations.
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	That the statistical model fits the data well.
((OPTION_C)) This is optional	That as the predictor variable increases, the likelihood of the outcome occurring decreases.
((OPTION_D)) This is optional	That the statistical model is a poor fit of the data.
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	B
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	Logistic regression assumes a:
((OPTION_A)) THIS IS MANDATORY OPTION	Linear relationship between continuous predictor variables and the outcome variable.
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	Linear relationship between continuous predictor variables and the logit of the outcome variable.
((OPTION_C)) This is optional	Linear relationship between continuous predictor variables.
((OPTION_D)) This is optional	Linear relationship between observations.
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	B
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	In binary logistic regression:
((OPTION_A)) THIS IS MANDATORY OPTION	The dependent variable is continuous.
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	The dependent variable is divided into two equal subcategories.
((OPTION_C)) This is optional	The dependent variable consists of two categories.
((OPTION_D)) This is optional	There is no dependent variable.
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	C
((EXPLANATION)) This is also optional	

((MARKS)) QUESTION IS OF HOW MANY MARKS? (1 OR 2 OR 3 UPTO 10)	1
((QUESTION)) ENTER CONTENT. QTN CAN HAVE IMAGES ALSO	The correlation coefficient is used to determine
((OPTION_A)) THIS IS MANDATORY OPTION	A specific value of the y-variable given a specific value of the x-variable
((OPTION_B)) THIS IS ALSO MANDATORY OPTION	A specific value of the x-variable given a specific value of the y-variable
((OPTION_C)) This is optional	The strength of the relationship between the x and y variables
((OPTION_D)) This is optional	none
((OPTION_E)) This is optional. If optional keep empty so that system will skip this option	
((CORRECT_CHOICE)) Either A or B or C or D or E	C
((EXPLANATION)) This is also optional	

Unit IV Naïve Bayes and Support Vector Machine

1. How many terms are required for building a bayes model?

- a) 1
- b) 2
- c) 3
- d) 4

Answer: c

2. What is needed to make probabilistic systems feasible in the world?

- a) Reliability
- b) Crucial robustness
- c) Feasibility
- d) None of the mentioned

Answer: b

3. Where does the bayes rule can be used?

- a) Solving queries
- b) Increasing complexity
- c) Decreasing complexity
- d) Answering probabilistic query

Answer: d

4. What does the bayesian network provides?

- a) Complete description of the domain
- b) Partial description of the domain
- c) Complete description of the problem
- d) None of the mentioned

Answer: a

5. How the entries in the full joint probability distribution can be calculated?

- a) Using variables
- b) Using information
- c) Both Using variables & information
- d) None of the mentioned

Answer: b

6. How the bayesian network can be used to answer any query?

- a) Full distribution
- b) Joint distribution
- c) Partial distribution
- d) All of the mentioned

Answer: b

7. How the compactness of the bayesian network can be described?

- a) Locally structured
- b) Fully structured
- c) Partial structure
- d) All of the mentioned

Answer: a

8. To which does the local structure is associated?

- a) Hybrid
- b) Dependant
- c) Linear
- d) None of the mentioned

Answer: c

9. Which condition is used to influence a variable directly by all the others?

- a) Partially connected
- b) Fully connected
- c) Local connected
- d) None of the mentioned

Answer: b

10. What is the consequence between a node and its predecessors while creating bayesian network?

- a) Functionally dependent
- b) Dependant
- c) Conditionally independent
- d) Both Conditionally dependant & Dependant

Answer: c

11.What do you mean by generalization error in terms of the SVM?

- A) How far the hyperplane is from the support vectors
- B) How accurately the SVM can predict outcomes for unseen data
- C) The threshold amount of error in an SVM

Solution: B

12. When the C parameter is set to infinite, which of the following holds true?

- A) The optimal hyperplane if exists, will be the one that completely separates the data
- B) The soft-margin classifier will separate the data
- C) None of the above

Solution: A

13.What do you mean by a hard margin?

- A) The SVM allows very low error in classification

B) The SVM allows high amount of error in classification

C) None of the above

Solution: A

14. The minimum time complexity for training an SVM is $O(n^2)$. According to this fact, what sizes of datasets are not best suited for SVM's?

A) Large datasets

B) Small datasets

C) Medium sized datasets

D) Size does not matter

Solution: A

15. The effectiveness of an SVM depends upon:

A) Selection of Kernel

B) Kernel Parameters

C) Soft Margin Parameter C

D) All of the above

Solution: D

16. Support vectors are the data points that lie closest to the decision surface.

A) TRUE

B) FALSE

Solution: A

17. The SVM's are less effective when:

A) The data is linearly separable

- B) The data is clean and ready to use
- C) The data is noisy and contains overlapping points

Solution: C

18. Suppose you are using RBF kernel in SVM with high Gamma value. What does this signify?

- A) The model would consider even far away points from hyperplane for modeling
- B) The model would consider only the points close to the hyperplane for modeling
- C) The model would not be affected by distance of points from hyperplane for modeling
- D) None of the above

Solution: B

19. The cost parameter in the SVM means:

- A) The number of cross-validations to be made
- B) The kernel to be used
- C) The tradeoff between misclassification and simplicity of the model
- D) None of the above

Solution: C

20. Suppose you are building a SVM model on data X. The data X can be error prone which means that you should not trust any specific data point too much. Now think that you want to build a SVM model which has quadratic kernel function of polynomial degree 2 that uses Slack variable C as one of its hyper parameter. Based upon that give the answer for following question.

What would happen when you use very large value of C($C \rightarrow \infty$)?

- A) We can still classify data correctly for given setting of hyper parameter C
- B) We can not classify data correctly for given setting of hyper parameter C
- C) Can't Say
- D) None of these

Solution: A

21. What would happen when you use very small C ($C \sim 0$)?

- A) Misclassification would happen
- B) Data will be correctly classified
- C) Can't say
- D) None of these

Solution: A

22. If I am using all features of my dataset and I achieve 100% accuracy on my training set, but ~70% on validation set, what should I look out for?

- A) Underfitting
- B) Nothing, the model is perfect
- C) Overfitting

Solution: C

23. Which of the following are real world applications of the SVM?

- A) Text and Hypertext Categorization
- B) Image Classification
- C) Clustering of News Articles
- D) All of the above

Solution: D

24. Which of the following option would you more likely to consider iterating SVM next time?

- A) You want to increase your data points
- B) You want to decrease your data points
- C) You will try to calculate more variables
- D) You will try to reduce the features

Solution: C

26. We usually use feature normalization before using the Gaussian kernel in SVM. What is true about feature normalization?

1. We do feature normalization so that new feature will dominate other
 2. Some times, feature normalization is not feasible in case of categorical variables
 3. Feature normalization always helps when we use Gaussian kernel in SVM
- A) 1
B) 1 and 2
C) 1 and 3
D) 2 and 3

Solution: B

Question context: 27 – 28

Suppose you are using SVM with linear kernel of polynomial degree 2, Now think that you have applied this on data and found that it perfectly fit the data that means, Training and testing accuracy is 100%.

27) Now, think that you increase the complexity(or degree of polynomial of this kernel). What would you think will happen?

- A) Increasing the complexity will overfit the data
B) Increasing the complexity will underfit the data
C) Nothing will happen since your model was already 100% accurate
D) None of these

Solution: A

28) In the previous question after increasing the complexity you found that training accuracy was still 100%. According to you what is the reason behind that?

1. Since data is fixed and we are fitting more polynomial term or parameters so the algorithm starts memorizing everything in the data

2. Since data is fixed and SVM doesn't need to search in big hypothesis space

A) 1

B) 2

C) 1 and 2

D) None of these

Solution: C

29. What is/are true about kernel in SVM?

1. Kernel function map low dimensional data to high dimensional space

2. It's a similarity function

A) 1

B) 2

C) 1 and 2

D) None of these

Solution: C

UNIT V

Decision Trees and Ensemble Learning

1. Predicting with trees evaluate _____ within each group of data.

- a) equality
- b) homogeneity
- c) heterogeneity
- d) all of the mentioned

Answer: b

2. Point out the wrong statement.

- a) Training and testing data must be processed in different way
- b) Test transformation would mostly be imperfect
- c) The first goal is statistical and second is data compression in PCA
- d) All of the mentioned

Answer: a

3. Which of the following method options is provided by train function for bagging?

- a) bagEarth
- b) treebag
- c) bagFDA
- d) all of the mentioned

Answer: d

4. Which of the following is correct with respect to random forest?

- a) Random forest are difficult to interpret but often very accurate
- b) Random forest are easy to interpret but often very accurate
- c) Random forest are difficult to interpret but very less accurate
- d) None of the mentioned

Answer: a

5. Point out the correct statement.

- a) Prediction with regression is easy to implement
- b) Prediction with regression is easy to interpret
- c) Prediction with regression performs well when linear model is correct

Answer: d

6. Which of the following library is used for boosting generalized additive models?

- a) gamBoost
- b) gbm
- c) ada
- d) all of the mentioned

Answer: a

7. The principal components are equal to left singular values if you first scale the variables.

- a) True
- b) False

Answer: b

8. Which of the following is statistical boosting based on additive logistic regression?

- a) gamBoost
- b) gbm
- c) ada
- d) mboost

Answer: a

9. Which of the following is one of the largest boost subclass in boosting?

- a) variance boosting
- b) gradient boosting
- c) mean boosting

d) all of the mentioned

Answer: b

10. PCA is most useful for non linear type models.

a) True

b) False

Answer: b

11. varImp is a wrapper around the evimp function in the _____ package.

a) numpy

b) earth

c) plot

d) none of the mentioned

Answer: b

12. Point out the wrong statement.

a) The trapezoidal rule is used to compute the area under the ROC curve

b) For regression, the relationship between each predictor and the outcome is evaluated

c) An argument, para, is used to pick the model fitting technique

d) All of the mentioned

Answer: c

13. Which of the following curve analysis is conducted on each predictor for classification?

a) NOC

b) ROC

c) COC

d) All of the mentioned

Answer: b

14. Which of the following function tracks the changes in model statistics?

- a) varImp
- b) varImpTrack
- c) findTrack
- d) none of the mentioned

Answer: a

15. Point out the correct statement.

- a) The difference between the class centroids and the overall centroid is used to measure the variable influence
- b) The Bagged Trees output contains variable usage statistics
- c) Boosted Trees uses different approach as a single tree
- d) None of the mentioned

Answer: a

16. The advantage of using a model-based approach is that is more closely tied to the model performance.

- a) True
- b) False

Answer: a

17. Which of the following model sums the importance over each boosting iteration?

- a) Boosted trees
- b) Bagged trees
- c) Partial least squares
- d) None of the mentioned

Answer: a

18. Which of the following argument is used to set importance values?

- a) scale
- b) set
- c) value
- d) all of the mentioned

Answer: a

19. For most classification models, each predictor will have a separate variable importance for each class.

- a) True
- b) False

Answer: a

20. A _____ is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.

- a) Decision tree
- b) Graphs
- c) Trees

d) Neural Networks

Answer: a

21. Decision Tree is a display of an algorithm.

a) True

b) False

Answer: a

22. What is Decision Tree?

a) Flow-Chart

b) Structure in which internal node represents test on an attribute, each branch represents outcome of test and each leaf node represents class label

c) Flow-Chart & Structure in which internal node represents test on an attribute, each branch represents outcome of test and each leaf node represents class label

d) None of the mentioned

Answer: c

23. Decision Trees can be used for Classification Tasks.

a) True

b) False

View Answer

Answer: a

24. Choose from the following that are Decision Tree nodes?

a) Decision Nodes

b) End Nodes

c) Chance Nodes

d) All of the mentioned

Answer: d

25. Decision Nodes are represented by _____

- a) Disks
- b) Squares
- c) Circles
- d) Triangles

Answer: b

26. Chance Nodes are represented by _____

- a) Disks
- b) Squares
- c) Circles
- d) Triangles

Answer: c

27. End Nodes are represented by _____

- a) Disks
- b) Squares
- c) Circles
- d) Triangles

Answer: d

28. Which of the following are the advantage/s of Decision Trees?

- a) Possible Scenarios can be added
- b) Use a white box model, If given result is provided by a model
- c) Worst, best and expected values can be determined for different scenarios
- d) All of the mentioned

Answer: d

29. Which of the following algorithm is not an example of an ensemble method?

- A. Extra Tree Regressor
- B. Random Forest
- C. Gradient Boosting
- D. Decision Tree

Solution: (D)

30. What is true about an ensembled classifier?

- 1. Classifiers that are more “sure” can vote with more conviction
 - 2. Classifiers can be more “sure” about a particular part of the space
 - 3. Most of the times, it performs better than a single classifier
- A. 1 and 2
 - B. 1 and 3
 - C. 2 and 3
 - D. All of the above

Solution: (D)

31. Which of the following option is / are correct regarding benefits of ensemble model?

- 1. Better performance
 - 2. Generalized models
 - 3. Better interpretability
- A. 1 and 3
 - B. 2 and 3
 - C. 1 and 2
 - D. 1, 2 and 3

Solution: (C)

32. Which of the following can be true for selecting base learners for an ensemble?

1. Different learners can come from same algorithm with different hyper parameters

2. Different learners can come from different algorithms

3. Different learners can come from different training spaces

A. 1

B. 2

C. 1 and 3

D. 1, 2 and 3

Solution: (D)

33. True or False: Ensemble learning can only be applied to supervised learning methods.

A. True

B. False

Solution: (B)

34. True or False: Ensembles will yield bad results when there is significant diversity among the models.

A. True

B. False

Solution: (B)

35. Which of the following is / are true about weak learners used in ensemble model?

1. They have low variance and they don't usually overfit

2. They have high bias, so they can not solve hard learning problems

3. They have high variance and they don't usually overfit

A. 1 and 2

B. 1 and 3

C. 2 and 3

D. None of these

Solution: (A)

36. True or False: Ensemble of classifiers may or may not be more accurate than any of its individual model.

A. True

B. False

Solution: (A)

37. parameters of all base models to improve the ensemble performance?

A. Yes

B. No

C. can't say

Solution: (B)

38. Generally, an ensemble method works better, if the individual base models have _____?

A. Less correlation among predictions

B. High correlation among predictions

C. Correlation does not have any impact on ensemble output

D. None of the above

Solution: (A)

39. Which of the following ensemble method works similar to above-discussed election procedure?

A. Bagging

B. Boosting

C. A Or B

D. None of these

Solution: (A)

40. Suppose you are given 'n' predictions on test data by 'n' different models (M_1, M_2, \dots, M_n) respectively. Which of the following method(s) can be used to combine the predictions of these models?

1. Median
 2. Product
 3. Average
 4. Weighted sum
 5. Minimum and Maximum
 6. Generalized mean rule
- A. 1, 3 and 4
B. 1,3 and 6
C. 1,3, 4 and 6
D. All of above

Solution: (D)

41. If you want to ensemble these models using majority voting method. What will be the maximum accuracy you can get?

- A. 100%
B. 78.38 %
C. 44%
D. 70

Solution: (A)

42. If you want to ensemble these models using majority voting. What will be the minimum accuracy you can get?

- A. Always greater than 70%
B. Always greater than and equal to 70%
C. It can be less than 70%

D. None of these

Solution: (C)

43. How can we assign the weights to output of different models in an ensemble?

1. Use an algorithm to return the optimal weights

2. Choose the weights using cross validation

3. Give high weights to more accurate models

A. 1 and 2

B. 1 and 3

C. 2 and 3

D. All of above

Solution: (D)

44. Which of the following is true about averaging ensemble?

A. It can only be used in classification problem

B. It can only be used in regression problem

C. It can be used in both classification as well as regression

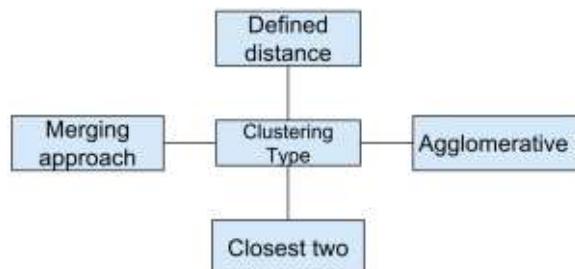
D. None of these

Solution: (C)

UNIT VI

Clustering Techniques

1. Which of the following clustering type has characteristic shown in the below figure?



- a) Partitional
- b) Hierarchical
- c) Naive bayes
- d) None of the mentioned

Answer: b

2. Point out the correct statement.

- a) The choice of an appropriate metric will influence the shape of the clusters
- b) Hierarchical clustering is also called HCA
- c) In general, the merges and splits are determined in a greedy manner
- d) All of the mentioned

Answer: d

3. Which of the following is finally produced by Hierarchical Clustering?

- a) final estimate of cluster centroids
- b) tree showing how close things are to each other
- c) assignment of each point to clusters
- d) all of the mentioned

Answer: b

4. Which of the following is required by K-means clustering?

- a) defined distance metric
- b) number of clusters
- c) initial guess as to cluster centroids
- d) all of the mentioned

Answer: d

5. Point out the wrong statement.

- a) k-means clustering is a method of vector quantization
- b) k-means clustering aims to partition n observations into k clusters
- c) k-nearest neighbor is same as k-means
- d) none of the mentioned

Answer: c

6. Hierarchical clustering should be primarily used for exploration.

a) True

b) False

Answer: a

8. Which of the following function is used for k-means clustering?

a) k-means

b) k-mean

c) heatmap

d) none of the mentioned

Answer: a

9. Which of the following clustering requires merging approach?

a) Partitional

b) Hierarchical

c) Naive Bayes

d) None of the mentioned

Answer: b

10. K-means is not deterministic and it also consists of number of iterations.

a) True

b) False

Answer: a

This sheet is for 1 Mark questions							
S r N o	Question	Image	a	b	c	d	Cor rec t An sw er
	Write down question	img.j pg	Option a	Optio n b	Option c	Option d	a/b /c/ d
1	In reinforcement learning if feedback is negative one it is defined as_____.		Penalty	Over earn ing	Reward	None of above	A
2	According to_____, it's a key success factor for the survival and evolution of all species.		Claude Shanno n's theory	Gini Index	Darwin's theory	None of above	C
3	How can you avoid overfitting ?		By using a lot of data	By using induc tive mach ine learn ing	By using validatio n only	None of above	A
4	What are the popular algorithms of Machine Learning?		Decisio n Trees and Neural Networ ks (back propag ation)	Proba bilsti c netw orks and Near est Neigh bor	Support vector machines	All	D
5	What is 'Training set'?		Trainin g set is used to test the accurac y of the hypoth eses generat ed by the learner.	A set of data is used to disco ver the poten tially predi ctive	Both A & B	None of above	B

				relationship.			
6	Common deep learning applications include_____		Image classification, Real-time visual tracking	Autonomous car driving, Logistic optimization	Bioinformatics, Speech recognition	All above	D
7	what is the function of 'Supervised Learning'?		Classifications, Predict time series, Annotate strings	Speech recognition, Regression	Both A & B	None of above	C
8	Commons unsupervised applications include		Object segmentation	Similarity detection	Automatic labeling	All above	D
9	Reinforcement learning is particularly efficient when_____.		the environment is not completely deterministic	it's often very dynamic	it's impossible to have a precise error measure	All above	D
10	if there is only a discrete number of possible outcomes (called categories), the process becomes a_____.		Regression	Classification	Modelfree	Categories	B
11	Which of the following are supervised learning applications		Spam detection, Pattern detection, Natural Language Processing	Image classification, Real-time visual tracking	Autonomous car driving, Logistic optimization	Bioinformatics, Speech recognition	A
12	During the last few years, many _____ algorithms have been applied to deep neural networks to learn the best policy		Logical	Classical	Classification	None of above	D

	for playing Atari video games and to teach an agent how to associate the right action with an input representing the state.					
1 3	Which of the following sentence is correct?	Machin e learnin g relates with the study, design and develo pment of the algorith ms that give comput ers the capabili ty to learn without being explicitl y progra mmed.	Data minin g can be defin ed as the proce ss in which the unstr uctur ed data tries to extra ct know ledge or unkn own inter estin g patte rns.	Both A & B	None of the above	C
1 4	What is 'Overfitting' in Machine learning?	when a statistic al model describ es random error or noise instead of underly ing relation ship 'overfit ting' occurs.	Robo ts are progr amed so that they can perfo rm the task based on data they gathe r	While involvin g the process of learning 'overfitti ng' occurs.	a set of data is used to discover the potentia lly predictiv e relations hip	A

			from senso rs.			
1 5	What is 'Test set'?	Test set is used to test the accuracy of the hypotheses generated by the learner.	It is a set of data is used to discover the potentially predictive relationship.	Both A & B	None of above	A
1 6	_____ is much more difficult because it's necessary to determine a supervised strategy to train a model for each feature and, finally, to predict their value	Removing the whole line	Creating sub-model to predict those features	Using an automatic strategy to input them according to the other known values	All above	B
1 7	How it's possible to use a different placeholder through the parameter_____.	regression	classification	random_state	missing_values	D
1 8	If you need a more powerful scaling feature, with a superior control on outliers and the possibility to select a quantile range, there's also the class_____.	RobustScaler	DictVectorizer	LabelBinarizer	FeatureHasher	A
1 9	scikit-learn also provides a class for per-sample normalization, Normalizer. It can apply_____ to each element of a dataset	max, l0 and l1 norms	max, l1 and l2 norms	max, l2 and l3 norms	max, l3 and l4 norms	B
2 0	There are also many univariate methods that can be used in order to select the best features according to specific criteria based on_____.	F-tests and p-values	chi-square	ANOVA	All above	A
2 1	Which of the following selects only a subset of features belonging to a certain percentile	SelectPercentile	FeatureSelector	SelectKBest	All above	A

2 2	_____ performs a PCA with non-linearly separable data sets.		SparsePCA	Kerne IPCA	SVD	None of the Mentioned	B
2 3	A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college. Which of the following statement is true in following case?		Feature F1 is an example of nominal variable.	Feature F1 is an example of ordinal variable.	It doesn't belong to any of the above category.	Both of these	B
2 4	What would you do in PCA to get the same projection as SVD?		Transform data to zero mean	Transform data to zero median	Not possible	None of these	A
2 5	What is PCA, KPCA and ICA used for?		Principal Components Analysis	Kernel based Principal Component Analysis	Independent Component Analysis	All above	D
2 6	Can a model trained for item based similarity also choose from a given set of items?		YES	NO			A
2 7	What are common feature selection methods in regression task?		correlation coefficient	Greedy algorithms	All above	None of these	C
2 8	The parameter _____ allows specifying the percentage of elements to put into the test/training set		test_size	training_size	All above	None of these	C
2 9	In many classification problems, the target _____ is made up of categorical labels which cannot immediately be processed by any algorithm.		random_state	dataset	test_size	All above	B
3 0	_____ adopts a dictionary-oriented approach, associating to each category label a progressive integer number.		LabelEncoder class	LabelBinarizer class	DictVectorizer	Feature Hasher	A
3 1	If Linear regression model perfectly fit i.e., train error is zero, then		a) Test error is also	b) Test error	c) Couldn't comment	d) Test error is equal to	c

			always zero	is non zero	t on Test error	Train error	
3 2	Which of the following metrics can be used for evaluating regression models? i) R Squared ii) Adjusted R Squared iii) F Statistics iv) RMSE / MSE / MAE		a) ii and iv	b) i and ii	c) ii, iii and iv	d) i, ii, iii and iv	d
3 3	How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?		a) 1	b) 2	c) 3	d) 4	b
3 4	In a simple linear regression model (One independent variable), If we change the input variable by 1 unit. How much output variable will change?		a) by 1	b) no change	c) by intercept	d) by its slope	d
3 5	Function used for linear regression in R is _____		a) lm(formula, data)	b) lr(formula, data)	c) lrm(formula, data)	d) regression.linear(formula, data)	a
3 6	In syntax of linear model lm(formula,data,...), data refers to _____		a) Matrix	b) Vector	c) Array	d) List	b
3 7	In the mathematical Equation of Linear Regression $Y = \beta_1 + \beta_2X + \epsilon$, (β_1, β_2) refers to _____		a) (X-intercept, Slope)	b) (Slope, X-Intercept, Slope)	c) (Y-Intercept, Slope)	d) (slope, Y-Intercept)	c
3 8	Linear Regression is a supervised machine learning algorithm.		A) TRUE	B) FALSE			a
3 9	It is possible to design a Linear regression algorithm using a neural network?		A) TRUE	B) FALSE			a
4 0	Which of the following methods do we use to find the best fit line for data in Linear Regression?		A) Least Square Error	B) Maximum Likelihood	C) Logarithmic Loss	D) Both A and B	a
4 1	Which of the following evaluation metrics can be used to evaluate a model while modeling a continuous output variable?		A) AUC-ROC	B) Accuracy	C) Logloss	D) Mean-Squared-Error	d
4 2	Which of the following is true about Residuals ?		A) Lower is better	B) Higher is better	C) A or B depend on the situation	D) None of these	a
4 3	Overfitting is more likely when you have huge amount of data to train?		A) TRUE	B) FALSE			b
4 4	Which of the following statement is true about outliers in Linear regression?		A) Linear regression	B) Linear regression	C) Can't say	D) None of these	a

			sensitive to outliers	is not sensitive to outliers			
4 5	Suppose you plotted a scatter plot between the residuals and predicted values in linear regression and you found that there is a relationship between them. Which of the following conclusion do you make about this situation?		A) Since there is a relationship means our model is not good	B) Since there is a relationship means our model is good	C) Can't say	D) None of these	a
4 6	Naive Bayes classifiers are a collection - -----of algorithms		Classification	Clustering	Regression	All	a
4 7	Naive Bayes classifiers is _____ Learning		Supervised	Unsupervised	Both	None	a
4 8	Features being classified is independent of each other in Naïve Bayes Classifier		False	TRUE			b
4 9	Features being classified is _____ of each other in Naïve Bayes Classifier		Independent	Dependent	Partial Dependent	None	a
5 0	Bayes Theorem is given by where 1. P(H) is the probability of hypothesis H being true. 2. P(E) is the probability of the evidence(regardless of the hypothesis). 3. P(E H) is the probability of the evidence given that hypothesis is true. 4. P(H E) is the probability of the hypothesis given that the evidence is there.	bayes.jpg	True	FALSE			a
5 1	In given image, P(H E) is _____ probability.	bayes.jpg	Posterior	Prior			a
5 2	In given image, P(H) is _____ probability.	bayes.jpg	Posterior	Prior			b
5 3	Conditional probability is a measure of the probability of an event given that another event has already occurred.		True	FALSE			a
5 4	Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event.		True	FALSE			a

5	Bernoulli Naïve Bayes Classifier is _____ distribution		Continuous	Discrete	Binary		c
5	Multinomial Naïve Bayes Classifier is _____ distribution		Continuous	Discrete	Binary		b
5	Gaussian Naïve Bayes Classifier is _____ distribution		Continuous	Discrete	Binary		a
5	Binarize parameter in BernoulliNB scikit sets threshold for binarizing of sample features.		True	FALSE			a
5	<u>Gaussian distribution when plotted, gives a bell shaped curve which is symmetric about the _____ of the feature values.</u>		Mean	Variance	Discrete	Random	a
6	SVMs directly give us the posterior probabilities $P(y = 1 jx)$ and $P(y = \bar{1} jx)$		True	FALSE			b
6	Any linear combination of the components of a multivariate Gaussian is a univariate Gaussian.		True	FALSE			a
6	Solving a non linear separation problem with a hard margin Kernelized SVM (Gaussian RBF Kernel) might lead to overfitting		True	FALSE			a
6	SVM is a ----- algorithm		Classification	Clustering	Regression	All	a
6	SVM is a ----- learning		Supervised	Unsupervised	Both	None	a
6	The linear SVM classifier works by drawing a straight line between two classes		True	FALSE			a
6	Which of the following function provides unsupervised prediction ?	--	cl_forecastB	cl_no wcast C	cl_prectestD	None of the Mentioned	D
6	Which of the following is characteristic of best machine learning method ?	--	fast	accuracy	scalable	All above	D
6	What are the different Algorithm techniques in Machine Learning?	--	Supervised Learning and Semi-supervised Learning	Unsupervised Learning and Transformation	Both A & B	None of the Mentioned	C

6 9	What is the standard approach to supervised learning?	--	split the set of example into the training set and the test	group the set of example into the training set and the test	a set of observed instances tries to induce a general rule	learns programs from data	A
7 0	Which of the following is not Machine Learning?	--	Artificial Intelligence	Rule based inference	Both A & B	None of the Mentioned	B
7 1	What is Model Selection in Machine Learning?	--	The process of selecting models among different mathematical models, which are used to describe the same data set	when a statistical model describes random error or noise instead of underlying relationship	Find interesting directions in data and find novel observations/ database cleaning	All above	A
7 2	Which are two techniques of Machine Learning ?	--	Genetic Programming and Inductive Learning	Speech recognition and Regression	Both A & B	None of the Mentioned	A
7 3	Even if there are no actual supervisors _____ learning is also based on feedback provided by the environment	--	Supervised	Reinforcement	Unsupervised	None of the above	B

7 4	What does learning exactly mean?	--	Robots are programmed so that they can perform the task based on data they gather from sensors .	A set of data is used to discover the potentially predictive relationship.	Learning is the ability to change according to external stimuli and remembering most of all previous experiences.	It is a set of data is used to discover the potentially predictive relationship.	C
7 5	When it is necessary to allow the model to develop a generalization ability and avoid a common problem called _____.	--	Overfitting	Overlearning	Classification	Regression	A
7 6	Techniques involve the usage of both labeled and unlabeled data is called _____.	--	Supervised	Semi-supervised	Unsupervised	None of the above	B
7 7	In reinforcement learning if feedback is negative one it is defined as _____.	--	Penalty	Overlearning	Reward	None of above	A
7 8	According to _____ , it's a key success factor for the survival and evolution of all species.	--	Claude Shannon's theory	Gini Index	Darwin's theory	None of above	C
7 9	A supervised scenario is characterized by the concept of a _____.	--	Programmer	Teacher	Author	Farmer	B
8 0	overlearning causes due to an excessive _____.	--	Capacity	Regression	Reinforcement	Accuracy	A
8 1	Which of the following is an example of a deterministic algorithm?	--	PCA	K-Means	None of the above		A
8 2	Which of the following model include a backwards elimination feature selection routine?	--	MCV	MARS	MCRS	All above	B
8 3	Can we extract knowledge without apply feature selection	--	YES	NO			A
8 4	While using feature selection on the data, is the number of features decreases.	--	NO	YES			B
8 5	Which of the following are several models for feature extraction	--	regression	classification	None of the above		C

8 6	_____ provides some built-in datasets that can be used for testing purposes.	--	scikit-learn	classification	regression	None of the above	A
8 7	While using _____ all labels are turned into sequential numbers.	--	LabelEncoder class	LabelBinarizer class	DictVectorizer	FeatureHasher	A
8 8	_____ produce sparse matrices of real numbers that can be fed into any machine learning model.	--	DictVectorizer	FeatureHasher	Both A & B	None of the Mentioned	C
8 9	scikit-learn offers the class _____, which is responsible for filling the holes using a strategy based on the mean, median, or frequency	--	LabelEncoder	LabelBinarizer	DictVectorizer	Imputer	D
9 0	Which of the following scale data by removing elements that don't belong to a given range or by considering a maximum absolute value.	--	MinMaxScaler	MaxAbsScaler	Both A & B	None of the Mentioned	C
9 1	scikit-learn also provides a class for per-sample normalization, _____	--	Normalizer	Imputer	Classifier	All above	A
9 2	_____ dataset with many features contains information proportional to the independence of all features and their variance.	--	normalized	unnormalized	Both A & B	None of the Mentioned	B
9 3	In order to assess how much information is brought by each component, and the correlation among them, a useful tool is the _____.	--	Concurrent matrix	Convergence matrix	Supportive matrix	Covariance matrix	D
9 4	The _____ parameter can assume different values which determine how the data matrix is initially processed.	--	run	start	init	stop	C
9 5	_____ allows exploiting the natural sparsity of data while extracting principal components.	--	SparsePCA	KerneIPCA	SVD	init parameter	A
9 6	Which of the following evaluation metrics can be used to evaluate a model while modeling a continuous output variable?	--	AUC-ROC	Accuracy	Logloss	Mean-Squared-Error	D
9 7	Which of the following is true about Residuals ?	--	Lower is better	Higher is better	A or B depend on the situation	None of these	A
9 8	Overfitting is more likely when you have huge amount of data to train?	--	TRUE	FALSE			B

9	Which of the following statement is true about outliers in Linear regression?	--	Linear regression is sensitive to outliers	Linear regression is not sensitive to outliers	Can't say	None of these	A
100	Suppose you plotted a scatter plot between the residuals and predicted values in linear regression and you found that there is a relationship between them. Which of the following conclusion do you make about this situation?	--	Since there is a relationship means our model is not good	Since there is a relationship means our model is good	Can't say	None of these	A
101	Let's say, a "Linear regression" model perfectly fits the training data (train error is zero). Now, Which of the following statement is true?	--	You will always have test error zero	You can not have test error zero	None of the above		C
102	In a linear regression problem, we are using "R-squared" to measure goodness-of-fit. We add a feature in linear regression model and retrain the same model.Which of the following option is true?	--	If R Squared increases, this variable is significant.	If R Squared decreases, this variable is not significant.	Individually R squared cannot tell about variable importance. We can't say anything about it right now.	None of these.	C
103	Which of the one is true about Heteroskedasticity?	--	Linear Regression with varying error terms	Linear Regression with constant error terms	Linear Regression with zero error terms	None of these	A

1 0 4	Which of the following assumptions do we make while deriving linear regression parameters?1. The true relationship between dependent y and predictor x is linear2. The model errors are statistically independent3. The errors are normally distributed with a 0 mean and constant standard deviation4. The predictor x is non-stochastic and is measured error-free	--	1,2 and 3.	1,3 and 4.	1 and 3.	All of above.	D
1 0 5	To test linear relationship of y(dependent) and x(independent) continuous variables, which of the following plot best suited?	--	Scatter plot	Barchart	Histograms	None of these	A
1 0 6	which of the following step / assumption in regression modeling impacts the trade-off between under-fitting and over-fitting the most.	--	The polynomial degree	Whether we learn the weights by matrix inversion or gradient descent	The use of a constant -term		A
1 0 7	Can we calculate the skewness of variables based on mean and median?	--	TRUE	FALSE			B
1 0 8	Which of the following is true about “Ridge” or “Lasso” regression methods in case of feature selection?	--	Ridge regression uses subset selection of features	Lasso regression uses subset selection of features	Both use subset selection of features	None of above	B
1 0 9	Which of the following statement(s) can be true post adding a variable in a linear regression model?1. R-Squared and Adjusted R-squared both increase2. R-Squared increases and Adjusted R-squared decreases3. R-	--	1 and 2	1 and 3	2 and 4	None of the above	A

	Squared decreases and Adjusted R-squared decreases 4. R-Squared decreases and Adjusted R-squared increases						
1 1 0	How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?	--	1	2	Can't Say		B
1 1 1	In given image, $P(H)$ is _____ probability.	baye s.jpg	Posterior	Prior			B
1 1 2	Conditional probability is a measure of the probability of an event given that another event has already occurred.	--	True	FALSE			A
1 1 3	<u>Gaussian distribution when plotted, gives a bell shaped curve which is symmetric about the _____ of the feature values.</u>	--	Mean	Variance	Discrete	Random	A
1 1 4	SVMs directly give us the posterior probabilities $P(y = 1 x)$ and $P(y = -1 x)$	--	True	FALSE			B
1 1 5	SVM is a ----- algorithm	--	Classification	Clustering	Regression	All	A
1 1 6	What is/are true about kernel in SVM? 1. Kernel function map low dimensional data to high dimensional space 2. It's a similarity function	--	1	2	1 and 2	None of these	C
1 1 7	Suppose you are building a SVM model on data X. The data X can be error prone which means that you should not trust any specific data point too much. Now think that you want to build a SVM model which has quadratic kernel function of polynomial degree 2 that uses Slack variable C as one of it's hyper parameter. What would happen when you use very small C ($C \sim 0$)?		Misclassification would happen	Data will be correctly classified	Can't say	None of these	A
1 1 8	The cost parameter in the SVM means:	--	The number of cross-validations to be made	The kernel to be used	The tradeoff between misclassification and simplicity of the model	None of the above	C
1 1 9	Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event.	--	True	FALSE			A

1 2 0	Bernoulli Naïve Bayes Classifier is _____ distribution	--	Continuous	Discrete	Binary		C
1 2 1	If you remove the non-red circled points from the data, the decision boundary will change?	svm.jpg	TRUE	FALSE			B
1 2 2	How do you handle missing or corrupted data in a dataset?	--	a. Drop missing rows or column s	b. Replace missing values with mean /median/mode	c. Assign a unique category to missing values	d. All of the above	D
1 2 3	Binarize parameter in BernoulliNB scikit sets threshold for binarizing of sample features.	--	True	FALSE			A
1 2 4	Which of the following statements about Naive Bayes is incorrect?	--	A. Attributes are equally important.	B. Attributes are statistically dependent of one another given the class value.	C. Attributes are statistically independent of one another given the class value.	D. Attributes can be nominal or numeric	B
1 2 5	The SVM's are less effective when:	--	The data is linearly separable	The data is clean and ready to use	The data is noisy and contains overlapping points		C
1 2 6	Naive Bayes classifiers is _____ Learning	--	Supervised	Unsupervised	Both	None	A
1 2 7	Features being classified is independent of each other in Naïve Bayes Classifier	--	False	TRUE			B

1 2 8	Features being classified is _____ of each other in Naïve Bayes Classifier	--	Independent	Dependent	Partial Dependent	None	A
1 2 9	Bayes Theorem is given by where 1. P(H) is the probability of hypothesis H being true. 2. P(E) is the probability of the evidence(regardless of the hypothesis). 3. P(E H) is the probability of the evidence given that hypothesis is true. 4. P(H E) is the probability of the hypothesis given that the evidence is there.	bayes.jpg	True	FALSE			A
1 3 0	Any linear combination of the components of a multivariate Gaussian is a univariate Gaussian.	--	True	FALSE			A

This sheet is for 2 Mark questions		Image	a	b	c	d	
S.r No	Question						Correct Answer
e.g 1	Write down question	img.jpg	Option a	Option b	Option c	Option d	a/b /c/ d
1	A supervised scenario is characterized by the concept of a _____.		Programmer	Teacher	Author	Farmer	B
2	overlearning causes due to an excessive _____.		Capacity	Regression	Reinforcement	Accuracy	A
3	If there is only a discrete number of _____		Modelfree	Categories	Prediction	None of above	B

	possible outcomes called _____.						
4	What is the standard approach to supervised learning?		split the set of example into the training set and the test	group the set of example into the training set and the test	a set of observed instances tries to induce a general rule	learns programs from data	A
5	Some people are using the term _____ instead of prediction only to avoid the weird idea that machine learning is a sort of modern magic.		Inference	Interference	Accuracy	None of above	A
6	The term _____ can be freely used, but with the same meaning adopted in physics or system theory.		Accuracy	Cluster	Regression	Prediction	D
7	Which are two techniques of Machine Learning ?		Genetic Programming and Inductive Learning	Speech recognition and Regression	Both A & B	None of the Mentioned	A
8	Even if there are no actual supervisors _____ learning is also based on feedback provided by		Supervised	Reinforcement	Unsupervised	None of the above	B

	the environment						
9	Common deep learning applications / problems can also be solved using _____		Real-time visual object identification	Classic approaches	Automatic labeling	Bio-inspired adaptive systems	B
10	Identify the various approaches for machine learning.		Concept Vs Classification Learning	Symbolic Vs Statistical Learning	Inductive Vs Analytical Learning	All above	D
11	what is the function of 'Unsupervised Learning'?		Find clusters of the data and find low-dimensional representations of the data	Find interesting directions in data and find novel observations/ database cleaning	Interesting coordinates and correlations	All	D
12	What are the two methods used for the calibration in Supervised Learning?		Platt Calibration and Isotonic Regression	Statistics and Informal Retrieval			A
13	What is the standard approach to supervised learning?		split the set of example into the training set and the test	group the set of example into the training set and the test	a set of observed instances tries to induce a general rule	learns programs from data	A
14	Which of the following is not Machine Learning?		Artificial Intelligence	Rule based inference	Both A & B	None of the Mentioned	B
15	What is Model Selection in		The process of	when a statistical model	Find interesting directions	All above	A

	Machine Learning?		selecting models among different mathematical models, which are used to describe the same data set	describes random error or noise instead of underlying relationship	in data and find novel observations/ database cleaning		
16	_____ provides some built-in datasets that can be used for testing purposes.		scikit-learn	classification	regression	None of the above	A
17	While using _____ all labels are turned into sequential numbers.		LabelEncoder class	LabelBinarizer class	DictVectorizer	FeatureHasher	A
18	_____ produce sparse matrices of real numbers that can be fed into any machine learning model.		DictVectorizer	FeatureHasher	Both A & B	None of the Mentioned	C
19	scikit-learn offers the class _____, which is responsible for filling the holes using a strategy based on the mean, median, or frequency		LabelEncoder	LabelBinarizer	DictVectorizer	Imputer	D

20	Which of the following scale data by removing elements that don't belong to a given range or by considering a maximum absolute value.		MinMax Scaler	MaxAbsScaler	Both A & B	None of the Mentioned	C
21	Which of the following model include a backwards elimination feature selection routine?		MCV	MARS	MCRS	All above	B
22	Can we extract knowledge without apply feature selection		YES	NO			A
23	While using feature selection on the data, is the number of features decreases.		NO	YES			B
24	Which of the following are several models for feature extraction		regression	classification	None of the above		C
25	scikit-learn also provides a class for per-sample		Normalizer	Imputer	Classifier	All above	A

	normalization, _____						
26	_____ data set with many features contains information proportional to the independence of all features and their variance.		normalized	unnormalized	Both A & B	None of the Mentioned	B
27	In order to assess how much information is brought by each component, and the correlation among them, a useful tool is the _____.		Concurrent matrix	Convergence matrix	Supportive matrix	Covariance matrix	D
28	The _____ parameter can assume different values which determine how the data matrix is initially processed.		run	start	init	stop	C
29	_____ allows exploiting the natural sparsity of data while extracting principal components .		SparsePCA	KernelPCA	SVD	init parameter	A
30	Which of the		PCA	K-Means	None of the above		A

	following is an example of a deterministic algorithm?						
31	Let's say, a "Linear regression" model perfectly fits the training data (train error is zero). Now, Which of the following statement is true?		A. You will always have test error zero	B. You can not have test error zero	C. None of the above		c
32	In a linear regression problem, we are using "R-squared" to measure goodness-of-fit. We add a feature in linear regression model and retrain the same model.Which of the following option is true?		A. If R Squared increases, this variable is significant.	B. If R Squared decreases, this variable is not significant.	C. Individually R squared cannot tell about variable importance . We can't say anything about it right now.	D. None of these.	c
33	Which of the one is true about Heteroskedasticity?		A. Linear Regression with varying error terms	B. Linear Regression with constant error terms	C. Linear Regression with zero error terms	D. None of these	a
34	Which of the following assumptions do we make		A. 1,2 and 3.	B. 1,3 and 4.	C. 1 and 3.	D. All of above.	d

	while deriving linear regression parameters ?1. The true relationship between dependent y and predictor x is linear2. The model errors are statistically independent3. The errors are normally distributed with a 0 mean and constant standard deviation4. The predictor x is non-stochastic and is measured error-free						
35	To test linear relationship of y(dependent) and x(independent) continuous variables, which of the following plot best suited?		A. Scatter plot	B. Barchart	C. Histograms	D. None of these	a
36	Generally, which of the following method(s) is used for		A. 1 and 2	B. only 1	C. only 2	D. None of these.	b

	predicting continuous dependent variable?1. Linear Regression2 . Logistic Regression						
37	Suppose you are training a linear regression model. Now consider these points.1. Overfitting is more likely if we have less data2. Overfitting is more likely when the hypothesis space is small.Which of the above statement(s) are correct?		A. Both are False	B. 1 is False and 2 is True	C. 1 is True and 2 is False	D. Both are True	c
38	Suppose we fit “Lasso Regression” to a data set, which has 100 features ($X_1, X_2 \dots X_{100}$). Now, we rescale one of these feature by multiplying with 10 (say that feature is X_1), and then refit Lasso		A. It is more likely for X_1 to be excluded from the model	B. It is more likely for X_1 to be included in the model	C. Can't say	D. None of these	b

	regression with the same regularization parameter. Now, which of the following option will be correct?						
39	Which of the following is true about “Ridge” or “Lasso” regression methods in case of feature selection?		A. Ridge regression uses subset selection of features	B. Lasso regression uses subset selection of features	C. Both use subset selection of features	D. None of above	b
40	Which of the following statement(s) can be true post adding a variable in a linear regression model? 1. R-Squared and Adjusted R-squared both increase 2. R-Squared increases and Adjusted R-squared decreases 3. R-Squared decreases and Adjusted R-squared decreases 4. R-Squared		A. 1 and 2	B. 1 and 3	C. 2 and 4	D. None of the above	a

	decreases and Adjusted R-squared increases						
41	We can also compute the coefficient of linear regression with the help of an analytical method called “Normal Equation”. Which of the following is/are true about “Normal Equation”? 1. We don't have to choose the learning rate. 2. It becomes slow when number of features is very large. 3. No need to iterate		A. 1 and 2	B. 1 and 3.	C. 2 and 3.	D. 1,2 and 3.	d
42	How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?		A. 1	B. 2	C. Can't Say		b
43	If two variables are correlated,		A. Yes	B. No			b

	is it necessary that they have a linear relationship ?						
44	Correlated variables can have zero correlation coefficient. True or False?		A. True	B. False			a
45	Which of the following option is true regarding "Regression" and "Correlation" ?Note: y is dependent variable and x is independent variable.		A. The relationship is symmetric between x and y in both.	B. The relationship is not symmetric between x and y in both.	C. The relationship is not symmetric between x and y in case of correlation but in case of regression it is symmetric.	D. The relationship is symmetric between x and y in case of correlation but in case of regression it is not symmetric.	d
46	What is/are true about kernel in SVM?1. Kernel function map low dimensional data to high dimensional space2. It's a similarity function		1	2	1 and 2	None of these	c
47	Suppose you are building a SVM model on data X. The data X can be error prone which		Classification would happen	Data will be correctly classified	Can't say	None of these	a

	means that you should not trust any specific data point too much. Now think that you want to build a SVM model which has quadratic kernel function of polynomial degree 2 that uses Slack variable C as one of its hyper parameter. What would happen when you use very small C ($C \sim 0$)?					
48	Suppose you are using a Linear SVM classifier with 2 class classification problem. Now you have been given the following data in which some points are circled red that are representing support vectors. If you remove the	svm.jpg	yes	no		a

	following any one red points from the data. Does the decision boundary will change?					
49	If you remove the non-red circled points from the data, the decision boundary will change?	svm.jpg	TRUE	FALSE		b
50	When the C parameter is set to infinite, which of the following holds true?		The optimal hyperplane if exists, will be the one that completely separates the data	The soft-margin classifier will separate the data	None of the above	a
51	Suppose you are building a SVM model on data X. The data X can be error prone which means that you should not trust any specific data point too much. Now think that you want to build a SVM model which has quadratic		We can still classify data correctly for given setting of hyper parameter C	We can not classify data correctly for given setting of hyper parameter C	Can't Say	None of these

	kernel function of polynomial degree 2 that uses Slack variable C as one of its hyper parameter. What would happen when you use very large value of C(C->infinity)?						
52	SVM can solve linear and non-linear problems		TRUE	FALSE			a
53	The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points.		TRUE	FALSE			a
54	Hyperplanes are _____ boundaries that help classify the data points.		usual	decision	parallel		b
55	The _____ of the hyperplane depends		dimension	classification	reduction		a

	upon the number of features.						
56	Hyperplanes are decision boundaries that help classify the data points.		TRUE	FALSE			a
57	SVM algorithms use a set of mathematical functions that are defined as the kernel.		TRUE	FALSE			a
58	In SVM, Kernel function is used to map a lower dimensional data into a higher dimensional data.		TRUE	FALSE			a
59	<i>In SVR we try to fit the error within a certain threshold.</i>		TRUE	FALSE			a
60	When the C parameter is set to infinite, which of the following holds true?		The optimal hyperplane if exists, will be the one that completely separates the data	The soft-margin classifier will separate the data	None of the above		a
61	How do you handle missing or corrupted data in a dataset?		a. Drop missing rows or columns	b. Replace missing values with mean/m	c. Assign a unique category to missing values	d. All of the above	d

				edian/mode			
62	What is the purpose of performing cross-validation?		a. To assess the predictive performance of the models	b. To judge how the trained model performs outside the sample on test data	c. Both A and B		c
63	Which of the following is true about Naive Bayes ?		a. Assumes that all the features in a dataset are equally important	b. Assumes that all the features in a dataset are independent	c. Both A and B	d. None of the above option	c
64	Which of the following statements about Naive Bayes is incorrect?		A. Attributes are equally important.	B. Attributes are statistically independent of one another given the class value.	C. Attributes are statistically independent of one another given the class value.	D. Attributes can be nominal or numeric	b
65	Which of the following is not supervised learning?		PCA	Decision Tree	Naive Bayesian	Linear regression	a
66	How can you avoid overfitting ?	--	By using a lot of data	By using inductive machine learning	By using validation only	None of above	A
67	What are the popular algorithms of Machine Learning?	--	Decision Trees and Neural Networks (back	Probabilistic networks and Nearest Neighbor	Support vector machines	All	D

			propagation)				
68	What is 'Training set'?	--	Training set is used to test the accuracy of the hypotheses generated by the learner.	A set of data is used to discover the potentially predictive relationship.	Both A & B	None of above	B
69	Identify the various approaches for machine learning.	--	Concept Vs Classification Learning	Symbolic Vs Statistical Learning	Inductive Vs Analytical Learning	All above	D
70	what is the function of 'Unsupervised Learning'?	--	Find clusters of the data and find low-dimensional representations of the data	Find interesting directions in data and find novel observations/ database cleaning	Interesting coordinates and correlations	All	D
71	What are the two methods used for the calibration in Supervised Learning?	--	Platt Calibration and Isotonic Regression	Statistics and Informal Retrieval			A
72	_____ can be adopted when it's necessary to categorize a large amount of data with a few complete examples or when there's the	--	Supervised	Semi-supervised	Reinforcement	Clusters	B

	need to impose some constraints to a clustering algorithm.						
73	In reinforcement learning, this feedback is usually called as ____.	--	Overfitting	Overlearning	Reward	None of above	C
74	In the last decade, many researchers started training bigger and bigger models, built with several different layers that's why this approach is called ____.	--	Deep learning	Machine learning	Reinforcement learning	Unsupervised learning	A
75	there's a growing interest in pattern recognition and associative memories whose structure and functioning are similar to what happens in the neocortex. Such an approach also allows	--	Regression	Accuracy	Modelfree	Scalable	C

	simpler algorithms called _____						
76	_____ showed better performance than other approaches, even without a context-based model	--	Machine learning	Deep learning	Reinforcement learning	Supervised learning	B
77	Common deep learning applications / problems can also be solved using _____	--	Real-time visual object identification	Classic approaches	Automatic labeling	Bio-inspired adaptive systems	B
78	Some people are using the term _____ instead of prediction only to avoid the weird idea that machine learning is a sort of modern magic.	--	Inference	Interference	Accuracy	None of above	A
79	The term _____ can be freely used, but with the same meaning adopted in physics or system theory.	--	Accuracy	Cluster	Regression	Prediction	D
80	If there is only a discrete	--	Modelfree	Categories	Prediction	None of above	B

	number of possible outcomes called _____.						
81	A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college. Which of the following statement is true in following case?	--	Feature F1 is an example of nominal variable.	Feature F1 is an example of ordinal variable.	It doesn't belong to any of the above category.	Both of these	B
82	What would you do in PCA to get the same projection as SVD?	--	Transform data to zero mean	Transform data to zero median	Not possible	None of these	A
83	What is PCA, KPCA and ICA used for?	--	Principal Components Analysis	Kernel based Principal Component Analysis	Independent Component Analysis	All above	D
84	Can a model trained for item based similarity also choose from a given set of items?	--	YES	NO			A
85	What are common feature selection methods in regression task?	--	correlation coefficient	Greedy algorithms	All above	None of these	C

86	The parameter _____ allows specifying the percentage of elements to put into the test/training set	--	test_size	training_size	All above	None of these	C
87	In many classification problems, the target _____ is made up of categorical labels which cannot immediately be processed by any algorithm.	--	random_state	dataset	test_size	All above	B
88	_____ adopts a dictionary-oriented approach, associating to each category label a progressive integer number.	--	LabelEncoder class	LabelBinarizer class	DictVectorizer	FeatureHas her	A
89	_____ is much more difficult because it's necessary to determine a supervised strategy to train a model for each feature and, finally, to	--	Removing the whole line	Creating sub-model to predict those features	Using an automatic strategy to input them according to the other known values	All above	B

	predict their value						
90	How it's possible to use a different placeholder through the parameter_____.	--	regression	classification	random_state	missing_values	D
91	If you need a more powerful scaling feature, with a superior control on outliers and the possibility to select a quantile range, there's also the class_____.	--	RobustScaler	DictVectorizer	LabelBinarizer	FeatureHas her	A
92	scikit-learn also provides a class for per-sample normalization, Normalizer. It can apply_____ to each element of a dataset	--	max, l0 and l1 norms	max, l1 and l2 norms	max, l2 and l3 norms	max, l3 and l4 norms	B
93	There are also many univariate methods that can be used in	--	F-tests and p-values	chi-square	ANOVA	All above	A

	order to select the best features according to specific criteria based on _____.						
94	Which of the following selects only a subset of features belonging to a certain percentile	--	SelectPercentile	FeatureHasher	SelectKBest	All above	A
95	_____ performs a PCA with non-linearly separable data sets.	--	SparsePCA	KernelPCA	SVD	None of the Mentioned	B
96	If two variables are correlated, is it necessary that they have a linear relationship ?	--	Yes	No			B
97	Correlated variables can have zero correlation coefficient. True or False?	--	TRUE	FALSE			A
98	Suppose we fit "Lasso Regression" to a data set, which has 100 features	--	It is more likely for X1 to be excluded from the model	It is more likely for X1 to be included in the model	Can't say	None of these	B

	(X1,X2...X100). Now, we rescale one of these feature by multiplying with 10 (say that feature is X1), and then refit Lasso regression with the same regularization parameter. Now, which of the following option will be correct?						
99	If Linear regression model perfectly fit i.e., train error is zero, then _____	--	Test error is also always zero	Test error is non zero	Couldn't comment on Test error	Test error is equal to Train error	C
100	Which of the following metrics can be used for evaluating regression models?i) R Squaredii) Adjusted R Squarediii) F Statisticsiv) RMSE / MSE / MAE	--	ii and iv	i and ii	ii, iii and iv	i, ii, iii and iv	D
101	In syntax of linear model lm(formula, data,..),	--	Matrix	Vector	Array	List	B

	data refers to _____						
102	Linear Regression is a supervised machine learning algorithm.	--	TRUE	FALSE			A
103	It is possible to design a Linear regression algorithm using a neural network?	--	TRUE	FALSE			A
104	Which of the following methods do we use to find the best fit line for data in Linear Regression?	--	Least Square Error	Maximum Likelihood	Logarithmic Loss	Both A and B	A
105	Suppose you are training a linear regression model. Now consider these points.1. Overfitting is more likely if we have less data.2. Overfitting is more likely when the hypothesis space is small.Which of the above statement(s)	--	Both are False	1 is False and 2 is True	1 is True and 2 is False	Both are True	C

) are correct?						
106	We can also compute the coefficient of linear regression with the help of an analytical method called "Normal Equation". Which of the following is/are true about "Normal Equation"? 1. We don't have to choose the learning rate2. It becomes slow when number of features is very large3. No need to iterate	--	1 and 2	1 and 3.	2 and 3.	1,2 and 3.	D
107	Which of the following option is true regarding "Regression" and "Correlation" ?Note: y is dependent variable and x is independent variable.	--	The relationship is symmetric between x and y in both.	The relationship is not symmetric between x and y in both.	The relationship is not symmetric between x and y in case of correlation but in case of regression it is symmetric.	The relationship is symmetric between x and y in case of correlation but in case of regression it is not symmetric.	D

108	In a simple linear regression model (One independent variable), If we change the input variable by 1 unit. How much output variable will change?	--	by 1	no change	by intercept	by its slope	D
109	Generally, which of the following method(s) is used for predicting continuous dependent variable?1. Linear Regression2 . Logistic Regression	--	1 and 2	only 1	only 2	None of these.	B
110	How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?	--	1	2	3	4	B
111	Suppose you are building a SVM model on data X. The data X can be error prone which means that you should not trust any specific data point	--	We can still classify data correctly for given setting of hyper parameter C	We can not classify data correctly for given setting of hyper parameter C	Can't Say	None of these	A

	too much. Now think that you want to build a SVM model which has quadratic kernel function of polynomial degree 2 that uses Slack variable C as one of it's hyper parameter. What would happen when you use very large value of C(C->infinity)?	--					
112	SVM can solve linear and non-linear problems	--	TRUE	FALSE			A
113	The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points.	--	TRUE	FALSE			A
114	Hyperplanes are _____	--	usual	decision	parallel		B

	___ boundaries that help classify the data points.						
115	When the C parameter is set to infinite, which of the following holds true?	--	The optimal hyperplane if exists, will be the one that completely separates the data	The soft-margin classifier will separate the data	None of the above		A
116	SVM is a ----- learning	--	Supervised	Unsupervised	Both	None	A
117	The linear SVM classifier works by drawing a straight line between two classes	--	True	FALSE			A
118	In a real problem, you should check to see if the SVM is separable and then include slack variables if it is not separable.	--	TRUE	FALSE			B
119	Which of the following are real world applications of the SVM?	--	Text and Hypertext Categorization	Image Classification	Clustering of News Articles	All of the above	D
120	The ___ of the hyperplane depends	--	dimension	classification	reduction		A

	upon the number of features.						
121	Hyperplanes are decision boundaries that help classify the data points.	--	TRUE	FALSE			A
122	SVM algorithms use a set of mathematical functions that are defined as the kernel.	--	TRUE	FALSE			A
123	Naive Bayes classifiers are a collection --- ----- of algorithms	--	Classification	Clustering	Regression	All	A
124	In given image, $P(H E)$ is _____ probability .	bayes.jpg	Posterior	Prior			A
125	Solving a non linear separation problem with a hard margin Kernelized SVM (Gaussian RBF Kernel) might lead to overfitting		True	FALSE			A
126	100 people are at party. Given data gives information about how many wear pink or not,	man.jpg	TRUE	FALSE			A

	and if a man or not. Imagine a pink wearing guest leaves, was it a man?						
127	For the given weather data, Calculate probability of playing	weather data.jpg	0.4	0.64	0.29	0.75	B
128	In SVM, Kernel function is used to map a lower dimensional data into a higher dimensional data.	--	TRUE	FALSE			A
129	In SVR we try to fit the error within a certain threshold.	--	TRUE	FALSE			A
130	When the C parameter is set to infinite, which of the following holds true?	--	The optimal hyperplane if exists, will be the one that completely separates the data	The soft-margin classifier will separate the data	None of the above		A
Question	Image	a	b	c	d	Correct Answer	
Write down question	img.jpg	Option a	Option b	Option c	Option d	a/b/c/d	
Which of the following is		fast	accuracy	scalable	All above	D	

characteristic of best machine learning method ?						
What are the different Algorithm techniques in Machine Learning?		Supervised Learning and Semi-supervised Learning	Unsupervised Learning and Transduction	Both A & B	None of the Mentioned	C
_____ can be adopted when it's necessary to categorize a large amount of data with a few complete examples or when there's the need to impose some constraints to a clustering algorithm.		Supervised	Semi-supervised	Reinforcement	Clusters	B
In reinforcement learning, this feedback is usually called as _____.		Overfitting	Overlearning	Reward	None of above	C
In the last decade, many researchers started training bigger and bigger models, built with several different layers that's why this approach is		Deep learning	Machine learning	Reinforcement learning	Unsupervised learning	A

called _____. .						
What does learning exactly mean?		Robots are programmed so that they can perform the task based on data they gather from sensors.	A set of data is used to discover the potentially predictive relationship.	Learning is the ability to change according to external stimuli and remembering most of all previous experiences.	It is a set of data used to discover the potentially predictive relationship.	C
When it is necessary to allow the model to develop a generalization ability and avoid a common problem called _____.		Overfitting	Overlearning	Classification	Regression	A
Techniques involve the usage of both labeled and unlabeled data is called _____.		Supervised	Semi-supervised	Unsupervised	None of the above	B
there's a growing interest in pattern recognition and associative memories whose structure and functioning are similar to what happens in the		Regression	Accuracy	Model-free	Scalable	C

neocortex. Such an approach also allows simpler algorithms called						
_____ showed better performance than other approaches, even without a context-based model		Machine learning	Deep learning	Reinforcement learning	Supervised learning	B
Which of the following sentence is correct?	--	Machine learning relates with the study, design and development of the algorithms that give computers the capability to learn without being explicitly programmed.	Data mining can be defined as the process in which the unstructured data tries to extract knowledge or unknown interesting patterns.	Both A & B	None of the above	C
What is 'Overfitting' in Machine learning?	--	when a statistical model describes random error or noise instead of underlying relationships 'overfitting' occurs.	Robots are programmed so that they can perform the task based on data they gather from sensors.	While involving the process of learning 'overfitting' occurs.	a set of data is used to discover the potentially predictive relationship	A

What is ‘Test set’?	--	Test set is used to test the accuracy of the hypotheses generated by the learner.	It is a set of data used to discover the potentially predictive relationship.	Both A & B	None of above	A
what is the function of ‘Supervised Learning’?	--	Classifications, Predict time series, Annotate strings	Speech recognition, Regression	Both A & B	None of above	C
Commons unsupervised applications include	--	Object segmentation	Similarity detection	Automatic labeling	All above	D
Reinforcement learning is particularly efficient when_____.	--	the environment is not completely deterministic	it's often very dynamic	it's impossible to have a precise error measure	All above	D
During the last few years, many _____ algorithms have been applied to deep neural networks to learn the best policy for playing Atari video games and to teach an agent how to associate the right action with an input representing the state.	--	Logical	Classical	Classification	None of above	D

Common deep learning applications include _____	--	Image classification, Real-time visual tracking	Autonomous car driving, Logistic optimization	Bioinformatics, Speech recognition	All above	D
if there is only a discrete number of possible outcomes (called categories), the process becomes a _____.	--	Regression	Classification.	Modeltree	Categories	B
Which of the following are supervised learning applications	--	Spam detection, Pattern detection, Natural Language Processing	Image classification, Real-time visual tracking	Autonomous car driving, Logistic optimization	Bioinformatics, Speech recognition	A
Let's say, you are working with categorical feature(s) and you have not looked at the distribution of the categorical variable in the test data. You want to apply one hot encoding (OHE) on the categorical feature(s). What challenges you may face if you have	--	All categories of categorical variable are not present in the test dataset.	Frequency distribution of categories is different in train as compared to the test dataset.	Train and Test always have same distribution.	Both A and B	D

applied OHE on a categorical variable of train dataset?						
Which of the following sentence is FALSE regarding regression?	--	It relates inputs to outputs.	It is used for prediction.	It may be used for interpretation.	It discovers causal relationships.	D
Which of the following method is used to find the optimal features for cluster analysis	--	k-Means	Density-Based Spatial Clustering	Spectral Clustering Find clusters	All above	D
scikit-learn also provides functions for creating dummy datasets from scratch:	--	make_classification()	make_regression()	make blobs()	All above	D
_____ which can accept a NumPy RandomState generator or an integer seed.	--	make blobs	random_state	test_size	training_size	B
In many classification problems, the target dataset is made up of categorical labels which cannot immediately be processed	--	1	2	3	4	B

by any algorithm. An encoding is needed and scikit-learn offers at least ____ valid options						
In which of the following each categorical label is first turned into a positive integer and then transformed into a vector where only one feature is 1 while all the others are 0.	--	LabelEncoder class	DictVectorizer	LabelBinarizer class	FeatureHas her	C
_____ is the most drastic one and should be considered only when the dataset is quite large, the number of missing features is high, and any prediction could be risky.	--	Removing the whole line	Creating sub-model to predict those features	Using an automatic strategy to input them according to the other known values	All above	A
It's possible to specify if the scaling process must include both mean and standard deviation using the	--	with_mean=True/False	with_std=True/False	Both A & B	None of the Mentioned	C

parameters_						
Which of the following selects the best K high-score features.	--	SelectPercentile	FeatureHasher	SelectKBest	All above	C
How does number of observations influence overfitting? Choose the correct answer(s). Note: Rest all parameters are same1. In case of fewer observations, it is easy to overfit the data.2. In case of fewer observations, it is hard to overfit the data.3. In case of more observations, it is easy to overfit the data.4. In case of more observations, it is hard to overfit the data.	--	1 and 4	2 and 3	1 and 3	None of theses	A

<p>Suppose you have fitted a complex regression model on a dataset.</p> <p>Now, you are using Ridge regression with tuning parameter lambda to reduce its complexity . Choose the option(s) below which describes relationship p of bias and variance with lambda.</p>	--	In case of very large lambda; bias is low, variance is low	In case of very large lambda; bias is low, variance is high	In case of very large lambda; bias is high, variance is low	In case of very large lambda; bias is high, variance is high	C
<p>What is/are true about ridge regression?</p> <p>1. When lambda is 0, model works like linear regression model2. When lambda is 0, model doesn't work like linear regression model3. When lambda goes to</p>	--	1 and 3	1 and 4	2 and 3	2 and 4	A

infinity, we get very, very small coefficient s approaching 0. When lambda goes to infinity, we get very, very large coefficient s approaching infinity						
Which of the following method(s) does not have closed form solution for its coefficient s?	--	Ridge regression	Lasso	Both Ridge and Lasso	None of both	B
Function used for linear regression in R is	--	lm(formula, data)	lr(formula, data)	lrm(formula, data)	regression.lm(formula, data)	A
In the mathematical Equation of Linear Regression $Y = \beta_1 + \beta_2X + \epsilon$, (β_1, β_2) refers to	--	(X-intercept, Slope)	(Slope, X-Intercept)	(Y-Intercept, Slope)	(slope, Y-Intercept)	C

<p>Suppose that we have N independent variables (X_1, X_2, \dots, X_n) and dependent variable is Y. Now Imagine that you are applying linear regression by fitting the best fit line using least square error on this data. You found that correlation coefficient for one of its variable(Say X_1) with Y is -0.95. Which of the following is true for X_1?</p>	--	Relation between the X_1 and Y is weak	Relation between the X_1 and Y is strong	Relation between the X_1 and Y is neutral	Correlation can't judge the relationship	B
<p>We have been given a dataset with n records in which we have input attribute as x and output attribute as y. Suppose we use a</p>	--	Increase	Decrease	Remain constant	Can't Say	D

<p>linear regression method to model this data. To test our linear regressor, we split the data in training set and test set randomly. Now we increase the training set size gradually. As the training set size increases, what do you expect will happen with the mean training error?</p>						
<p>We have been given a dataset with n records in which we have input attribute as x and output attribute as y. Suppose we use a linear regression method to model this data. To test our linear</p>	--	<p>Bias increases and Variance increases</p>	<p>Bias decreases and Variance increases</p>	<p>Bias decreases and Variance decreases</p>	<p>Bias increases and Variance decreases</p>	D

<p>regressor, we split the data in training set and test set randomly. What do you expect will happen with bias and variance as you increase the size of training data?</p>	--	1 and 2	2 and 3	1 and 3	1, 2 and 3	A

Problem: Players will play if weather is sunny. Is this statement is correct?	weather data.jpg	TRUE	FALSE			A
Multinomial Naïve Bayes Classifier is _____ distribution		Continuous	Discrete	Binary		B
For the given weather data, Calculate probability of not playing	weather data.jpg	0.4	0.64	0.36	0.5	C
Suppose you have trained an SVM with linear decision boundary after training SVM, you correctly infer that your SVM model is under fitting. Which of the following option would you more likely to consider iterating SVM next time?	--	You want to increase your data points	You want to decrease your data points	You will try to calculate more variables	You will try to reduce the features	C

The minimum time complexity for training an SVM is O(n ²). According to this fact, what sizes of datasets are not best suited for SVM's?	--	Large datasets	Small datasets	Medium sized datasets	Size does not matter	A
The effectiveness of an SVM depends upon:	--	Selection of Kernel	Kernel Parameters	Soft Margin Parameter C	All of the above	D
What do you mean by generalization error in terms of the SVM?	--	How far the hyperplane is from the support vectors	How accurately the SVM can predict outcomes for unseen data	The threshold amount of error in an SVM		B
What do you mean by a hard margin?	--	The SVM allows very low error in classification	The SVM allows high amount of error in classification	None of the above		A
We usually use feature normalization before using the Gaussian kernel in SVM. What is true about feature	--	1	1 and 2	1 and 3	2 and 3	B

normalization? 1. We do feature normalization so that new feature will dominate other 2. Some times, feature normalization is not feasible in case of categorical variables 3. Feature normalization always helps when we use Gaussian kernel in SVM						
Support vectors are the data points that lie closest to the decision surface.	--	TRUE	FALSE			A
Which of the following is not supervised learning?	--	PCA	Decision Tree	Naive Bayesian	Linear regression	A
Suppose you are using RBF kernel in SVM with high Gamma value. What does	--	The model would consider even far away points from hyperplane	The model would consider only the points close to the hyperplane for	The model would not be affected by distance of points from hyperplane	None of the above	B

this signify?		ne for modeling	modeling	ne for modeling		
Gaussian Naïve Bayes Classifier is <u> </u> distribution	--	Continuous	Discrete	Binary		A
If I am using all features of my dataset and I achieve 100% accuracy on my training set, but ~70% on validation set, what should I look out for?	--	Underfitting	Nothing, the model is perfect	Overfitting		C
What is the purpose of performing cross-validation?	--	a. To assess the predictive performance of the models	b. To judge how the trained model performs outside the sample on test data	c. Both A and B		C

Which of the following is true about Naive Bayes ?	--	a. Assumes that all the features in a dataset are equally important	b. Assumes that all the features in a dataset are independent	c. Both A and B	d. None of the above option	C
Suppose you are using a Linear SVM classifier with 2 class classification problem. Now you have been given the following data in which some points are circled red that are representing support vectors.If you remove the following any one red points from the data. Does the decision boundary will change?	svm.jpg	yes	no			A
Linear SVMs have no	--	TRUE	FALSE			B

hyperparameters that need to be set by cross-validation						
For the given weather data, what is the probability that players will play if weather is sunny	weather data.jpg	0.5	0.26	0.73	0.6	D
100 people are at party. Given data gives information about how many wear pink or not, and if a man or not. Imagine a pink wearing guest leaves, what is the probability of being a man	man.jpg	0.4	0.2	0.6	0.45	B
Problem: Players will play if weather is sunny. Is this statement is correct?	weather data.jpg	TRUE	FALSE			a
For the given weather	weather data.jpg	0.4	0.64	0.29	0.75	b

data, Calculate probability of playing						
For the given weather data, Calculate probability of not playing	weather data.jpg	0.4	0.64	0.36	0.5	c
For the given weather data, what is the probability that players will play if weather is sunny	weather data.jpg	0.5	0.26	0.73	0.6	d
100 people are at party. Given data gives informatio n about how many wear pink or not, and if a man or not. Imagine a pink wearing guest leaves, what is the probability of being a man	man.jpg	0.4	0.2	0.6	0.45	b
100 people are at party.	man.jpg	TRUE	FALSE			a

Given data gives information about how many wear pink or not, and if a man or not. Imagine a pink wearing guest leaves, was it a man?						
What do you mean by generalization error in terms of the SVM?		How far the hyperplane is from the support vectors	How accurately the SVM can predict outcomes for unseen data	The threshold amount of error in an SVM		b
What do you mean by a hard margin?		The SVM allows very low error in classification	The SVM allows high amount of error in classification	None of the above		a
The minimum time complexity for training an SVM is O(n ²). According to this fact, what sizes of datasets are not best suited for SVM's?		Large datasets	Small datasets	Medium sized datasets	Size does not matter	a
The effectiveness of an SVM		Selection of Kernel	Kernel Parameters	Soft Margin Parameter C	All of the above	d

depends upon:						
Support vectors are the data points that lie closest to the decision surface.		TRUE	FALSE			a
The SVM's are less effective when:		The data is linearly separable	The data is clean and ready to use	The data is noisy and contains overlapping points		c
Suppose you are using RBF kernel in SVM with high Gamma value. What does this signify?		The model would consider even far away points from hyperplane for modeling	The model would consider only the points close to the hyperplane for modeling	The model would not be affected by distance of points from hyperplane for modeling	None of the above	b
The cost parameter in the SVM means:		The number of cross-validation s to be made	The kernel to be used	The tradeoff between misclassification and simplicity of the model	None of the above	c
If I am using all features of my dataset and I achieve 100% accuracy on my training set, but ~70% on validation set, what		Underfitting	Nothing, the model is perfect	Overfitting		c

should I look out for?						
Which of the following are real world applications of the SVM?		Text and Hypertext Categorization	Image Classification	Clustering of News Articles	All of the above	d
Suppose you have trained an SVM with linear decision boundary after training SVM, you correctly infer that your SVM model is under fitting. Which of the following option would you more likely to consider iterating SVM next time?		You want to increase your data points	You want to decrease your data points	You will try to calculate more variables	You will try to reduce the features	c
We usually use feature normalization before using the Gaussian kernel in SVM. What is true about feature normalization? 1. We do feature normalization so that new feature will	1	1 and 2	1 and 3	2 and 3		b

This sheet is for 1 Mark questions							
S.r No	Question	Image	a	b	c	d	Cor rec t An sw er
	Write down question	img.jpg	Option a	Option b	Option c	Option d	a/b /c/ d
1	In reinforcement learning if feedback is negative one it is defined as _____.		Penalty	Overlearning	Reward	None of above	A
2	According to _____, it's a key success factor for the survival and evolution of all species.		Claude Shannon's theory	Gini Index	Darwin's theory	None of above	C
3	How can you avoid overfitting ?		By using a lot of data	By using inductive machine learning	By using validation only	None of above	A
4	What are the popular algorithms of Machine Learning?		Decision Trees and Neural Networks (back propagation)	Probabilistic networks and Nearest Neighbor	Support vector machines	All	D
5	What is 'Training set'?		Training set is used to test the accuracy of the hypotheses generated by the learner.	A set of data is used to discover the potentially predictive relationship.	Both A & B	None of above	B

6	Common deep learning applications include _____		Image classification, Real-time visual tracking	Autonomous car driving, Logistic optimization	Bioinformatics, Speech recognition	All above	D
7	what is the function of ‘Supervised Learning’?		Classifications, Predict time series, Annotate strings	Speech recognition, Regression	Both A & B	None of above	C
8	Common unsupervised applications include		Object segmentation	Similarity detection	Automatic labeling	All above	D
9	Reinforcement learning is particularly efficient when_____.		the environment is not completely deterministic	it's often very dynamic	it's impossible to have a precise error measure	All above	D
10	if there is only a discrete number of possible outcomes (called categories), the process becomes a_____.		Regression	Classification.	Modelfree	Categories	B
11	Which of the following are supervised learning applications		Spam detection , Pattern detection , Natural Language Processing	Image classification, Real-time visual tracking	Autonomous car driving, Logistic optimization	Bioinformatics, Speech recognition	A
12	During the last few years, many _____ algorithms		Logical	Classical	Classification	None of above	D

	have been applied to deep neural networks to learn the best policy for playing Atari video games and to teach an agent how to associate the right action with an input representing the state.					
13	Which of the following sentence is correct?		Machine learning relates with the study, design and development of the algorithms that give computer s the capability to learn without being explicitly programmed.	Data mining can be defined as the process in which the unstructured data tries to extract knowledge or unknown interesting patterns.	Both A & B	None of the above C
14	What is ‘Overfitting’ in Machine learning?		when a statistical model describes random error or noise instead of underlying relations hip ‘overfitti	Robots are programme d so that they can perform the task based on data they gather from sensors.	While involving the process of learning ‘overfitting ’ occurs.	a set of data is used to discover the potentially predictive relationship A

			ng' occurs.				
15	What is 'Test set'?		Test set is used to test the accuracy of the hypotheses generated by the learner.	It is a set of data used to discover the potentially predictive relationship.	Both A & B	None of above	A
16	_____ is much more difficult because it's necessary to determine a supervised strategy to train a model for each feature and, finally, to predict their value		Removing the whole line	Creating sub-model to predict those features	Using an automatic strategy to input them according to the other known values	All above	B
17	How it's possible to use a different placeholder through the parameter ____.		regression	classification	random_state	missing_values	D
18	If you need a more powerful scaling feature, with a superior control on outliers and the possibility to select a quantile range, there's also the class ____.		RobustScaler	DictVectorizer	LabelBinarizer	FeatureHas	A

19	scikit-learn also provides a class for per-sample normalization, Normalizer. It can apply _____ to each element of a dataset		max, l0 and l1 norms	max, l1 and l2 norms	max, l2 and l3 norms	max, l3 and l4 norms	B
20	There are also many univariate methods that can be used in order to select the best features according to specific criteria based on_____.		F-tests and p-values	chi-square	ANOVA	All above	A
21	Which of the following selects only a subset of features belonging to a certain percentile		SelectPercentile	FeatureHasher	SelectKBest	All above	A
22	_____performs a PCA with non-linearly separable data sets.		SparsePCA	KernelPCA	SVD	None of the Mentioned	B
23	A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college.		Feature F1 is an example of nominal variable.	Feature F1 is an example of ordinal variable.	It doesn't belong to any of the above category.	Both of these	B

	Which of the following statement is true in following case?						
24	What would you do in PCA to get the same projection as SVD?		Transform data to zero mean	Transform data to zero median	Not possible	None of these	A
25	What is PCA, KPCA and ICA used for?		Principal Components Analysis	Kernel based Principal Component Analysis	Independent Component Analysis	All above	D
26	Can a model trained for item based similarity also choose from a given set of items?		YES	NO			A
27	What are common feature selection methods in regression task?		correlation coefficient	Greedy algorithms	All above	None of these	C
28	The parameter _____ allows specifying the percentage of elements to put into the test/training set		test_size	training_size	All above	None of these	C
29	In many classification problems, the target _____ is made up of categorical labels which cannot immediately		random_state	dataset	test_size	All above	B

	be processed by any algorithm.						
30	_____adopts a dictionary-oriented approach, associating to each category label a progressive integer number.		LabelEncoder class	LabelBinarizer class	DictVectorizer	FeatureHas her	A
31	If Linear regression model perfectly fits i.e., train error is zero, then _____		a) Test error is also always zero	b) Test error is non zero	c) Couldn't comment on Test error	d) Test error is equal to Train error	c
32	Which of the following metrics can be used for evaluating regression models? i) R Squared ii) Adjusted R Squared iii) F Statistics iv) RMSE / MSE / MAE		a) ii and iv	b) i and ii	c) ii, iii and iv	d) i, ii, iii and iv	d
33	How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?		a) 1	b) 2	c) 3	d) 4	b
34	In a simple linear		a) by 1	b) no change	c) by intercept	d) by its slope	d

	regression model (One independent variable), If we change the input variable by 1 unit. How much output variable will change?						
35	Function used for linear regression in R is _____		a) lm(formula, data)	b) lr(formula, data)	c) lrm(formula, data)	d) regression.linear(formula, data)	a
36	In syntax of linear model lm(formula, data,...), data refers to _____		a) Matrix	b) Vector	c) Array	d) List	b
37	In the mathematical Equation of Linear Regression $Y = \beta_1 + \beta_2X + \epsilon$, (β_1, β_2) refers to _____		a) (X-intercept, Slope)	b) (Slope, X-Intercept)	c) (Y-Intercept, Slope)	d) (slope, Y-Intercept)	c
38	Linear Regression is a supervised machine learning algorithm.		A) TRUE	B) FALSE			a
39	It is possible to design a Linear regression algorithm using a neural network?		A) TRUE	B) FALSE			a

40	Which of the following methods do we use to find the best fit line for data in Linear Regression?		A) Least Square Error	B) Maximum Likelihood	C) Logarithmic Loss	D) Both A and B	a
41	Which of the following evaluation metrics can be used to evaluate a model while modeling a continuous output variable?		A) AUC-ROC	B) Accuracy	C) Logloss	D) Mean-Squared-Error	d
42	Which of the following is true about Residuals ?		A) Lower is better	B) Higher is better	C) A or B depend on the situation	D) None of these	a
43	Overfitting is more likely when you have huge amount of data to train?		A) TRUE	B) FALSE			b
44	Which of the following statement is true about outliers in Linear regression?		A) Linear regression is sensitive to outliers	B) Linear regression is not sensitive to outliers	C) Can't say	D) None of these	a
45	Suppose you plotted a scatter plot between the residuals and		A) Since there is a relationship means our	B) Since there is a relationship means our	C) Can't say	D) None of these	a

	predicted values in linear regression and you found that there is a relationship between them. Which of the following conclusion do you make about this situation?		model is not good	model is good			
46	Naive Bayes classifiers are a collection --- ----- of algorithms		Classification	Clustering	Regression	All	a
47	Naive Bayes classifiers is _____ Learning		Supervised	Unsupervised	Both	None	a
48	Features being classified is independent of each other in Naïve Bayes Classifier		False	TRUE			b
49	Features being classified is _____ of each other in Naïve Bayes Classifier		Independent	Dependent	Partial Dependent	None	a
50	Bayes Theorem is given by where 1. $P(H)$ is the	bayes.jpg	True	FALSE			a

	<p>probability of hypothesis H being true.</p> <p>2. $P(E)$ is the probability of the evidence (regardless of the hypothesis).</p> <p>3. $P(E H)$ is the probability of the evidence given that hypothesis is true.</p> <p>4. $P(H E)$ is the probability of the hypothesis given that the evidence is there.</p>					
51	In given image, $P(H E)$ is _____ probability .	bayes.jpg	Posterior	Prior		a
52	In given image, $P(H)$ is _____ probability .	bayes.jpg	Posterior	Prior		b
53	Conditional probability is a measure of the probability of an event given that another event has already occurred.		True	FALSE		a

54	Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event.		True	FALSE			a
55	Bernoulli Naïve Bayes Classifier is _____ distribution		Continuous	Discrete	Binary		c
56	Multinomial Naïve Bayes Classifier is _____ distribution		Continuous	Discrete	Binary		b
57	Gaussian Naïve Bayes Classifier is _____ distribution		Continuous	Discrete	Binary		a
58	Binarize parameter in BernoulliNB scikit sets threshold for binarizing of sample features.		True	FALSE			a
59	Gaussian distribution when plotted, gives a bell shaped curve which is symmetric		Mean	Variance	Discrete	Random	a

	about the _____ of the feature values.						
60	SVMs directly give us the posterior probabilities $P(y = 1 x)$ and $P(y = -1 x)$		True	FALSE			b
61	Any linear combination of the components of a multivariate Gaussian is a univariate Gaussian.		True	FALSE			a
62	Solving a non linear separation problem with a hard margin Kernelized SVM (Gaussian RBF Kernel) might lead to overfitting		True	FALSE			a
63	SVM is a ---- ----- algorithm		Classification	Clustering	Regression	All	a
64	SVM is a ---- ----- learning		Supervised	Unsupervised	Both	None	a
65	The linear SVM classifier works by drawing a straight line between two classes		True	FALSE			a
66	Which of the following	--	cl_forecastB	cl_nowcastC	cl_prestcastD	None of the Mentioned	D

	function provides unsupervised prediction ?						
67	Which of the following is characteristic of best machine learning method ?	--	fast	accuracy	scalable	All above	D
68	What are the different Algorithm techniques in Machine Learning?	--	Supervised Learning and Semi-supervised Learning	Unsupervised Learning and Transduction	Both A & B	None of the Mentioned	C
69	What is the standard approach to supervised learning?	--	split the set of example into the training set and the test	group the set of example into the training set and the test	a set of observed instances tries to induce a general rule	learns programs from data	A
70	Which of the following is not Machine Learning?	--	Artificial Intelligence	Rule based inference	Both A & B	None of the Mentioned	B
71	What is Model Selection in Machine Learning?	--	The process of selecting models among different mathematical models, which	when a statistical model describes random error or noise instead of underlying	Find interesting directions in data and find novel observations/ database cleaning	All above	A

			are used to describe the same data set	relationship			
72	Which are two techniques of Machine Learning ?	--	Genetic Programming and Inductive Learning	Speech recognition and Regression	Both A & B	None of the Mentioned	A
73	Even if there are no actual supervisors learning is also based on feedback provided by the environment	--	Supervised	Reinforcement	Unsupervised	None of the above	B
74	What does learning exactly mean?	--	Robots are programmed so that they can perform the task based on data they gather from sensors.	A set of data is used to discover the potentially predictive relationship.	Learning is the ability to change according to external stimuli and remembering most of all previous experiences.	It is a set of data is used to discover the potentially predictive relationship.	C
75	When it is necessary to allow the model to develop	--	Overfitting	Overlearning	Classification	Regression	A

	a generalization ability and avoid a common problem called _____.						
76	Techniques involve the usage of both labeled and unlabeled data is called _____.	--	Supervised	Semi-supervised	Unsupervised	None of the above	B
77	In reinforcement learning if feedback is negative one it is defined as _____.	--	Penalty	Overlearning	Reward	None of above	A
78	According to _____ , it's a key success factor for the survival and evolution of all species.	--	Claude Shannon's theory	Gini Index	Darwin's theory	None of above	C
79	A supervised scenario is characterized by the concept of a _____.	--	Programmer	Teacher	Author	Farmer	B
80	overlearning causes due to an excessive _____.	--	Capacity	Regression	Reinforcement	Accuracy	A
81	Which of the following	--	PCA	K-Means	None of the above		A

	is an example of a deterministic algorithm?						
82	Which of the following model include a backwards elimination feature selection routine?	--	MCV	MARS	MCRS	All above	B
83	Can we extract knowledge without apply feature selection	--	YES	NO			A
84	While using feature selection on the data, is the number of features decreases.	--	NO	YES			B
85	Which of the following are several models for feature extraction	--	regression	classification	None of the above		C
86	_____ provides some built-in datasets that can be used for testing purposes.	--	scikit-learn	classification	regression	None of the above	A

87	While using _____ all labels are turned into sequential numbers.	--	LabelEncoder class	LabelBinarizer class	DictVectorizer	FeatureHasher	A
88	_____ produce sparse matrices of real numbers that can be fed into any machine learning model.	--	DictVectorizer	FeatureHasher	Both A & B	None of the Mentioned	C
89	scikit-learn offers the class _____, which is responsible for filling the holes using a strategy based on the mean, median, or frequency	--	LabelEncoder	LabelBinarizer	DictVectorizer	Imputer	D
90	Which of the following scale data by removing elements that don't belong to a given range or by considering a maximum absolute value.	--	MinMaxScaler	MaxAbsScaler	Both A & B	None of the Mentioned	C

91	scikit-learn also provides a class for per-sample normalization, _____	--	Normalizer	Imputer	Classifier	All above	A
92	_____ data set with many features contains information proportional to the independence of all features and their variance.	--	normalized	unnormalized	Both A & B	None of the Mentioned	B
93	In order to assess how much information is brought by each component, and the correlation among them, a useful tool is the _____.	--	Concurrent matrix	Convergence matrix	Supportive matrix	Covariance matrix	D
94	The _____ parameter can assume different values which determine how the data matrix is initially processed.	--	run	start	init	stop	C
95	_____ allows exploiting	--	SparsePCA	KernelPCA	SVD	init parameter	A

	the natural sparsity of data while extracting principal components.						
96	Which of the following evaluation metrics can be used to evaluate a model while modeling a continuous output variable?	--	AUC-ROC	Accuracy	Logloss	Mean-Squared-Error	D
97	Which of the following is true about Residuals ?	--	Lower is better	Higher is better	A or B depend on the situation	None of these	A
98	Overfitting is more likely when you have huge amount of data to train?	--	TRUE	FALSE			B
99	Which of the following statement is true about outliers in Linear regression?	--	Linear regression is sensitive to outliers	Linear regression is not sensitive to outliers	Can't say	None of these	A
100	Suppose you plotted a scatter plot between the residuals and predicted values in linear regression and you found that there is a relationship	--	Since the there is a relationship means our model is not good	Since the there is a relationship means our model is good	Can't say	None of these	A

	between them. Which of the following conclusion do you make about this situation?						
101	Let's say, a "Linear regression" model perfectly fits the training data (train error is zero). Now, Which of the following statement is true?	--	You will always have test error zero	You can not have test error zero	None of the above		C
102	In a linear regression problem, we are using "R-squared" to measure goodness-of-fit. We add a feature in linear regression model and retrain the same model.Which of the following option is true?	--	If R Squared increases , this variable is significant.	If R Squared decreases , this variable is not significant.	Individualy R squared cannot tell about variable importance . We can't say anything about it right now.	None of these.	C
103	Which of the one is true about Heteroskedasticity?	--	Linear Regression with varying error terms	Linear Regression with constant error terms	Linear Regression with zero error terms	None of these	A
104	Which of the following assumptions	--	1,2 and 3.	1,3 and 4.	1 and 3.	All of above.	D

	<p>do we make while deriving linear regression parameters?</p> <ol style="list-style-type: none"> 1. The true relationship between dependent y and predictor x is linear 2. The model errors are statistically independent 3. The errors are normally distributed with a 0 mean and constant standard deviation 4. The predictor x is non-stochastic and is measured error-free 						
105	To test linear relationship of y(dependent) and x(independent) continuous variables, which of the following plot best suited?	--	Scatter plot	Barchart	Histograms	None of these	A

106	which of the following step / assumption in regression modeling impacts the trade-off between under-fitting and over-fitting the most.	--	The polynomial degree	Whether we learn the weights by matrix inversion or gradient descent	The use of a constant-term		A
107	Can we calculate the skewness of variables based on mean and median?	--	TRUE	FALSE			B
108	Which of the following is true about “Ridge” or “Lasso” regression methods in case of feature selection?	--	Ridge regression uses subset selection of features	Lasso regression uses subset selection of features	Both use subset selection of features	None of above	B
109	Which of the following statement(s) can be true post adding a variable in a linear regression model?1. R-Squared and Adjusted R-squared both increase2. R-Squared increases and	--	1 and 2	1 and 3	2 and 4	None of the above	A

	Adjusted R-squared decreases3. R-Squared decreases and Adjusted R-squared decreases4. R-Squared decreases and Adjusted R-squared increases						
110	How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?	--	1	2	Can't Say		B
111	In given image, $P(H)$ is _____ probability.	bayes.jpg	Posterior	Prior			B
112	Conditional probability is a measure of the probability of an event given that another event has already occurred.	--	True	FALSE			A
113	Gaussian distribution when plotted, gives a bell shaped curve which is	--	Mean	Variance	Discrete	Random	A

	symmetric about the _____ of the feature values.						
114	SVMs directly give us the posterior probabilities $P(y = 1 jx)$ and $P(y = -1 jx)$	--	True	FALSE			B
115	SVM is a - ----- --- algorithm	--	Classification	Clustering	Regression	All	A
116	What is/are true about kernel in SVM?1. Kernel function map low dimensional data to high dimensional space2. It's a similarity function	--	1	2	1 and 2	None of these	C
117	Suppose you are building a SVM model on data X. The data X can be error prone which means that you should not trust any specific data point too much. Now think	--	Misclassification would happen	Data will be correctly classified	Can't say	None of these	A

	that you want to build a SVM model which has quadratic kernel function of polynomial degree 2 that uses Slack variable C as one of it's hyper parameter. What would happen when you use very small C ($C \sim 0$)?						
118	The cost parameter in the SVM means:	--	The number of cross-validations to be made	The kernel to be used	The tradeoff between misclassification and simplicity of the model	None of the above	C
119	Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event.	--	True	FALSE			A
120	Bernoulli Naïve	--	Continuous	Discrete	Binary		C

	Bayes Classifier is _____ distribution						
121	If you remove the non-red circled points from the data, the decision boundary will change?	svm.jpg	TRUE	FALSE			B
122	How do you handle missing or corrupted data in a dataset?	--	a. Drop missing rows or columns b. Replace missing values with mean/median/mode c. Assign a unique category to missing values d. All of the above				D
123	Binarize parameter in BernoulliNB scikit sets threshold for binarizing of sample features.	--	True	FALSE			A
124	Which of the following statements about Naive Bayes is incorrect?	--	A. Attributes are equally important. B. Attributes are statistically dependent of one another given the class value. C. Attributes are statistically independent of one another given the class value. D. Attributes can be nominal or numeric				B

125	The SVM's are less effective when:	--	The data is linearly separable	The data is clean and ready to use	The data is noisy and contains overlapping points		C
126	Naive Bayes classifiers is _____ Learning	--	Supervised	Unsupervised	Both	None	A
127	Features being classified is independent of each other in Naïve Bayes Classifier	--	False	TRUE			B
128	Features being classified is _____ of each other in Naïve Bayes Classifier	--	Independent	Dependent	Partial Dependent	None	A
129	Bayes Theorem is given by where 1. P(H) is the probability of hypothesis H being true. 2. P(E) is the probability of the evidence(regardless of the hypothesis).	bayes.jpg	True	FALSE			A

	3. $P(E H)$ is the probability of the evidence given that hypothesis is true. 4. $P(H E)$ is the probability of the hypothesis given that the evidence is there.					
130	Any linear combination of the components of a multivariate Gaussian is a univariate Gaussian.	--	True	FALSE		A

Multiple Choice Questions

1. Supervised learning and unsupervised clustering both require at least one
 - A. **Hidden attribute**
 - B. Output attribute
 - C. Input attribute
 - D. Categorical attribute

2. Supervised learning and unsupervised clustering in that supervised learning requires
 - A. At least one input attribute
 - B. **Input attribute to be categorical**
 - C. At least one output attribute
 - D. output attribute to be categorical

3. Select all multi class classification techniques from given techniques
 - (A) One-versus-all (OVA)
 - (B) All-versus-all (AVA)
 - (C) Error-Correcting Output-Coding (ECOC)
 - (D) None of these

4. Causes of overfitting
 1. Small training dataset
 2. Large number of features in a dataset
 3. Noise in the dataset
 4. None of the Above

5. Which of the following is characteristic of best machine learning method?
 1. **Fast**
 2. **Accuracy**
 3. **Scalable**
 4. **All of the Mentioned**

6. Different learning methods does not include?
 - a) Memorization
 - b) Analogy
 - c) Deduction
 - d) **Introduction**

7. Which of the factors affect the performance of learner system does not include?
 - a) Representation scheme used
 - b) Training scenario
 - c) Type of feedback
 - d) **Good data structures**

- 8.. Which of the following is a categorical outcome?
 - a) **RMSE**
 - b) RSquared

- c) Accuracy
- d) All of the mentioned

9. Point out the wrong combination.

- a) True negative=correctly rejected
- b) False negative=correctly rejected
- c) False positive=correctly identified**
- d) All of the mentioned

10. Which of the following is a common error measure?

- a) Sensitivity
- b) Median absolute deviation
- c) Specificity
- d) All of the mentioned**

11. Predictive analytics is same as forecasting.

- a) True
- b) False**

12 Maximum a posteriori classifier is also known as:

- A. Decision tree classifier
- B. Bayes classifier**
- C. Gaussian classifier
- D. Maximum margin classifier

13. Which of the following is an example of continuous attribute?

- A. Weight of a person**
- B. Shoe size of a person
- C. Gender of a person
- D. None of the above

14. Rows of a data matrix storing record data usually represents?

- A. Metadata
- B. Objects**
- C. Attributes
- D. Aggregates

15. Sales database of items in a supermarket can be considered as an example of:

- A. Record data**
- B. Tree data
- C. Graph data
- D. None of the above

16. User rating given to a movie in a scale 1-10, can be considered as an attribute of type?

- A. Nominal
- B. Ordinal**
- C. Interval

D. Ratio

17.. Name of a movie, can be considered as an attribute of type?

- A. Nominal
- B. Ordinal
- C. Interval
- D. Ratio

18 Average squared difference between classifier predicted output and actual output is

- A. Mean Squared error
- B. Root mean squared error
- C. Mean absolute error
- D. Mean Relative error

19. Machine learning differs from statistical techniques in that machine learning methods

- A. Typically assume an underlying distribution of data
- B. are better able to deal with missing and noisy data
- C. are not able to explain their behavior
- D. None of the above.

20. Suppose a model is overfitting, which is not a valid way to reduce overfitting

- A. Increase the amount of training data
- B. Decrease the model complexity
- C. Reduce the noise of training data
- D. Improve the optimization algorithm being used for error minimization

21. WHICH OF THE FOLLOWING STATEMENTS IS/ARE TRUE ABOUT "TYPE-1" AND "TYPE-2" ERRORS?

- 1 TYPE1 IS KNOWN AS FALSE POSITIVE AND TYPE2 IS KNOWN AS FALSE NEGATIVE.
- 2 TYPE1 IS KNOWN AS FALSE NEGATIVE AND TYPE2 IS KNOWN AS FALSE POSITIVE.
- 3 TYPE1 ERROR OCCURS WHEN WE REJECT A NULL HYPOTHESIS WHEN IT IS ACTUALLY

- A 1 and 2
- B 1 and 3
- C 1

22. How many types are available in machine learning

- A 2
- B 4
- C 1
- D 3

23.Which of the following is the model used for learning:

- A Neural networks
- B Prepositional and FOL rules
- C All of the above
- D Decision tree

24.Automated vehicle is an example of

- A Unsupervised learning
- B None of above
- C Active learning
- D Supervised learning

25. Which Statement is true about prediction problems

- A. Only one independent variable
- B. More than one independent variable
- C. . More than one dependent variable
- D. None of the above

26. Adaptive system

1. Uses machine-learning techniques in which program can learn from past experience and adapt themselves to new situations.
2. Is a computational procedure that takes some value as input and produces some value as output.
3. Is a science of making machines performs tasks that would require intelligence when performed by humans
4. None of these

ANSWER: 1

27. In the representation of machine learning algorithm with $Y = f(X) + e$, e represents ;

1. Reducible error specifying, model not having enough attributes to sufficiently characterize the best mapping from X to Y.
2. Irreducible error specifying, model not having enough attributes to sufficiently characterize the best mapping from X to Y.
3. Propagation error specifying, model not having enough attributes to sufficiently characterize the best mapping from X to Y.
4. Transmission error specifying, model not having enough attributes to sufficiently characterize the best mapping from X to Y.

ANSWER: 2

28. Which of the following is a supervised learning problem? (multiple options may be correct)

1. Predicting credit approval based on historical data
2. Grouping people in a social network.
3. Predicting the gender of a person from his/her image. You are given the data of 1 Million

4. Images along the gender.
5. Given the class labels of old news articles, predicting the class of a new news article from its content. Class of a news article can be such as sports, politics, technology, etc.

ANSWER: (1), (3), (4)

29. Which of the following are classification problems? (multiple options may be correct)
1. Predicting the temperature (in Celsius) of a room from other environmental features (such as atmospheric pressure, humidity etc).
 2. Predicting if a cricket player is a batsman or bowler given his playing records.
 3. Finding the shorter route between two existing routes between two points.
 4. Predicting if a particular route between two points has traffic jam or not based on the travel time of vehicles.
 5. Filtering of spam messages

ANSWER : (2),(4), (5)

30. Which of the following is an unsupervised task?
1. Learning to play chess.
 2. Predicting if an edible item is sweet or spicy based on the information of the ingredients and their quantities.
 3. Grouping related documents from an unannotated corpus.
 4. all of the above

ANSWER : (3)

31. Which of the following are true about bias and variance of overfitted and underfitted models? (multiple options may be correct)
1. Underfitted models have high bias.
 2. Underfitted models have low bias.
 3. Overfitted models have high variance.
 4. Overfitted models have low variance.
 5. none of these

ANSWER : (1), (3)

32. Which of the following is a categorical feature?
1. Number of legs of an animal
 2. Number of hours you study in a day
 3. Branch of an engineering student
 4. Your weekly expenditure in rupees.
 5. Ethnicity of a person
 6. Height of a person in inches

ANSWER : (3) and (5)

33. Which of the following is a regression task? (multiple options may be correct)
1. Predicting the monthly sales of a cloth store in rupees.
 2. Predicting if a user would like to listen to a newly released song or not based on historical data.
 3. Predicting the confirmation probability (in fraction) of your train ticket whose current status is waiting list based on historical data.
 4. Predicting if a patient has diabetes or not based on historical medical records.
 5. Predicting the gender of a human

ANSWER : (1) and (3)

34. What happens when your model complexity increases?

1. Model bias increases
2. Model bias decreases
3. Variance of the model increases
4. Variance of the model decreases

ANSWER : 1

35. Supervised learning problems can be further grouped into

1. Regression and classification problems
2. Clustering and association problems
3. Both of the mentioned
4. None of the mentioned

ANSWER : 1

36. Unsupervised learning problems can be further grouped into

1. Regression and classification problems
2. Clustering and association problems.
3. Both of the mentioned
4. None of the mentioned

ANSWER : 2

37. Problems where you have a large amount of input data (X) and only some of the data is labeled (Y) are called

1. Un supervised learning problems
2. Supervised learning problems.
3. Semi supervised learning problems
4. None of the mentioned

ANSWER : 3

38. Reinforcement learning is based on _____ provided by the environment. This _____ is usually called reward .

1. Positive feedback, Feedback
2. Feedback, Positive feedback
3. Feedback, Negative feedback

ANSWER : 2

39. Machine learning concerned with algorithms inspired by the structure and function of the brain called

1. Neural networks.
2. Deep neural networks.
3. Artificial neural networks.
4. Option 1 & 2

ANSWER : 3

40. _____ are typically feedforward networks in which data flows from the input layer to the output layer without looping back.

1. Convolutional Neural Networks
2. Deep Neural Network
3. Recurrent Neural Networks
4. Artificial Neural Networks

ANSWER : 2

41. _____ are the networks in which data can flow in any direction and are used for applications for language purposes.

1. Convolutional Neural Networks
2. Deep Neural Network
3. Recurrent Neural Networks
4. Artificial Neural Networks

ANSWER : 3

42. Common Deep learning applications include:

1. Real-time visual tracking
2. Logistic optimization
3. Bioinformatics, Speech recognition
4. Only 1 & 3
5. All of the mentioned

ANSWER : 5

43. The most widely used metrics and tools to assess a classification model are:

1. Confusion matrix
2. Cost-sensitive accuracy
3. Area under the ROC curve
4. All of the above

ANSWER :4

44. Which of the following is a good test dataset characteristic?

1. Large enough to yield meaningful results
2. Is representative of the dataset as a whole
3. Both A and B
4. None of the above

ANSWER : 3

45. How do you handle missing or corrupted data in a dataset?

1. Drop missing rows or columns
2. Replace missing values with mean/median/mode
3. Assign a unique category to missing values
4. All of the above

ANSWER : 4

46. Choose the options that are correct regarding machine learning (ML) and artificial intelligence (AI),

1. ML is an alternate way of programming intelligent machines.
2. ML and AI have very different goals.
3. ML is a set of techniques that turns a dataset into a software.
4. AI is a software that can emulate the human mind.

Answer: (1), (3), (4)

47. Unsupervised learning is where you only have input data (X) and

1. No corresponding output variables.
2. No corresponding model to define the set of variables
3. No corresponding training dataset.
4. None of the above.

ANSWER: 1

48. Self-organizing maps are the examples of

1. Supervised learning
2. Unsupervised learning
3. Reinforcement learning
4. Missing data imputation

ANSWER: 2

49. Task of inferring a model from the labeled training data is called as :

1. Supervised learning
2. Unsupervised learning
3. Reinforcement learning
4. Missing data imputation

ANSWER: 1

50. A classification problem is when the output variable

1. Is a real value, such as “dollars” or “weight”.
2. Discovers the inherent groupings in the data.
3. Is a category, such as “red” or “blue” or “disease” and “no disease”
4. Discovers rules that describe large portions of your data.

ANSWER: 3

Subject: Machine Learning (BE-A&B)

UNIT II- SYLLABUS : Feature Selection

Scikit- learn Dataset, Creating training and test sets, managing categorical data, Managing missing features, Data scaling and normalization.

Feature selection and Filtering, Principle Component Analysis(PCA)-non negative matrix factorization, Sparse PCA, Kernel PCA. Atom Extraction and Dictionary Learning.

1. Tick the odd one out

- A. Scikit-learn dataset provide some built-in datasets that can be used for testing purposes.
- B. Built-in datasets are available in the package sklearn.datasets and have a common structure
- C. scikit-learn comes with a few standard datasets, for instance the iris and digits datasets for regression and the boston house prices dataset for classification
- D. Data instance variable contains the whole input set X while target contains the labels for classification or target values for regression.

ANSWER: C

2. Arrange in correct sequence

- A. Loading the dataset.
- B. Installing the Python and SciPy platform.-Download Miniconda
- C. Summarizing the dataset.
- D. Evaluating some algorithms.
- E. Making some predictions
- F. Visualizing the dataset.

ANSWER: B-A-C-F-D-E

3. Requirements for working with data in scikit-learn

- A. Features and response are separate objects
- B. Features and response should be numeric
- C. Features and response should have specific shapes
- D. All of the options mentioned
- E. Only option 1to 3

ANSWER: D

4. Correct project name for project name for scikit learn dataset

- A. scikit-learn
- B. scikit_learn
- C. SciKit learn
- D. sci-kit learn.

ANSWER: A

5. **ndarray is [TICK THE CORRECT OPTION]**

- A. a fast and space-efficient multidimensional array providing vectorized arithmetic operation
- B. a generic multidimensional container for homogeneous data that is all of the elements must be the same type.
- C. Every array has a shape a tuple indicating the size of each dimension, and a dtype, an object describing the *data type* of the array

ANSWER: A,B,C

6. When a dataset is large enough, it's a good practice to split it into training and test sets by rules such as [tick the odd one out]
- A. Both datasets must reflect the original distribution
 - B. The original dataset must be randomly shuffled before the split phase in order to avoid a correlation between consequent elements
 - C. Percentage of elements to put into the test/training set can be of a ratio as 90 percent for training and 10 percent for the test phase.
 - D. All of the mentioned

ANSWER: C

7. To convert categorical features to such integer codes, we can use the
- E. OrdinalEncoder
 - F. LabelEncoder
 - G. Both A and B can be used alternatively
 - H. None of the above

ANSWER: A

8. LabelEncoder can be used to normalize labels
- A. TRUE
 - B. FALSE

ANSWER: A

9 Different ways of encoding categorical features are [tick the odd one out]

- A. OneHotEncoder
- B. DictVectorizer
- C. Pandas get_dummies
- D. OrdinalEncoder

ANSWER: D

10. One-Hot Encoding is

- A. Applied for categorical variables where no ordinal relationship exists
- B. Applied where the integer encoding is not enough.
- C. Used to add new binary variable for each unique integer value.
- D. None of the options mentioned
- E. All of the options mentioned

ANSWER: E

11. A dataset can contain missing features, with the options to be considered as:

- A. Removing the whole line- dataset is quite large, the number of missing features is high, and any prediction could be risky.
- B. Creating sub-model to predict those features-more difficult because it's necessary to determine a supervised strategy to train a model for each feature and, finally, to predict their value.
- C. Using an automatic strategy to input them according to the other known values-likely to be the best choice.
- D. Option 1 & 2 only
- E. All of the options mentioned

ANSWER: E

12. The command to create an environment in sklearn with a specific version of Python is
- \$ conda create -n myenv python=3.4
 - \$ conda create -n myenv scipy
 - \$ conda create -n myenv scipy=0.15.0
 - None of the mentioned

ANSWER: A

13. To create an environment in sklearn with a specific package:
- \$ conda create -n myenv python=3.4
 - \$ conda create -n myenv scipy
 - \$ conda create -n myenv scipy=0.15.0
 - None of the mentioned

ANSWER: B

14. To create an environment with a specific version of a package in sklearn
- \$ conda create -n myenv python=3.4
 - \$ conda create -n myenv scipy
 - \$ conda create -n myenv scipy=0.15.0
 - None of the mentioned

ANSWER: C

15. Command to load boston dataset in python is
- from sklearn.datasets import load_boston
 - from sklear_datasets import load_boston
 - from sklearn.datasets.import_load_boston
 - All of the above

ANSWER: A

- 16 .Which of the following are true about forward subset selection?

O($2d$) models must be trained during the algorithm, where d is the number of features

It greedily adds the feature that most improves cross-validation accuracy

It finds the subset of features that give the lowest test error

Forward selection is faster than backward selection if few features are relevant to prediction

17. Principal component analysis (PCA) can be used with variables of any mathematical types: quantitative, qualitative, or a mixture of these types. – True, False

18. Principal component analysis (PCA) requires quantitative multivariate data. – True, False.

19. For variables with physical dimensions (e.g. kg), their variances also have physical dimensions. – True, False

20. The variables subjected to PCA must all have the same physical dimensions. – True, False.

21. When the variables have different physical dimensions, they must be made dimensionless by standardization or ranging before PCA. – True, False.

22. Which of the following is the second goal of PCA?

- a) data compression
- b) statistical analysis
- c) data dredging
- d) all of the mentioned

Answer: a

23. T or F The goal of PCA is to interpret the underlying structure of the data in terms of the principal components that are best at predicting the output variable.

24 T or F The output of PCA is a new representation of the data that is always of lower dimensionality than the original feature representation.

25. In principal component analysis, a smaller eigenvalue indicates that

- A. A given variable in the original data set, say X_j , is more important
- B. A given variable in the original data set, say X_j , is less important
- C. A given principal component, say Y_j , is more important
- D. A given principal component, say Y_j , is less important

26. Why do we often pick just the first two principal components?

- A. Because we can graph them in a scatterplot
- B. Because they explain most of the variance
- C. Because they are uncorrelated
- D. Because of the Kaiser criterion

27. Taking a bootstrap sample of n data points in p dimensions means:

- A. Sampling p features with replacement.
- B. Sampling \sqrt{p} features without replacement.
- C. Sampling n samples with replacement.
- D. Sampling $k < n$ samples without replacement

28. PCA properties are

- A. Unsupervised dimensionality reduction
- B. Linear representation that gives best squared error fit
- C. No local minima (exact)
- D. Orthogonal vectors

29. NMF properties are

- A. Unsupervised dimensionality reduction
- B. Non-negative coefficients
- C. Iterative (the presented algorithm)
- D. “Parts-based”; easier to interpret

30. Kernel PCA applications including

- A. denoising,
- B. compression
- C. structured prediction

31. Which of the following is an example of a deterministic algorithm?

- A) PCA
- B) K-Means
- C) None of the above

32. [True or False] A Pearson correlation between two variables is zero but, still their values can still be related to each other.

- A) TRUE
- B) FALSE

33. What would you do in PCA to get the same projection as SVD?

- A) Transform data to zero mean
- B) Transform data to zero median
- C) Not possible
- D) None of these

34. Which of the following is an example of feature extraction?

- A all of the above
- B Constructing bag of words vector from an email
- C Applying PCA projects to a large high-dimensional data
- D Removing stopwords in a sentence

35. When performing regression or classification, which of the following is the correct way to preprocess the data?

- A none of the above
- B Normalize the data → PCA → normalize PCA output → training
- C Normalize the data → PCA → training
- D PCA → normalize PCA output → training

36. Which of the following techniques would perform better for reducing dimensions of a data set?

- A.none of these
- B Removing columns with dissimilar data trends
- C Removing columns which have high variance in data
- D Removing columns which have too many missing values

37. Dimensionality reduction algorithms are one of the possible ways to reduce the computation time required to build a model.

TRUE/False

38. Which of the following is/are one of the important step(s) to pre-process the text in NLP based projects?

- 1 STEMMING
- 2 STOP WORD REMOVAL
- 3 OBJECT STANDARDIZATION

- A: 1,2
- B1,2,3
- C 2,3
- D 1,3

39. Which of the following is an example of a deterministic algorithm?

- A PCA
- B none of the above
- C K means

40. What is pca.components_ in Sklearn?

- A None of the above
- B Result of the multiplication matrix
- C Set of all eigen vectors for the projection space
- D Matrix of principal components

Subject: Machine Learning (BE-A&B)

UNIT III- SYLLABUS :Regression

Linear regression- Linear models, A bi-dimensional example, Linear Regression and higher dimensionality, Ridge, Lasso and ElasticNet, Robust regression with random sample consensus, Polynomial regression, Isotonic regression.

Logistic regression-Linear classification, Logistic regression, Implementation and Optimizations, Stochastic gradient descendent algorithms, Finding the optimal hyper-parameters through grid search, Classification metric, ROC Curve.

[Home](#)

SECTION A: MCQS

1. Which one of the statement is true regarding residuals in regression analysis?
A Mean of residuals is always zero
B Mean of residuals is always less than zero
C Mean of residuals is always greater than zero
D There is no such rule for residuals.

ANSWER: A

2. The correlation coefficient is used to determine:
A A specific value of the y-variable given a specific value of the x-variable
B A specific value of the x-variable given a specific value of the y-variable
C The strength of the relationship between the x and y variables
D All of the above

ANSWER: C

3. To test linear relationship of y(dependent) and x(independent) continuous variables, which of the following plot best suited?
A Scatter plot
B Barchart
C Histograms
D All of the above

ANSWER: A

4. Which of the following method(s) does not have closed form solution for its coefficients?
A Ridge regression
B Lasso
C Both Ridge and Lasso
D None of both

ANSWER: B

5. Suppose we fit “Lasso Regression” to a data set, which has 100 features ($X_1, X_2 \dots X_{100}$). Now, we rescale one of these feature by multiplying with 10 (say that feature is X_1), and then refit Lasso regression with the same regularization parameter. Now, which of the following option will be correct?
A It is more likely for X_1 to be excluded from the model
B It is more likely for X_1 to be included in the model
C Can't say
D All of these

ANSWER: B

6. In a simple linear regression model (One independent variable), If we change the input variable by 1 unit. How much output variable will change?

- A By 1
- B No change
- C By its Slope
- D None of the above

ANSWER: C

7. It is possible to design a Linear regression algorithm using a neural network?

- A. TRUE
- B. FALSE

ANSWER: A

8. What will happen when you apply very large penalty in case of Lasso?

- A. Some of the coefficient will become zero
- B. Some of the coefficient will be approaching to zero but not absolute zero
- C. Both A and B depending on the situation
- D. None of these

ANSWER: A

9. Suppose you plotted a scatter plot between the residuals and predicted values in linear regression and you found that there is a relationship between them. Which of the following conclusion do you make about this situation?

- A. Since the there is a relationship means our model is not good
- B. Since the there is a relationship means our model is good
- C. Can't say
- D. None of these

ANSWER: A

10. If the slope of the regression equation $y = b_0 + b_1x$ is positive, then;

- A. as x increases y decreases
- B. as x increases so does y
- C. Either a or b is correct
- D. as x decreases y increases

ANSWER: B

11. Ridge regression uses _____ regularization which adds the penalty term to the OLS equation

- A. L2
- B. L1
- C. Uses L1 and L2 alternatively
- D. None of the mentioned

ANSWER: A

12. TICK THE ODD ONE OUT

- A. Linear regression is a linear model which assumes a linear relationship between the input variables (x) and the single output variable (y).
- B. In linear regression y can be calculated from a linear combination of the input variables (x).

- C. The regression analysis helps us to understand how much the dependent variable changes with a change in one or more independent variables.
- D. Linear regression is similar to Kernel PCA.

ANSWER: A

13. Match the following

1. Simple linear regression	a. 1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)
2. Multiple linear regression	b. 1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous)
3. Logistic regression	c. 1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)
4. Multinomial regression	d. 1 dependent variable (interval or ratio), 2+ independent variables (interval or ratio or dichotomous)

- A. 1-b, 2-a, 3-d, 4-c
- B. 1-d, 2-b, 3-a, 4-c
- C. 1-b, 2-d, 3-a, 4-c
- D. 1-c, 2-d, 3-a, 4-b

ANSWER: C

14. Let's say, a "Linear regression" model perfectly fits the training data (train error is zero). Which of the following statement is true?

- A. You will always have test error zero
- B. You cannot have test error zero
- C. None of the above

ANSWER: C

15. What would be then consequences for the OLS estimator if heteroscedasticity is present in a regression model but ignored?

- A. It will be biased
- B. It will be inconsistent
- C. It will be inefficient
- D. All of the options mentioned

ANSWER: C

16. TICK THE ODD ONE OUT

- A. Multiple linear regression requires at least two independent variables, which can be nominal, ordinal, or interval/ratio level variables.
- B. Multiple linear regression requires the relationship between the independent and dependent variables to be linear.
- C. Multiple linear regression analysis requires that the errors between observed and predicted values should be normally distributed.
- D. Multiple linear regression assumes that there is multicollinearity in the data

ANSWER: D

17. Multicollinearity may be checked multiple as

- A. Using correlation matrix
- B. Using Variance Inflation Factor (VIF)
- C. By centering the data.
- D. Option 1 and 2 only
- E. All of the options mentioned

Prepared by: Mrs. DeepaliGohil & Mrs. NilamPatil(Subject Teachers)

ANSWER: E

18. To check the accuracy of a regression, scikit-learn provides _____ function which evaluates the model on test data.

- A. lr.score(X_test, Y_test)
- B. lr.accute(X_test, Y_test)
- C. lr.accuracy(X_test, Y_test)
- D. lr.acc.reg(X_test, Y_test)

ANSWER: A

19. If the points on the scatter diagram indicate that as one variable increases the other variable tends to decrease the value of r will be:

- A. Perfect positive
- B. Perfect negative
- C. Negative
- D. Zero

ANSWER: C

20. General equation of a polynomial regression is

- A. $Y = \theta_0 + \theta_1 X + \theta_2 X^2 + \dots + \theta_m X^m + \text{transmission error}$
- B. $Y = \theta_0 + \theta_1 X + \theta_2 X^2 + \dots + \theta_m X^m + \text{residual error}$
- C. $Y = \theta_0 + \theta_1 X + \theta_2 X^2 + \dots + \theta_m X^m + \text{gross error}$
- D. None of the options mentioned

ANSWER: B

21. Ridge Regression performs L2 regularization which

- A. Adds penalty equivalent to the magnitude of coefficients
- B. Adds penalty equivalent to squareroot of the magnitude of coefficients
- C. Adds penalty equivalent to square of the magnitude of coefficients
- D. Adds penalty equivalent to cube root of the magnitude of coefficients

ANSWER: C

22. **ElasticNet**, which combines both Lasso and Ridge into a single model with two penalty factors, one proportional to $L1$ norm and the other to $L2$ norm.

- A. TRUE
- B. FALSE

ANSWER: A

23. L1 Regularization which is _____ adds regularization terms in the model which are function of _____ of the coefficients of parameters.

- A. Lasso Regularization, absolute value
- B. Ridge Regularization, square
- C. Lasso Regularization, square
- D. Ridge Regularization, absolute value

ANSWER: A

24. Steps followed in RANSAC algorithm

- a. Select a random subset of the original data. Call this subset the hypothetical inliers.
- b. All other data are then tested against the fitted model. Those points that fit the estimated model well, according to some model-specific loss function, are considered as part of the consensus set.
- c. A model is fitted to the set of hypothetical inliers.
- d. The estimated model is reasonably good if sufficiently many points have been classified as part of the consensus set.
- e. Afterwards, the model may be improved by reestimating it using all members of the consensus set.

ANSWER: a-c-b-d-e

25. TICK THE ODD ONE OUT

- A. The isotonic regression finds a non-decreasing approximation of a function while minimizing the mean squared error on the training data.
- B. The benefit of isotonic regression model is that it does not assume any form for the target function such as linearity.
- C. Isotonic regression equation becomes polynomial regression equation if the power of independent variable is more than 1.
- D. Isotonic regression produces a piecewise interpolating functions minimizing the functional.

ANSWER: C

1) True-False: Is Logistic regression a supervised machine learning algorithm?

- A) TRUE
B) FALSE

2) True-False: Is Logistic regression mainly used for Regression?

- A) TRUE
B) FALSE

3) True-False: Is it possible to design a logistic regression algorithm using a Neural Network Algorithm?

- A) TRUE
B) FALSE

4) True-False: Is it possible to apply a logistic regression algorithm on a 3-class Classification problem?

- A) TRUE
B) FALSE

5) [True-False] Standardisation of features is required before training a Logistic Regression.

- A) TRUE
B) FALSE

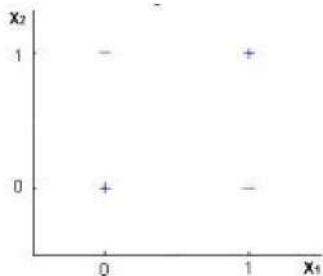
ANSWER: B

6) Choose which of the following options is true regarding One-Vs-All method in Logistic Regression.

- A) We need to fit n models in n-class classification problem
B) We need to fit n-1 models to classify into n classes
C) We need to fit only 1 model to classify into n classes
D) None of these

ANSWER: A

7) Can a Logistic Regression classifier do a perfect classification on the below data?



Note: You can use only X1 and X2 variables where X1 and X2 can take only two binary values(0,1).

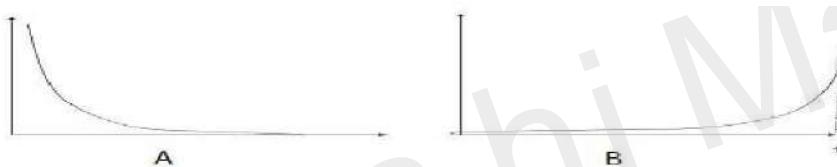
- A) TRUE
- B) FALSE
- C) Can't say
- D) None of these

ANSWER: B

8) Which of the following image is showing the cost function for $y = 1$.

Following is the loss function in logistic regression(Y-axis loss function and x axis log probability) for two class classification problem.

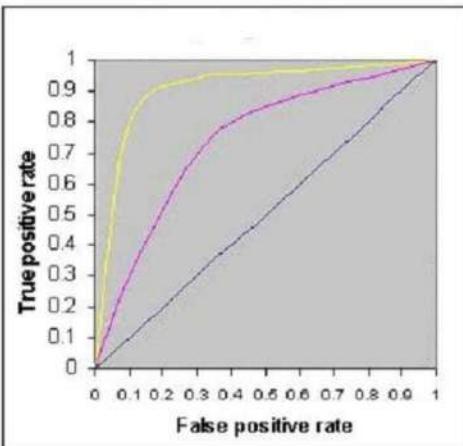
Note: Y is the target class



- A) A
- B) B
- C) Both
- D) None of these

ANSWER: A

9) The below figure shows AUC-ROC curves for three logistic regression models. Different colors show curves for different hyper parameters values. Which of the following AUC-ROC will give best result?



- A) Yellow
 B) Pink
 C) Black
 D) All are same

ANSWER: A

Q10. Generally, which of the following method(s) is used for predicting continuous dependent variable?

1. Linear Regression
 2. Logistic Regression
- A. 1 and 2
 B. only 1
 C. only 2
 D. None of these.

ANSWER: B

Q11. Suppose I applied a logistic regression model on data and got training accuracy X and testing accuracy Y. Now I want to add few new features in data. Select option(s) which are correct in such case.
 Note: Consider remaining parameters are same.

1. Training accuracy always decreases.
 2. Training accuracy always increases or remain same.
 3. Testing accuracy always decreases
 4. Testing accuracy always increases or remain same
- A. Only 2
 B. Only 1
 C. Only 3
 D. Only 4

ANSWER: A

Q12. Suppose, we are using Logistic regression model for n-class classification problem. In this case, we can use One-vs-rest method. Choose which of the following option is true regarding this?

- A. We need to fit n model in n-class classification problem.
 B. We need to fit n-1 models to classify into n classes.
 C. We need to fit only 1 model to classify into n classes.
 D. None of these.

ANSWER: A

13) Which of the following hyper parameter(s), when increased may cause random forest to over fit the data?

1. Number of Trees
2. Depth of Tree
3. Learning Rate

- A) Only 1
- B) Only 2
- C) Only 3
- D) 1 and 2
- E) 2 and 3
- F) 1,2 and 3

ANSWER: (B)

Usually, if we increase the depth of tree it will cause overfitting. Learning rate is not an hyperparameter in random forest. Increase in the number of tree will cause under fitting.

14) Imagine, you are solving a classification problems with highly imbalanced class. The majority class is observed 99% of times in the training data.

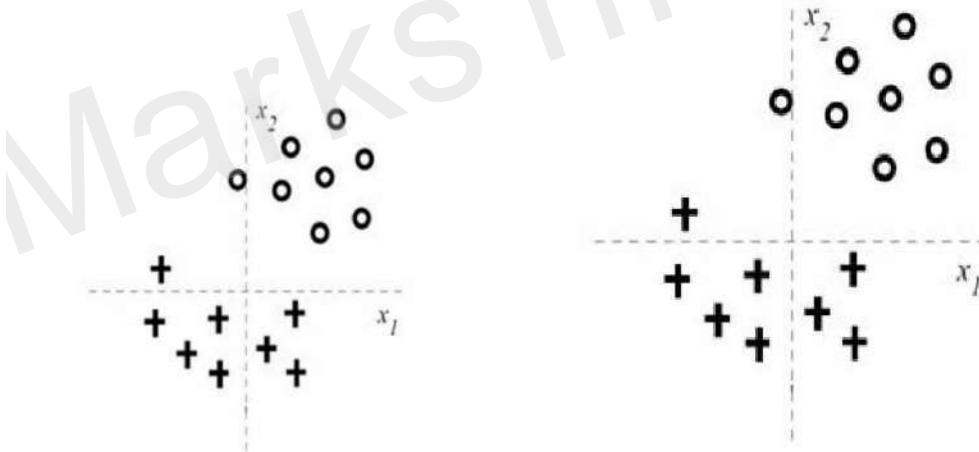
Your model has 99% accuracy after taking the predictions on test data. Which of the following is true in such a case?

1. Accuracy metric is not a good idea for imbalanced class problems.
2. Accuracy metric is a good idea for imbalanced class problems.
3. Precision and recall metrics are good for imbalanced class problems.
4. Precision and recall metrics aren't good for imbalanced class problems.

- A) 1 and 3
- B) 1 and 4
- C) 2 and 3
- D) 2 and 4

ANSWER: (A)

15. Suppose you are given the below data and you want to apply a logistic regression model for classifying it in two given classes.



You are using logistic regression with L1 regularization.

$$\sum_{i=1}^n \log P(y_i|x_i, w_0, w_1, w_2) - C(|w_1| + |w_2|).$$

Where C is the regularization parameter and w1 & w2 are the coefficients of x1 and x2.

Which of the following option is correct when you increase the value of C from zero to a very large value?
A) First w2 becomes zero and then w1 becomes zero

- B) First w1 becomes zero and then w2 becomes zero
C) Both becomes zero at the same time
D) Both cannot be zero even after very large value of C

ANSWER: (B)

16.If searching among a large number of hyperparameters, you should try values in a grid rather than random values, so that you can carry out the search more systematically and not rely on chance. True or False?
FALSE

17.Every hyperparameter, if set poorly, can have a huge negative impact on training, and so all hyperparameters are about equally important to tune well. True or False?
FALSE

18.During hyperparameter search, whether you try to babysit one model ("Panda" strategy) or train a lot of models in parallel ("Caviar") is largely determined by:

- A. Whether you use batch or mini-batch optimization
- B. The presence of local minima (and saddle points) in your neural network
- C. The amount of computational power you can access
- D. The number of hyperparameters you have to tune

19.Finding good hyperparameter values is very time-consuming. So typically you should do it once at the start of the project, and try to find very good hyperparameters so that you don't ever have to revisit tuning them again. True or false?

FALSE

20.In the normalization formula, why do we use epsilon?
To avoid division by zero. True or false?

TRUE

SECTION B:Short question and answer

1. What is linear regression?

In simple terms, linear regression is a method of finding the best straight line fitting to the given data, i.e. finding the best linear relationship between the independent and dependent variables. In technical terms, linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method.

2. State the assumptions in a linear regression model.

There are three main assumptions in a linear regression model:

1. The assumption about the form of the model:

It is assumed that there is a linear relationship between the dependent and independent variables.
It is known as the 'linearity assumption'.

2. Assumptions about the residuals:

Prepared by: Mrs. DeepaliGohil& Mrs. NilamPatil(Subject Teachers)

Subject: Machine Learning (BE-A&B)

UNIT IV- SYLLABUS: Naïve Bayes and Support Vector Machine

Bayes" Theorum, Naïve Bayes" Classifiers, Naïve Bayes in Scikit- learn- Bernoulli Naïve Bayes, Multinomial Naïve Bayes, and Gaussian Naïve Bayes.

Support Vector Machine (SVM)- Linear Support Vector Machines, Scikit- learn implementation- Linear Classification, Kernel based classification, Non- linear Examples. Controlled Support Vector Machines, Support Vector Regression.

SECTION A: MCQS

1. Naive Bayes classifier
 - A. Assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature
 - B. Is easy to build and particularly useful for very large data sets.
 - C. Is easy to build and particularly useful for small data sets.
 - D. Only option 1 & 2
 - E. Only option 1 & 3

ANSWER: D

2. Match the following:

1. Bernoulli	a. Can be used for both classification or regression challenges
2. Multinomial	b. Binary distribution, useful when a feature can be present or absent
3. Gaussian.	c. Discrete distribution and is used whenever a feature must be represented by a whole number
4. SVM	d. Continuous distribution characterized by its mean and variance.

ANSWER: 1-b,2-c,3-d,4-a

3. Kernels makes _____ work in _____ by mapping data to _____ where it exhibits _____ patterns.

- A. Nonlinear models, Linear settings, Nonzero dimensions , Nonlinear patterns
- B. Linear models, Nonlinear settings, Higher dimensions , Linear patterns
- C. None of the mentioned
- D. Both of the mentioned

ANSWER: B

4. Tick the odd one

- A. from sklearn.naive_bayes import BernoulliNB
- B. from sklearn.feature_extraction import DictVectorizer
- C. from sklearn.naive_bayes import GaussianNB
- D. Only B
- E. None of the mentioned

ANSWER: E

5. Which of the following are applications of the SVM?

- A. Text and Hypertext Categorization
- B. Image Classification
- C. Clustering of News Articles
- D. All of the above

ANSWER: D

6. The minimum time complexity for training an SVM is $O(n^2)$. According to this fact, what sizes of datasets are not best suited for SVM's?

- A. Large datasets
- B. Small datasets
- C. Medium sized datasets
- D. Size does not matter

ANSWER: A

7. Steps in Naïve Bayes algorithm are:

- a. Use theorem equation to calculate the posterior probability for each class.
- b. Convert the data set into a frequency table.
- c. Create Likelihood table by finding the probabilities.

Correct sequence of these steps is:

- A. a,b,c
- B. b,c,a**
- C. a,c,b
- D. b,a,c

ANSWER: B

8. In Naïve Bayes formula , $P(c|x)$ and $P(x|c)$ represents,

- A. $P(c|x)$ is the likelihood which is the probability of predictor given class and $P(x|c)$ is the posterior probability of class
- B. $P(c|x)$ is the posterior probability of class and $P(x|c)$ is the likelihood which is the probability of predictor given class.
- C. None of the above.
- D. Both of the mentioned

ANSWER: B

9. Performance of the SVM depends upon no of training instances, Linear vs. non linear problems and

- A. Input scale of features
- B. The chosen hyperparameter
- C. How you evaluate the model
- D. All of the mentioned

ANSWER: D

10. SVM can be widely used in robotics and in computer vision for classifying objects and sensor data .

- A. TRUE
- B. FALSE

ANSWER: A

11. At a certain university, 4% of men are over 6 feet tall and 1% of women are over 6 feet tall. The total student population is divided in the ratio 3:2 in favour of women. If a student is selected at random from among all those over six feet tall, what is the probability that the student is a woman?

- A. 2/5
- B. 3/5
- C. 3/11
- D. 1/100

ANSWER: C

12. Support vector machines, like logistic regression models, give a probability distribution over the possible labels given an input example.

- A. TRUE
- B. FALSE

ANSWER: B

13. Tick the odd one out with respect to advantages of Naïve Bayes theorem

- A. Can successfully train on large data set
- B. Good for text classification, good for multiclass classification
- C. Quick and simple calculation since it is naïve
- D. All of the mentioned

ANSWER: A

14. Naive Bayes can be used in

- A. Face Recognition, as a classifier to identify the faces or its other features, like nose, mouth, eyes etc
- B. Weather Prediction to predict if the weather will be good or bad.
- C. Medical Diagnosis to indicate if a patient is at high risk for certain diseases and conditions, such as heart disease, cancer, and other ailments.
- D. News Classification to predict whether the news is political, world news, and so on.
- E. All of the options mentioned
- F. Only option A,B,C

ANSWER: E

15. In Gaussian Naive Bayes, continuous values associated with each feature are assumed to be distributed according to a

- A. Gaussian distribution.
- B. Normal distribution.
- C. Both A & B
- D. None of the options mentioned

ANSWER: C

16. The previous probabilities in Bayes theorem that are changed with the help of new available information are classified as

- A. Independent probabilities
- B. Posterior probabilities
- C. Interior probabilities
- D. Independent probabilities

ANSWER: B

17. The formula for Bayes' theorem is

- A. $P(H|E) = P(E|H) * P(H) / P(E)$
- B. $P(H|E) = P(H) / P(E) * P(E|H)$
- C. $P(H|E) = P(E) / P(E|H) * P(H)$
- D. All the options mentioned are correct

ANSWER: A

18. Which of the following loss functions are not convex? (Multiple options may be correct)

- A. loss (sometimes referred as mis-classification loss)
- B. Hinge loss
- C. Logistic loss
- D. Squared error loss

ANSWER: A

19. Which of the following properties is false in the case of a Bayesian Network?

- A. The edges are directed
- B. Contains cycles
- C. Represents conditional independence relations among random variables
- D. All of the above

ANSWER: B

20. Which of the following is/are not true regarding an SVM?

- A. For two dimensional data points, the separating hyperplane learnt by a linear SVM will be a straight line.
- B. In theory, a Gaussian kernel SVM can model any complex separating hyperplane.
- C. For every kernel function used in a SVM, one can obtain a equivalent closed form basis expansion.
- D. Over fitting in an SVM is a function of number of support vectors.

ANSWER: C

21. Select the correct statement Support Vector Machine

- A. SVM is a supervised machine learning algorithm which can be used for both classification and regression challenges.
- B. SVM is a supervised machine learning algorithm which can be used for both classification and clustering challenges.
- C. Both of the options mentioned
- D. None of the options mentioned

ANSWER: A

22. Hyperplane is

- A. Hyperplanes are decision boundaries that help classify the data points.
- B. A hyperplane in an n-dimensional Euclidean space is a flat, n-1 dimensional subset of that space that divides the space into two disconnected parts.
- C. The most common example of hyperplanes in practice is with support vector machines
- D. Option A & C only
- E. All of the options mentioned

ANSWER: E

23. The Naive Bayes algorithm is called “naive” because it makes the assumption that the occurrence of a certain feature is independent of the occurrence of other features.

- A. TRUE
- B. FALSE

ANSWER: A

24. The components of the above statement are [TICK THE ODD ONE OUT]

- A. $P(A|B)$: Probability (conditional probability) of occurrence of event A given the event B is true
- B. $P(A)$ and $P(B)$: Probabilities of the occurrence of event A and B respectively
- C. $P(B|A)$: Probability of the occurrence of event B given the event A is false

ANSWER: C

25. Steps followed in Naïve Bayes theorem are (arrange the correct sequence)

- a. Draw the likelihood table for the features against the classes
- b. Create a frequency table for all the features against the different classes.
- c. Calculate $\max_i P(C_i|x_1, x_2, \dots, x_n)$
- d. Calculate the conditional probabilities for all the classes,

ANSWER: b-a-d-c

1. Support vector machine (SVM) is a _____ classifier?

- A. Discriminative
- B. Generative

ANSWER: A

2. SVM can be used to solve _____ problems.

- A. Classification
- B. Regression
- C. Clustering
- D. Both Classification and Regression

ANSWER: D

3. SVM is a _____ learning algorithm

- A. Supervised
- B. Unsupervised

ANSWER: A

4. SVM is termed as _____ classifier

- A. Minimum margin
- B. Maximum margin

ANSWER: B

5. The training examples closest to the separating hyperplane are called as _____

- A. Training vectors
- B. Test vectors
- C. Support vectors

ANSWER: C

6. Which of the following is a type of SVM?

- A. Maximum margin classifier
- B. Soft margin classifier
- C. Support vector regression
- D. All of the above

ANSWER: D

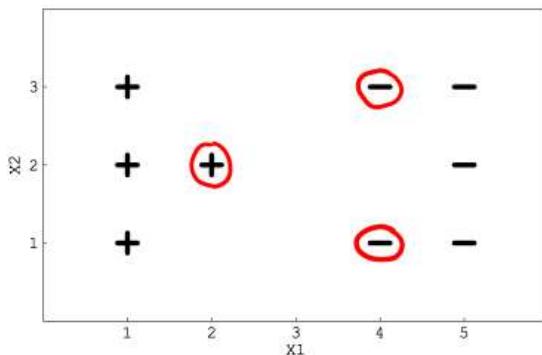
7. The goal of the SVM is to _____

- A. Find the optimal separating hyperplane which minimizes the margin of training data
- B. Find the optimal separating hyperplane which maximizes the margin of training data

ANSWER: B

Context for 8-9

Suppose you are using a Linear SVM classifier with 2 class classification problem. Now you have been given the following data in which some points are circled red that are representing support vectors.



8 If you remove the following any one red points from the data. Does the decision boundary will change?

A) Yes

B) No

ANSWER: A

9 [True or False] If you remove the non-red circled points from the data, the decision boundary will change?

A) True

B) False

ANSWER: B

10 What do you mean by generalization error in terms of the SVM?

A) How far the hyperplane is from the support vectors

B) How accurately the SVM can predict outcomes for unseen data

C) The threshold amount of error in an SVM

ANSWER: B

11 The cost parameter in the SVM means:

A) The number of cross-validations to be made

B) The kernel to be used

C) The tradeoff between misclassification and simplicity of the model

D) None of the above

ANSWER: C

12. Suppose you are building a SVM model on data X. The data X can be error prone which means that you should not trust any specific data point too much. Now think that you want to build a SVM model which has quadratic kernel function of polynomial degree 2 that uses Slack variable C as one of its hyper parameter. Based upon that give the answer for following question.

What would happen when you use very large value of C(C->infinity)?

Note: For small C was also classifying all data points correctly

A) We can still classify data correctly for given setting of hyper parameter C

B) We can not classify data correctly for given setting of hyper parameter C

C) Can't Say

D) None of these

ANSWER: A

13 What would happen when you use very small C (C~0)?

A) Misclassification would happen

B) Data will be correctly classified

C) Can't say

D) None of these

ANSWER: A

14. If I am using all features of my dataset and I achieve 100% accuracy on my training set, but ~70% on validation set, what should I look out for?

A) Underfitting

B) Nothing, the model is perfect

C) Overfitting

ANSWER: C

15 Which of the following are real world applications of the SVM?

A) Text and Hypertext Categorization

B) Image Classification

C) Clustering of News Articles

D) All of the above

ANSWER: D

Question Context: 16 – 18

Suppose you have trained an SVM with linear decision boundary after training SVM, you correctly infer that your SVM model is under fitting.

16) Which of the following option would you more likely to consider iterating SVM next time?

- A) You want to increase your data points
- B) You want to decrease your data points
- C) You will try to calculate more variables
- D) You will try to reduce the features

ANSWER: C

17) Suppose you gave the correct answer in previous question. What do you think that is actually happening?

- 1. We are lowering the bias
- 2. We are lowering the variance
- 3. We are increasing the bias
- 4. We are increasing the variance

- A) 1 and 2
- B) 2 and 3
- C) 1 and 4
- D) 2 and 4

ANSWER: C

18 In above question suppose you want to change one of it's(SVM) hyperparameter so that effect would be same as previous questions i.e model will not under fit?

- A) We will increase the parameter C
- B) We will decrease the parameter C
- C) Changing in C don't effect
- D) None of these

ANSWER: A

19. We usually use feature normalization before using the Gaussian kernel in SVM. What is true about feature normalization?

- 1. We do feature normalization so that new feature will dominate other
- 2. Some times, feature normalization is not feasible in case of categorical variables
- 3. Feature normalization always helps when we use Gaussian kernel in SVM

- A) 1
- B) 1 and 2
- C) 1 and 3
- D) 2 and 3

ANSWER: B

Question Context: 20-22

Suppose you are dealing with 4 class classification problem and you want to train a SVM model on the data for that you are using One-vs-all method. Now answer the below questions?

20) How many times we need to train our SVM model in such case?

- A) 1
- B) 2
- C) 3
- D) 4

ANSWER: D

21 Suppose you have same distribution of classes in the data. Now, say for training 1 time in one vs all setting the SVM is taking 10 second. How many seconds would it require to train one-vs-all method end to end?

- A) 20
- B) 40

- C) 60
- D) 80

ANSWER: B

22. Suppose your problem has changed now. Now, data has only 2 classes. What would you think how many times we need to train SVM in such case?

- A) 1
- B) 2
- C) 3
- D) 4

ANSWER: A

Question context: 23 – 24

Suppose you are using SVM with linear kernel of polynomial degree 2, Now think that you have applied this on data and found that it perfectly fit the data that means, Training and testing accuracy is 100%.

23 Now, think that you increase the complexity(or degree of polynomial of this kernel). What would you think will happen?

- A) Increasing the complexity will overfit the data
- B) Increasing the complexity will underfit the data
- C) Nothing will happen since your model was already 100% accurate
- D) None of these

ANSWER: A

24 In the previous question after increasing the complexity you found that training accuracy was still 100%. According to you what is the reason behind that?

1. Since data is fixed and we are fitting more polynomial term or parameters so the algorithm starts memorizing everything in the data

2. Since data is fixed and SVM doesn't need to search in big hypothesis space

- A) 1
- B) 2
- C) 1 and 2
- D) None of these

ANSWER: C

25) What is/are true about kernel in SVM?

1. Kernel function map low dimensional data to high dimensional space

2. It's a similarity function

- A) 1
- B) 2
- C) 1 and 2
- D) None of these

ANSWER: C

UNIT IV- SYLLABUS: Decision Trees and Ensemble Learning

Decision Trees- Impurity measures, Feature Importance. Decision Tree Classification with Scikitlearn, Ensemble Learning-Random Forest, AdaBoost, Gradient Tree Boosting, Voting Classifier. Introduction to Meta Classifier: Concepts of Weak and eager learner, Ensemble methods, Bagging, Boosting, Random Forests. Clustering Fundamentals- Basics, K-means: Finding optimal number of clusters, DBSCAN, Spectral Clustering. Evaluation methods based on Ground Truth- Homogeneity, Completeness, Adjusted Rand Index.

SECTION A: MCQS

1. Given that we can select the same feature multiple times during the recursive partitioning of the input space, is it always possible to achieve 100% accuracy on the training data (given that we allow for trees to grow to their maximum size) when building decision trees?
 - A. Yes
 - B. No

ANSWER: B

2. Which of the following statements are true with respect to the application of Cost-Complexity Pruning and Reduced Error Pruning with Cross-Validation?
 - A. In Reduced Error Pruning, the pruned tree error can never be less than the original tree
 - B. on the training dataset.
 - C. In Cost Complexity Pruning, the pruned tree error can never be less than the original tree
 - D. on the validation dataset.
 - E. In Reduced Error Pruning, the pruned tree error can never be less than the original tree
 - F. on the validation dataset.
 - G. both (b) and (c)

ANSWER: A

3. Suppose on performing reduced error pruning, we collapsed a node and observed an improvement in the prediction accuracy on the validation set. Which among the following statements re possible in light of the performance improvement observed? (multiple options may be correct)
 - A. The collapsed node helped overcome the effect of one or more noise affected data points
 - B. in the training set
 - C. The validation set had one or more noise affected data points in the region corresponding
 - D. to the collapsed node
 - E. The validation set did not have any data points along at least one of the collapsed branches
 - F. The validation set did have data points adversely affected by the collapsed node

ANSWER: A

4. Which of these classifiers do not require any additional modifications to their original descriptions (as seen in the lectures) to use them when we have more than 2 classes? (multiple options may be correct)
 - A. decision trees
 - B. logistic regression
 - C. support vector machines
 - D. k nearest neighbors

ANSWER: A & D

5. In a random forest model let $m << p$ be the number of randomly selected features that are used to identify the best split at any node of a tree. Which of the following are true? (p is the original number of features) (Multiple options may be correct)
- A. increasing m reduces the correlation between any two trees in the forest
 - B. decreasing m reduces the correlation between any two trees in the forest
 - C. increasing m increases the performance of individual trees in the forest
 - D. decreasing m increases the performance of individual trees in the forest

ANSWER: B & C

6. In AdaBoost, we re-weight points giving points misclassified in previous iterations more weight. Suppose we introduced a limit or cap on the weight that any point can take (for example, say we introduce a restriction that prevents any point's weight from exceeding a value of 10). Which among the following would be an effect of such a modification? (Multiple options may be correct)
- A. We may observe the performance of the classifier reduce as the number of stages increase
 - B. It makes the final classifier robust to outliers
 - C. It may result in lower overall performance

ANSWER: B & C

7. Which of the following method(s) is not inherently sequential?
- A. Gradient Boosting
 - B. Committee machines
 - C. AdaBoost

ANSWER: B

8. Boosting techniques typically give very high accuracy classifiers by sequentially training a collection of similar low-accuracy classifiers. Which of the following statements are true with respect to Boosting? (multiple options may be correct)
- A. LogitBoost (like AdaBoost, but with Logistic Loss instead of Exponential Loss) is less susceptible to overfitting than AdaBoost.
 - C. Boosting techniques tend to have low bias and high variance
 - D. Boosting techniques tend to have low variance and high bias
 - E. For basic linear regression classifiers, there is no effect of using Gradient Boosting.

ANSWER: A,B ,D

9. In a tournament classifier with N classes. What is the complexity of the number of classifiers we require?
- A. $O(N)$
 - B. $O(N^2)$
 - C. $O(N \cdot \log(N))$
 - D. $O(\log(N))$

ANSWER: A

10. Which of the following statements are true about ensemble classifiers? (multiple options may be correct)
- A. The different learners in boosting based ensembles can be trained in parallel
 - B. The different learners in bagging based ensembles can be trained in parallel
 - C. Boosting based algorithms which iteratively re-weight training points, such as AdaBoost, are more sensitive to noise than bagging based methods.
 - E. Boosting methods generally use strong learners as individual classifiers
 - F. Boosting methods generally use weak learners as individual classifiers.
 - G. An individual classifier in a boosting based ensemble is trained with every point in the training set.

ANSWER: B.C.E.F

11. In which approach do the classification models train on data sets whose distribution are modified in comparison to the distribution of the original training data set
- A. bagging
 - B. boosting
 - C. both
 - D. neither

ANSWER: C

(Explanation: In bagging, each classifier is trained on a data set generated by sampling from the original training data set with replacement, resulting in stochastic modification of the original data distribution. Similarly, in boosting, each classifier is trained on a data set generated by modifying the weights of the data instances based on the performance of the previous classifier, resulting again in modification of the original data distribution.)

12. Which of the following is/are false about bagging?
- A. Bagging reduces variance of the classifier
 - B. Bagging increases the variance of the classifier
 - C. Bagging can help make robust classifiers from unstable classifiers
 - D. Bagging results in increased bias

ANSWER : B.D

(Explanation : In bagging we combine the outputs of multiple classifiers trained on different samples of the training data. This helps in reducing overall variance. Due to the reduction in variance, normally unstable classifiers can be made robust with the help of bagging.)

13. Considering the AdaBoost algorithm, which among the following statements is false?
- A. In each stage, we try to train a classifier which makes accurate predictions on any subset
 - B. of the data points where the subset size is at least half the size of the data set
 - C. In each stage, we try to train a classifier which makes accurate predictions on a subset of
 - D. the data points where the subset contains more of the data points which were misclassified
 - E. in earlier stages
 - F. The weight assigned to an individual classifier depends upon the number of data points
 - G. correctly classified by the classifier
 - H. The weight assigned to an individual classifier depends upon the weighted sum error of
 - I. misclassified points for that classifier

ANSWER: A,C

14. Which of the following measure best analyze the performance of a classifier?
- A. Precision
 - B. Recall
 - C. Accuracy
 - D. Time complexity
 - E. Depends on the application

ANSWER: E

(Explanation Different applications might need to optimize different performance measures. Applications of machine learning span over playing games to very critical domains(such as health and security). Measures like accuracy for instance cannot be reliable when we have a dataset with significant class imbalance. So there cannot be a single measure to analyze the effectiveness of a classifier in all environments.)

15. Which of the following is required by K-means clustering?
- A. defined distance metric
 - B. number of clusters
 - C. initial guess as to cluster centroids
 - D. all of the mentioned

ANSWER: D

16. Point out the wrong statement.

- A. k-means clustering is a method of vector quantization
- B. k-means clustering aims to partition n observations into k clusters
- C. k-nearest neighbor is same as k-means
- D. none of the mentioned

ANSWER: C

17. K-means is not deterministic and it also consists of number of iterations.

- A. True
- B. False

ANSWER: A

18. Which points are eliminated by the DBSCAN algorithm?

- A. Core points
- B. Border points
- C. Noise points

ANSWER: C

19. How is the density of point p at the density based clustering defined?

- A. MinPts minus number of data points in an epsilon-neighbourhood
- B. Number of data points in an epsilon-neighbourhood of p
- C. Reciprocal value of the distance from p to the nearest neighbour

ANSWER: B

20 Which of the following tasks can be best solved using Clustering.

- A. Predicting the amount of rainfall based on various cues
- B. Detecting fraudulent credit card transactions
- C. Training a robot to solve a maze

ANSWER: B

21. Which of the following properties are characteristic of decision trees?

- A. High bias
- B. High variance
- C. Lack of smoothness of prediction surfaces
- D. Unbounded parameter set

ANSWER : B.C.D

22. Which among the following prevents overfitting when we perform bagging?

- A. The use of sampling with replacement as the sampling technique
- B. The use of weak classifiers
- C. The use of classification algorithms which are not prone to overfitting
- D. The practice of validation performed on every classifier trained

ANSWER: B

23. Consider an alternative way of learning a Random Forest where instead of randomly sampling the attributes at each node, we sample a subset of attributes for each tree and build the tree on these features. Would you prefer this method over the original or not, and why?

- A. Yes, because it reduces the correlation between the resultant trees
- B. Yes, because it reduces the time taken to build the trees due to the decrease in the attributes considered
- C. No, because many of the trees will be bad classifiers due to the absence of critical features considered in the construction of some of the trees

ANSWER: C

24. In case of limited training data, which technique, bagging or stacking, would be preferred, and why?

- A. Bagging, because we can combine as many classifier as we want by training each on a different sample of the training data
- B. Bagging, because we use the same classification algorithms on all samples of the training data
- C. Stacking, because each classifier is trained on all of the available data
- D. Stacking, because we can use different classification algorithms on the training data

ANSWER: C

25. Is AdaBoost sensitive to outliers?

- A. Yes
- B. No

ANSWER: A

26. Which of the following is/are true about bagging trees?

- 1. In bagging trees, individual trees are independent of each other
- 2. Bagging is the method for improving the performance by aggregating the results of weak learners
 - A) 1
 - B) 2
 - C) 1 and 2
 - D) None of these

Answer: C

27. Which of the following is/are true about Random Forest and Gradient Boosting ensemble methods?

- 1. Both methods can be used for classification task
- 2. Random Forest is use for classification whereas Gradient Boosting is use for regression task
- 3. Random Forest is use for regression whereas Gradient Boosting is use for Classification task
- 4. Both methods can be used for regression task
 - A) 1
 - B) 2
 - C) 3
 - D) 4
 - E) 1 and 4

Answer: E

28. In Random forest you can generate hundreds of trees (say T₁, T₂T_n) and then aggregate the results of these tree. Which of the following is true about individual(T_k) tree in Random Forest?

1. Individual tree is built on a subset of the features
 2. Individual tree is built on all the features
 3. Individual tree is built on a subset of observations
 4. Individual tree is built on full set of observations
- A) 1 and 3
B) 1 and 4
C) 2 and 3
D) 2 and 4

Answer: A

29. Which of the following is true about “max_depth” hyperparameter in Gradient Boosting?

1. Lower is better parameter in case of same validation accuracy
 2. Higher is better parameter in case of same validation accuracy
 3. Increase the value of max_depth may overfit the data
 4. Increase the value of max_depth may underfit the data
- A) 1 and 3
B) 1 and 4
C) 2 and 3
D) 2 and 4

Answer: A

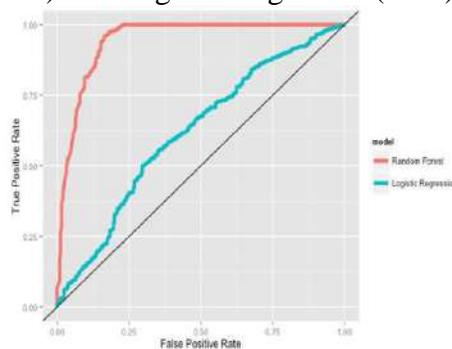
30. Which of the following algorithm doesn't uses learning Rate as of one of its hyperparameter?

1. Gradient Boosting
 2. Extra Trees
 3. AdaBoost
 4. Random Forest
- A) 1 and 3
B) 1 and 4
C) 2 and 3
D) 2 and 4

Answer: D

31. Which of the following algorithm would you take into the consideration in your final model building on the basis of performance?

Suppose you have given the following graph which shows the ROC curve for two different classification algorithms such as Random Forest(Red) and Logistic Regression(Blue)



- A) Random Forest
- B) Logistic Regression
- C) Both of the above
- D) None of these

Answer: A

32. Which of the following is true about training and testing error in such case?

Suppose you want to apply AdaBoost algorithm on Data D which has T observations. You set half the data for training and half for testing initially. Now you want to increase the number of data points for training $T_1, T_2 \dots T_n$ where $T_1 < T_2 \dots T_{n-1} < T_n$.

- A) The difference between training error and test error increases as number of observations increases
- B) The difference between training error and test error decreases as number of observations increases
- C) The difference between training error and test error will not change
- D) None of These

Answer: B

33. In random forest or gradient boosting algorithms, features can be of any type. For example, it can be a continuous feature or a categorical feature. Which of the following option is true when you consider these types of features?

- A) Only Random forest algorithm handles real valued attributes by discretizing them
- B) Only Gradient boosting algorithm handles real valued attributes by discretizing them
- C) Both algorithms can handle real valued attributes by discretizing them
- D) None of these

Answer: C

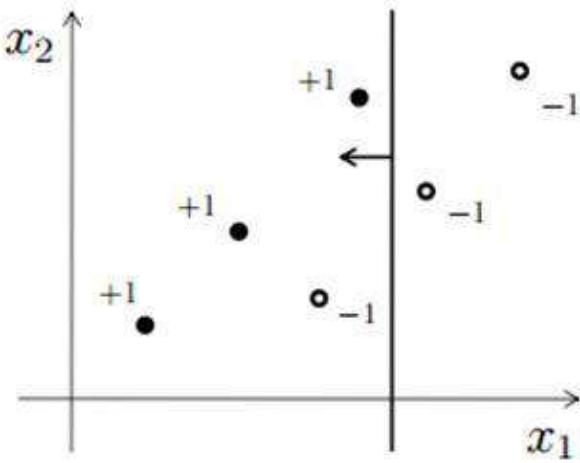
34. Suppose you are using a bagging based algorithm say a RandomForest in model building. Which of the following can be true?

1. Number of tree should be as large as possible
 2. You will have interpretability after using RandomForest
- A) 1
 - B) 2
 - C) 1 and 2
 - D) None of these

Answer: A

Context 11-14

Consider the following figure for answering the next few questions. In the figure, X_1 and X_2 are the two features and the data point is represented by dots (-1 is negative class and +1 is a positive class). And you first split the data based on feature X_1 (say splitting point is x_{11}) which is shown in the figure using vertical line. Every value less than x_{11} will be predicted as positive class and greater than x will be predicted as negative class.



35. How many data points are misclassified in above image?

- A) 1
- B) 2
- C) 3
- D) 4

Answer: A

36. Which of the following splitting point on feature x_1 will classify the data correctly?

- A) Greater than x_{11}
- B) Less than x_{11}
- C) Equal to x_{11}
- D) None of above

Answer: D

37. If you consider only feature X_2 for splitting. Can you now perfectly separate the positive class from negative class for any one split on X_2 ?

- A) Yes
- B) No

Answer: B

38. Now consider only one splitting on both (one on X_1 and one on X_2) feature. You can split both features at any point. Would you be able to classify all data points correctly?

- A) TRUE
- B) FALSE

Answer: B

Context 15-16

Suppose, you are working on a binary classification problem with 3 input features. And you chose to apply a bagging algorithm(X) on this data. You chose max_features = 2 and the n_estimators =3. Now, Think that each estimators have 70% accuracy.

Note: Algorithm X is aggregating the results of individual estimators based on maximum voting
39. What will be the maximum accuracy you can get?

- A) 70%
- B) 80%
- C) 90%
- D) 100%

Answer: D

Refer below table for models M1, M2 and M3.

Actual predictions	M1	M2	M3	Output
1	1	0	1	1
1	1	0	1	1
1	1	0	1	1
1	0	1	1	1
1	0	1	1	1
1	0	1	1	1
1	1	1	1	1
1	1	1	0	1
1	1	1	0	1
1	1	1	0	1

40. What will be the minimum accuracy you can get?

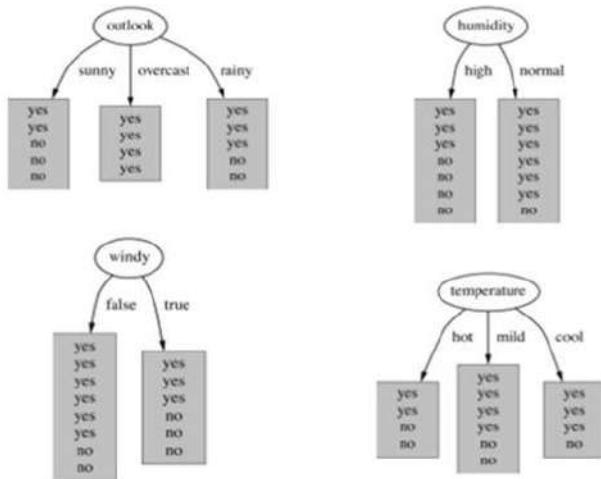
- A) Always greater than 70%
- B) Always greater than and equal to 70%
- C) It can be less than 70%
- D) None of these

Answer: C

Refer below table for models M1, M2 and M3.

Actual predictions	M1	M2	M3	Output
1	1	0	0	0
1	1	1	1	1
1	1	0	0	0
1	0	1	0	0
1	0	1	1	1
1	0	0	1	0
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1

41. Suppose you are building random forest model, which split a node on the attribute, that has highest information gain. In the below image, select the attribute which has the highest information gain?



- A) Outlook
- B) Humidity
- C) Windy
- D) Temperature

Answer: A

42. Which of the following is true about the Gradient Boosting trees?

1. In each stage, introduce a new regression tree to compensate the shortcomings of existing model
 2. We can use gradient decent method for minimize the loss function
- A) 1
 - B) 2
 - C) 1 and 2
 - D) None of these

Answer: C

43. Which of the following is true when you choose fraction of observations for building the base learners in tree based algorithm?

- A) Decrease the fraction of samples to build a base learners will result in decrease in variance
- B) Decrease the fraction of samples to build a base learners will result in increase in variance
- C) Increase the fraction of samples to build a base learners will result in decrease in variance
- D) Increase the fraction of samples to build a base learners will result in Increase in variance

Answer: A

44. In gradient boosting it is important use learning rate to get optimum output. Which of the following is true about choosing the learning rate?

- A) Learning rate should be as high as possible
- B) Learning Rate should be as low as possible
- C) Learning Rate should be low but it should not be very low
- D) Learning rate should be high but it should not be very high

Answer: C

45. [True or False] Cross validation can be used to select the number of iterations in boosting; this procedure may help reduce overfitting.

- A) TRUE
- B) FALSE

Answer: A

46. Which one of these is not a tree based learner?

- A. CART
- B. ID3
- C. Bayesian classifier
- D. Random Forest

ANSWER: C

47. Which one of these is a tree based learner?

- A. Rule based
- B. Bayesian Belief Network
- C. Bayesian classifier
- D. Random Forest

ANSWER: d

48. What is the approach of basic algorithm for decision tree induction?

- A. Greedy
- B. Top Down
- C. Procedural
- D. Step by Step

ANSWER : A

49. In the root node how many classes will be there?

- A. 3
- B. 2
- C. 4
- D. 14

ANSWER : B

50. Which among the following is/are some of the assumptions made by the k-means algorithm (assuming Euclidean distance measure)?

- A. Clusters are spherical in shape
- B. Clusters are of similar sizes
- C. Data points in one cluster are well separated from data points of other clusters
- D. There is no wide variation in density among the data points

ANSWER : A & B

Subject: Machine Learning (BE-A&B)

UNIT IV- SYLLABUS: Clustering Techniques

Hierarchical Clustering, Expectation maximization clustering, Agglomerative Clustering- Dendograms, Agglomerative clustering in Scikit- learn, Connectivity Constraints.

Introduction to Recommendation Systems- Naïve User based systems, Content based Systems, Model free collaborative filtering-singular value decomposition, alternating least squares. Fundamentals of Deep Networks-Defining Deep learning, common architectural principles of deep networks, building blocks of deep networks.

SECTION A: MCQS

1. In hierarchical clustering, is it possible for a point to be closer to points in other clusters than to points in its own cluster? If so, in which approach will this tend to be observed?

- A. Single-link
- B. Complete-link
- C. Centroid-based
- D. None of the above

ANSWER: B AND C

2. What assumption does the CURE clustering algorithm make with regards to the shape of the clusters?

- A. No assumption
- B. Spherical
- C. Elliptical

ANSWER: A

3. What would, in general, be the effect of increasing MinPts in DBSCAN while retaining the same Eps parameter? (Note that more than one statement may be correct)

- A. Increase in the sizes of individual clusters
- B. Decrease in the sizes of individual clusters
- C. Increase in the number of clusters
- D. Decrease in the number of clusters

ANSWER: B AND C

4. Considering single-link and complete-link hierarchical clustering, is it possible for a point to be closer to points in other clusters than to points in its own cluster? If so, in which approach will this tend to be observed?

- A. No
- B. Yes, single-link clustering
- C. Yes, complete-link clustering
- D. Yes, both single-link and complete-link clustering

ANSWER: D

5. A graph is said to be k-connected if there does not exist a set of k-1 vertices whose removal disconnects the graph. If we define clusters as comprising of k-connected components of the thresholded graphs, does this result in a well-defined clustering algorithm?

- A. Yes
- B. No

ANSWER: A

6. A set of nodes forms a p-cluster, if at least p percentage of the edges from the nodes in the set go to another node in the set. If we define clusters as comprising of p-clusters of the thresholded graphs, does this result in a well-defined clustering algorithm?

- A. Yes
- B. No

ANSWER: B

7. In the CURE clustering algorithm, representative points of a cluster are moved a fraction of the distance between their original location and the centroid of the cluster. Would it make more sense to move them all a fixed distance towards the centroid instead? Why or why not?

- A. Yes, because this approach will ensure that the original cluster shape is preserved.
- B. No, because this approach will not be as effective against outliers as the original approach.

ANSWER: B

8. Suppose while performing DBSCAN we randomly choose a point which has less than MinPts number of points in its neighbourhood. Which among the following is true for such a point?

- A. It is treated as noise, and not considered further in the algorithm
- B. It becomes part of its own cluster
- C. Depending upon other points, it may later turn out to be a core point
- D. Depending upon other points, it may be density connected to other points

ANSWER: D

9. Which of the following statements are true about similarity graph based representations which are used for spectral clustering? (Note that more than one statements may be correct)

- A. One can give a tighter upper bound than $O(n)$ (where n is the number of data points) on the maximum degree of the vertex corresponding to a point in its kNN based similarity graph representation
- B. One can give a tighter upper bound than $O(n)$ (where n is the number of data points) on the maximum degree of the vertex corresponding to a point in its epsilon neighborhood based similarity graph representation
- C. If a is in the k nearest neighbors of b, then b is in the k nearest neighbors of a
- D. If a is in the epsilon neighborhood of b, then b is in the epsilon neighborhood of a

ANSWER: A AND D

10. Which of the following is untrue regarding Expectation Maximization algorithm?

- A. An initial guess is made as to the location and size of the site of interest in each of the sequences, and these parts of the sequence are aligned
- B. The alignment provides an estimate of the base or amino acid composition of each column in the site
- C. The column-by-column composition of the site already available is used to estimate the probability of finding the site at any position in each of the sequences
- D. The row-by-column composition of the site already available is used to estimate the probability

ANSWER: D

11. Out of the two repeated steps in EM algorithm, the step 2 is _____

- A. the maximization step
- B. the minimization step
- C. the optimization step
- D. the normalization step

ANSWER:A

12. Point out the correct statement.

- A. The choice of an appropriate metric will influence the shape of the clusters
- B. Hierarchical clustering is also called HCA
- C. In general, the merges and splits are determined in a greedy manner
- D. All of the mentioned

ANSWER:D

13. Which of the following is finally produced by Hierarchical Clustering?

- A. final estimate of cluster centroids
- B. tree showing how close things are to each other
- C. assignment of each point to clusters
- D. all of the mentioned

ANSWER:B

14. Which of the following is required by K-means clustering?

- A. defined distance metric
- B. number of clusters
- C. initial guess as to cluster centroids
- D. all of the mentioned

ANSWER:D

15. Hierarchical clustering should be primarily used for exploration.

- A. True
- B. False

ANSWER:A

16. _____ clustering approach initially assumes that each data instance represents a single cluster

- A. Hierarchical Clustering
- B. Expectation maximization clustering
- C. Agglomerative Clustering
- D. K means clustering

ANSWER:C

17. Sentiment Analysis is an example of:

- a. Regression
- b. Classification
- c. Clustering
- d. Reinforcement Learning

- A. a Only
- B. a and b
- C. a and c
- D. a, b and c
- E. a, b and d

ANSWER:E

18. Can decision trees be used for performing clustering?

- A. True
- B. False

ANSWER: A

19. After performing K-Means Clustering analysis on a dataset, you observed the following dendrogram. Which of the following conclusion can be drawn from the dendrogram?

- A. There were 28 data points in clustering analysis
- B. The best no. of clusters for the analyzed data points is 4
- C. The proximity function used is Average-link clustering
- D. The above dendrogram interpretation is not possible for K-Means clustering analysis

ANSWER: D

20. What could be the possible reason(s) for producing two different dendograms using agglomerative clustering algorithm for the same dataset?

- A. Proximity function used
- B. of data points used
- C. of variables used
- D. B and c only
- E. All of the above

ANSWER: E

21. What is the most appropriate no. of clusters for the data points represented by the following dendrogram:

- A. 2
- B. 4
- C. 6
- D. 8

ANSWER: B

22. Which of the following metrics, do we have for finding dissimilarity between two clusters in hierarchical clustering?

- a. Single-link
- b. Complete-link
- c. Average-link
- d. Options:
 - A. a and b
 - B. a and c
 - C. b and c
 - D. a,b and c

ANSWER: D

23. Which of the following is/are valid iterative strategy for treating missing values before clustering analysis?
- A. Imputation with mean
 - B. Nearest Neighbor assignment
 - C. Imputation with Expectation Maximization algorithm
 - D. All of the above

ANSWER:C

24. If you are using Multinomial mixture models with the expectation-maximization algorithm for clustering a set of data points into two clusters, which of the assumptions are important:
- A. All the data points follow two Gaussian distribution
 - B. All the data points follow n Gaussian distribution ($n > 2$)
 - C. All the data points follow two multinomial distribution
 - D. All the data points follow n multinomial distribution ($n > 2$)

ANSWER: C

25. Chameleon is
- A. Density based algorithm
 - B. Partitioning based algorithm
 - C. Model based algorithm
 - D. Hierarchical clustering algorithm

ANSWER: D

26. Which of the following clustering algorithms suffers from the problem of convergence at local optima?
- 1. K- Means clustering algorithm
 - 2. Agglomerative clustering algorithm
 - 3. Expectation-Maximization clustering algorithm
 - 4. Diverse clustering algorithm
- A. 1 only
 - B. 2 and 3
 - C. 2 and 4
 - D. 1 and 3
 - E. 1,2 and 4
 - F. All of the above

ANSWER: D.

27. What could be the possible reason(s) for producing two different dendograms using agglomerative clustering algorithm for the same dataset?
- A. Proximity function used
 - B. of data points used
 - C. of variables used
 - D. B and c only
 - E. All of the above

ANSWER: E

28. Collaborative Filtering and Content Based Models are the two popular recommendation engines, what role does NLP play in building such algorithms.
- A. Feature Extraction from text
 - B. Measuring Feature Similarity

C Engineering Features for vector space learning model

D. All of these

ANSWER: D

29..The singular value decomposition of a real matrix is unique.

A.True

B.False

ANSWER: B

30. All but one of these techniques can be used for building a content filtering profile for a user. Which of these techniques is NOT used for building a content filtering profile?

- A. Provide an interface where users can specify and edit their own vector.
- B. Build an attribute preference vector from explicit user ratings.
- C. Build an attribute preference vector based on the most popular items in the catalog.
- D. Build an attribute preference vector based on user

ANSWER: C

31. Each of the following statements describes Entrée Style recommenders except one.

Which of these statements DOES NOT describe the Entrée Style Recommenders?

- A. They don't use individual users' ratings of the items anywhere in the recommendation process.
- B. They build a model of user preferences that can be used to provide personalized recommendations.
- C. They require a substantial collection of information about the items being recommended.
- D. They provide an interface that allows the user to refine recommendations by requesting items that differ in a certain way from the current recommendation.

ANSWER: B

32. When is "term-frequency" most useful as part of a content-filtering recommender?

- A. When certain items are much more popular than other items.
- B. When the attributes of the items can apply in different degrees to different items.
- C. When users are unlikely to have experienced many of the items in the system.
- D. When certain terms aren't very useful because they apply to too many different items.

ANSWER: B

33.TRUE/FALSE things that are *similar* about model based clustering and kmeans clustering .

- (i) both methods are used to assign data to clusters
- (ii) both methods work well with spherical clusters

ANSWER:TRUE

34.TRUE/FALSE things that are *different* about model based clustering and kmeans clustering .

(i) kmeans has no model, mclust uses a model

(ii) kmeans has no objective criterion for choosing number of clusters; mclust uses the objective BIC criterion.
ANSWER:TRUE

35. In model-based clustering, when do observations come from the same true cluster?

A. When they come from the same distribution

B. When they have the highest posterior probability of belonging to the same cluster

C. When they are close to each other in terms of Mahalanobis distance

D. When they are close to each other in terms of Euclidean distance

ANSWER:A

36.Which of the followings is responsible for indexing content in a manner that permits fast retrieval through multiple search mechanisms?

- A. legacy integration
- B. system architecture
- C. Scalability
- D. Collaborative filtering

Answers: D

37.Which of the following is an example of active learning?

- A. News Recommender system
- B. Dust cleaning machine
- C. Automated vehicle
- D. None of the mentioned

ANSWER:A

38. Autonomous Question/Answering systems are _____

- A. Expert Systems
- B. Rule Based Expert Systems
- C. Decision Tree Based Systems
- D. All of the mentioned

ANSWER: D

39.You have collected a data of about 10,000 rows of tweet text and no other information. You want to create a tweet classification model that categorizes each of the tweets in three buckets – positive, negative and neutral. Which of the following models can perform tweet classification with regards to context mentioned above?

- A. Naive Bayes
- B. SVM
- C. None of the above

ANSWER: C

40. Classes of Collaborative Filtering includes

- A. User Based
- B. Item Based
- C. Both User Based and Item Based
- D. None of them

ANSWER: C

Question	a	b	c	d	Answer
Which of the following methods do we use to find the best fit line for data in Linear Regression?	Least Square Error	Maximum Likelihood	Logarithmic Loss	Both A and B	a
Which of the following is a good test dataset characteristic?	Large enough to yield meaningful results	Is representative of the dataset as a whole	none of any	Both A and B - answer	d
following is the example of which learning . A child trying to take his/her first steps, then he will follows to start Observing others walking and supervised learning	unsupervised learning	reinforcement learning	deep learning		c
what is true about underfitting 1.statistical model describes random error or noise instead of the underlying relationship. 2.model is excessiv	1 and 2	1, 2 and 3	3		4 d
what is true about overfitting 1.statistical model describes random error or noise instead of the underlying relationship. 2.model is excessiv	1 and 2	2 and 3	1, 2 and 3		4 c
What should be the best choice for number of clusters based on the following results:	5	6	14 greater than 14		b
In which of the following cases will K-Means clustering fail to give good results? 1) Data points with outliers	1 and 2	2 and 3	2 and 4	1, 2 and 4	d
How can Clustering (Unsupervised Learning) be used to improve the accuracy of Linear Regression model (Supervised Learning)?	1 and 4	3 only	2 and 4	All of the above	d
1. Why is an RNN (Recurrent Neural Network) used for machine translation, say translating English to French? (Check all that apply.) a. It can predict next word in a sentence	a and c	ab and c	a and d	a,b and d	a
which of the following is not application of Autoencoder	detecting anomalies in a signal	removing noise from an image, audio or scanned document	lowering the dimensions for better visualizations		a
Which of the following is true about Naive Bayes ?	Assumes that all the features in a dataset are equally important	Assumes that all the features in a dataset are independent	None of the above options		c
In which of the following cases will K-means clustering fail to give good results? 1) Data points with outliers 2) Data points with different densities	1 and 2	2 and 3	1, 2, and 3		c
Which of the following techniques would perform better for reducing dimensions of a data set?	Removing columns which have too many missing values	Removing columns which have high variance in data	Removing columns with dissimilar data trends	None of these	a
Which of the following algorithms cannot be used for reducing the dimensionality of data?	t-SNE	PCA	LDA False	None of these	d
The most popularly used dimensionality reduction algorithm is Principal Component Analysis (PCA). Which of the following is/are true about PCA?	1 and 2	2 and 3	1, 2 and 4	All of the above	d
Suppose we are using dimensionality reduction as pre-processing technique, i.e., instead of using all the features, we reduce the data to k dimensions.	Higher λ means more regularization	Higher λ means less regularization	Can't Say	None of these	b
PCA works better if there is	1. linear structure in the data	2. If the data lies on a curved surface and non-linear structure	1 and 3	1, 2 and 3	c
What happens when you get features in lower dimensions using PCA?	1. The features will still have interpretability	1 and 3	2 and 3	2 and 4	d
A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college.	Feature F1 is an example of nominal variable.	Feature F1 is an example of ordinal variable.	It doesn't belong to any of the above category.	Both of these	b
Which of the following is an example of a deterministic algorithm?	PCA	K-Means	None of the above	All of the above	a
When performing regression classification, which of the following is the correct way to preprocess the data?	Normalize the data \hat{A} PCA \hat{A} training	PCA \hat{A} normalize PCA output \hat{A} training	Normalize the data \hat{A} PCA \hat{A} normalize PCA output \hat{A}	None of the above	a
Which of the following is an example of feature extraction?	Constructing bag of words vector from an email	Applying PCA projects to a large high-dimensional data	Removing stopwords in a sentence	All of the above	d
What is pca.components_ in Sklearn?	Set of all eigen vectors for the projection space	Matrix of principal components	Result of the multiplication matrix	None of the above options	a
Which of the following is contained in NumPy library?	n-dimensional array object	tools for integrating C/C++ and Fortran code	fourier transform	all of the mentioned	d
To create sequences of numbers, NumPy provides a function	arange	aspace	affine	all of the mentioned	a
Point out the wrong statement	ipython is an enhanced interactive Python shell	matplotlib will enable you to plot graphics	rPy provides a lot of scientific routines that work on top	all of the mentioned	c
Point out the correct statement	NumPy main object is the homogeneous multidimensional array	NumPy, dimensions are called axes	Numpy array class is called ndarray	All of the mentioned	d
How we install Numpy in the system ?	install numpy	pip install python numpy	pip install numpy	pip install numpy python	c
Numpy in the Python provides the	Function	Lambda function	Type casting	Array	d
Best way to import the pandas module in your program ?	import pandas	import pandas as p	from pandas import *	All of the above	d
For what purpose a Pandas is used ?	To create a GUI programming	To create a database	To create a High level array	All of the above	c
In data science, which of the python library are more popular ?	Numpy	Pandas	OpenCV	Django	b
How do we load the iris dataset into scikit-learn?	from sklearn.datasets import load_iris	from sklearn.datasets access load_iris	from sklearn.datasets load load_iris	none of these	a
A model of language consists of the categories which does not include?	Language units	Role structure of units	System constraints	Structural units	d
Which of the following statement is true in following case?	Feature F1 is an example of nominal variable.	Feature F1 is an example of ordinal variable.	It doesn't belong to any of the above category.	Both of these	b
Imagine, you are working with Analytics Vidhya and you want to develop a machine learning algorithm which predicts the number of visitors per day. Your analysis is based on features like audience, number of articles written by the same author, etc.	True	False	True	True	that case?
Let's say, you are working with categorical feature(s) and you have not looked at the distribution of the categorical variable in the test data set. All categories of categorical variable are not present in the test data set.	Frequency distribution of categories is different in train and test data sets.	Train and Test always have same distribution.	Both A and B		d
Which of the following statements is/are true about Type-1• and Type-2• errors? 1. Type1 is known as false positive and Type2 is known as false negative. 2. Type1 is known as negative and Type2 is known as positive. 3. Type1 error occurs when we reject a null hypothesis when it is actually true.	1 and 2	2 and 3			d
Which of the following can be used to impute data sets based only on information in the training set?	postProcess	preProcess	process	all of the mentioned	b
Which of the following can also be used to find new variables that are linear combinations of the original set with independent components?	ICA	SCA	PCA	None of the mentioned	a
Which of the following function is used to generate the class distances?	preprocess.classDist	predict.classDist	predict.classDistance	all of the mentioned	b
Which of the following statements about regularization is not correct?	Using too large a value of lambda can cause your hypothesis to	Using too large a value of lambda can cause your hypothesis to	Using a very large value of lambda cannot hurt the perf	None of the above	d
Q- How can you prevent a clustering algorithm from getting stuck in bad local optima?	Set the same seed value for each run	Use multiple random initializations	Both A and B	None of the above	b
Which of the following techniques can be used for normalization in text mining?	Stemming	Lemmatization	Stop Word Removal	Both A and B -	d
Which of the following model include a backwards elimination feature selection routine?	MCV	MARS	MCRS	all of the mentioned	b
Which of the following is correct use of cross validation?	Selecting variables to include in a model	Comparing predictors	Selecting parameters in prediction function	All of the mentioned	d
Point out the wrong combination.	True negative=correctly rejected	False negative=correctly rejected	False positive=correctly identified	All of the mentioned	c
Which of the following is a common error measure?	Sensitivity	Median absolute deviation	Specificity	All of the mentioned	d
Which of the following is not a machine learning algorithm?	SVG	SVM	Random forest	None of the mentioned	a
Which of the following is a categorical outcome?	RMSE	R Squared	Accuracy	All of the mentioned	c
Point out the wrong statement.	ROC curve stands for receiver operating characteristic	Foretime series, data must be in chunks	Random sampling must be done with replacement	None of the mentioned	d
Which of the following can be used to impute data sets based only on information in the training set. ?	postProcess	Preprocess	process	all of the mentioned	b
Given below are three scatter plots for two features.In the above images, which of the following is/are example of multi-collinear features?	Features in Image 1	Features in Image 2	Features in Image 3	Features in Image 1 & 2	d
suppose you have identified multi-collinear features. Which of the following action(s) would you perform next? 1. Remove both collinear var	Only 2	Only 3	Either 1 or 3	Either 2 or 3	d
What is the limit of Factor Loadings?	-1	-1 < 1.0 to 1.0	No limit		c
What helps in determining the optimal number of factors in factor analysis	Screen Plot	Bell curve	Pie chart	Box plot	a
The percentage of the variance in the original variables that is captured by the system of factor equations together is called?	Factor Loading	Screen Plot	Variance Summarized	Community	d
Which of these about a dictionary is false?	The values of a dictionary can be accessed using keys	The keys of a dictionary can be accessed using values	c) Dictionaries aren't ordered	d) Dictionaries are mutable	b
Which of the following is not a declaration of the dictionary?	{1: 'A', 2: 'B'}	dict([(1, 'A'), (2, 'B')])	(1, 'A', 2, 'B')	()	c
What will be the output of the following Python code snippet? a=[1:"A",2:"B",3:"C"]	1 A 2 B 3 C	1 2 3	A B C	1:A 2:B 3:C	a
What will be the output of the following Python code snippet? a=[1:"A",2:"B",3:"C"] print(a.get(1,4))	1 A			4 Invalid syntax for get method	b
What will be the output of the following Python code snippet? a=[1:"A",2:"B",3:"C"]	Error, invalid syntax	A		5	d
What will be the output of the following Python code snippet? a=[1:"A",2:"B",3:"C"]	(1: 'A', 2: 'B', 3: 'C')	C	[1: 3, 2: 3, 3: 3]	No method called setdefault() exists for dictionary	b
What will be the output of the following Python code snippet? a=[1:"A",2:"B",3:"C"]	(1: 'A', 2: 'B', 3: 'C')	None	Error	[1,3,6,10]	a
What will be the output of the following Python code snippet? a=[1:"A",2:"B",3:"C"]	(1: 'A', 2: 'B', 3: 'C')	None	Keys must be immutable	Keys must be integers	c
If a is a dictionary with some key-value pairs, what does a.popitem() do?	Removes an arbitrary element	Removes all the key-value pairs	Removes the key-value pair for the key given as an arg	Invalid method for dictionary	a
Which of the following is characteristic of best machine learning method ?	Fast	Accuracy	Scalable	All of the mentioned	d
What will be the output of the following code :print type(type(int))	type 'int'	type 'type'	error		b
Which of the following function can be used to identify near zero-variance variables?	zeroVar	nearZeroVar		all of the mentioned	c
Which of the following function can be used to flag predictors for removal?	searchCorrelation	findCausation	findCorrelation	none of the mentioned	c
Point out the correct statement.	findLinearColumns will also return a vector of column position	findLinearCombos will return a list that enumerates dep	the function findLinearRows can be used to generate a	none of the mentioned	b

MCQs on Unit 1(One mark questions)

1. A system which performs following tasks:

Taking Input , Processing input & Providing output, can be called as:

- a. Adaptive System
- b. Classic System- answer**
- c. Reinforced Learning
- d. None of the above

2. A system which performs following tasks:

Taking Input , Processing input, Providing output & Tuning Parameters through Feedback from environment can be termed as:

- a. Adaptive System.**
- b. Classic System.
- c. Non Adaptive System
- d. None of the above

3. Classification and Regression techniques fall under the category of

- a. Supervised Learning**
- b. Unsupervised Learning
- c. Semi-supervised Learning
- d. Reinforcement Learning

4. Supervised Learning algorithms are accompanied by both Input and Expected Output?

- a. True- answer**
- b. False

5. Linear Regression, Random Forest , SVM are examples of

- a. Supervised Learning- answer**
- b. Unsupervised Learning
- c. Semi-Supervised Learning
- d. Reinforcement Learning

6. Decision Tree algorithm can work on

- a. Only Categorical values
- b. Only Continuous values
- c. Both Categorical and Continuous values- answer**
- d. None of the above

7. If the input and output variables are continuous in nature, which technique is more preferred?

- a. **Regression- answer**
- b. Classification
- c. Association Rule mining
- d. All of these

8. k-NN algorithm does more computation on ‘test’ time rather than ‘train’ time.

- a. **True- answer**
- b. False

9. Which of the following distance metric can be used in k-NN?

- a. Manhattan
- b. Minkowski
- c. Jaccard
- d. Mahalanobis
- e. **All can be used- answer**

10. Which of the following machine learning algorithm can be used for imputing missing values of both categorical and continuous variables?

- a. **K-NN- answer**
- b. Linear Regression
- c. Logistic Regression
- d. Decision Tree

11. Which of the following algorithm isNOT an example of ensemble learning algorithm

- a. Random Forest
- b. Adaboost
- c. Gradient Boosting
- d. **Decision Trees**

12. Spam detection, pattern detection, NLP are examples of

- a. Semi-supervised learning.
- b. **Supervised Learning**
- c. Unsupervised Learning
- d. All of these

13. Clustering technique & Association rule mining are examples of

- a. Supervised Learning
- b. Semi-supervised Learning
- c. **Unsupervised Learning- answer**
- d. Reinforcement Learning

14. Unsupervised Learning algorithms are accompanied by both Input and Expected Output?

- a. True
- b. False (Only Input) - answer**

15. K-Means technique is an example of

- a. Clustering- answer**
- b. Classification
- c. Regression
- d. Association

16. Which of the following is/are types of clustering

- a. Centroid-based Clustering
- b. Density-based Clustering
- c. Hierarchical Clustering
- d. All of the above- answer**

17. Learning algorithms that use both labelled and unlabelled data can be categorised as

- a. Supervised Algorithms
- b. Unsupervised Algorithms
- c. Semi-supervised Algorithms- answer**
- d. Reinforcement Learning

18. Reinforcement learning is particularly efficient when the environment is NOT completely deterministic

- a. True- answer**
- b. False

19. When the number of output classes is greater than one, which is / are the possible strategy used to handle them

- a. One-vs-All
- b. One-vs-One
- c. Both of them- answer**
- d. None of the above

20. In One-vs-All strategy how many classifiers are trained for n classes

- a. 1
- b. n- answer**
- c. n/2
- d. None of the above

21. In One-vs-One strategy how many classifiers are trained for n classes

- a. 1
- b. n
- c. n*(n-1)/2- answer**
- d. n/2

22. When the model isn't able to capture the dynamics shown by the same training set, such situation is called as

- a. Underfitting- answer**
- b. Overfitting
- c. Normal fitting
- d. Regularization

23. When the model can associate almost perfectly all the known samples to the corresponding output values, but when an unknown input is presented, the corresponding prediction error can be very high, such situation is called as

- a. Underfitting
- b. Overfitting- answer**
- c. Normal fitting
- d. None of these

24. The formula given below is to calculate _____

$$P(c | x) = \frac{P(x | c)P(c)}{P(x)}$$

Likelihood Class Prior Probability
↓ ↓
Posterior Probability Predictor Prior Probability

$$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \dots \times P(x_n | c) \times P(c)$$

- a. Posterior Probability in Naïve Bayes Classifier- answer**
- b. Prior Probability in Naïve Bayes Classifier

- c. Entropy in Decision Tree classifier
 - d. None of the above
- 25.
- The following formula is used to calculate _____

$$-\sum p(X) \log p(X)$$

- a. Information Gain
 - b. Entropy- answer**
 - c. Probability of an event
 - d. None of the above
26. Which algorithm is not a type of Parametric Learning?

- a. Logistic Regression
- b. Naïve Bayes
- c. K-Nearest Neighbors- answer**
- d. Simple Neural Networks

27. What is Machine learning?

- a. The autonomous acquisition of knowledge through the use of computer programs- answer**
- b. The autonomous acquisition of knowledge through the use of manual programs
- c. The selective acquisition of knowledge through the use of computer programs
- d. The selective acquisition of knowledge through the use of manual programs

28. Which of the factors affect the performance of learner system does not include?

- a. Representation scheme used
- b. Training scenario
- c. Type of feedback
- d. Good data structures- answer**

29. Which system is based on static or permanent structures?

- a. Adaptive system
- b. Non-adaptive system- answer**
- c. Both
- d. None of the above

30. Which is not a type of supervised learning algorithm?

- a. K-Nearest Neighbor
- b. Decision Tree
- c. K-means- answer**

- d. Linear Regression
31. From following, which are the approaches to Machine Learning?
- Supervised Learning
 - Unsupervised Learning
 - Reinforcement Learning
 - All of the above- answer
32. In which type of Learning, both features and labels are given to an algorithm?
- Supervised Learning- answer**
 - Unsupervised Learning
 - Reinforcement Learning
 - None of the above
33. In which type of learning, the algorithm maps input variable to output variable?
- Supervised Learning- answer**
 - Unsupervised Learning
 - Reinforcement Learning
 - None of the above
34. Which is not a type of Supervised Learning?
- Classification
 - Regression
 - Clustering- answer**
 - None of the above
35. Which approach should be used to e-mail spam filtering?
- Classification- answer**
 - Clustering
 - Regression
 - Association
36. Which approach should be used to predict sales of a supermarket?
- Classification
 - Clustering
 - Regression- answer**
 - Association
37. In which learning technique, the system discovers patterns from dataset?
- Supervised Learning
 - Unsupervised Learning- answer**
 - Reinforcement Learning
 - None of the above

38. In which type of learning, the problem can be solved without knowing labels?

- a. Supervised Learning
- b. Unsupervised Learning- answer**
- c. All of the above
- d. None of the above

39. Which type of problem discovers groups of data based on similarities?

- a. Clustering- answer**
- b. Association
- c. Regression
- d. None of the above

40. Which type of problem discovers rules to describe large data?

- a. Clustering
- b. Association- answer**
- c. Regression
- d. None of the above

41. From the following, which is best suited to build a game of chess?

- a. Supervised Learning
- b. Unsupervised Learning
- c. Deep Learning- answer**
- d. None of the above

42. In which type, rewards and punishments are given as a feedback?

- a. Supervised Learning
- b. Unsupervised Learning
- c. Reinforcement Learning- answer**
- d. None of the above

43. Which approach should be used for automatic labelling?

- a. Supervised Learning
- b. Unsupervised Learning- answer**
- c. Reinforcement Learning
- d. None of the above

44. From the options, which application you should solve by deep learning for the best performance?

- a. Spam filtering
- b. Image classification- answer**
- c. Sales prediction
- d. Automatic labelling

45. A neural network model is said to be inspired from the human brain. Which of the following statement(s) correctly represents a real neuron?

- a. A neuron has a single input and a single output only
- b. A neuron has multiple inputs but a single output only
- c. A neuron has a single input but multiple outputs
- d. **All of the above statements are valid- answer**

46. What is unsupervised learning?

- a. Features of group explicitly stated
- b. Number of groups may be known
- c. **Neither feature & nor number of groups is known- answer**
- d. None of the mentioned

47. Which is not a correct statement with respect to Deep Learning?

- a. Large computing power is required
- b. **Less complex than machine learning- answer**
- c. Difficulty in interpreting the resulting models
- d. Requires large amount of labelled data

48. Which algorithm is not a type of Non-parametric learning?

- a. **Naïve bayes- answer**
- b. C4.5
- c. K-Nearest Neighbor
- d. Support Vector Machines

49. In which type, the training data is modelled very well?

- a. Underfitting
- b. **Overfitting- answer**
- c. Both
- d. Not a and b

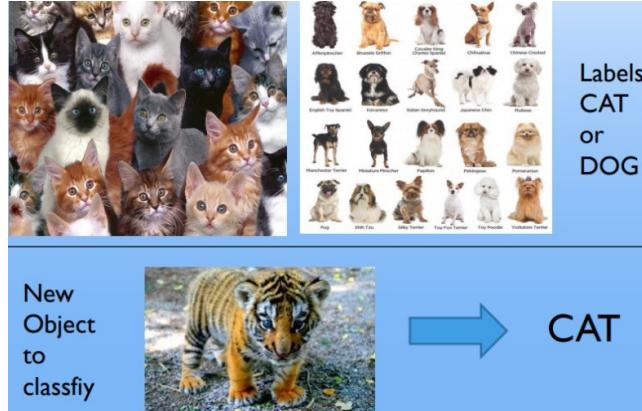
50. Which model gives poor performance on training data?

- a. **Underfitting- answer**
- b. Overfitting
- c. Both
- d. None of the above

Unit 1: Two marks questions

1. The goal(s) of the supervised learning system is (are) _____
 - a. Training a system that must also work with samples never seen before.
 - b. To allow the model to develop a generalization ability and avoid a common problem called over fitting
 - c. Supervisor: to provide the agent with a precise measure of its error
 - d. All of the above- answer**

2. Identify the type of model for the given problem



- a. Reinforcement learning
 - b. Supervised learning- answer**
 - c. Un supervised learning
 - d. Semi supervised learning
3. The goal (s) of Classification techniques is (are) _____
 - a. Try to find the best separating hyperplane (in this case, it's a linear problem).
 - b. Reduce the number of misclassifications
 - c. Increasing the noise-robustness
 - d. All of these- answer**
4. Consider D be a training set of n samples , each sample is represented by X of m features , $X = (x_1, x_2, x_3 \dots, x_n)$, Consider C classes : C1, C2..... Cc.
Bayesian classifier predicts that tuple X belongs to class Ci iff.
 - a. **P(Ci/X) > P(Cj/X) for i<= j<=c , j != i Thus we maximize P(Ci/X) - answer**
 - b. $P(Ci/X) < P(Cj/X)$ for i<= j<=c , j != i Thus we maximize P(Ci/X)
 - c. $P(Ci/X) > P(Cj/X)$ for i<= j<=c , j != i Thus we maximize P(Cj/X)
 - d. None of the above
5. The problem of high variance and low bias is called _____
 - a. Over-fitting- answer**
 - b. Underfitting
 - c. Normal fitting
 - d. Best fitting
6. Identify the type of Machine learning approach to solve the given problems:
Decision Support System to predict the decision to play Match or not to play
 - a. Reinforcement learning
 - b. Supervised learning- answer**

- c. Un supervised learning
- d. Semi supervised learning

7. Identify the type of Machine learning approach to solve the given problems:

Grouping of documents retrieved by Google Search Engine

- a. Reinforcement learning
- b. Supervised learning
- c. Un supervised learning- answer**
- d. Semi supervised learning

8. Identify the type of Machine learning approach to solve the given problems:

System to predict price of product in next year

- a. Reinforcement learning
- b. Supervised learning- answer**
- c. Unsupervised learning
- d. Semi supervised learning

9. Identify the type of Machine learning approach to solve the given problems:

System to predict the suitable treatment

- a. Reinforcement learning
- b. Supervised learning**
- c. Un supervised learning
- d. Semi supervised learning

10. Identify the type of Machine learning approach to solve the given problems:

System for Driverless Car

- a. Reinforcement learning- answer**
- b. Supervised learning
- c. Unsupervised learning
- d. Semi supervised learning

11. Which is true for AI, ML and DP

- a. AI>ML>DP- answer**
- b. DP>ML>AI
- c. ML>AI>DP
- d. DP>ML>AI

MCQ's on Unit 2: Feature selection (Two marks)

1. For creating Training and Test datasets which statements are true?
 - a. Both datasets must reflect the original distribution
 - b. The original dataset must be randomly shuffled before the split phase in order to avoid correlation between consequent elements
 - c. Both a and b - answer**
 - d. None of the above
2. SK-Learn provides which function to create train and test data:
 - a. **train_test_split- answer**
 - b. test_train_split
 - c. TestTrainSplit
 - d. Split_test_train
3. In scikit-learn LabelEncoder class:
 - a. Adopts a dictionary-oriented approach,
 - b. Associating to each category label a progressive integer number,
 - c. That is an index of an instance array called classes_
 - d. All of the above- answer**
4. Scikit-learn class Imputer fills the holes using a strategy based on the:
 - a. mean
 - b. median
 - c. frequency (the most frequent entry)
 - d. All of the above- answer**
5. Consider 3 dimensional dataset given below

x	y	z
1	Nan	2
2	3	nan
-1	4	2

SK-Learn Imputer mean strategy will fill missing values with

- a. 3, 2
- b. 4, 2

c. 3.5, 2 - answer

d. Difficult to tell

6. Consider 3 dimensional dataset given below

x	y	z
1	Nan	2
2	3	nan
-1	4	2

SK-Learn Imputer median strategy will fill missing values with

a. 3, 2

b. 4, 2

c. 3.5, 2 - answer

d. Difficult to tell

7. Consider 3 dimensional dataset given below

x	y	z
1	Nan	2
2	3	nan
-1	4	2

SK-Learn Imputer most_frequent strategy will fill missing values with

a. 3, 2- answer

b. 4, 2

c. 3.5, 2

d. Difficult to tell

8. Which statement(s) is (are) true for SK-Learn MinMaxScaler ?

a. Works well for cases when the distribution is *not Gaussian*

b. Works well when the standard deviation is very small

c. It is sensitive to outliers

d. All of these- answer

9. _____ uses the interquartile range , which makes it *robust* to outliers.

a. MonMaxScaler

b. Standard Scaler

c. Robust Scaler- answer

d. None of these

10. Consider Q1=31 and Q3=119. The inter quartile range (IQR) will be _____

a. 88 - answer

- b. -88
- c. 150
- d. -150

MCQs on unit 2 (One mark question)

- 1) Which of the following contains train_test_split() function
 - A) sklearn.feature_extraction
 - B) sklearn.preprocessing
 - C) sklearn.model_selection- answer**
 - D) sklearn.decomposition
- 2) Default value of test_size in train_test_split() when both test_size and train_size are none
 - A) 0.33
 - B) 0.25 - answer**
 - C) 0.50
 - D) 0.20
- 3) The LabelEncoder class, adopts which approach?
 - A) **Dictionary-oriented- answer**
 - B) List-oriented
 - C) Tree-oriented
 - D) Map-oriented
- 4) FeatureHasher class in scikit-learn adopts which hashing technique:
 - A) SHA256
 - B) MD5
 - C) MurmurHash 3- answer**
 - D) BLAKE3
- 5) Which of the following is best option to handle missing data?
 - A) Removing the whole line
 - B) Creating sub-model to predict those features
 - C) Using an automatic strategy to input them according to the other known values- answer**
 - D) Inserting random values
- 6) When performing regression or classification, which of the following is the correct way to preprocess the data?
 - A) Normalize the data → PCA → training - answer**
 - B) PCA → normalize PCA output → training
 - C) Normalize the data → PCA → normalize PCA output → training
 - D) None of the above

- 7) What is `pca.components_` in Sklearn?
- A) Set of all eigen vectors for the projection space - answer
 - B) Matrix of principal components
 - C) Result of the multiplication matrix
 - D) None of the above options
- 8) How do you handle missing or corrupted data in a dataset?
- A) Drop missing rows or columns
 - B) Replace missing values with mean/median/mode
 - C) Assign a unique category to missing values
 - D) All of the above - answer
- 9) The class KernelPCA, which performs a PCA with?
- A) non-linearly separable data sets - answer
 - B) linearly separable data sets
 - C) categorical data sets
 - D) Heterogeneous data sets
- 10) Principal component analysis is a method to select only a subset of features which contain the largest amount of?
- A) Total covariance
 - B) Total variance - answer
 - C) Total count
 - D) Mean
- 11) In the following loss function which parameter controls the level of sparsity?
- $$L(\mathbf{D}, \mathbf{A}) = \frac{1}{2} \sum_i \|\mathbf{x}_i - \mathbf{D}\bar{\alpha}_i\|_2^2 + c \|\bar{\alpha}_i\|_1$$
- A) \mathbf{x}_i
 - B) c - answer
 - C) D
 - D) α_i
- 12) Which parameter determines the number of atoms in scikit-learn DictionaryLearning class?
- A) alpha
 - B) n_jobs
 - C) n_components - answer
 - D) tol
- 13) In KernelPCA the default value for gamma is?
- A) 1.0/number of features - answer
 - B) 2.0/number of features
 - C) 10/number of features
 - D) None of above
- 14) Non negative matrix factorization algorithm optimizes a loss function based on?
- A) L1 Norm
 - B) Frobenius norm - answer

- C) linalg.norm
D) matrix norm
- 15) Which of the following encoding technique is efficient to deal with large number of possible categories?
- A) Effect Encoding
 - B) Feature Hashing
 - C) One Hot Encoding
 - D) Bin counting scheme - answer**
- 16) Which scaling technique scales data without being affected by outliers?
- A) Robust Scaling - answer**
 - B) Min Max Scaling
 - C) Standardized Scaling
 - D) Z-score Scaling
- 17) Which feature selection technique use recursive approach?
- A) Filter Methods
 - B) Wrapper Methods - answer**
 - C) Embedded Methods
 - D) Subset Methods
- 18) From the following which can be applied on dataset with more than one dimension?
- A) Mean
 - B) Standard Deviation
 - C) Covariance - answer**
 - D) Variance
- 19) In principal component analysis the sparse loadings can be obtained by imposing which constraint on regression coefficients:
- A) Ridge
 - B) Lasso - answer**
 - C) Linear
 - D) Logistic
- 20) What provides better statistical regularization?
- A) Sparse PCA - answer**
 - B) Kernel PCA
 - C) Non-negative Matrix Factorization
 - D) Atom Extraction
- 21) Eigen vector with _____ Eigen value is the principle component of dataset.
- A) Lowest
 - B) Highest - answer**
 - C) Mean
 - D) Zero

22) Trace is equal to the ___ of the Eigen values.

- A) Difference
- B) Sum - answer**
- C) Product
- D) Mean

23) In which scaling technique the upper and lower can be specified by user?

- A) Robust Scaling
- B) Min Max Scaling - answer**
- C) Standardized Scaling
- D) Z-score Scaling

24) Principal component analysis (PCA) can be used with variables of any mathematical types: quantitative, qualitative, or a mixture of these types.

- A) True
- B) False - answer**

25) Variances and covariances can be computed for variables of any mathematical types: quantitative, qualitative, or a mixture of these types.

- A) True
- B) False - answer**

Unit- 3: Regression (One mark)

1. A process by which we estimate the value of dependent variable on the basis of one or more independent variables is called:
 - a. Correlation
 - b. Regression - answer**
 - c. Residual
 - d. Slope
2. All data points falling along a straight line is called:
 - a. **Linear relationship - answer**
 - b. Non linear relationship
 - c. Residual
 - d. Scatter diagram
3. A relationship where the flow of the data points is best represented by a curve is called:
 - a. Linear relationship
 - b. Nonlinear relationship - answer**
 - c. Linear positive
 - d. Linear negative
4. The value we would predict for the dependent variable when the independent variables are all equal to zero is called:
 - (a) Slope
 - (b) Sum of residual
 - (c) Intercept - answer**
 - (d) Difficult to tell
5. The predicted rate of response of the dependent variable to changes in the independent variable is called:
 - (a) Slope - answer**
 - (b) Intercept
 - (c) Error
 - (d) Regression equation
6. The slope of the regression line of Y on X is also called the:
 - (a) Correlation coefficient of X on Y
 - (b) Correlation coefficient of Y on X
 - (c) Regression coefficient of X on Y
 - (d) Regression coefficient of Y on X - answer**

8. In simple linear regression, the numbers of unknown constants are:

- (a) One
- (b) Two - answer**
- (c) Three
- (d) Four

9. In simple regression equation, the numbers of variables involved are:

- (a) 0
- (b) 1
- (c) 2 - answer**
- (d) 3

10. If the value of any regression coefficient is zero, then two variables are:

- (a) Qualitative
- (b) Correlation
- (c) Dependent
- (d) Independent- answer**

11. In SK-Learn Linear Regression offers two instance variables, _____ and _____

- a) intercept_and_coef_ - answer**
- b) Intercept and coef
- c) Slope and Intercept
- d) Slope and Coef

12. _____ regression imposes an additional shrinkage penalty to the ordinary least squares loss function to limit its squared $L2$ norm:

$$L(\bar{w}) = \|\mathbf{X}\bar{w} - \bar{y}\|_2^2 + \alpha\|\bar{w}\|_2^2$$

- a) Lasso
- b) LassoCV
- c) Ridge - answer**
- d) ElasticNet

13. _____ regressor imposes a penalty on the $L1$ norm of w to determine a potentially higher number of null coefficients:

$$L(\bar{w}) = \frac{1}{2n} \|\mathbf{X}\bar{w} - \bar{y}\|_2^2 + \alpha\|\bar{w}\|_1$$

- a) Lasso - answer**
- b) RidgeCV
- c) Ridge
- d) ElasticNet

14. A Regression approach to avoid the problem of outliers is offered by _____

- a) Linear Regression
- b) Logistic Regression
- c) RANSAC Regressor - answer**
- d) Polynomial Regressor

15. Model with high variance and low bias is called _____

- a) Over-fitted model - answer**
- b) Under-fitted model
- c) Best fitted
- d) None of the above

16. _____ occurs when our model neither fits the training data nor generalizes on the new data.

- a) Over-fitting
- b) Under-fitting - answer**
- c) Best fitting
- d) None of the above

17. _____ is the process of adding information in order to solve an ill-posed problem or to prevent overfitting

- a) Under-fitting
- b) Regularization - answer**
- c) Best fitting
- d) None of the above

18. _____ selects the only some feature while reduces the coefficients of others to zero.
This property is known as feature selection

- a) Lasso - answer**
- b) RidgeCV
- c) Ridge
- d) ElasticNet

19. _____ combines both Lasso and Ridge Regression into one model with two penalty factors, one proportional to L1 norm and other proportional to L2 norm.

- a) LassoCV
- b) RidgeCV
- c) ElasticNet - answer**
- d) None of the above

20. _____ minimizes the cost function by gradually updating the weight values.

a. Gradient Descent - answer

b. Perceptron

C. Grid search

d. None of the above

21. _____ is a technique allows using linear models even when the dataset has strong non-linearities. The idea is to add some extra variables computed from the existing ones and using (in this case) only polynomial combinations.

- a) Linear Regression
- b) Logistic Regression
- c) RANSAC Regressor
- d) Polynomial Regressor - answer**

22. The Regression technique that uses sigmoid function is called _____

- a) Linear Regression
- b) Logistic Regression - answer**
- c) RANSAC Regressor
- d) Polynomial Regressor

23. Confusion Matrix can be used to measure the performance of _____ model.

- a) Linear Regression
- b) Logistic Regression - answer**
- c) RANSAC Regressor
- d) Polynomial Regressor

24. The residual is defined as the difference between the:

- a) actual value of y and the estimated value of y - answer**
- b) actual value of x and the estimated value of x
- c) actual value of y and the estimated value of x
- d) actual value of x and the estimated value of y

25) Which of the following methods do we use to find the best fit line for data in Linear Regression?

- A) Least Square Error
- B) Maximum Likelihood
- C) Logarithmic Loss
- D) Both A and B

Answer: (A)

26) True- False: Overfitting is more likely when you have a huge amount of data to train.

- A) TRUE
 - B) FALSE
- Solution: (B)

27) What will happen when you apply very large penalty in the case of Lasso?

- A) Some of the coefficients will become zero
- B) Some of the coefficients will be approaching to zero but not absolute zero
- C) Both A and B depending on the situation
- D) None of these

Solution: (A)

28) Generally, which of the following method(s) is used for predicting continuous dependent

variable?

1. Linear Regression 2. Logistic Regression

- A) 1 and 2
- B)only 1
- C)only 2
- D)None of these

Solution:(B)

29)Full form of ROC is

- A)Regression Operation Characteristics Curve
- B)Receiver Operating Characteristics Curve
- C)Regression Operating Characteristics Curve
- D)Ridge Operation Characteristics Curve

Solution:(B)

30.F score is given by :

- A) $F=2*(precision+recall)/precision*recall$
- B) $F=(precision+recall)/precision*recall$
- C) $F=2*(precision*recall)/(precision+recall)$
- D) $F=precision+recall$

31)Which is L1 regression

- A)Lasso
- B)Ridge
- C)polynomial
- D)Isotonic

Answer A

32)Which of the following is true about “Ridge” or “Lasso” regression methods in case of feature selection?

- A) Ridge regression uses subset selection of features
- B)Lasso regression uses subset selection of features
- C)Both use subset selection of features
- D)None of the above

Solution:(B)

33)SSE can never be

- (A) larger than SST
- (B) smaller than SST
- (C)equal to 1
- (D)equal to zero

Solution:(A)

34) 1. Which of the following is correct about regularized regression?

- a) Can help with bias trade-off
- b) Cannot help with model selection
- c) Cannot help with variance trade-off
- d) All of the mentioned

Solution:(A)

35) Which of the following statement is true about outliers in Linear regression?

- A) Linear regression is sensitive to outliers
- B) Linear regression is not sensitive to outliers
- C) Can't say
- D) None of these

Solution: (A)

36) What do you expect will happen with bias and variance as you increase the size of training data?

- A) Bias increases and Variance increases
- B) Bias decreases and Variance increases
- C) Bias decreases and Variance decreases
- D) Bias increases and Variance decreases

Solution: (D)

37) A Pearson correlation between two variables is zero but, still, their values can still be related to each other.

- A) TRUE
- B) FALSE

Solution: (A)

38) Which of the following statement(s) is / are true for Gradient Decent (GD) and Stochastic Gradient Decent (SGD)?

1. In GD and SGD, you update a set of parameters in an iterative manner to minimize the error function.
2. In SGD, you have to run through all the samples in your training set for a single update of a parameter in each iteration.
3. In GD, you either use the entire data or a subset of training data to update a parameter in each iteration.

- A) Only 1
- B) Only 2
- C) Only 3
- D) 1 and 2

Solution:(A)

39) When hypothesis tests and confidence limits are to be used, the residuals are assumed to follow the _____ distribution.

- A) Formal
- B) Mutual
- C) Normal
- D) Abnormal

Solution:(C)

40) The error due to simplistic assumptions made by the model in fitting the data is called as

- A)variance
- B)bias
- C)MSE
- D)none of these

Solution:(B)

41) ROC curves show the trade-off between which parameters

- A)TPR and FPR
- B)TNR And TPR
- C)FPR and TNR
- D)FPR and FNR

Solution:(A)

42) The accuracy of the model can be measured by

- A)The area above ROC curve
- B)The area under ROC curve
- C)All of the above

D)None of the above

Solution:(B)

43) Least square method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the _____ deviations.

- a) Vertical
- b) Horizontal
- c) Both of these
- d) None of these

Solution:(A)

Unit-3 (Two marks)

1. The regression line $\hat{y} = 3 + 2x$ has been fitted to the data points (4,8), (2,5), and (1,2). The residual sum of squares will be:

- a) 10
- b) 15
- c) 13
- d) 22 - answer**

2. Suppose you have trained a logistic regression classifier and it outputs a new example x with a prediction $h_0(x) = 0.2$. This means

- a. Our estimate for $P(y=1 | x)$
- b. Our estimate for $P(y=0 | x)$ - answer**
- c. Our estimate for $P(y=1 | x)$
- d. Our estimate for $P(y=0 | x)$

3. A regression analysis between sales (in \$1000) and advertising (in \$100) resulted in the following least squares line: $\hat{y} = 75 + 6x$. This implies that if advertising is \$800, then the predicted amount of sales (in dollars) is:

- a. \$4875 - answer**
- b. \$123,000
- c. \$487,500
- d. \$12,300

4. The value for SSE equals zero. This means that the coefficient of determination (r^2) must equal:

- a. 0.0.
- b. -1.0.
- c. 2.3.
- d. 1.0 - answer**

5. Below equation shows the loss function of _____

$$L = \frac{1}{2} \sum_{i=1}^n \|\tilde{y}_i - y_i\|_2^2 \text{ which becomes } L = \frac{1}{2} \sum_{i=1}^n (\alpha + \beta x_i - y_i)^2$$

- a) Logistic Regression Model
- b) Linear Regression Model - answer**

- c) Gaussian Naïve Bayes Model
- d) Polynomial Model

6. For the given results of a recently conducted study on the correlation of the number of hours spent driving with the risk of developing acute backache. The Intercept of the line is _____.

Number of hours spent driving (x)	Risk score on a scale of 0-100 (y)
10	95
9	80
2	10
15	50
10	45
16	98
11	38
16	93

- a) **12.58 - answer**
- b) 10.58
- c) 11.85
- d) 10.85

7. For the given results of a recently conducted study on the correlation of the number of hours spent driving with the risk of developing acute backache. The slope of the line is _____.

Number of hours spent driving (x)	Risk score on a scale of 0-100 (y)
10	95
9	80
2	10
15	50
10	45
16	98
11	38
16	93

- a) **4.59 - answer**
- b) 10.58
- c) 5.85
- d) 10.85

8. for the given vector of outputs the Mean squared error is _____.

$$\begin{aligned} \mathbf{y_true} &= [3, -0.5, 2, 7] \\ \mathbf{y_pred} &= [2.5, 0.0, 2, 8] \end{aligned}$$

- a) 0.45
- b) **0.375 - answer**
- c) 0.56
- d) None of the above

9)The correct relationship between SST, SSR, and SSE is given by;

- a) $\text{SSR} = \text{SST} + \text{SSE}$
- b) $\text{SST} = \text{SSR} + \text{SSE}$
- c) $\text{SSE} = \text{SSR} - \text{SST}$
- d) all of the above

Solution:(B)

10)Stochastic gradient descent performs less computation per update than batch gradient descent.

A)True

B)False

Solution:(A)

11)A parameter that is external to model and whose value cannot be estimated from data is called as

A)Hyperparameter

B)Model Parameter

C)Outlier

D)Regularization constant

Solution:(A)

12)Which strategy is used for tuning hyperparameters

A)Gradient Descent

B)Feature Scaling

C)Regularization

D)Grid Search

Solution:(D)

13) Which is another term for true positive rate

A)precision

B)Recall

C)Specificity

D)Fscore

Solution:(B)

14)The most widely used metrics and tools to assess a classification model are:

A)Confusion matrix

B)Cost-sensitive accuracy

C)Area under the ROC curve

D>All of the above

Solution:(D)

15)Regularization term in ridge regression is

A) λ (sum of the absolute value of coefficients)

B) λ (sum of the square of coefficients)

C) λ square

D)None on these

Solution:(B)

16) In practice, Line of best fit or regression line is found when _____

a) Sum of residuals ($\Sigma(Y - h(X))$) is minimum

b) Sum of the absolute value of residuals ($\Sigma|Y-h(X)|$) is maximum

c) Sum of the square of residuals ($\Sigma (Y-h(X))^2$) is minimum

d) Sum of the square of residuals ($\Sigma (Y-h(X))^2$) is maximum

Solution:(C)

Unit- 4 : Naïve Bayes and SVM

(one mark)

1. Naive bayes falls under which category-

- a. Unsupervised classification learning
- b. Supervised classification learning
- c. Semi- supervised classification learning
- d. Reinforcement learning

Ans - b

2. What machine learning task is the Naive Bayes algorithm used for?

- a. dimensionality reduction
- b. clustering
- c. classification
- d. regression

Ans - c

3. Naive Bayes assumption about data is-

- a. input is independent, conditional on the output label.
- b. input is dependent, conditional on the output label.
- c. input is independent, not conditional on the output label.
- d. input is dependent, not conditional on the output label.

Ans - a

4. Bayes rule:

- a. $P(A | B) = P(B|A) . P(B) / P(A)$
- b. $P(A | B) = P(B|A) . P(A) / P(B)$
- c. $P(A | B) = P(B|A) . P(A)$
- d. $P(A | B) = P(B|A) . P(B)$

Ans - b

5. Which is not a main type of naive bayes classifier -

- a. Bernoulli naive bayes
- b. Multinomial naive bayes
- c. Gaussian naive bayes
- d. Complement Naive bayes

Ans - d

6. Which type of naive bayes classifier is suited for imbalanced datasets -

- a. Bernoulli naive bayes
- b. Multinomial naive bayes
- c. Gaussian naive bayes
- d. Complement Naive bayes

Ans - b

7. Which type of naive bayes classifier is best suited for document classification problem -

- a. Bernoulli naive bayes
- b. Multinomial naive bayes
- c. Gaussian naive bayes
- d. Complement Naive bayes

Ans - b

8. Which type of naive bayes classifiers is usually used for yes/no type boolean predictores-

- a. Bernoulli naive bayes
- b. Multinomial naive bayes
- c. Gaussian naive bayes
- d. Complement Naive bayes

Ans - a

9. Naive Bayes is termed as 'Naive' because it assumes-

- a. Dependence between every pair of feature in the data.
- b. It is multiclass classifier
- c. It is not multiclass classifier
- d. Independence between every pair of feature in the data.

Ans- d

10. SVM Classifiers and Linear Classifiers are strictly:

- a. Probabilistic Binary Linear Classifier
- b. Probabilistic Multiclass classifier
- c. Non Probabilistic Binary Linear Classifier
- d. Non Probabilistic Multiclass classifier

Ans - c

11. SVM falls under which category-

- a. Unsupervised classification learning
- b. Supervised classification learning
- c. Semi- supervised classification learning
- d. Reinforcement learning

Ans - b

12. The effectiveness of an SVM depends upon:

- a. Selection of Kernel
- b. Kernel Parameters
- c. Soft Margin Parameter C
- d. All of the above

Ans- D

9. Which of the following is true about Naive Bayes ?

- a. Assumes that all the features in a dataset are equally important
- b. Assumes that all the features in a dataset are independent
- c. **Both A and B - answer**
- d. None of the above options

(Two marks)

1. One marble jar has several different colored marbles inside of it. It has 1 red, 2 green, 4 blue, and 8 yellow marbles. All the marbles are the same size and shape. If Peter takes out a marble from the jar without looking, what is the probability that he will NOT choose a yellow marble.

- a. 7/15
- b. 8/15
- c. 7/8
- d. 5/8

Ans- a

2. If we train a Naive Bayes classifier using infinite training data that satisfies all of its modeling assumptions , then in general, what can we say about the training error (error in training data) and test error (error in held-out test data)?

- a. It may not achieve either zero training error or zero test error
- b. It will always achieve zero training error and zero test error.
- c. It will always achieve zero training error but may not achieve zero test error.
- d. It may not achieve zero training error but will always achieve zero test error.

Ans - a

3. If $P(A) = 0.10$, $P(B) = 0.05$.and $P(B|A) = 7\%$. Find $P(A|B)$ -

- a. 0.35
- b. 0.34
- c. 0.14
- d. 0.15

Ans - c

4. Which method is provided by scikit learn to tackle large scale classification for which full training set might not fit in memory-

- a. Memory_manage method
- b. Partial_manage method
- c. Partial_fit method
- d. None of the above

Ans - c

5. If I am using all features of my dataset and I achieve 100% accuracy on my training set, but ~70% on validation set, what should I look out for?

- a. Underfitting

- b. Nothing, the model is perfect
- c. Overfitting
- d. None of the above

Ans- C

6. What is/are true about kernel in SVM?

- 1.. Kernel function map low dimensional data to high dimensional space
- 2. It's a similarity function
 - a. 1
 - b. 2
 - c. 1 and
 - d. None of these

Ans- C

7. The performance of SVM depends on which factors

- a. the number of training instances
- b. the distribution of the data
- c. linear vs. non-linear problems
- d. input scale of the features
- e. All of the above

Ans - e

8. What do you mean by generalization error in terms of the SVM?

- a. How far the hyperplane is from the support vectors
- b. How accurately the SVM can predict outcomes for unseen data
- c. How much you want to avoid misclassification of each training example
- d. How far the influence of a single training example reaches.

Ans- b

9. What is regularisation parameter tells in SVM-

- a. How far the hyperplane is from the support vectors
- b. How accurately the SVM can predict outcomes for unseen data
- c. How much you want to avoid misclassification of each training example
- d. How far the influence of a single training example reaches.

Ans - c

10. What is gamma parameter tells in SVM-

- a. How far the hyperplane is from the support vectors
- b. How accurately the SVM can predict outcomes for unseen data

- c. How much you want to avoid misclassification of each training example
- d. How far the influence of a single training example reaches.

Ans - d

11. The SVM's are less effective when:

- a. The data is linearly separable
- b. The data is clean and ready to use
- c. The data is noisy and contains overlapping points
- d. None of the above

Ans- c

12. Which of the following are real world applications of the SVM?

- a. Text and Hypertext Categorization
- b. Image Classification
- c. Clustering of News Articles
- d. All of the above

Ans- d

13. What is the kernel trick -

- a. Polynomial and exponential kernels calculate the separation line in lower dimensions.
- b. Polynomial and exponential kernels calculate the separation line in higher dimensions.
- c. Polynomial or exponential kernels calculate the separation line in lower dimensions.
- d. Polynomial or exponential kernels calculate the separation line in higher dimensions.

Ans - b

Unit-V (One mark)

2. SK-Learn provides _____ in built class for Decision Tree Classifier?
 - a) DTClassifier
 - b) DecisionTreeClassifier - answer**
 - c) Tree
 - d) None of the above
3. What approach is taken by Decision Tree for Knowledge Engineering?
 - a) Inductive - answer**
 - b) Association Rules
 - c) Statistical
 - d) Substitutive
4. Which of the following is a widely used and effective machine learning algorithm based on the idea of bagging?
 - a. Decision Tree
 - b. Regression

- c. Classification
- d. **Random Forest** - answer
5. In the given formula of Decision Tree family , what A and D represents?
Gain(A) = Cross_Entropy(D) – EntropyA(D)
- Attribute, Decision
 - Attribute, Dataset- answer**
 - Probability, Dataset
 - None of the above
6. In the given formula of Decision Tree family , which are the given statements are true?
Gain(A) = Cross_Entropy(D) – EntropyA(D)
- Gain(A) should be maximum.
 - The attribute A with highest gain is chosen as the splitting attribute
 - Both a and b-answer**
 - None of the above
7. A _____ is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.
- Decision tree- answer**
 - Graphs
 - Trees
 - Neural Networks
8. 3. What is Decision Tree?
- Flow-Chart
 - Structure in which internal node represents test on an attribute, each branch represents outcome of test and each leaf node represents class label
 - Flow-Chart & Structure in which internal node represents test on an attribute, each branch represents outcome of test and each leaf node represents class label- answer**
 - None of the mentioned
9. Decision Trees can be used for Classification Tasks.
- True- answer**
 - False
10. The most widely used metrics and tools to assess a classification model are:
- Confusion matrix
 - Cost-sensitive accuracy
 - Area under the ROC curve
 - All of the above - answer**
11. Which of the following is a good test dataset characteristic?
- Large enough to yield meaningful results
 - Is representative of the dataset as a whole
 - Both A and B - answer**
 - None of the above
12. Which of the following is a disadvantage of decision trees?

- a. Factor analysis
 - b. Decision trees are robust to outliers
 - c. **Decision trees are prone to be overfit - answer**
 - d. None of the above
13. What is the purpose of performing cross-validation?
- a. To assess the predictive performance of the models
 - b. To judge how the trained model performs outside the sample on test data
 - c. **Both A and B – answer**
 - d. None of the above
14. **Which of the following is/are true about bagging trees?**
- 1. In bagging trees, individual trees are independent of each other
 - 2. Bagging is the method for improving the performance by aggregating the results of weak learners
 - A) 1
 - B) 2
 - C) 1 and 2- answer**
 - D) None of these
15. **Which of the following is/are true about boosting trees?**
- 1. In boosting trees, individual weak learners are independent of each other
 - 2. It is the method for improving the performance by aggregating the results of weak learners
 - A) 1
 - B) 2- answer
 - C) 1 and 2
 - D) None of these
16. Which of the following algorithm are not an example of ensemble learning algorithm?
- A) Random Forest
 - B) Adaboost
 - C) Extra Trees
 - D) Gradient Boosting
 - E) Decision Trees- answer**
17. Suppose you are using a bagging based algorithm say a RandomForest in model building. Which of the following can be true?
- 1. Number of tree should be as large as possible
 - 2. You will have interpretability after using RandomForest
 - A) 1- answer**
 - B) 2
 - C) 1 and 2
 - D) None of these

18. True-False: The bagging is suitable for high variance low bias models?

- A) TRUE- answer
- B) FALSE

19. In which of the following scenario a gain ratio is preferred over Information Gain?

- A) When a categorical variable has very large number of category - answer**
- B) When a categorical variable has very small number of category
- C) Number of categories is the not the reason
- D) None of these

20. In K-means clustering, the distance between each sample and each centroid is computed and the sample is assigned to the cluster where the distance is minimum. This approach is often called ----

- a. **Minimizing the inertia** of the clusters- answer
- b. Minimizing no. of clusters
- c. Maximizing the inertia of the clusters
- d. None of the above

21. Which statements are true about K-means method of clustering?

- 1)The process is iterative
- 2)All the distances are recomputed.
- 3)The algorithm stops when the centroids become stable and, therefore, the inertia is minimized
- 4) All of these- answer**

22. [True or False] k-NN algorithm does more computation on test time rather than train time.

- A) TRUE - answer
- B) FALSE

23. Which of the following statements is true for k-NN classifiers?

- A) The classification accuracy is better with larger values of k
- B) The decision boundary is smoother with smaller values of k
- C) The decision boundary is linear
- D) k-NN does not require an explicit training step- **answer**

Unit-5 (Two marks)

1. Which of the following is/are true about Random Forest and Gradient Boosting ensemble methods?

- 1.Both methods can be used for classification task
- 2.Random Forest is use for classification whereas Gradient Boosting is use for regression task
- 3.Random Forest is use for regression whereas Gradient Boosting is use for Classification task
- 4.Both methods can be used for regression task

A) 1

B) 2

C) 3

D) 4

E) 1 and 4 – answer

2. In Random forest you can generate hundreds of trees (say T₁, T₂T_n) and then aggregate the results of these tree. Which of the following is true about individual(T_k) tree in Random Forest?

1. Individual tree is built on a subset of the features
2. Individual tree is built on all the features
3. Individual tree is built on a subset of observations
4. Individual tree is built on full set of observations

A) 1 and 3 - answer

B) 1 and 4

C) 2 and 3

D) 2 and 4

3. Which of the following algorithm doesn't uses learning Rate as of one of its hyperparameter?

1. Gradient Boosting
2. Extra Trees
3. AdaBoost
4. Random Forest

A) 1 and 3

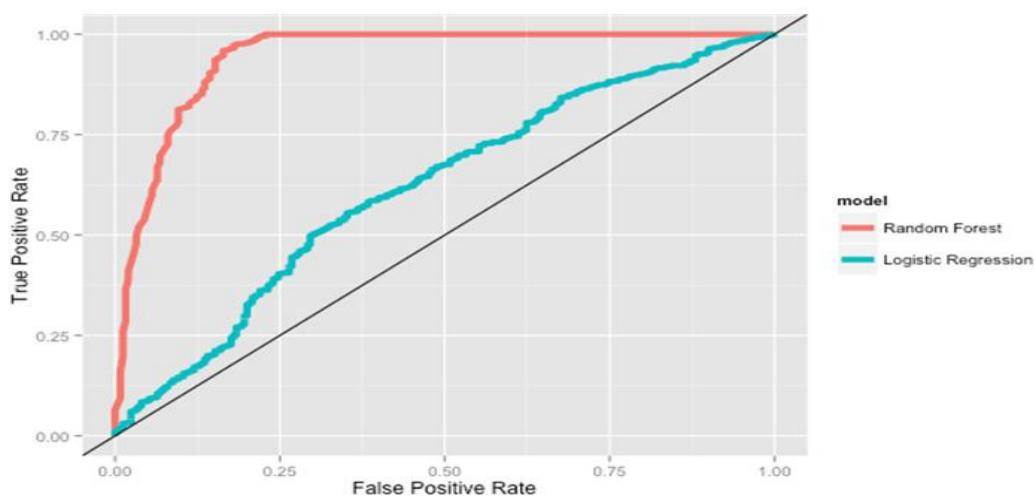
B) 1 and 4

C) 2 and 3

D) 2 and 4 - answer

4. Which of the following algorithm would you take into the consideration in your final model building on the basis of performance?

Suppose you have given the following graph which shows the ROC curve for two different classification algorithms such as Random Forest(Red) and Logistic Regression(Blue)



A) Random Forest- answer

- B) Logistic Regression
- C) Both of the above
- D) None of these

5. Which of the following is true about training and testing error in such case?

Suppose you want to apply AdaBoost algorithm on Data D which has T observations. You set half the data for training and half for testing initially. Now you want to increase the number of data points for training $T_1, T_2 \dots T_n$ where $T_1 < T_2 \dots T_{n-1} < T_n$.

E) The difference between training error and test error increases as number of observations increases

B) The difference between training error and test error decreases as number of observations increases- answer

C) The difference between training error and test error will not change

D) None of These

6. In random forest or gradient boosting algorithms, features can be of any type. For example, it can be a continuous feature or a categorical feature. Which of the following option is true when you consider these types of features?

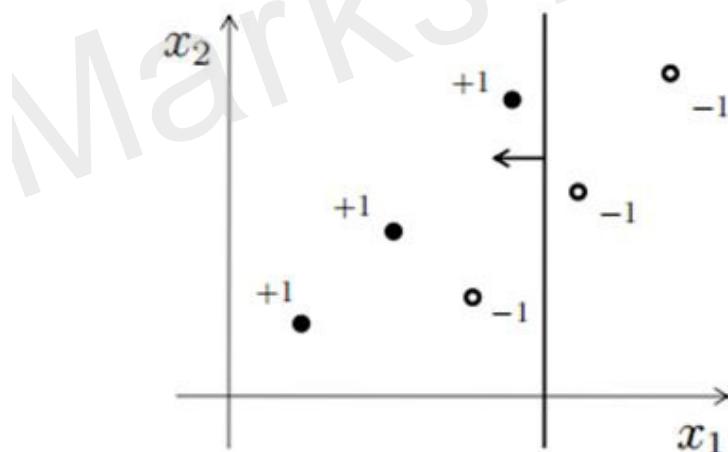
A) Only Random forest algorithm handles real valued attributes by discretizing them

B) Only Gradient boosting algorithm handles real valued attributes by discretizing them

C) Both algorithms can handle real valued attributes by discretizing them- answer

D) None of these

7. Consider the following figure for answering the next few questions. In the figure, X_1 and X_2 are the two features and the data point is represented by dots (-1 is negative class and +1 is a positive class). And you first split the data based on feature X_1 (say splitting point is x_{11}) which is shown in the figure using vertical line. Every value less than x_{11} will be predicted as positive class and greater than x_{11} will be predicted as negative class.



How many data points are misclassified in above image?

- A) 1- answer**
- B) 2
- C) 3
- D) 4

8. Suppose, you are working on a binary classification problem with 3 input features. And you chose to apply a bagging algorithm(X) on this data. You chose `max_features = 2` and the `n_estimators = 3`. Now, Think that each estimators have 70% accuracy.

Note: Algorithm X is aggregating the results of individual estimators based on maximum voting

What will be the maximum accuracy you can get?

- A) 70%
- B) 80%
- C) 90%
- D) 100%- answer**

9. Which of the following is true about the Gradient Boosting trees?

- 1. In each stage, introduce a new regression tree to compensate the shortcomings of existing model
- 2. We can use gradient decent method for minimize the loss function

- A) 1
- B) 2
- C) 1 and 2- answer**
- D) None of these

9. In SK-Learn which below parameters are in built in KMeans method

- a. cluster_centers_
- b. inertia_
- c. n_clusters
- d. all of the above**

10. In which of the following cases will K-means clustering fail to give good results?

- 1) Data points with outliers
- 2) Data points with different densities
- 3) Data points with nonconvex shapes

- 1. 1 and 2
- 2. 2 and 3
- 3. 1, 2, and 3 - answer**
- 4. 1 and 3

11. Which of the following is a reasonable way to select the number of clusters "k"?

- 1. Choose k to be the smallest value so that at least 99% of the variance is retained.
- 2. Choose k to be 99% of m ($k = 0.99*m$, rounded to the nearest integer).
- 3. Choose k to be the largest value so that 99% of the variance is retained.
- 4. Use the elbow method- answer**

12. A company has build a kNN classifier that gets 100% accuracy on training data. When they deployed this model on client side it has been found that the model is not at all accurate. Which of the following thing might gone wrong?

Note: Model has successfully deployed and no technical issues are found at client side except the model performance

- A) It is probably a overfitted model - answer**
- B) It is probably a underfitted model

- C) Can't say
- D) None of these

13. In k-NN it is very likely to overfit due to the curse of dimensionality. Which of the following option would you consider to handle such problem?

- 1. Dimensionality Reduction
- 2. Feature selection

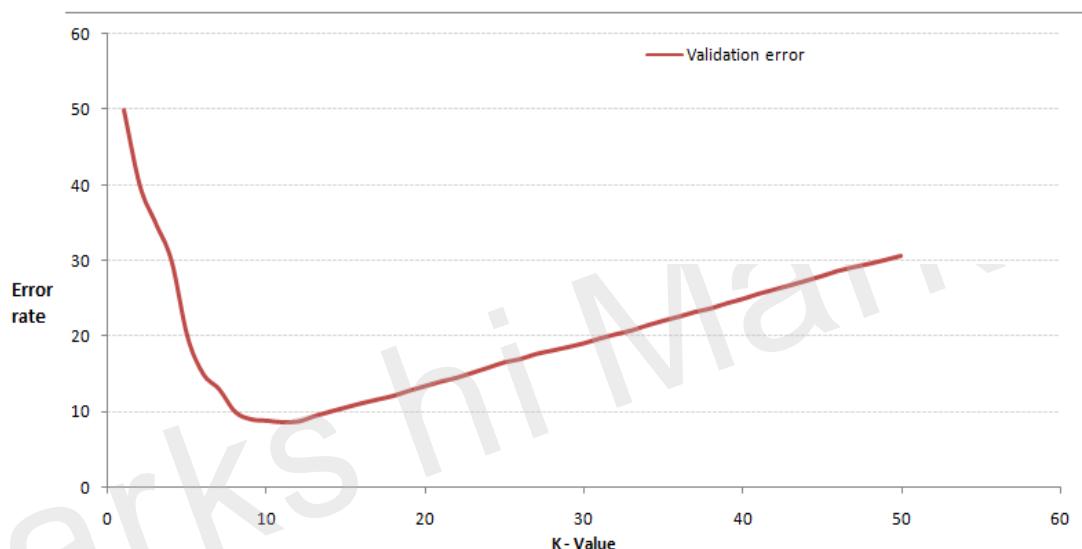
A) 1

B) 2

C) 1 and 2 - answer

D) None of these

14. In the image below, which would be the best value for k assuming that the algorithm you are using is k-Nearest Neighbor.



A) 3

B) 10 - answer

C) 20

D 50

15. Which of the following is/are not true about DBSCAN clustering algorithm:

- 1. For data points to be in a cluster, they must be in a distance threshold to a core point
- 2. It has strong assumptions for the distribution of data points in dataspace
- 3. It has substantially high time complexity of order $O(n^3)$
- 4. It does not require prior knowledge of the no. of desired clusters
- 5. It is robust to outliers

Options:

A. 1 only

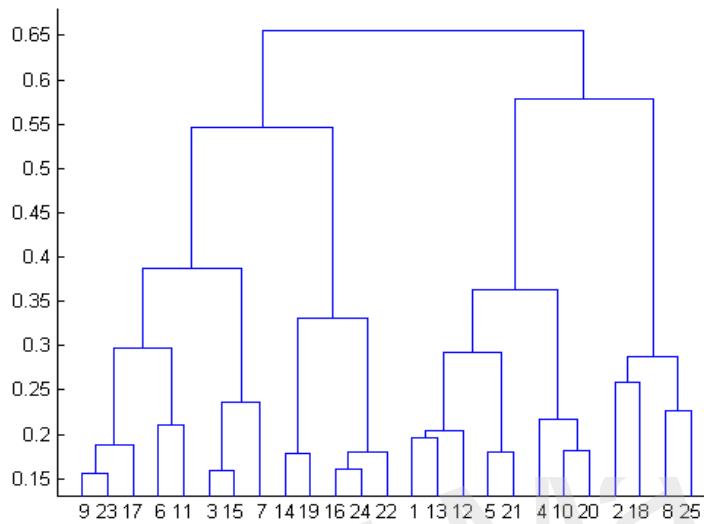
B. 2 only

C. 4 only

D. 2 and 3 - answer

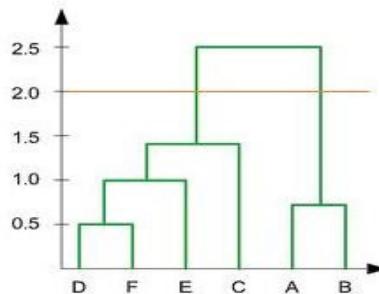
Unit-6 (Two marks)

1. After performing K-Means Clustering analysis on a dataset, you observed the following dendrogram. Which of the following conclusion can be drawn from the dendrogram?



- A. There were 28 data points in clustering analysis
B. The best no. of clusters for the analyzed data points is 4
C. The proximity function used is Average-link clustering
D. The above dendrogram interpretation is not possible for K-Means clustering analysis - answer

3. In the figure below, if you draw a horizontal line on y-axis for $y=2$. What will be the number



of clusters formed?

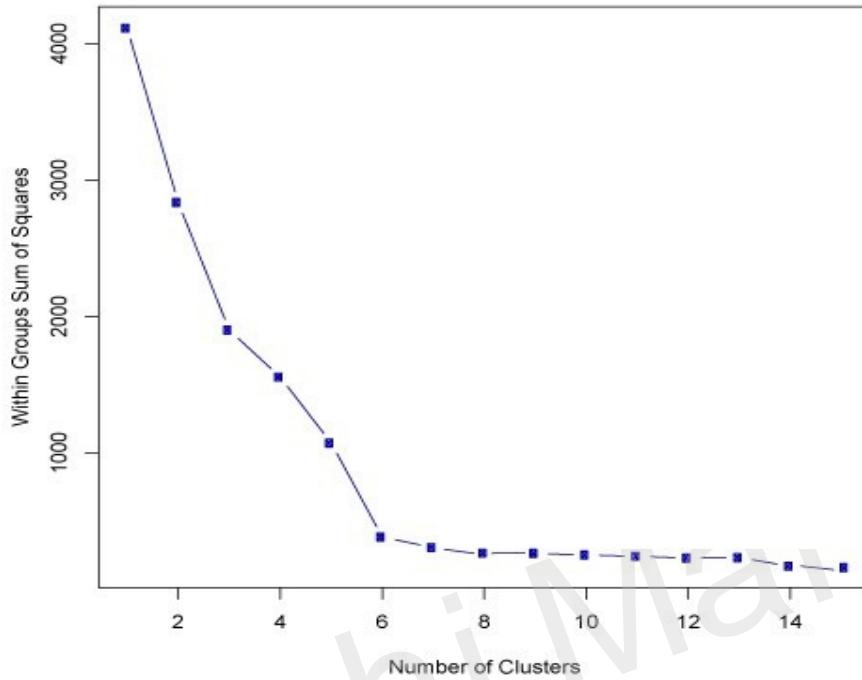
A. 1

B. 2 - answer

C. 3

D. 4

4. What should be the best choice for number of clusters based on the following results:



A. 5

B. 6 - answer

C. 14

D. Greater than 14

5. Which of the following is/are not true about Centroid based K-Means clustering algorithm and Distribution based expectation-maximization clustering algorithm:

1. Both starts with random initializations
2. Both are iterative algorithms
3. Both have strong assumptions that the data points must fulfill
4. Both are sensitive to outliers
5. Expectation maximization algorithm is a special case of K-Means
6. Both requires prior knowledge of the no. of desired clusters
7. The results produced by both are non-reproducible.

Options:

A. 1 only

B. 5 only - **answer**

C. 1 and 3

D. 6 and 7

7. If you are using Multinomial mixture models with the expectation-maximization algorithm for clustering a set of data points into two clusters, which of the assumptions are important:

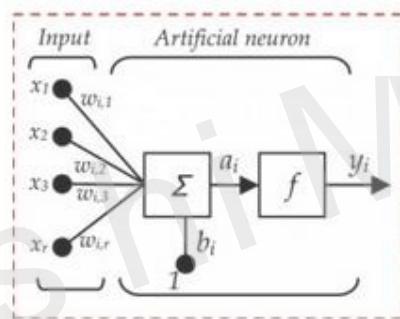
A. All the data points follow two Gaussian distribution

B. All the data points follow n Gaussian distribution ($n > 2$)

C. All the data points follow two multinomial distribution - answer

D. All the data points follow n multinomial distribution ($n > 2$)

8. Below is a mathematical representation of a neuron.



The different components of the neuron are denoted as:

- x_1, x_2, \dots, x_N : These are inputs to the neuron. These can either be the actual observations from input layer or an intermediate value from one of the hidden layers.
- w_1, w_2, \dots, w_N : The Weight of each input.
- b_i : Is termed as Bias units. These are constant values added to the input of the activation function corresponding to each weight. It works similar to an intercept term.
- a : Is termed as the activation of the neuron which can be represented as
- and y : is the output of the neuron

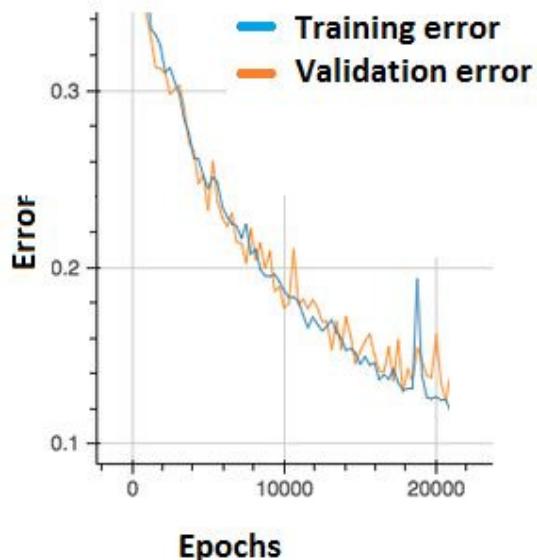
$$a = f\left(\sum_{i=0}^N w_i x_i\right)$$

Considering the above notations, will a line equation ($y = mx + c$) fall into the category of a neuron?

A. Yes- answer

B. No

9. In the graph below, we observe that the error has many “ups and downs”



Should we be worried?

- A. Yes, because this means there is a problem with the learning rate of neural network.
B. No, as long as there is a cumulative decrease in both training and validation error, we don't need to worry - answer

Unit 6 (One mark)

1. Which of the following metrics, do we have for finding dissimilarity between two clusters in hierarchical clustering?

1. Single-link
2. Complete-link
3. Average-link

Options:

- A. 1 and 2
B. 1 and 3
C. 2 and 3
D. 1, 2 and 3 - answer

2. Which of the following statement(s) correctly represents a real neuron?

- A. A neuron has a single input and a single output only

B. A neuron has multiple inputs but a single output only

C. A neuron has a single input but multiple outputs

D. A neuron has multiple inputs and multiple outputs

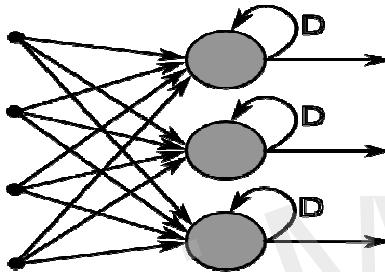
E. All of the above statements are valid - **answer**

3. If you increase the number of hidden layers in a Multi Layer Perceptron, the classification error of test data always decreases. True or False?

A. True

B. False - **answer**

4. You are building a neural network where it gets input from the previous layer as well as from itself.



Which of the following architecture has feedback connections?

A. Recurrent Neural network - **answer**

B. Convolutional Neural Network

C. Restricted Boltzmann Machine

D. None of these

5. In which neural net architecture, does weight sharing occur?

A. Convolutional neural Network

B. Recurrent Neural Network

C. Fully Connected Neural Network

D. Both A and B - **answer**

6. In a neural network, which of the following techniques is used to deal with overfitting?

A. Dropout

B. Regularization

C. Batch Normalization

D. All of these - **answer**

7. What is a dead unit in a neural network?

A. A unit which doesn't update during training by any of its neighbour - **answer**

- B. A unit which does not respond completely to any of the training patterns
 - C. The unit which produces the biggest sum-squared error
 - D. None of these
8. Suppose a convolutional neural network is trained on ImageNet dataset (Object recognition dataset). This trained model is then given a completely white image as an input. The output probabilities for this input would be equal for all classes. True or False?
- A. True
 - B. False - **answer**
9. For an image recognition problem (recognizing a cat in a photo), which architecture of neural network would be better suited to solve the problem?
- A. Multi Layer Perceptron
 - B. Convolutional Neural Network - answer**
 - C. Recurrent Neural network
 - D. Perceptron

10. What are the factors to select the depth of neural network?

- 1. Type of neural network (eg. MLP, CNN etc)
 - 2. Input data
 - 3. Computation power, i.e. Hardware capabilities and software capabilities
 - 4. Learning Rate
 - 5. The output function to map
- A. 1, 2, 4, 5
 - B. 2, 3, 4, 5
 - C. 1, 3, 4, 5
 - D. All of these - **answer**

11. Movie Recommendation systems are an example of:

- 1. Classification
- 2. Clustering
- 3. Reinforcement Learning
- 4. Regression

Options:

- 1. 2 Only
 - 2. 1 only
 - C. 1 and 2
 - D. 2 and 3 - **answer**
13. Recommendation systems are used in which of the following applications:

- a. Banking
- b. Shopping
- c. Search Engine
- d. **All of the above – answer**

14. Which of the following are methods of Recommendation Systems-

- a. Naïve User based systems,
- b. Content based Systems,
- c. Model free collaborative filtering
- d. All of the above – **answer**

15. Select correct option related to Hierarchical clustering.

- a. Creates sets of clusters
- b. Uses A tree data structure Dendrogram
- c. Only b
- d. **Both a and b- answer**

16. Agglomerative clustering is based on _____ approach

- a. Top Down
- b. **Bottom Up- answer**
- c. Linear
- d. Partition

17. For each pair of clusters, which algorithm computes the maximum distance between the clusters using below formula?

$$\forall C_i, C_j \ L_{ij} = \max\{d(x_a, x_b) \ \forall x_a \in C_i \text{ and } x_b \in C_j\}$$

- a. Single link
- b. **Complete link -answer**
- c. Average link
- d. Ward's Linkage

18. _____ Graphical method to better understand the agglomeration process shows in a static way how the aggregations are performed ,starting from the bottom (where all samples are separated) till the top (where the linkage is complete).

- a. Flow chart
- b. Histo graph
- c. **Dendrogram –answer**
- d. Decision tree

19. Which of the following functions are activation function?

- a. ReLU
- b. Tanh
- c. Sigmoid
- d. **All of the above- answer**

20. Which activation function is used by most of the Deep networks nowadays?

- a. ReLU - answer
- b. Tanh
- c. Sigmoid
- d. All of the above

21. _____ are general computers which can learn algorithms to map input sequences to output sequences

- a. CNN
- b. RNN- answer**
- c. Deep Q-Learning
- d. All of these

MCQ for unit 1: **Introduction to Machine learning**

1) Adaptive system management is

- A) It uses machine-learning techniques. Here program can learn from past experience and adapt themselves to new situations.
- B) Computational procedure that takes some value as input and produces some value as output.
- C) Science of making machines performs tasks that would require intelligence when performed by humans.
- D) None of these

Answer: A

2) Bayesian classifiers is

- A) A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory.
- B) Any mechanism employed by a learning system to constrain the search space of a hypothesis.
- C) An approach to the design of learning algorithms that is inspired by the fact that when people encounter new situations, they often explain them by reference to familiar experiences, adapting the explanations to fit the new situation.
- D) None of these

Answer: A

3) Algorithm is

- A) It uses machine-learning techniques. Here program can learn from past experience and adapt themselves to new situations.
- B) Computational procedure that takes some value as input and produces some value as output.
- C) Science of making machines performs tasks that would require intelligence when performed by humans.
- D) None of these

Answer: B

4) Bias is

- A) A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory.

- B) Any mechanism employed by a learning system to constrain the search space of a hypothesis.
- C) An approach to the design of learning algorithms that is inspired by the fact that when people encounter new situations, they often explain them by reference to familiar experiences, adapting the explanations to fit the new situation.
- D) None of these

Answer: B

5) Background knowledge referred to

- A) Additional acquaintance used by a learning algorithm to facilitate the learning process.
- B) A neural network that makes use of a hidden layer.
- C) It is a form of automatic learning.
- D) None of these

Answer: A

6) Case-based learning is

- A) A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory.
- B) Any mechanism employed by a learning system to constrain the search space of a hypothesis.
- C) An approach to the design of learning algorithms that is inspired by the fact that when people encounter new situations, they often explain them by reference to familiar experiences, adapting the explanations to fit the new situation.
- D) None of these

Answer: C

7) Classification is

- A) A subdivision of a set of examples into a number of classes.
- B) A measure of the accuracy, of the classification of a concept that is given by a certain theory.
- C) The task of assigning a classification to a set of examples
- D) None of these

Answer: A

8) Binary attribute are

- A) This takes only two values. In general, these values will be 0 and 1 and they can be coded as one bit
- B) The natural environment of a certain species.
- C) Systems that can be used without knowledge of internal operations.
- D) None of these

Answer: A

9) Classification accuracy is

- A) A subdivision of a set of examples into a number of classes
- B) Measure of the accuracy, of the classification of a concept that is given by a certain theory.
- C) The task of assigning a classification to a set of examples
- D) None of these

Answer: B

10) Biotope are

- A) This takes only two values. In general, these values will be 0 and 1 and they can be coded as one bit.
- B) The natural environment of a certain species
- C) Systems that can be used without knowledge of internal operations
- D) None of these

Answer: B

11) Cluster is

- A) Group of similar objects that differ significantly from other objects
- B) Operations on a database to transform or simplify data in order to prepare it for a machine-learning algorithm
- C) Symbolic representation of facts or ideas from which information can potentially be extracted
- D) None of these

Answer: A

12) Black boxes are

- A) This takes only two values. In general, these values will be 0 and 1 and they can be coded as one bit.
- B) The natural environment of a certain species

- C) Systems that can be used without knowledge of internal operations
- D) None of these

Answer: C

13) A definition of a concept is-----if it recognizes all the instances of that concept

- A) Complete
- B) Consistent
- C) Constant
- D) None of these

Answer: A

14) Data mining is

- A) The actual discovery phase of a knowledge discovery process
- B) The stage of selecting the right data for a KDD process
- C) A subject-oriented integrated time variant non-volatile collection of data in support of management
- D) None of these

Answer: A

15) A definition or a concept is----- if it classifies any examples as coming within the concept

- A) Complete
- B) Consistent
- C) Constant
- D) None of these

Answer: B

16) Data selection is

- A) The actual discovery phase of a knowledge discovery process
- B) The stage of selecting the right data for a KDD process
- C) A subject-oriented integrated time variant non-volatile collection of data in support of management
- D) None of these

Answer: B

17) Classification task referred to

- A) A subdivision of a set of examples into a number of classes
- B) A measure of the accuracy, of the classification of a concept that is given by a certain theory.
- C) The task of assigning a classification to a set of examples
- D) None of these

Answer: C

18) DNA (Deoxyribonucleic acid)

- A) It is hidden within a database and can only be recovered if one is given certain clues (an example IS encrypted information).
- B) The process of executing implicit previously unknown and potentially useful information from data
- C) An extremely complex molecule that occurs in human chromosomes and that carries genetic information in the form of genes.
- D) None of these

Answer: C

19) Hybrid is

- A) Combining different types of method or information
- B) Approach to the design of learning algorithms that is structured along the lines of the theory of evolution.
- C) Decision support systems that contain an information base filled with the knowledge of an expert formulated in terms of if-then rules.
- D) None of these

Answer: A

20) Discovery is

- A) It is hidden within a database and can only be recovered if one is given certain clues (an example IS encrypted information).
- B) The process of executing implicit previously unknown and potentially useful information from data.
- C) An extremely complex molecule that occurs in human chromosomes and that carries genetic information in the form of genes.
- D) None of these

Answer: B

21) Euclidean distance measure is

- A) A stage of the KDD process in which new data is added to the existing selection.
- B) The process of finding a solution for a problem simply by enumerating all possible solutions according to some pre-defined order and then testing them
- C) The distance between two points as calculated using the Pythagoras theorem.
- D) None of these

Answer: C

22) Hidden knowledge referred to

- A) A set of databases from different vendors, possibly using different database paradigms
- B) An approach to a problem that is not guaranteed to work but performs well in most cases
- C) Information that is hidden in a database and that cannot be recovered by a simple SQL query.
- D) None of these

Answer: C

23) Enrichment is

- A) A stage of the KDD process in which new data is added to the existing selection
- B) The process of finding a solution for a problem simply by enumerating all possible solutions according to some pre-defined order and then testing them
- C) The distance between two points as calculated using the Pythagoras theorem.
- D) None of these

Answer: A

24) Heterogeneous databases referred to

- A) A set of databases from different b vendors, possibly using different database paradigms
- B) An approach to a problem that is not guaranteed to work but performs well in most cases.

- C) Information that is hidden in a database and that cannot be recovered by a simple SQL query.
- D) None of these

Answer: A

25) Enumeration is referred to

- A) A stage of the KDD process in which new data is added to the existing selection.
- B) The process of finding a solution for a problem simply by enumerating all possible solutions according to some pre-defined order and then testing them
- C) The distance between two points as calculated using the Pythagoras theorem.
- D) None of these

Answer: B

26) Heuristic is

- A) A set of databases from different vendors, possibly using different database paradigms
- B) An approach to a problem that is not guaranteed to work but performs well in most cases
- C) Information that is hidden in a database and that cannot be recovered by a simple SQL query.
- D) None of these

Answer: B

27) Hybrid learning is

- A) Machine-learning involving different techniques
- B) The learning algorithmic analyzes the examples on a systematic basis 2nd makes incremental adjustments to the theory that is learned
- C) Learning by generalizing from examples
- D) None of these

Answer: A

28) Kohonen self-organizing map referred to

- A) The process of finding the right formal representation of a certain body of knowledge in order to represent it in a knowledge-based system

- B) It automatically maps an external signal space into a system's internal representational space. They are useful in the performance of classification tasks
- C) A process where an individual learns how to carry out a certain task when making a transition from a situation in which the task cannot be carried out to a situation in which the same task under the same circumstances can be carried out.
- D) None of these

Answer: B

29) Incremental learning referred to

- A) Machine-learning involving different techniques
- B) The learning algorithmic analyzes the examples on a systematic basis and makes incremental adjustments to the theory that is learned
- C) Learning by generalizing from examples
- D) None of these

Answer: B

30) Knowledge engineering is

- A) The process of finding the right formal representation of a certain body of knowledge in order to represent it in a knowledge-based system
- B) It automatically maps an external signal space into a system's internal representational space. They are useful in the performance of classification tasks.
- C) A process where an individual learns how to carry out a certain task when making a transition from a situation in which the task cannot be carried out to a situation in which the same task under the same circumstances can be carried out.
- D) None of these

Answer: A

31) Information content is

- A) The amount of information with in data as opposed to the amount of redundancy or noise.
- B) One of the defining aspects of a data warehouse
- C) Restriction that requires data in one column of a database table to the a subset of another-column.
- D) None of these

Answer: A

32) Inductive learning is

- A) Machine-learning involving different techniques
- B) The learning algorithmic analyzes the examples on a systematic basis and makes incremental adjustments to the theory that is learned
- C) Learning by generalizing from examples
- D) None of these

Answer: C

33) Inclusion dependencies

- A) The amount of information with in data as opposed to the amount of redundancy or noise
- B) One of the defining aspects of a data warehouse
- C) Restriction that requires data in one column of a database table to the a subset of another-column
- D) None of these

Answer: C

34) KDD (Knowledge Discovery in Databases) is referred to

- A) Non-trivial extraction of implicit previously unknown and potentially useful information from data
- B) Set of columns in a database table that can be used to identify each record within this table uniquely.
- C) Collection of interesting and useful patterns in a database
- D) none of these

Answer: A

35) Learning is

- A) The process of finding the right formal representation of a certain body of knowledge in order to represent it in a knowledge-based system
- B) It automatically maps an external signal space into a system's internal representational space. They are useful in the performance of classification tasks.
- C) A process where an individual learns how to carry out a certain task when making a transition from a situation in which the task cannot be carried out to a situation in which the same task under the same circumstances can be carried out.

D) None of these

Answer: C

36) Naive prediction is

- A) A class of learning algorithms that try to derive a Prolog program from examples.
- B) A table with n independent attributes can be seen as an n -dimensional space.
- C) A prediction made using an extremely simple method, such as always predicting the same output.
- D) None of these

Answer: C

37) Learning algorithm referrs to

- A) An algorithm that can learn
- B) A sub-discipline of computer science that deals with the design and implementation of learning algorithms.
- C) A machine-learning approach that abstracts from the actual strategy of an individual algorithm and can therefore be applied to any other form of machine learning.
- D) None of these

Answer: A

38) Knowledge is referred to

- A) Non-trivial extraction of implicit previously unknown and potentially useful information from data
- B) Set of columns in a database table that can be used to identify each record within this table uniquely
- C) Collection of interesting and useful patterns in a database
- D) none of these

Answer: C

39) Node is

- A) A component of a network
- B) In the context of KDD and data mining, this refers to random errors in a database table.
- C) One of the defining aspects of a data warehouse
- D) None of these

Answer: A

40) Machine learning is

- A) An algorithm that can learn
- B) A sub-discipline of computer science that deals with the design and implementation of learning algorithms
- C) An approach that abstracts from the actual strategy of an individual algorithm and can therefore be applied to any other form of machine learning.
- D) None of these

Answer: B

41) Projection pursuit is

- A) The result of the application of a theory or a rule in a specific case
- B) One of several possible entries within a database table that is chosen by the designer as the primary means of accessing the data in the table.
- C) Discipline in statistics that studies ways to find the most interesting projections of multi-dimensional spaces
- D) None of these

Answer: C

42) Inductive logic programming is

- A) A class of learning algorithms that try to derive a Prolog program from examples
- B) A table with n independent attributes can be seen as an n-dimensional space
- C) A prediction made using an extremely simple method, such as always predicting the same output
- D) None of these

Answer: A

43) Statistical significance is

- A) The science of collecting, organizing, and applying numerical facts
- B) Measure of the probability that a certain hypothesis is incorrect given certain observations.

- C) One of the defining aspects of a data warehouse, which is specially built around all the existing applications of the operational data
- D) None of these

Answer: B

44) Multi-dimensional knowledge is

- A) A class of learning algorithms that try to derive a Prolog program from examples
- B) A table with n independent attributes can be seen as an n -dimensional space
- C) A prediction made using an extremely simple method, such as always predicting the same output.
- D) None of these

Answer: B

45) Prediction is

- A) The result of the application of a theory or a rule in a specific case
- B) One of several possible entries within a database table that is chosen by the designer as the primary means of accessing the data in the table.
- C) Discipline in statistics that studies ways to find the most interesting projections of multi-dimensional spaces.
- D) None of these

Answer: A

46) Query tools are

- A) A reference to the speed of an algorithm, which is quadratically dependent on the size of the data
- B) Attributes of a database table that can take only numerical values.
- C) Tools designed to query a database.
- D) None of these

Answer: C

47) Operational database is

- A) A measure of the desired maximal complexity of data mining algorithms

- B) A database containing volatile data used for the daily operation of an organization
- C) Relational database management system
- D) None of these

Answer: B

48) Which of the following is/are the Data mining tasks?

- (a) Regression
- (b) Classification
- (c) Clustering
- (d) inference of associative rules
- (e) All (a), (b), (c) and (d) above.

Answer: E

Explanation: Regression, Classification and Clustering are the data mining tasks.

49) In a data warehouse, if D1 and D2 are two conformed dimensions, then

- (a) D1 may be an exact replica of D2
- (b) D1 may be at a rolled up level of granularity compared to D2
- (c) Columns of D1 may be a subset of D2 and vice versa
- (d) Rows of D1 may be a subset of D2 and vice versa
- (e) All (a), (b), (c) and (d) above.

Answer: A

Explanation: In a data warehouse, if D1 and D2 are two conformed dimensions, then D1 may be an exact replica of D2.

50. Which of the following is not an ETL tool?

- (a) Informatica
- (b) Oracle warehouse builder
- (c) Datastage
- (d) Visual studio
- (e) DT/studio.

Answer: D

Explanation: Visual Studio is not an ETL tool.

51) is an essential process where intelligent methods are applied to extract data patterns.

- A) Data warehousing
- B) Data mining
- C) Text mining
- D) Data selection

Answer: B) Data mining

52) Data mining can also applied to other forms such as

- i) Data streams
- ii) Sequence data
- iii) Networked data
- iv) Text data
- v) Spatial data

- A) i, ii, iii and v only
- B) ii, iii, iv and v only
- C) i, iii, iv and v only
- D) All i, ii, iii, iv and v

Answer: D) All i, ii, iii, iv and v

53) Which of the following is not a data mining functionality?

- A) Characterization and Discrimination
- B) Classification and regression
- C) Selection and interpretation
- D) Clustering and Analysis

Answer: C) Selection and interpretation

54) is a summarization of the general characteristics or features of a target class of data.

- A) Data Characterization
- B) Data Classification
- C) Data discrimination
- D) Data selection

Answer: A) Data Characterization

55) is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes.

- A) Data Characterization
- B) Data Classification
- C) Data discrimination
- D) Data selection

Answer: C) Data discrimination

56) Strategic value of data mining is

- A) cost-sensitive
- B) work-sensitive
- C) time-sensitive
- D) technical-sensitive

Answer: C) time-sensitive

57) is the process of finding a model that describes and distinguishes data classes or concepts.

- A) Data Characterization
- B) Data Classification
- C) Data discrimination
- D) Data selection

Answer: B) Data Classification

58. The various aspects of data mining methodologies is/are

i) Mining various and new kinds of knowledge
ii) Mining knowledge in multidimensional space
iii) Pattern evaluation and pattern or constraint-guided mining.
iv) Handling uncertainty, noise, or incompleteness of data

- A) i, ii and iv only
- B) ii, iii and iv only
- C) i, ii and iii only
- D) All i, ii, iii and iv

Answer: D) All i, ii, iii and iv

59) The full form of KDD is

- A) Knowledge Database
- B) Knowledge Discovery Database

- C) Knowledge Data House
- D) Knowledge Data Definition

Answer: B) Knowledge Discovery Database

60) The out put of KDD is

- A) Data
- B) Information
- C) Query
- D) Useful information

Answer: D) Useful information

61. The full form of OLAP is

- A) Online Analytical Processing
- B) Online Advanced Processing
- C) Online Advanced Preparation
- D) Online Analytical Performance

Answer: A) Online Analytical Processing

62) is a subject-oriented, integrated, time-variant, nonvolatile collection or data in support of management decisions.

- A) Data Mining
- B) Data Warehousing
- C) Document Mining
- D) Text Mining

Answer: B) Data Warehousing

63) The data is stored, retrieved and updated in

- A) OLAP
- B) OLTP
- C) SMTP
- D) FTP

Answer: B) OLTP

64) An system is market-oriented and is used for data analysis by knowledge workers, including managers, executives, and analysts.

- A) OLAP
- B) OLTP
- C) Both of the above
- D) None of the above

Answer: A) OLAP

65) is a good alternative to the star schema.

- A) Star schema
- B) Snowflake schema
- C) Fact constellation
- D) Star-snowflake schema

Answer: C) Fact constellation

66) The exposes the information being captured, stored, and managed by operational systems.

- A) top-down view
- B) data warehouse view
- C) data source view
- D) business query view

Answer: C) data source view

67) The type of relationship in star schema is

- A) many to many
- B) one to one
- C) one to many
- D) many to one

Answer: C) one to many

68) The allows the selection of the relevant information necessary for the data warehouse.

- A) top-down view
- B) data warehouse view

- C) data source view
- D) business query view

Answer: A) top-down view

69) Which of the following is not a component of a data warehouse?

- A) Metadata
- B) Current detail data
- C) Lightly summarized data
- D) Component Key

Answer: D) Component Key

70) Which of the following is not a kind of data warehouse application?

- A) Information processing
- B) Analytical processing
- C) Data mining
- D) Transaction processing

Answer: D) Transaction processing

71) Data warehouse architecture is based on

- A) DBMS
- B) RDBMS
- C) Sybase
- D) SQL Server

Answer:B) RDBMS

72) supports basic OLAP operations, including slice and dice, drill-down, roll-up and pivoting.

- A) Information processing
- B) Analytical processing
- C) Data mining
- D) Transaction processing

Answer: B) Analytical processing

73) The core of the multidimensional model is the , which consists of a large set of facts and a number of dimensions.

- A) Multidimensional cube
- B) Dimensions cube
- C) Data cube
- D) Data model

Answer: C) Data cube

74) The data from the operational environment enter of data warehouse.

- A) Current detail data
- B) Older detail data
- C) Lightly Summarized data
- D) Highly summarized data

Answer: A) Current detail data

75) A data warehouse is

- A) updated by end users.
- B) contains numerous naming conventions and formats
- C) organized around important subject areas
- D) contain only current data

Answer: C) organized around important subject areas

76) Business Intelligence and data warehousing is used for

- A) Forecasting
- B) Data Mining
- C) Analysis of large volumes of product sales data
- D) All of the above

Answer: D) All of the above

77) Data warehouse contains data that is never found in the operational environment.

- A) normalized
- B) informational
- C) summary
- D) denormalized

Answer: C) summary

78) are responsible for running queries and reports against data warehouse tables.

- A) Hardware
- B) Software
- C) End users
- D) Middle ware

Answer: C) End users

79) The biggest drawback of the level indicator in the classic star schema is that is limits

- A) flexibility
- B) quantify
- C) qualify
- D) ability

Answer: A) flexibility

80) are designed to overcome any limitations placed on the warehouse by the nature of the relational data model.

- A) Operational database
- B) Relational database
- C) Multidimensional database
- D) Data repository

Answer: C) Multidimensional database

81) Which of the following is the most important when deciding on the data structure of a data mart?

- (a) XML data exchange standards
- (b) Data access tools to be used
- (c) Metadata naming conventions
- (d) Extract, Transform, and Load (ETL) tool to be used
- (e) All (a), (b), (c) and (d) above.

Answer: B

Explanation: Data access tools to be used when deciding on the data structure of a data mart.

82) The process of removing the deficiencies and loopholes in the data is called as

- (a) Aggregation of data
- (b) Extracting of data
- (c) Cleaning up of data.
- (d) Loading of data
- (e) Compression of data.

Answer: C

Explanation: The process of removing the deficiencies and loopholes in the data is called as cleaning up of data.

83) Which one manages both current and historic transactions?

- (a) OLTP
- (b) OLAP
- (c) Spread sheet
- (d) XML
- (e) All (a), (b), (c) and (d) above.

Answer: B

Explanation: Online Analytical Processing (OLAP) manages both current and historic transactions.

84) Which of the following is the collection of data objects that are similar to one another within the same group?

- (a) Partitioning
- (b) Grid
- (c) Cluster
- (d) Table
- (e) Data source.

Answer: C

Explanation: Cluster is the collection of data objects that are similar to one another within the same group.

85) Which of the following employs data mining techniques to analyze the intent of a user query, provided additional generalized or associated information relevant to the query?

- (a) Iceberg query method
- (b) Data analyzer
- (c) Intelligent query answering
- (d) DBA
- (e) Query parser.

Answer: C

Explanation: Intelligent Query Answering employee's data mining techniques to analyze the intent of a user query provided additional generalized or associated information relevant to the query.

86) Which of the following process includes data cleaning, data integration, data selection, data transformation, data mining, pattern evolution and knowledge presentation?

- (a) KDD process
- (b) ETL process
- (c) KTL process
- (d) MDX process
- (e) None of the above.

Answer: A

Explanation: KDD Process includes data cleaning, data integration, data selection, data transformation, data mining, pattern evolution, and knowledge presentation.

87. At which level we can create dimensional models?

- (a) Business requirements level
- (b) Architecture models level
- (c) Detailed models level
- (d) Implementation level
- (e) Testing level.

Answer: B

Explanation: Dimensional models can be created at Architecture models level.

88) Which of the following is not related to dimension table attributes?

- (a) Verbose
- (b) Descriptive
- (c) Equally unavailable
- (d) Complete
- (e) Indexed.

Answer: C

Explanation: Equally unavailable is not related to dimension table attributes.

89) Data warehouse bus matrix is a combination of

- (a) Dimensions and data marts
- (b) Dimensions and facts
- (c) Facts and data marts
- (d) Dimensions and detailed facts
- (e) All (a), (b), (c) and (d) above.

Answer: A

Explanation: Data warehouse bus matrix is a combination of Dimensions and data marts.

90) Which of the following is not the managing issue in the modeling process?

- (a) Content of primary units column
- (b) Document each candidate data source
- (c) Do regions report to zones
- (d) Walk through business scenarios
- (e) Ensure that the transaction edit flat is used for analysis.

Answer: E

Explanation: Ensure that the transaction edit flat is used for analysis is not the managing issue in the modeling process.

91) Data modeling technique used for data marts is

- (a) Dimensional modeling
- (b) ER – model
- (c) Extended ER – model
- (d) Physical model
- (e) Logical model.

Answer: A

Explanation: Data modeling technique used for data marts is Dimensional modeling.

92) A warehouse architect is trying to determine what data must be included in the warehouse. A meeting has been arranged with a business analyst to understand the data requirements, which of the following should be included in the agenda?

- (a) Number of users
- (b) Corporate objectives
- (c) Database design
- (d) Routine reporting
- (e) Budget.

Answer: D

Explanation: Routine reporting should be included in the agenda.

93. An OLAP tool provides for

- (a) Multidimensional analysis
- (b) Roll-up and drill-down
- (c) Slicing and dicing
- (d) Rotation
- (e) Setting up only relations.

Answer: C

Explanation: An OLAP tool provides for Slicing and dicing.

94. The Synonym for data mining is

- (a) Data warehouse
- (b) Knowledge discovery in database
- (c) ETL
- (d) Business intelligence
- (e) OLAP.

Answer: C

Explanation: The synonym for data mining is Knowledge discovery in Database.

95) Which of the following statements is true?

- (a) A fact table describes the transactions stored in a DWH
- (b) A fact table describes the granularity of data held in a DWH
- (c) The fact table of a data warehouse is the main store of descriptions of the transactions stored in a DWH
- (d) The fact table of a data warehouse is the main store of all of the recorded transactions over time
- (e) A fact table maintains the old records of the database.

Answer: D

Explanation: The fact table of a data warehouse is the main store of all of the recorded transactions over time is the correct statement.

96) Most common kind of queries in a data warehouse

- (a) Inside-out queries

- (b) Outside-in queries
- (c) Browse queries
- (d) Range queries
- (e) All (a), (b), (c) and (d) above.

Answer: A

Explanation: The Most common kind of queries in a data warehouse is Inside-out queries.

97) Concept description is the basic form of the

- (a) Predictive data mining
- (b) Descriptive data mining
- (c) Data warehouse
- (d) Relational data base
- (e) Proactive data mining.

Answer: B

Explanation: Concept description is the basis form of the descriptive data mining.

98) The apriori property means

- (a) If a set cannot pass a test, all of its supersets will fail the same test as well
- (b) To improve the efficiency the level-wise generation of frequent item sets
- (c) If a set can pass a test, all of its supersets will fail the same test as well
- (d) To decrease the efficiency the level-wise generation of frequent item sets
- (e) All (a), (b), (c) and (d) above.

Answer: B

Explanation: The apriori property means to improve the efficiency the level-wise generation of frequent item sets.

99) Which of following form the set of data created to support a specific short lived business situation?

- (a) Personal data marts
- (b) Application models
- (c) Downstream systems
- (d) Disposable data marts
- (e) Data mining models.

Answer: D

Explanation: Disposable Data Marts is the form the set of data created to support a specific short lived business situation.

100) What is/are the different types of Meta data?

I. Administrative.

II. Business.

III. Operational.

(a) Only (I) above

(b) Both (II) and (III) above

(c) Both (I) and (II) above

(d) Both (I) and (III) above

(e) All (I), (II) and (III) above.

Answer: E

Explanation: The different types of Meta data are Administrative, Business and Operational.

101) Multiple Regression means

(a) Data are modeled using a straight line

(b) Data are modeled using a curve line

(c) Extension of linear regression involving only one predictor value

(d) Extension of linear regression involving more than one predictor value

(e) All (a), (b), (c) and (d) above.

Answer: D

Explanation: Multiple Regression means extension of linear regression involving more than one predictor value.

102) Which of the following should not be considered for each dimension attribute?

(a) Attribute name

(b) Rapid changing dimension policy

(c) Attribute definition

(d) Sample data

(e) Cardinality.

Answer: B

Explanation: Rapid changing dimension policy should not be considered for each dimension attribute.

103) A Business Intelligence system requires data from:

- (a) Data warehouse
- (b) Operational systems
- (c) All possible sources within the organization and possibly from external sources
- (d) Web servers
- (e) Database servers.

Answer: A

Explanation: A business Intelligence system requires data from Data warehouse

104) Data mining application domains are

- (a) Biomedical
- (b) DNA data analysis
- (c) Financial data analysis
- (d) Retail industry and telecommunication industry
- (e) All (a), (b), (c) and (d) above.

Answer: E

Explanation: Data mining application domains are Biomedical, DNA data analysis, Financial data analysis and Retail industry and telecommunication industry

105. The generalization of multidimensional attributes of a complex object class can be performed by examining each attribute, generalizing each attribute to simple-value data and constructing a multidimensional data cube is called as

- (a) Object cube
- (b) Relational cube
- (c) Transactional cube
- (d) Tuple
- (e) Attribute.

Answer: A

Explanation: The generalization of multidimensional attributes of a complex object class can be performed by examining each attribute, generalizing each attribute to simple-value data and constructing a multidimensional data cube is called as object cube.

106. Which of the following project is a building a data mart for a business process/department that is very critical for your organization?

- (a) High risk high reward
- (b) High risk low reward
- (c) Low risk low reward
- (d) Low risk high reward
- (e) Involves high risks.

Answer: A

Explanation: High risk high reward project is a building a data mart for a business process/department that is very critical for your organization

107. Which of the following tools a business intelligence system will have?

- (a) OLAP tool
- (b) Data mining tool
- (c) Reporting tool
- (d) Both(a) and (b) above
- (e) (a), (b) and (c) above.

Answer: A

Explanation: Business intelligence system will have OLAP, Data mining and reporting tolls.

108. A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college.

1) Which of the following statement is true in following case?

- A) Feature F1 is an example of nominal variable.
- B) Feature F1 is an example of ordinal variable.
- C) It doesn't belong to any of the above category.
- D) Both of these

Solution: (B)

Ordinal variables are the variables which has some order in their categories. For example, grade A should be consider as high grade than grade B.

2) Which of the following is an example of a deterministic algorithm?

- A) PCA

- B) K-Means
- C) None of the above

Solution: (A)

A deterministic algorithm is that in which output does not change on different runs. PCA would give the same result if we run again, but not k-means.

3) [True or False] A Pearson correlation between two variables is zero but, still their values can still be related to each other.

- A) TRUE
- B) FALSE

Solution: (A)

$Y=X^2$. Note that, they are not only associated, but one is a function of the other and Pearson correlation between them is 0.

4) Which of the following statement(s) is / are true for Gradient Decent (GD) and Stochastic Gradient Decent (SGD)?

1. In GD and SGD, you update a set of parameters in an iterative manner to minimize the error function.
2. In SGD, you have to run through all the samples in your training set for a single update of a parameter in each iteration.
3. In GD, you either use the entire data or a subset of training data to update a parameter in each iteration.

- A) Only 1
- B) Only 2
- C) Only 3

- D) 1 and 2
- E) 2 and 3
- F) 1,2 and 3

Solution: (A)

In SGD for each iteration you choose the batch which is generally contain the random sample of data But in case of GD each iteration contain the all of the training observations.

5) Which of the following hyper parameter(s), when increased may cause random forest to over fit the data?

- 1. Number of Trees
 - 2. Depth of Tree
 - 3. Learning Rate
- A) Only 1
 - B) Only 2
 - C) Only 3
 - D) 1 and 2
 - E) 2 and 3
 - F) 1,2 and 3

Solution: (B)

Usually, if we increase the depth of tree it will cause overfitting. Learning rate is not an hyperparameter in random forest. Increase in the number of tree will cause under fitting.

6) Imagine, you are working with “Analytics Vidhya” and you want to develop a machine learning algorithm which predicts the number of views on the articles.

Your analysis is based on features like author name, number of articles written by the same author on Analytics Vidhya in past and a few other features. Which of the following evaluation metric would you choose in that case?

1. Mean Square Error

2. Accuracy

3. F1 Score

A) Only 1

B) Only 2

C) Only 3

D) 1 and 3

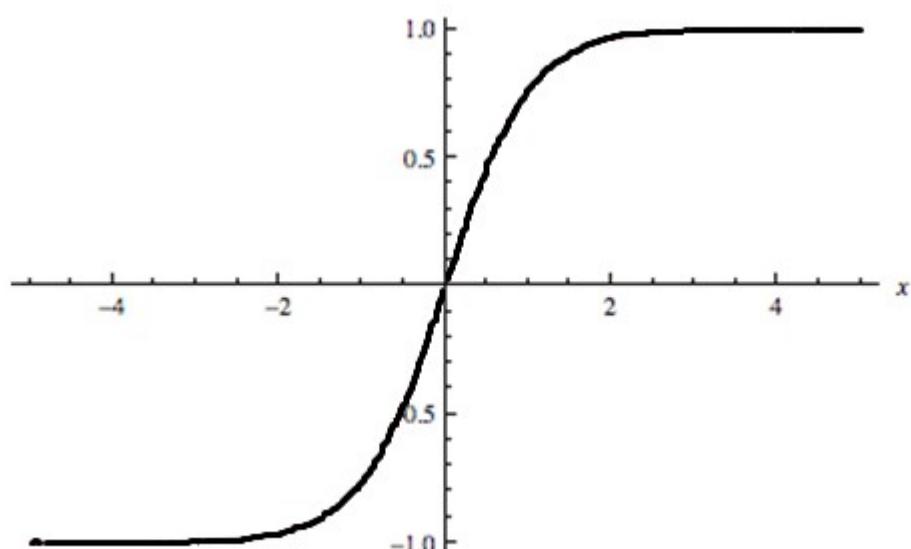
E) 2 and 3

F) 1 and 2

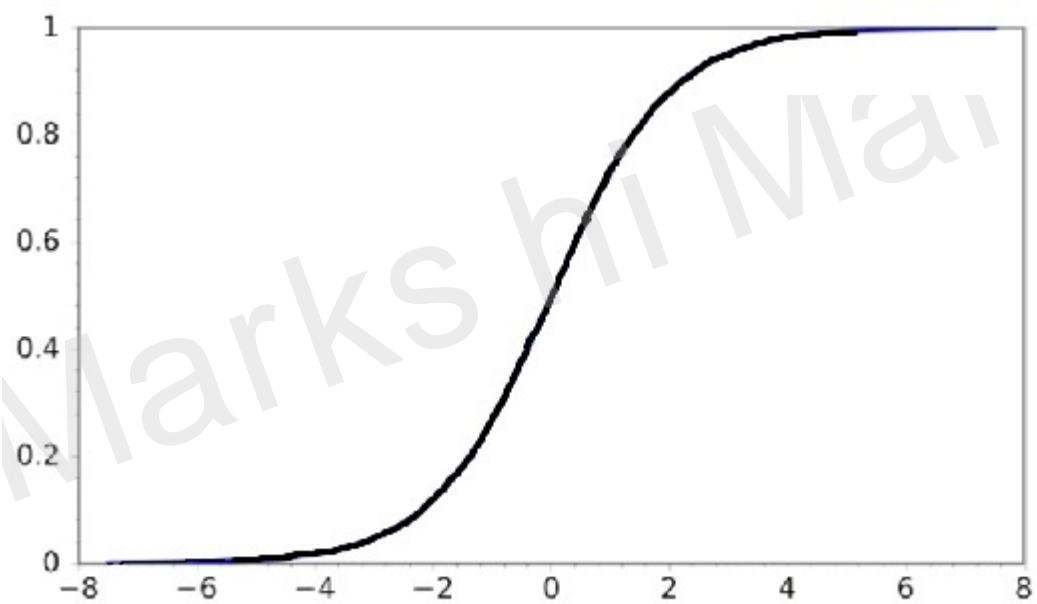
Solution: (A)

You can think that the number of views of articles is the continuous target variable which fall under the regression problem. So, mean squared error will be used as an evaluation metrics.

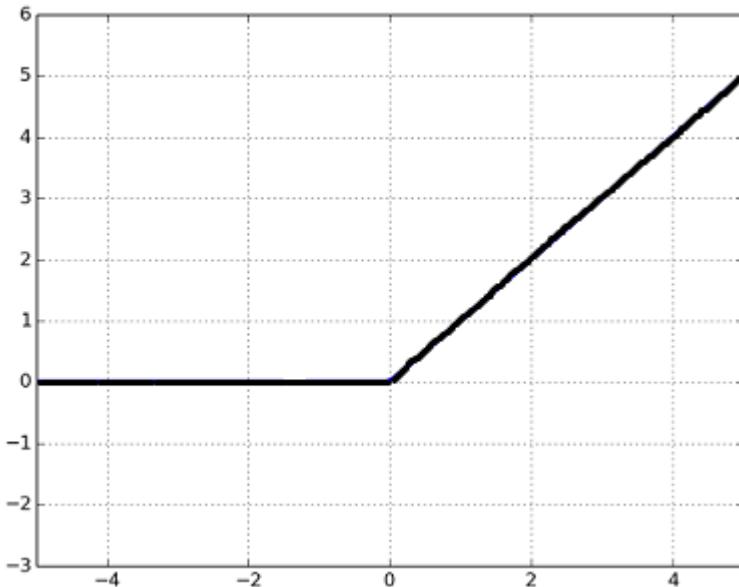
7) Given below are three images (1,2,3). Which of the following option is correct for these images?



A)



B)



C)

- A) 1 is tanh, 2 is ReLU and 3 is SIGMOID activation functions.
- B) 1 is SIGMOID, 2 is ReLU and 3 is tanh activation functions.
- C) 1 is ReLU, 2 is tanh and 3 is SIGMOID activation functions.
- D) 1 is tanh, 2 is SIGMOID and 3 is ReLU activation functions.

Solution: (D)

The range of SIGMOID function is $[0,1]$.

The range of the tanh function is $[-1,1]$.

The range of the RELU function is $[0, \text{infinity}]$.

So Option D is the right answer.

8) Below are the 8 actual values of target variable in the train file.

[0,0,0,1,1,1,1,1]

What is the entropy of the target variable?

A) $-(5/8 \log(5/8) + 3/8 \log(3/8))$

B) $5/8 \log(5/8) + 3/8 \log(3/8)$

C) $\frac{3}{8} \log(\frac{5}{8}) + \frac{5}{8} \log(\frac{3}{8})$

D) $\frac{5}{8} \log(\frac{3}{8}) - \frac{3}{8} \log(\frac{5}{8})$

Solution: (A)

The formula for entropy is $-\sum p(x) * \log p(x)$

So the answer is A.

9) Let's say, you are working with categorical feature(s) and you have not looked at the distribution of the categorical variable in the test data.

You want to apply one hot encoding (OHE) on the categorical feature(s). What challenges you may face if you have applied OHE on a categorical variable of train dataset?

A) All categories of categorical variable are not present in the test dataset.

B) Frequency distribution of categories is different in train as compared to the test dataset.

C) Train and Test always have same distribution.

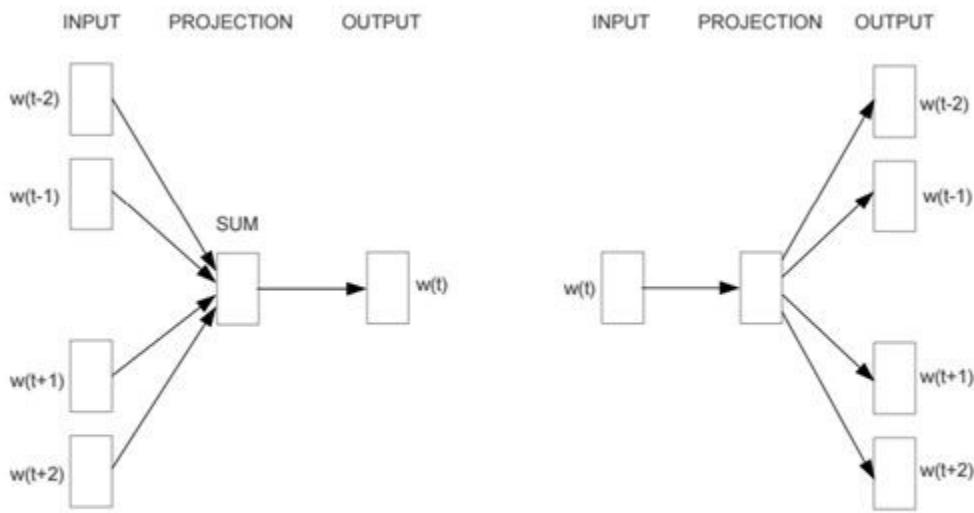
D) Both A and B

E) None of these

Solution: (D)

Both are true, The OHE will fail to encode the categories which is present in test but not in train so it could be one of the main challenges while applying OHE. The challenge given in option B is also true you need to more careful while applying OHE if frequency distribution doesn't same in train and test.

10) Skip gram model is one of the best models used in Word2vec algorithm for words embedding. Which one of the following models depict the skip gram model?



Model A

Model B

- A) A
- B) B
- C) Both A and B
- D) None of these

Solution: (B)

Both models (model1 and model2) are used in Word2vec algorithm. The model1 represent a CBOW model whereas Model2 represent the Skip gram model.

11) Let's say, you are using activation function X in hidden layers of neural network. At a particular neuron for any given input, you get the output as "-0.0001". Which of the following activation function could X represent?

- A) ReLU
- B) tanh
- C) SIGMOID

D) None of these

Solution: (B)

The function is a tanh because the this function output range is between (-1, 1).

12) [True or False] LogLoss evaluation metric can have negative values.

- A) TRUE
- B) FALSE

Solution: (B)

Log loss cannot have negative values.

13) Which of the following statements is/are true about "Type-1" and "Type-2" errors?

1. Type1 is known as false positive and Type2 is known as false negative.
2. Type1 is known as false negative and Type2 is known as false positive.
3. Type1 error occurs when we reject a null hypothesis when it is actually true.

A) Only 1

B) Only 2

C) Only 3

D) 1 and 2

E) 1 and 3

F) 2 and 3

Solution: (E)

In statistical hypothesis testing, a type I error is the incorrect rejection of a true null hypothesis (a "false positive"), while a type II error is incorrectly retaining a false null hypothesis (a "false negative").

14) Which of the following is/are one of the important step(s) to pre-process the text in NLP based projects?

1. **Stemming**
 2. **Stop word removal**
 3. **Object Standardization**
- A) 1 and 2
B) 1 and 3
C) 2 and 3
D) 1,2 and 3

Solution: (D)

Stemming is a rudimentary rule-based process of stripping the suffixes ("ing", "ly", "es", "s" etc) from a word.

Stop words are those words which will have not relevant to the context of the data for example is/am/are.

Object Standardization is also one of the good way to pre-process the text.

15) Suppose you want to project high dimensional data into lower dimensions. The two most famous dimensionality reduction algorithms used here are PCA and t-SNE. Let's say you have applied both algorithms respectively on data "X" and you got the datasets "X_projected_PCA" , "X_projected_tSNE".

Which of the following statements is true for "X_projected_PCA" & "X_projected_tSNE" ?

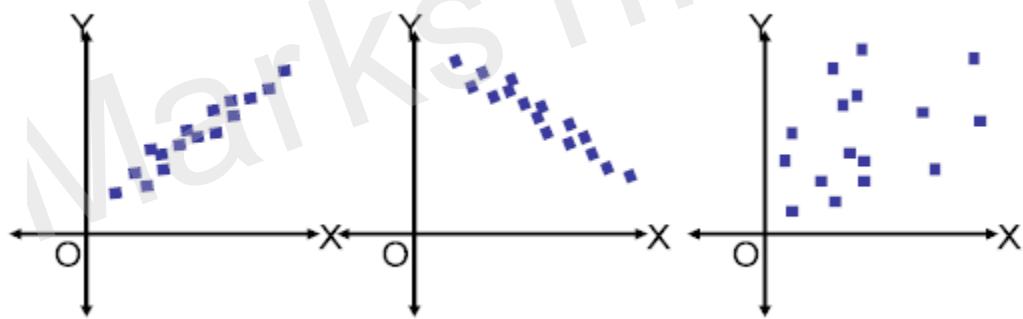
- A) X_projected_PCA will have interpretation in the nearest neighbour space.
- B) X_projected_tSNE will have interpretation in the nearest neighbour space.
- C) Both will have interpretation in the nearest neighbour space.
- D) None of them will have interpretation in the nearest neighbour space.

Solution: (B)

t-SNE algorithm consider nearest neighbour points to reduce the dimensionality of the data. So, after using t-SNE we can think that reduced dimensions will also have interpretation in nearest neighbour space. But in case of PCA it is not the case.

Context: 16-17

Given below are three scatter plots for two features (Image 1, 2 & 3 from left to right).



16) In the above images, which of the following is/are example of multi-collinear features?

- A) Features in Image 1
- B) Features in Image 2
- C) Features in Image 3
- D) Features in Image 1 & 2

E) Features in Image 2 & 3

F) Features in Image 3 & 1

Solution: (D)

In Image 1, features have high positive correlation where as in Image 2 has high negative correlation between the features so in both images pair of features are the example of multicollinear features.

17) In previous question, suppose you have identified multi-collinear features. Which of the following action(s) would you perform next?

1. Remove both collinear variables.
2. Instead of removing both variables, we can remove only one variable.
3. Removing correlated variables might lead to loss of information. In order to retain those variables, we can use penalized regression models like ridge or lasso regression.

A) Only 1

B) Only 2

C) Only 3

D) Either 1 or 3

E) Either 2 or 3

Solution: (E)

You cannot remove the both features because after removing the both features you will lose all of the information so you should either remove the only 1 feature or you can use the regularization algorithm like L1 and L2.

18) Adding a non-important feature to a linear regression model may result in.

1. Increase in R-square
2. Decrease in R-square

- A) Only 1 is correct
- B) Only 2 is correct
- C) Either 1 or 2
- D) None of these

Solution: (A)

After adding a feature in feature space, whether that feature is important or unimportant features the R-squared always increase.

19) Suppose, you are given three variables X, Y and Z. The Pearson correlation coefficients for (X, Y), (Y, Z) and (X, Z) are C1, C2 & C3 respectively.

Now, you have added 2 in all values of X (i.e. new values become $X+2$), subtracted 2 from all values of Y (i.e. new values are $Y-2$) and Z remains the same. The new coefficients for (X,Y), (Y,Z) and (X,Z) are given by D1, D2 & D3 respectively. How do the values of D1, D2 & D3 relate to C1, C2 & C3?

- A) D1= C1, D2 < C2, D3 > C3
- B) D1 = C1, D2 > C2, D3 > C3
- C) D1 = C1, D2 > C2, D3 < C3
- D) D1 = C1, D2 < C2, D3 < C3
- E) D1 = C1, D2 = C2, D3 = C3
- F) Cannot be determined

Solution: (E)

Correlation between the features won't change if you add or subtract a value in the features.

20) Imagine, you are solving a classification problems with highly imbalanced class. The majority class is observed 99% of times in the training data.

Your model has 99% accuracy after taking the predictions on test data. Which of the following is true in such a case?

1. Accuracy metric is not a good idea for imbalanced class problems.
 2. Accuracy metric is a good idea for imbalanced class problems.
 3. Precision and recall metrics are good for imbalanced class problems.
 4. Precision and recall metrics aren't good for imbalanced class problems.
- A) 1 and 3
- B) 1 and 4
- C) 2 and 3
- D) 2 and 4

Solution: (A)

Refer the question number 4 from in [this](#) article.

21) In ensemble learning, you aggregate the predictions for weak learners, so that an ensemble of these models will give a better prediction than prediction of individual models.

Which of the following statements is / are true for weak learners used in ensemble model?

1. They don't usually overfit.

2. They have high bias, so they cannot solve complex learning problems
 3. They usually overfit.
- A) 1 and 2
- B) 1 and 3
- C) 2 and 3
- D) Only 1
- E) Only 2
- F) None of the above

Solution: (A)

Weak learners are sure about particular part of a problem. So, they usually don't overfit which means that weak learners have low variance and high bias.

22) Which of the following options is/are true for K-fold cross-validation?

1. Increase in K will result in higher time required to cross validate the result.
 2. Higher values of K will result in higher confidence on the cross-validation result as compared to lower value of K.
 3. If $K=N$, then it is called Leave one out cross validation, where N is the number of observations.
- A) 1 and 2
- B) 2 and 3
- C) 1 and 3
- D) 1, 2 and 3

Solution: (D)

Larger k value means less bias towards overestimating the true expected error (as training folds will be closer to the total dataset) and higher running time (as you are getting closer to the limit case: Leave-One-Out CV). We also need to consider the variance between the k folds accuracy while selecting the k.

Question Context 23-24

Cross-validation is an important step in machine learning for hyper parameter tuning. Let's say you are tuning a hyper-parameter "max_depth" for GBM by selecting it from 10 different depth values (values are greater than 2) for tree based model using 5-fold cross validation.

Time taken by an algorithm for training (on a model with max_depth 2) 4-fold is 10 seconds and for the prediction on remaining 1-fold is 2 seconds.

Note: Ignore hardware dependencies from the equation.

23) Which of the following option is true for overall execution time for 5-fold cross validation with 10 different values of "max_depth"?

- A) Less than 100 seconds
- B) 100 - 300 seconds
- C) 300 - 600 seconds
- D) More than or equal to 600 seconds
- C) None of the above
- D) Can't estimate

Solution: (D)

Each iteration for depth "2" in 5-fold cross validation will take 10 secs for training and 2 second for testing. So, 5 folds will take $12 \times 5 = 60$ seconds. Since we are searching over the 10 depth values so the algorithm would take $60 \times 10 = 600$ seconds. But training and testing a model on depth greater than 2 will

take more time than depth "2" so overall timing would be greater than 600.

24) In previous question, if you train the same algorithm for tuning 2 hyper parameters say "max_depth" and "learning_rate".

You want to select the right value against "max_depth" (from given 10 depth values) and learning rate (from given 5 different learning rates). In such cases, which of the following will represent the overall time?

- A) 1000-1500 second
- B) 1500-3000 Second
- C) More than or equal to 3000 Second
- D) None of these

Solution: (D)

Same as question number 23.

25) Given below is a scenario for training error TE and Validation error VE for a machine learning algorithm M1. You want to choose a hyperparameter (H) based on TE and VE.

H	TE	VE
1	105	90
2	200	85
3	250	96
4	105	85
5	300	100

Which value of H will you choose based on the above table?

- A) 1

- B) 2
- C) 3
- D) 4
- E) 5

Solution: (D)

Looking at the table, option D seems the best

26) What would you do in PCA to get the same projection as SVD?

- A) Transform data to zero mean
- B) Transform data to zero median
- C) Not possible
- D) None of these

Solution: (A)

When the data has a zero mean vector PCA will have same projections as SVD, otherwise you have to centre the data first before taking SVD.

Question Context 27-28

Assume there is a black box algorithm, which takes training data with multiple observations ($t_1, t_2, t_3, \dots, t_n$) and a new observation (q_1). The black box outputs the nearest neighbor of q_1 (say t_i) and its corresponding class label c_i .

You can also think that this black box algorithm is same as 1-NN (1-nearest neighbor).

27) It is possible to construct a k-NN classification algorithm based on this black box alone.

Note: Where n (number of training observations) is very large compared to k .

A) TRUE

B) FALSE

Solution: (A)

In first step, you pass an observation (q_1) in the black box algorithm so this algorithm would return a nearest observation and its class.

In second step, you pass it through the nearest observation from train data and again input the observation (q_1). The black box algorithm will again return the a nearest observation and its class.

You need to repeat this procedure k times

28) Instead of using 1-NN black box we want to use the j -NN ($j > 1$) algorithm as black box. Which of the following option is correct for finding k -NN using j -NN?

1. **J must be a proper factor of k**
2. **$J > k$**
3. **Not possible**

A) 1

B) 2

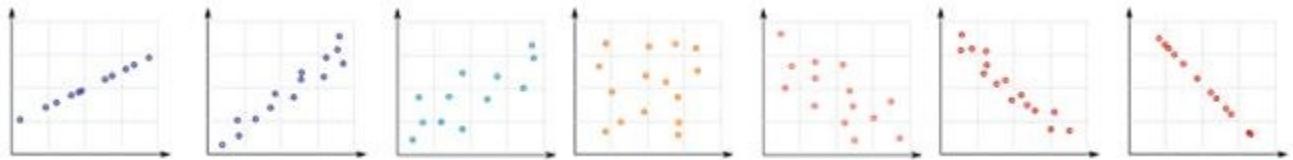
C) 3

Solution: (A)

Same as question number 27

29) Suppose you are given 7 Scatter plots 1-7 (left to right) and you want to compare Pearson correlation coefficients between variables of each scatterplot.

Which of the following is in the right order?



1. 1<2<3<4
 2. 1>2>3 > 4
 3. 7<6<5<4
 4. 7>6>5>4
- A) 1 and 3
 B) 2 and 3
 C) 1 and 4
 D) 2 and 4

Solution: (B)

from image 1to 4 correlation is decreasing (absolute value). But from image 4 to 7 correlation is increasing but values are negative (for example, 0, -0.3, -0.7, -0.99).

30) You can evaluate the performance of a binary class classification problem using different metrics such as accuracy, log-loss, F-Score. Let's say, you are using the log-loss function as evaluation metric.

Which of the following option is / are true for interpretation of log-loss as an evaluation metric?

$$\text{logLoss} = \frac{-1}{N} \sum_{i=1}^N (y_i(\log p_i) + (1 - y_i)\log(1 - p_i))$$

1. If a classifier is confident about an incorrect classification, then log-loss will penalise it heavily.
2. For a particular observation, the classifier assigns a very small probability for the correct class then the

corresponding contribution to the log-loss will be very large.

3. Lower the log-loss, the better is the model.

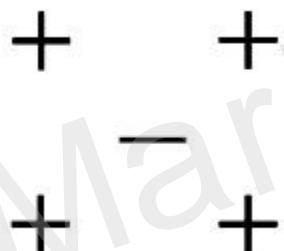
- A) 1 and 3
- B) 2 and 3
- C) 1 and 2
- D) 1,2 and 3

Solution: (D)

Options are self-explanatory.

Question 31-32

Below are five samples given in the dataset.



Note: Visual distance between the points in the image represents the actual distance.

31) Which of the following is leave-one-out cross-validation accuracy for 3-NN (3-nearest neighbor)?

- A) 0
- D) 0.4
- C) 0.8

D) 1

Solution: (C)

In Leave-One-Out cross validation, we will select $(n-1)$ observations for training and 1 observation of validation. Consider each point as a cross validation point and then find the 3 nearest point to this point. So if you repeat this procedure for all points you will get the correct classification for all positive class given in the above figure but negative class will be misclassified. Hence you will get 80% accuracy.

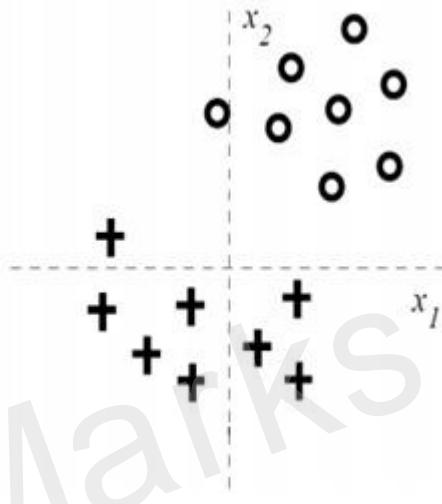
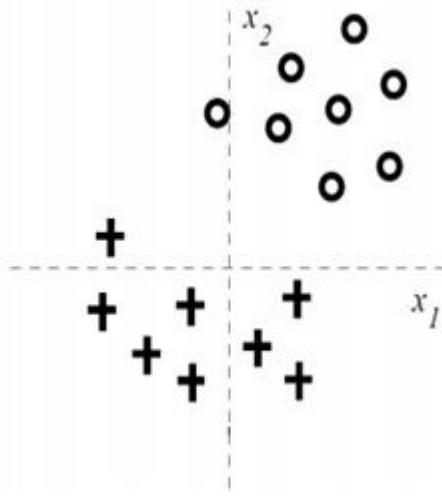
32) Which of the following value of K will have least leave-one-out cross validation accuracy?

- A) 1NN
- B) 3NN
- C) 4NN
- D) All have same leave one out error

Solution: (A)

Each point which will always be misclassified in 1-NN which means that you will get the 0% accuracy.

33) Suppose you are given the below data and you want to apply a logistic regression model for classifying it in two given classes.



You are using logistic regression with L1 regularization.

$$\sum_{i=1}^n \log P(y_i|x_i, w_0, w_1, w_2) - C(|w_1| + |w_2|).$$

Where C is the regularization parameter and w1 & w2 are the coefficients of x1 and x2.

Which of the following option is correct when you increase the value of C from zero to a very large value?

- A) First w2 becomes zero and then w1 becomes zero

- B) First w_1 becomes zero and then w_2 becomes zero
- C) Both becomes zero at the same time
- D) Both cannot be zero even after very large value of C

Solution: (B)

By looking at the image, we see that even on just using x_2 , we can efficiently perform classification. So at first w_1 will become 0. As regularization parameter increases more, w_2 will come more and more closer to 0.

34) Suppose we have a dataset which can be trained with 100% accuracy with help of a decision tree of depth 6. Now consider the points below and choose the option based on these points.

Note: All other hyper parameters are same and other factors are not affected.

- 1. Depth 4 will have high bias and low variance
- 2. Depth 4 will have low bias and low variance

- A) Only 1
- B) Only 2
- C) Both 1 and 2
- D) None of the above

Solution: (A)

If you fit decision tree of depth 4 in such data means it will more likely to underfit the data. So, in case of underfitting you will have high bias and low variance.

35) Which of the following options can be used to get global minima in k-Means Algorithm?

- 1. Try to run algorithm for different centroid initialization
- 2. Adjust number of iterations

3. Find out the optimal number of clusters

- A) 2 and 3
- B) 1 and 3
- C) 1 and 2
- D) All of above

Solution: (D)

All of the option can be tuned to find the global minima.

36) Imagine you are working on a project which is a binary classification problem. You trained a model on training dataset and get the below confusion matrix on validation dataset.

		Predicted: NO	Predicted: YES	
n=165				
Actual: NO	50	10		
Actual: YES	5	100		
n=165				

Based on the above confusion matrix, choose which option(s) below will give you correct predictions?

- 1. Accuracy is ~0.91
- 2. Misclassification rate is ~ 0.91
- 3. False positive rate is ~0.95
- 4. True positive rate is ~0.95

- A) 1 and 3
- B) 2 and 4

C) 1 and 4

D) 2 and 3

Solution: (C)

The Accuracy (correct classification) is $(50+100)/165$ which is nearly equal to 0.91.

The true Positive Rate is how many times you are predicting positive class correctly so true positive rate would be $100/105 = 0.95$ also known as "Sensitivity" or "Recall"

37) For which of the following hyperparameters, higher value is better for decision tree algorithm?

1. Number of samples used for split

2. Depth of tree

3. Samples for leaf

A) 1 and 2

B) 2 and 3

C) 1 and 3

D) 1, 2 and 3

E) Can't say

Solution: (E)

For all three options A, B and C, it is not necessary that if you increase the value of parameter the performance may increase. For example, if we have a very high value of depth of tree, the resulting tree may overfit the data, and would not generalize well. On the other hand, if we have a very low value, the tree may underfit the data. So, we can't say for sure that "higher is better".

Context 38-39

Imagine, you have a 28 * 28 image and you run a 3 * 3 convolution neural network on it with the input depth of 3 and output depth of 8.

Note: Stride is 1 and you are using same padding.

38) What is the dimension of output feature map when you are using the given parameters.

- A) 28 width, 28 height and 8 depth
- B) 13 width, 13 height and 8 depth
- C) 28 width, 13 height and 8 depth
- D) 13 width, 28 height and 8 depth

Solution: (A)

The formula for calculating output size is

$$\text{output size} = (N - F)/S + 1$$

where, N is input size, F is filter size and S is stride.

Read this [article](#) to get a better understanding.

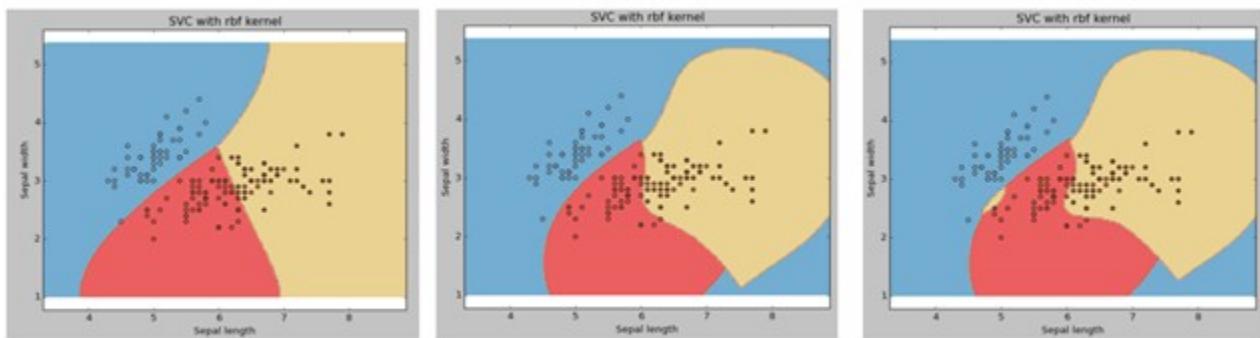
39) What is the dimensions of output feature map when you are using following parameters.

- A) 28 width, 28 height and 8 depth
- B) 13 width, 13 height and 8 depth
- C) 28 width, 13 height and 8 depth
- D) 13 width, 28 height and 8 depth

Solution: (B)

Same as above

40) Suppose, we were plotting the visualization for different values of C (Penalty parameter) in SVM algorithm. Due to some reason, we forgot to tag the C values with visualizations. In that case, which of the following option best explains the C values for the images below (1,2,3 left to right, so C values are C1 for image1, C2 for image2 and C3 for image3) in case of rbf kernel.



- A) $C_1 = C_2 = C_3$
- B) $C_1 > C_2 > C_3$
- C) $C_1 < C_2 < C_3$
- D) None of these

Solution: (C)

Penalty parameter C of the error term. It also controls the trade-off between smooth decision boundary and classifying the training points correctly. For large values of C, the optimization will choose a smaller-margin hyperplane.

1. Which of the following is a widely used and effective machine learning algorithm based on the idea of bagging?

- A. Decision Tree
- B. Regression
- C. Classification
- D. Random Forest

ANSWER: D

2. The most widely used metrics and tools to assess a classification model is:

- A. Confusion matrix
- B. Cost-sensitive accuracy
- C. Area under the ROC curve
- D. All of these

ANSWER: D

3. Which of the following is a good test dataset characteristic?

- A. Large enough to yield meaningful results
- B. Is representative of the dataset as a whole
- C. Both A and B
- D. None of these

ANSWER: C

4. How do you handle missing or corrupted data in a dataset?

- A. Drop missing rows or columns
- B. Replace missing values with mean/median/mode
- C. Assign a unique category to missing values
- D. All of these

ANSWER: D

5. What is the purpose of performing cross-validation?

- A. To assess the predictive performance of the models
- B. To judge how the trained model performs outside the sample on test data
- C. Both A and B
- D. None of these

ANSWER: C

6. Statistical significance is

- A. The science of collecting, organizing and applying numerical facts
- B. Measure of the probability that a certain hypothesis is incorrect given certain observations
- C. One of the defining aspects of a data warehouse, which is specially built around all the existing applications of the operational data
- D. None of these

ANSWER: B

7. Which of the following is an example of feature extraction?
- A. Constructing bag of words vector from an email
 - B. Applying PCA projects to a large high-dimensional data
 - C. Removing stopwords in a sentence
 - D. All of these

ANSWER: D

8. How can you prevent a clustering algorithm from getting stuck in bad local optima?
- A. Set the same seed value for each run
 - B. Use multiple random initializations
 - C. Both A and B
 - D. None of these

ANSWER: B

9. Adaptive system management is
- A. It uses machine learning technique and program can learn from past experience and adapt themselves to new situation
 - B. Computational procedure that takes some value as input and produces some value as output
 - C. Science of making machines performs tasks that would require intelligence when performed by humans
 - D. None of these

ANSWER: A

10. Binary attribute are
- A. This takes only two values. In general, these values will be 0 and 1 and they can be coded as one bit
 - B. The natural environment of a certain species
 - C. Systems that can be used without knowledge of internal operations
 - D. None of these

ANSWER: A

11. Background knowledge referred to
- A. Additional acquaintance used by a learning algorithm to facilitate the learning process
 - B. Neural network that makes use of a hidden layer
 - C. It is a form of automatic learning
 - D. None of these

ANSWER: A

12. Classification is
- A. Subdivision of a set of examples into a number of classes

- B. Measure of the accuracy, of the classification of a concept that is given by a certain theory
- C. The task of assigning a classification to a set of examples
- D. None of these

ANSWER: A

13. Classification accuracy is
- A. Subdivision of a set of examples into a number of classes
 - B. Measure of the accuracy, of the classification of a concept that is given by a certain theory
 - C. The task of assigning a classification to a set of examples
 - D. None of these

ANSWER: B

14. Cluster is
- A. Group of similar objects that differ significantly from other objects
 - B. Operations on a database to transform or simplify data in order to prepare it for a machine-learning algorithm
 - C. Symbolic representation of facts or ideas from which information can potentially be extracted
 - D. None of these

ANSWER: A

15. Suppose you are given an EM algorithm that finds maximum likelihood estimates for a model with latent variables. You are asked to modify the algorithm so that it finds MAP estimates instead. Which step or steps do you need to modify?

- A. Expectation
- B. Maximization
- C. No modification necessary
- D. Both A & B

ANSWER: B

16. Compared to the variance of the Maximum Likelihood Estimate (MLE), the variance of the Maximum A Posteriori (MAP) estimate is _____

- A. Higher
- B. Same
- C. Lower
- D. It could be any of the above

ANSWER: C

17. Incremental learning referred to
- A. Machine-learning involving different techniques
 - B. The learning algorithmic analyzes the examples on a systematic basis and makes incremental adjustments to the theory that is learned

C. Learning by generalizing from examples

D. None of these

ANSWER: B

18. Inductive learning is

A. Machine-learning involving different techniques

B. The learning algorithm analyzes the examples on a systematic basis and makes incremental adjustments to the theory that is learned

C. Learning by generalizing from examples

D. None of these

ANSWER: C

19. Predicting on whether will it rain or not tomorrow evening at a particular time is a type of _____ problem.

A. Classification

B. Regression

C. Unsupervised learning

D. All of these

ANSWER: A

20. Machine learning is

A. An algorithm that can learn

B. Sub-discipline of computer science that deals with the design and implementation of learning algorithms

C. An approach that abstracts from the actual strategy of an individual algorithm and can therefore be applied to any other form of machine learning.

D. None of these

ANSWER: B

21. A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college. Which of the following statement is true in following case?

A. Feature F1 is an example of nominal variable.

B. Feature F1 is an example of ordinal variable.

C. It doesn't belong to any of the above category.

D. Both of A & B

ANSWER: B

22. If your training loss increases with number of epochs, which of the following could be a possible issue with the learning process?

A. Regularization is too low and model is overfitting

B. Regularization is too high and model is underfitting

C. Step size is too large

D. Step size is too small

ANSWER: C

23. Given a large dataset of medical records from patients suffering from heart disease, try to learn whether there might be different clusters of such patients for which we might tailor separate treatments. What kind of learning problem is this?

- A. Supervised learning
- B. Unsupervised learning
- C. Both A and B
- D. None of these

ANSWER: B

24. Multi-dimensional knowledge is

- A. A class of learning algorithms that try to derive a Prolog program from examples
- B. A table with n independent attributes can be seen as an n-dimensional space
- C. A prediction made using an extremely simple method, such as always predicting the same output
- D. None of these

ANSWER: B

25. The mutual information

- A. Is symmetric
- B. Always non negative
- C. Both A and B
- D. None of these

ANSWER: C

26. Classifying email as a spam, labeling webpages based on their content, voice recognition are the example of ____.

- A. Supervised learning
- B. Unsupervised learning
- C. Machine learning
- D. Deep learning

ANSWER: A

27. Deep learning is a subfield of machine learning where concerned algorithms are inspired by the structure and function of the brain called ____.

- A. Machine learning
- B. Artificial neural networks
- C. Deep learning
- D. Robotics

ANSWER: B

28. Machine learning invented by ____.

- A. John McCarthy

- B. Nicklaus Wirth
- C. Joseph Weizenbaum
- D. Arthur Samuel

ANSWER: D

29. When the number of output classes is greater than one, there are main possibilities to manage a classification problem:

- A. One-vs-all, One-vs-one
- B. One-vs-one, Many-vs-one
- C. One-vs-many, Many-vs-one
- D. None of these

ANSWER: A

30. For a neural network, which one of these structural assumptions is the one that most affects the trade-off between underfitting (i.e. a high bias model) and overfitting (i.e. a high variance model):

- A. The learning rate
- B. The number of hidden nodes
- C. The initial choice of weights
- D. The use of a constant-term unit input

ANSWER: B

31. _____ refers to a model that can neither model the training data nor generalize to new data.

- A. Good fitting
- B. Overfitting
- C. Underfitting
- D. All of the these

ANSWER: C

32. Given two Boolean random variables, A and B, where $P(A) = 1/2$, $P(B) = 1/3$, and $P(A | \neg B) = 1/4$, what is $P(A | B)$?

- A. 1/6
- B. 1/4
- C. 3/4
- D. 1

ANSWER: D

33. Suppose your model is overfitting. Which of the following is NOT a valid way to try and reduce the overfitting?

- A. Increase the amount of training data
- B. Improve the optimization algorithm being used for error minimization
- C. Decrease the model complexity
- D. Reduce the noise in the training data

ANSWER: B

34. Predicting on whether will it rain or not tomorrow evening at a particular time is a type of _____ problem.

- A. Classification
- B. Regression
- C. Unsupervised learning
- D. All of these

ANSWER: A

35. Given a large dataset of medical records from patients suffering from heart disease, try to learn whether there might be different clusters of such patients for which we might tailor separate treatments. What kind of learning problem is this?

- A. Supervised learning
- B. Unsupervised learning
- C. Both A and B
- D. Neither A nor B

ANSWER: B

36. Given a large dataset of medical records from patients suffering from heart disease, try to learn whether there might be different clusters of such patients for which we might tailor separate treatments. What kind of learning problem is this?

- A. Supervised learning
- B. Unsupervised learning
- C. Both A and B
- D. Neither A nor B

ANSWER: B

37. Which of the following is NOT supervised learning?

- A. Decision Tree
- B. PCA
- C. Linear Regression
- D. Naive Bayesian

ANSWER: B

38. In 1984, the computer scientist _____ proposed a mathematical approach to determine whether a problem is learnable by a computer.

- A. John McCarthy
- B. Nicklaus Wirth
- C. L. Valiant
- D. Arthur Samuel

ANSWER: C

39. In binary classification which error measure or loss function is used?
- A. Non-negative error measure
 - B. Mean square error
 - C. Zero-one-loss
 - D. None of these

ANSWER: C

40. Benefits of Parametric Machine Learning Algorithms:
- A. Complex, slow, more training data
 - B. Simpler, faster, less training Data
 - C. Both A and B
 - D. Neither A nor B

ANSWER: B

41. Limitations of Parametric Machine Learning Algorithms is:
- A. Highly Constrained
 - B. Limited Complexity
 - C. Poor Fit
 - D. All of these

ANSWER: D

42. Artificial Neural Networks is example of:
- A. Nonparametric model
 - B. Parametric models
 - C. Both A and B
 - D. None of these

ANSWER: A

43. Benefits of Non-parametric Machine Learning Algorithms:
- A. More data, Slower, Overfitting
 - B. Flexibility, Power, Performance
 - C. Both A and B
 - D. Neither A nor B

ANSWER: B

44. Limitations of Non-parametric Machine Learning Algorithms:
- A. More data, Slower, Overfitting
 - B. Flexibility, Power, Performance
 - C. Both A and B
 - D. Neither A nor B

ANSWER: A

45. Naive Bayes is example of:

- A. Nonparametric model
- B. Parametric models
- C. Both A and B
- D. Neither A nor B

ANSWER: B

46. Which of the following is wrong statement about the maximum likelihood approach?
- A. This method doesn't always involve probability calculations
 - B. It finds a tree that best accounts for the variation in a set of sequences
 - C. The method is similar to the maximum parsimony method
 - D. The analysis is performed on each column of a multiple sequence alignment

ANSWER: A

47. The main disadvantage of maximum likelihood methods is that they are ____
- A. Mathematically less folded
 - B. Mathematically less complex
 - C. Computationally lucid
 - D. Computationally intense

ANSWER: B

48. Which learning is often preferable to MAP learning?
- A. Expectation-maximization
 - B. Log-likelihood (L)
 - C. Maximum-likelihood (ML)
 - D. None of these

ANSWER: C

49. Which is measure used in information theory?
- A. Entropy
 - B. Cross-entropy
 - C. Conditional entropy
 - D. All of these

ANSWER: C

50. Which measure uses bits in information theory?
- A. Entropy
 - B. Cross-entropy
 - C. Conditional entropy
 - D. All of these

ANSWER: A

MCQ questions for unit 3: Regression

Multiple choice questions

1) True-False: Linear Regression is a supervised machine learning algorithm.

- A) TRUE
- B) FALSE

Solution: (A)

Yes, Linear regression is a supervised learning algorithm because it uses true labels for training. Supervised learning algorithm should have input variable (x) and an output variable (Y) for each example.

2) True-False: Linear Regression is mainly used for Regression.

- A) TRUE
- B) FALSE

Solution: (A)

Linear Regression has dependent variables that have continuous values.

3) True-False: It is possible to design a Linear regression algorithm using a neural network?

- A) TRUE
- B) FALSE

Solution: (A)

True. A Neural network can be used as a *universal* approximator, so it can definitely implement a linear regression algorithm.

4) Which of the following methods do we use to find the best fit line for data in Linear Regression?

- A) Least Square Error
- B) Maximum Likelihood
- C) Logarithmic Loss
- D) Both A and B

Solution: (A)

In linear regression, we try to minimize the least square errors of the model to identify the line of best fit.

5) Which of the following evaluation metrics can be used to evaluate a model while modeling a continuous output variable?

- A) AUC-ROC
- B) Accuracy
- C) Logloss
- D) Mean-Squared-Error

Solution: (D)

Since linear regression gives output as continuous values, so in such case we use mean squared error metric to evaluate the model performance. Remaining options are use in case of a classification problem.

6) True-False: Lasso Regularization can be used for variable selection in Linear Regression.

- A) TRUE
- B) FALSE

Solution: (A)

True, In case of lasso regression we apply absolute penalty which makes some of the coefficients zero.

7) Which of the following is true about Residuals ?

- A) Lower is better
- B) Higher is better
- C) A or B depend on the situation
- D) None of these

Solution: (A)

Residuals refer to the error values of the model. Therefore lower residuals are desired.

8) Suppose that we have N independent variables ($X_1, X_2 \dots X_n$) and dependent variable is Y. Now Imagine that you are applying linear regression by fitting the best fit line using least square error on this data.

You found that correlation coefficient for one of its variable (Say X_1) with Y is -0.95.

Which of the following is true for X_1 ?

- A) Relation between the X_1 and Y is weak
- B) Relation between the X_1 and Y is strong**
- C) Relation between the X_1 and Y is neutral
- D) Correlation can't judge the relationship

Solution: (B)

The absolute value of the correlation coefficient denotes the strength of the relationship. Since absolute correlation is very high it means that the relationship is strong between X_1 and Y.

9) Looking at above two characteristics, which of the following option is the correct for Pearson correlation between V1 and V2?

If you are given the two variables V1 and V2 and they are following below two characteristics.

1. If V1 increases then V2 also increases
 2. If V1 decreases then V2 behavior is unknown
- A) Pearson correlation will be close to 1
 - B) Pearson correlation will be close to -1
 - C) Pearson correlation will be close to 0
 - D) None of these**

Solution: (D)

We cannot comment on the correlation coefficient by using only statement 1. We need to consider the both of these two statements. Consider V1 as x and V2 as $|x|$. The correlation coefficient would not be close to 1 in such a case.

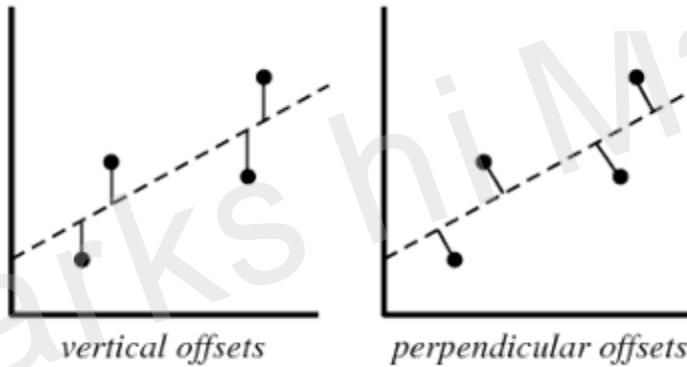
10) Suppose Pearson correlation between V1 and V2 is zero. In such case, is it right to conclude that V1 and V2 do not have any relation between them?

- A) TRUE
- B) FALSE

Solution: (B)

Pearson correlation coefficient between 2 variables might be zero even when they have a relationship between them. If the correlation coefficient is zero, it just means that that they don't move together. We can take examples like $y=|x|$ or $y=x^2$.

11) Which of the following offsets, do we use in linear regression's least square line fit? Suppose horizontal axis is independent variable and vertical axis is dependent variable.



- A) Vertical offset
- B) Perpendicular offset
- C) Both, depending on the situation
- D) None of above

Solution: (A)

We always consider residuals as vertical offsets. We calculate the direct differences between actual value and the Y labels. Perpendicular offset are useful in case of PCA.

12) True- False: Overfitting is more likely when you have huge amount of data to train?

- A) TRUE
- B) FALSE

Solution: (B)

With a small training dataset, it's easier to find a hypothesis to fit the training data exactly i.e. overfitting.

13) We can also compute the coefficient of linear regression with the help of an analytical method called “Normal Equation”. Which of the following is/are true about Normal Equation?

- 1. We don't have to choose the learning rate
- 2. It becomes slow when number of features is very large
- 3. There is no need to iterate

- A) 1 and 2
- B) 1 and 3
- C) 2 and 3
- D) 1,2 and 3

Solution: (D)

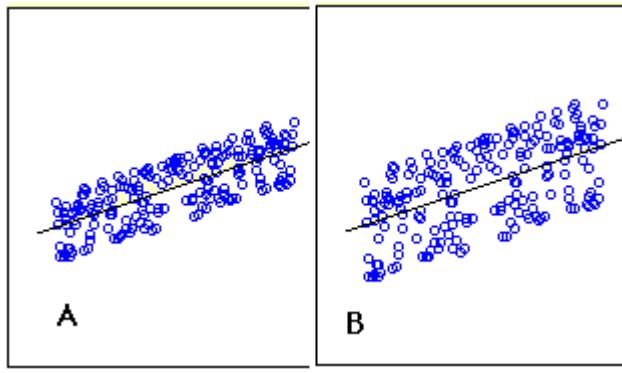
Instead of gradient descent, Normal Equation can also be used to find coefficients. Refer this [article](#) for read more about normal equation.

14) Which of the following statement is true about sum of residuals of A and B?

Below graphs show two fitted regression lines (A & B) on randomly generated data. Now, I want to find the sum of residuals in both cases A and B.

Note:

- 1. Scale is same in both graphs for both axis.
- 2. X axis is independent variable and Y-axis is dependent variable.



- A) A has higher sum of residuals than B
- B) A has lower sum of residual than B
- C) Both have same sum of residuals
- D) None of these

Solution: (C)

Sum of residuals will always be zero, therefore both have same sum of residuals

Question Context 15-17:

Suppose you have fitted a complex regression model on a dataset. Now, you are using Ridge regression with penalty x .

15) Choose the option which describes bias in best manner.

- A) In case of very large x ; bias is low
- B) In case of very large x ; bias is high
- C) We can't say about bias
- D) None of these

Solution: (B)

If the penalty is very large it means model is less complex, therefore the bias would be high.

16) What will happen when you apply very large penalty?

- A) Some of the coefficient will become absolute zero
- B) Some of the coefficient will approach zero but not absolute zero
- C) Both A and B depending on the situation
- D) None of these

Solution: (B)

In lasso some of the coefficient value become zero, but in case of Ridge, the coefficients become close to zero but not zero.

17) What will happen when you apply very large penalty in case of Lasso?

- A) Some of the coefficient will become zero
- B) Some of the coefficient will be approaching to zero but not absolute zero
- C) Both A and B depending on the situation
- D) None of these

Solution: (A)

As already discussed, lasso applies absolute penalty, so some of the coefficients will become zero.

18) Which of the following statement is true about outliers in Linear regression?

- A) Linear regression is sensitive to outliers
- B) Linear regression is not sensitive to outliers
- C) Can't say
- D) None of these

Solution: (A)

The slope of the regression line will change due to outliers in most of the cases. So Linear Regression is sensitive to outliers.

19) Suppose you plotted a scatter plot between the residuals and predicted values in linear regression and you found that there is a relationship between them. Which of the following conclusion do you make about this situation?

- A) Since the there is a relationship means our model is not good
- B) Since the there is a relationship means our model is good
- C) Can't say
- D) None of these

Solution: (A)

There should not be any relationship between predicted values and residuals. If there exists any relationship between them, it means that the model has not perfectly captured the information in the data.

Question Context 20-22:

Suppose that you have a dataset D1 and you design a linear regression model of degree 3 polynomial and you found that the training and testing error is “0” or in another terms it perfectly fits the data.

20) What will happen when you fit degree 4 polynomial in linear regression?

- A) There are high chances that degree 4 polynomial will over fit the data
- B) There are high chances that degree 4 polynomial will under fit the data
- C) Can't say
- D) None of these

Solution: (A)

Since more degree 4 will be more complex(overfit the data) than the degree 3 model so it will again perfectly fit the data. In such case training error will be zero but test error may not be zero.

21) What will happen when you fit degree 2 polynomial in linear regression?

- A) It is high chances that degree 2 polynomial will over fit the data
- B) It is high chances that degree 2 polynomial will under fit the data
- C) Can't say
- D) None of these

Solution: (B)

If a degree 3 polynomial fits the data perfectly, it's highly likely that a simpler model(degree 2 polynomial) might under fit the data.

22) In terms of bias and variance. Which of the following is true when you fit degree 2 polynomial?

- A) Bias will be high, variance will be high
- B) Bias will be low, variance will be high

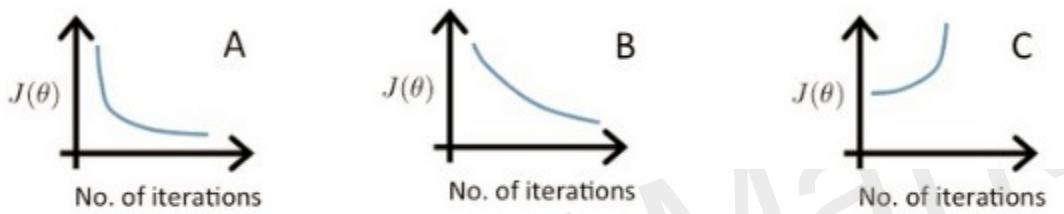
- C) Bias will be high, variance will be low
 D) Bias will be low, variance will be low

Solution: (C)

Since a degree 2 polynomial will be less complex as compared to degree 3, the bias will be high and variance will be low.

Question Context 23:

Which of the following is true about below graphs(A,B, C left to right) between the cost function and Number of iterations?



23) Suppose α_1, α_2 and α_3 are the three learning rates for A,B,C respectively. Which of the following is true about α_1, α_2 and α_3 ?

- A) $\alpha_2 < \alpha_1 < \alpha_3$
 B) $\alpha_1 > \alpha_2 > \alpha_3$
 C) $\alpha_1 = \alpha_2 = \alpha_3$
 D) None of these

Solution: (A)

In case of high learning rate, step will be high, the objective function will decrease quickly initially, but it will not find the global minima and objective function starts increasing after a few iterations.

In case of low learning rate, the step will be small. So the objective function will decrease slowly

Question Context 24-25:

We have been given a dataset with n records in which we have input attribute as x and output attribute as y . Suppose we use a linear regression method to model this data. To test our linear regressor, we split the data in training set and test set randomly.

24) Now we increase the training set size gradually. As the training set size increases, what do you expect will happen with the mean training error?

- A) Increase
- B) Decrease
- C) Remain constant
- D) Can't Say

Solution: (D)

Training error may increase or decrease depending on the values that are used to fit the model. If the values used to train contain more outliers gradually, then the error might just increase.

25) What do you expect will happen with bias and variance as you increase the size of training data?

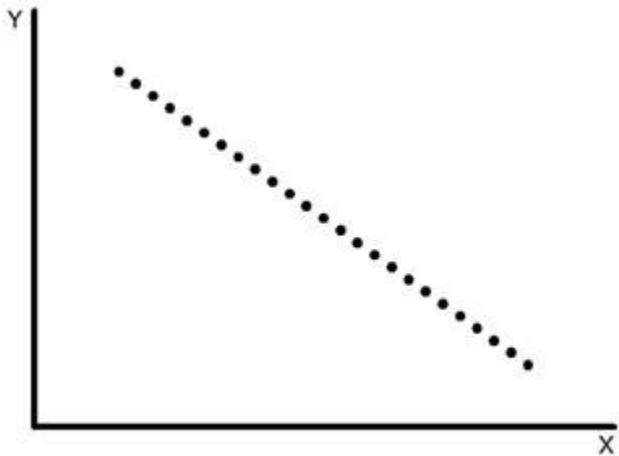
- A) Bias increases and Variance increases
- B) Bias decreases and Variance increases
- C) Bias decreases and Variance decreases
- D) Bias increases and Variance decreases
- E) Can't Say False

Solution: (D)

As we increase the size of the training data, the bias would increase while the variance would decrease.

Question Context 26:

Consider the following data where one input(X) and one output(Y) is given.



26) What would be the root mean square training error for this data if you run a Linear Regression model of the form ($Y = A_0 + A_1X$)?

- A) Less than 0
- B) Greater than zero
- C) Equal to 0
- D) None of these

Solution: (C)

We can perfectly fit the line on the following data so mean error will be zero.

Question Context 27-28:

Suppose you have been given the following scenario for training and validation error for Linear Regression.

Scenario	Learning Rate	Number of iterations	Training Error	Validation Error
1	0.1	1000	100	110
2	0.2	600	90	105
3	0.3	400	110	110
4	0.4	300	120	130
5	0.4	250	130	150

27) Which of the following scenario would give you the right hyper parameter?

- A) 1
- B) 2
- C) 3
- D) 4

Solution: (B)

Option B would be the better option because it leads to less training as well as validation error.

28) Suppose you got the tuned hyper parameters from the previous question. Now, Imagine you want to add a variable in variable space such that this added feature is important. Which of the following thing would you observe in such case?

- A) Training Error will decrease and Validation error will increase
- B) Training Error will increase and Validation error will increase
- C) Training Error will increase and Validation error will decrease
- D) Training Error will decrease and Validation error will decrease
- E) None of the above

Solution: (D)

If the added feature is important, the training and validation error would decrease.

Question Context 29-30:

Suppose, you got a situation where you find that your linear regression model is under fitting the data.

29) In such situation which of the following options would you consider?

1. I will add more variables
2. I will start introducing polynomial degree variables
3. I will remove some variables

- A) 1 and 2
- B) 2 and 3
- C) 1 and 3
- D) 1, 2 and 3

Solution: (A)

In case of under fitting, you need to induce more variables in variable space or you can add some polynomial degree variables to make the model more complex to be able to fit the data better.

30) Now situation is same as written in previous question(under fitting).Which of following regularization algorithm would you prefer?

- A) L1
- B) L2
- C) Any
- D) None of these

Solution: (D)

I won't use any regularization methods because regularization is used in case of overfitting.

MCQs ON Linear Regression

1) True-False: Is Logistic regression a supervised machine learning algorithm?

- A) TRUE
- B) FALSE

Solution: A

True, Logistic regression is a supervised learning algorithm because it uses true labels for training. Supervised learning algorithm should have input variables (x) and a target variable (Y) when you train the model .

2) True-False: Is Logistic regression mainly used for Regression?

- A) TRUE
- B) FALSE

Solution: B

Logistic regression is a classification algorithm, don't confuse with the name regression.

3) True-False: Is it possible to design a logistic regression algorithm using a Neural Network Algorithm?

- A) TRUE
- B) FALSE

Solution: A

True, Neural network is a universal approximator so it can implement linear regression algorithm.

4) True-False: Is it possible to apply a logistic regression algorithm on a 3-class Classification problem?

- A) TRUE
- B) FALSE

Solution: A

Yes, we can apply logistic regression on 3 classification problem, We can use One Vs all method for 3 class classification in logistic regression.

5) Which of the following methods do we use to best fit the data in Logistic Regression?

- A) Least Square Error
- B) Maximum Likelihood
- C) Jaccard distance
- D) Both A and B

Solution: B

Logistic regression uses maximum likelihood estimate for training a logistic regression.

6) Which of the following evaluation metrics can not be applied in case of logistic regression output to compare with target?

- A) AUC-ROC
- B) Accuracy
- C) Logloss
- D) Mean-Squared-Error

Solution: D

Since, Logistic Regression is a classification algorithm so its output can not be real time value so mean squared error can not be used for evaluating it

7) One of the very good methods to analyze the performance of Logistic Regression is AIC, which is similar to R-Squared in Linear Regression. Which of the following is true about AIC?

- A) We prefer a model with minimum AIC value
- B) We prefer a model with maximum AIC value
- C) Both but depend on the situation
- D) None of these

Solution: A

We select the best model in logistic regression which has least AIC. For more information refer this source: <http://www4.ncsu.edu/~shu3/Presentation/AIC.pdf>

8) [True-False] Standardisation of features is required before training a Logistic Regression.

- A) TRUE
- B) FALSE

Solution: B

Standardization isn't required for logistic regression. The main goal of standardizing features is to help convergence of the technique used for optimization.

9) Which of the following algorithms do we use for Variable Selection?

- A) LASSO
- B) Ridge

- C) Both
- D) None of these

Solution: A

In case of lasso we apply a absolute penalty, after increasing the penalty in lasso some of the coefficient of variables may become zero.

Context: 10-11

Consider a following model for logistic regression: $P(y=1|x, w) = g(w_0 + w_1x)$ where $g(z)$ is the logistic function.

In the above equation the $P(y=1|x; w)$, viewed as a function of x , that we can get by changing the parameters w .

10) What would be the range of p in such case?

- A) (0, inf)
- B) (-inf, 0)
- C) (0, 1)
- D) (-inf, inf)

Solution: C

For values of x in the range of real number from $-\infty$ to $+\infty$ Logistic function will give the output between (0,1)

11) In above question what do you think which function would make p between (0,1)?

- A) logistic function
- B) Log likelihood function
- C) Mixture of both
- D) None of them

Solution: A

Explanation is same as question number 10

Context: 12-13

Suppose you train a logistic regression classifier and your hypothesis function H is

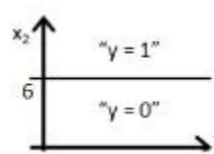
$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

Where

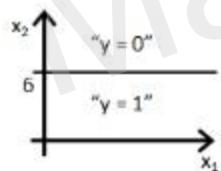
$$\theta_0 = 6, \theta_1 = 0, \theta_2 = -1.$$

12) Which of the following figure will represent the decision boundary as given by above classifier?

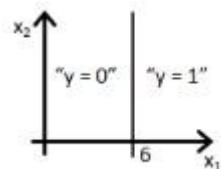
A)



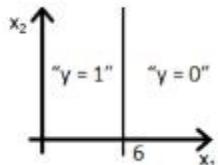
B)



C)



D)

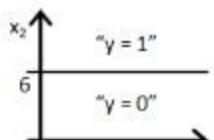


Solution: B

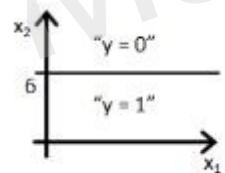
Option B would be the right answer. Since our line will be represented by $y = g(-6+x_2)$ which is shown in the option A and option B. But option B is the right answer because when you put the value $x_2 = 6$ in the equation then $y = g(0)$ you will get that means $y= 0.5$ will be on the line, if you increase the value of x_2 greater than 6 you will get negative values so output will be the region $y=0$.

13) If you replace coefficient of x_1 with x_2 what would be the output figure?

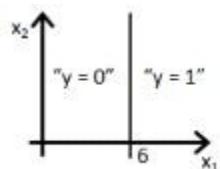
A)



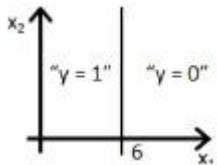
B)



C)



D)



Solution: D

Same explanation as in previous question.

14) Suppose you have been given a fair coin and you want to find out the odds of getting heads. Which of the following option is true for such a case?

- A) odds will be 0
- B) odds will be 0.5
- C) odds will be 1
- D) None of these

Solution: C

Odds are defined as the ratio of the probability of success and the probability of failure. So in case of fair coin probability of success is $1/2$ and the probability of failure is $1/2$ so odd would be 1

15) The logit function(given as $l(x)$) is the log of odds function. What could be the range of logit function in the domain $x=[0,1]$?

- A) $(-\infty, \infty)$
- B) $(0,1)$
- C) $(0, \infty)$
- D) $(-\infty, 0)$

Solution: A

For our purposes, the odds function has the advantage of transforming the probability function, which has values from 0 to 1, into an equivalent function with values between 0 and ∞ . When we take the natural log of the odds function, we get a range of values from $-\infty$ to ∞ .

16) Which of the following option is true?

- A) Linear Regression errors values has to be normally distributed but in case of Logistic Regression it is not the case
- B) Logistic Regression errors values has to be normally distributed but in case of Linear Regression it is not the case
- C) Both Linear Regression and Logistic Regression error values have to be normally distributed
- D) Both Linear Regression and Logistic Regression error values have not to be normally distributed

Solution:A

Only A is true.

17) Which of the following is true regarding the logistic function for any value “x”?

Note:

Logistic(x): is a logistic function of any number “x”

Logit(x): is a logit function of any number “x”

Logit_inv(x): is a inverse logit function of any number “x”

- A) $\text{Logistic}(x) = \text{Logit}(x)$
- B) $\text{Logistic}(x) = \text{Logit_inv}(x)$
- C) $\text{Logit_inv}(x) = \text{Logit}(x)$
- D) None of these

Solution: B

MCQ For UNIT 2

1) Which of the following statement is true in following case?

- A) Feature F1 is an example of nominal variable.
- B) Feature F1 is an example of ordinal variable.
- C) It doesn't belong to any of the above category.
- D) Both of these

Solution: (B)

Ordinal variables are the variables which has some order in their categories. For example, grade A should be consider as high grade than grade B.

2) Which of the following is an example of a deterministic algorithm?

- A) PCA
- B) K-Means
- C) None of the above

Solution: (A)

A deterministic algorithm is that in which output does not change on different runs. PCA would give the same result if we run again, but not k-means.

3) [True or False] A Pearson correlation between two variables is zero but, still their values can still be related to each other.

- A) TRUE
- B) FALSE

Solution: (A)

$Y=X^2$. Note that, they are not only associated, but one is a function of the other and Pearson correlation between them is 0.

4) Which of the following statement(s) is / are true for Gradient Decent (GD) and Stochastic Gradient Decent (SGD)?

- 1. In GD and SGD, you update a set of parameters in an iterative manner to minimize the error function.**
- 2. In SGD, you have to run through all the samples in your training set for a single update of a parameter in each iteration.**
- 3. In GD, you either use the entire data or a subset of training data to update a parameter in each iteration.**

A) Only 1

B) Only 2

C) Only 3

D) 1 and 2

E) 2 and 3

F) 1,2 and 3

Solution: (A)

In SGD for each iteration you choose the batch which is generally contain the random sample of data But in case of GD each iteration contain the all of the training observations.

5) Which of the following hyper parameter(s), when increased may cause random forest to over fit the data?

- 1. Number of Trees**
- 2. Depth of Tree**
- 3. Learning Rate**

A) Only 1

B) Only 2

C) Only 3

- D) 1 and 2
- E) 2 and 3
- F) 1,2 and 3

Solution: (B)

Usually, if we increase the depth of tree it will cause overfitting. Learning rate is not an hyperparameter in random forest. Increase in the number of tree will cause under fitting.

6) Imagine, you are working with “Analytics Vidhya” and you want to develop a machine learning algorithm which predicts the number of views on the articles.

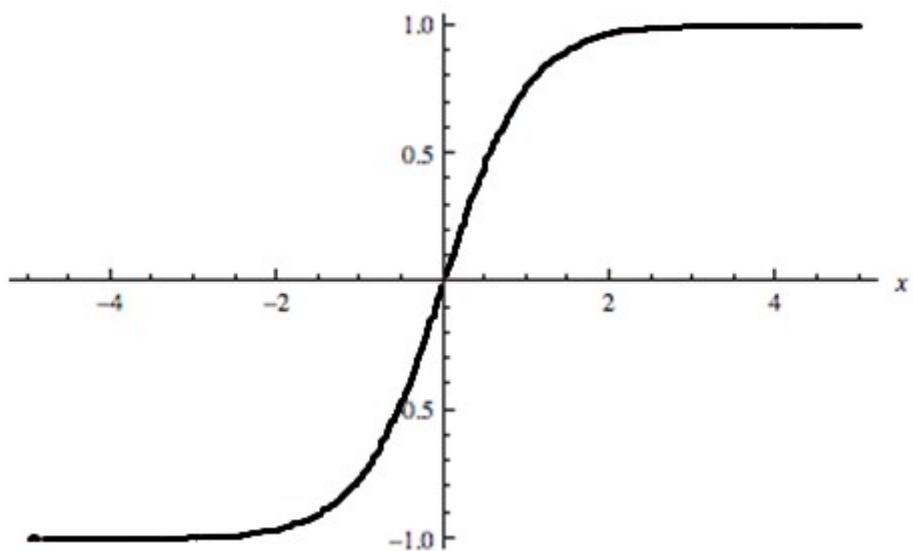
Your analysis is based on features like author name, number of articles written by the same author on Analytics Vidhya in past and a few other features. Which of the following evaluation metric would you choose in that case?

- 1. Mean Square Error
 - 2. Accuracy
 - 3. F1 Score
- A) Only 1
 - B) Only 2
 - C) Only 3
 - D) 1 and 3
 - E) 2 and 3
 - F) 1 and 2

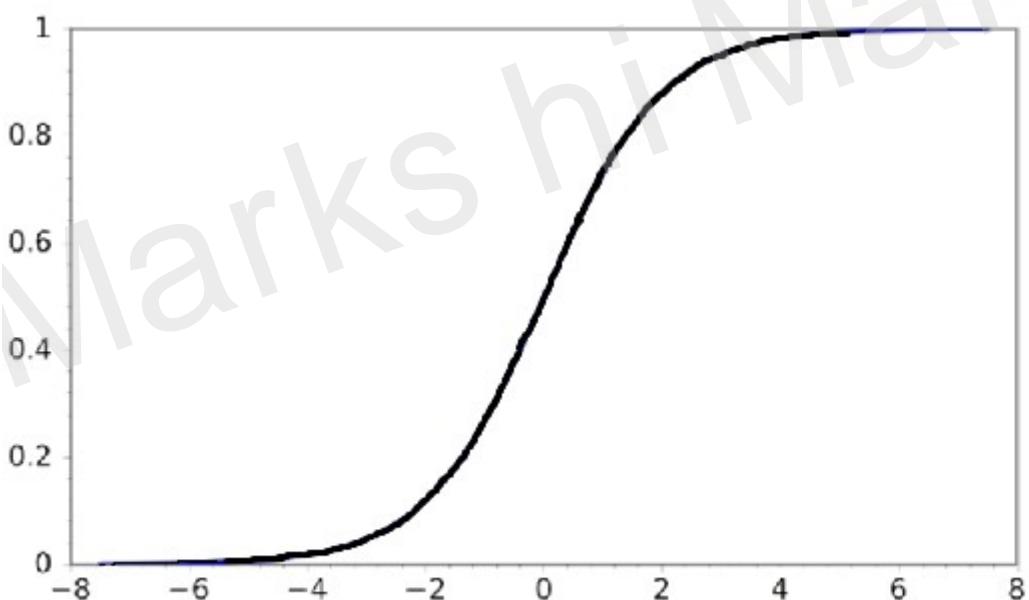
Solution:(A)

You can think that the number of views of articles is the continuous target variable which fall under the regression problem. So, mean squared error will be used as an evaluation metrics.

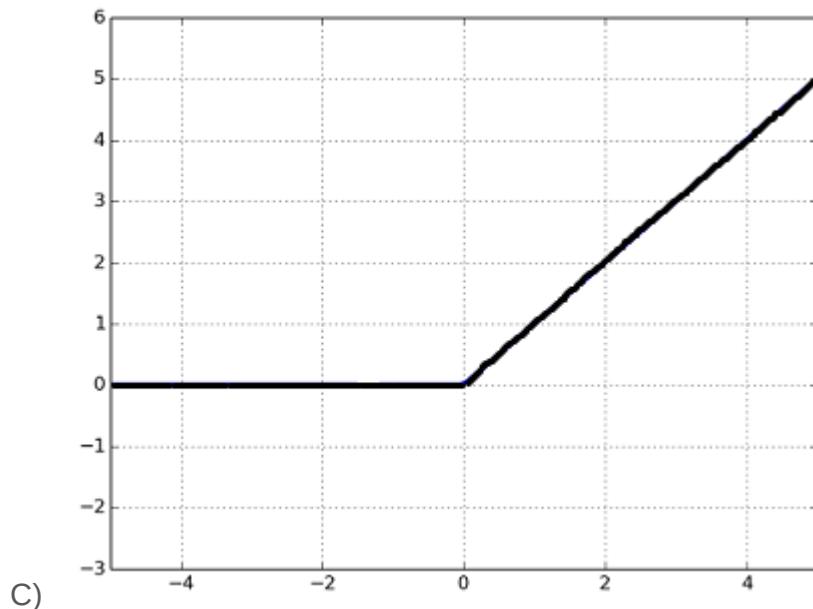
7) Given below are three images (1,2,3). Which of the following option is correct for these images?



A)



B)



C)

- A) 1 is tanh, 2 is ReLU and 3 is SIGMOID activation functions.
- B) 1 is SIGMOID, 2 is ReLU and 3 is tanh activation functions.
- C) 1 is ReLU, 2 is tanh and 3 is SIGMOID activation functions.
- D) 1 is tanh, 2 is SIGMOID and 3 is ReLU activation functions.

Solution: (D)

The range of SIGMOID function is [0,1].

The range of the tanh function is [-1,1].

The range of the RELU function is [0, infinity].

So Option D is the right answer.

8) Below are the 8 actual values of target variable in the train file.

[0,0,0,1,1,1,1,1]

What is the entropy of the target variable?

- A) -(5/8 log(5/8) + 3/8 log(3/8))
- B) 5/8 log(5/8) + 3/8 log(3/8)

C) $\frac{3}{8} \log(\frac{5}{8}) + \frac{5}{8} \log(\frac{3}{8})$

D) $\frac{5}{8} \log(\frac{3}{8}) - \frac{3}{8} \log(\frac{5}{8})$

Solution: (A)

$$-\sum p(x) * \log p(x)$$

The formula for entropy is

So the answer is A.

9) Let's say, you are working with categorical feature(s) and you have not looked at the distribution of the categorical variable in the test data.

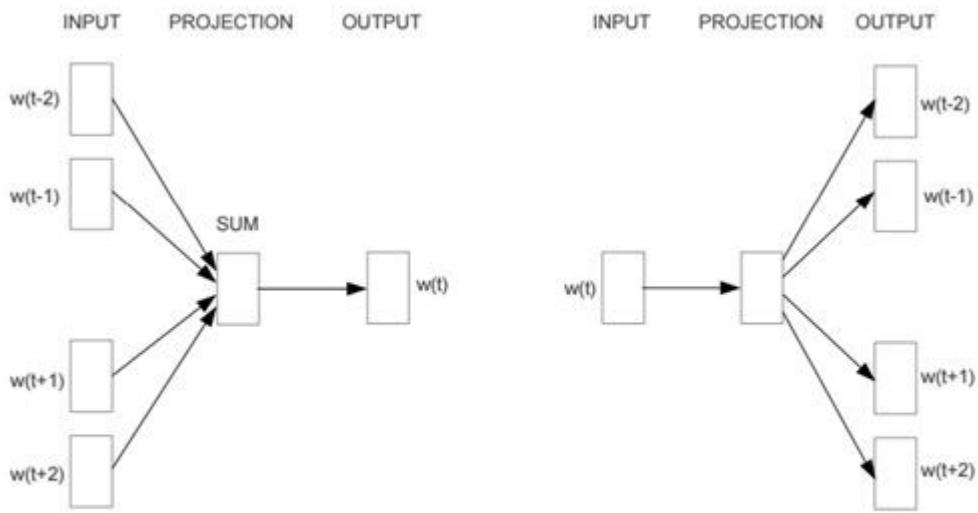
You want to apply one hot encoding (OHE) on the categorical feature(s). What challenges you may face if you have applied OHE on a categorical variable of train dataset?

- A) All categories of categorical variable are not present in the test dataset.
- B) Frequency distribution of categories is different in train as compared to the test dataset.
- C) Train and Test always have same distribution.
- D) Both A and B
- E) None of these

Solution: (D)

Both are true, The OHE will fail to encode the categories which is present in test but not in train so it could be one of the main challenges while applying OHE. The challenge given in option B is also true you need to more careful while applying OHE if frequency distribution doesn't same in train and test.

10) Skip gram model is one of the best models used in Word2vec algorithm for words embedding. Which one of the following models depict the skip gram model?



Model A

Model B

- A) A
- B) B
- C) Both A and B
- D) None of these

Solution: (B)

Both models (model1 and model2) are used in Word2vec algorithm. The model1 represent a CBOW model where as Model2 represent the Skip gram model.

11) Let's say, you are using activation function X in hidden layers of neural network. At a particular neuron for any given input, you get the output as “-0.0001”. Which of the following activation function could X represent?

- A) ReLU
- B) tanh
- C) SIGMOID
- D) None of these

Solution: (B)

The function is a tanh because the this function output range is between (-1,-1).

12) [True or False] LogLoss evaluation metric can have negative values.

- A) TRUE
- B) FALSE

Solution: (B)

Log loss cannot have negative values.

13) Which of the following statements is/are true about “Type-1” and “Type-2” errors?

1. Type1 is known as false positive and Type2 is known as false negative.
2. Type1 is known as false negative and Type2 is known as false positive.
3. Type1 error occurs when we reject a null hypothesis when it is actually true.

- A) Only 1
- B) Only 2
- C) Only 3
- D) 1 and 2
- E) 1 and 3
- F) 2 and 3

Solution: (E)

In statistical hypothesis testing, a type I error is the incorrect rejection of a true null hypothesis (a “false positive”), while a type II error is incorrectly retaining a false null hypothesis (a “false negative”).

14) Which of the following is/are one of the important step(s) to pre-process the text in NLP based projects?

- 1. Stemming**
- 2. Stop word removal**
- 3. Object Standardization**

A) 1 and 2

B) 1 and 3

C) 2 and 3

D) 1,2 and 3

Solution: (D)

Stemming is a rudimentary rule-based process of stripping the suffixes ("ing", "ly", "es", "s" etc) from a word.

Stop words are those words which will have no relevance to the context of the data for example is/am/are.

Object Standardization is also one of the good way to pre-process the text.

15) Suppose you want to project high dimensional data into lower dimensions. The two most famous dimensionality reduction algorithms used here are PCA and t-SNE. Let's say you have applied both algorithms respectively on data "X" and you got the datasets "X_projected_PCA" , "X_projected_tSNE".

Which of the following statements is true for "X_projected_PCA" & "X_projected_tSNE" ?

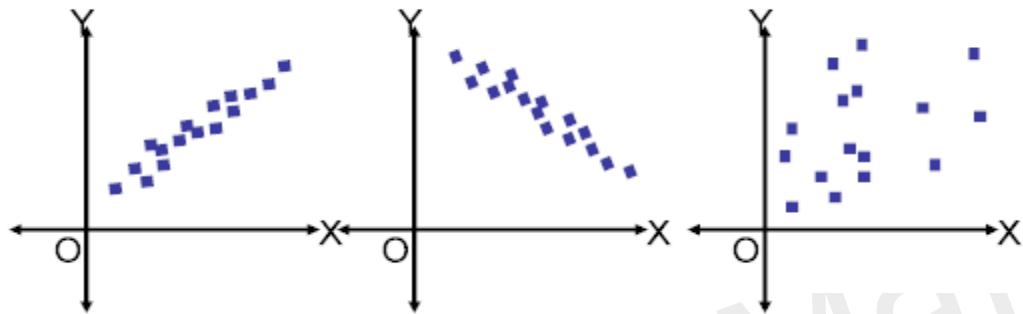
- A) X_projected_PCA will have interpretation in the nearest neighbour space.
- B) X_projected_tSNE will have interpretation in the nearest neighbour space.
- C) Both will have interpretation in the nearest neighbour space.
- D) None of them will have interpretation in the nearest neighbour space.

Solution: (B)

t-SNE algorithm consider nearest neighbour points to reduce the dimensionality of the data. So, after using t-SNE we can think that reduced dimensions will also have interpretation in nearest neighbour space. But in case of PCA it is not the case.

Context: 16-17

Given below are three scatter plots for two features (Image 1, 2 & 3 from left to right).



16) In the above images, which of the following is/are example of multi-collinear features?

- A) Features in Image 1
- B) Features in Image 2
- C) Features in Image 3
- D) Features in Image 1 & 2
- E) Features in Image 2 & 3
- F) Features in Image 3 & 1

Solution: (D)

In Image 1, features have high positive correlation where as in Image 2 has high negative correlation between the features so in both images pair of features are the example of multicollinear features.

17) In previous question, suppose you have identified multi-collinear features. Which of the following action(s) would you perform next?

- 1. Remove both collinear variables.**
- 2. Instead of removing both variables, we can remove only one variable.**
- 3. Removing correlated variables might lead to loss of information. In order to retain those variables, we can use penalized regression models like ridge or lasso regression.**

A) Only 1

B) Only 2

C) Only 3

D) Either 1 or 3

E) Either 2 or 3

Solution: (E)

You cannot remove the both features because after removing the both features you will lose all of the information so you should either remove the only 1 feature or you can use the regularization algorithm like L1 and L2.

18) Adding a non-important feature to a linear regression model may result in.

- 1. Increase in R-square**
- 2. Decrease in R-square**

A) Only 1 is correct

B) Only 2 is correct

C) Either 1 or 2

D) None of these

Solution: (A)

After adding a feature in feature space, whether that feature is important or unimportant features the R-squared always increase.

19) Suppose, you are given three variables X, Y and Z. The Pearson correlation coefficients for (X, Y), (Y, Z) and (X, Z) are C₁, C₂ & C₃ respectively.

Now, you have added 2 in all values of X (i.e. new values become X+2), subtracted 2 from all values of Y (i.e. new values are Y-2) and Z remains the same. The new coefficients for (X,Y), (Y,Z) and (X,Z) are given by D₁, D₂ & D₃ respectively. How do the values of D₁, D₂ & D₃ relate to C₁, C₂ & C₃?

- A) D₁= C₁, D₂ < C₂, D₃ > C₃
- B) D₁ = C₁, D₂ > C₂, D₃ > C₃
- C) D₁ = C₁, D₂ > C₂, D₃ < C₃
- D) D₁ = C₁, D₂ < C₂, D₃ < C₃
- E) D₁ = C₁, D₂ = C₂, D₃ = C₃
- F) Cannot be determined

Solution: (E)

Correlation between the features won't change if you add or subtract a value in the features.

20) Imagine, you are solving a classification problems with highly imbalanced class. The majority class is observed 99% of times in the training data.

Your model has 99% accuracy after taking the predictions on test data. Which of the following is true in such a case?

1. Accuracy metric is not a good idea for imbalanced class problems.
2. Accuracy metric is a good idea for imbalanced class problems.
3. Precision and recall metrics are good for imbalanced class problems.
4. Precision and recall metrics aren't good for imbalanced class problems.

A) 1 and 3

B) 1 and 4

C) 2 and 3

D) 2 and 4

Solution: (A)

Refer the question number 4 from in [this](#) article.

21) In ensemble learning, you aggregate the predictions for weak learners, so that an ensemble of these models will give a better prediction than prediction of individual models.

Which of the following statements is / are true for weak learners used in ensemble model?

1. They don't usually overfit.
2. They have high bias, so they cannot solve complex learning problems
3. They usually overfit.

A) 1 and 2

B) 1 and 3

C) 2 and 3

D) Only 1

E) Only 2

F) None of the above

Solution: (A)

Weak learners are sure about particular part of a problem. So, they usually don't overfit which means that weak learners have low variance and high bias.

22) Which of the following options is/are true for K-fold cross-validation?

1. Increase in K will result in higher time required to cross validate the result.
2. Higher values of K will result in higher confidence on the cross-validation result as compared to lower value of K.
3. If $K=N$, then it is called Leave one out cross validation, where N is the number of observations.

A) 1 and 2

B) 2 and 3

C) 1 and 3

D) 1,2 and 3

Solution: (D)

Larger k value means less bias towards overestimating the true expected error (as training folds will be closer to the total dataset) and higher running time (as you are getting closer to the limit case: Leave-One-Out CV). We also need to consider the variance between the k folds accuracy while selecting the k.

Question Context 23-24

Cross-validation is an important step in machine learning for hyper parameter tuning. Let's say you are tuning a hyper-parameter "max_depth" for GBM by selecting it from 10 different depth values (values are greater than 2) for tree based model using 5-fold cross validation.

Time taken by an algorithm for training (on a model with max_depth 2) 4-fold is 10 seconds and for the prediction on remaining 1-fold is 2 seconds.

Note: Ignore hardware dependencies from the equation.

23) Which of the following option is true for overall execution time for 5-fold cross validation with 10 different values of "max_depth"?

A) Less than 100 seconds

- B) 100 – 300 seconds
- C) 300 – 600 seconds
- D) More than or equal to 600 seconds
- C) None of the above
- D) Can't estimate

Solution: (D)

Each iteration for depth “2” in 5-fold cross validation will take 10 secs for training and 2 second for testing. So, 5 folds will take $12*5 = 60$ seconds. Since we are searching over the 10 depth values so the algorithm would take $60*10 = 600$ seconds. But training and testing a model on depth greater than 2 will take more time than depth “2” so overall timing would be greater than 600.

24) In previous question, if you train the same algorithm for tuning 2 hyper parameters say “max_depth” and “learning_rate”.

You want to select the right value against “max_depth” (from given 10 depth values) and learning rate (from given 5 different learning rates). In such cases, which of the following will represent the overall time?

- A) 1000-1500 second
- B) 1500-3000 Second
- C) More than or equal to 3000 Second
- D) None of these

Solution: (D)

Same as question number 23.

25) Given below is a scenario for training error TE and Validation error VE for a machine learning algorithm M1. You want to choose a hyperparameter (H) based on TE and VE.

H	TE	VE
1	105	90
2	200	85
3	250	96
4	105	85
5	300	100

Which value of H will you choose based on the above table?

- A) 1
- B) 2
- C) 3
- D) 4
- E) 5

Solution: (D)

Looking at the table, option D seems the best

26) What would you do in PCA to get the same projection as SVD?

- A) Transform data to zero mean
- B) Transform data to zero median
- C) Not possible
- D) None of these

Solution: (A)

When the data has a zero mean vector PCA will have same projections as SVD, otherwise you have to centre the data first before taking SVD.

Question Context 27-28

Assume there is a black box algorithm, which takes training data with multiple observations ($t_1, t_2, t_3, \dots, t_n$) and a new observation (q_1). The black box outputs the nearest neighbor of q_1 (say t_i) and its corresponding class label c_i .

You can also think that this black box algorithm is same as 1-NN (1-nearest neighbor).

27) It is possible to construct a k-NN classification algorithm based on this black box alone.

Note: Where n (number of training observations) is very large compared to k .

A) TRUE

B) FALSE

Solution: (A)

In first step, you pass an observation (q_1) in the black box algorithm so this algorithm would return a nearest observation and its class.

In second step, you pass it through the black box again. The black box algorithm will again return the nearest observation and its class.

You need to repeat this procedure k times

28) Instead of using 1-NN black box we want to use the j-NN ($j > 1$) algorithm as black box. Which of the following option is correct for finding k-NN using j-NN?

1. J must be a proper factor of k
2. J > k
3. Not possible

A) 1

B) 2

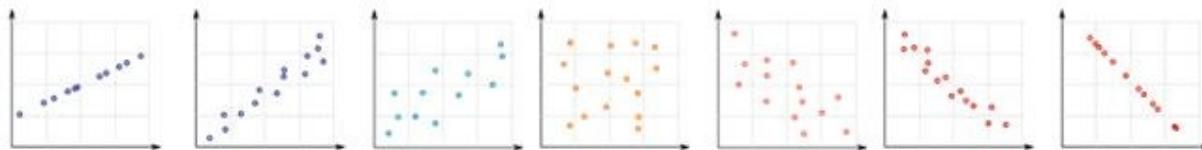
C) 3

Solution: (A)

Same as question number 27

29) Suppose you are given 7 Scatter plots 1-7 (left to right) and you want to compare Pearson correlation coefficients between variables of each scatterplot.

Which of the following is in the right order?



1. $1 < 2 < 3 < 4$
2. $1 > 2 > 3 > 4$
3. $7 < 6 < 5 < 4$
4. $7 > 6 > 5 > 4$

A) 1 and 3

B) 2 and 3

C) 1 and 4

D) 2 and 4

Solution: (B)

from image 1to 4 correlation is decreasing (absolute value). But from image 4 to 7 correlation is increasing but values are negative (for example, 0, -0.3, -0.7, -0.99).

30) You can evaluate the performance of a binary class classification problem using different metrics such as accuracy, log-loss, F-Score. Let's say, you are using the log-loss function as evaluation metric.

Which of the following option is / are true for interpretation of log-loss as an evaluation metric?

$$\text{logLoss} = \frac{-1}{N} \sum_{i=1}^N (y_i (\log p_i) + (1 - y_i) \log(1 - p_i))$$

1.

- If a classifier is confident about an incorrect classification, then log-loss will penalise it heavily.
- 2. For a particular observation, the classifier assigns a very small probability for the correct class then the corresponding contribution to the log-loss will be very large.
- 3. Lower the log-loss, the better is the model.

A) 1 and 3

B) 2 and 3

C) 1 and 2

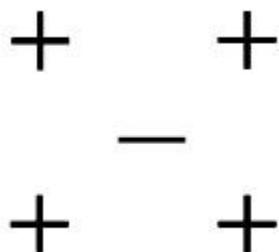
D) 1,2 and 3

Solution: (D)

Options are self-explanatory.

Question 31-32

Below are five samples given in the dataset.



Note: Visual distance between the points in the image represents the actual distance.

31) Which of the following is leave-one-out cross-validation accuracy for 3-NN (3-nearest neighbor)?

A) 0

D) 0.4

C) 0.8

D) 1

Solution: (C)

In Leave-One-Out cross validation, we will select $(n-1)$ observations for training and 1 observation of validation. Consider each point as a cross validation point and then find the 3 nearest point to this point. So if you repeat this procedure for all points you will get the correct classification for all positive class given in the above figure but negative class will be misclassified. Hence you will get 80% accuracy.

32) Which of the following value of K will have least leave-one-out cross validation accuracy?

A) 1NN

B) 3NN

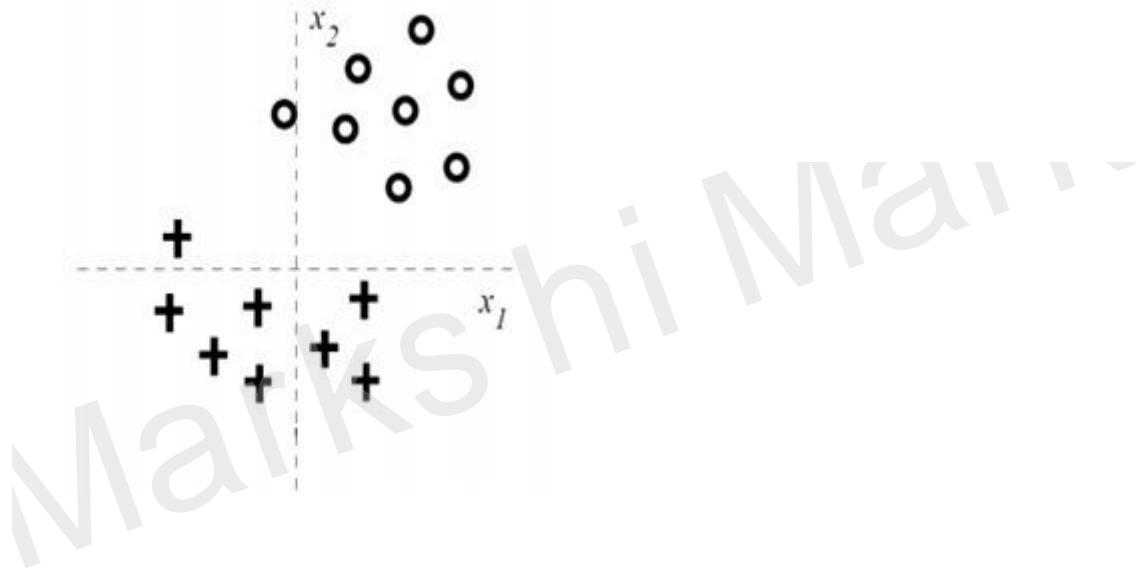
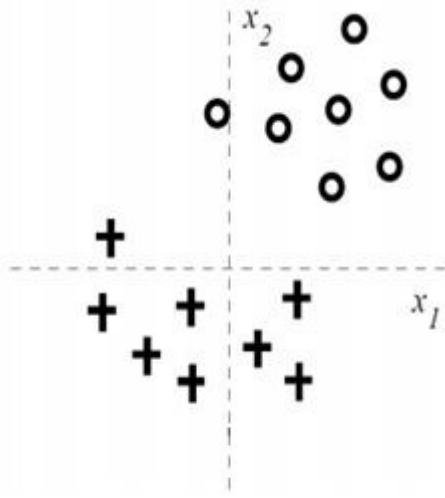
C) 4NN

D) All have same leave one out error

Solution: (A)

Each point which will always be misclassified in 1-NN which means that you will get the 0% accuracy.

33) Suppose you are given the below data and you want to apply a logistic regression model for classifying it in two given classes.



You are using logistic regression with L1 regularization.

$$\sum_{i=1}^n \log P(y_i|x_i, w_0, w_1, w_2) - C(|w_1| + |w_2|).$$

Where C is the regularization parameter and w1 & w2 are the coefficients of x1 and x2.

Which of the following option is correct when you increase the value of C from zero to a very large value?

- A) First w2 becomes zero and then w1 becomes zero
- B) First w1 becomes zero and then w2 becomes zero

- C) Both becomes zero at the same time
- D) Both cannot be zero even after very large value of C

Solution: (B)

By looking at the image, we see that even on just using x_2 , we can efficiently perform classification. So at first w_1 will become 0. As regularization parameter increases more, w_2 will come more and more closer to 0.

34) Suppose we have a dataset which can be trained with 100% accuracy with help of a decision tree of depth 6. Now consider the points below and choose the option based on these points.

Note: All other hyper parameters are same and other factors are not affected.

- 1. Depth 4 will have high bias and low variance**
- 2. Depth 4 will have low bias and low variance**

- A) Only 1
- B) Only 2
- C) Both 1 and 2
- D) None of the above

Solution: (A)

If you fit decision tree of depth 4 in such data means it will more likely to underfit the data. So, in case of underfitting you will have high bias and low variance.

35) Which of the following options can be used to get global minima in k-Means Algorithm?

- 1. Try to run algorithm for different centroid initialization**
- 2. Adjust number of iterations**
- 3. Find out the optimal number of clusters**

- A) 2 and 3

- B) 1 and 3
- C) 1 and 2
- D) All of above

Solution: (D)

All of the option can be tuned to find the global minima.

36) Imagine you are working on a project which is a binary classification problem. You trained a model on training dataset and get the below confusion matrix on validation dataset.

n=165	Predicted:	
	NO	YES
Actual: NO	50	10
Actual: YES	5	100

Based on the above confusion matrix, choose which option(s) below will give you correct predictions?

- 1. Accuracy is ~0.91
 - 2. Misclassification rate is ~ 0.91
 - 3. False positive rate is ~0.95
 - 4. True positive rate is ~0.95
- A) 1 and 3
 - B) 2 and 4
 - C) 1 and 4

D) 2 and 3

Solution: (C)

The Accuracy (correct classification) is $(50+100)/165$ which is nearly equal to 0.91.

The true Positive Rate is how many times you are predicting positive class correctly so true positive rate would be $100/105 = 0.95$ also known as "Sensitivity" or "Recall"

37) For which of the following hyperparameters, higher value is better for decision tree algorithm?

1. Number of samples used for split
2. Depth of tree
3. Samples for leaf

A) 1 and 2

B) 2 and 3

C) 1 and 3

D) 1, 2 and 3

E) Can't say

Solution: (E)

For all three options A, B and C, it is not necessary that if you increase the value of parameter the performance may increase. For example, if we have a very high value of depth of tree, the resulting tree may overfit the data, and would not generalize well. On the other hand, if we have a very low value, the tree may underfit the data. So, we can't say for sure that "higher is better".

Context 38-39

Imagine, you have a $28 * 28$ image and you run a $3 * 3$ convolution neural network on it with the input depth of 3 and output depth of 8.

Note: Stride is 1 and you are using same padding.

38) What is the dimension of output feature map when you are using the given parameters.

- A) 28 width, 28 height and 8 depth
- B) 13 width, 13 height and 8 depth
- C) 28 width, 13 height and 8 depth
- D) 13 width, 28 height and 8 depth

Solution: (A)

The formula for calculating output size is

$$\text{output size} = (N - F)/S + 1$$

where, N is input size, F is filter size and S is stride.

Read this [article](#) to get a better understanding.

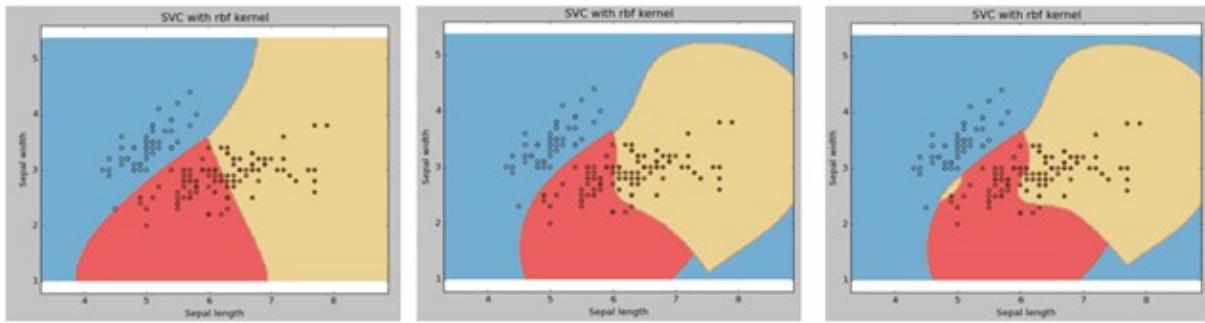
39) What is the dimensions of output feature map when you are using following parameters.

- A) 28 width, 28 height and 8 depth
- B) 13 width, 13 height and 8 depth
- C) 28 width, 13 height and 8 depth
- D) 13 width, 28 height and 8 depth

Solution: (B)

Same as above

40) Suppose, we were plotting the visualization for different values of C (Penalty parameter) in SVM algorithm. Due to some reason, we forgot to tag the C values with visualizations. In that case, which of the following option best explains the C values for the images below (1,2,3 left to right, so C values are C1 for image1, C2 for image2 and C3 for image3) in case of rbf kernel.



- A) $C_1 = C_2 = C_3$
- B) $C_1 > C_2 > C_3$
- C) $C_1 < C_2 < C_3$
- D) None of these

Solution: (C)

MCQ questions for unit 4: Naïve Bayes and Support Vector Machine

1. 1. How many terms are required for building a bayes model?

- a) 1
- b) 2
- c) 3

d) 4 Answer: c

Explanation: The three required terms are a conditional probability and two unconditional probability.

2. 2. What is needed to make probabilistic systems feasible in the world?

- a) Reliability
- b) Crucial robustness
- c) Feasibility

d) None of the mentioned Answer: b

Explanation: On a model-based knowledge provides the crucial robustness needed to make probabilistic system feasible in the real world.

3. 3. Where does the bayes rule can be used?

- a) Solving queries
- b) Increasing complexity
- c) Decreasing complexity

d) Answering probabilistic query Answer: d

Explanation: Bayes rule can be used to answer the probabilistic queries conditioned on one piece of evidence.

4. 4. What does the bayesian network provides?

- a) Complete description of the domain
- b) Partial description of the domain
- c) Complete description of the problem

d) None of the mentioned Answer: a

Explanation: A Bayesian network provides a complete description of the domain.

5. 5. How the entries in the full joint probability distribution can be calculated?

- a) Using variables
- b) Using information

c) Both Using variables & information

d) None of the mentioned Answer: b

Explanation: Every entry in the full joint probability distribution can be calculated from the information in the network

6. 6. How the bayesian network can be used to answer any query?

- a) Full distribution
- b) Joint distribution

c) Partial distribution

d) All of the mentioned Answer: b

Explanation: If a bayesian network is a representation of the joint distribution, then it can solve any query, by summing all the relevant joint entries.

7. 7. How the compactness of the bayesian network can be described?

a) Locally structured

b) Fully structured

c) Partial structure

d) All of the mentioned Answer: a

Explanation: The compactness of the bayesian network is an example of a very general property of a locally structured system.

8. 8. To which does the local structure is associated?

a) Hybrid

b) Dependant

c) Linear

d) None of the mentioned Answer: c

Explanation: Local structure is usually associated with linear rather than exponential growth in complexity.

9. 9. Which condition is used to influence a variable directly by all the others?

a) Partially connected

b) Fully connected

c) Local connected

d) None of the mentioned Answer: b

Explanation: None.

10. 10. What is the consequence between a node and its predecessors while creating bayesian network?

a) Functionally dependent

b) Dependant

c) Conditionally independent

d) Both Conditionally dependant & Dependant Answer: c

Explanation: The semantics to derive a method for constructing bayesian networks were led to the consequence that a node can be conditionally independent of its predecessors.

11. What do you mean by generalization error in terms of the SVM?

12. A) How far the hyperplane is from the support vectors

B) How accurately the SVM can predict outcomes for unseen data

C) The threshold amount of error in an SVM

Solution: B

Generalisation error in statistics is generally the out-of-sample error which is the measure of how accurately a model can predict values for previously unseen data.

13. The minimum time complexity for training an SVM is $O(n^2)$. According to this fact, what sizes of datasets are not best suited for SVM's?

- A) Large datasets
- B) Small datasets
- C) Medium sized datasets
- D) Size does not matter

Solution: A

Datasets which have a clear classification boundary will function best with SVM's.

14. The effectiveness of an SVM depends upon:

- A) Selection of Kernel
- B) Kernel Parameters
- C) Soft Margin Parameter C
- D) All of the above

Solution: D

The SVM effectiveness depends upon how you choose the basic 3 requirements mentioned above in such a way that it maximises your efficiency, reduces error and overfitting.

14. The SVM's are less effective when:

- A) The data is linearly separable
- B) The data is clean and ready to use
- C) The data is noisy and contains overlapping points

Solution: C

When the data has noise and overlapping points, there is a problem in drawing a clear hyperplane without misclassifying.

15. Suppose you are using RBF kernel in SVM with high Gamma value. What does this signify?

- A) The model would consider even far away points from hyperplane for modeling
- B) The model would consider only the points close to the hyperplane for modeling
- C) The model would not be affected by distance of points from hyperplane for modeling
- D) None of the above

Solution: B

The gamma parameter in SVM tuning signifies the influence of points either near or far away from the hyperplane.

For a low gamma, the model will be too constrained and include all points of the training dataset, without really capturing the shape.

For a higher gamma, the model will capture the shape of the dataset well.

16. The cost parameter in the SVM means:

- A) The number of cross-validations to be made
- B) The kernel to be used
- C) The tradeoff between misclassification and simplicity of the model
- D) None of the above

Solution: C

The cost parameter decides how much an SVM should be allowed to “bend” with the data. For a low cost, you aim for a smooth decision surface and for a higher cost, you aim to classify more points correctly. It is also simply referred to as the cost of misclassification.

17. Which of the following are real world applications of the SVM?

- A) Text and Hypertext Categorization
- B) Image Classification
- C) Clustering of News Articles
- D) All of the above

Solution: D

SVM's are highly versatile models that can be used for practically all real world problems ranging from regression to clustering and handwriting recognitions.

18. We usually use feature normalization before using the Gaussian kernel in SVM. What is true about feature normalization?

- 1. We do feature normalization so that new feature will dominate other
- 2. Some times, feature normalization is not feasible in case of categorical variables
- 3. Feature normalization always helps when we use Gaussian kernel in SVM

- A) 1
- B) 1 and 2
- C) 1 and 3
- D) 2 and 3

Solution: B

Statements one and two are correct.

19. What is/are true about kernel in SVM?

1. Kernel function map low dimensional data to high dimensional space
2. It's a similarity function

- A) 1
- B) 2
- C) 1 and 2
- D) None of these

Solution: C

Both the given statements are correct.

1. Which of the following is a widely used and effective machine learning algorithm based on the idea of bagging?
 - a. Decision Tree
 - b. Regression
 - c. Classification
 - d. Random Forest - answer
2. To find the minimum or the maximum of a function, we set the gradient to zero because:
 - a. The value of the gradient at extrema of a function is always zero - answer
 - b. Depends on the type of problem
 - c. Both A and B
 - d. None of the above
3. The most widely used metrics and tools to assess a classification model are:
 - a. Confusion matrix
 - b. Cost-sensitive accuracy
 - c. Area under the ROC curve
 - d. All of the above - answer
4. Which of the following is a good test dataset characteristic?
 - a. Large enough to yield meaningful results
 - b. Is representative of the dataset as a whole
 - c. Both A and B - answer
 - d. None of the above
5. Which of the following is a disadvantage of decision trees?
 - a. Factor analysis
 - b. Decision trees are robust to outliers
 - c. Decision trees are prone to be overfit - answer
 - d. None of the above
6. How do you handle missing or corrupted data in a dataset?
 - a. Drop missing rows or columns
 - b. Replace missing values with mean/median/mode
 - c. Assign a unique category to missing values

- d. All of the above - answer
7. What is the purpose of performing cross-validation?
- To assess the predictive performance of the models
 - To judge how the trained model performs outside the sample on test data
 - Both A and B - answer
8. Why is second order differencing in time series needed?
- To remove stationarity
 - To find the maxima or minima at the local point
 - Both A and B - answer
 - None of the above
9. When performing regression or classification, which of the following is the correct way to preprocess the data?
- Normalize the data → PCA → training - answer
 - PCA → normalize PCA output → training
 - Normalize the data → PCA → normalize PCA output → training
 - None of the above
10. Which of the following is an example of feature extraction?
- Constructing bag of words vector from an email
 - Applying PCA projects to a large high-dimensional data
 - Removing stopwords in a sentence
 - All of the above - answer
11. What is pca.components_ in Sklearn?
- Set of all eigen vectors for the projection space - answer
 - Matrix of principal components
 - Result of the multiplication matrix
 - None of the above options
12. Which of the following is true about Naive Bayes ?
- Assumes that all the features in a dataset are equally important
 - Assumes that all the features in a dataset are independent
 - Both A and B - answer
 - None of the above options
13. Which of the following statements about regularization is not correct?
- Using too large a value of lambda can cause your hypothesis to underfit the data.
 - Using too large a value of lambda can cause your hypothesis to overfit the data.
 - Using a very large value of lambda cannot hurt the performance of your hypothesis.
 - None of the above - answer
14. How can you prevent a clustering algorithm from getting stuck in bad local optima?
- Set the same seed value for each run
 - Use multiple random initializations - answer
 - Both A and B
 - None of the above
15. Which of the following techniques can be used for normalization in text mining?
- Stemming
 - Lemmatization
 - Stop Word Removal
 - Both A and B - answer

16. In which of the following cases will K-means clustering fail to give good results? 1) Data points with outliers 2) Data points with different densities 3) Data points with nonconvex shapes
- 1 and 2
 - 2 and 3
 - 1, 2, and 3 - answer
 - 1 and 3
17. Which of the following is a reasonable way to select the number of principal components "k"?
- Choose k to be the smallest value so that at least 99% of the variance is retained. - answer
 - Choose k to be 99% of m ($k = 0.99*m$, rounded to the nearest integer).
 - Choose k to be the largest value so that 99% of the variance is retained.
 - Use the elbow method
18. You run gradient descent for 15 iterations with $\alpha=0.3$ and compute $J(\theta)$ after each iteration. You find that the value of $J(\theta)$ decreases quickly and then levels off. Based on this, which of the following conclusions seems most plausible?
- Rather than using the current value of α , use a larger value of α (say $\alpha=1.0$)
 - Rather than using the current value of α , use a smaller value of α (say $\alpha=0.1$)
 - $\alpha=0.3$ is an effective choice of learning rate- answer
 - None of the above
19. What is a sentence parser typically used for?
- It is used to parse sentences to check if they are utf-8 compliant.
 - It is used to parse sentences to derive their most likely syntax tree structures. - answer
 - It is used to parse sentences to assign POS tags to all tokens.
 - It is used to check if sentences can be parsed into meaningful tokens.
20. Suppose you have trained a logistic regression classifier and it outputs a new example x with a prediction $h_0(x) = 0.2$. This means
- Our estimate for $P(y=1 | x)$
 - Our estimate for $P(y=0 | x)$ - answer
 - Our estimate for $P(y=1 | x)$
 - Our estimate for $P(y=0 | x)$
- 1) If you remove the following any one red points from the data. Does the decision boundary will change?
- A) Yes
B) No
- Solution: A
- These three examples are positioned such that removing any one of them introduces slack in the constraints. So the decision boundary would completely change.
21. [True or False] If you remove the non-red circled points from the data, the decision boundary will change?
- A) True
B) False

Solution: B

On the other hand, rest of the points in the data won't affect the decision boundary much.

22. What do you mean by generalization error in terms of the SVM?

- A) How far the hyperplane is from the support vectors
- B) How accurately the SVM can predict outcomes for unseen data**
- C) The threshold amount of error in an SVM

Solution: B

Generalisation error in statistics is generally the out-of-sample error which is the measure of how accurately a model can predict values for previously unseen data.

23. When the C parameter is set to infinite, which of the following holds true?

- A) The optimal hyperplane if exists, will be the one that completely separates the data**
- B) The soft-margin classifier will separate the data
- C) None of the above

Solution: A

At such a high level of misclassification penalty, soft margin will not hold existence as there will be no room for error.

24. What do you mean by a hard margin?

- A) The SVM allows very low error in classification**
- B) The SVM allows high amount of error in classification
- C) None of the above

Solution: A

A hard margin means that an SVM is very rigid in classification and tries to work extremely well in the training set, causing overfitting.

25. The minimum time complexity for training an SVM is $O(n^2)$. According to this fact, what sizes of datasets are not best suited for SVM's?

- A) Large datasets**
- B) Small datasets
- C) Medium sized datasets
- D) Size does not matter

Solution: A

Datasets which have a clear classification boundary will function best with SVM's.

26. The effectiveness of an SVM depends upon:

- A) Selection of Kernel
- B) Kernel Parameters
- C) Soft Margin Parameter C
- D) All of the above**

Solution: D

The SVM effectiveness depends upon how you choose the basic 3 requirements mentioned above in such a way that it maximises your efficiency, reduces error and overfitting.

27. Support vectors are the data points that lie closest to the decision surface.

A) TRUE

B) FALSE

Solution: A

They are the points closest to the hyperplane and the hardest ones to classify. They also have a direct bearing on the location of the decision surface.

28. The SVM's are less effective when:

A) The data is linearly separable

B) The data is clean and ready to use

C) The data is noisy and contains overlapping points

Solution: C

When the data has noise and overlapping points, there is a problem in drawing a clear hyperplane without misclassifying.

29. Suppose you are using RBF kernel in SVM with high Gamma value. What does this signify?

A) The model would consider even far away points from hyperplane for modeling

B) The model would consider only the points close to the hyperplane for modeling

C) The model would not be affected by distance of points from hyperplane for modeling

D) None of the above

Solution: B

The gamma parameter in SVM tuning signifies the influence of points either near or far away from the hyperplane

For a low gamma, the model will be too constrained and include all points of the training dataset, without really capturing the shape.

For a higher gamma, the model will capture the shape of the dataset well.

30. The cost parameter in the SVM means:

A) The number of cross-validations to be made

B) The kernel to be used

C) The tradeoff between misclassification and simplicity of the model

D) None of the above

Solution: C

The cost parameter decides how much an SVM should be allowed to “bend” with the data. For a low cost, you aim for a smooth decision surface and for a higher cost, you aim to classify more points correctly. It is also simply referred to as the cost of misclassification.

31. Suppose you are building a SVM model on data X. The data X can be error prone which means that you should not trust any specific data point too much. Now think that you want to build a SVM model which has quadratic kernel function of polynomial degree 2 that uses Slack variable C as one of its hyper parameter. Based upon that give the answer for following question.

What would happen when you use very large value of C(C->infinity)?

Note: For small C was also classifying all data points correctly

- A) We can still classify data correctly for given setting of hyper parameter C
- B) We can not classify data correctly for given setting of hyper parameter C
- C) Can't Say
- D) None of these

Solution: A

For large values of C, the penalty for misclassifying points is very high, so the decision boundary will perfectly separate the data if possible.

32. What would happen when you use very small C (C~0)?

- A) Misclassification would happen
- B) Data will be correctly classified
- C) Can't say
- D) None of these

Solution: A

The classifier can maximize the margin between most of the points, while misclassifying a few points, because the penalty is so low.

33. If I am using all features of my dataset and I achieve 100% accuracy on my training set, but ~70% on validation set, what should I look out for?

- A) Underfitting
- B) Nothing, the model is perfect
- C) Overfitting

Solution: C

If we're achieving 100% training accuracy very easily, we need to check to verify if we're overfitting our data.

34. Which of the following are real world applications of the SVM?

- A) Text and Hypertext Categorization
- B) Image Classification
- C) Clustering of News Articles
- D) All of the above

Solution: D

SVM's are highly versatile models that can be used for practically all real world problems ranging from regression to clustering and handwriting recognitions.

Question Context: 16 – 18

Suppose you have trained an SVM with linear decision boundary after training SVM, you correctly infer that your SVM model is under fitting.

35. Which of the following option would you more likely to consider iterating SVM next time?

- A) You want to increase your data points
- B) You want to decrease your data points
- C) You will try to calculate more variables

D) You will try to reduce the features

Solution: C

The best option here would be to create more features for the model.

36. Suppose you gave the correct answer in previous question. What do you think that is actually happening?

1. We are lowering the bias
2. We are lowering the variance
3. We are increasing the bias
4. We are increasing the variance

A) 1 and 2

B) 2 and 3

C) 1 and 4

D) 2 and 4

Solution: C

Better model will lower the bias and increase the variance

37. In above question suppose you want to change one of it's(SVM) hyperparameter so that effect would be same as previous questions i.e model will not under fit?

- A) We will increase the parameter C
- B) We will decrease the parameter C
- C) Changing in C don't effect
- D) None of these

Solution: A

Increasing C parameter would be the right thing to do here, as it will ensure regularized model

38. We usually use feature normalization before using the Gaussian kernel in SVM. What is true about feature normalization?

1. We do feature normalization so that new feature will dominate other
2. Some times, feature normalization is not feasible in case of categorical variables
3. Feature normalization always helps when we use Gaussian kernel in SVM

A) 1

B) 1 and 2

C) 1 and 3

D) 2 and 3

Solution: B

Statements one and two are correct.

Question Context: 20-22

Suppose you are dealing with 4 class classification problem and you want to train a SVM model on the data for that you are using One-vs-all method. Now answer the below questions?

39. How many times we need to train our SVM model in such case?

A) 1

- B) 2
- C) 3
- D) 4

Solution: D

For a 4 class problem, you would have to train the SVM at least 4 times if you are using a one-vs-all method.

40. Suppose you have same distribution of classes in the data. Now, say for training 1 time in one vs all setting the SVM is taking 10 second. How many seconds would it require to train one-vs-all method end to end?

- A) 20
- B) 40**
- C) 60
- D) 80

Solution: B

It would take $10 \times 4 = 40$ seconds

41. Suppose your problem has changed now. Now, data has only 2 classes. What would you think how many times we need to train SVM in such case?

- A) 1**
- B) 2
- C) 3
- D) 4

Solution: A

Training the SVM only one time would give you appropriate results

1. Support Vector Machine works well with,
 - a) Linear Scenarios
 - b) Non-linear Scenarios
 - c) Both of these
 - d) None of these

Answer: c) Both of these

2. Which of the following is best for MNIST dataset classification,
 - a) Naïve Bayes
 - b) Support Vector Machines
 - c) Random forest
 - d) Decision tree

Answer: b) Support Vector Machines

3. Two classes separated by a margin with two boundaries are called as,
 - a) Linear Vectors
 - b) Support Vectors
 - c) Test Vectors
 - d) None of these

Answer: b) Support Vectors

4. Scikit-learn supports which kernels,
 - a) Polynomial kernels
 - b) Sigmoid kernels
 - c) Custom kernels
 - d) All of these

Answer: d) All of these

5. Which of the following is the default kernel used in SVM,
 - a) Polynomial kernel
 - b) Sigmoid kernel
 - c) Custom kernel
 - d) Radial Basis Function

Answer: d) Radial Basis Function

6. The gamma parameter in RBF determines,
 - a) Amplitude of the function
 - b) Altitude of the function
 - c) Complexity of the function
 - d) None of these

Answer: a) Amplitude of the function

7. Scikit-learn allows us to create which kernel as a normal python function,
- Polynomial kernel
 - Custom kernel
 - Sigmoid kernel
 - All of these

Answer: b) Custom kernel

8. To find out a trade-off between precision and number of support vectors, scikit-learn provides an implementation called as,
- NuSVC
 - BuSVC
 - MuSVC
 - AuSVC

Answer: a) NuSVC

9. The RBF kernel is based on the function:

- $K(\bar{x}_i, \bar{x}_j) = e^{-\gamma |\bar{x}_i - \bar{x}_j|^2}$
- $K(\bar{x}_i, \bar{x}_j) = (\gamma \bar{x}_i^T \cdot \bar{x}_j + r)^c$
- $$K(\bar{x}_i, \bar{x}_j) = \frac{1 - e^{-2(\gamma \bar{x}_i^T \cdot \bar{x}_j + r)}}{1 + e^{-2(\gamma \bar{x}_i^T \cdot \bar{x}_j + r)}}$$
- None of these

Answer: a) $K(\bar{x}_i, \bar{x}_j) = e^{-\gamma |\bar{x}_i - \bar{x}_j|^2}$

10. The polynomial kernel is based on the function:

- $K(\bar{x}_i, \bar{x}_j) = e^{-\gamma |\bar{x}_i - \bar{x}_j|^2}$
- $K(\bar{x}_i, \bar{x}_j) = (\gamma \bar{x}_i^T \cdot \bar{x}_j + r)^c$
- $$K(\bar{x}_i, \bar{x}_j) = \frac{1 - e^{-2(\gamma \bar{x}_i^T \cdot \bar{x}_j + r)}}{1 + e^{-2(\gamma \bar{x}_i^T \cdot \bar{x}_j + r)}}$$
- None of these

Answer: b) $K(\bar{x}_i, \bar{x}_j) = (\gamma \bar{x}_i^T \cdot \bar{x}_j + r)^c$

11. The sigmoid kernel is based on this function:

- a) $K(\bar{x}_i, \bar{x}_j) = e^{-\gamma |\bar{x}_i - \bar{x}_j|^2}$
- b) $K(\bar{x}_i, \bar{x}_j) = (\gamma \bar{x}_i^T \cdot \bar{x}_j + r)^c$
- c) $K(\bar{x}_i, \bar{x}_j) = \frac{1 - e^{-2(\gamma \bar{x}_i^T \cdot \bar{x}_j + r)}}{1 + e^{-2(\gamma \bar{x}_i^T \cdot \bar{x}_j + r)}}$
- d) None of these

$$K(\bar{x}_i, \bar{x}_j) = \frac{1 - e^{-2(\gamma \bar{x}_i^T \cdot \bar{x}_j + r)}}{1 + e^{-2(\gamma \bar{x}_i^T \cdot \bar{x}_j + r)}}$$

Answer: c)

12. What is/are true about kernel in SVM,

- 1. It maps low dimensional data to high dimensional data.
- 2. It is a similarity function.
- a) 1
- b) 2
- c) Both 1 and 2
- d) None of these

Answer: c) Both 1 and 2

13. Which type of classifier is SVM,

- a) Discriminative
- b) Generative
- c) Both
- d) None of these

Answer: a) Discriminative

14. SVM is used to solve which type of problems,

- a) Classification
- b) Regression
- c) Clustering
- d) Both Classification and Regression

Answer: d) Both Classification and Regression

15. SVM is which type of learning algorithm,

- a) Supervised
- b) Unsupervised
- c) Both
- d) None of these

Answer: a) Supervised

16. The goal of SVM is to,

- a) Find the optimal separating hyperplane which minimizes the margin of training data.
- b) Find the optimal separating hyperplane which maximizes the margin of training data.
- c) Both
- d) None of these

Answer: b) Find the optimal separating hyperplane which maximizes the margin of training data.

17. The equation for hyperplane is,

a) $\bar{w}^T \bar{x} + b = 0$ where $\bar{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_m \end{pmatrix}$ and $\bar{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}$

b) $\bar{w}^T \bar{x} - b = 0$ where $\bar{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_m \end{pmatrix}$ and $\bar{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}$

c) $\bar{w}^T \bar{x} * b = 0$ where $\bar{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_m \end{pmatrix}$ and $\bar{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}$

- d) None of these

Answer: a) $\bar{w}^T \bar{x} + b = 0$ where $\bar{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_m \end{pmatrix}$ and $\bar{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}$

18. What is a kernel in SVM?

- a) SVM algorithms use a set of mathematical functions that are defined as the kernel
- b) SVM algorithms use a set of logarithmic functions that are defined as the kernel
- c) SVM algorithms use a set of exponential functions that are defined as the kernel
- d) SVM algorithms use a set of algebraic functions that are defined as the kernel

Answer: a) SVM algorithms use a set of mathematical functions that are defined as the kernel

19. Which of the following is false,

- a) SVM's are very good when we have no idea on the data.
- b) It works well with unstructured and semi structured data.
- c) The kernel trick is real strength of SVM.
- d) It scales relatively well to low dimensional data.

Answer: d) It scales relatively well to low dimensional data.

20. Which of the following is false,

- a) SVM algorithm is suitable for large data sets.
- b) It does not perform well when the data has more noise.
- c) SVM algorithm is not suitable for large data sets.
- d) None of these

Answer: a) SVM algorithm is suitable for large data sets.

1. The Naive Bayes Classifier is a _____ in probability.
A. Technique.
B. Process.
C. Classification.
D. None of these answers are correct.

ANSWER: D

2. How many terms are required for building a bayes model?
A. 1
B. 2
C. 3
D. 4

ANSWER: C

3. Where does the bayes rule can be used
A. Solving queries
B. Increasing complexity
C. Decreasing complexity
D. Answering probabilistic query

ANSWER: D

4. _____ is the mathematical likelihood that something will occur.
A. Classification
B. Probability
C. NAive Bayes Classifier
D. None

ANSWER: B

5. _____ binary distribution, useful when a feature can be present or absent
A. Bernoulli
B. multinomial
C. Gaussian
D. None

ANSWER: A

6. Naive Bayes Algorithm is a _____ learning algorithm.
A. Supervised
B. Reinforcement
C. Unsupervised
D. None of these

ANSWER: A

7. Examples of Naïve Bayes Algorithm is/are
A. Spam filtration
B. Sentimental analysis
C. Classifying articles
D. All of the above

ANSWER: D

8. Why it is needed to make probabilistic systems feasible in the world
- A. Feasibility
 - B. Reliability
 - C. Crucial robustness
 - D. None of the above

ANSWER: C

9. Probability provides a way of summarizing the _____ that comes from our laziness and ignorances.

- A. Belief
- B. Uncertainty
- C. Joint probability distributions
- D. Randomness

ANSWER: B

10. The entries in the full joint probability distribution can be calculated as

- A. Using variables
- B. Both Using variables & information
- C. Using information
- D. All of the above

ANSWER: C

11. Which of the following is correct about the Naïve Bayes?

- A. Assumes that all the features in a dataset are independent
- B. Assumes that all the features in a dataset are equally important
- C. None
- D. All of the above

ANSWER: C

12. Naïve Bayes algorithm is based on _____ and used for solving classification problems.

- A. Bayes Theorem
- B. Candidate elimination algorithm
- C. EM algorithm
- D. None of the above

ANSWER: A

13. Types of Naïve Bayes Model:

- A. Bernoulli
- B. multinomial
- C. Gaussian
- D. All of above

ANSWER: D

14. Disadvantages of Naïve Bayes Classifier

- A. Naive Bayes assumes that all features are independent or unrelated, so it cannot learn the relationship between features.
- B. It performs well in Multi-class predictions as compared to the other Algorithms.
- C. Naïve Bayes is one of the fast and easy ML algorithms to predict a class of datasets.
- D. It is the most popular choice for text classification problems.

15. The benefit of Naïve Bayes
- A. Naïve Bayes is one of the fast and easy ML algorithms to predict a class of datasets.
 - B. It is the most popular choice for text classification problems.
 - C. It can be used for Binary as well as Multi-class Classifications.
 - D. All of the above

ANSWER: D

16. How can SVM be classified
- A. It is a model trained using unsupervised learning. It can be used for classification and regression.
 - B. It is a model trained using unsupervised learning. It can be used for classification but not for regression.
 - C. It is a model trained using supervised learning. It can be used for classification and regression.
 - D. It is a model trained using unsupervised learning. It can be used for classification but not for regression.

ANSWER: C

17. What do you mean by a hard margin
- A. The SVM allows very low error in classification
 - B. The SVM allows high amount of error in classification
 - C. None of the above
 - D. All of above

ANSWER: A

18. The effectiveness of an SVM depends upon:
- A. Selection of Kernel
 - B. Kernel Parameters
 - C. Soft Margin Parameter C
 - D. All of the above

ANSWER: D

19. Support vectors are the data points that lie closest to the decision surface.
- A. TRUE
 - B. FALSE

ANSWER: A

20. The SVM's are less effective when:
- A. The data is linearly separable
 - B. The data is clean and ready to use
 - C. The data is noisy and contains overlapping points

ANSWER: C

21. The cost parameter in the SVM means:
- A. The number of cross-validations to be made
 - B. The kernel to be used
 - C. The tradeoff between misclassification and simplicity of the model

D. None of the above

ANSWER: C

22. Which of the following are real world applications of the SVM?

- A. Text and Hypertext Categorization
- B. Image Classification
- C. Clustering of News Articles
- D. All of the above

ANSWER:D

23. Gaussian naive Bayes is useful when working with continuous values whose probabilities can be modeled using a Gaussian distribution

- A. Bernoulli
- B. multinomial
- C. Gaussian
- D. All of above

ANSWER: C

24. A multinomial distribution is useful to model feature vectors where each value represents, the number of occurrences of a term or its relative frequency

- A. Bernoulli
- B. multinomial
- C. Gaussian
- D. All of above

ANSWER: B

25. Gaussian naive Bayes is limited due to

- A. Mean and variance
- B. Mean and Median
- C. Median and covariance
- D. Mean and standard deviation

ANSWER:A

26. The two classes are normally separated by a margin with two boundaries where a few elements lie. Those elements are called

- A. principal components
- B. support vectors
- C. factors
- D. None

ANSWER: B

27. What is/are true about kernel in SVM? 1. Kernel function map low dimensional data to high dimensional space. 2. It's a similarity function

- A. 1
- B. 2
- C. 1 and 2
- D. None of these

ANSWER: C

28. Support vector machine (SVM) is a _____ classifier

- A. Descriptive
- B. Generative

ANSWER: A

29. SVM is termed as _____ classifier

- A. Maximum margin
- B. Manimum margin

ANSWER: A

30. The training examples closest to the separating hyperplane are called as _____

- A. Training vector
- B. Testing Vector
- C. Support margin
- D. Support vector

ANSWER: D

31. Which of the following is a type of SVM?

- A. Maximum margin classifier
- B. Soft margin classifier
- C. Support vector regression
- D. All of the above

ANSWER: D

32. The goal of the SVM is to _____

- A. Find the optimal separating hyperplane which minimizes the margin of training data
- B. Find the optimal separating hyperplane which maximizes the margin of training data

ANSWER: B

33. When using R, which of the following package is used for SVM?

- A. b1072
- B. c1071
- C. d2012
- D. e1071

ANSWER: D

34. What are the different kernels functions in SVM ?

- A. Linear Kernel
- B. Polynomial kernel
- C. Radial basis kernel
- D. Sigmoid kernel
- E. ALL of the above

ANSWER: E

35. Which of the following might be valid reasons for preferring an SVM over a neural network?

- A. An SVM can automatically learn to apply a non-linear transformation on the input space; a neural net cannot.

- B. An SVM can effectively map the data to an infinite-dimensional space; a neural net cannot.
- C. An SVM should not get stuck in local minima, unlike a neural net.
- D. The transformed (basis function) representation constructed by an SVM is usually easier to visualise/interpret than for a neural net.

ANSWER: B, C

36. You are given a labeled binary classification data set with N data points and D features. Suppose that $N < D$. In training an SVM on this data set, which of the following kernels is likely to be most appropriate?

- A. Linear kernel
- B. Quadratic kernel
- C. Higher-order polynomial kernel
- D. RBF kernel

ANSWER: A

1. Techniques of feature engineering involve:
 - A. Clean dataset
 - B. Increase their signal-noise ratio
 - C. Reduce dimensionality
 - D. All of these

ANSWER: D

2. Correlated features provide additional pieces of information
 - A. True
 - B. False
 - C. None of these
 - D. All of these

ANSWER: B

3. Training set is used to test performance of system
 - A. True
 - B. False
 - C. None of these
 - D. All of these

ANSWER: B

4. The original dataset must be randomly shuffled before the split phase
 - A. avoid a correlation between consequent elements
 - B. avoid a sequencing between consequent elements
 - C. build a relation between consequent elements
 - D. None of these

ANSWER: A

5. NumPy RandomState generator or an integer seed is used to
 - A. randomize data
 - B. reproduce experiments
 - C. benchmark dataset
 - D. None of these

ANSWER: B

6. A good ratio of training and testing split is
 - A. 50–50
 - B. 60–40
 - C. 70–20
 - D. 80–20

ANSWER: D

7. If dataset has categorical data, then following action needs to be taken
 - A. encode categorical data
 - B. drop the column containing categorical data
 - C. Convert it to number
 - D. keep it as it is, no effect

ANSWER: A.

8. Which of the following is a categorical variable?

- A. Gender
- B. Object
- C. Number
- D. Alphabet

ANSWER: A

9. which adopts a dictionary-oriented approach, associating to each category label a progressive integer number, that is an index of an instance array called classes_:

- A. Hashing
- B. Dict method
- C. LabelEncoder
- D. None of these

ANSWER: C

10. Drawback of label encoder class is

- A. All labels are turned into binary numbers
- B. All labels are turned into sequential numbers
- C. All labels are turned into random numbers
- D. All of these

ANSWER: B

11. Label encoder class

- A. provides position of data
- B. does not concern about semantics
- C. preserve semantics
- D. all of the above

ANSWER: B

12. One-hot encoding method converts the data into binary

- A. True
- B. False
- C. None of these
- D. All of these

ANSWER: A

13. For converting labels we use

- A. Label encoder
- B. Label Binarizer
- C. One hot encoding
- D. One label encoding

ANSWER: B

14. In which method, each categorical label is first turned into a positive integer and then transformed into a vector where only one feature is 1 while all the others are 0

- A. LabelBinarizer
- B. Encoder
- C. One hot encoding
- D. None of these

ANSWER: A

15. Methods to manage categorical variable

- A. One hot encoding
- B. LabelEncoder
- C. LabelBinarizer
- D. All of these

ANSWER: D

16. DictVectorizer and FeatureHasher produce

- A. Sparse Matrices
- B. Inverse Matrices
- C. Identity Matrices
- D. None of these

ANSWER: A

17. While managing missing values, which method are available

- A. Removing the whole line
- B. Creating sub-model to predict those features
- C. Using an automatic strategy to input them according to the other known values
- D. All of these

ANSWER: D

18. Which option should be considered only when the dataset is quite large, the number of missing features is high, and any prediction could be risky

- A. Removing the whole line
- B. Creating sub-model to predict those features
- C. Using an automatic strategy to input them according to the other known values
- D. All of these

ANSWER: A

19. While managing missing values, which method is said as best choice

- A. Removing the whole line
- B. Creating sub-model to predict those features
- C. Using an automatic strategy to input them according to the other known values
- D. All of these

ANSWER: C

20. What are data preprocessing techniques to handle outliers
- A. Winsorize (cap at threshold).
 - B. Transform to reduce skew (using Box-Cox or similar).
 - C. Remove outliers if you're certain they are anomalies or measurement errors.
 - D. All of above

ANSWER: D

21. Which of the following model model include a backwards elimination feature selection routine?
- A. MCV
 - B. MARS
 - C. MCFS
 - D. All of the Mentioned

ANSWER: B

22. Which of the following is a categorical outcome
- A. RMSE
 - B. RSquared
 - C. Accuracy
 - D. All of the Mentioned

ANSWER: C

23. What are data preprocessing techniques to handle outliers
- A. Winsorize (cap at threshold).
 - B. Transform to reduce skew (using Box-Cox or similar).
 - C. Remove outliers if you're certain they are anomalies or measurement errors.
 - D. All of above

ANSWER: D

24. What are ways of reducing dimensionality
- A. Removing collinear features.
 - B. Performing PCA, ICA, or other forms of algorithmic dimensionality reduction.
 - C. Combining features with feature engineering.
 - D. All of above

ANSWER: D

25. If you split your data into train/test splits, is it still possible to overfit your model?
- A. True
 - B. False
 - C. None of these
 - D. All of these

ANSWER: A

26. How do you handle missing or corrupted data in a dataset
- A. Drop missing rows or columns
 - B. Replace missing values with mean/median/mode
 - C. Assign a unique category to missing values
 - D. All of the above

ANSWER: D

27. Techniques to perform Feature Scaling are
- A. Min-Max Normalization
 - B. Standardization
 - C. Both a and b
 - D. None of above

ANSWER: B

28. Technique to re-scales a feature or observation value with distribution value between 0 and 1 is known as
- A. Mean Normalization
 - B. Max Normalization
 - C. Mode Normalization
 - D. Min-Max Normalization

ANSWER: D

29. Feature Scaling is a technique to standardize the independent features present in the data in a fixed range
- A. True
 - B. False
 - C. None of these
 - D. All of these

ANSWER: A

30. To calculate the distance between centroid and data point which method is used
- A. Euclidean Distance
 - B. Manhattan Distance
 - C. Minkowski Distance
 - D. All of the above

ANSWER: D

31. Feature Scaling is a technique to standardize the independent features present in the data in a fixed range
- A. True
 - B. False
 - C. None of these

D. All of these

ANSWER: A

32. ----- performed during the data pre-processing to handle highly varying magnitudes or values or units

- A. Label encoding
- B. Feature Scaling
- C. Feature extraction
- D. Normalization

ANSWER: B

33. Techniques to perform feature scaling are

- A. standardization
- B. Min Normalization
- C. Max Normalization
- D. Minmax Normalization and standardization

ANSWER: D

34. Normalization is generally required when we are dealing with attributes on a different scale

- A. True
- B. False
- C. None of these
- D. All of these

ANSWER: A

35. Why do we perform feature scaling?

- A. The range of all features should be normalized so that each feature contributes approximately proportionately to the final distance
- B. Gradient descent converges much faster with feature scaling than without it.
- C. Both A and B
- D. None

ANSWER: C

36. What is the difference between normalization and scaling

- A. Normalization devides the value and scaling multiplies the value
- B. Normalization converts data in range and scaling multiplies data by weight
- C. Both are same
- D. In scaling we change the range of your data while in normalization we change the shape of the distribution of your data

ANSWER: D

37. For normalization, the maximum value and minimum value is

- A. 1 and 0
- B. 0 and 1
- C. 0 to 0
- D. 1 and 1

ANSWER: A

38. An unnormalized dataset with many features contains
- A. information which is proportional to the independence of all features and their variance
 - B. information which is inversely proportional to the independence of all features and their variance
 - C. information proportional to the mean of all features and their Covariance
 - D. None

ANSWER: A

39. A -----is a useful approach to remove all those elements whose contribution is under a predefined level
- A. correlation threshold
 - B. covariance threshold
 - C. variance threshold
 - D. None of these

ANSWER: C

40. Imagine, you have 1000 input features and 1 target feature in a machine learning problem. You have to select 100 most important features based on the relationship between input features and the target features. Do you think, this is an example of dimensionality reduction?
- A. Yes
 - B. No
 - C. None of these
 - D. All of these

ANSWER: A

41. When performing regression or classification, which of the following is the correct way to preprocess the data
- A. Normalize the data → PCA → training
 - B. PCA → normalize PCA output → training
 - C. Normalize the data → PCA → normalize PCA output → training
 - D. None of the above

ANSWER: A

42. What is `pca.components_` in Sklearn
- A. Set of all eigen vectors for the projection space
 - B. Matrix of principal components

C. Result of the multiplication matrix

D. None of the above options

ANSWER: A

43. Which of the following is a reasonable way to select the number of principal components "k"

- A. Choose k to be the smallest value so that at least 99% of the variance is retained.
- B. Choose k to be 99% of m ($k = 0.99*m$, rounded to the nearest integer).
- C. Choose k to be the largest value so that 99% of the variance is retained.
- D. Use the elbow method

ANSWER: A

44. Dimensionality reduction algorithms are one of the possible ways to reduce the computation time required to build a model.

- A. TRUE
- B. FALSE
- C. None of these
- D. All of these

ANSWER: A

45. When a dataset is made up of non-negative elements can we use non-negative matrix factorization (NNMF) instead of standard PCA

- A. Yes
- B. NO
- C. None of these
- D. All of these

ANSWER: A

46. PCA can be used for projecting and visualizing data in lower dimensions.

- A. TRUE
- B. FALSE
- C. None of these
- D. All of these

ANSWER: A

47. Which of the following is/are true about PCA? 1. PCA is an unsupervised method. 2. It searches for the directions that data have the largest variance 3. Maximum number of principal components \leq number of features 4. All principal components are orthogonal to each other

- A. 1 and 2
- B. 1 and 3
- C. 2 and 3
- D. All of these

ANSWER: D

48. ----- allows exploiting the natural sparsity of data while extracting principal components

- A. Standard PCA
- B. Kernal PCA
- C. Sparse PCA
- D. All of the above

ANSWER: C

49. ----- performs a PCA with non-linearly separable data sets

- A. Standard PCA
- B. Kernal PCA
- C. Sparse PCA
- D. All of the above

ANSWER: B

50. Dictionary learning is a technique which allows rebuilding a sample starting from a sparse dictionary of atoms

- A. True
- B. False
- C. None of these
- D. All of these

ANSWER: A

1. In practice, Line of best fit or regression line is found when _____

- a) Sum of residuals ($\sum(Y - h(X))$) is minimum
- b) Sum of the absolute value of residuals ($\sum|Y-h(X)|$) is maximum
- c) Sum of the square of residuals ($\sum(Y-h(X))^2$) is minimum
- d) Sum of the square of residuals ($\sum(Y-h(X))^2$) is maximum

[View Answer](#)

Answer: c

Explanation: Here we penalize higher error value much more as compared to the smaller one, such that there is a significant difference between making big errors and small errors, which makes it easy to differentiate and select the best fit line.

2. If Linear regression model perfectly first i.e., train error is zero, then

- a) Test error is also always zero
- b) Test error is non zero
- c) Couldn't comment on Test error
- d) Test error is equal to Train error

[View Answer](#)

Answer: c

Explanation: Test Error depends on the test data. If the Test data is an exact representation of train data then test error is always zero. But this may not be the case.

3. Which of the following metrics can be used for evaluating regression models?

- i) R Squared
 - ii) Adjusted R Squared
 - iii) F Statistics
 - iv) RMSE / MSE / MAE
- a) ii and iv
 - b) i and ii
 - c) ii, iii and iv
 - d) i, ii, iii and iv

[View Answer](#)

Answer: d

Explanation: These (R Squared, Adjusted R Squared, F Statistics, RMSE / MSE / MAE) are some metrics which you can use to evaluate your regression model.

4. How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?

- a) 1
- b) 2
- c) 3
- d) 4

[View Answer](#)

Answer: b

Explanation: In simple linear regression, there is one independent variable so 2 coefficients ($Y=a+bx+\text{error}$).

5. In a simple linear regression model (One independent variable), If we change the input variable by 1 unit. How much output variable will change?

- a) by 1
- b) no change
- c) by intercept
- d) by its slope

[View Answer](#)

Answer: d

Explanation: For linear regression $Y=a+bx+\text{error}$. If neglect error then $Y=a+bx$. If x increases by 1, then $Y = a+b(x+1)$ which implies $Y=a+bx+b$. So Y increases by its slope.

6. Function used for linear regression in R is _____

- a) lm(formula, data)
- b) lr(formula, data)
- c) lrm(formula, data)
- d) regression.linear(formula, data)

[View Answer](#)

Answer: a

Explanation: lm(formula, data) refers to a linear model in which formula is the object of the class "formula", representing the relation between variables. Now this formula is applied on the data to create a relationship model.

7. In syntax of linear model lm(formula, data, ...), data refers to _____

- a) Matrix
- b) Vector
- c) Array
- d) List

[View Answer](#)

Answer: b

Explanation: Formula is just a symbol to show the relationship and is applied on data which is a vector. In General, data.frame are used for data.

8. In the mathematical Equation of Linear Regression $Y = \beta_1 + \beta_2X + \epsilon$, (β_1, β_2) refers to _____

- a) (X-intercept, Slope)
- b) (Slope, X-Intercept)
- c) (Y-Intercept, Slope)
- d) (slope, Y-Intercept)

[View Answer](#)

Answer: c

9. True–False: Linear Regression is a supervised machine learning algorithm.

- A) TRUE
- B) FALSE

Solution: (A)

Yes, Linear regression is a supervised learning algorithm because it uses true labels for training. Supervised learning algorithm should have input variable (x) and an output variable (Y) for each example.

10. True–False: Linear Regression is mainly used for Regression.

- A) TRUE
- B) FALSE

Solution: (A)

Linear Regression has dependent variables that have continuous values.

11. True–False: It is possible to design a Linear regression algorithm using a neural network?

- A) TRUE
- B) FALSE

Solution: (A)

True. A Neural network can be used as a universal approximator, so it can definitely implement a linear regression algorithm.

12. Which of the following methods do we use to find the best fit line for data in Linear Regression?

- A) Least Square Error
- B) Maximum Likelihood
- C) Logarithmic Loss
- D) Both A and B

Solution: (A)

In linear regression, we try to minimize the least square errors of the model to identify the line of best fit.

13. Which of the following evaluation metrics can be used to evaluate a model while modeling a continuous output variable?

- A) AUC–ROC
- B) Accuracy
- C) Logloss
- D) Mean–Squared–Error

Solution: (D)

Since linear regression gives output as continuous values, so in such case we use mean squared error metric to evaluate the model performance. Remaining options are use in case of a classification problem.

14. True-False: Lasso Regularization can be used for variable selection in Linear Regression.

- A) TRUE
- B) FALSE

Solution: (A)

True, In case of lasso regression we apply absolute penalty which makes some of the coefficients zero.

15. Which of the following is true about Residuals ?

- A) Lower is better
- B) Higher is better
- C) A or B depend on the situation
- D) None of these

Solution: (A)

Residuals refer to the error values of the model. Therefore lower residuals are desired.

16. Suppose that we have N independent variables ($X_1, X_2 \dots X_n$) and dependent variable is Y. Now Imagine that you are applying linear regression by fitting the best fit line using least square error on this data.

You found that correlation coefficient for one of it's variable(Say X_1) with Y is -0.95.

Which of the following is true for X_1 ?

- A) Relation between the X_1 and Y is weak
- B) Relation between the X_1 and Y is strong
- C) Relation between the X_1 and Y is neutral
- D) Correlation can't judge the relationship

Solution: (B)

The absolute value of the correlation coefficient denotes the strength of the relationship. Since absolute correlation is very high it means that the relationship is strong between X_1 and Y.

17. Looking at above two characteristics, which of the following option is the correct for Pearson correlation between V1 and V2?

If you are given the two variables V1 and V2 and they are following below two characteristics.

1. If V1 increases then V2 also increases
2. If V1 decreases then V2 behavior is unknown
 - A) Pearson correlation will be close to 1
 - B) Pearson correlation will be close to -1
 - C) Pearson correlation will be close to 0
 - D) None of these

Solution: (D)

We cannot comment on the correlation coefficient by using only statement 1. We need to consider the both of these two statements. Consider V1 as x and V2 as $|x|$. The correlation coefficient would not be close to 1 in such a case.

18. Suppose Pearson correlation between V1 and V2 is zero. In such case, is it right to conclude that V1 and V2 do not have any relation between them?

- A) TRUE
- B) FALSE

Solution: (B)

Pearson correlation coefficient between 2 variables might be zero even when they have a relationship between them. If the correlation coefficient is zero, it just means that that they don't move together. We can take examples like $y=|x|$ or $y=x^2$.

19. Which of the following offsets, do we use in linear regression's least square line fit? Suppose horizontal axis is independent variable and vertical axis is dependent variable.

- A) Vertical offset
- B) Perpendicular offset
- C) Both, depending on the situation
- D) None of above

Solution: (A)

We always consider residuals as vertical offsets. We calculate the direct differences between actual value and the Y labels. Perpendicular offset are useful in case of PCA.

20. True- False: Overfitting is more likely when you have huge amount of data to train?

- A) TRUE
- B) FALSE

Solution: (B)

With a small training dataset, it's easier to find a hypothesis to fit the training data exactly i.e. overfitting.

21. We can also compute the coefficient of linear regression with the help of an analytical method called "Normal Equation". Which of the following is/are true about Normal Equation?

We don't have to choose the learning rate
It becomes slow when number of features is very large
There is no need to iterate

- A) 1 and 2
- B) 1 and 3
- C) 2 and 3
- D) 1, 2 and 3

Solution: (D)

Instead of gradient descent, Normal Equation can also be used to find coefficients. Refer this article for read more about normal equation.

22. Which of the following statement is true about sum of residuals of A and B?

Below graphs show two fitted regression lines (A & B) on randomly generated data. Now, I want to find the sum of residuals in both cases A and B.

Note:

Scale is same in both graphs for both axis.
X axis is independent variable and Y-axis is dependent variable.

- A) A has higher sum of residuals than B
- B) A has lower sum of residual than B
- C) Both have same sum of residuals
- D) None of these

Solution: (C)

Sum of residuals will always be zero, therefore both have same sum of residuals

23. Suppose you have fitted a complex regression model on a dataset. Now, you are using Ridge regression with penalty λ . Choose the option which describes bias in best manner.

- A) In case of very large λ ; bias is low
- B) In case of very large λ ; bias is high

- C) We can't say about bias
- D) None of these

Solution: (B)

24. If the penalty is very large it means model is less complex, therefore the bias would be high. What will happen when you apply very large penalty?

- A) Some of the coefficient will become absolute zero
- B) Some of the coefficient will approach zero but not absolute zero
- C) Both A and B depending on the situation
- D) None of these

Solution: (B)

In lasso some of the coefficient value become zero, but in case of Ridge, the coefficients become close to zero but not zero.

25. What will happen when you apply very large penalty in case of Lasso?

- A) Some of the coefficient will become zero
- B) Some of the coefficient will be approaching to zero but not absolute zero
- C) Both A and B depending on the situation
- D) None of these

Solution: (A)

As already discussed, lasso applies absolute penalty, so some of the coefficients will become zero.

26. Which of the following statement is true about outliers in Linear regression?

- A) Linear regression is sensitive to outliers
- B) Linear regression is not sensitive to outliers
- C) Can't say
- D) None of these

Solution: (A)

The slope of the regression line will change due to outliers in most of the cases. So Linear Regression is sensitive to outliers.

27. Suppose you plotted a scatter plot between the residuals and predicted values in linear regression and you found that there is a relationship between them. Which of the following conclusion do you make about this situation?

- A) Since there is a relationship means our model is not good

- B) Since there is a relationship means our model is good
- C) Can't say
- D) None of these

Solution: (A)

There should not be any relationship between predicted values and residuals. If there exists any relationship between them, it means that the model has not perfectly captured the information in the data.

Question Context 28-30:

Suppose that you have a dataset D1 and you design a linear regression model of degree 3 polynomial and you found that the training and testing error is "0" or in another terms it perfectly fits the data.

28. What will happen when you fit degree 4 polynomial in linear regression?
- A) There are high chances that degree 4 polynomial will over fit the data
 - B) There are high chances that degree 4 polynomial will under fit the data
 - C) Can't say
 - D) None of these

Solution: (A)

Since degree 4 will be more complex (overfit the data) than the degree 3 model so it will again perfectly fit the data. In such case training error will be zero but test error may not be zero.

29. What will happen when you fit degree 2 polynomial in linear regression?
- A) It is high chances that degree 2 polynomial will over fit the data
 - B) It is high chances that degree 2 polynomial will under fit the data
 - C) Can't say
 - D) None of these

Solution: (B)

If a degree 3 polynomial fits the data perfectly, it's highly likely that a simpler model (degree 2 polynomial) might under fit the data.

30. In terms of bias and variance. Which of the following is true when you fit degree 2 polynomial?

- A) Bias will be high, variance will be high
- B) Bias will be low, variance will be high
- C) Bias will be high, variance will be low

D) Bias will be low, variance will be low

Solution: (C)

Since a degree 2 polynomial will be less complex as compared to degree 3, the bias will be high and variance will be low.

Question Context 31:

Which of the following is true about below graphs (A, B, C left to right) between the cost function and Number of iterations?

31. Suppose l_1 , l_2 and l_3 are the three learning rates for A, B, C respectively. Which of the following is true about l_1 , l_2 and l_3 ?

- A) $l_2 < l_1 < l_3$
- B) $l_1 > l_2 > l_3$
- C) $l_1 = l_2 = l_3$
- D) None of these

Solution: (A)

In case of high learning rate, step will be high, the objective function will decrease quickly initially, but it will not find the global minima and objective function starts increasing after a few iterations.

In case of low learning rate, the step will be small. So the objective function will decrease slowly

Question Context 32- 33:

We have been given a dataset with n records in which we have input attribute as x and output attribute as y. Suppose we use a linear regression method to model this data. To test our linear regressor, we split the data in training set and test set randomly.

32. Now we increase the training set size gradually. As the training set size increases, what do you expect will happen with the mean training error?

- A) Increase
- B) Decrease
- C) Remain constant
- D) Can't Say

Solution: (D)

Training error may increase or decrease depending on the values that are used to fit the model. If the values used to train contain more outliers gradually, then the error might just increase.

33. What do you expect will happen with bias and variance as you increase the size of training data?

- A) Bias increases and Variance increases
- B) Bias decreases and Variance increases
- C) Bias decreases and Variance decreases
- D) Bias increases and Variance decreases
- E) Can't Say False

Solution: (D)

As we increase the size of the training data, the bias would increase while the variance would decrease.

Question Context 34:

Consider the following data where one input (X) and one output (Y) is given.

34. What would be the root mean square training error for this data if you run a Linear Regression model of the form ($Y = A_0 + A_1X$)?

- A) Less than 0
- B) Greater than zero
- C) Equal to 0
- D) None of these

Solution: (C)

We can perfectly fit the line on the following data so mean error will be zero.

Question Context 35-36:

Suppose you have been given the following scenario for training and validation error for Linear Regression.

Scenario	Learning Rate	Number of iterations	Training Error	Validation Error
1	0.1	1000	100	110
2	0.2	600	90	105
3	0.3	400	110	110
4	0.4	300	120	130
5	0.4	250	130	150

35. Which of the following scenario would give you the right hyper parameter?

- A) 1
- B) 2
- C) 3
- D) 4

Solution: (B)

Option B would be the better option because it leads to less training as well as validation error.

36. Suppose you got the tuned hyper parameters from the previous question. Now, Imagine you want to add a variable in variable space such that this added feature is important. Which of the following thing would you observe in such case?

- A) Training Error will decrease and Validation error will increase
- B) Training Error will increase and Validation error will increase
- C) Training Error will increase and Validation error will decrease
- D) Training Error will decrease and Validation error will decrease
- E) None of the above

Solution: (D)

If the added feature is important, the training and validation error would decrease.

Question Context 37-38:

Suppose, you got a situation where you find that your linear regression model is under fitting the data.

37. In such situation which of the following options would you consider?

- I will add more variables
- I will start introducing polynomial degree variables
- I will remove some variables

- A) 1 and 2
- B) 2 and 3
- C) 1 and 3
- D) 1, 2 and 3

Solution: (A)

In case of under fitting, you need to induce more variables in variable space or you can add some polynomial degree variables to make the model more complex to be able to fit the data better.

38. Now situation is same as written in previous question (under fitting). Which of following regularization algorithm would you prefer?

- A) L1
- B) L2
- C) Any
- D) None of these

Solution: (D)

39. Which of the following step / assumption in regression modeling impacts the trade-off between under-fitting and over-fitting the most.

- A. The polynomial degree
- B. Whether we learn the weights by matrix inversion or gradient descent
- C. The use of a constant-term

Solution: A

Choosing the right degree of polynomial plays a critical role in fit of regression. If we choose higher degree of polynomial, chances of overfit increase significantly.

40. Suppose you have the following data with one real-value input variable & one real-value output variable. What is leave-one out cross validation mean square error in case of linear regression ($Y = bX+c$)?

- A. 10/27
- B. 20/27
- C. 50/27
- D. 49/27

Solution: D

We need to calculate the residuals for each cross validation point. After fitting the line with 2 points and leaving 1 point for cross validation.

Leave one out cross validation mean square error = $(2^2 + (2/3)^2 + 1^2) / 3 = 49/27$

41. Which of the following is/ are true about "Maximum Likelihood estimate (MLE)"?

- MLE may not always exist
- MLE always exists
- If MLE exist, it (they) may not be unique

If MLE exist, it (they) must be unique

- A. 1 and 4
- B. 2 and 3
- C. 1 and 3
- D. 2 and 4

Solution: C

The MLE may not be a turning point i.e. may not be a point at which the first derivative of the likelihood (and log-likelihood) function vanishes.

* The MLE may not be unique.

42. Let's say, a "Linear regression" model perfectly fits the training data (train error is zero). Now, Which of the following statement is true?

- A. You will always have test error zero
- B. You can not have test error zero
- C. None of the above

Solution: C

Test error may be zero if there no noise in test data. In other words, it will be zero, if the test data is perfect representative of train data but not always.

43. In a linear regression problem, we are using "R-squared" to measure goodness-of-fit. We add a feature in linear regression model and retrain the same model.

Which of the following option is true?

- A. If R Squared increases, this variable is significant.
- B. If R Squared decreases, this variable is not significant.
- C. Individually R squared cannot tell about variable importance. We can't say anything about it right now.
- D. None of these.

Solution: C

"R squared" individually can't tell whether a variable is significant or not because each time when we add a feature, "R squared" can either increase or stay constant. But, it is not true in case of "Adjusted R squared" (increases when features found to be significant).

44. Which one of the statement is true regarding residuals in regression analysis?

- A. Mean of residuals is always zero
- B. Mean of residuals is always less than zero
- C. Mean of residuals is always greater than zero
- D. There is no such rule for residuals.

Solution: A

Sum of residual in regression is always zero. If the sum of residuals is zero, the 'Mean' will also be zero.

45. Which of the one is true about Heteroskedasticity?

- A. Linear Regression with varying error terms
- B. Linear Regression with constant error terms
- C. Linear Regression with zero error terms
- D. None of these

Solution: A

The presence of non-constant variance in the error terms results in heteroskedasticity. Generally, non-constant variance arises because of presence of outliers or extreme leverage values.

You can refer this article for more detail about regression analysis.

46. Which of the following indicates a fairly strong relationship between X and Y?

- A. Correlation coefficient = 0.9
- B. The p-value for the null hypothesis Beta coefficient = 0 is 0.0001
- C. The t-statistic for the null hypothesis Beta coefficient=0 is 30
- D. None of these

Solution: A

Correlation between variables is 0.9. It signifies that the relationship between variables is fairly strong.

On the other hand, p-value and t-statistics merely measure how strong is the evidence that there is non zero association. Even a weak effect can be extremely significant given enough data.

47. Which of the following assumptions do we make while deriving linear regression parameters?

- The true relationship between dependent y and predictor x is linear
- The model errors are statistically independent
- The errors are normally distributed with a 0 mean and constant standard deviation
- The predictor x is non-stochastic and is measured error-free

A. 1, 2 and 3.

B. 1, 3 and 4.

C. 1 and 3.

D. All of above.

Solution: D

When deriving regression parameters, we make all the four assumptions mentioned above. If any of the assumptions is violated, the model would be misleading.

48. To test linear relationship of y(dependent) and x(independent) continuous variables, which of the following plot best suited?

A. Scatter plot

B. Barchart

C. Histograms

D. None of these

Solution: A

To test the linear relationship between continuous variables Scatter plot is a good option. We can find out how one variable is changing w.r.t. another variable. A scatter plot displays the relationship between two quantitative variables.

49. Generally, which of the following method(s) is used for predicting continuous dependent variable?

- Linear Regression
- Logistic Regression

- A. 1 and 2
- B. only 1
- C. only 2
- D. None of these.

Solution: B

Logistic Regression is used for classification problems. Regression term is misleading here.

50. A correlation between age and health of a person found to be -1.09 . On the basis of this you would tell the doctors that:

- A. The age is good predictor of health
- B. The age is poor predictor of health
- C. None of these

Solution: C

Correlation coefficient range is between $[-1, 1]$. So -1.09 is not possible.

51. Which of the following offsets, do we use in case of least square line fit? Suppose horizontal axis is independent variable and vertical axis is dependent variable.

- A. Vertical offset
- B. Perpendicular offset
- C. Both but depend on situation
- D. None of above

Solution: A

We always consider residual as vertical offsets. Perpendicular offset are useful in case of PCA.

52. Suppose we have generated the data with help of polynomial regression of degree 3 (degree 3 will perfectly fit this data). Now consider below points and choose the option based on these points.

Simple Linear regression will have high bias and low variance
Simple Linear regression will have low bias and high variance
polynomial of degree 3 will have low bias and high variance
Polynomial of degree 3 will have low bias and Low variance

- A. Only 1
- B. 1 and 3
- C. 1 and 4
- D. 2 and 4

Solution: C

If we fit higher degree polynomial greater than 3, it will overfit the data because model will become more complex. If we fit the lower degree polynomial less than 3 which means that we have less complex model so in this case high bias and low variance. But in case of degree 3 polynomial it will have low bias and low variance.

53. Suppose you are training a linear regression model. Now consider these points.

Overfitting is more likely if we have less data
Overfitting is more likely when the hypothesis space is small

Which of the above statement(s) are correct?

- A. Both are False
- B. 1 is False and 2 is True
- C. 1 is True and 2 is False
- D. Both are True

Solution: C

1. With small training dataset, it's easier to find a hypothesis to fit the training data exactly i.e. overfitting.

2. We can see this from the bias-variance trade-off. When hypothesis space is small, it has higher bias and lower variance. So with a small hypothesis space, it's less likely to find a hypothesis to fit the data exactly i.e. underfitting.

54. Suppose we fit "Lasso Regression" to a data set, which has 100 features ($X_1, X_2 \dots X_{100}$). Now, we rescale one of these feature by multiplying with 10 (say that feature is X_1), and then refit Lasso regression with the same regularization parameter.

Now, which of the following option will be correct?

- A. It is more likely for X_1 to be excluded from the model
- B. It is more likely for X_1 to be included in the model
- C. Can't say
- D. None of these

Solution: B

Big feature values \Rightarrow smaller coefficients \Rightarrow less lasso penalty \Rightarrow more likely to have been kept

55. Which of the following is true about "Ridge" or "Lasso" regression methods in case of feature selection?

- A. Ridge regression uses subset selection of features
- B. Lasso regression uses subset selection of features
- C. Both use subset selection of features
- D. None of above

Solution: B

"Ridge regression" will use all predictors in final model whereas "Lasso regression" can be used for feature selection because coefficient values can be zero. For more detail click [here](#).

56. Which of the following statement(s) can be true post adding a variable in a linear regression model?

R-Squared and Adjusted R-squared both increase

R-Squared increases and Adjusted R-squared decreases

R-Squared decreases and Adjusted R-squared decreases

R-Squared decreases and Adjusted R-squared increases

- A. 1 and 2
- B. 1 and 3
- C. 2 and 4
- D. None of the above

Solution: A

Each time when you add a feature, R squared always either increase or stays constant, but it is not true in case of Adjusted R squared. If it increases, the feature would be significant.

57. The following visualization shows the fit of three different models (in blue line) on same training data. What can you conclude from these visualizations?

The training error in first model is higher when compared to second and third model.

The best model for this regression problem is the last (third) model, because it has minimum training error.

The second model is more robust than first and third because it will perform better on unseen data.

The third model is overfitting data as compared to first and second model.

All models will perform same because we have not seen the test data.

- A. 1 and 3
- B. 1 and 3
- C. 1, 3 and 4
- D. Only 5

Solution: C

The trend of the data looks like a quadratic trend over independent variable X. A higher degree (Right graph) polynomial might have a very high accuracy on the train population but is expected to fail badly on test dataset. But if you see in left graph we will have training error maximum because it under-fits the training data.

58. Which of the following metrics can be used for evaluating regression models?

R Squared

Adjusted R Squared

F Statistics

RMSE / MSE / MAE

- A. 2 and 4.
- B. 1 and 2.
- C. 2, 3 and 4.
- D. All of the above.

Solution: D

These (R Squared, Adjusted R Squared, F Statistics , RMSE / MSE / MAE) are some metrics which you can use to evaluate your regression model.

59. We can also compute the coefficient of linear regression with the help of an analytical method called "Normal Equation". Which of the following is/are true about "Normal Equation"?

We don't have to choose the learning rate
It becomes slow when number of features is very large
No need to iterate

- A. 1 and 2
- B. 1 and 3.
- C. 2 and 3.
- D. 1, 2 and 3.

Solution: D

Instead of gradient descent, Normal Equation can also be used to find coefficients. Refer this article for read more about normal equation.

60. The expected value of Y is a linear function of the X(X1, X2.... Xn) variables and regression line is defined as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

Which of the following statement(s) are true?

If X_i changes by an amount ΔX_i , holding other variables constant, then the expected value of Y changes by a proportional amount $\beta_i \Delta X_i$, for some constant β_i (which in general could be a positive or negative number).

The value of β_i is always the same, regardless of values of the other X 's.

The total effect of the X 's on the expected value of Y is the sum of their separate effects.

Note: Features are independent of each others(zero interaction).

- A. 1 and 2
- B. 1 and 3
- C. 2 and 3
- D. 1, 2 and 3

Solution: D

The expected value of Y is a linear function of the X variables. This means:

If X_i changes by an amount ΔX_i , holding other variables fixed, then the expected value of Y changes by a proportional amount $\beta_i \Delta X_i$, for some constant β_i (which in general could be a positive or negative number).

The value of β_i is always the same, regardless of values of the other X 's.

The total effect of the X 's on the expected value of Y is the sum of their separate effects.

The unexplained variations of Y are independent random variables (in particular, not "auto correlated" if the variables are time series)

They all have the same variance ("homoscedasticity").

They are normally distributed.

61. How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?

- A. 1
- B. 2
- C. Can't Say

Solution: B

In simple linear regression, there is one independent variable so 2 coefficients ($Y = a + bx$).

62. Below graphs show two fitted regression lines (A & B) on randomly generated data. Now, I want to find the sum of residuals in both cases A and B.

Note:

Scale is same in both graphs for both axis.

X axis is independent variable and Y-axis is dependent variable.

Which of the following statement is true about sum of residuals of A and B?

- A) A has higher than B
- B) A has lower than B
- C) Both have same
- D) None of these

Solution: C

Sum of residuals always zero.

63. If two variables are correlated, is it necessary that they have a linear relationship?

- A. Yes
- B. No

Solution: B

It is not necessary. They could have non linear relationship

64. Correlated variables can have zero correlation coefficient. True or False?

- A. True
- B. False

Solution: A

65. Suppose I applied a logistic regression model on data and got training accuracy X and testing accuracy Y. Now I want to add few new features in data. Select option(s) which are correct in such case.

Note: Consider remaining parameters are same.

Training accuracy always decreases.
Training accuracy always increases or remain same.
Testing accuracy always decreases
Testing accuracy always increases or remain same

- A. Only 2
- B. Only 1
- C. Only 3
- D. Only 4

Solution: A

Adding more features to model will always increase the training accuracy i.e. low bias.
But testing accuracy increases if feature is found to be significant.

66. The graph below represents a regression line predicting Y from X. The values on the graph shows the residuals for each predictions value. Use this information to compute the SSE.

- A. 3.02
- B. 0.75
- C. 1.01
- D. None of these

Solution: A

SSE is the sum of the squared errors of prediction, so $\text{SSE} = (-.2)^2 + (.4)^2 + (-.8)^2 + (1.3)^2 + (-.7)^2 = 3.02$

67. Height and weight are well known to be positively correlated. Ignoring the plot scales (the variables have been standardized), which of the two scatter plots (plot1, plot2) is more likely to be a plot showing the values of height (Var1 – X axis) and weight (Var2 – Y axis).

- A. Plot2
- B. Plot1
- C. Both
- D. Can't say

Solution: A

Plot 2 is definitely a better representation of the association between height and weight. As individuals get taller, they take up more volume, which leads to an increase in height, so a positive relationship is expected. The plot on the right has this positive relationship while the plot on the left shows a negative relationship.

68. Suppose the distribution of salaries in a company X has median \$35,000, and 25th and 75th percentiles are \$21,000 and \$53,000 respectively.

Would a person with Salary \$1 be considered an Outlier?

- A. Yes
- B. No
- C. More information is required
- D. None of these.

Solution: C

69. Which of the following option is true regarding "Regression" and "Correlation" ?

Note: y is dependent variable and x is independent variable.

- A. The relationship is symmetric between x and y in both.
- B. The relationship is not symmetric between x and y in both.
- C. The relationship is not symmetric between x and y in case of correlation but in case of regression it is symmetric.
- D. The relationship is symmetric between x and y in case of correlation but in case of regression it is not symmetric.

Solution: D

Correlation is a statistic metric that measures the linear association between two variables. It treats y and x symmetrically.

Regression is setup to predict y from x. The relationship is not symmetric.

70. Can we calculate the skewness of variables based on mean and median?

- A. True

B. False

Solution: B

The skewness is not directly related to the relationship between the mean and median.

71. Suppose you have n datasets with two continuous variables (y is dependent variable and x is independent variable). We have calculated summary statistics on these datasets. All of them give the following result:

Are all the given datasets same?

A. Yes

B. No

C. Can't Say

Solution: C

To answer this question, you should know about Anscombe's quartet. Refer this link [to read more about this.](#)

72. How does number of observations influence overfitting? Choose the correct answer(s).

Note: Rest all parameters are same

In case of fewer observations, it is easy to overfit the data.

In case of fewer observations, it is hard to overfit the data.

In case of more observations, it is easy to overfit the data.

In case of more observations, it is hard to overfit the data.

A. 1 and 4

B. 2 and 3

C. 1 and 3

D. None of theses

Solution: A

In particular, if we have very few observations and it's small, then our models can rapidly overfits data. Because we have only a few points and as we're increasing in our model complexity like the order of the polynomial, it becomes very easy to hit all of our observations.

On the other hand, if we have lots and lots of observations, even with really, really complex models, it is difficult to overfit because we have dense observations across our input.

73. Suppose you have fitted a complex regression model on a dataset. Now, you are using Ridge regression with tuning parameter lambda to reduce its complexity. Choose the option(s) below which describes relationship of bias and variance with lambda.

- A. In case of very large lambda; bias is low, variance is low
- B. In case of very large lambda; bias is low, variance is high
- C. In case of very large lambda; bias is high, variance is low
- D. In case of very large lambda; bias is high, variance is high

Solution: C

If lambda is very large it means model is less complex. So in this case bias is high and variance is low.

74. Suppose you have fitted a complex regression model on a dataset. Now, you are using Ridge regression with tuning parameter lambda to reduce its complexity. Choose the option(s) below which describes relationship of bias and variance with lambda.

- A. In case of very small lambda; bias is low, variance is low
- B. In case of very small lambda; bias is low, variance is high
- C. In case of very small lambda; bias is high, variance is low
- D. In case of very small lambda; bias is high, variance is high

Solution: B

If lambda is very small it means model is complex. So in this case bias is low and variance is high because model will overfit the data.

75. What is/are true about ridge regression?

When lambda is 0, model works like linear regression model

When lambda is 0, model doesn't work like linear regression model

When lambda goes to infinity, we get very, very small coefficients approaching 0

When lambda goes to infinity, we get very, very large coefficients approaching infinity

- A. 1 and 3

- B. 1 and 4
- C. 2 and 3
- D. 2 and 4

Solution: A

Specifically, we can see that when lambda is 0, we get our least square solution. When lambda goes to infinity, we get very, very small coefficients approaching 0.

76. Out of the three residual plots given below, which of the following represent worse model(s) compared to others?

Note:

All residuals are standardized.
The plots are between predicted values Vs. residuals

- A. 1
- B. 2
- C. 3
- D. 1 and 2

Solution: C

There should not be any relationship between predicted values and residuals. If there exist any relationship between them means model has not perfectly capture the information in data.

77. Which of the following method(s) does not have closed form solution for its coefficients?

- A. Ridge regression
- B. Lasso
- C. Both Ridge and Lasso
- D. None of both

Solution: B

The Lasso does not admit a closed-form solution. The L1-penalty makes the solution non-linear. So we need to approximate the solution.

78. Consider the following dataset

Which bold point, if removed will have the largest effect on fitted regression line as shown in above figure(dashed)?

- A) a
- B) b
- C) c
- D) d

Solution: D

Linear regression is sensitive to outliers in the data. Although c is also an outlier in given data space but it is closed to the regression line(residual is less) so it will not affect much.

79. In a simple linear regression model (One independent variable), If we change the input variable by 1 unit. How much output variable will change?

- A: By 1
- B. No change
- C. By intercept
- D. By its Slope

Solution: D

Equation for simple linear regression: $Y=a+bx$. Now if we increase the value of x by 1 then the value of y would be $a+b(x+1)$ i.e. value of y will get incremented by b .

80. Logistic Regression transforms the output probability to be in a range of [0, 1]. Which of the following function is used by logistic regression to convert the probability in the range between [0, 1].

- A. Sigmoid
- B. Mode
- C. Square

D. Probit

Solution: A

Sigmoid function is used to convert output probability between [0, 1] in logistic regression.

81: Which of the following statement is true about partial derivative of the cost functions w.r.t weights / coefficients in linear-regression and logistic-regression?

- A. Both will be different
- B. Both will be same
- C. Can't say
- D. None of these

Solution: B

82. Suppose, we are using Logistic regression model for n-class classification problem. In this case, we can use One-vs-rest method. Choose which of the following option is true regarding this?

- A. We need to fit n model in n-class classification problem.
- B. We need to fit n-1 models to classify into n classes.
- C. We need to fit only 1 model to classify into n classes.
- D. None of these.

Solution: A

If there are n classes, then n separate logistic regression has to fit, where the probability of each category is predicted over the rest of the categories combined.

Take a example of 3-class (-1, 0, 1) classification. Then need to train 3 logistic regression classifiers.

-1 vs 0 and 1
0 vs -1 and 1
1 vs 0 and -1

83. Below are two different logistic models with different values for β_0 and β_1 .

Which of the following statement(s) is true about β_0 and β_1 values of two logistics models (Green, Black)?

Note: consider $Y = \beta_0 + \beta_1*X$. Here, β_0 is intercept and β_1 is coefficient.

- A. β_1 for Green is greater than Black
- B. β_1 for Green is lower than Black
- C. β_1 for both models is same
- D. Can't Say.

Solution: B

Name of Faculty	Dr Roshani Raut							
Name of Subject	Machine Learning & Applications							
Year	BE							
Branch	IT							
Q.no	Description Question	Choice A	Choice B	Choice C	Choice D	Unit No	Difficulty Level (Easy-1/Medium-2/Hard-3)	Blooms Taxonomy Level
1	In multiclass classification number of classes must be	Less than two	Equals to two	Greater than two	option 1 and option 2	2	1	1
2	Application of machine learning methods to large databases is called	Data Mining.	Artificial Intelligence	Big Data Computing	Internet of Things	1	2	1
3	If machine learning model output involves target variable then that model is called as	Descriptive model	Predictive Model	Reinforcement Learning	All of the above	1	1	1
4	In what type of learning labelled training data is used	Unsupervised Learning	Supervised Learning	Reinforcement Learning	Active Learning	1	1	1
5	In following type of feature selection method we start with empty feature set	Forward Feature selection	Backword Feature selection	Both A and B	None of the above	1	1	1
6	In PCA the number of input dimensions are equal to principal components	True	FALSE			1	1	1
7	PCA can be used for projecting and visualizing data in lower dimensions.	True	FALSE			1	1	1
8	Which of the following is the best machine learning method?	Scalable	Accuracy	Fast	All of the above	1	1	1
9	What characterize unlabeled examples in machine learning	There is no prior knowledge	There is no confusing knowledge	There is prior knowledge	There is plenty of confusing knowledge	1	2	1
10	What does dimensionality reduction reduce?	stochastics	collinearity	performance	Entropy	1	1	1
11	Data used to build a data mining model.	Training data	Validation data	test data	hidden data	1	1	1
12	The problem of finding hidden structure in unlabeled data is called...	Supervised learning	Unsupervised learning	Reinforcement learning	None of the above	1	1	1
13	The difference between the actual Y value and the predicted Y value found using a regression equation is called the	slope	residual	outlier	scatter plot	3	1	1
14	Which of the following can only be used when training data are linearlyseparable?	Linear hard-margin SVM	Linear Logistic Regression	Linear Soft margin SVM	The centroid method	2	1	1
15	Impact of high variance on the training set ?	overfitting	underfitting	both underfitting & overfitting	Depends upon the dataset	2	1	1
16	What do you mean by a hard margin?	The SVM allows very low error in classification	The SVM allows high amount of error in classification	Both 1 & 2	None of the above	2	1	1
17	The effectiveness of an SVM depends upon:	Selection of Kernel	Kernel Parameters	Soft Margin Parameter C	All of the above	2	1	1
18	What is back propagation?	It is another name given to the curvy function in the perceptron	It is the transmission of error back through the network to adjust the inputs	It is the transmission of error back through the network to allow weights to be adjusted so that the network can learn	None of the mentioned	6	2	1

19	What are support vectors?	All the examples that have a non-zero weight w_k in a SVM	The only examples necessary to compute $f(x)$ in an SVM.	All of the above	None of the above	2	2	1
20	Neural networks	optimize a convex cost function	always output values between 0 and 1	can be used for regression as well as classification	All of the above	3	2	1
21	Of the Following Examples, Which would you address using an supervised learning Algorithm?	Given email labeled as Spam or not Spam, learn a spam filter	Given a set of news articles found on the web, group them into set of articles about the same story.	Given a database of customer data, automatically discover market segments and group customers into different market segments.	Find the patterns in Market Basket Analysis	1	1	2
22	Dimensionality Reduction Algorithms are one of the possible ways to reduce the computation time required to build a model	TRUE	FALSE			1	1	2
23	You are given reviews of few netflix series marked as positive, negative and neutral. Classifying reviews of a new netflix series is an example of	Supervised Learning	Unsupervised Learning	Semisupervised Learning	Reinforcement Learning	1	1	2
24	Which of the following is a good test dataset characteristic?	Large enough to yield meaningful results	Is representative of the dataset as a whole	Both A and B	None of the above	1	2	2
25	Following are the types of supervised learning	Classification	Regression	subgroup discovery	All of the above	1	1	2
26	Type of matrix decomposition model is	Descriptive model	Predictive Model	Logical model	None of the above	1	3	2
27	Following is powerful distance metrics used by Geometric model	Euclidean distance	Manhattan distance	Both A and B	square distance	1	2	2
28	The output of training process in machine learning is	machine learning model	machine learning algorithm	null	accuracy	1	1	2
29	A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college. Here feature type is	nominal	ordinal	categorical	boolean	1	3	2
30	PCA is	Forward Feature selection	Backward Feature selection	Feature Extraction	All of the above	1	1	2
31	Dimensionality reduction algorithms are one of the possible ways to reduce the computation time required to build a model.	True	FALSE			1	1	2
32	Which of the following techniques would perform better for reducing dimensions of a data set?	Removing columns which have too many missing values	Removing columns which have high variance in data	Removing columns with dissimilar data trends	None of these	1	3	2
33	Supervised learning and unsupervised clustering both require which is correct according to the statement.	output attribute.	hidden attribute.	input attribute.	categorical attribute	1	2	2
34	What characterize is hyperplane in geometrical model of machine learning?	A plane with 1 dimensional fewer than number of input attributes	A plane with 2 dimensional fewer than number of input attributes	A plane with 1 dimensional more than number of input attributes	A plane with 2 dimensional more than number of input attributes	1	2	2
35	Like the probabilistic view, the _____ view allows us to associate a probability of membership with each classification.	exemplar	deductive	classical	inductive	1	2	2
36	Database query is used to uncover this type of knowledge.	deep	hidden	shallow	multidimensional	1	2	2
37	A person trained to interact with a human expert in order to capture their knowledge.	knowledge programmer	knowledge developer	knowledge engineer	knowledge extractor	1	2	2
38	Some telecommunication company wants to segment their customers into distinct groups ,this is an example of	Supervised learning	Reinforcement learning	Unsupervised learning	Data extraction	1	2	2

39	In the example of predicting number of babies based on stork's population ,Number of babies is	outcome	feature	observation	attribute	1	3	2
40	Linear Regression is a _____ machine learning algorithm.	Supervised	Unsupervised	Semi-Supervised	Can't say	3	1	2
41	A perceptron adds up all the weighted inputs it receives, and if it exceeds a certain value, it outputs a 1, otherwise it just outputs a 0.	TRUE	False	Sometimes – it can also output intermediate values as well	Can't say	2	1	2
42	What is the purpose of the Kernel Trick?	To transform the data from nonlinearly separable to linearly separable	To transform the problem from regression to classification	To transform the problem from supervised to unsupervised learning.	All of the above	2	1	2
43	Which of the following can only be used when training data are linearlyseparable?	Linear hard-margin SVM	Linear Logistic Regression	Linear Soft margin SVM	Parzen windows	2	1	2
44	The firing rate of a neuron	determines how strongly the dendrites of the neuron stimulate axons of neighboring neurons	is more analogous to the output of a unit in a neural net than the output voltage of the neuron	only changes very slowly, taking a period of several seconds to make large adjustments	can sometimes exceed 30,000 action potentials per second	2	1	2
45	Which of the following methods/methods do we use to find the best fit line for data in Linear Regression?	Least Square Error	Maximum Likelihood	Logarithmic Loss	Both A and B	3	2	2
46	Which of the following methods do we use to best fit the data in Logistic Regression?	Least Square Error	Maximum Likelihood	Jaccard distance	Both A and B	3	2	2
47	Which of the following evaluation metrics can not be applied in case of logistic regression output to compare with target?	AUC-ROC	Accuracy	Logloss	Mean-Squared-Error	2	2	2
48	Which of the following is an application of NN (Neural Network)?	Sales forecasting	Data validation	Risk management	All of the mentioned	6	2	2
49	Neural Networks are complex _____ with many parameters.	Linear Functions	Nonlinear Functions	Discrete Functions	Exponential Functions	6	2	2
50	The cost parameter in the SVM means:	The number of cross-validations to be made	The kernel to be used	The tradeoff between misclassification and simplicity of the model	None of the above	2	2	2
51	Lasso can be interpreted as least-squares linear regression where	weights are regularized with the L1 norm	the weights have a Gaussian prior	weights are regularized with the L2 norm	the solution algorithm is simpler	3	2	2
52	The kernel trick	can be applied to every classification algorithm	is commonly used for dimensionality reduction	changes ridge regression so we solve a $d \times d$ linear system instead of an $n \times n$ system, given n sample points with d features	exploits the fact that in many learning algorithms, the weights can be written as a linear combination of input points	2	2	2
53	How does the bias-variance decomposition of a ridge regression estimator compare with that of ordinary least squares regression?	Ridge has larger bias, larger variance	Ridge has smaller bias, larger variance	Ridge has larger bias, smaller variance	Ridge has smaller bias, smaller variance	2	2	2
54	Which of the following evaluation metrics can be used to evaluate a model while modeling a continuous output variable?	AUC-ROC	Accuracy	Logloss	Mean-Squared-Error	3	3	2
55	What are tree based classifiers?	Classifiers which form a tree with each attribute at one level	Classifiers which perform series of condition checking with one attribute at a time	Both options except none	None of the options	4	1	2
56	What is gini index?	It is a type of index structure	It is a measure of purity	Both options except none	None of the options	4	1	2

57	Which of the following sentences are correct in reference to Information gain? a. It is biased towards single-valued attributes b. It is biased towards multi-valued attributes c. ID3 makes use of information gain d. The approach used by ID3 is greedy	a and b	a and d	b, c and d	All of the above	4	1	2
58	Multivariate split is where the partitioning of tuples is based on a combination of attributes rather than on a single attribute.	TRUE	FALSE			4	1	2
59	Gain ratio tends to prefer unbalanced splits in which one partition is much smaller than the other	TRUE	FALSE			4	1	2
60	The gini index is not biased towards multivalued attributes.	TRUE	FALSE			4	1	2
61	Gini index does not favour equal sized partitions.	TRUE	FALSE			4	1	2
62	When the number of classes is large Gini index is not a good choice.	TRUE	FALSE			4	1	2
63	Attribute selection measures are also known as splitting rules.	TRUE	FALSE			4	1	2
64	This clustering approach initially assumes that each data instance represents a single cluster.	expectation maximization	K-Means clustering	agglomerative clustering	conceptual clustering	4	1	2
65	Which statement is true about the K-Means algorithm?	The output attribute must be categorical	All attribute values must be categorical	All attributes must be numeric	Attribute values may be either categorical or numeric	4	1	2
66	The probability of a hypothesis before the presentation of evidence.	priori	posterior	conditional	subjective	5	1	2
67	KDD represents extraction of	data	knowledge	rules	model	4	1	2
68	The most general form of distance is	Manhattan	Eucledian	Mean	Minkowski	4	1	2
69	With Bayes theorem the probability of hypothesis H_i specified by $P(H_i)$ is referred to as	a conditional probability	an a priori probability	a bidirectional probability	a posterior probability	5	1	2
70	Simple regression assumes a _____ relationship between the input attribute and output attribute.	quadratic	inverse	linear	reciprocal	3	1	2
71	Which of the following algorithm comes under the classification	Apriori	Brute force	DBSCAN	K-nearest neighbor	4	1	2
72	Hierarchical agglomerative clustering is typically visualized as?	Dendrogram	Binary trees	Block diagram	Graph	4	1	2
73	The _____ step eliminates the extensions of $(k-1)$ -itemsets which are not found to be frequent, from being considered for counting support	Partitioning	Candidate generation	Itemset eliminations	Pruning	4	1	2
74	The distance between two points calculated using Pythagoras theorem is	Supremum distance	Eucledian distance	Linear distance	Manhattan Distance	4	1	2
75	Which learning Requires Self Assessment to identify patterns within data?	Unsupervised Learning	Supervised Learning	Semisupervised Learning	Reinforced Learning	1	1	3
76	Select the correct answers for following statements. 1. Filter methods are much faster compared to wrapper methods. 2. Wrapper methods use statistical methods for evaluation of a subset of features while Filter methods use cross validation.	Both are True	1 is True and 2 is False	Both are False	1 is False and 2 is True	1	2	3

77	The "curse of dimensionality" refers	All the problems that arise when working with data in the higher dimensions, that did not exist in the lower dimensions.	All the problems that arise when working with data in the lower dimensions, that did not exist in the higher dimensions.	All the problems that arise when working with data in the lower dimensions, that did not exist in the higher dimensions.	All the problems that arise when working with data in the higher dimensions, that did not exist in the higher dimensions.	1	2	3
78	In simple term, machine learning is	Training based on historical data	Prediction to answer a query	Both A and B	Automization of complex tasks	1	1	3
79	If machine learning model output doesnot involves target variable then that model is called as	Descriptive model	Predictive Model	Reinforcement Learning	All of the above	1	1	3
80	Following are the descriptive models	Clustering	Classification	Association rule	Both a and c	1	1	3
81	Different learning methods does not include?	Memorization	Analogy	Deduction	Introduction	1	3	3
82	A measurable property or parameter of the data-set is	training data	feature	test data	validation data	1	2	3
83	Feature can be used as a	Binary split	Predictor	Both A and B	None of the above	1	1	3
84	It is not necessary to have a target variable for applying dimensionality reduction algorithms	True	FALSE			1	1	3
85	The most popularly used dimensionality reduction algorithm is Principal Component Analysis (PCA). Which of the following is/are true about PCA? 1. PCA is an unsupervised method 2. It searches for the directions that data have the largest variance 3. Maximum number of principal components <= number of features 4. All principal components are orthogonal to each other	1 & 2	2 & 3	3 & 4	All of the above	1	3	3
86	Which of the following is a reasonable way to select the number of principal components "k"?	Choose k to be the smallest value so that at least 99% of the variance is retained. - answer	Choose k to be 99% of m (k = 0.99*m, rounded to the nearest integer).	Choose k to be the largest value so that 99% of the variance is retained.	Use the elbow method	1	3	3
87	Which of the following is an example of feature extraction?	Construction bag of words from an email	Applying PCA to project high dimensional data	Removing stop words	Forward selection	1	3	3
88	Prediction is	The result of application of specific theory or rule in a specific case	Discipline in statistics used to find projections in multidimensional data	Value entered in database by expert	Independent of data	1	2	3
89	You are given sesimic data and you want to predict next earthquake , this is an example of	Supervised learning	Reinforcement learning	Unsupervised learning	Dimensionality reduction	1	3	3
90	In the regression equation Y = 75.65 + 0.50X, the intercept is	0.5	75.65	1	indeterminable	3	1	3
91	The selling price of a house depends on many factors. For example, it depends on the number of bedrooms, number of kitchen, number of bathrooms, the year the house was built, and the square footage of the lot. Given these factors, predicting the selling price of the house is an example of task.	Binary Classification	Multilabel Classification	Simple Linear Regression	Multiple Linear Regression	3	1	3
92	Suppose, you got a situation where you find that your linear regression model is under fitting the data. In such situation which of the following options would you consider?	You will add more features	You will remove some features	All of the above	None of the above	3	1	3

93	Which of the following are real world applications of the SVM?	Text and Hypertext Categorization	Image Classification	Clustering of News Articles	All of the above	2	1	3
94	How can SVM be classified?	It is a model trained using unsupervised learning. It can be used for classification and regression.	It is a model trained using unsupervised learning. It can be used for classification but not for regression.	It is a model trained using supervised learning. It can be used for classification and regression.	t is a model trained using unsupervised learning. It can be used for classification but not for regression.	2	1	3
95	Which of the following can help to reduce overfitting in an SVM classifier?	Use of slack variables	High-degree polynomial features	Normalizing the data	Setting a very low learning rate	2	1	3
96	We have been given a dataset with n records in which we have input attribute as x and output attribute as y. Suppose we use a linear regression method to model this data. To test our linear regressor, we split the data in training set and test set randomly. Now we increase the training set size gradually. As the training set size increases, What do you expect will happen with the mean training error?	Increase	Decrease	Remain constant	Can't Say	3	2	3
97	We have been given a dataset with n records in which we have input attribute as x and output attribute as y. Suppose we use a linear regression method to model this data. To test our linear regressor, we split the data in training set and test set randomly. What do you expect will happen with bias and variance as you increase the size of training data?	Bias increases and Variance increases	Bias decreases and Variance increases	Bias decreases and Variance decreases	Bias increases and Variance decreases	3	2	3
98	Regarding bias and variance, which of the following statements are true? (Here 'high' and 'low' are relative to the ideal model. (i) Models which overfit are more likely to have high bias (ii) Models which overfit are more likely to have low bias (iii) Models which overfit are more likely to have high variance (iv) Models which overfit are more likely to have low variance	(i) and (ii)	(ii) and (iii)	(iii) and (iv)	None of these	3	2	3
99	Which of the following indicates the fundamental of least squares?	arithmetic mean should be maximized	arithmetic mean should be zero	arithmetic mean should be neutralized	arithmetic mean should be minimized	3	2	3
100	Having multiple perceptrons can actually solve the XOR problem satisfactorily: this is because each perceptron can partition off a linear part of the space itself, and they can then combine their results.	True – this works always, and these multiple perceptrons learn to classify even complex problems	False – perceptrons are mathematically incapable of solving linearly inseparable functions, no matter what you do	True – perceptrons can do this but are unable to learn to do it – they have to be explicitly hand-coded	False – just having a single perceptron is enough	6	2	3
101	Suppose you have trained an SVM with linear decision boundary after training SVM, you correctly infer that your SVM model is under fitting. Which of the following is best option would you more likely to consider iterating SVM next time?	You want to increase your data points	You want to decrease your data points	You will try to calculate more variables	You will try to reduce the features	2	2	3

102	What is/are true about kernel in SVM? 1. Kernel function map low dimensional data to high dimensional space 2. It's a similarity function	1	2	1 and 2	None of these	2	2	3
103	You trained a binary classifier model which gives very high accuracy on the training data, but much lower accuracy on validation data. Which is false.	This is an instance of overfitting	This is an instance of underfitting	The training was not well regularized	The training and testing examples are sampled from different distributions	2	2	3
104	Suppose that we have N independent variables ($X_1, X_2 \dots X_n$) and dependent variable is Y. Now imagine that you are applying linear regression by fitting the best fit line using least square error on this data. You found that correlation coefficient for one of its variable (Say X_1) with Y is 0.95.	Relation between the X_1 and Y is weak	Relation between the X_1 and Y is strong	Relation between the X_1 and Y is neutral	Correlation can't judge the relationship	3	3	3
105	In terms of bias and variance. Which of the following is true when you fit degree 2 polynomial?	Bias will be high, variance will be high	Bias will be low, variance will be high	Bias will be high, variance will be low	Bias will be low, variance will be low	3	3	3
106	Which of the following statements are true for a design matrix $X \in \mathbb{R}^{n \times d}$ with $d > n$? (The rows are n sample points and the columns represent d features.)	Least-squares linear regression computes the weights $w = (XTX)^{-1}XTy$	The sample points are linearly separable	X has exactly $d - n$ eigenvectors with eigenvalue zero	At least one principal component direction is orthogonal to a hyperplane that contains all the sample points	3	3	3
107	Suppose your model is demonstrating high variance across the different training sets. Which of the following is NOT valid way to try and reduce the variance?	Increase the amount of training data in each training set	Improve the optimization algorithm being used for error minimization.	Decrease the model complexity	Reduce the noise in the training data	2	3	3
108	Point out the wrong statement.	Regression through the origin yields an equivalent slope if you center the data first	Normalizing variables results in the slope being the correlation	Least squares is not an estimation tool	None of the mentioned	3	3	3
109	Suppose you are using RBF kernel in SVM with high Gamma value. What does this signify?	The model would consider even far away points from hyperplane for modeling	The model would consider only the points close to the hyperplane for modeling	The model would not be affected by distance of points from hyperplane for modeling	None of the above	2	3	3
110	We usually use feature normalization before using the Gaussian kernel in SVM. What is true about feature normalization? 1. We do feature normalization so that new feature will dominate other 2. Some times, feature normalization is not feasible in case of categorical variables 3. Feature normalization always helps when we use Gaussian kernel in SVM	1	1 and 2	1 and 3	2 and 3	2	3	3
111	Wrapper methods are hyper-parameter selection methods that	Should be used whenever possible because they are computationally efficient	Should be avoided unless there are no other options because they are always prone to overfitting.	Are useful mainly when the learning machines are "black boxes"	Should be avoided altogether.	2	3	3
112	Which of the following methods can not achieve zero training error on any linearly separable dataset?	Decision tree	15-nearest neighbors	Hard-margin SVM	Perceptron	2	3	3

113	Suppose we train a hard-margin linear SVM on $n > 100$ data points in R^2 , yielding a hyperplane with exactly 2 support vectors. If we add one more data point and retrain the classifier, what is the maximum possible number of support vectors for the new hyperplane (assuming the $n + 1$ points are linearly separable)?	2	3	n	$n+1$	2	3	3
114	Let S_1 and S_2 be the set of support vectors and w_1 and w_2 be the learnt weight vectors for a linearly separable problem using hard and soft margin linear SVMs respectively. Which of the following are correct?	$S_1 \subset S_2$	S_1 may not be a subset of S_2	$w_1 = w_2$	All of the above	2	3	3
115	Which one of these is not a tree based learner?	CART	ID3	Bayesian classifier	Random Forest	4	2	3
116	Which one of these is a tree based learner?	Rule based	Bayesian Belief Network	Bayesian classifier	Random Forest	4	2	3
117	What is the approach of basic algorithm for decision tree induction?	Greedy	Top Down	Procedural	Step by Step	4	2	3
118	Which of the following classifications would best suit the student performance classification systems?	If...then... Analysis	Market-basket analysis	Regression analysis	Cluster analysis	4	2	3
119	Given that we can select the same feature multiple times during the recursive partitioning of the input space, is it always possible to achieve 100% accuracy on the training data (given that we allow for trees to grow to their maximum size) when building decision trees?	Yes	No			4	2	3
120	This clustering algorithm terminates when mean values computed for the current iteration of the algorithm are identical to the computed mean values for the previous iteration	K-Means clustering	conceptual clustering	expectation maximization	agglomerative clustering	4	2	3
121	The number of iterations in apriori _____ Select one: a. b. c. d.	increases with the size of the data	decreases with the increase in size of the data	increases with the size of the maximum frequent set	decreases with increase in size of the maximum frequent set	4	2	3
122	Frequent item sets is	Superset of only closed frequent item sets	Superset of only maximal frequent item sets	Subset of maximal frequent item sets	Superset of both closed frequent item sets and maximal frequent item sets	4	2	3
123	A good clustering method will produce high quality clusters with	high inter class similarity	low intra class similarity	high intra class similarity	no inter class similarity	4	2	3
124	Which statement is true about neural network and linear regression models?	Both techniques build models whose output is determined by a linear sum of weighted input attribute values	The output of both models is a categorical attribute value	Both models require numeric attributes to range between 0 and 1	Both models require input attributes to be numeric	4	2	3
125	Which statement about outliers is true?	Outliers should be part of the training dataset but should not be present in the test data	Outliers should be identified and removed from a dataset	The nature of the problem determines how outliers are used	Outliers should be part of the test dataset but should not be present in the training data	2	2	3
126	Which Association Rule would you prefer	High support and medium confidence	High support and low confidence	Low support and high confidence	Low support and low confidence	4	2	3
127	In a Rule based classifier, If there is a rule for each combination of attribute values, what do you called that rule set R	Exhaustive	Inclusive	Comprehensive	Mutually exclusive	4	2	3
128	The apriori property means	If a set cannot pass a test, its supersets will also fail the same test	To decrease the efficiency, do level-wise generation of frequent item sets	To improve the efficiency, do level-wise generation of frequent item sets d.	If a set can pass a test, its supersets will fail the same test	4	2	3
129	If an item set 'XYZ' is a frequent item set, then all subsets of that frequent item set are	Undefined	Not frequent	Frequent	Can not say	4	2	3

130	Clustering is _____ and is example of learning	Predictive and supervised	Predictive and unsupervised	Descriptive and supervised	Descriptive and unsupervised	4	2	3
131	To determine association rules from frequent item sets	Only minimum confidence needed	Neither support nor confidence needed	Both minimum support and confidence are needed	Minimum support is needed	4	2	3
132	If {A,B,C,D} is a frequent itemset, candidate rules which is not possible is	C → A	D → ABCD	A → BC	B → ADC	4	2	3
133	Which Association Rule would you prefer	High support and low confidence	Low support and high confidence	Low support and low confidence	High support and medium confidence	4	2	3
134	The probability that a person owns a sports car given that they subscribe to automotive magazine is 40%. We also know that 3% of the adult population subscribes to automotive magazine. The probability of a person owning a sports car given that they don't subscribe to automotive magazine is 30%. Use this information to compute the probability that a person subscribes to automotive magazine given that they own a sports car	0.0398	0.0389	0.0368	0.0396	5	3	3
135	This clustering algorithm terminates when mean values computed for the current iteration of the algorithm are identical to the computed mean values for the previous iteration	conceptual clustering	K-Means clustering	expectation maximization	agglomerative clustering	4	2	3
136	Classification rules are extracted from	decision tree	root node	branches	siblings	4	2	3
137	What does K refers in the K-Means algorithm which is a non-hierarchical clustering approach?	Complexity	Fixed value	No of iterations	number of clusters	4	2	3
138	PCA works better if there is 1. A linear structure in the data 2. If the data lies on a curved surface and not on a flat surface 3. If variables are scaled in the same unit	1 and 2	2 and 3	1 and 3	1,2 and 3	1	3	4
139	If TP=9 FP=6 FN=26 TN=70 then Error rate will be	45 percentage	99 percentage	28 percentage	20 percentage	2	3	4
140	Imagine, you are solving a classification problems with highly imbalanced class. The majority class is observed 99% of times in the training data. Your model has 99% accuracy after taking the predictions on test data. Which of the following is true in such a case? 1. Accuracy metric is not a good idea for imbalanced class problems. 2. Accuracy metric is a good idea for imbalanced class problems. 3. Precision and recall metrics are good for imbalanced class problems. 4. Precision and recall metrics aren't good for imbalanced class problems.	1 and 3	1 and 4	2 and 3	2 and 4	2	3	4
141	he minimum time complexity for training an SVM is O(n ²). According to this fact, what sizes of datasets are not best suited for SVM's?	Large datasets	Small datasets	Medium sized datasets	Size does not matter	2	1	4
142	How will you counter over-fitting in decision tree?	By pruning the longer rules	By creating new rules	Both By pruning the longer rules' and ' By creating new rules'	None of the options	4	3	4
143	What are two steps of tree pruning work?	Pessimistic pruning and Optimistic pruning	Postpruning and Prepruning	Cost complexity pruning and time complexity pruning	None of the options	4	3	4

144	Which of the following sentences are true?	In pre-pruning a tree is 'pruned' by halting its construction early	A pruning set of class labelled tuples is used to estimate cost complexity	The best pruned tree is the one that minimizes the number of encoding bits	All of the above	4	3	4
145	Assume that you are given a data set and a neural network model trained on the data set. You are asked to build a decision tree model with the sole purpose of understanding/interpreting the built neural network model. In such a scenario, which among the following measures would you concentrate most on optimising?	Accuracy of the decision tree model on the given data set	F1 measure of the decision tree model on the given data set	Fidelity of the decision tree model, which is the fraction of instances on which the neural network and the decision tree give the same output	Comprehensibility of the decision tree model, measured in terms of the size of the corresponding rule set	4	3	4
146	Which of the following properties are characteristic of decision trees? (a) High bias (b) High variance (c) Lack of smoothness of prediction surfaces (d) Unbounded parameter set	a and b	a and d	b, c and d	All of the above	4	3	4
147	To control the size of the tree, we need to control the number of regions. One approach to do this would be to split tree nodes only if the resultant decrease in the sum of squares error exceeds some threshold. For the described method, which among the following are true? (a) It would, in general, help restrict the size of the trees (b) It has the potential to affect the performance of the resultant regression/classification model (c) It is computationally infeasible	a and b	a and d	b, c and d	All of the above	4	3	4
148	Which among the following statements best describes our approach to learning decision trees	Identify the best partition of the input space and response per partition to minimise sum of squares error	Identify the best approximation of the above by the greedy approach (to identifying the partitions)	Identify the model which gives the best performance using the greedy approximation (option (b)) with the smallest partition scheme	Identify the model which gives performance close to the best greedy approximation performance (option (b)) with the smallest partition scheme	4	3	4
149	Having built a decision tree, we are using reduced error pruning to reduce the size of the tree. We select a node to collapse. For this particular node, on the left branch, there are 3 training data points with the following outputs: 5, 7, 9.6 and for the right branch, there are four training data points with the following outputs: 8.7, 9.8, 10.5, 11. What were the original responses for data points along the two branches (left & right respectively) and what is the new response after collapsing the node?	10.8, 13.33, 14.48	10.8, 13.33, 12.06	7.2, 10, 8.8	7.2, 10, 8.6	4	3	4

	Suppose on performing reduced error pruning, we collapsed a node and observed an improvement in the prediction accuracy on the validation set. Which among the following statements are possible in light of the performance improvement observed? (a) The collapsed node helped overcome the effect of one or more noise affected data points in the training set (b) The validation set had one or more noise affected data points in the region corresponding to the collapsed node (c) The validation set did not have any data points along at least one of the collapsed branches (d) The validation set did have data points adversely affected by the collapsed node	a and b	a and d	b, c and d	All of the above	4	3	4
151	Time Complexity of k-means is given by	$O(mn)$	$O(tkn)$	$O(kn)$	$O(t2kn)$	4	3	4
152	Which one of the following is not a major strength of the neural network approach?	Neural network learning algorithms are guaranteed to converge to an optimal solution	Neural networks work well with datasets containing noisy data	Neural networks can be used for both supervised learning and unsupervised clustering	Neural networks can be used for applications that require a time element to be included in the data	6	3	4
153	In Apriori algorithm, if 1 item-sets are 100, then the number of candidate 2 item-sets are	100	200	4950	5000	4	3	4
154	Significant Bottleneck in the Apriori algorithm is	Finding frequent itemsets	Pruning	Candidate generation	Number of iterations	4	3	4
155	Machine learning techniques differ from statistical techniques in that machine learning methods	are better able to deal with missing and noisy data	typically assume an underlying distribution for the data	have trouble with large-sized datasets	are not able to explain their behavior	4	3	4
156	The probability that a person owns a sports car given that they subscribe to automotive magazine is 40%. We also know that 3% of the adult population subscribes to automotive magazine. The probability of a person owning a sports car given that they don't subscribe to automotive magazine is 30%. Use this information to compute the probability that a person subscribes to automotive magazine given that they own a sports car	0.0368	0.0396	0.0389	0.0398	4	3	4
157	What is the final resultant cluster size in Divisive algorithm, which is one of the hierarchical clustering approaches?	Zero	Three	singleton	Two	4	3	4
158	Given a frequent itemset L, If $ L = k$, then there are	$2k - 1$ candidate association rules	$2k$ candidate association rules	$2k - 2$ candidate association rules	$2k - 2$ candidate association rules	4	3	5
159	A student Grade is a variable F1 which takes a value from A,B,C and D. Which of the following is True in the following case?	Variable F1 is an example of nominal variable	Variable F1 is an example of ordinal variable	It doesn't belong to any of the mentioned categories	It belongs to both ordinal and nominal category	1	2	3
160	What can be major issue in Leave-One-Out-Cross-Validation(LOOCV)?	Low Variance	High Variance	Faster Runtime Compared to K-Fold Cross Validation	Slower Runtime Compared to normal Validation	1	2	3
161	Imagine a Newly-Born starts to learn walking. It will try to find a suitable policy to learn walking after repeated falling and getting up.specify what type of machine learning is best suited?	classification	regression	Kmeans algorithm	Reinforcement Learning	1	2	3
162	Perceptron Classifier is	Unsupervised learning algorithm	Semi-Supervised Learning Algorithm	Supervised learning algorithm	Soft margin classifier	2	1	2

163	Type of dataset available in Supervised Learning is	Unlabeled dataset	Labeled Dataset	CSV file	Excel file	2	2	3
164	which among the following is the most appropriate kernel that can be used with SVM to separate the classes.	Linear kernel	Gaussian RBF kernel	Polynomial kernel	Option 1 and option 3	2	2	3
165	The SVMs are less effective when	The data is linearly separable	The data is clean and ready to use	The data is noisy and contains overlapping points	option 1 and option 2	2	2	3
166	Suppose you are using RBF kernel in SVM with high Gamma value. What does this signify?	The model would consider even far away points from hyperplane for modeling	The model would consider only the points close to the hyperplane for modeling	The model would not be affected by distance of points from hyperplane for modeling	option 1 and option 2	2	2	3
167	What is the precision value for following confusion matrix of binary classification?	0.91	0.09	0.9	0.95	2	3	4
168	Which of the following are components of generalization Error?	Bias	Variance	Both of them	None of them	2	1	2
169	Which of the following is not a kernel method in SVM?	Linear Kernel	Polynomial Kernel	RBF Kernel	Nonlinear Kernel	2	2	3
170	During the treatment of cancer patients , the doctor needs to be very careful about which patients need to be given chemotherapy.Which metric should we use in order to decide the patients who should be given chemotherapy?	Precision	Recall	call	score	2	3	4
171	Which one of the following is suitable? 1. When the hypothesis space is richer, overfitting is more likely. 2. when the feature space is larger , overfitting is more likely.	True, False	False, True	True, True	False, False	2	2	3
172	Which of the following is a categorical data?	Branch of Bank	Expenditure in rupees	prize of house	Weight of a person	2	2	3
173	The soft margin SVM is more preferred than the hard-margin SVM when-	The data is linearly separable	The data is noisy and contains overlapping points	The data is not noisy and linearly separable	The data is noisy and linearly separable	2	2	3
174	In SVM which has quadratic kernel function of polynomial degree 2 that has slack variable C as one hyper parameter. What would happen if we use very large value for C	We can still classify the data correctly for given setting of hyper parameter C	We can not classify the data correctly for given setting of hyper parameter C	We can not classify the data at all	Data can be classified correctly without any impact of C	2	3	4
175	In SVM, RBF kernel with appropriate parameters to perform binary classification where the data is non-linearly separable. In this scenario	The Decision boundary in the transformed feature space is non-linear	The Decision boundary in the transformed feature space is linear	The Decision boundary in the original feature space is not considered	The Decision boundary in the original feature space is linear	2	2	3
176	Which of the following is true about SVM? 1. Kernel function map low dimensional data to high dimensional space. 2. It is a similarity Function	1 is True, 2 is False	1 is False, 2 is True	1 is True, 2 is True	1 is False, 2 is False	2	1	2
177	What is the Accuracy in percentage based on following confusion matrix of three class classification. Confusion Matrix C= [14 0 0] [1 15 0] [0 0 6]	75%	97%	95%	85%	2	3	4
178	Which of the following method is used for multiclass classification?	One Vs Rest	LOOCV	All vs One	One vs Another	2	1	2
179	What is the precision value for following confusion matrix of binary classification?	0.91	0.09	0.9	0.95	2	3	4
180	Which of the following is not a kernel method in SVM?	Linear Kernel	Polynomial Kernel	RBF Kernel	Nonlinear Kernel	2	1	2

181	Based on survey , it was found that the probability that person like to watch serials is 0.25 and the probability that person like to watch netflix series is 0.43. Also the probability that person like to watch serials and netflix series is 0.12. what is the probability that a person doesn't like to watch either?	0.32	0.2	0.44	0.56	2	2	3
182	A machine learning problem involves four attributes plus a class. The attributes have 3, 2, 2, and 2 possible values each. The class has 3 possible values. How many maximum possible different examples are there?	12	24	48	72	2	3	4
183	MLE estimates are often undesirable because	they are biased	they have high variance	they are not consistent estimators	None of the above	2	1	2
184	Linear Regression is a _____ machine learning algorithm.	Supervised	Unsupervised	Semi-Supervised	Can't say	3	1	2
185	In the regression equation $Y = 75.65 + 0.50X$, the intercept is	0.5	75.65	1	indeterminable	3	1	2
186	The difference between the actual Y value and the predicted Y value found using a regression equation is called the	slope	residual	outlier	scatter plot	3	2	3
187	The selling price of a house depends on many factors. For example, it depends on the number of bedrooms, number of kitchen, number of bathrooms, the year the house was built, and the square footage of the lot. Given these factors, predicting the selling price of the house is an example of _____ task.	Binary Classification	Multilabel Classification	Simple Linear Regression	Multiple Linear Regression	3	3	4
188	Suppose, you got a situation where you find that your linear regression model is under fitting the data. In such situation which of the following options would you consider?	You will add more features	You will remove some features	All of the above	None of the above	3	2	3
189	Which of the following methods/methods do we use to find the best fit line for data in Linear Regression?	Least Square Error	Maximum Likelihood	Logarithmic Loss	Both A and B	3	2	3
190	We have been given a dataset with n records in which we have input attribute as x and output attribute as y. Suppose we use a linear regression method to model this data. To test our linear regressor, we split the data in training set and test set randomly. Now we increase the training set size gradually. As the training set size increases, What do you expect will happen with the mean training error?	Increase	Decrease	Remain constant	Can't Say	3	2	3
191	We have been given a dataset with n records in which we have input attribute as x and output attribute as y. Suppose we use a linear regression method to model this data. To test our linear regressor, we split the data in training set and test set randomly. What do you expect will happen with bias and variance as you increase the size of training data?	Bias increases and Variance increases	Bias decreases and Variance increases	Bias decreases and Variance decreases	Bias increases and Variance decreases	3	2	3
192	If X and Y in a regression model are totally unrelated,	the correlation coefficient would be -1	the coefficient of determination would be 0	the coefficient of determination would be 1	the SSE would be 0	3	2	3

	Regarding bias and variance, which of the following statements are true? (Here 'high' and 'low' are relative to the ideal model. (i) Models which overfit are more likely to have high bias (ii) Models which overfit are more likely to have low bias (iii) Models which overfit are more likely to have high variance (iv) Models which overfit are more likely to have low variance	(i) and (ii)	(ii) and (iii)	(iii) and (iv)	None of these	3	2	3
193	Which of the following evaluation metrics can be used to evaluate a model while modeling a continuous output variable?	AUC-ROC	Accuracy	Logloss	Mean-Squared-Error	3	3	4
194	Suppose that we have N independent variables ($X_1, X_2 \dots X_n$) and dependent variable is Y. Now Imagine that you are applying linear regression by fitting the best fit line using least square error on this data. You found that correlation coefficient for one of it's variable(Say X_1) with Y is 0.95.	Relation between the X_1 and Y is weak	Relation between the X_1 and Y is strong	Relation between the X_1 and Y is neutral	Correlation can't judge the relationship	3	3	4
195	In terms of bias and variance. Which of the following is true when you fit degree 2 polynomial?	Bias will be high, variance will be high	Bias will be low, variance will be high	Bias will be high, variance will be low	Bias will be low, variance will be low	3	3	4
196	Which of the following statements are true for a design matrix $X \in R^{n \times d}$ with $d > n$? (The rows are n sample points and the columns represent d features.)	Least-squares linear regression computes the weights $w = (XTX)^{-1} XTy$	The sample points are linearly separable	X has exactly $d - n$ eigenvectors with eigenvalue zero	At least one principal component direction is orthogonal to a hyperplane that contains all the sample points	3	3	4
197	Suppose your model is demonstrating high variance across the different training sets. Which of the following is NOT valid way to try and reduce the variance?	Increase the amount of training data in each training set	Improve the optimization algorithm being used for error minimization.	Decrease the model complexity	Reduce the noise in the training data	3	3	3
198	Point out the wrong statement.	Regression through the origin yields an equivalent slope if you center the data first	Normalizing variables results in the slope being the correlation	Least squares is not an estimation tool	None of the mentioned	3	3	4
199	Which of the following are components of generalization Error?	Bias	Variance	Both of them	None of them	3	1	2
200	Problem in multi regression is ?	multicollinearity	overfitting	both multicollinearity & overfitting	underfitting	3	1	2
201	How can we best represent 'support' for the following association rule: "If X and Y, then Z".	$\{X,Y\}/(\text{Total number of transactions})$	$\{Z\}/(\text{Total number of transactions})$	$\{Z\}/\{X,Y\}$	$\{X,Y,Z\}/(\text{Total number of transactions})$	3	2	3
202	Choose the correct statement with respect to 'confidence' metric in association rules	It is the conditional probability that a randomly selected transaction will include all the items in the consequent given that the transaction includes all the items in the antecedent.	A high value of confidence suggests a weak association rule	It is the probability that a randomly selected transaction will include all the items in the consequent as well as all the items in the antecedent.	Confidence is not measured in terms of (estimated) conditional probability.	3	2	3
203	Which Statement is not true statement.	k-means clustering is a linear clustering algorithm.	k-means clustering aims to partition n observations into k clusters	k-nearest neighbor is same as k-means	k-means is sensitive to outlier	4	1	2
204								

205	which of the following cases will K-Means clustering give poor results? 1. Data points with outliers 2. Data points with different densities 3. Data points with round shapes 4. Data points with non-convex shapes	1 and 2	2 and 3	2 and 4	1, 2 and 4	4	1	2
206	What is Decision Tree?	Flow-Chart	Structure in which internal node represents test on an attribute, each branch represents outcome of test and each leaf node represents class label	Flow-Chart like Structure in which internal node represents test on an attribute, each branch represents outcome of test and each leaf node represents class label	None of the above	4	1	2
207	8 observations are clustered into 3 clusters using K-Means clustering algorithm. After first iteration clusters, C1, C2, C3 has following observations: C1: {(2,2), (4,4), (6,6)} C2: {(0,4), (4,0),(2,5)} C3: {(5,5), (9,9)} What will be the cluster centroids if you want to proceed for second iteration?	C1: (4,4), C2: (2,2), C3: (7,7)	C1: (6,6), C2: (4,4), C3: (9,9)	C1: (2,2), C2: (0,0), C3: (5,5)	C1: (4,4), C2: (3,3), C3: (7,7)	4	2	3
208	Choose the correct statement with respect to 'confidence' metric in association rules	It is the conditional probability that a randomly selected transaction will include all the items in the consequent given that the transaction includes all the items in the antecedent.	A high value of confidence suggests a weak association rule	It is the probability that a randomly selected transaction will include all the items in the consequent as well as all the items in the antecedent.	Confidence is not measured in terms of (estimated) conditional probability.	4	2	3
209	What are two steps of tree pruning work?	Pessimistic pruning and Optimistic pruning	Postpruning and Prepruning	Cost complexity pruning and time complexity pruning	None of the options	4	2	3
210	A database has 5 transactions. Of these, 4 transactions include milk and bread. Further, of the given 4 transactions, 2 transactions include cheese. Find the support percentage for the following association rule "if milk and bread are purchased, then cheese is also purchased".	0.4	0.6	0.8	0.42	4	2	3
211	Which of the following option is true about k-NN algorithm?	It can be used for classification	It can be used for regression	It can be used in both classification and regression	Not useful in ML algorithm	4	1	2
212	How to select best hyperparameters in tree based models?	Measure performance over training data	Measure performance over validation data	Both of these	Random selection of hyper parameters	4	1	2
213	What is true about K-Mean Clustering? 1. K-means is extremely sensitive to cluster center initializations 2. Bad initialization can lead to Poor convergence speed 3. Bad initialization can lead to bad overall clustering	1 and 3	1 and 2	2 and 3	1, 2 and 3	4	1	2
214	What are tree based classifiers?	Classifiers which form a tree with each attribute at one level	Classifiers which perform series of condition checking with one attribute at a time	Both options except none	Not possible	4	1	2
215	What is gini index?	Gini index operates on the categorical target variables	It is a measure of purity	Gini index performs only binary split	All (1,2 and 3)	4	1	2
216	Tree/Rule based classification algorithms generate ... rule to perform the classification.	if-then.	while.	do while	switch.	4	1	2

217	Decision Tree is	Flow-Chart	Structure in which internal node represents test on an attribute, each branch represents outcome of test and each leaf node represents class label	Both a & b	Class of instance	4	1	2
218	Which of the following is true about Manhattan distance?	It can be used for continuous variables	It can be used for categorical variables	It can be used for categorical as well as continuous	It can be used for constants	4	2	3
219	A company has build a KNN classifier that gets 100% accuracy on training data. When they deployed this model on client side it has been found that the model is not at all accurate. Which of the following thing might gone wrong? Note: Model has successfully deployed and no technical issues are found at client side except the model performance	It is probably a overfitted model	It is probably a underfitted model	Can't say	Wrong Client data	4	2	3
220	Which of the following classifications would best suit the student performance classification systems?	If...then... analysis	Market-basket analysis	Regression analysis	Cluster analysis	4	3	4
221	Which statement is true about the K-Means algorithm? Select one:	The output attribute must be categorical.	All attribute values must be categorical.	All attributes must be numeric	Attribute values may be either categorical or numeric	4	2	3
222	Which of the following can act as possible termination conditions in K-Means? 1. For a fixed number of iterations. 2. Assignment of observations to clusters does not change between iterations. Except for cases with a bad local minimum. 3. Centroids do not change between successive iterations. 4. Terminate when RSS falls below a threshold.	1, 3 and 4	1, 2 and 3	1, 2 and 4	1,2,3,4	4	3	4
223	Which of the following statement is true about k-NN algorithm? 1) k-NN performs much better if all of the data have the same scale 2) k-NN works well with a small number of input variables (p), but struggles when the number of inputs is very large 3) k-NN makes no assumptions about the functional form of the problem being solved	1 and 2	1 and 3	Only 1	1,2 and 3	4	3	4
224	In which of the following cases will K-means clustering fail to give good results? 1) Data points with outliers 2) Data points with different densities 3) Data points with nonconvex shapes	1 and 2	2 and 3	1, 2, and 3	1 and 3	4	3	4
225	How will you counter over-fitting in decision tree?	By pruning the longer rules	By creating new rules	Both By pruning the longer rules' and ' By creating new rules'	Over-fitting is not possible	4	3	4
226	This clustering algorithm terminates when mean values computed for the current iteration of the algorithm are identical to the computed mean values for the previous iteration Select one:	K-Means clustering	conceptual clustering	expectation maximization	agglomerative clustering	4	3	4
227	Which one of the following is the main reason for pruning a Decision Tree?	To save computing time during testing	To save space for storing the Decision Tree	To make the training set error smaller	To avoid overfitting the training set	4	3	4

228	You've just finished training a decision tree for spam classification, and it is getting abnormally bad performance on both your training and test sets. You know that your implementation has no bugs, so what could be causing the problem?	Your decision trees are too shallow.	You need to increase the learning rate.	You are overfitting.	Incorrect data	4	3	4
229	The K-means algorithm:	Requires the dimension of the feature space to be no bigger than the number of samples	Has the smallest value of the objective function when K = 1	Minimizes the within class variance for a given number of clusters	Converges to the global optimum if and only if the initial means are chosen as some of the samples themselves	4	3	4
230	Which of the following metrics, do we have for finding dissimilarity between two clusters in hierarchical clustering? 1. Single-link 2. Complete-link 3. Average-link	1 and 2	1 and 3	2 and 3	1, 2 and 3	4	3	4
231	In which of the following cases will K-Means clustering fail to give good results? 1. Data points with outliers 2. Data points with different densities 3. Data points with round shapes 4. Data points with non-convex shapes	1 and 2	2 and 3	2 and 4	1, 2 and 4	4	2	3
232	Hierarchical clustering is slower than non-hierarchical clustering?	TRUE	FALSE	Depends on data	Cannot say	4	2	3
233	High entropy means that the partitions in classification are	pure	not pure	useful	useless	4	2	3
234	Suppose we would like to perform clustering on spatial data such as the geometrical locations of houses. We wish to produce clusters of many different sizes and shapes. Which of the following methods is the most appropriate?	Decision Trees	Density-based clustering	Model-based clustering	K-means clustering	4	3	4
235	The main disadvantage of maximum likelihood methods is that they are	mathematically less folded	mathematically less complex	mathematically less complex	computationally intense	4	1	2
236	The maximum likelihood method can be used to explore relationships among more diverse sequences, conditions that are not well handled by maximum parsimony methods.	TRUE	FALSE	-	-	4	2	3
237	Which Statement is not true statement.	k-means clustering is a linear clustering algorithm.	k-means clustering aims to partition n observations into k clusters	k-nearest neighbor is same as k-means	k-means is sensitive to outlier	4	1	2
238	what is Feature scaling done before applying K-Mean algorithm?	In distance calculation it will give the same weights for all features	You always get the same clusters. If you use or don't use feature scaling	In Manhattan distance it is an important step but in Euclidian it is not	None of these	4	1	2
239	which of the following cases will K-Means clustering give poor results? 1. Data points with outliers 2. Data points with different densities 3. Data points with round shapes 4. Data points with non-convex shapes	1 and 2	2 and 3	2 and 4	1, 2 and 4	4	1	2
240	What is the naïve assumption in a Naïve Bayes Classifier.	All the classes are independent of each other	All the features of a class are independent of each other	The most probable feature for a class is the most important feature to be considered for classification	All the features of a class are conditionally dependent on each other	5	1	2

241	Based on survey , it was found that the probability that person like to watch serials is 0.25 and the probability that person like to watch netflix series is 0.43. Also the probability that person like to watch serials and netflix series is 0.12. what is the probability that a person doesn't like to watch either?	0.32	0.2	0.44	0.56	5	2	3
242	What is the actual number of independent parameters which need to be estimated in P dimensional Gaussian distribution model?	P	2P	$P(P+1)/2$	$P(P+3)/2$	5	1	2
243	Give the correct Answer for following statements. 1. It is important to perform feature normalization before using the Gaussian kernel. 2. The maximum value of the Gaussian kernel is 1.	1 is True, 2 is False	1 is False, 2 is True	1 is True, 2 is True	1 is False, 2 is False	5	3	4
244	What is the naïve assumption in a Naïve Bayes Classifier.	All the classes are independent of each other	All the features of a class are independent of each other	The most probable feature for a class is the most important feature to be considered for classification	All the features of a class are conditionally dependent on each other	5	1	2
245	What is the actual number of independent parameters which need to be estimated in P dimensional Gaussian distribution model?	P	2P	$P(P+1)/2$	$P(P+3)/2$	5	1	2
246	Which of the following quantities are minimized directly or indirectly during parameter estimation in Gaussian distribution Model?	Negative Log-likelihood	Log-liklihood	Cross Entropy	Residual Sum of Square	5	2	3
247	In Naive Bayes equation $P(C / X) = (P(X / C) * P(C)) / P(X)$ which part considers "likelihood"?	$P(X/C)$	$P(C/X)$	$P(C)$	$P(X)$	5	1	2
248	Consider the following dataset. x,y,z are the features and T is a class(1/0). Classify the test data (0,0,1) as values of x,y,z respectively.	0	1	0.1	0.9	5	3	4
249	Given a rule of the form IF X THEN Y, rule confidence is defined as the conditional probability that Select one:	Y is false when X is known to be false.	Y is true when X is known to be true.	X is true when Y is known to be true	X is false when Y is known to be false.	5	3	4
250	Which of the following statements about Naive Bayes is incorrect?	Attributes are equally important.	Attributes are statistically dependent of one another given the class value.	Attributes are statistically independent of one another given the class value.	Attributes can be nominal or numeric	5	2	3
251	How the entries in the full joint probability distribution can be calculated?	Using variables	Using information	Both Using variables & information	None of the mentioned	5	2	3
252	How many terms are required for building a bayes model?	1	2	3	4	5	2	3
253	Skewness of Normal distribution is _____	Negative	Positive	0	Undefined	5	1	2
254	The shape of the Normal Curve is _____	Bell Shaped	flat	circular	spiked	5	1	2
255	The correlation coefficient for two real-valued attributes is -0.85. What does this value tell you?	The attributes are not linearly related.	As the value of one attribute increases the value of the second attribute also increases	As the value of one attribute decreases the value of the second attribute increases	The attributes show a linear relationship	5	1	2
256	8 observations are clustered into 3 clusters using K-Means clustering algorithm. After first iteration clusters, C1, C2, C3 has following observations: C1: {(2,2), (4,4), (6,6)} C2: {(0,4), (4,0),(2,5)} C3: {(5,5), (9,9)} What will be the cluster centroids if you want to proceed for second iteration?	C1: (4,4), C2: (2,2), C3: (7,7)	C1: (6,6), C2: (4,4), C3: (9,9)	C1: (2,2), C2: (0,0), C3: (5,5)	C1: (4,4), C2: (3,3), C3: (7,7)	5	3	4

257	Which of the following quantities are minimized directly or indirectly during parameter estimation in Gaussian distribution Model?	Negative Log-likelihood	Log-liklihood	Cross Entropy	Residual Sum of Square	5	2	3
258	In Naive Bayes equation $P(C/X) = (P(X/C) * P(C)) / P(X)$ which part considers "likelihood"?	P(X/C)	$P(C/X)$	$P(C)$	$P(X)$	5	1	2
259	Consider the following dataset. x,y,z are the features and T is a class(1/0). Classify the test data (0,0,1) as values of x,y,z respectively.	0	1	0.1	0.9	5	3	4
260	Which of the following option is / are correct regarding benefits of ensemble model? 1. Better performance 2. Generalized models 3. Better interpretability	1 and 3	2 and 3	1, 2 and 3	1 and 2	5	1	2
261	The network that involves backward links from output to the input and hidden layers is called	Self organizing maps	Perceptrons	Recurrent neural network	Multi layered perceptron	6	1	2
262	Which of the following parameters can be tuned for finding good ensemble model in bagging based algorithms? 1. Max number of samples 2. Max features 3. Bootstrapping of samples 4. Bootstrapping of features	1	2	3&4	1,2,3&4	6	1	2
263	What is back propagation? a) It is another name given to the curvy function in the perceptron b) It is the transmission of error back through the network to adjust the inputs c) It is the transmission of error back through the network to allow weights to be adjusted so that the network can learn d) None of the mentioned	a	b	c	b&c	6	1	2
264	In an election for the head of college, N candidates are competing against each other and people are voting for either of the candidates. Voters don't communicate with each other while casting their votes. which of the following ensembles method works similar to the discussed elction Procedure?	Bagging	Boosting	Stacking	Randomization	6	2	3
265	What is the sequence of the following tasks in a perceptron? Initialize weights of perceptron randomly Go to the next batch of dataset If the prediction does not match the output, change the weights For a sample input, compute an output	1, 4, 3, 2	3, 1, 2, 4	4, 3, 2, 1	1, 2, 3, 4	6	2	3
266	In which neural net architecture, does weight sharing occur?	Recurrent Neural Network	Convolutional neural Network	. Fully Connected Neural Network	Both A and B	6	2	3

	Which of the following are correct statement(s) about stacking? 1. A machine learning model is trained on predictions of multiple machine learning models 2. A Logistic regression will definitely work better in the second stage as compared to other classification methods 3. First stage models are trained on full / partial feature space of training data	1 and 2	2 and 3	1 and 3	1,2 and 3	6	2	3
268	Given above is a description of a neural network. When does a neural network model become a deep learning model?	When you add more hidden layers and increase depth of neural network	When there is higher dimensionality of data	When the problem is an image recognition problem	When there is lower dimensionality of data	6	2	3
269	What are the steps for using a gradient descent algorithm? 1)Calculate error between the actual value and the predicted value 2)Reiterate until you find the best weights of network 3)Pass an input through the network and get values from output layer 4)Initialize random weight and bias 5)Go to each neurons which contributes to the error and change its respective values to reduce the error	1, 2, 3, 4, 5	4, 3, 1, 5, 2	3, 2, 1, 5, 4	5, 4, 3, 2, 1	6	3	4
270	A 4-input neuron has weights 1, 2, 3 and 4. The transfer function is linear with the constant of proportionality being equal to 2. The inputs are 4, 10, 10 and 30 respectively. What will be the output?	238	76	248	348	6	3	4
271	Which of the following option is / are correct regarding benefits of ensemble model? 1. Better performance 2. Generalized models 3. Better interpretability	1 and 3	2 and 3	1, 2 and 3	1 and 2	6	1	2
272	The network that involves backward links from output to the input and hidden layers is called	Self organizing maps	Perceptrons	Recurrent neural network	Multi layered perceptron	6	1	2
273	Which of the following parameters can be tuned for finding good ensemble model in bagging based algorithms? 1. Max number of samples 2. Max features 3. Bootstrapping of samples 4. Bootstrapping of features	1	2	3&4	1,2,3&4	6	1	2
274	Increase in size of a convolutional kernel would necessarily increase the performance of a convolutional network.	TRUE	FALSE			6	1	2
275	The F-test	an omnibus test	considers the reduction in error when moving from the complete model to the reduced model	considers the reduction in error when moving from the reduced model to the complete model	can only be conceptualized as a reduction in error	6	1	2

276	What is back propagation? a) It is another name given to the curvy function in the perceptron b) It is the transmission of error back through the network to adjust the inputs c) It is the transmission of error back through the network to allow weights to be adjusted so that the network can learn d) None of the mentioned	a	b	c	b&c	6	1	2
277	In an election for the head of college, N candidates are competing against each other and people are voting for either of the candidates. Voters don't communicate with each other while casting their votes. Which of the following ensembles method works similar to the discussed election Procedure?	Bagging	Boosting	Stacking	Randomization	6	1	2
278	Which of the following is NOT supervised learning?	PCA	Decision tree	Linear Regression	Naive Bayesian	2,3, 4,5	1	2
279	Which of the following algorithm is not an example of an ensemble method?	Extra Tree Regressor	Random Forest	Gradient Boosting	Decision Tree	6	2	3
280	What is true about an ensembled classifier? 1. Classifiers that are more "sure" can vote with more conviction 2. Classifiers can be more "sure" about a particular part of the space 3. Most of the times, it performs better than a single classifier	1 and 2	1 and 3	2 and 3	All of the above	6	2	3
281	Which of the following option is / are correct regarding benefits of ensemble model? 1. Better performance 2. Generalized models 3. Better interpretability	1 and 3	2 and 3	1 and 2	1, 2 and 3	6	1	2
282	Which of the following can be true for selecting base learners for an ensemble? 1. Different learners can come from same algorithm with different hyper parameters 2. Different learners can come from different algorithms 3. Different learners can come from different training spaces	1	2	1 and 3	1, 2 and 3	6	2	3
283	True or False: Ensemble learning can only be applied to supervised learning methods.	TRUE	FALSE			6	1	2
284	True or False: Ensembles will yield bad results when there is significant diversity among the models. Note: All individual models have meaningful and good predictions.	TRUE	FALSE			6	1	2
285	Which of the following is / are true about weak learners used in ensemble model? 1. They have low variance and they don't usually overfit 2. They have high bias, so they can not solve hard learning problems 3. They have high variance and they don't usually overfit	1 and 2	1 and 3	2 and 3	None of these	6	3	4

286	True or False: Ensemble of classifiers may or may not be more accurate than any of its individual model.	TRUE	False			6	1	2
287	If you use an ensemble of different base models, is it necessary to tune the hyper parameters of all base models to improve the ensemble performance?	Yes	No	can't say		6	1	2
288	Generally, an ensemble method works better, if the individual base models have _____? Note: Suppose each individual base models have accuracy greater than 50%.	Less correlation among predictions	High correlation among predictions	Correlation does not have any impact on ensemble output	None of the above	6	3	4
289	In an election, N candidates are competing against each other and people are voting for either of the candidates. Voters don't communicate with each other while casting their votes. Which of the following ensemble method works similar to above-discussed election procedure? Hint: Persons are like base models of ensemble method.	Bagging	Boosting	A Or B	None of these	6	3	4
290	Suppose there are 25 base classifiers. Each classifier has error rates of $e = 0.35$. Suppose you are using averaging as ensemble technique. What will be the probabilities that ensemble of above 25 classifiers will make a wrong prediction? Note: All classifiers are independent of each other	0.05	0.06	0.07	0.09	6	3	4
291	In machine learning, an algorithm (or learning algorithm) is said to be unstable if a small change in training data cause the large change in the learned classifiers. True or False: Bagging of unstable classifiers is a good idea	TRUE	FALSE			6	2	3
292	Which of the following parameters can be tuned for finding good ensemble model in bagging based algorithms? 1. Max number of samples 2. Max features 3. Bootstrapping of samples 4. Bootstrapping of features	1 and 3	2 and 3	1 and 2	All of above	6	2	3
293	How is the model capacity affected with dropout rate (where model capacity means the ability of a neural network to approximate complex functions)?	Model capacity increases in increase in dropout rate	Model capacity decreases in increase in dropout rate	Model capacity is not affected on increase in dropout rate	None of these	6	3	4
294	True or False: Dropout is computationally expensive technique w.r.t. bagging	TRUE	FALSE			6	1	2

295	Suppose, you want to apply a stepwise forward selection method for choosing the best models for an ensemble model. Which of the following is the correct order of the steps? Note: You have more than 1000 models predictions 1. Add the models predictions (or in another term take the average) one by one in the ensemble which improves the metrics in the validation set. 2. Start with empty ensemble 3. Return the ensemble from the nested set of ensembles that has maximum performance on the validation set	1-2-3	1-3-4	2-1-3	None of above	6	3	4
296	Suppose, you have 2000 different models with their predictions and want to ensemble predictions of best x models. Now, which of the following can be a possible method to select the best x models for an ensemble?	Step wise forward selection	Step wise backward elimination	Both	None of above	6	2	3
297	Below are the two ensemble models: 1. E1(M1, M2, M3) and 2. E2(M4, M5, M6) Above, M_x is the individual base models. Which of the following are more likely to choose if following conditions for E1 and E2 are given? E1: Individual Models accuracies are high but models are of the same type or in another term less diverse E2: Individual Models accuracies are high but they are of different types in another term high diverse in nature	E1	E2	Any of E1 and E2	None of these	6	3	4
298	True or False: In boosting, individual base learners can be parallel.	TRUE	FALSE			6	1	2
299	Which of the following is true about bagging? 1. Bagging can be parallel 2. The aim of bagging is to reduce bias not variance 3. Bagging helps in reducing overfitting	1 and 2	2 and 3	1 and 3	All of these	6	2	3
300	Suppose you are using stacking with n different machine learning algorithms with k folds on data. Which of the following is true about one level (m base models + 1 stacker) stacking? Note: Here, we are working on binary classification problem All base models are trained on all features You are using k folds for base models	You will have only k features after the first stage	You will have only m features after the first stage	You will have k+m features after the first stage	You will have k^*n features after the first stage	6	3	4
301	Which of the following is the difference between stacking and blending?	Stacking has less stable CV compared to Blending	In Blending, you create out of fold prediction	Stacking is simpler than Blending	None of these	6	2	3

	Which of the following can be one of the steps in stacking? 1. Divide the training data into k folds 2. Train k models on each k-1 folds and get the out of fold predictions for remaining one fold 3. Divide the test data set in "k" folds and get individual fold predictions by different algorithms	1 and 2	2 and 3	1 and 3	All of above	6	2	3
302	Q25. Which of the following are advantages of stacking? 1) More robust model 2) better prediction 3) Lower time of execution	1 and 2	2 and 3	1 and 3	All of the above	6	2	3
303	Which of the following are correct statement(s) about stacking? A machine learning model is trained on predictions of multiple machine learning models A Logistic regression will definitely work better in the second stage as compared to other classification methods First stage models are trained on full / partial feature space of training data	1 and 2	2 and 3	1 and 3	All of above	6	2	3
304	Which of the following is true about weighted majority votes? 1. We want to give higher weights to better performing models 2. Inferior models can overrule the best model if collective weighted votes for inferior models is higher than best model 3. Voting is special case of weighted voting	1 and 2	2 and 3	1 and 3	All of above	6	2	3
305	Which of the following is true about averaging ensemble?	1 and 3	2 and 3	1 and 2	1, 2 and 3	6	2	3
306	How can we assign the weights to output of different models in an ensemble?	It can only be used in classification problem	It can only be used in regression problem	It can be used in both classification as well as regression	None of these	6	1	2
307	1. Use an algorithm to return the optimal weights 2. Choose the weights using cross validation 3. Give high weights to more accurate models	1 and 2	1 and 3	2 and 3	All of above	6	2	3
308	Suppose you are given 'n' predictions on test data by 'n' different models (M1, M2, ..., Mn) respectively. Which of the following method(s) can be used to combine the predictions of these models? Note: We are working on a regression problem 1. Median 2. Product 3. Average 4. Weighted sum 5. Minimum and Maximum 6. Generalized mean rule	1, 3 and 4	1,3 and 6	1,3, 4 and 6	All of above	6	3	4

309	In an election, N candidates are competing against each other and people are voting for either of the candidates. Voters don't communicate with each other while casting their votes. Which of the following ensemble method works similar to above-discussed election procedure? Hint: Persons are like base models of ensemble method.	Bagging	Boosting	A Or B	None of these	6	3	4
310	Generally, an ensemble method works better, if the individual base models have _____? Note: Suppose each individual base models have accuracy greater than 50%.	Less correlation among predictions	High correlation among predictions	Correlation does not have any impact on ensemble output	None of the above	6	3	4
311	If you use an ensemble of different base models, is it necessary to tune the hyper parameters of all base models to improve the ensemble performance?	Yes	No	can't say		6	2	3
312	Support Vector Machine is	Logical Model	Probabilistic Model	Geometric Model	None of the above	1	2	3
313	If X and Y in a regression model are totally unrelated,	the correlation coefficient would be -1	the coefficient of determination would be 0	the coefficient of determination would be 1	the SSE would be 0	3	2	4

Machine Learning Questions & Solutions

Question Context

A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college.

1) Which of the following statement is true in following case?

- A) Feature F1 is an example of nominal variable.
- B) Feature F1 is an example of ordinal variable.
- C) It doesn't belong to any of the above category.
- D) Both of these

Solution: (B)

Ordinal variables are the variables which has some order in their categories. For example, grade A should be consider as high grade than grade B.

2) Which of the following is an example of a deterministic algorithm?

- A) PCA
- B) K-Means
- C) None of the above

Solution: (A) A deterministic algorithm is that in which output does not change on different runs. PCA would give the same result if we run again, but not k-means.

3) [True or False] A Pearson correlation between two variables is zero but, still their values can still be related to each other.

- A) TRUE
- B) FALSE

Solution: (A)

$Y=X^2$. Note that, they are not only associated, but one is a function of the other and Pearson correlation between them is 0.

4) Which of the following statement(s) is / are true for Gradient Decent (GD) and Stochastic Gradient Decent (SGD)?

1. In GD and SGD, you update a set of parameters in an iterative manner to minimize the error function.
2. In SGD, you have to run through all the samples in your training set for a single update of a parameter in each iteration.
3. In GD, you either use the entire data or a subset of training data to update a parameter in each iteration.

A) Only 1

B) Only 2

C) Only 3

D) 1 and 2

E) 2 and 3

F) 1,2 and 3

Solution: (A)In SGD for each iteration you choose the batch which is generally contain the random sample of data But in case of GD each iteration contain the all of the training observations.

5) Which of the following hyper parameter(s), when increased may cause random forest to over fit the data?

1. Number of Trees
2. Depth of Tree
3. Learning Rate

A) Only 1

B) Only 2

C) Only 3

D) 1 and 2

E) 2 and 3

F) 1,2 and 3

Solution: (B) Usually, if we increase the depth of tree it will cause overfitting. Learning rate is not an hyperparameter in random forest. Increase in the number of tree will cause under fitting.

6) Imagine, you are working with “Analytics Vidhya” and you want to develop a machine learning algorithm which predicts the number of views on the articles.

Your analysis is based on features like author name, number of articles written by the same author on Analytics Vidhya in past and a few other features. Which of the following evaluation metric would you choose in that case?

- 1. Mean Square Error
- 2. Accuracy
- 3. F1 Score

A) Only 1

B) Only 2

C) Only 3

D) 1 and 3

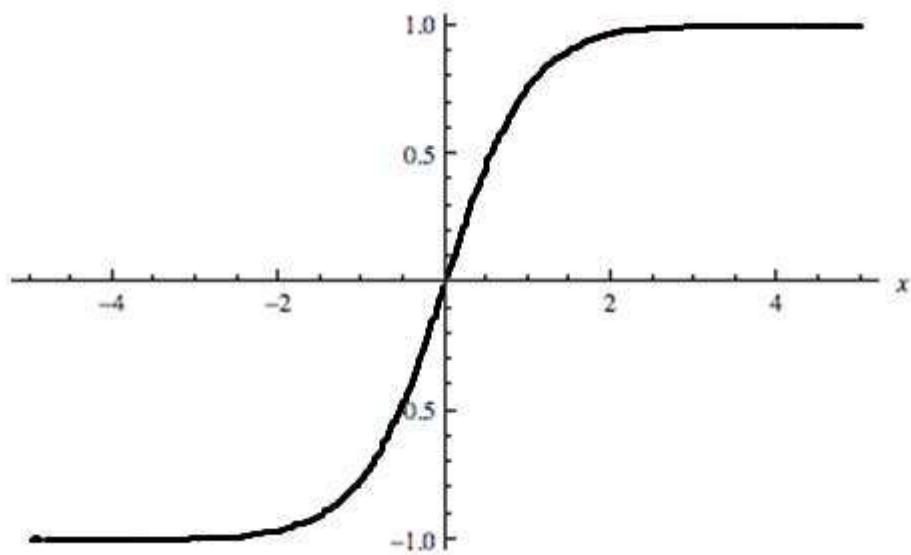
E) 2 and 3

F) 1 and 2

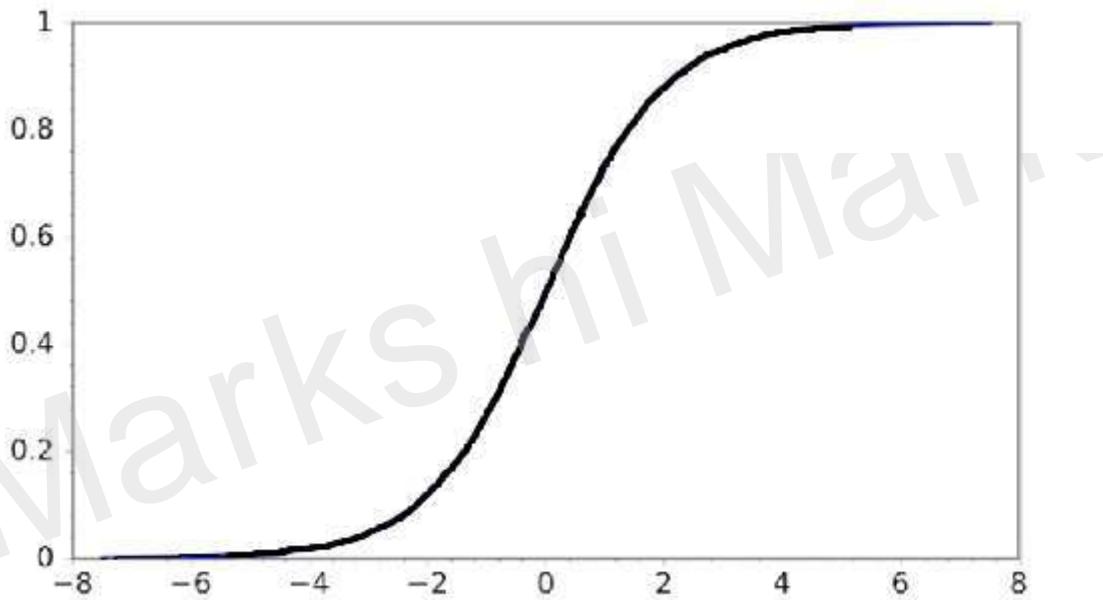
Solution:(A)

You can think that the number of views of articles is the continuous target variable which fall under the regression problem. So, mean squared error will be used as an evaluation metrics.

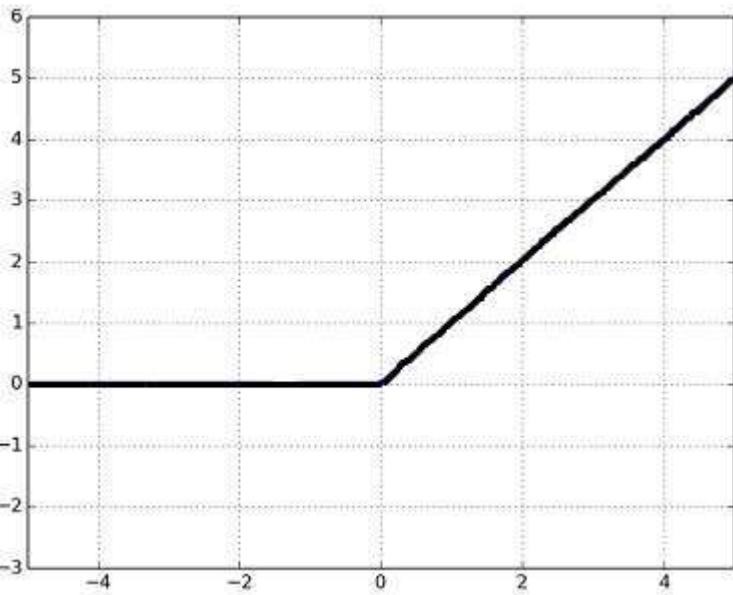
7) Given below are three images (1,2,3). Which of the following option is correct for these images?



A)



B)



C)
 A) 1 is tanh, 2 is ReLU and 3 is SIGMOID activation functions.

B) 1 is SIGMOID, 2 is ReLU and 3 is tanh activation functions.

C) 1 is ReLU, 2 is tanh and 3 is SIGMOID activation functions.

D) 1 is tanh, 2 is SIGMOID and 3 is ReLU activation functions.

Solution: (D)

The range of SIGMOID function is [0,1].

The range of the tanh function is [-1,1].

The range of the RELU function is [0, infinity].

So Option D is the right answer.

8) Below are the 8 actual values of target variable in the train file.

[0,0,0,1,1,1,1,1]

What is the entropy of the target variable?

A) $-(5/8 \log(5/8) + 3/8 \log(3/8))$

B) $5/8 \log(5/8) + 3/8 \log(3/8)$

C) $3/8 \log(5/8) + 5/8 \log(3/8)$

D) $5/8 \log(3/8) - 3/8 \log(5/8)$

$$-\sum p(x) * \log p(x)$$

Solution: (A) The formula for entropy is

So the answer is A.

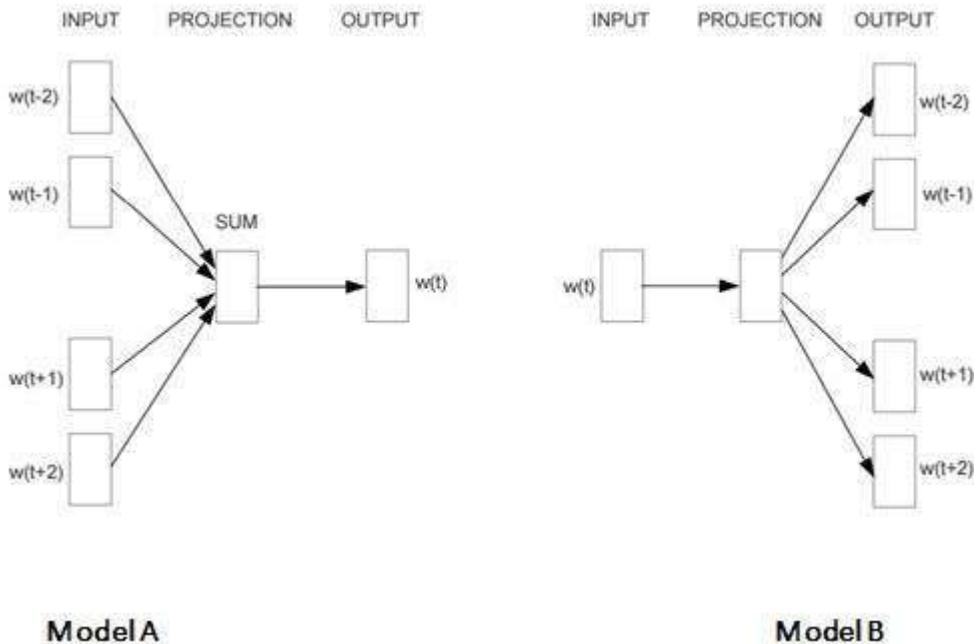
9) Let's say, you are working with categorical feature(s) and you have not looked at the distribution of the categorical variable in the test data.

You want to apply one hot encoding (OHE) on the categorical feature(s). What challenges you may face if you have applied OHE on a categorical variable of train dataset?

- A) All categories of categorical variable are not present in the test dataset.
- B) Frequency distribution of categories is different in train as compared to the test dataset.
- C) Train and Test always have same distribution.
- D) Both A and B
- E) None of these

Solution: (D) Both are true, The OHE will fail to encode the categories which is present in test but not in train so it could be one of the main challenges while applying OHE. The challenge given in option B is also true you need to more careful while applying OHE if frequency distribution doesn't same in train and test.

10) Skip gram model is one of the best models used in Word2vec algorithm for words embedding. Which one of the following models depict the skip gram model?



- A) A
 B) B
 C) Both A and B
 D) None of these

Solution: (B)

Both models (model1 and model2) are used in Word2vec algorithm. The model1 represent a CBOW model where as Model2 represent the Skip gram model.

11) Let's say, you are using activation function X in hidden layers of neural network. At a particular neuron for any given input, you get the output as “-0.0001”. Which of the following activation function could X represent?

- A) ReLU
 B) tanh
 C) SIGMOID
 D) None of these

Solution: (B) The function is a tanh because the this function output range is between (-1, -1).

12) [True or False] LogLoss evaluation metric can have negative values.

- A) TRUE
- B) FALSE

Solution: (B) Log loss cannot have negative values.

13) Which of the following statements is/are true about “Type-1” and “Type-2” errors?

- 1. Type1 is known as false positive and Type2 is known as false negative.
- 2. Type1 is known as false negative and Type2 is known as false positive.
- 3. Type1 error occurs when we reject a null hypothesis when it is actually true.

- A) Only 1
- B) Only 2
- C) Only 3
- D) 1 and 2
- E) 1 and 3
- F) 2 and 3

Solution: (E)

In statistical hypothesis testing, a type I error is the incorrect rejection of a true null hypothesis (a “false positive”), while a type II error is incorrectly retaining a false null hypothesis (a “false negative”).

14) Which of the following is/are one of the important step(s) to pre-process the text in NLP based projects?

- 1. Stemming
- 2. Stop word removal
- 3. Object Standardization

- A) 1 and 2
- B) 1 and 3

C) 2 and 3

D) 1,2 and 3

Solution: (D)

Stemming is a rudimentary rule-based process of stripping the suffixes ("ing", "ly", "es", "s" etc) from a word.

Stop words are those words which will have no relevance to the context of the data for example is/am/are.

Object Standardization is also one of the good way to pre-process the text.

15) Suppose you want to project high dimensional data into lower dimensions. The two most famous dimensionality reduction algorithms used here are PCA and t-SNE. Let's say you have applied both algorithms respectively on data "X" and you got the datasets "X_projected_PCA" , "X_projected_tSNE".

Which of the following statements is true for "X_projected_PCA" & "X_projected_tSNE" ?

A) X_projected_PCA will have interpretation in the nearest neighbour space.

B) X_projected_tSNE will have interpretation in the nearest neighbour space.

C) Both will have interpretation in the nearest neighbour space.

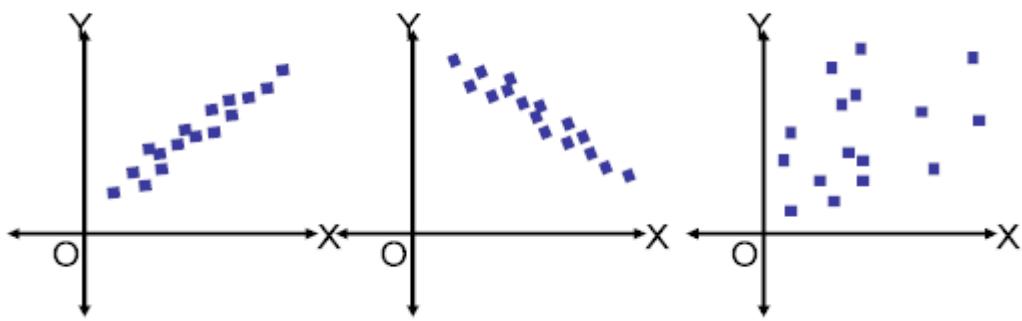
D) None of them will have interpretation in the nearest neighbour space.

Solution: (B)

t-SNE algorithm considers nearest neighbour points to reduce the dimensionality of the data. So, after using t-SNE we can think that reduced dimensions will also have interpretation in nearest neighbour space. But in the case of PCA it is not the case.

Context: 16-17

Given below are three scatter plots for two features (Image 1, 2 & 3 from left to right).



16) In the above images, which of the following is/are examples of multi-collinear features?

- A) Features in Image 1
- B) Features in Image 2
- C) Features in Image 3
- D) Features in Image 1 & 2
- E) Features in Image 2 & 3
- F) Features in Image 3 & 1

Solution: (D)

In Image 1, features have high positive correlation whereas in Image 2 has high negative correlation between the features so in both images pair of features are the example of multicollinear features.

17) In previous question, suppose you have identified multi-collinear features. Which of the following action(s) would you perform next?

1. Remove both collinear variables.
2. Instead of removing both variables, we can remove only one variable.
3. Removing correlated variables might lead to loss of information. In order to retain those variables, we can use penalized regression models like ridge or lasso regression.

- A) Only 1
- B) Only 2

- C) Only 3
- D) Either 1 or 3
- E) Either 2 or 3

Solution: (E)

You cannot remove both features because after removing the both features you will lose all of the information so you should either remove the only 1 feature or you can use the regularization algorithm like L1 and L2.

18) Adding a non-important feature to a linear regression model may result in.

- 1. Increase in R-square
- 2. Decrease in R-square

- A) Only 1 is correct
- B) Only 2 is correct
- C) Either 1 or 2
- D) None of these

Solution: (A)

After adding a feature in feature space, whether that feature is important or unimportant features the R-squared always increase.

19) Suppose, you are given three variables X, Y and Z. The Pearson correlation coefficients for (X, Y), (Y, Z) and (X, Z) are C1, C2 & C3 respectively.

Now, you have added 2 in all values of X (i.e. new values become X+2), subtracted 2 from all values of Y (i.e. new values are Y-2) and Z remains the same. The new coefficients for (X,Y), (Y,Z) and (X,Z) are given by D1, D2 & D3 respectively. How do the values of D1, D2 & D3 relate to C1, C2 & C3?

- A) D1 = C1, D2 < C2, D3 > C3
- B) D1 = C1, D2 > C2, D3 > C3
- C) D1 = C1, D2 > C2, D3 < C3
- D) D1 = C1, D2 < C2, D3 < C3

E) $D_1 = C_1$, $D_2 = C_2$, $D_3 = C_3$

F) Cannot be determined

Solution: (E) Correlation between the features won't change if you add or subtract a value in the features.

20) Imagine, you are solving a classification problems with highly imbalanced class. The majority class is observed 99% of times in the training data.

Your model has 99% accuracy after taking the predictions on test data. Which of the following is true in such a case?

1. Accuracy metric is not a good idea for imbalanced class problems.
2. Accuracy metric is a good idea for imbalanced class problems.
3. Precision and recall metrics are good for imbalanced class problems.
4. Precision and recall metrics aren't good for imbalanced class problems.

A) 1 and 3

B) 1 and 4

C) 2 and 3

D) 2 and 4

Solution: (A) Refer the question number 4 from in [this](#) article.

21) In ensemble learning, you aggregate the predictions for weak learners, so that an ensemble of these models will give a better prediction than prediction of individual models.

Which of the following statements is / are true for weak learners used in ensemble model?

1. They don't usually overfit.
2. They have high bias, so they cannot solve complex learning problems
3. They usually overfit.

A) 1 and 2

B) 1 and 3

C) 2 and 3

- D) Only 1
- E) Only 2
- F) None of the above

Solution: (A)

Weak learners are sure about particular part of a problem. So, they usually don't overfit which means that weak learners have low variance and high bias.

22) Which of the following options is/are true for K-fold cross-validation?

1. Increase in K will result in higher time required to cross validate the result.
2. Higher values of K will result in higher confidence on the cross-validation result as compared to lower value of K.
3. If $K=N$, then it is called Leave one out cross validation, where N is the number of observations.

- A) 1 and 2
- B) 2 and 3
- C) 1 and 3
- D) 1,2 and 3

Solution: (D)

Larger k value means less bias towards overestimating the true expected error (as training folds will be closer to the total dataset) and higher running time (as you are getting closer to the limit case: Leave-One-Out CV). We also need to consider the variance between the k folds accuracy while selecting the k.

Question Context 23-24

Cross-validation is an important step in machine learning for hyper parameter tuning. Let's say you are tuning a hyper-parameter "max_depth" for GBM by selecting it from 10 different depth values (values are greater than 2) for tree based model using 5-fold cross validation.

Time taken by an algorithm for training (on a model with max_depth 2) 4-fold is 10 seconds and for the prediction on remaining 1-fold is 2 seconds.

Note: Ignore hardware dependencies from the equation.

23) Which of the following option is true for overall execution time for 5-fold cross validation with 10 different values of “max_depth”?

- A) Less than 100 seconds
- B) 100 – 300 seconds
- C) 300 – 600 seconds
- D) More than or equal to 600 seconds
- C) None of the above
- D) Can't estimate

Solution: (D)

Each iteration for depth “2” in 5-fold cross validation will take 10 secs for training and 2 second for testing. So, 5 folds will take $12 \times 5 = 60$ seconds. Since we are searching over the 10 depth values so the algorithm would take $60 \times 10 = 600$ seconds. But training and testing a model on depth greater than 2 will take more time than depth “2” so overall timing would be greater than 600.

24) In previous question, if you train the same algorithm for tuning 2 hyper parameters say “max_depth” and “learning_rate”.

You want to select the right value against “max_depth” (from given 10 depth values) and learning rate (from given 5 different learning rates). In such cases, which of the following will represent the overall time?

- A) 1000-1500 second
- B) 1500-3000 Second
- C) More than or equal to 3000 Second
- D) None of these

Solution: (D)Same as question number 23.

25) Given below is a scenario for training error TE and Validation error VE for a machine learning algorithm M1. You want to choose a hyperparameter (H) based on TE and VE.

H	TE	VE
1	105	90
2	200	85
3	250	96
4	105	85
5	300	100

Which value of H will you choose based on the above table?

- A) 1
- B) 2
- C) 3
- D) 4
- E) 5

Solution: (D) Looking at the table, option D seems the best

26) What would you do in PCA to get the same projection as SVD?

- A) Transform data to zero mean
- B) Transform data to zero median
- C) Not possible
- D) None of these

Solution: (A) When the data has a zero mean vector PCA will have same projections as SVD, otherwise you have to centre the data first before taking SVD.

Question Context 27-28

Assume there is a black box algorithm, which takes training data with multiple observations ($t_1, t_2, t_3, \dots, t_n$) and a new observation (q_1). The black box outputs the nearest neighbor of q_1 (say t_i) and its corresponding class label c_i .

You can also think that this black box algorithm is same as 1-NN (1-nearest neighbor).

27) It is possible to construct a k-NN classification algorithm based on this black box alone.

Note: Where n (number of training observations) is very large compared to k.

A) TRUE

B) FALSE

Solution: (A)

In first step, you pass an observation (q_1) in the black box algorithm so this algorithm would return a nearest observation and its class.

In second step, you pass it out nearest observation from train data and again input the observation (q_1). The black box algorithm will again return the a nearest observation and it's class.

You need to repeat this procedure k times

28) Instead of using 1-NN black box we want to use the j-NN ($j > 1$) algorithm as black box. Which of the following option is correct for finding k-NN using j-NN?

1. J must be a proper factor of k
2. J > k
3. Not possible

A) 1

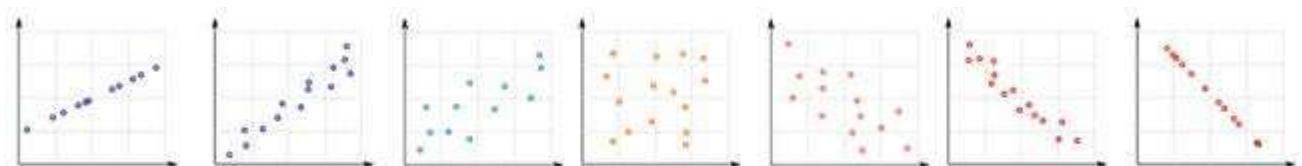
B) 2

C) 3

Solution: (A) Same as question number 27

29) Suppose you are given 7 Scatter plots 1-7 (left to right) and you want to compare Pearson correlation coefficients between variables of each scatterplot.

Which of the following is in the right order?



1. $1 < 2 < 3 < 4$
2. $1 > 2 > 3 > 4$
3. $7 < 6 < 5 < 4$
4. $7 > 6 > 5 > 4$

A) 1 and 3

B) 2 and 3

C) 1 and 4

D) 2 and 4

Solution: (B)

from image 1 to 4 correlation is decreasing (absolute value). But from image 4 to 7 correlation is increasing but values are negative (for example, 0, -0.3, -0.7, -0.99).

30) You can evaluate the performance of a binary class classification problem using different metrics such as accuracy, log-loss, F-Score. Let's say, you are using the log-loss function as evaluation metric.

Which of the following option is / are true for interpretation of log-loss as an evaluation metric?

$$\text{logLoss} = \frac{-1}{N} \sum_{i=1}^N (y_i(\log p_i) + (1 - y_i)\log(1 - p_i))$$

1. If a classifier is confident about an incorrect classification, then log-loss will penalise it heavily.
2. For a particular observation, the classifier assigns a very small probability for the correct class then the corresponding contribution to the log-loss will be very large.
3. Lower the log-loss, the better is the model.

A) 1 and 3

B) 2 and 3

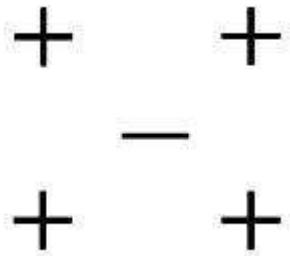
C) 1 and 2

D) 1,2 and 3

Solution: (D) Options are self-explanatory.

Context Question 31-32

Below are five samples given in the dataset.



Note: Visual distance between the points in the image represents the actual distance.

31) Which of the following is leave-one-out cross-validation accuracy for 3-NN (3-nearest neighbor)?

- A) 0
- D) 0.4
- C) 0.8
- D) 1

Solution: (C)

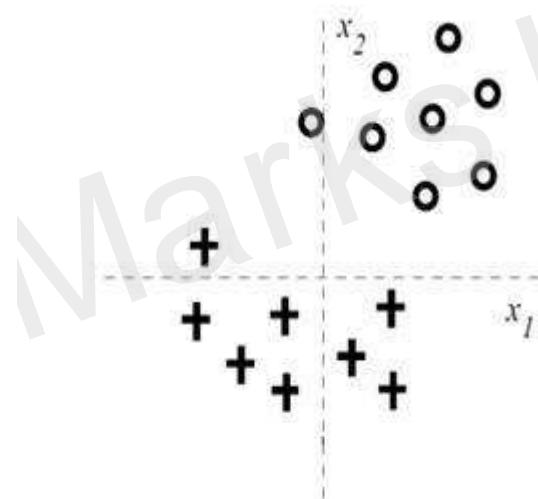
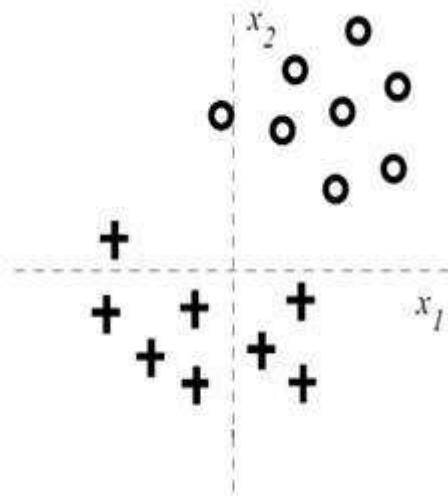
In Leave-One-Out cross validation, we will select $(n-1)$ observations for training and 1 observation of validation. Consider each point as a cross validation point and then find the 3 nearest point to this point. So if you repeat this procedure for all points you will get the correct classification for all positive class given in the above figure but negative class will be misclassified. Hence you will get 80% accuracy.

32) Which of the following value of K will have least leave-one-out cross validation accuracy?

- A) 1NN
- B) 3NN
- C) 4NN
- D) All have same leave one out error

Solution: (A) Each point which will always be misclassified in 1-NN which means that you will get the 0% accuracy.

33) Suppose you are given the below data and you want to apply a logistic regression model for classifying it in two given classes.



You are using logistic regression with L1 regularization.

$$\sum_{i=1}^n \log P(y_i|x_i, w_0, w_1, w_2) - C(|w_1| + |w_2|).$$

Where C is the regularization parameter and w1 & w2 are the coefficients of x1 and x2.

Which of the following option is correct when you increase the value of C from zero to a very large value?

- A) First w2 becomes zero and then w1 becomes zero
- B) First w1 becomes zero and then w2 becomes zero
- C) Both becomes zero at the same time
- D) Both cannot be zero even after very large value of C

Solution: (B)

By looking at the image, we see that even on just using x_2 , we can efficiently perform classification. So at first w_1 will become 0. As regularization parameter increases more, w_2 will come more and more closer to 0.

34) Suppose we have a dataset which can be trained with 100% accuracy with help of a decision tree of depth 6. Now consider the points below and choose the option based on these points.

Note: All other hyper parameters are same and other factors are not affected.

- 1. Depth 4 will have high bias and low variance
- 2. Depth 4 will have low bias and low variance

- A) Only 1
- B) Only 2
- C) Both 1 and 2
- D) None of the above

Solution: (A) If you fit decision tree of depth 4 in such data means it will more likely to underfit the data. So, in case of underfitting you will have high bias and low variance.

35) Which of the following options can be used to get global minima in k-Means Algorithm?

- 1. Try to run algorithm for different centroid initialization
- 2. Adjust number of iterations
- 3. Find out the optimal number of clusters

- A) 2 and 3
- B) 1 and 3

C) 1 and 2

D) All of above

Solution: (D) All of the option can be tuned to find the global minima.

36) Imagine you are working on a project which is a binary classification problem. You trained a model on training dataset and get the below confusion matrix on validation dataset.

		Predicted: NO	Predicted: YES
n=165	Actual: NO	50	10
Actual: YES	5	100	

Based on the above confusion matrix, choose which option(s) below will give you correct predictions?

1. Accuracy is ~0.91
2. Misclassification rate is ~ 0.91
3. False positive rate is ~0.95
4. True positive rate is ~0.95

A) 1 and 3

B) 2 and 4

C) 1 and 4

D) 2 and 3

Solution: (C)

The Accuracy (correct classification) is $(50+100)/165$ which is nearly equal to 0.91.

The true Positive Rate is how many times you are predicting positive class correctly so true positive rate would be $100/105 = 0.95$ also known as "Sensitivity" or "Recall"

37) For which of the following hyperparameters, higher value is better for decision tree algorithm?

1. Number of samples used for split
2. Depth of tree
3. Samples for leaf

A) 1 and 2

B) 2 and 3

C) 1 and 3

D) 1, 2 and 3

E) Can't say

Solution: (E)

For all three options A, B and C, it is not necessary that if you increase the value of parameter the performance may increase. For example, if we have a very high value of depth of tree, the resulting tree may overfit the data, and would not generalize well. On the other hand, if we have a very low value, the tree may underfit the data. So, we can't say for sure that "higher is better".

Context 38-39

Imagine, you have a $28 * 28$ image and you run a $3 * 3$ convolution neural network on it with the input depth of 3 and output depth of 8.

Note: Stride is 1 and you are using same padding.

38) What is the dimension of output feature map when you are using the given parameters.

A) 28 width, 28 height and 8 depth

B) 13 width, 13 height and 8 depth

C) 28 width, 13 height and 8 depth

D) 13 width, 28 height and 8 depth

Solution: (A) The formula for calculating output size is

$$\text{output size} = (N - F)/S + 1$$

where, N is input size, F is filter size and S is stride.

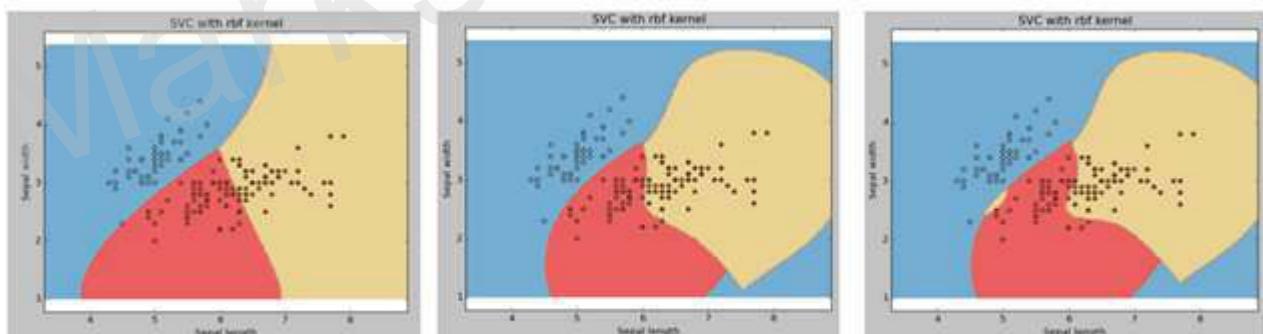
Read this [article](#) to get a better understanding.

39) What is the dimensions of output feature map when you are using following parameters.

- A) 28 width, 28 height and 8 depth
- B) 13 width, 13 height and 8 depth
- C) 28 width, 13 height and 8 depth
- D) 13 width, 28 height and 8 depth

Solution: (B) Same as above

40) Suppose, we were plotting the visualization for different values of C (Penalty parameter) in SVM algorithm. Due to some reason, we forgot to tag the C values with visualizations. In that case, which of the following option best explains the C values for the images below (1,2,3 left to right, so C1 for image1, C2 for image2 and C3 for image3) in case of rbf kernel.



- A) $C_1 = C_2 = C_3$
- B) $C_1 > C_2 > C_3$
- C) $C_1 < C_2 < C_3$
- D) None of these

Solution: (C)

Questions & Answers

Q1. Movie Recommendation systems are an example of:

1. Classification
2. Clustering
3. Reinforcement Learning
4. Regression

Options:

B. A. 2 Only

C. 1 and 2

D. 1 and 3

E. 2 and 3

F. 1, 2 and 3

H. 1, 2, 3 and 4

Solution: (E)

Generally, movie recommendation systems cluster the users in a finite number of similar groups based on their previous activities and profile. Then, at a fundamental level, people in the same cluster are made similar recommendations.

In some scenarios, this can also be approached as a classification problem for assigning the most appropriate movie class to the user of a specific group of users. Also, a movie recommendation system can be viewed as a reinforcement learning problem where it learns by its previous recommendations and improves the future recommendations.

Q2. Sentiment Analysis is an example of:

1. Regression
2. Classification
3. Clustering
4. Reinforcement Learning

Options:

A. 1 Only

- B. 1 and 2
- C. 1 and 3
- D. 1, 2 and 3
- E. 1, 2 and 4
- F. 1, 2, 3 and 4

Solution: (E)

Sentiment analysis at the fundamental level is the task of classifying the sentiments represented in an image, text or speech into a set of defined sentiment classes like happy, sad, excited, positive, negative, etc. It can also be viewed as a regression problem for assigning a sentiment score of say 1 to 10 for a corresponding image, text or speech.

Another way of looking at sentiment analysis is to consider it using a reinforcement learning perspective where the algorithm constantly learns from the accuracy of past sentiment analysis performed to improve the future performance.

Q3. Can decision trees be used for performing clustering?

- A. True
- B. False

Solution: (A)

Decision trees can also be used to find clusters in the data but clustering often generates natural clusters and is not dependent on any objective function.

Q4. Which of the following is the most appropriate strategy for data cleaning before performing clustering analysis, given less than desirable number of data points:

1. Capping and flooring of variables
2. Removal of outliers

Options:

- A. 1 only
- B. 2 only

- C. 1 and 2
- D. None of the above

Solution: (A)

Removal of outliers is not recommended if the data points are few in number. In this scenario, capping and flouring of variables is the most appropriate strategy.

Q5. What is the minimum no. of variables/ features required to perform clustering?

- A. 0
- B. 1
- C. 2
- D. 3

Solution: (B)

At least a single variable is required to perform clustering analysis. Clustering analysis with a single variable can be visualized with the help of a histogram.

Q6. For two runs of K-Mean clustering is it expected to get same clustering results?

- A. Yes
- B. No

Solution: (B)

K-Means clustering algorithm instead converges on local minima which might also correspond to the global minima in some cases but not always. Therefore, it's advised to run the K-Means algorithm multiple times before drawing inferences about the clusters.

However, note that it's possible to receive same clustering results from K-means by setting the same seed value for each run. But that is done by simply making the algorithm choose the set of same random no. for each run.

Q7. Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means

- A. Yes
- B. No
- C. Can't say
- D. None of these

Solution: (A)

When the K-Means algorithm has reached the local or global minima, it will not alter the assignment of data points to clusters for two successive iterations.

Q8. Which of the following can act as possible termination conditions in K-Means?

- 1. For a fixed number of iterations.
- 2. Assignment of observations to clusters does not change between iterations. Except for cases with a bad local minimum.
- 3. Centroids do not change between successive iterations.
- 4. Terminate when RSS falls below a threshold.

Options:

- A. 1, 3 and 4
- B. 1, 2 and 3
- C. 1, 2 and 4
- D. All of the above

Solution: (D)

All four conditions can be used as possible termination condition in K-Means clustering:

- 1. This condition limits the runtime of the clustering algorithm, but in some cases the quality of the clustering will be poor because of an insufficient number of iterations.
- 2. Except for cases with a bad local minimum, this produces a good clustering, but runtimes may be unacceptably long.
- 3. This also ensures that the algorithm has converged at the minima.

4. Terminate when RSS falls below a threshold. This criterion ensures that the clustering is of a desired quality after termination. Practically, it's a good practice to combine it with a bound on the number of iterations to guarantee termination.

Q9. Which of the following clustering algorithms suffers from the problem of convergence at local optima?

1. K-Means clustering algorithm
2. Agglomerative clustering algorithm
3. Expectation-Maximization clustering algorithm
4. Diverse clustering algorithm

Options:

- A. 1 only
- B. 2 and 3
- C. 2 and 4
- D. 1 and 3
- E. 1,2 and 4
- F. All of the above

Solution: (D)

Out of the options given, only K-Means clustering algorithm and EM clustering algorithm has the drawback of converging at local minima.

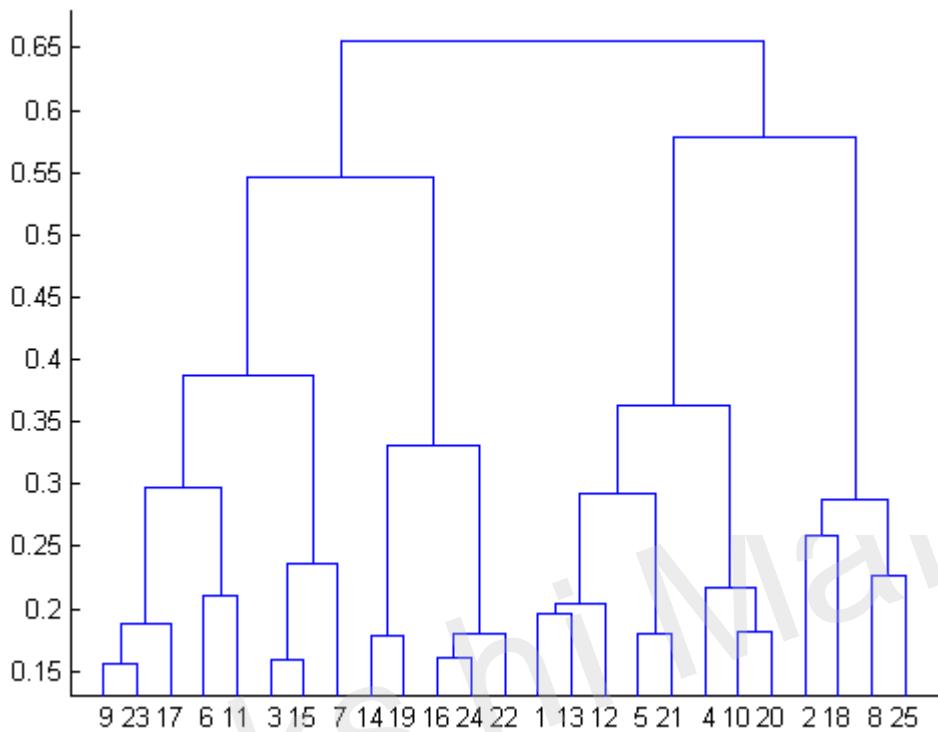
Q10. Which of the following algorithm is most sensitive to outliers?

- A. K-means clustering algorithm
- B. K-medians clustering algorithm
- C. K-modes clustering algorithm
- D. K-medoids clustering algorithm

Solution: (A)

Out of all the options, K-Means clustering algorithm is most sensitive to outliers as it uses the mean of cluster data points to find the cluster center.

Q11. After performing K-Means Clustering analysis on a dataset, you observed the following dendrogram. Which of the following conclusion can be drawn from the dendrogram?



- A. There were 28 data points in clustering analysis
- B. The best no. of clusters for the analyzed data points is 4
- C. The proximity function used is Average-link clustering
- D. The above dendrogram interpretation is not possible for K-Means clustering analysis

Solution: (D)

A dendrogram is not possible for K-Means clustering analysis. However, one can create a cluster gram based on K-Means clustering analysis.

Q12. How can Clustering (Unsupervised Learning) be used to improve the accuracy of Linear Regression model (Supervised Learning):

1. Creating different models for different cluster groups.

2. Creating an input feature for cluster ids as an ordinal variable.
3. Creating an input feature for cluster centroids as a continuous variable.
4. Creating an input feature for cluster size as a continuous variable.

Options:

- A. 1 only
- B. 1 and 2
- C. 1 and 4
- D. 3 only
- E. 2 and 4
- F. All of the above

Solution: (F)

Creating an input feature for cluster ids as ordinal variable or creating an input feature for cluster centroids as a continuous variable might not convey any relevant information to the regression model for multidimensional data. But for clustering in a single dimension, all of the given methods are expected to convey meaningful information to the regression model. For example, to cluster people in two groups based on their hair length, storing clustering ID as ordinal variable and cluster centroids as continuous variables will convey meaningful information.

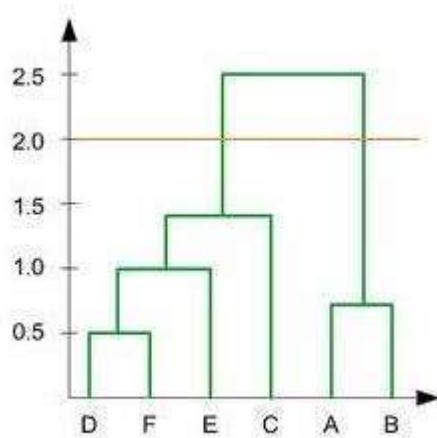
Q13. What could be the possible reason(s) for producing two different dendograms using agglomerative clustering algorithm for the same dataset?

- A. Proximity function used
- B. of data points used
- C. of variables used
- D. B and c only
- E. All of the above

Solution: (E)

Change in either of Proximity function, no. of data points or no. of variables will lead to different clustering results and hence different dendograms.

Q14. In the figure below, if you draw a horizontal line on y-axis for $y=2$. What will be the number of clusters formed?

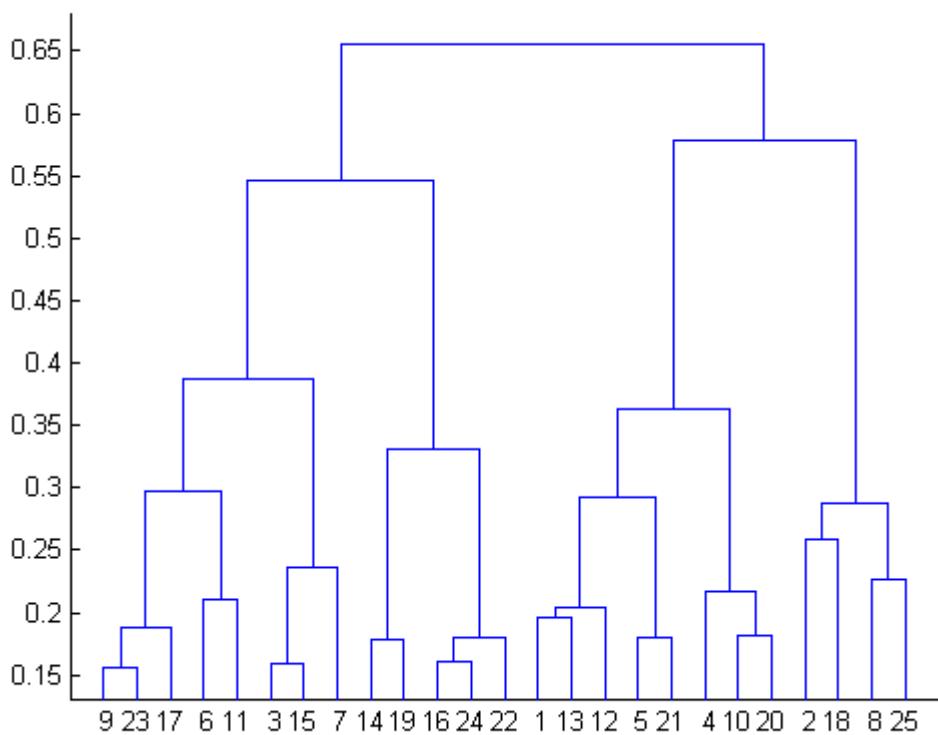


- A. 1
- B. 2
- C. 3
- D. 4

Solution: (B)

Since the number of vertical lines intersecting the red horizontal line at $y=2$ in the dendrogram are 2, therefore, two clusters will be formed.

Q15. What is the most appropriate no. of clusters for the data points represented by the following dendrogram:



A. 2

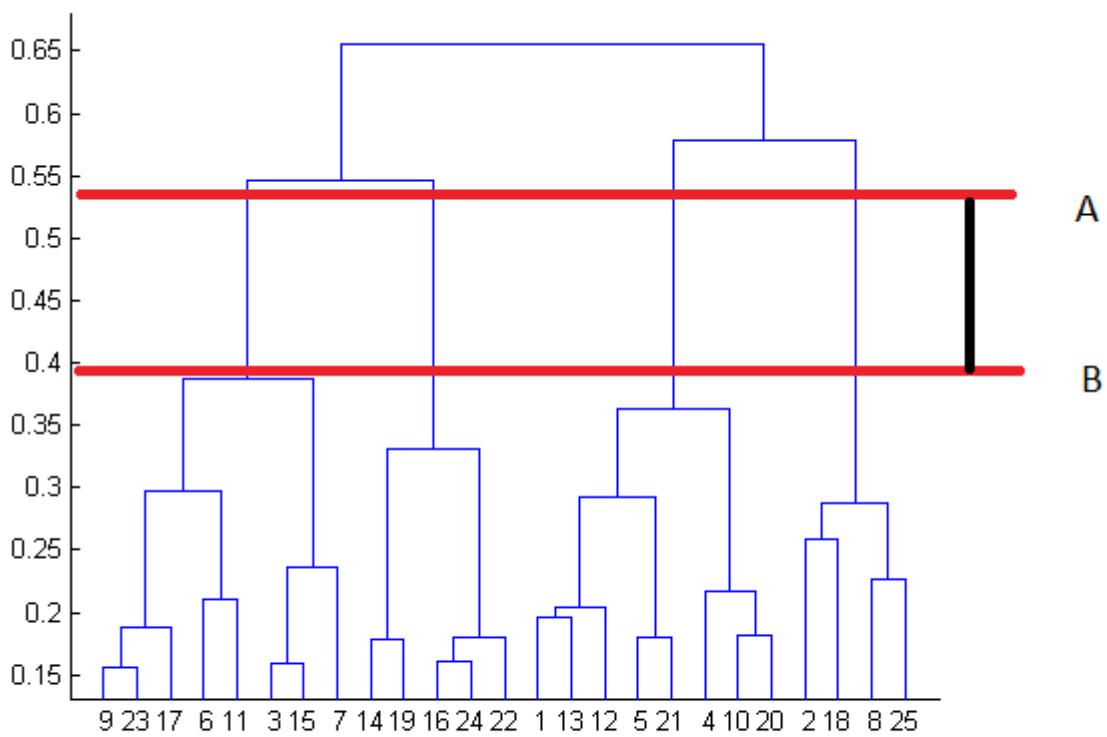
B. 4

C. 6

D. 8

Solution: (B)

The decision of the no. of clusters that can best depict different groups can be chosen by observing the dendrogram. The best choice of the no. of clusters is the no. of vertical lines in the dendrogram cut by a horizontal line that can transverse the maximum distance vertically without intersecting a cluster.



In the above example, the best choice of no. of clusters will be 4 as the red horizontal line in the dendrogram below covers maximum vertical distance AB.

Q16. In which of the following cases will K-Means clustering fail to give good results?

1. Data points with outliers
2. Data points with different densities
3. Data points with round shapes
4. Data points with non-convex shapes

Options:

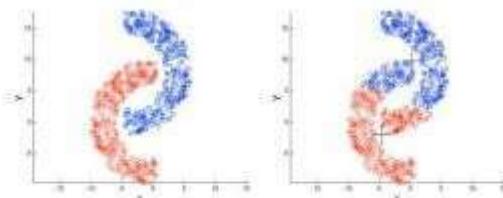
- A. 1 and 2
- B. 2 and 3
- C. 2 and 4
- D. 1, 2 and 4

E. 1, 2, 3 and 4

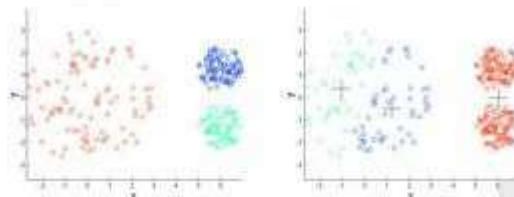
Solution: (D)

K-Means clustering algorithm fails to give good results when the data contains outliers, the density spread of data points across the data space is different and the data points follow non-convex shapes.

Non-convex/non-round-shaped clusters: Standard K-means fails!



Clusters with different densities



Q17. Which of the following metrics, do we have for finding dissimilarity between two clusters in hierarchical clustering?

1. Single-link
2. Complete-link
3. Average-link

Options:

- A. 1 and 2
- B. 1 and 3
- C. 2 and 3
- D. 1, 2 and 3

Solution: (D)

All of the three methods i.e. single link, complete link and average link can be used for finding dissimilarity between two clusters in hierarchical clustering.

Q18. Which of the following are true?

1. Clustering analysis is negatively affected by multicollinearity of features
2. Clustering analysis is negatively affected by heteroscedasticity

Options:

- A. 1 only
- B. 2 only
- C. 1 and 2
- D. None of them

Solution: (A)

Clustering analysis is not negatively affected by heteroscedasticity but the results are negatively impacted by multicollinearity of features/ variables used in clustering as the correlated feature/ variable will carry extra weight on the distance calculation than desired.

Q19. Given, six points with the following attributes:

point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

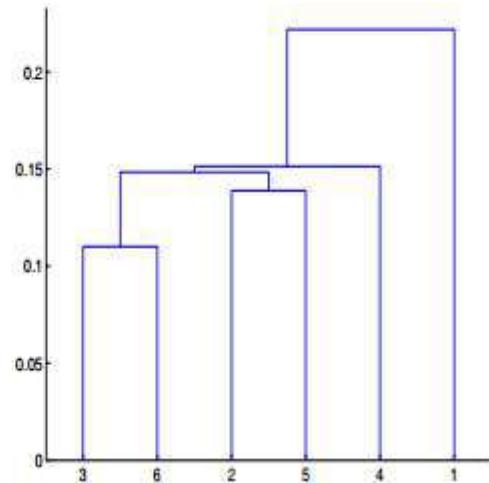
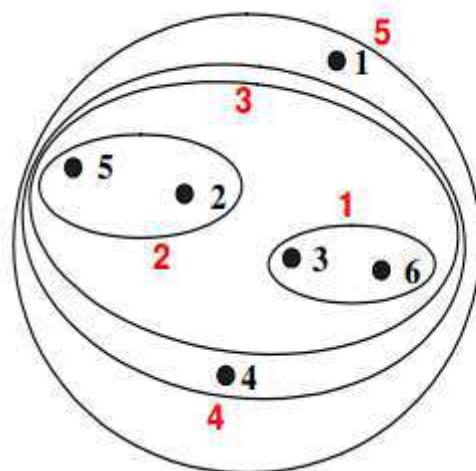
Table : X-Y coordinates of six points.

	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

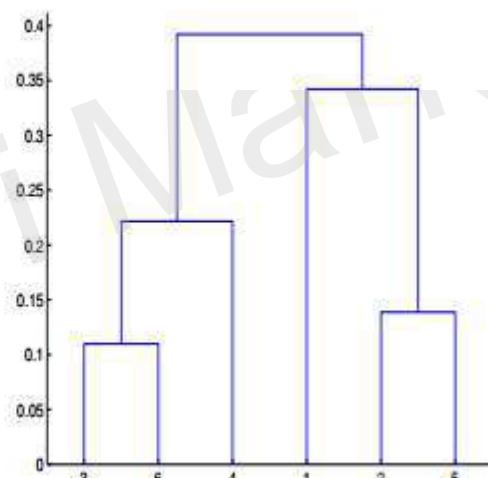
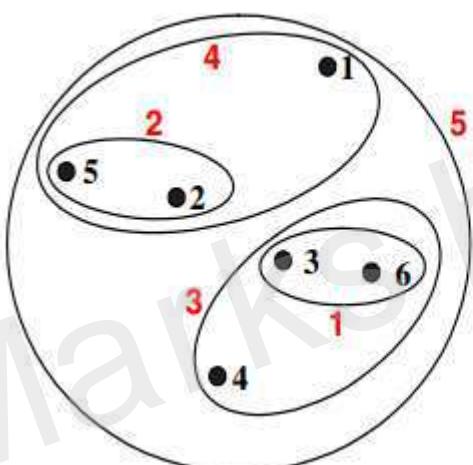
Table : Distance Matrix for Six Points

Which of the following clustering representations and dendrogram depicts the use of MIN or Single link proximity function in hierarchical clustering:

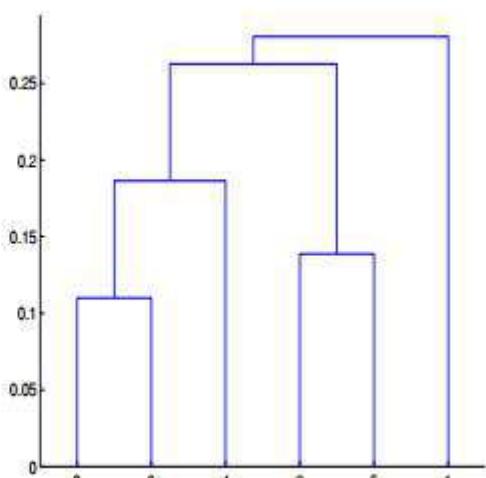
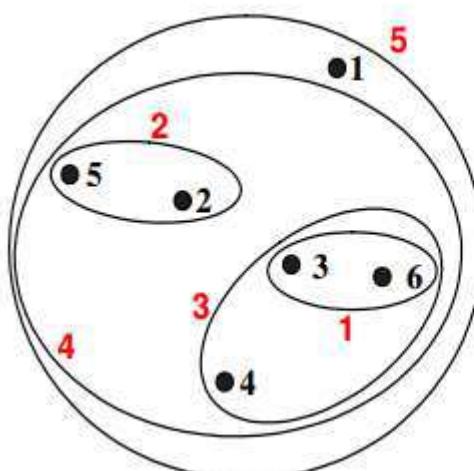
A.



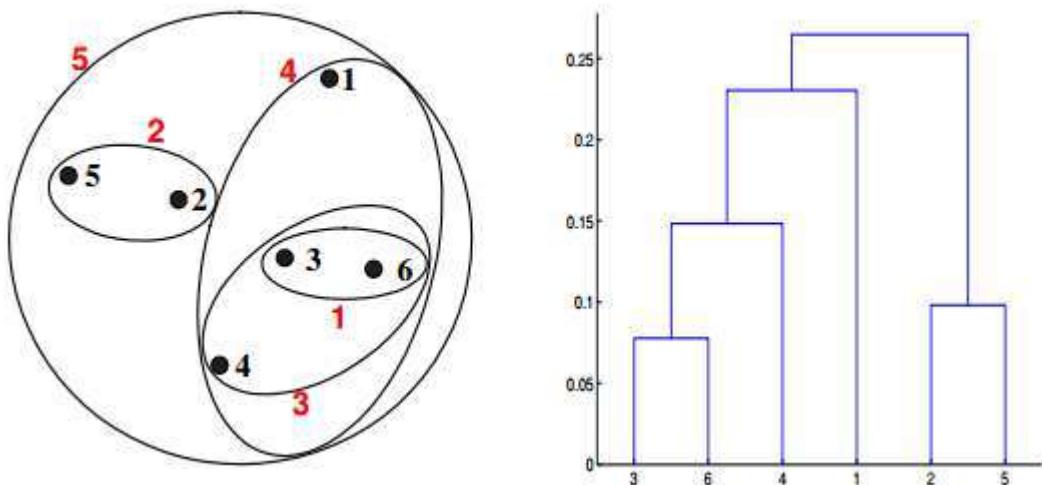
B.



C.



D.



Solution: (A)

For the single link or MIN version of hierarchical clustering, the proximity of two clusters is defined to be the minimum of the distance between any two points in the different clusters. For instance, from the table, we see that the distance between points 3 and 6 is 0.11, and that is the height at which they are joined into one cluster in the dendrogram. As another example, the distance between clusters {3, 6} and {2, 5} is given by $\text{dist}(\{3, 6\}, \{2, 5\}) = \min(\text{dist}(3, 2), \text{dist}(6, 2), \text{dist}(3, 5), \text{dist}(6, 5)) = \min(0.1483, 0.2540, 0.2843, 0.3921) = 0.1483$.

Q20 Given, six points with the following attributes:

point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

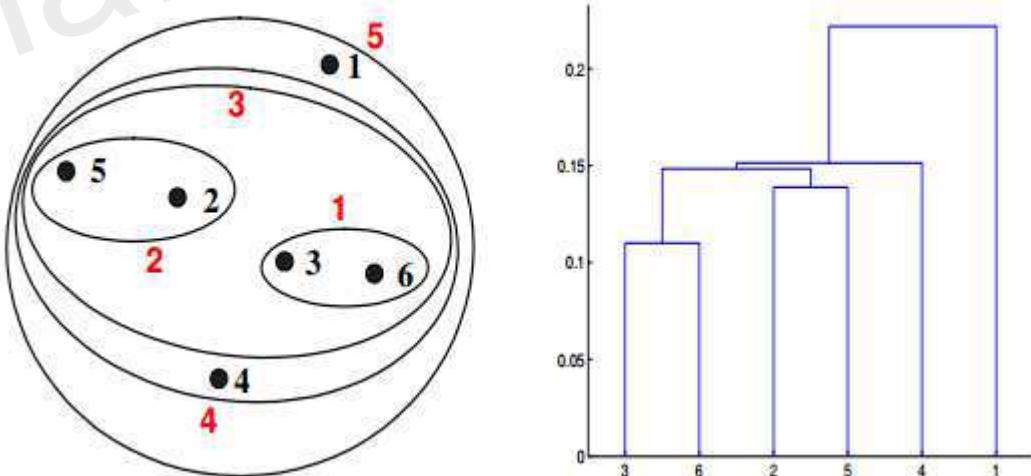
Table : X-Y coordinates of six points.

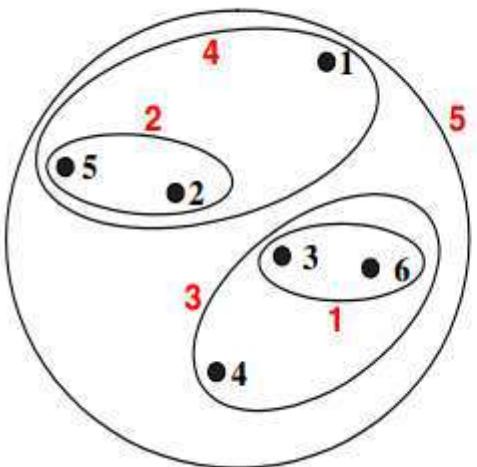
	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

Table : Distance Matrix for Six Points

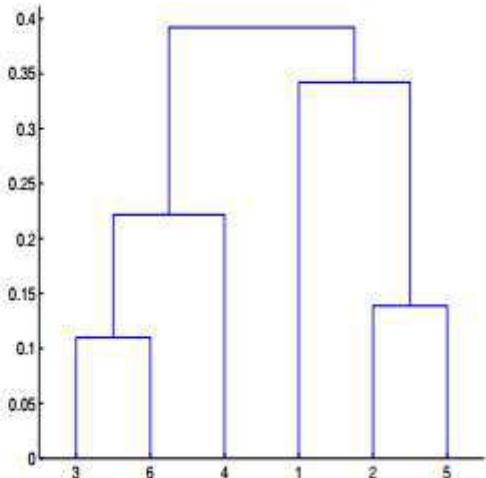
Which of the following clustering representations and dendrogram depicts the use of MAX or Complete link proximity function in hierarchical clustering:

A.

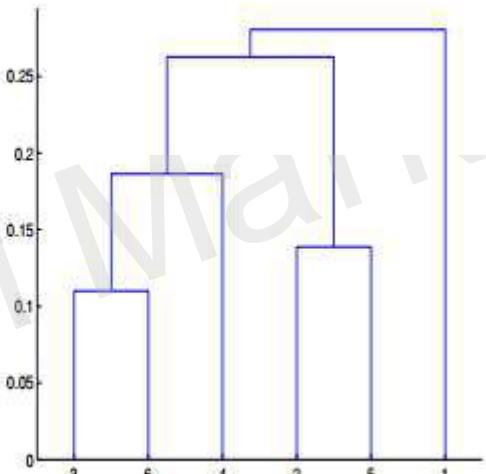
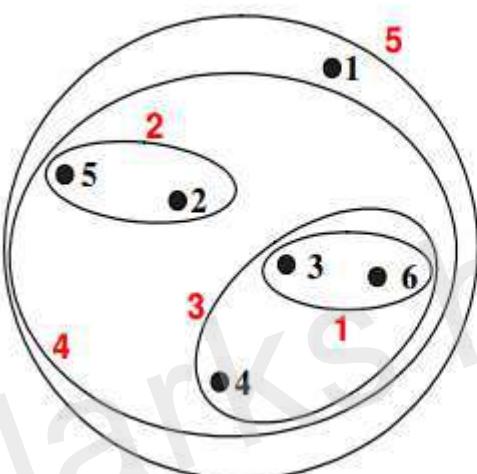




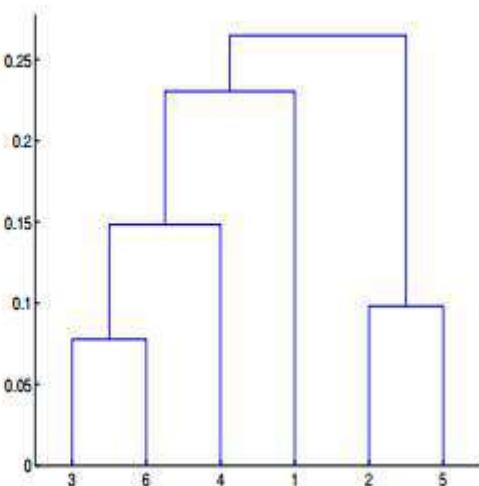
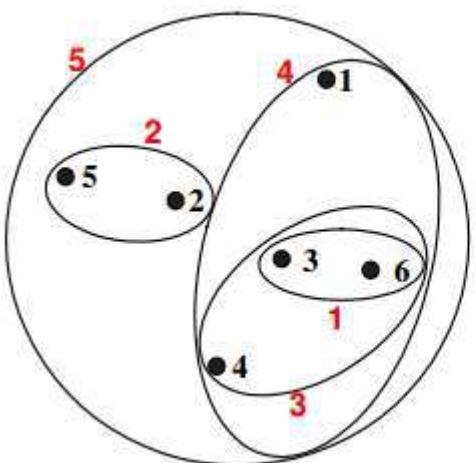
B.



C.



D.



Solution: (B)

For the single link or MAX version of hierarchical clustering, the proximity of two clusters is defined to be the maximum of the distance between any two points in the different clusters. Similarly, here points 3 and 6 are merged first. However, $\{3, 6\}$ is merged with $\{4\}$, instead of $\{2, 5\}$. This is because the $\text{dist}(\{3, 6\}, \{4\}) = \max(\text{dist}(3, 4), \text{dist}(6, 4)) = \max(0.1513, 0.2216) = 0.2216$, which is smaller than $\text{dist}(\{3, 6\}, \{2, 5\}) = \max(\text{dist}(3, 2), \text{dist}(6, 2), \text{dist}(3, 5), \text{dist}(6, 5)) = \max(0.1483, 0.2540, 0.2843, 0.3921) = 0.3921$ and $\text{dist}(\{3, 6\}, \{1\}) = \max(\text{dist}(3, 1), \text{dist}(6, 1)) = \max(0.2218, 0.2347) = 0.2347$.

Q21 Given, six points with the following attributes:

point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

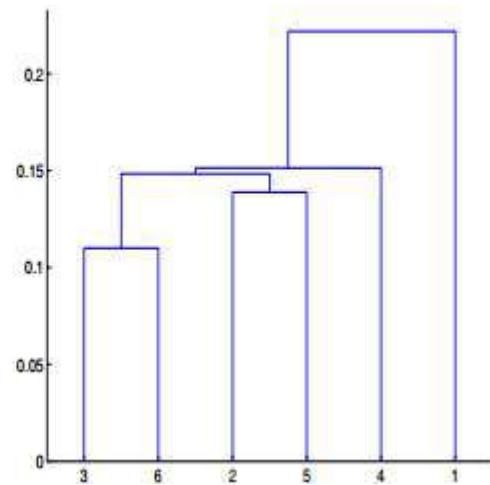
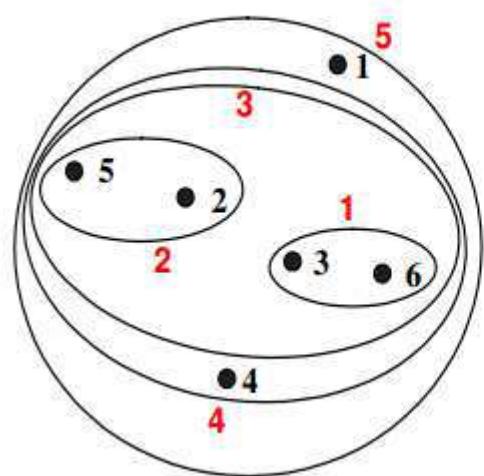
Table : X-Y coordinates of six points.

	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

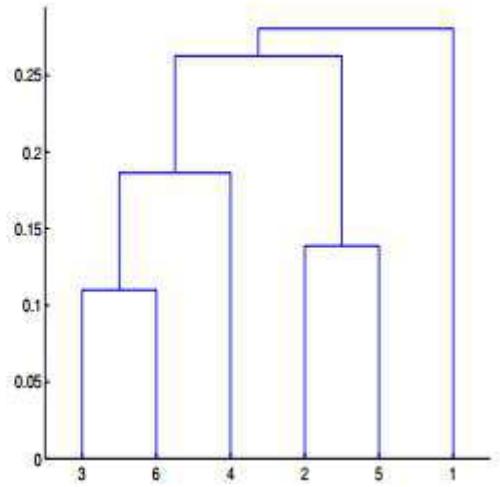
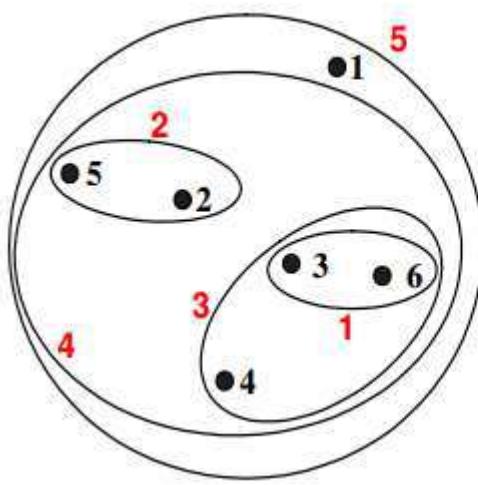
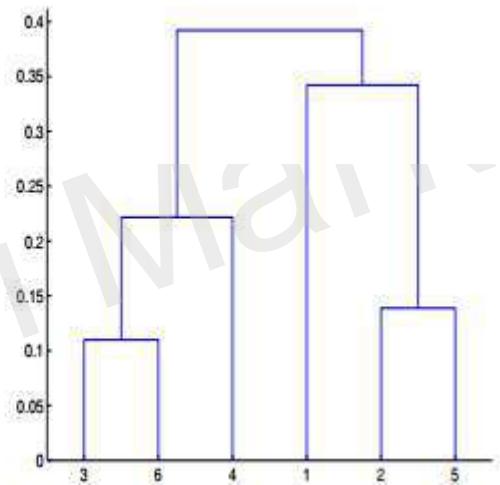
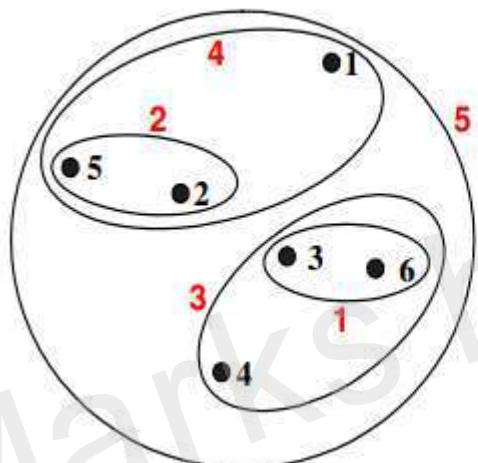
Table : Distance Matrix for Six Points

Which of the following clustering representations and dendrogram depicts the use of Group average proximity function in hierarchical clustering:

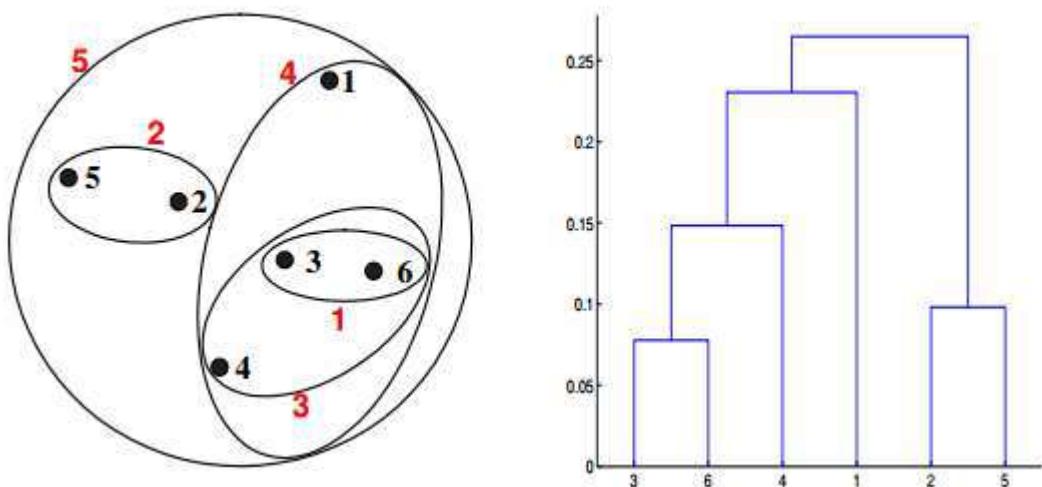
A.



B.
C.



D.



Solution: (C)

For the group average version of hierarchical clustering, the proximity of two clusters is defined to be the average of the pairwise proximities between all pairs of points in the different clusters. This is an intermediate approach between MIN and MAX. This is expressed by the following equation:

$$\text{proximity}(\text{cluster}_1, \text{cluster}_2) = \sum_{\substack{p_1 \in \text{cluster}_1 \\ p_2 \in \text{cluster}_2}} \frac{\text{proximity}(p_1, p_2)}{\text{size}(\text{cluster}_1) * \text{size}(\text{cluster}_2)}$$

Here, the distance between some clusters. $\text{dist}(\{3, 6, 4\}, \{1\}) = (0.2218 + 0.3688 + 0.2347)/(3 * 1) = 0.2751$. $\text{dist}(\{2, 5\}, \{1\}) = (0.2357 + 0.3421)/(2 * 1) = 0.2889$. $\text{dist}(\{3, 6, 4\}, \{2, 5\}) = (0.1483 + 0.2843 + 0.2540 + 0.3921 + 0.2042 + 0.2932)/(6 * 1) = 0.2637$. Because $\text{dist}(\{3, 6, 4\}, \{2, 5\})$ is smaller than $\text{dist}(\{3, 6, 4\}, \{1\})$ and $\text{dist}(\{2, 5\}, \{1\})$, these two clusters are merged at the fourth stage

Q22. Given, six points with the following attributes:

point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

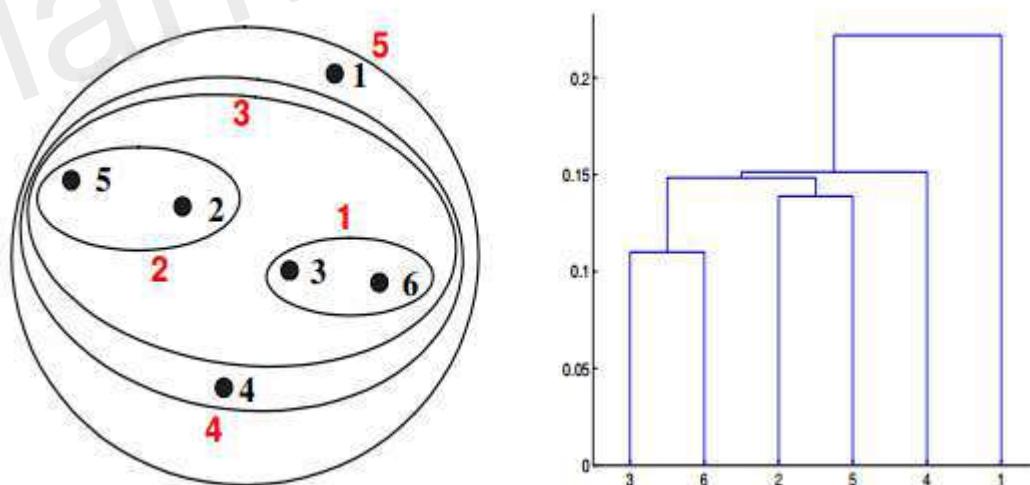
Table : X-Y coordinates of six points.

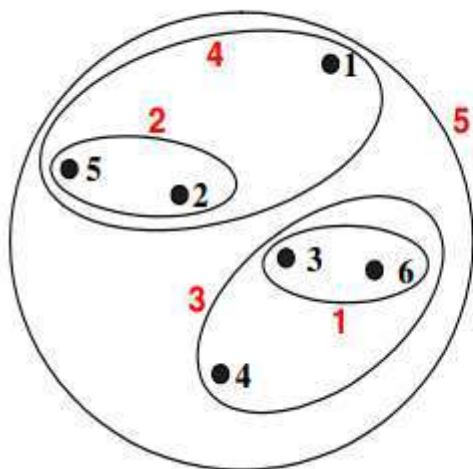
	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

Table : Distance Matrix for Six Points

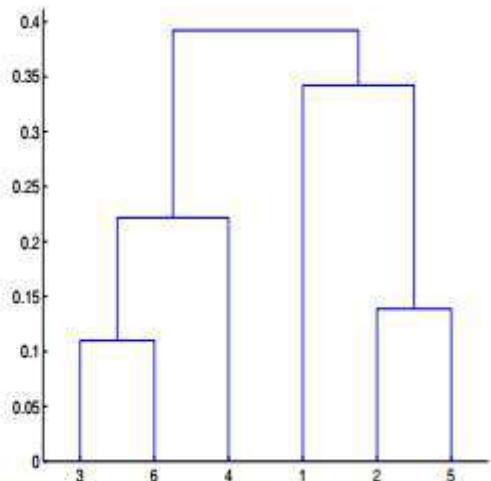
Which of the following clustering representations and dendrogram depicts the use of Ward's method proximity function in hierarchical clustering:

A.

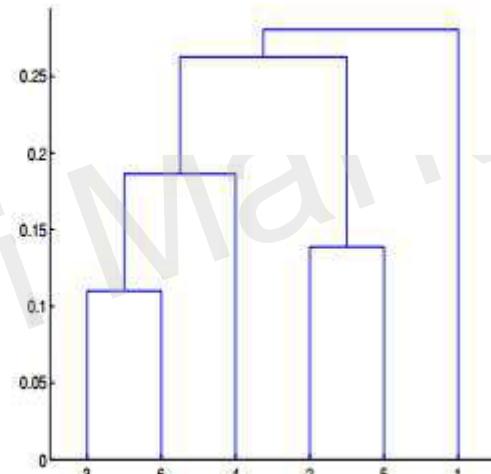
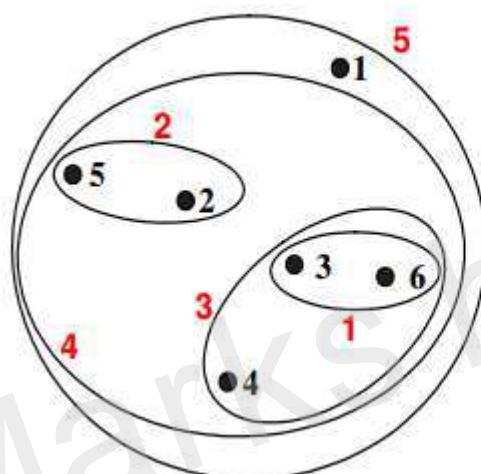




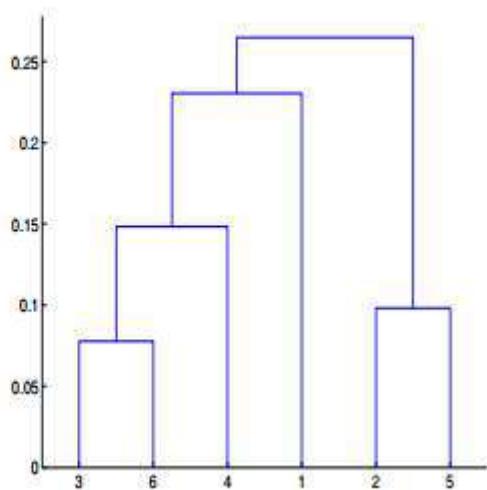
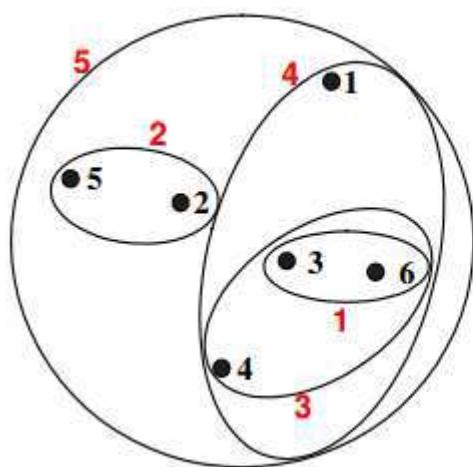
B.



C.



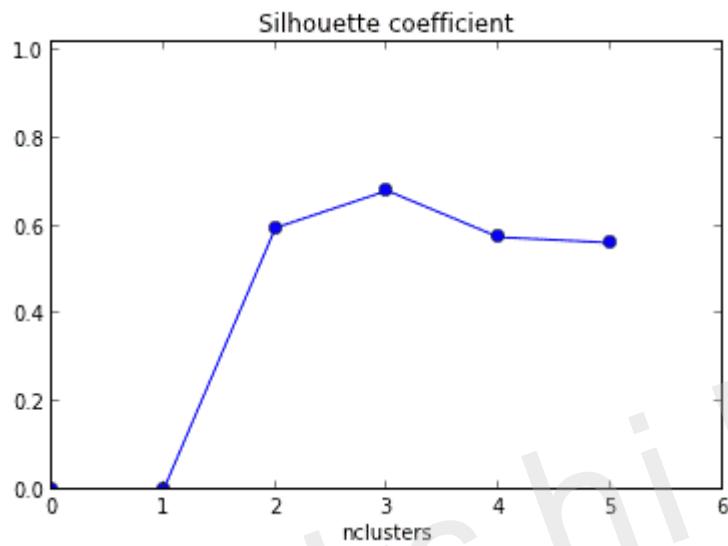
D.



Solution: (D)

Ward method is a centroid method. Centroid method calculates the proximity between two clusters by calculating the distance between the centroids of clusters. For Ward's method, the proximity between two clusters is defined as the increase in the squared error that results when two clusters are merged. The results of applying Ward's method to the sample data set of six points. The resulting clustering is somewhat different from those produced by MIN, MAX, and group average.

Q23. What should be the best choice of no. of clusters based on the following results:



- A. 1
- B. 2
- C. 3
- D. 4

Solution: (C)

The silhouette coefficient is a measure of how similar an object is to its own cluster compared to other clusters. Number of clusters for which silhouette coefficient is highest represents the best choice of the number of clusters.

Q24. Which of the following is/are valid iterative strategy for treating missing values before clustering analysis?

- A. Imputation with mean

- B. Nearest Neighbor assignment
- C. Imputation with Expectation Maximization algorithm
- D. All of the above

Solution: (C)

All of the mentioned techniques are valid for treating missing values before clustering analysis but only imputation with EM algorithm is iterative in its functioning.

Q25. K-Mean algorithm has some limitations. One of the limitation it has is, it makes hard assignments(A point either completely belongs to a cluster or not belongs at all) of points to clusters.

Note: Soft assignment can be consider as the probability of being assigned to each cluster: say K = 3 and for some point x_n , $p_1 = 0.7$, $p_2 = 0.2$, $p_3 = 0.1$)

Which of the following algorithm(s) allows soft assignments?

1. Gaussian mixture models
2. Fuzzy K-means

Options:

- A. 1 only
- B. 2 only
- C. 1 and 2
- D. None of these

Solution: (C)

Both, Gaussian mixture models and Fuzzy K-means allows soft assignments.

Q26. Assume, you want to cluster 7 observations into 3 clusters using K-Means clustering algorithm. After first iteration clusters, C1, C2, C3 has following observations:

C1: {(2,2), (4,4), (6,6)}

C2: {(0,4), (4,0)}

C3: {(5,5), (9,9)}

What will be the cluster centroids if you want to proceed for second iteration?

- A. C1: (4,4), C2: (2,2), C3: (7,7)
- B. C1: (6,6), C2: (4,4), C3: (9,9)
- C. C1: (2,2), C2: (0,0), C3: (5,5)
- D. None of these

Solution: (A)

Finding centroid for data points in cluster C1 = $((2+4+6)/3, (2+4+6)/3) = (4, 4)$

Finding centroid for data points in cluster C2 = $((0+4)/2, (4+0)/2) = (2, 2)$

Finding centroid for data points in cluster C3 = $((5+9)/2, (5+9)/2) = (7, 7)$

Hence, C1: (4,4), C2: (2,2), C3: (7,7)

Q27. Assume, you want to cluster 7 observations into 3 clusters using K-Means clustering algorithm. After first iteration clusters, C1, C2, C3 has following observations:

C1: {(2,2), (4,4), (6,6)}

C2: {(0,4), (4,0)}

C3: {(5,5), (9,9)}

What will be the Manhattan distance for observation (9, 9) from cluster centroid C1. In second iteration.

- A. 10
- B. $5\sqrt{2}$
- C. $13\sqrt{2}$
- D. None of these

Solution: (A)

Manhattan distance between centroid C1 i.e. (4, 4) and (9, 9) = $(9-4) + (9-4) = 10$

Q28. If two variables V1 and V2, are used for clustering. Which of the following are true for K means clustering with k =3?

1. If V1 and V2 has a correlation of 1, the cluster centroids will be in a straight line
2. If V1 and V2 has a correlation of 0, the cluster centroids will be in straight line

Options:

- A. 1 only
- B. 2 only
- C. 1 and 2
- D. None of the above

Solution: (A)

If the correlation between the variables V1 and V2 is 1, then all the data points will be in a straight line. Hence, all the three cluster centroids will form a straight line as well.

Q29. Feature scaling is an important step before applying K-Mean algorithm. What is reason behind this?

- A. In distance calculation it will give the same weights for all features
- B. You always get the same clusters. If you use or don't use feature scaling
- C. In Manhattan distance it is an important step but in Euclidian it is not
- D. None of these

Solution; (A)

Feature scaling ensures that all the features get same weight in the clustering analysis. Consider a scenario of clustering people based on their weights (in KG) with range 55-110 and height (in inches) with range 5.6 to 6.4. In this case, the clusters produced without scaling can be very misleading as the range of weight is much higher than that of height. Therefore, its necessary to bring them to same scale so that they have equal weightage on the clustering result.

Q30. Which of the following method is used for finding optimal of cluster in K-Mean algorithm?

- A. Elbow method
- B. Manhattan method
- C. Ecludian mehthod
- D. All of the above
- E. None of these

Solution: (A)

Out of the given options, only elbow method is used for finding the optimal number of clusters. The elbow method looks at the percentage of variance explained as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't give much better modeling of the data.

Q31. What is true about K-Mean Clustering?

- 1. K-means is extremely sensitive to cluster center initializations
- 2. Bad initialization can lead to Poor convergence speed
- 3. Bad initialization can lead to bad overall clustering

Options:

- A. 1 and 3
- B. 1 and 2
- C. 2 and 3
- D. 1, 2 and 3

Solution: (D)

All three of the given statements are true. K-means is extremely sensitive to cluster center initialization. Also, bad initialization can lead to Poor convergence speed as well as bad overall clustering.

Q32. Which of the following can be applied to get good results for K-means algorithm corresponding to global minima?

1. Try to run algorithm for different centroid initialization
2. Adjust number of iterations
3. Find out the optimal number of clusters

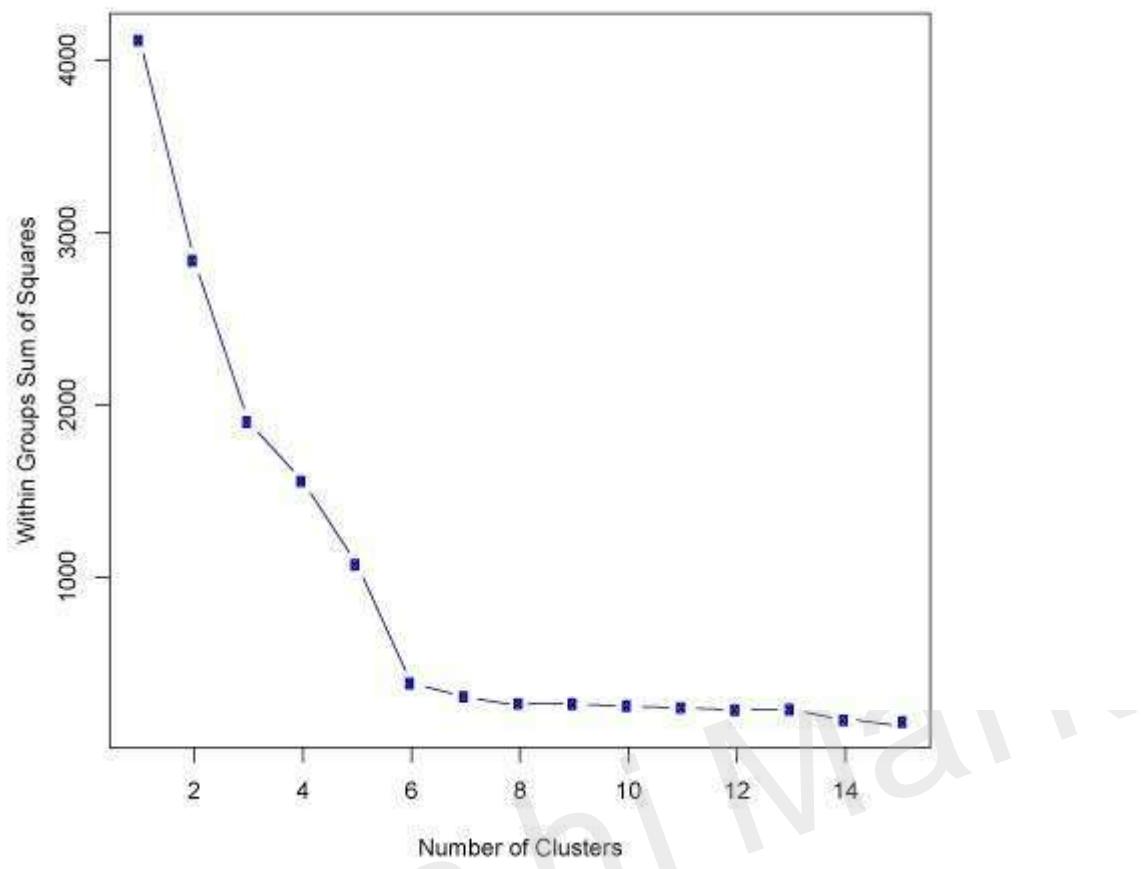
Options:

- A. 2 and 3
- B. 1 and 3
- C. 1 and 2
- D. All of above

Solution: (D)

All of these are standard practices that are used in order to obtain good clustering results.

Q33. What should be the best choice for number of clusters based on the following results:

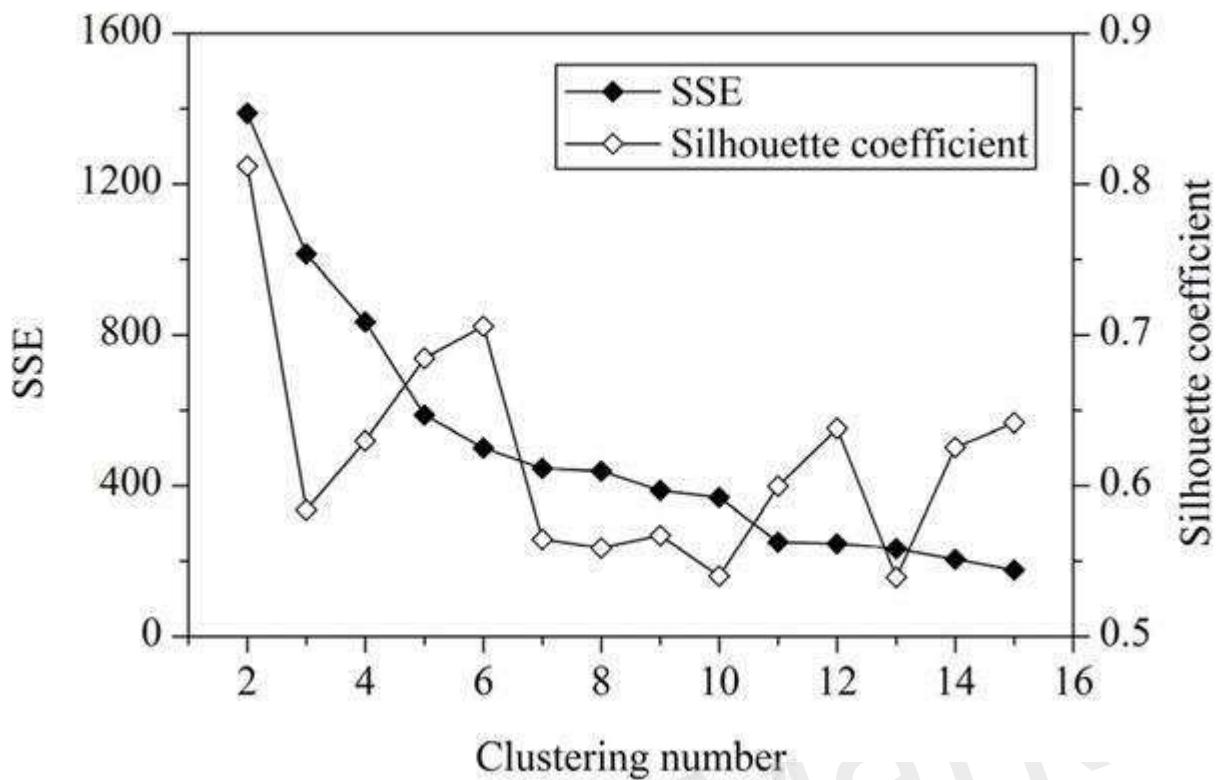


- A. 5
- B. 6
- C. 14
- D. Greater than 14

Solution: (B)

Based on the above results, the best choice of number of clusters using elbow method is 6.

Q34. What should be the best choice for number of clusters based on the following results:



- A. 2
 B. 4
 C. 6
 D. 8

Solution: (C)

Generally, a higher average silhouette coefficient indicates better clustering quality. In this plot, the optimal clustering number of grid cells in the study area should be 2, at which the value of the average silhouette coefficient is highest. However, the SSE of this clustering solution ($k = 2$) is too large. At $k = 6$, the SSE is much lower. In addition, the value of the average silhouette coefficient at $k = 6$ is also very high, which is just lower than $k = 2$. Thus, the best choice is $k = 6$.

Q35. Which of the following sequences is correct for a K-Means algorithm using Forgy method of initialization?

1. Specify the number of clusters
2. Assign cluster centroids randomly
3. Assign each data point to the nearest cluster centroid

4. Re-assign each point to nearest cluster centroids
5. Re-compute cluster centroids

Options:

- A. 1, 2, 3, 5, 4
- B. 1, 3, 2, 4, 5
- C. 2, 1, 3, 4, 5
- D. None of these

Solution: (A)

The methods used for initialization in K means are Forgy and Random Partition. The Forgy method randomly chooses k observations from the data set and uses these as the initial means. The Random Partition method first randomly assigns a cluster to each observation and then proceeds to the update step, thus computing the initial mean to be the centroid of the cluster's randomly assigned points.

Q36. If you are using Multinomial mixture models with the expectation-maximization algorithm for clustering a set of data points into two clusters, which of the assumptions are important:

- A. All the data points follow two Gaussian distribution
- B. All the data points follow n Gaussian distribution ($n > 2$)
- C. All the data points follow two multinomial distribution
- D. All the data points follow n multinomial distribution ($n > 2$)

Solution: (C)

In EM algorithm for clustering its essential to choose the same no. of clusters to classify the data points into as the no. of different distributions they are expected to be generated from and also the distributions must be of the same type.

Q37. Which of the following is/are not true about Centroid based K-Means clustering algorithm and Distribution based expectation-maximization clustering algorithm:

1. Both starts with random initializations
2. Both are iterative algorithms

3. Both have strong assumptions that the data points must fulfill
4. Both are sensitive to outliers
5. Expectation maximization algorithm is a special case of K-Means
6. Both requires prior knowledge of the no. of desired clusters
7. The results produced by both are non-reproducible.

Options:

- A. 1 only
- B. 5 only
- C. 1 and 3
- D. 6 and 7
- E. 4, 6 and 7
- F. None of the above

Solution: (B)

All of the above statements are true except the 5th as instead K-Means is a special case of EM algorithm in which only the centroids of the cluster distributions are calculated at each iteration.

Q38. Which of the following is/are not true about DBSCAN clustering algorithm:

1. For data points to be in a cluster, they must be in a distance threshold to a core point
2. It has strong assumptions for the distribution of data points in dataspace
3. It has substantially high time complexity of order $O(n^3)$
4. It does not require prior knowledge of the no. of desired clusters
5. It is robust to outliers

Options:

- A. 1 only
- B. 2 only
- C. 4 only
- D. 2 and 3
- E. 1 and 5

F. 1, 3 and 5

Solution: (D)

- DBSCAN can form a cluster of any arbitrary shape and does not have strong assumptions for the distribution of data points in the dataspace.
- DBSCAN has a low time complexity of order $O(n \log n)$ only.

Q39. Which of the following are the high and low bounds for the existence of F-Score?

- A. [0,1]
- B. (0,1)
- C. [-1,1]
- D. None of the above

Solution: (A)

The lowest and highest possible values of F score are 0 and 1 with 1 representing that every data point is assigned to the correct cluster and 0 representing that the precision and/ or recall of the clustering analysis are both 0. In clustering analysis, high value of F score is desired.

Q40. Following are the results observed for clustering 6000 data points into 3 clusters: A, B and C:

		Actual			
		A	B	C	SUM
Predicted	A	600	400	200	1200
	B	1000	1200	200	2400
	C	400	400	1600	2400
	SUM	2000	2000	2000	

What is the F_1 -Score with respect to cluster B?

A. 3

B. 4

C. 5

D. 6

Solution: (D)

Here,

True Positive, $TP = 1200$

True Negative, $TN = 600 + 1600 = 2200$

False Positive, $FP = 1000 + 200 = 1200$

False Negative, $FN = 400 + 400 = 800$

Therefore,

$$\text{Precision} = TP / (TP + FP) = 0.5$$

$$\text{Recall} = TP / (TP + FN) = 0.6$$

Hence,

$$F_1 = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{recall}) = 0.54 \sim 0.5$$

Skill test Questions and Answers

1) [True or False] k-NN algorithm does more computation on test time rather than train time.

- A) TRUE
- B) FALSE

Solution: A

The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples.

In the testing phase, a test point is classified by assigning the label which are most frequent among the k training samples nearest to that query point – hence higher computation.

2) In the image below, which would be the best value for k assuming that the algorithm you are using is k-Nearest Neighbor.

- A) 3
- B) 10
- C) 20
- D) 50

Solution: B

Validation error is the least when the value of k is 10. So it is best to use this value of k

3) Which of the following distance metric can not be used in k-NN?

- A) Manhattan
- B) Minkowski
- C) Tanimoto
- D) Jaccard
- E) Mahalanobis
- F) All can be used

Solution: F

All of these distance metric can be used as a distance metric for k-NN.

4) Which of the following option is true about k-NN algorithm?

- A) It can be used for classification
- B) It can be used for regression
- C) It can be used in both classification and regression

Solution: C

We can also use k-NN for regression problems. In this case the prediction can be based on the mean or the median of the k-most similar instances.

5) Which of the following statement is true about k-NN algorithm?

1. k-NN performs much better if all of the data have the same scale
2. k-NN works well with a small number of input variables (p), but struggles when the number of inputs is very large
3. k-NN makes no assumptions about the functional form of the problem being solved

- A) 1 and 2
- B) 1 and 3
- C) Only 1
- D) All of the above

Solution: D

The above mentioned statements are assumptions of kNN algorithm

6) Which of the following machine learning algorithm can be used for imputing missing values of both categorical and continuous variables?

- A) K-NN
- B) Linear Regression
- C) Logistic Regression

Solution: A

k-NN algorithm can be used for imputing missing value of both categorical and continuous variables.

7) Which of the following is true about Manhattan distance?

- A) It can be used for continuous variables
- B) It can be used for categorical variables
- C) It can be used for categorical as well as continuous
- D) None of these

Solution: A

Manhattan Distance is designed for calculating the distance between real valued features.

8) Which of the following distance measure do we use in case of categorical variables in k-NN?

- 1. Hamming Distance
 - 2. Euclidean Distance
 - 3. Manhattan Distance
- A) 1
 - B) 2
 - C) 3
 - D) 1 and 2
 - E) 2 and 3
 - F) 1,2 and 3

Solution: A

Both Euclidean and Manhattan distances are used in case of continuous variables, whereas hamming distance is used in case of categorical variable.

9) Which of the following will be Euclidean Distance between the two data point A(1,3) and B(2,3)?

- A) 1
- B) 2
- C) 4
- D) 8

Solution: A

$$\sqrt{(1-2)^2 + (3-3)^2} = \sqrt{1^2 + 0^2} = 1$$

10) Which of the following will be Manhattan Distance between the two data point A(1,3) and B(2,3)?

- A) 1
- B) 2
- C) 4
- D) 8

Solution: A

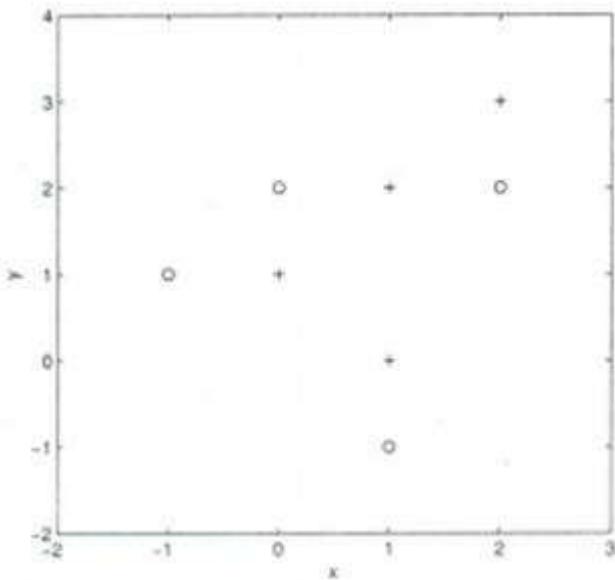
$$\sqrt{\text{mod}((1-2)) + \text{mod}((3-3))} = \sqrt{1 + 0} = 1$$

Context: 11-12

Suppose, you have given the following data where x and y are the 2 input variables and Class is the dependent variable.

x	y	Class
-1	1	-
0	1	+
0	2	-
1	-1	-
1	0	+
1	2	+
2	2	-
2	3	+

Below is a scatter plot which shows the above data in 2D space.



11) Suppose, you want to predict the class of new data point $x=1$ and $y=1$ using euclidian distance in 3-NN. In which class this data point belong to?

- A) + Class
- B) – Class
- C) Can't say
- D) None of these

Solution: A

All three nearest point are of +class so this point will be classified as +class.

12) In the previous question, you are now want use 7-NN instead of 3-KNN which of the following $x=1$ and $y=1$ will belong to?

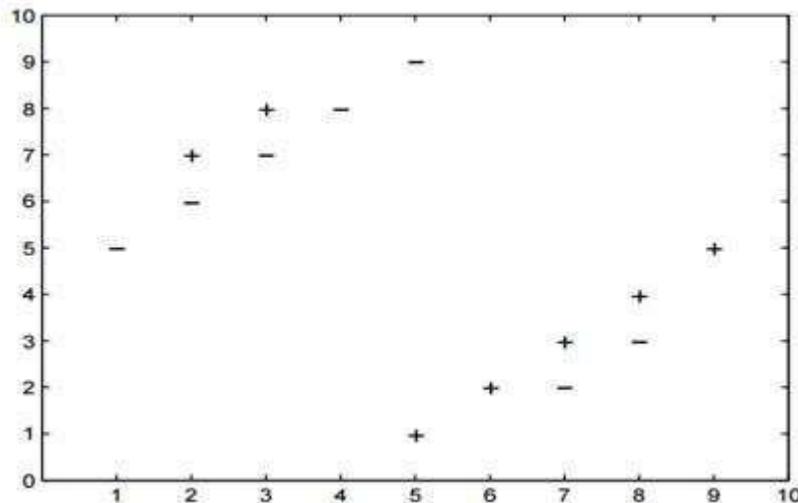
- A) + Class
- B) – Class
- C) Can't say

Solution: B

Now this point will be classified as – class because there are 4 – class and 3 +class point are in nearest circle.

Context 13-14:

Suppose you have given the following 2-class data where "+" represent a positive class and "-" is represent negative class.



13) Which of the following value of k in k-NN would minimize the leave one out cross validation accuracy?

- A) 3
- B) 5
- C) Both have same
- D) None of these

Solution: B

5-NN will have least leave one out cross validation error.

14) Which of the following would be the leave one out cross validation accuracy for k=5?

- A) 2/14
- B) 4/14
- C) 6/14
- D) 8/14
- E) None of the above

Solution: E

In 5-NN we will have 10/14 leave one out cross validation accuracy.

15) Which of the following will be true about k in k-NN in terms of Bias?

- A) When you increase the k the bias will be increases
- B) When you decrease the k the bias will be increases
- C) Can't say
- D) None of these

Solution: A

large K means simple model, simple model always consider as high bias

16) Which of the following will be true about k in k-NN in terms of variance?

- A) When you increase the k the variance will increases
- B) When you decrease the k the variance will increases
- C) Can't say
- D) None of these

Solution: B

Simple model will be consider as less variance model

17) The following two distances(Euclidean Distance and Manhattan Distance) have given to you which generally we used in K-NN algorithm. These distance are between two points A(x_1, y_1) and B(x_2, Y_2).

Your task is to tag the both distance by seeing the following two graphs. Which of the following option is true about below graph ?



- A) Left is Manhattan Distance and right is euclidean Distance
- B) Left is Euclidean Distance and right is Manhattan Distance
- C) Neither left or right are a Manhattan Distance
- D) Neither left or right are a Euclidian Distance

Solution: B

Left is the graphical depiction of how euclidean distance works, whereas right one is of Manhattan distance.

18) When you find noise in data which of the following option would you consider in k-NN?

- A) I will increase the value of k
- B) I will decrease the value of k
- C) Noise can not be dependent on value of k
- D) None of these

Solution: A

To be more sure of which classifications you make, you can try increasing the value of k.

19) In k-NN it is very likely to overfit due to the curse of dimensionality. Which of the following option would you consider to handle such problem?

- 1. Dimensionality Reduction
 - 2. Feature selection
-
- A) 1
 - B) 2
 - C) 1 and 2
 - D) None of these

Solution: C

In such case you can use either dimensionality reduction algorithm or the feature selection algorithm

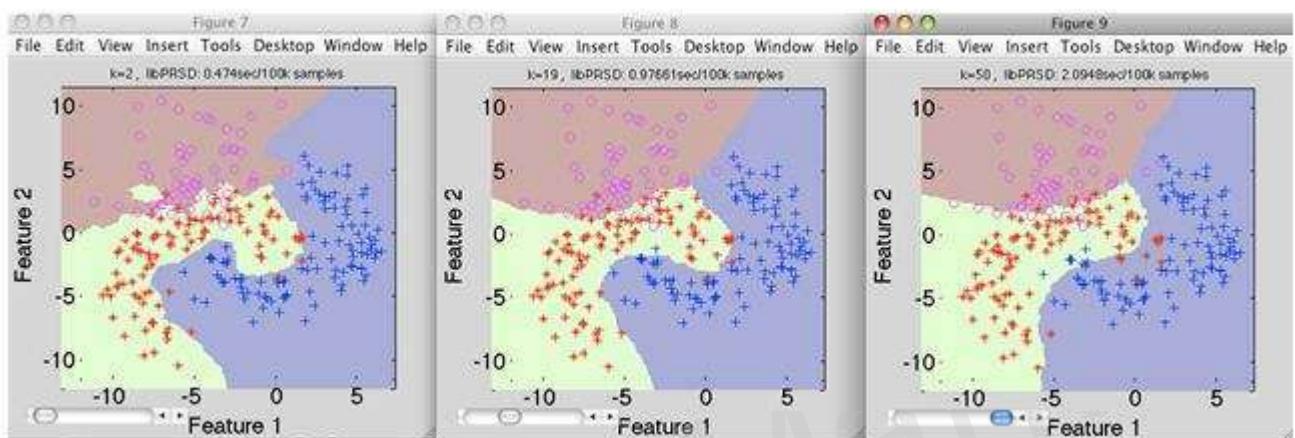
20) Below are two statements given. Which of the following will be true both statements?

- 1. k-NN is a memory-based approach is that the classifier immediately adapts as we collect new training data.
 - 2. The computational complexity for classifying new samples grows linearly with the number of samples in the training dataset in the worst-case scenario.
-
- A) 1
 - B) 2
 - C) 1 and 2
 - D) None of these

Solution: C

Both are true and self explanatory

21) Suppose you have given the following images(1 left, 2 middle and 3 right), Now your task is to find out the value of k in k-NN in each image where k1 is for 1st, k2 is for 2nd and k3 is for 3rd figure.

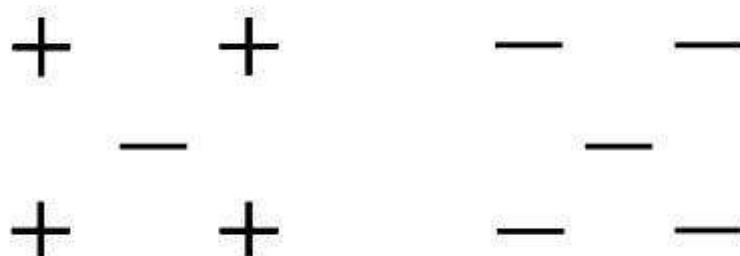


- A) $k_1 > k_2 > k_3$
- B) $k_1 < k_2$
- C) $k_1 = k_2 = k_3$
- D) None of these

Solution: D

Value of k is highest in k3, whereas in k1 it is lowest

22) Which of the following value of k in the following graph would you give least leave one out cross validation accuracy?



- A) 1

- B) 2
- C) 3
- D) 5

Solution: B

If you keep the value of k as 2, it gives the lowest cross validation accuracy. You can try this out yourself.

23) A company has build a kNN classifier that gets 100% accuracy on training data. When they deployed this model on client side it has been found that the model is not at all accurate. Which of the following thing might gone wrong?

Note: Model has successfully deployed and no technical issues are found at client side except the model performance

- A) It is probably a overfitted model
- B) It is probably a underfitted model
- C) Can't say
- D) None of these

Solution: A

In an overfitted module, it seems to be performing well on training data, but it is not generalized enough to give the same results on a new data.

24) You have given the following 2 statements, find which of these option is/are true in case of k-NN?

1. In case of very large value of k, we may include points from other classes into the neighborhood.
2. In case of too small value of k the algorithm is very sensitive to noise

- A) 1
- B) 2
- C) 1 and 2
- D) None of these

Solution: C

Both the options are true and are self explanatory.

25) Which of the following statements is true for k-NN classifiers?

- A) The classification accuracy is better with larger values of k
- B) The decision boundary is smoother with smaller values of k
- C) The decision boundary is linear
- D) k-NN does not require an explicit training step

Solution: D

Option A: This is not always true. You have to ensure that the value of k is not too high or not too low.

Option B: This statement is not true. The decision boundary can be a bit jagged

Option C: Same as option B

Option D: This statement is true

26) True-False: It is possible to construct a 2-NN classifier by using the 1-NN classifier?

- A) TRUE
- B) FALSE

Solution: A

You can implement a 2-NN classifier by ensembling 1-NN classifiers

27) In k-NN what will happen when you increase/decrease the value of k?

- A) The boundary becomes smoother with increasing value of K
- B) The boundary becomes smoother with decreasing value of K
- C) Smoothness of boundary doesn't depend on value of K
- D) None of these

Solution: A

The decision boundary would become smoother by increasing the value of K

28) Following are the two statements given for k-NN algorithm, which of the statement(s)

is/are true?

1. We can choose optimal value of k with the help of cross validation

2. Euclidean distance treats each feature as equally important
- A) 1
 - B) 2
 - C) 1 and 2
 - D) None of these

Solution: C

Both the statements are true

Context 29-30:

Suppose, you have trained a k-NN model and now you want to get the prediction on test data. Before getting the prediction suppose you want to calculate the time taken by k-NN for predicting the class for test data.

Note: Calculating the distance between 2 observation will take D time.

29) What would be the time taken by 1-NN if there are N(Very large) observations in test data?

- A) $N \cdot D$
- B) $N \cdot D^2$
- C) $(N \cdot D)/2$
- D) None of these

Solution: A

The value of N is very large, so option A is correct

30) What would be the relation between the time taken by 1-NN,2-NN,3-NN.

- A) 1-NN > 2-NN > 3-NN
- B) 1-NN < 2-NN < 3-NN
- C) 1-NN ~ 2-NN ~ 3-NN
- D) None of these

Solution: C

The training time for any value of k in kNN algorithm is the same.

1. A project team performed a feature selection procedure on the full data set and reduced their large

feature set to a smaller set. Then they split the data into test and training portions. They built their model on training data using several different model settings, and report the best test error they achieved. Which of the following is TRUE about the given experimental setup?

- a) Best setup
- b) Problematic setup
- c) Invalid setup
- d) Cannot be decided

Answer: (b) Problematic setup

(a) Using the full data for feature selection will leak information from the test examples into the model. The feature selection should be done exclusively using training and validation data not on test data.

(b) The best parameter setting should not be chosen based on the test error; this has the danger of overfitting to the test data. They should have used validation data and use the test data only in the final evaluation step.

2. If we increase the k value in k-nearest neighbor, the model will ____ the bias and ____ the variance.

- a) Decrease, Decrease
- b) Increase, Decrease
- c) Decrease, Increase
- d) Increase, Increase

Answer: (b) Increase, Decrease

When K increases to a large value, the model becomes simplest. All test data point will belong to the same class: the majority class. This is under-fit, that is, high bias and low variance.

Bias-Variance tradeoff

The bias is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs. In other words, model with high bias pays very little attention to the training data and oversimplifies the model.

The variance is an error from sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the

random noise in the training data, rather than the intended outputs. In other words, model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before. [Source: [Refer here](#)]

3. For a large k value the k-nearest neighbor model becomes _____ and _____ .

- a) Complex model, Overfit
- b) Complex model, Underfit
- c) Simple model, Underfit
- d) Simple model, Overfit

Answer: (c) Simple model, Underfit

When K increases to inf, the model is simplest. All test data point will belong to the same class: the majority class. This is under-fit, that is, high bias and low variance.

knn classification is an averaging operation. To come to a decision, the labels of K nearest neighbour samples are averaged. The standard deviation (or the variance) of the output of averaging decreases as the number of samples increases. In the case K==N (you select K as large as the size of the dataset), variance becomes zero.

Underfitting means the model does not fit, in other words, does not predict, the (training) data very well.

Overfitting means that the model predicts the (training) data too well. It is too good to be true. If the new data point comes in, the prediction may be wrong.

4. When we have a real-valued input attribute during decision-tree learning, what would be the impact multi-way split with one branch for each of the distinct values of the attribute?

- a) It is too computationally expensive.
- b) It would probably result in a decision tree that scores badly on the training set and a test set.
- c) It would probably result in a decision tree that scores well on the training set but badly on a test set.
- d) It would probably result in a decision tree that scores well on a test set but badly on a training set.

Answer: (c) It would probably result in a decision tree that scores well on the training set but badly on a test set

It is usual to make only binary splits because multiway splits break the data into small subsets too quickly. This causes a bias towards splitting predictors with many classes since they are more likely to produce relatively pure child nodes, which results in overfitting. [For more, [refer here](#)]

5. The VC dimension of a Perceptron is ____ the VC dimension of a simple linear SVM.

- a) Larger than
- b) Smaller than
- c) Same as
- d) Not at all related

Answer: (c) Same as

Both Perceptron and linear SVM are linear discriminators (i.e. a line in 2D space or a plane in 3D space.), so they should have the same VC dimension.

VC dimension

The Vapnik-Chervonenkis (VC) dimension is a measure of the capacity (complexity, expressive power, richness, or flexibility) of a space of functions that can be learned by a statistical binary classification algorithm. It is defined as the cardinality of the largest set of points that the algorithm can shatter. [[Wikipedia](#)]

1. The process of forming general concept definitions from examples of concepts to be learned.

- A. Deduction
- B. abduction
- C. induction**
- D. conjunction

2. Computers are best at learning

- A. facts.**
- B. concepts.**
- C. procedures.
- D. principles.

3. Data used to build a data mining model.

- A. validation data
- B. training data**
- C. test data
- D. hidden data

4. Supervised learning and unsupervised clustering both require at least one

- A. hidden attribute.**
- B. output attribute.
- C. input attribute.**
- D. categorical attribute.

5. Supervised learning differs from unsupervised clustering in that supervised learning requires

- A. at least one input attribute.
- B. input attributes to be categorical.**
- C. at least one output attribute.**
- D. ouput attributes to be categorical.

6. A regression model in which more than one independent variable is used to predict the dependent variable is called

- A. a simple linear regression model
- B. a multiple regression models**
- C. an independent model**
- D. none of the above

7. A term used to describe the case when the independent variables in a multiple regression model are correlated is
- A. regression
 - B. correlation
 - C. multicollinearity
 - D. none of the above
8. A multiple regression model has the form: $y = 2 + 3x_1 + 4x_2$. As x_1 increases by 1 unit (holding x_2 constant), y will
- A. increase by 3 units
 - B. decrease by 3 units
 - C. increase by 4 units
 - D. decrease by 4 units
9. A multiple regression model has
- A. only one independent variable
 - B. more than one dependent variable
 - C. more than one independent variable
 - D. none of the above
10. A measure of goodness of fit for the estimated regression equation is the
- A. multiple coefficient of determination
 - B. mean square due to error
 - C. mean square due to regression
 - D. none of the above
11. The adjusted multiple coefficient of determination accounts for
- A. the number of dependent variables in the model
 - B. the number of independent variables in the model
 - C. unusually large predictors
 - D. none of the above
12. The multiple coefficient of determination is computed by
- A. dividing SSR by SST
 - B. dividing SST by SSR
 - C. dividing SST by SSE
 - D. none of the above
13. For a multiple regression model, $SST = 200$ and $SSE = 50$. The multiple coefficient of determination is
- A. 0.25
 - B. 4.00
 - C. 0.75
 - D. none of the above

14. A nearest neighbor approach is best used

- A. with large-sized datasets.
- B. when irrelevant attributes have been removed from the data.
- C. when a generalized model of the data is desireable.
- D. when an explanation of what has been found is of primary importance.

15. Determine which is the best approach for each problem.

- A. *supervised learning*
- B. *unsupervised clustering*
- C. *data query*

1. What is the average weekly salary of all female employees under forty years of age? **(C)**
2. Develop a profile for credit card customers likely to carry an average monthly balance of more than \$1000.00. **(A)**
3. Determine the characteristics of a successful used car salesperson. **(A)**
4. What attribute similarities group customers holding one or several insurance policies? **(A)**
5. Do meaningful attribute relationships exist in a database containing information about credit card customers? **(B)**
6. Do single men play more golf than married men? **(C)**
7. Determine whether a credit card transaction is valid or fraudulent **(A)**

16. Another name for an output attribute.

- A. predictive variable
- B. **independent variable**
- C. estimated variable
- D. dependent variable

17. Classification problems are distinguished from estimation problems in that

- A. classification problems require the output attribute to be numeric.
- B. classification problems require the output attribute to be categorical.
- C. **classification problems do not allow an output attribute.**
- D. classification problems are designed to predict future outcome.

18. Which statement is true about prediction problems?

- A. The output attribute must be categorical.
- B. The output attribute must be numeric.
- C. The resultant model is designed to determine future outcomes.
- D. **The resultant model is designed to classify current behavior.**

19. Which statement about outliers is true?

- A. Outliers should be identified and removed from a dataset.
- B. Outliers should be part of the training dataset but should not be present in the test data.
- C. Outliers should be part of the test dataset but should not be present in the training data.
- D. **The nature of the problem determines how outliers are used.**
- E. More than one of a,b,c or d is true.

20. Which statement is true about neural network and linear regression models?

- A. **Both models require input attributes to be numeric.**
- B. Both models require numeric attributes to range between 0 and 1.
- C. The output of both models is a categorical attribute value.
- D. Both techniques build models whose output is determined by a linear sum of weighted input attribute values.
- E. More than one of a,b,c or d is true.

21. Which of the following is a common use of unsupervised clustering?

- A. **detect outliers**
- B. determine a best set of input attributes for supervised learning
- C. evaluate the likely performance of a supervised learner model
- D. determine if meaningful relationships can be found in a dataset
- E. All of a,b,c, and d are common uses of unsupervised clustering.

22. The average positive difference between computed and desired outcome values.

- A. root mean squared error
- B. mean squared error
- C. mean absolute error
- D. **mean positive error**

23. Selecting data so as to assure that each class is properly represented in both the training and test set.

- A. cross validation
- B. **stratification**
- C. verification
- D. bootstrapping

24. The standard error is defined as the square root of this computation.

- A. **The sample variance divided by the total number of sample instances.**
- B. The population variance divided by the total number of sample instances.
- C. The sample variance divided by the sample mean.
- D. The population variance divided by the sample mean.

25. Data used to optimize the parameter settings of a supervised learner model.

- A. training
- B. test
- C. verification
- D. validation

26. Bootstrapping allows us to

- A. choose the same training instance several times.
- B. choose the same test set instance several times.
- C. build models with alternative subsets of the training data several times.
- D. test a model with alternative subsets of the test data several times.

27. The correlation between the number of years an employee has worked for a company and the salary of the employee is 0.75. What can be said about employee salary and years worked?

- A. There is no relationship between salary and years worked.
- B. Individuals that have worked for the company the longest have higher salaries.
- C. Individuals that have worked for the company the longest have lower salaries.
- D. The majority of employees have been with the company a long time.
- E. The majority of employees have been with the company a short period of time.

28. The correlation coefficient for two real-valued attributes is -0.85 . What does this value tell you?

- A. The attributes are not linearly related.
- B. As the value of one attribute increases the value of the second attribute also increases.
- C. As the value of one attribute decreases the value of the second attribute increases.
- D. The attributes show a curvilinear relationship.

29. The average squared difference between classifier predicted output and actual output.

- A. mean squared error
- B. root mean squared error
- C. mean absolute error
- D. mean relative error

30. Simple regression assumes a _____ relationship between the input attribute and output attribute.

- A. linear
- B. quadratic
- C. reciprocal
- D. inverse

31. Regression trees are often used to model _____ data.

- A. linear
- B. nonlinear
- C. categorical
- D. symmetrical

32. The leaf nodes of a model tree are

- A. averages of numeric output attribute values.
- B. nonlinear regression equations.
- C. linear regression equations.
- D. sums of numeric output attribute values.

33. Logistic regression is a _____ regression technique that is used to model data having a _____ outcome.

- A. linear, numeric
- B. linear, binary
- C. nonlinear, numeric
- D. nonlinear, binary

34. This technique associates a conditional probability value with each data instance.

- A. linear regression
- B. logistic regression
- C. simple regression
- D. multiple linear regression

35. This supervised learning technique can process both numeric and categorical input attributes.

- A. linear regression
- B. Bayes classifier
- C. logistic regression
- D. backpropagation learning

36. With Bayes classifier, missing data items are

- A. treated as equal compares.
- B. treated as unequal compares.
- C. replaced with a default value.
- D. ignored.

37. This clustering algorithm merges and splits nodes to help modify nonoptimal partitions.

- A. agglomerative clustering
- B. expectation maximization
- C. conceptual clustering
- D. K-Means clustering

38. This clustering algorithm initially assumes that each data instance represents a single cluster.

- A. agglomerative clustering
- B. conceptual clustering
- C. K-Means clustering
- D. expectation maximization

39. This unsupervised clustering algorithm terminates when mean values computed for the current iteration of the algorithm are identical to the computed mean values for the previous iteration.

- A. agglomerative clustering
- B. conceptual clustering
- C. K-Means clustering
- D. expectation maximization

40. Machine learning techniques differ from statistical techniques in that machine learning methods

- A. typically assume an underlying distribution for the data.
- B. are better able to deal with missing and noisy data.
- C. are not able to explain their behavior.
- D. have trouble with large-sized datasets.

1. Which of the following would be more appropriate to be replaced with question mark in the following figure?

- a) Data Analysis
- b) Data Science**
- c) Descriptive Analytics
- d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: Data Science is a multidisciplinary which involves extraction of knowledge from large volumes of data that are structured or unstructured.

2. Point out the correct statement.

- a) Raw data is original source of data**
- b) Preprocessed data is original source of data
- c) Raw data is the data obtained after processing steps
- d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: Accounting programs are prototypical examples of data processing applications.

3. Which of the following is performed by Data Scientist?

- a) Define the question
- b) Create reproducible code
- c) Challenge results
- d) All of the mentioned**

[View Answer](#)

Answer: d

Explanation: A data scientist is a job title for an employee or business intelligence (BI) consultant who excels at analyzing data, particularly large amounts of data.

4. Which of the following is the most important language for Data Science?

- a) Java
- b) Ruby
- c) R**
- d) None of the mentioned

[View Answer](#)

Answer: c

Explanation: R is free software for statistical computing and analysis.

5. Point out the wrong statement.

- a) Merging concerns combining datasets on the same observations to produce a result with more variables
- b) Data visualization** is the organization of information according to preset specifications
- c) Subsetting can be used to select and exclude variables and observations
- d) All of the mentioned

[View Answer](#)

Answer: b

Explanation: Data formatting is the organization of information according to preset specifications.

6. Which of the following approach should be used to ask Data Analysis question?

- a) Find only one solution for particular problem
- b) Find out the question which is to be answered**

- c) Find out answer from dataset without asking question
d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: Data analysis has multiple facets and approaches.

7. Which of the following is one of the key data science skills?

- a) Statistics
b) Machine Learning
c) Data Visualization
d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: Data visualization is the presentation of data in a pictorial or graphical format.

8. Which of the following is a key characteristic of a hacker?

- a) Afraid to say they don't know the answer
b) Willing to find answers on their own
c) Not Willing to find answers on their own
d) All of the mentioned

[View Answer](#)

Answer: b

Explanation: Hacker is an expert at programming and solving problems with a computer.

9. Which of the following is characteristic of Processed Data?

- a) Data is not ready for analysis
b) All steps should be noted
c) Hard to use for data analysis
d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: Processing includes merging, summarizing and subsetting data.

10. Raw data should be processed only one time.

- a) True
b) False

[View Answer](#)

Answer: b

Explanation: Raw data may only need to be processed once.

Sanfoundry Global Education & Learning Series – Data Science.

This set of Data Science Multiple Choice Questions & Answers (MCQs) focuses on “ToolBox Overview”.

1. Which of the following principle is incorrectly represented in the below figure?

- a) Show Comparisons
- b) Integrate Evidence
- c) Describe Evidence
- d) None of the mentioned

[View Answer](#)

Answer: d

Explanation: Principles of Analytical graphs are sequentially shown in the stepwise manner.

2. Point out the correct statement.

- a) Least square is an estimation tool
- b) Least square problems falls in to three categories
- c) Compound least square is one of the category of least square
- d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: The Method of Least Squares is a procedure to determine the best fit line to data.

3. How many principles of analytical graphs exist?

- a) 3
- b) 4
- c) 6
- d) None of the mentioned

[View Answer](#)

Answer: c

Explanation: Six Principles of Analytical Graphs are useful for data analysis.

4. Which of the following is not a step in data analysis?

- a) Obtain the data
- b) Clean the data
- c) EDA
- d) None of the mentioned

[View Answer](#)

Answer: d

Explanation: EDA stands for Exploratory Data Analysis.

5. Point out the wrong statement.

- a) Simple linear regression is equipped to handle more than one predictor
- b) Compound linear regression is not equipped to handle more than one predictor
- c) Linear regression consists of finding the best-fitting straight line through the points
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: Simple linear regression is equipped to handle more than one predictor.

6. Which of the following technique comes under practical machine learning?

- a) Bagging
- b) Boosting
- c) Forecasting
- d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: Boosting is an approach to machine learning based on the idea of creating a highly accurate predictor.

7. Data Products shown in the below figure is built using which programming language?

- a) S
- b) Python
- c) R
- d) Java

[View Answer](#)

Answer: c

Explanation: Products mentioned in the figure are web application frameworks written in R.

8. Which of the following technique is also referred to as Bagging?

- a) Bootstrap aggregating
- b) Bootstrap subsetting
- c) Bootstrap predicting
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: Bagging is used in statistical classification and regression.

9. Which of the following is characteristic of Raw Data?

- a) Data is ready for analysis
- b) Original version of data
- c) Easy to use for data analysis
- d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: Raw data is data that has not been processed for use.

10. Standard normal RVs are always labelled as Z.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: Standard normal RVs are often labelled as Z.

Q1. Which of the following CLI command can also be used to rename files?

- a) rm
- b) mv
- c) rm -r
- d) none of the mentioned

[View Answer](#)

Answer: b
Explanation: mv stands for move.

2. Point out the correct statement.

- a) CLI can help you to organize messages
- b) CLI can help you to organize files and folders
- c) Navigation of directory is possible using CLI
- d) None of the mentioned

[View Answer](#)

Answer: b
Explanation: CLI stands for Command Line Interface.

3. Which of the following command allows you to change directory to one level above your parent directory?

- a) cd
- b) cd..
- c) cd.
- d) none of the mentioned

[View Answer](#)

Answer: c
Explanation: cd stands for change directory.

4. Which of the following is not a CLI command?

- a) delete
- b) rm
- c) clear
- d) none of the mentioned

[View Answer](#)

Answer: a
Explanation: rm can be used to remove files and directories.

5. Point out the wrong statement.

- a) Command is the CLI command which does a specific task
- b) There is one and only flag for every command in CLI
- c) Flags are the options given to command for activating particular behaviour
- d) All of the mentioned

[View Answer](#)

Answer: b
Explanation: Depending on the command, there can be zero or more flags and arguments.

6. Which of the following systems record changes to a file over time?

- a) Record Control
- b) Version Control
- c) Forecast Control
- d) None of the mentioned

[View Answer](#)

Answer: b
Explanation: Version control is also known as revision control.

7. Which of the following is a revision control system?

- a) Git
- b) NumPy
- c) Slidify
- d) None of the mentioned

[View Answer](#)

Answer: a
Explanation: Git is a free and open source distributed version control system designed to handle everything from small to very large projects with speed and efficiency.

8. Which of the following command line environment is used for interacting with Git?

- a) GitHub
- b) Git Bash
- c) Git Boot
- d) All of the mentioned

[View Answer](#)

Answer: b

Explanation: Git for Windows provides a BASH emulation used to run Git from the command line.

9. Which of the following web hosting service use Git control system?

- a) GitHub
- b) Open Hash
- c) Git Bash
- d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: GitHub is a Web-based Git repository hosting service, which offers all of the distributed revision control and source code management (SCM) functionality of Git.

10. cp command can be used to copy the content of directories.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: -r flag should be used for copying the content.

1. Which of the following adds all new files to local repository?

- a) git add .
- b) git add -u
- c) git add -A
- d) none of the mentioned

[View Answer](#)

Answer: a

Explanation: You should do this before committing.

2. Point out the correct statement.

- a) You don't need GitHub to use Git
- b) CLI can help you to organize files and folders
- c) Navigation of directory is possible using CLI
- d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: CLI stands for Command Line Interface.

3. Which of the following command updates tracking for files that are modified?

- a) git add .
- b) git add -u
- c) git add -A
- d) none of the mentioned

[View Answer](#)

Answer: b

Explanation: The git add command adds a change in the working directory to the staging area.

4. Which of the following command is used to give a message description?

- a) git command -m
- b) git command -d
- c) git command -message
- d) none of the mentioned

[View Answer](#)

Answer: a

Explanation: This only updates your local repository.

5. Point out the wrong statement.

- a) You need GitHub to use Git
- b) GitHub allows you to share repositories with others
- c) GitHub allows you to access others repositories
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: GitHub can store a remote copy of your repository.

6. Which of the following command allows you to update the repository?

- a) push
- b) pop
- c) update
- d) none of the mentioned

[View Answer](#)

Answer: a

Explanation: The git branch command is your general-purpose branch administration tool.

7. Which of the following is the correct way of creating GitHub repository in to well labelled commits?

- a) Fork another user's repository
- b) Pop another user's repository
- c) Zip another user's repository
- d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: A fork is a copy of a repository.

8. Which of the following command is used to squash the commits?

- a) rebase
- b) squash
- c) boot
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: In Git, there are two main ways to integrate changes from one branch into another: the merge and the rebase.

9. Which of the following statement would create branch named as 'sanfoundry'?

- a) git checkout -b sanfoundry
- b) git checkout -c sanfoundry
- c) git check -b sanfoundry
- d) none of the mentioned

[View Answer](#)

Answer: a

Explanation: A branch in Git is simply a lightweight movable pointer to one of these commits.

10. branch command is used to determine which branch you are currently in.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: -r flag should be used for copying the content.

1. Which of the following principle characteristic is odd man out in the below figure?

- a) Principle 1
- b) Principle 2
- c) Principle 3
- d) Principle 4

[View Answer](#)

Answer: c

Explanation: Multivariate Data is the only characteristic related to Principle 3.

2. Point out the correct statement.

- a) Descriptive analysis is first kind of data analysis performed
- b) Descriptions can be generalized without statistical modelling
- c) Description and Interpretation are same in descriptive analysis
- d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: Descriptive analysis describe a set of data.

3. Which of the following allows you to find the relationship you didn't about?

- a) Inferential
- b) Exploratory
- c) Causal
- d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: In statistics, exploratory data analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.

4. Which of the following command help us to give message description?

- a) git command -m
- b) git command -d
- c) git command -message
- d) none of the mentioned

[View Answer](#)

Answer: a

Explanation: This only updates your local repository.

5. Point out the wrong statement.

- a) Exploratory analyses are usually the final way
- b) Exploratory models are useful for discovering new connection
- c) Exploratory analysis alone should not be used for predicting
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: Exploratory analyses are usually not the final way.

6. Which of the following uses data on some object to predict values for other object?

- a) Inferential
- b) Exploratory
- c) Predictive

- d) None of the mentioned
[View Answer](#)

Answer: c

Explanation: A prediction is a forecast, but not only about the weather.

7. Which of the following is the common goal of statistical modelling?
a) **Inference**
b) Summarizing
c) Subsetting
d) None of the mentioned
[View Answer](#)

Answer: a

Explanation: Inference is the act or process of deriving logical conclusions from premises known or assumed to be true.

8. Which of the following model is usually a gold standard for data analysis?
a) Inferential
b) Descriptive
c) Causal
d) All of the mentioned
[View Answer](#)

Answer: c

Explanation: A causal model is an abstract model that describes the causal mechanisms of a system.

9. Which of the following analysis should come in place of question mark in the below figure?

- a) Inferential
b) Exploratory
c) Causal
d) None of the mentioned
[View Answer](#)

Answer: a

Explanation: Inferential statistics is concerned with making predictions or inferences about a population from observations and analyses of a sample.

10. Causal analysis is commonly applied to census data.
a) True
b) False
[View Answer](#)

Answer: b

Explanation: Descriptive analysis is commonly applied to census data.

1. Which of the following type of data science question is missing in the figure?

- a) Correlative
b) Exploratory

- c) Relative
d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: Exploratory analysis is used to find relationships about you didn't know about.

2. Point out the correct statement.

- a) Descriptive analysis can be more useful for defining future studies
 b) Correlation does imply causation
c) Inference is commonly the goal of statistical model
d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: Inference depends heavily on the sampling scheme.

3. Which of the following uses relatively small amount of data to estimate about bigger population?

- a) Inferential
b) Exploratory
c) Causal
d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: Inferential statistics is concerned with making predictions or inferences about a population from observations and analyses of a sample.

4. Which of the following analysis helps out to find the effect of variable change?

- a) Inferential
b) Exploratory
 c) Causal
d) None of the mentioned

[View Answer](#)

Answer: c

Explanation: Causal Analysis provides the real reason why things happen and hence allows focused change activity.

5. Point out the correct statement.

- a) Exploratory analyses are not usually the final way
b) Inferential models are useful for discovering new connection
 c) Inference involves estimating uncertainty
d) All of the mentioned

[View Answer](#)

Answer: c

Explanation: Statistical inference is the process of deducing properties of an underlying distribution by analysis of data.

6. Which of the following relationship are usually identified as average effects?

- a) Descriptive
 b) Causal
c) Predictive
d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: A correlation is a measure or degree of relationship between two variables.

7. Which of the following is more applicable to the below figure?

- a) Descriptive
- b) Causal
- c) Predictive
- d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: Google trends helps to describe the set of data.

8. Which of the following analysis is usually modeled by deterministic set of equations?

- a) Predictive
- b) Causal
- c) Mechanistic
- d) All of the mentioned

[View Answer](#)

Answer: c

Explanation: Equations are based on physical/engineering science.

9. Which of the following analysis are incredibly hard to infer?

- a) Inferential
- b) Exploratory
- c) Causal
- d) Mechanistic

[View Answer](#)

Answer: d

Explanation: Mechanistic analysis are hard to infer except for simple simulations.

10. Accurate prediction depends heavily on measuring the right variables.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: Prediction is very hard, especially for future references.

1. Which of the following term is appropriate to the below figure?

- a) Large Data
- b) Big Data
- c) Dark Data
- d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate.

2. Point out the correct statement.

- a) Machine learning focuses on prediction, based on known properties learned from the training data
- b) Data Cleaning focuses on prediction, based on known properties learned from the training data
- c) Representing data in a form which both mere mortals can understand and get valuable insights is as much a science as it is art
- d) None of the mentioned

[View Answer](#)

Answer: d

Explanation: Visualization is becoming a very important aspect.

3. Which of the following characteristic of big data is relatively more concerned to data science?

- a) Velocity
- b) Variety
- c) Volume
- d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: Big data enables organizations to store, manage, and manipulate vast amounts of disparate data at the right speed and at the right time.

4. Which of the following analytical capabilities are provided by information management company?

- a) Stream Computing
- b) Content Management
- c) Information Integration
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: With stream computing, store less, analyze more and make better decisions faster.

5. Point out the wrong statement.

- a) The big volume indeed represents Big Data
- b) The data growth and social media explosion have changed how we look at the data
- c) Big Data is just about lots of data
- d) All of the mentioned

[View Answer](#)

Answer: c

Explanation: Big Data is actually a concept providing an opportunity to find new insight into your existing data as well guidelines to capture and analysis your future data.

6. Which of the following step is performed by data scientist after acquiring the data?

- a) Data Cleansing
- b) Data Integration
- c) Data Replication
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: Data cleansing, data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.

7. 3V's are not sufficient to describe big data.

a) True

b) False

[View Answer](#)

Answer: a

Explanation: IBM data scientists break big data into four dimensions: volume, variety, velocity and veracity.

8. Which of the following focuses on the discovery of (previously) unknown properties on the data?

- a) Data mining
- b) Big Data
- c) Data wrangling

- d) Machine Learning
[View Answer](#)

Answer: a

Explanation: Data munging or data wrangling is loosely the process of manually converting or mapping data from one “raw” form into another format that allows for more convenient consumption of the data with the help of semi-automated tools.

9. Which of the following language should be replaced with the question mark in the below figure?

- a) Java
b) PHP
c) COBOL
d) None of the mentioned
[View Answer](#)

Answer: a

Explanation: Java is used for processing data in Big data Analytics.

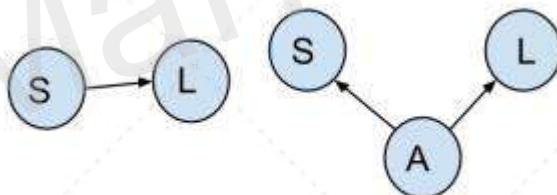
10. Beyond Volume, variety and velocity are the issues of big data veracity.

- a) True
b) False
[View Answer](#)

Answer: a

Explanation: Data Veracity is uncertain or imprecise data.

1. Which of the following design term is perfectly applicable to the below figure?



- a) Correlation
 b) Confounding
c) Causation
d) None of the mentioned
[View Answer](#)

Answer: b

Explanation: Confounding can be dealt with either at the study design stage, or at the analysis stage.

2. Point out the correct statement.

- a) If equations are known but the parameters are not, they may be inferred with data analysis
b) If equations are not known but the parameters are, they may be inferred with data analysis
c) If equations and parameter are not, they may be inferred with data analysis
d) None of the mentioned
[View Answer](#)

Answer: a

Explanation: Usually the random component of data is measurement error.

3. Which of the following is the top most important thing in data science?

- a) answer
- b) question
- c) data
- d) none of the mentioned

[View Answer](#)

Answer: b

Explanation: The second most important is the data.

4. Which of the following approach should be used if you can't fix the variable?

- a) randomize it
- b) non stratify it
- c) generalize it
- d) none of the mentioned

[View Answer](#)

Answer: a

Explanation: If you can't fix the variable, stratify it.

5. Point out the wrong statement.

- a) Randomized studies are not used to identify causation
- b) Complication approached exist for inferring causation
- c) Causal relationships may not apply to every individual
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: Randomized studies are usually used to identify causation.

6. Which of the following is a good way of performing experiments in data science?

- a) Measure variability
- b) Generalize to the problem
- c) Have Replication
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: Experiments on causal relationships investigate the effect of one or more variables on one or more outcome variables.

7. Which of the following is commonly referred to as 'data fishing'?

- a) Data bagging
- b) Data booting
- c) Data merging
- d) None of the mentioned

[View Answer](#)

Answer: d

Explanation: Data dredging is sometimes referred to as "data fishing".

8. Which of the following data mining technique is used to uncover patterns in data?

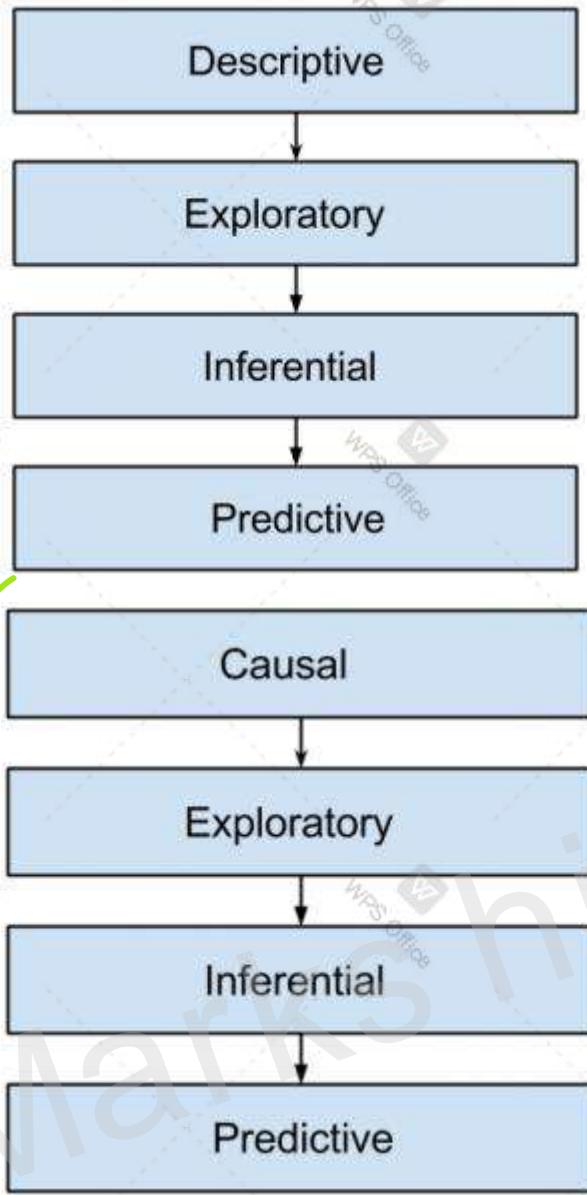
- a) Data bagging
- b) Data booting
- c) Data merging
- d) Data Dredging

[View Answer](#)

Answer: d

Explanation: Data dredging, also called as data snooping, refers to the practice of misusing data mining techniques to show misleading scientific 'research'.

9. Which of the following figure correctly shows approximate order of difficulty?



c)

- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: Predictive analysis is the practice of extracting information from existing data sets.

10. If X predicts Y, it does mean X causes Y.

- a) True

- b) False

[View Answer](#)

Answer: b

Explanation: If X predicts Y, it does not mean X causes Y.

1. Which of the following operations are supported on Time Frames?

- a) idxmax

- b) ixmax

- c) ixmin

- d) none of the mentioned

[View Answer](#)

Answer: a

Explanation: Operands can also appear in a reversed order.

2. Point out the correct statement.

- a) Timedeltas are differences in times, expressed in difference units

- b) You can construct a Timedelta scalar through various argument

- c) DateOffsets cannot be used in construction

- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: Timedeltas can be both positive and negative.

3. Numeric reduction operation for timedelta64[ns] will return _____ objects.

- a) Timeseries

- b) Timeplus

- c) Timedelta

- d) None of the mentioned

[View Answer](#)

Answer: c

Explanation: NaT are skipped during evaluation.

4. Which of the following scalars can be converted to other ‘frequencies’ by as typing to a specific timedelta type?

- a) Timedelta Series

- b) TimedeltaIndex

- c) Timedelta

- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: These operations yield Series and propagate NaT -> nan.

5. Point out the wrong statement.

- a) min, max, idxmin, idxmax operations are supported on Series

- b) You cannot pass a timedelta to get a particular value

- c) Division by the numpy scalar is true division

- d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: Dividing or multiplying a timedelta64[ns] Series by an integer or integer Series yields another timedelta64[ns] dtypes Series.

6. Which of the following is used to generate an index with time delta?

- a) TimeIndex
- b) TimedeltaIndex
- c) LeadIndex
- d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: Using TimedeltaIndex you can pass string-like, Timedelta, timedelta, or np.timedelta64 objects.

7. Combination of TimedeltaIndex with DatetimeIndex allow certain combination operations that are NaT preserving.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: You can also convert indices to yield another index.

8. Using _____ on categorical data will produce similar output to a Series or DataFrame of type string.

- a) .desc()
- b) .describe()
- c) .rank()
- d) none of the mentioned

[View Answer](#)

Answer: b

Explanation: Categorical data has a categories and a ordered property.

9. Which of the following method can be used to rename categorical data?

- a) Categorical.rename_categories()
- b) Categorical.rename()
- c) Categorical.mv_categories()
- d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: Renaming categories is done by assigning new values to the Series.cat.categories property.

10. All values of categorical data are either in categories or np.nan.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: Categoricals are pandas data type.

1. The plot method on Series and DataFrame is just a simple wrapper around _____

- a) gplt.plot()
- b) plt.plot()
- c) plt.plotgraph()
- d) none of the mentioned

[View Answer](#)

Answer: b

Explanation: If the index consists of dates, it calls gcf().autofmt_xdate() to try to format the x-axis nicely.

2. Point out the correct combination with regards to kind keyword for graph plotting.

- a) 'hist' for histogram
- b) 'box' for boxplot
- c) 'area' for area plots
- d) all of the mentioned

[View Answer](#)

Answer: d

Explanation: The kind keyword argument of plot() accepts a handful of values for plots other than the default Line plot.

3. Which of the following value is provided by kind keyword for barplot?

- a) bar
- b) kde
- c) hexbin
- d) none of the mentioned

[View Answer](#)

Answer: a

Explanation: bar can also be used for barplot.

4. You can create a scatter plot matrix using the _____ method in pandas.tools.plotting.

- a) sca_matrix
- b) scatter_matrix
- c) DataFrame.plot
- d) all of the mentioned

[View Answer](#)

Answer: b

Explanation: You can create density plots using the Series/DataFrame.plot.

5. Point out the wrong combination with regards to kind keyword for graph plotting.

- a) 'scatter' for scatter plots
- b) 'kde' for hexagonal bin plots
- c) 'pie' for pie plots
- d) none of the mentioned

[View Answer](#)

Answer: b

Explanation: kde is used for density plots.

6. Which of the following plots are used to check if a data set or time series is random?

- a) Lag
- b) Random
- c) Lead
- d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: Random data should not exhibit any structure in the lag plot.

7. Plots may also be adorned with error bars or tables.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: There are several plotting functions in pandas.tools.plotting.

8. Which of the following plots are often used for checking randomness in time series?

- a) Autocausation
- b) Autorank
- c) Autocorrelation
- d) None of the mentioned

[View Answer](#)

Answer: c

Explanation: If the time series is random, such autocorrelations should be near zero for any and all time-lag separations.

9. _____ plots are used to visually assess the uncertainty of a statistic.

- a) Lag

- b) RadViz
 c) Bootstrap
d) None of the mentioned

[View Answer](#)

Answer: c

Explanation: Resulting plots and histograms are what constitutes the bootstrap plot.

10. Andrews curves allow one to plot multivariate data.

- a) True
b) False

[View Answer](#)

Answer: a

Explanation: Curves belonging to samples of the same class will usually be closer together and form larger structures.

1. Which of the following is used to compute the percent change over a given number of periods?

- a) pct_change
b) percent_change
c) per_change
d) none of the mentioned

[View Answer](#)

Answer: a

Explanation: Series, DataFrame, and Panel all have a method pct_change.

2. Point out the correct statement.

- a) Pandas represents timestamps in microsecond resolution
b) Pandas is 100% thread safe
 c) For Series and DataFrame objects, var normalizes by N-1 to produce unbiased estimates
d) All of the mentioned

[View Answer](#)

Answer: c

Explanation: Pandas represents timestamps in nanosecond resolution.

3. Which of the following object has a method cov to compute covariance between series?

- a) Series
b) DataFrame
c) Panel
d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: DataFrame has a method cov to compute pairwise covariances among the series in the DataFrame, also excluding NA/null values.

4. Which of the following specifies the required minimum number of observations for each column pair in order to have a valid result?

- a) min_periods
b) max_periods
c) minimum_periods
d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: DataFrame.cov also supports an optional min_periods.

5. Point out the wrong statement.

- a) lxml is very fast
b) lxml requires Cython to install correctly
 c) lxml does not make any guarantees about the results of its parse
d) none of the mentioned

[View Answer](#)

Answer: c

Explanation: There are some versioning issues surrounding the libraries that are used to parse HTML tables in the top-level pandas io function read_html.

6. Which of the following is implemented on DataFrame to compute the correlation between like-labeled Series contained in different DataFrame objects?

- a) corrwith
- b) corwith
- c) corwt
- d) none of the mentioned

[View Answer](#)

Answer: a

Explanation: A score close to 1 means their tastes are very similar.

7. rolling_count function gives the number of non-null observations.

- a) True

- b) False

[View Answer](#)

Answer: b

Explanation: The binary operators take two Series or DataFrames.

8. Which of the following method produces a data ranking with ties being assigned the mean of the ranks for the group?

- a) rank
- b) dense_rank
- c) partition_rank
- d) none of the mentioned

[View Answer](#)

Answer: a

Explanation: rank is also a DataFrame method.

9. Which of the following can potentially change the dtype of a series?

- a) reindex_like
- b) index_like
- c) itime_like
- d) none of the mentioned

[View Answer](#)

Answer: a

Explanation: reindex_like silently inserts NaNs and the dtype changes accordingly.

10. cov and corr supports the optional min_periods keyword.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: Non-numeric columns will be automatically excluded from the correlation calculation.

1. Which of the following thing can be data in Pandas?

- a) a python dict
- b) an ndarray
- c) a scalar value
- d) all of the mentioned

[View Answer](#)

Answer: d

Explanation: The passed index is a list of axis labels.

2. Point out the correct statement.

- a) If data is a list, if index is passed the values in data corresponding to the labels in the index will be pulled out

- b) NaN is the standard missing data marker used in pandas
- c) Series acts very similarly to a array
- d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: If data is a dict, if index is passed the values in data corresponding to the labels in the index will be pulled out.

3. The result of an operation between unaligned Series will have the _____ of the indexes involved.

- a) intersection
- b) union
- c) total
- d) all of the mentioned

[View Answer](#)

Answer: b

Explanation: If a label is not found in one Series or the other, the result will be marked as missing NaN.

4. Which of the following input can be accepted by DataFrame?

- a) Structured ndarray
- b) Series
- c) DataFrame
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: DataFrame is a 2-dimensional labeled data structure with columns of potentially different types.

5. Point out the wrong statement.

- a) A DataFrame is like a fixed-size dict in that you can get and set values by index label
- b) Series can be passed into most NumPy methods expecting an ndarray
- c) A key difference between Series and ndarray is that operations between Series automatically align the data based on label
- d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: A Series is like a fixed-size dict in that you can get and set values by index label.

6. Which of the following takes a dict of dicts or a dict of array-like sequences and returns a DataFrame?

- a) DataFrame.from_items
- b) DataFrame.from_records
- c) DataFrame.from_dict
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: DataFrame.from_dict operates like the DataFrame constructor except for the orient parameter which is 'columns' by default.

7. Series is a one-dimensional labeled array capable of holding any data type.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: The axis labels are collectively referred to as the index.

8. Which of the following works analogously to the form of the dict constructor?

- a) DataFrame.from_items
- b) DataFrame.from_records
- c) DataFrame.from_dict
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: DataFrame.from_records takes a list of tuples or an ndarray with structured dtype.

9. Which of the following operation works with the same syntax as the analogous dict operations?

- a) Getting columns
- b) Setting columns
- c) Deleting columns
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: You can treat a DataFrame semantically like a dict of like-indexed Series objects.

10. If data is an ndarray, index must be the same length as data.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: If no index is passed, one will be created having values [0, ..., len(data) - 1].

1. All pandas data structures are ____ mutable but not always _____ mutable.

- a) size, value
- b) semantic, size
- c) value, size
- d) none of the mentioned

[View Answer](#)

Answer: c

Explanation: The length of a Series cannot be changed.

2. Point out the correct statement.

- a) Pandas consist of set of labeled array data structures
- b) Pandas consist of an integrated group by engine for aggregating and transforming data sets
- c) Pandas consist of moving window statistics
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: Some elements may be close to one another according to one distance and farther away according to another.

3. Which of the following statement will import pandas?

- a) import pandas as pd
- b) import panda as py
- c) import pandaspy as pd
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: You can read data from a CSV file using the read_csv function.

4. Which of the following object you get after reading CSV file?

- a) DataFrame
- b) Character Vector
- c) Panel
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: You get columns out of a DataFrame the same way you get elements out of a dictionary.

5. Point out the wrong statement.

- a) Series is 1D labeled homogeneously-typed array

- b) DataFrame is general 2D labeled, size-mutable tabular structure with potentially heterogeneously-typed columns
c) Panel is generally 2D labeled, also size-mutable array
d) None of the mentioned

[View Answer](#)

Answer: c

Explanation: Panel is generally 3D labeled.

6. Which of the following library is similar to Pandas?

- a) NumPy
b) RPy
c) OutPy
d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: NumPy is the fundamental package for scientific computing with Python.

7. Panel is a container for Series, and DataFrame is a container for DataFrame objects.

- a) True
b) False

[View Answer](#)

Answer: b

Explanation: DataFrame is a container for Series, and panel is a container for DataFrame objects.

8. Which of the following is prominent python “statistics and econometrics library”?

- a) Bokeh
b) Seaborn
c) Statsmodels
d) None of the mentioned

[View Answer](#)

Answer: c

Explanation: Bokeh is a Python interactive visualization library for large datasets that natively uses the latest web technologies.

9. Which of the following is a foundational exploratory visualization package for the R language in pandas ecosystem?

- a) yhat
b) Seaborn
c) Vincent
d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: It has great support for pandas data objects.

10. Pandas consist of static and moving window linear and panel regression.

- a) True
b) False

[View Answer](#)

Answer: a

Explanation: Time series and cross-sectional data are special cases of panel data.

1. Quandl API for Python wraps the _____ REST API to return Pandas DataFrames with time series indexes.

- a) Quandl
b) PyDatastream
c) PyData
d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: PyDatastream is a Python interface to the Thomson Dataworks Enterprise (DWE/Datastream) SOAP API to return indexed pandas DataFrames or panels with financial data.

2. Point out the correct statement.

- a) Statsmodels provides powerful statistics, econometrics, analysis and modeling functionality that is out of pandas' scope
- b) Vintage leverages pandas objects as the underlying data container for computation
- c) Bokeh is a Python interactive visualization library for small datasets
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: Bokeh goal is to provide elegant, concise construction of novel graphics in the style of D3.

3. Which of the following library is used to retrieve and acquire statistical data and metadata disseminated in SDMX 2.1?

- a) pandaSDMX
- b) freedapi
- c) geopandas
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: Geopandas extends pandas data objects to include geographic information which supports geometric operations.

4. Which of the following provides a standard API for doing computations with MongoDB?

- a) Blaze
- b) Geopandas
- c) FRED
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: If your work entails maps and geographical coordinates, and you love pandas, you should take a close look at Geopandas.

5. Point out the wrong statement.

- a) qgrid is an interactive grid for sorting and filtering DataFrames
- b) Pandas DataFrames implement `_repr_html_` methods which are utilized by IPython Notebook
- c) Spyder is a cross-platform Qt-based open-source R IDE
- d) None of the mentioned

[View Answer](#)

Answer: c

Explanation: Spyder is a cross-platform Qt-based open-source Python IDE.

6. Which of the following makes use of pandas and returns data in a series or DataFrame?

- a) pandaSDMX
- b) freedapi
- c) OutPy
- d) none of the mentioned

[View Answer](#)

Answer: b

Explanation: freedapi module requires a FRED API key that you can obtain for free on the FRED website.

7. Spyder can introspect and display Pandas DataFrames.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: Spyder shows both "column wise min/max and global min/max coloring."

8. Which of the following is used for machine learning in python?

- a) scikit-learn
- b) seaborn-learn
- c) stats-learn
- d) none of the mentioned

[View Answer](#)

Answer: a
Explanation: scikit-learn is built on NumPy, SciPy, and matplotlib.

9. The _____ project builds on top of pandas and matplotlib to provide easy plotting of data.
a) yhat
b) Seaborn
c) Vincent
d) None of the mentioned

[View Answer](#)

Answer: b
Explanation: Seaborn has great support for pandas data objects.

10. x-ray brings the labeled data power of pandas to the physical sciences.
a) True
b) False

[View Answer](#)

Answer: a
Explanation: It aims to provide a pandas-like and pandas-compatible toolkit for analytics on multi-dimensional arrays

1. Which of the following is the base layer for all of the sparse indexed data structures?
a) SArray
b) SparseArray
c) PyArray
d) None of the mentioned

[View Answer](#)

Answer: b
Explanation: SparseArray is a 1-dimensional ndarray-like object storing only values distinct from the fill_value.

2. Point out the correct statement.
a) All of the standard pandas data structures have a to_sparse method
b) Any sparse object can be converted back to the standard dense form by calling to_dense
c) The sparse objects exist for memory efficiency reasons
d) All of the mentioned

[View Answer](#)

Answer: d
Explanation: The to_sparse method takes a kind argument and a fill_value.

3. Which of the following is not an indexed object?
a) SparseSeries
b) SparseDataFrame
c) SparsePanel
d) None of the mentioned

[View Answer](#)

Answer: d
Explanation: SparseArray can be converted back to a regular ndarray by calling to_dense.

4. Which of the following list-like data structure is used for managing a dynamic collection of SparseArrays?
a) SparseList
b) GeoList
c) SparseSeries
d) All of the mentioned

[View Answer](#)

Answer: a
Explanation: To create one, simply call the SparseList constructor with a fill_value.

5. Point out the wrong statement.
a) to_array.append can accept scalar values or any 2-dimensional sequence

- b) Two kinds of SparseIndex are implemented
c) The integer format keeps an arrays of all of the locations where the data are not equal to the fill value
d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: to_array. append can accept scalar values or any 1-dimensional sequence.

6. Which of the following method is used for transforming a SparseSeries indexed by a MultiIndex to a scipy.sparse.coo_matrix?
a) SparseSeries.to_coo()
b) Series.to_coo()
c) SparseSeries.to_cooer()
d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: Experimental api to transform between sparse pandas and scipy.sparse structures.

7. The integer format tracks only the locations and sizes of blocks of data.
a) True
b) False

[View Answer](#)

Answer: b

Explanation: The block format tracks only the locations and sizes of blocks of data.

8. Which of the following is used for testing for membership in the list of column names?
a) in
b) out
c) elseif
d) none of the mentioned

[View Answer](#)

Answer: a

Explanation: For DataFrames, likewise, in applies to the column axis.

9. Which of the following indexing capabilities is used as a concise means of selecting data from a pandas object?
a) In
b) ix
c) ipy
d) none of the mentioned

[View Answer](#)

Answer: b

Explanation: ix and reindex are 100% equivalent.

10. Pandas follow the NumPy convention of raising an error when you try to convert something to a bool.
a) True
b) False

[View Answer](#)

Answer: a

Explanation: This happens in an if or when using the boolean operations, and, or, or not.

1. Which of the following block information is odd man out?

- a) Subsetting
- b) Raw data
- c) Ready for analysis
- d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: Characteristics mentioned in the diagram are traits of processed data.

2. Point out the correct statement.

- a) Data has only qualitative value
- b) Data has only quantitative value
- c) Data has both qualitative and quantitative value
- d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: Data belongs to the set of items.

3. Data that summarize all observations in a category are called _____ data.

- a) frequency
- b) summarized
- c) raw
- d) none of the mentioned

[View Answer](#)

Answer: b

Explanation: The summary could be the sum of the observations, the number of occurrences, their mean value, and so on.

4. Which of the following is an example of raw data?

- a) original swath files generated from a sonar system
- b) initial time-series file of temperature values
- c) a real-time GPS-encoded navigation file
- d) all of the mentioned

[View Answer](#)

Answer: d

Explanation: Raw data refers to data that have not been changed since acquisition.

5. Point out the correct statement.

- a) Primary data is original source of data
- b) Secondary data is original source of data
- c) Questions are obtained after data processing steps
- d) None of the Mentioned

[View Answer](#)

Answer: a

Explanation: Primary data is also referred to as raw data.

6. Which of the following data is put into a formula to produce commonly accepted results?

- a) Raw
- b) Processed
- c) Synchronized

d) All of the Mentioned

[View Answer](#)

Answer: b

Explanation: Raw data came from direct measurements.

7. Processing data includes subsetting, formatting and merging only.

a) True

b) False

[View Answer](#)

Answer: b

Explanation: There are many other techniques applied to raw data.

8. Which of the following is another name for raw data?

a) destination data

b) eggy data

c) secondary

d) machine learning

[View Answer](#)

Answer: b

Explanation: Although raw data has the potential to become “information,” extraction, organization, and sometimes analysis and formatting for presentation are required for that to occur.

9. Which type of data is generated by POS terminal in a busy supermarket each day?

a) Source

b) Processed

c) Synchronized

d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: Raw data is sometimes referred to as source data.

10. Following figure represents correct sequence of steps in performing data analysis.

a) True

b) False

[View Answer](#)

Answer: a

Explanation: Data analysis is not a goal in itself; the goal is to enable the business to make better decisions.

1. Which of the following is an example of tidy data?

- a) complicated JSON from facebook API
- b) complicated JSON from Twitter API
- c) unformatted excel file
- d) all of the mentioned

[View Answer](#)

Answer: d

Explanation: Tidy data is obtained after processing script.

2. Point out the correct statement.

- a) Nearly 80% of data analysis is spent on wrangling data
- b) Nearly 20% of data analysis is spent on data dredging
- c) Nearly 80% of data analysis is spent on the cleaning and preparing data
- d) None of the mentioned

[View Answer](#)

Answer: c

Explanation: Data cleansing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.

3. Which of the following is a trait of tidy data?

- a) each variable in one column
- b) each observation in different row
- c) one table for each kind of variable
- d) none of the mentioned

[View Answer](#)

Answer: b

Explanation: The summary could be the sum of the observations, the number of occurrences, their mean value, and so on.

4. Which of the following package is used for tidy data?

- a) tidyR
- b) souryr
- c) NumPy
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: tidyR is used for tidy data with spread and gather functions.

5. Point out the wrong statement.

- a) Tidy datasets are all alike but every messy dataset is messy in its own way
- b) Most statistical datasets are data frames made up of rows and columns
- c) Tidy datasets provide a standardized way to link the structure of a dataset with its semantics
- d) None of the mentioned

[View Answer](#)

Answer: d

Explanation: The tidy data standard has been designed to simplify the development of data analysis tools that work well together.

6. Which of the following process involves structuring datasets to facilitate analysis?

- a) Data tidying
- b) Data mining
- c) Data booting
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: The principles of tidy data provide a standard way to organize data values within a dataset.

7. Strange binary file generated from machines is an example of tidy data.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: Data sets stored in spreadsheets, such as Microsoft's Excel, are binary, not raw ASCII data files.

8. Which of the following is the most common problem with messy data?

- a) Column headers are values
- b) Variables are stored in both rows and columns
- c) A single observational unit is stored in multiple tables
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: Real datasets can, and often do, violate the three precepts of tidy data in almost every way imaginable.

9. tidyr is a reframing of _____ designed to accompany the tidy data framework.

- a) reshape5
- b) dplyr
- c) reshape2
- d) all of the mentioned

[View Answer](#)

Answer: c

Explanation: tidyr does less reframing than reshape2.

10. Raw data in the real-world is tidy and properly formatted.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: Data analysis is not a goal in itself; the goal is to enable the business to make better decisions.

1. Which of the following function is used for loading flat files?

- a) read.data
- b) read.sheet
- c) read.table
- d) none of the mentioned

[View Answer](#)

Answer: c

Explanation: This reads data in to the RAM.

2. Point out the correct statement.

- a) XLConnect package has more options for manipulating access files
- b) XLConnect vignette package can also be used for manipulating excel files
- c) write.xlsx write out an excel file with different argument
- d) None of the mentioned

[View Answer](#)

Answer: c

Explanation: write.xlsx write out an excel file with similar argument.

3. Which of the following is an important parameter of read.table function?

- a) file
- b) header
- c) sep
- d) all of the mentioned

[View Answer](#)

Answer: d

Explanation: More parameters are required for loading the data.

4. Which of the following will set the character that represents missing value?

- a) na.quote

- b) na.strings
- c) nrows
- d) all of the mentioned

[View Answer](#)

Answer: b

Explanation: na.strings takes a character vector.

5. Point out the wrong statement.

- a) data.table inherits from data.frame
- b) data.table is written in Java
- c) data.table is faster at subsetting and updating data
- d) none of the mentioned

[View Answer](#)

Answer: b

Explanation: data.table is written in C.

6. Which of the following package is used for reading excel data?

- a) xlsx
- b) xlsc
- c) read.sheet
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: read.xlsx and read.xls functions are part of xlsx package.

7. Which of the following can be used to view all the tables in memory?

- a) tables
- b) alitable
- c) table
- d) none of the mentioned

[View Answer](#)

Answer: a

Explanation: The table function is a very basic, but essential, function to master while performing interactive data analyses.

8. Which of the following function programmatically extract parts of XML file?

- a) XmlSApply
- b) XmlApply
- c) XmlSApplyData
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: xmlSApply are simple wrappers for tapply and lapply functions.

9. Which of the following package is used for reading JSON data?

- a) jsonlite
- b) json
- c) jsondata
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: The jsonlite package is a JSON generator optimized for the web.

10. Extracting XML is the basis for most web scraping.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: XML is particularly used in web applications.

1. Which of the following package is used to connect MySQL RDBMS with R?

- a) RMySQL vignette
- b) MySQL vignette
- c) RSQL vignette
- d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: This package contains meta information and index.

2. Point out the correct statement.

- a) HDF5 is a hierarchical format
- b) HDF5 does not support range of different data types
- c) HDF5 is used for storing small datasets
- d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: HDF5 is used for storing large datasets.

3. Which of the following is used to extract data from HTML code of websites?

- a) Webscraping
- b) Webdredging
- c) Webcleaning
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: Webscraping is a great way to get data.

4. Which of the following function is used to read data off the webpages?

- a) read.web
- b) read.Lines
- c) read.Line
- d) all of the mentioned

[View Answer](#)

Answer: b

Explanation: read.Lines function will extract the web page data.

5. Point out the wrong statement.

- a) hdf5 can be used to reading/writing from disc in Python
- b) rhdf5 is an interface for hdf5 format
- c) maximum size of an HDF5 dataset is fixed when it is created
- d) all of the mentioned

[View Answer](#)

Answer: b

Explanation: hdf5 can be used to reading/writing from disc in R.

6. Which of the following package is used for reading HTML and XML data?

- a) httr
- b) http
- c) httx
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: httr contains tools for Working with URLs and HTTP.

7. httr package does not work well with facebook and twitter API.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: Most modern APIs use something like oauth.

8. Which of the following request can be issued from httr package?

- a) GET
- b) PUT
- c) DELETE
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: Authentication is necessary for issuing a request.

9. Which of the following package loads data from SPSS?

- a) read.spss(SPSS)
- b) read.oct(SPSS)
- c) read.xpot(SPSS)
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: SPSS is a comprehensive and flexible statistical analysis and data management solution.

10. Which of the following package is used for reading GIS data?

- a) rgdal
- b) rgeos
- c) raster
- d) all of the mentioned

[View Answer](#)

Answer: d

Explanation: A geographic information system is a system designed to capture, store, manipulate, analyze, manage, and present all types of spatial or geographical data.

1. Which of the following function gives information about top level data?

- a) head
- b) tail
- c) summary
- d) none of the mentioned

[View Answer](#)

Answer: a

Explanation: The function head is very useful for working with lists, tables, data frames and even functions.

2. Point out the correct statement.

- a) head function work on string
- b) tail function work on string
- c) head function work on string but tail function do not
- d) none of the mentioned

[View Answer](#)

Answer: d

Explanation: Both head and tail function do not work on strings.

3. Which of the following function is used for quantiles of quantitative values?

- a) quantile
- b) quantity
- c) quantiles
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: In probability and statistics, the quantile function specifies, for a given probability in the probability distribution of a random variable, the value at which the probability of the random variable will be less than or equal to that probability.

4. Which of the following function is used for determining missing values?

- a) any
- b) all
- c) is
- d) all of the mentioned

[View Answer](#)

Answer: d

Explanation: In R, missing values are represented by the symbol NA.

5. Point out the wrong statement.

- a) Common variables are used to create missingness vector
- b) Common variables are used to cutting up quantitative variables
- c) Common variables are not used to apply transforms
- d) All of the mentioned

[View Answer](#)

Answer: c

Explanation: Common variables are not used to apply transforms.

6. Which of the following transforms can be performed with data value?

- a) log2
- b) cos
- c) log10
- d) all of the mentioned

[View Answer](#)

Answer: d

Explanation: Many common transforms can be applied to the data with R.

7. Each observation forms a column in tidy data.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: Each variable forms a column in tidy data.

8. Which of the following function is used for casting data frames?

- a) decast
- b) ucast
- c) rcast
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: Use acast or decast depending on whether you want vector/matrix/array output or data frame output.

9. Which of the following join is by default used in plyr package?

- a) left
- b) right
- c) full
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: Join is faster in plyr package.

10. mutate function is used for casting as multi dimensional arrays.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: mutate is used for adding new variables.

1. Which of the following function is good for the automatic splitting of names?

- a) split
- b) strsplit
- c) autsplit
- d) none of the mentioned

[View Answer](#)

Answer: b

Explanation: strsplit split a character string or vector of character strings using a regular expression or a literal string.

2. Point out the correct statement.

- a) gsub is used for fixing character vectors
- b) sub is used for finding values like grep
- c) grep is used for fixing character vectors
- d) none of the mentioned

[View Answer](#)

Answer: a

Explanation: sub and gsub is used for fixing character vectors.

3. Which of the following function is used for fixing character vectors?

- a) tolower
- b) toUPPER
- c) toLOWER
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: It translates character to lowercase.

4. Which of the following metacharacter is used to refer to any character?

- a) %
- b) @
- c) .
- d) All of the mentioned

[View Answer](#)

Answer: c

Explanation: A dot in function name can mean any of the following: nothing at all; a separator between method and class in S3 method.

5. Point out the wrong statement.

- a) Variables with character values should be made less descriptive
- b) Variables with character values should usually be made into factor variable
- c) Common variables are used to apply transforms
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: Variables with character values should be made more descriptive.

6. Which of the following is used for specifying character class with metacharacter?

- a) []
- b) {}
- c) /+
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: You can list set of characters to accept a given point in the match.

7. Regular expressions can be thought of as a combination of literals and metacharacters.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: Regular expressions have rich set of metacharacters.

8. Which of the following signs are used to indicate repetition?

- a) #
- b) *
- c) -
- d) All of the mentioned

[View Answer](#)

Answer: b

Explanation: * and + are metacharacters for repetition of data.

9. Which of the following function is used for searching text strings by means of regular expression?

- a) grep
- b) grep1
- c) grepexpr
- d) all of the mentioned

[View Answer](#)

Answer: b

Explanation: grep, grep1, regexpr, gregexpr and regexec search for matches to argument pattern within each element of a character vector.

10. merge function is used for merging data frames.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: To merge two data frames horizontally, use the merge function.

1. Which of the the following graphic device information is odd man out in the below figure?

- a) quartz
- b) window
- c) unix
- d) x11

[View Answer](#)

Answer: c

Explanation: unix keyword does not exist with regards to graphics device.

2. Point out the correct statement.

- a) On Mac, the screen device is launched with quartz
- b) On Windows, the screen device is launched with wind

- c) On Unix, the screen device is launched with x12
d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: On Windows, the screen device is launched with window function.

3. Which of the following is an example of graphics device?

- a) PDF
b) SVG
c) JPEG
d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: When the plot() function is invoked, R sends the data corresponding to the plot over, and the graphics device generates the plot.

4. Which of the following file format is graphic device only for windows?

- a) pdf
b) svg
c) win.metafile
d) all of the mentioned

[View Answer](#)

Answer: c

Explanation: Exporting graphics to a Windows MetaFile can be achieved via the win.metafile.

5. Point out the wrong statement.

- a) For quick visualizations and exploratory analysis, usually you want to use the screen device
b) Functions like xyplot in lattice will not default to sending a plot to the screen device
c) Not all graphics devices are available on all platforms
d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: window function cannot be used on Mac.

6. Which of the following system most often don't have postscript viewer?

- a) Windows
b) Linux
c) Mac
d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: postscript is older format but it resizes well.

7. There are mainly three types of file devices.

- a) True
b) False

[View Answer](#)

Answer: b

Explanation: There are mainly basic types of file devices-vector and bitmap.

8. Which of the following is a bitmap file type?

- a) tiff
b) svg
c) pdf
d) none of the mentioned

[View Answer](#)

Answer: c

Explanation: TIFF is a computer file format for storing raster graphics images.

9. Which of the following function displays currently active graphics device?

- a) dev.present
- b) dev.cur
- c) pre.cur
- d) all of the mentioned

[View Answer](#)

Answer: b

Explanation: You can change the active graphics device with dev.set.

10. The most familiar place for a plot to be “sent” is screen device.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: On Linux, the screen device is launched with x11 function.

Sanfoundry Global Education & Learning Series – Data Science.

1. Which of the following function has parameters shown in the below figure?

- a) par
- b) bar
- c) base
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: R makes it easy to combine multiple plots into one overall graph, using either the par() or layout() function.

2. Point out the correct statement.

- a) Vector formats are good for line drawings and plots with solid colors using a modest number of points
- b) Vector formats are good for plots with a large number of points, natural scenes or web based plots
- c) The default graphics device is always the screen device
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: Bitmap formats are good for plots with a large number of points, natural scenes or web based plots.

3. Which of the following will copy the plot from one device to another?

- a) dev.copy
- b) dev.copypdf
- c) dev.device
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: Copying a plot to another device can be useful because some plots require a lot of code and it can be a pain to type all that in again for a different device.

4. Which of the following is used to change active graphic device?

- a) dev.set
- b) dev.int
- c) dev.win
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: You can change the active graphics device with `dev.set(<integer>)` where `<integer>` is the number associated with the graphics device you want to switch to.

5. Point out the wrong statement.

- a) File devices are useful for creating plots that can be included in other documents or sent to other people
- b) Plots must be created on a graphics device
- c) For file devices, there are vector and bitmap formats
- d) None of the mentioned

[View Answer](#)

Answer: d

Explanation: For file devices, there are vector and bitmap formats.

6. Which of the following is the second goal of PCA?

- a) data compression
- b) statistical analysis
- c) data dredging
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: The principal components are equal to the right singular values if you first scale the variables.

7. `dev.copy2pdf` specifically copy a plot to a PDF file.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: Copying a plot is not an exact operation, so the result may not be identical to the original.

8. Which of the following is a vector file device?

- a) png
- b) svg
- c) bmp
- d) none of the mentioned

[View Answer](#)

Answer: b

Explanation: svg stands for scalable vector graphics.

9. Which of the following is alternative technique to principal component analysis?

- a) Factor analysis
- b) Independent components analysis
- c) Latent semantic analysis
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: PC's may mix real patterns.

10. Every open graphics device is assigned an integer greater than 2.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: Every open graphics device is assigned an integer greater than equal to 2.

1. Which of the following block information is odd man out in the below figure?

- a) Scatterplots
- b) 5 number summary
- c) 2D Graph
- d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: 5 number summary is one dimensional graph.

2. Which type of graph is shown in the following figure?

- a) Scatterplot
- b) Barplot
- c) Overlaying
- d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: A bar plot represents an estimate of central tendency for a numeric variable with the height of each rectangle.

3. Which of the following annotation function is used to add or modify text?

- a) word
- b) graph
- c) lines
- d) all of the mentioned

[View Answer](#)

Answer: d

Explanation: points and axis are other well known annotation function.

4. Which of the following package is implemented by lattice plotting system?

- a) grDevices
- b) grid
- c) graphics
- d) all of the mentioned

[View Answer](#)

Answer: b

Explanation: Use grid on to display the major grid lines.

5. Point out the wrong statement.

- a) Plot are created with multiple functions only
- b) Plots are created with both single and multiple function calls
- c) Annotation in plot is not especially intuitive
- d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: Plots are created with single function also.

6. Which of the following parameter defines line type such as dashed and dotted?

- a) lty
- b) pch
- c) lwd
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: lwd is used for line width.

7. The core plotting engine is encapsulated in graphics package.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: graphics package contain plotting functions.

8. Which of the following argument specifies margin size with regards to par function?

- a) las
- b) bg
- c) mar
- d) all of the mentioned

[View Answer](#)

Answer: c

Explanation: par function is used to specify global parameters.

9. How many stages commonly occurs in creation of plot?

- a) 2
- b) 5
- c) 8
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: The base plotting system is highly flexible.

10. Base graphics are used most commonly for creating 2D graphics.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: Base graphics is a very powerful system for creating 2D graphics.

1. Which of the following clustering type has characteristic shown in the below figure?

- a) Partitional
- b) Hierarchical
- c) Naive bayes
- d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: Hierarchical clustering groups data over a variety of scales by creating a cluster tree or dendrogram.

2. Point out the correct statement.

- a) The choice of an appropriate metric will influence the shape of the clusters
- b) Hierarchical clustering is also called HCA
- c) In general, the merges and splits are determined in a greedy manner
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: Some elements may be close to one another according to one distance and farther away according to another.

3. Which of the following is finally produced by Hierarchical Clustering?

- a) final estimate of cluster centroids
- b) tree showing how close things are to each other
- c) assignment of each point to clusters
- d) all of the mentioned

[View Answer](#)

Answer: b

Explanation: Hierarchical clustering is an agglomerative approach.

4. Which of the following is required by K-means clustering?

- a) defined distance metric
- b) number of clusters
- c) initial guess as to cluster centroids
- d) all of the mentioned

[View Answer](#)

Answer: d

Explanation: K-means clustering follows partitioning approach.

5. Point out the wrong statement.

- a) k-means clustering is a method of vector quantization
- b) k-means clustering aims to partition n observations into k clusters
- c) k-nearest neighbor is same as k-means
- d) none of the mentioned

[View Answer](#)

Answer: c

Explanation: k-nearest neighbor has nothing to do with k-means.

6. Which of the following combination is incorrect?

- a) Continuous – euclidean distance
- b) Continuous – correlation similarity
- c) Binary – manhattan distance

- a) None of the mentioned

[View Answer](#)

Answer: d

Explanation: You should choose a distance/similarity that makes sense for your problem.

7. Hierarchical clustering should be primarily used for exploration.

- a) True

- b) False

[View Answer](#)

Answer: a

Explanation: Hierarchical clustering is deterministic.

8. Which of the following function is used for k-means clustering?

- a) k-means

- b) k-mean

- c) heatmap

- d) none of the mentioned

[View Answer](#)

Answer: a

Explanation: K-means requires a number of clusters.

9. Which of the following clustering requires merging approach?

- a) Partitional

- b) Hierarchical

- c) Naive Bayes

- d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: Hierarchical clustering requires a defined distance as well.

10. K-means is not deterministic and it also consists of number of iterations.

- a) True

- b) False

[View Answer](#)

Answer: a

Explanation: K-means clustering produces the final estimate of cluster centroids.

1. Which of the following graphs has properties in the below figure?

- a) Exploratory

- b) Inferential

- c) Causal

- d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: Making plots of the data reveals various interesting features.

2. Which of the following dimension type graph is shown in the below figure?

- a) one-dimensional
- b) two-dimensional
- c) three-dimensional
- d) none of the mentioned

[View Answer](#)

Answer: b

Explanation: A two-dimensional graph is a set of points in two-dimensional space.

3. Which of the following gave rise to need of graphs in data analysis?

- a) Data visualization
- b) Communicating results
- c) Decision making
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: A picture can tell better story than data.

4. Which of the following is characteristic of exploratory graph?

- a) Made slowly
- b) Axes are not cleaned up
- c) Color is used for personal information
- d) All of the mentioned

[View Answer](#)

Answer: c

Explanation: A large number of exploratory graphs are made.

5. Point out the correct statement.

- a) coplots are one dimensional data graph
- b) Exploratory graphs are made quickly
- c) Exploratory graphs are made relatively less in number
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: coplot is used for two dimensional representation.

6. Which of the following graph can be used for simple summarization of data?

- a) Scatterplot
- b) Overlaying
- c) Barplot
- d) All of the mentioned

[View Answer](#)

Answer: c

Explanation: A bar chart or bar graph is a chart that presents Grouped data with rectangular bars with lengths proportional to the values that they represent.

7. Color and shape are used to add dimensions to graph data.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: Graphs are commonly used by print and electronic media.

8. Which of the following information is not given by five-number summary?

- a) Mean
- b) Median
- c) Mode
- d) All of the mentioned

[View Answer](#)

Answer: c

Explanation: The mode is the value that appears most often in a set of data.

9. Which of the following is also referred to as overlayed 1D plot?

- a) lattice
- b) barplot
- c) gplot
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: lattice is an add-on package that implements Trellis graphics.

10. Spinning plots can be used for two dimensional data.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: There are many ways to create a 3D spinning plot as well.

1. Which of the following problem is solved by reproducibility?

- a) Scalability
- b) Data availability
- c) Improved data analysis
- d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: More transparency is achieved with reproducibility.

2. Point out the correct statement with respect to replication.

- a) Focuses on the validity of the data analysis
- b) Focuses on the validity of the scientific claim
- c) Arguably a minimum standard for any scientific study
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: Data replication if the same data is stored on multiple storage device.

3. Which of the following is effective way of checking validity of data analysis?

- a) Re-run the analysis
- b) Review the code
- c) Check the sensitivity
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: Reproducibility addresses the most “downstream” aspect of the research process.

4. Which of the following is similar to a pre-specified clinical trial protocol?

- a) Caching-based Data Analysis

- b) Evidence-based Data Analysis
- c) Markdown-based Data Analysis
- d) All of the mentioned

[View Answer](#)

Answer: b

Explanation: Evidence-based Data Analysis a deterministic statistical machine.

5. Point out the wrong statement with respect to reproducibility.

- a) Focuses on the validity of the data analysis
- b) The ultimate standard for strengthening scientific evidence
- c) Important when replication is impossible
- d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: Replication is particularly important in studies that can impact broad policy or regulatory decisions.

6. Which of the following can be used for data analysis model?

- a) CRAN
- b) CPAN
- c) CTAN
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: Different problems require different approaches and expertise.

7. Reproducibility determines correctness of data analysis.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: Reproducibility has nothing to do with validity of data analysis.

8. Which of the following step is not required in data analysis?

- a) Synthesize results
- b) Create reproducible code
- c) Interpret results
- d) None of the mentioned

[View Answer](#)

Answer: d

Explanation: The data set may depend on your goal.

9. Which of the following gives reviewers an important tool without dramatically increasing the burden?

- a) Quality research
- b) Replication research
- c) Reproducible research
- d) None of the mentioned

[View Answer](#)

Answer: c

Explanation: Reproducible research is important, but does not necessarily solve the critical question of whether a data analysis is trustworthy.

10. Result analysis are relatively easy to replicate or reproduce.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: Complicated analyses should not be trusted.

1. Which of the following is suitable for knitr?

- a) Reports
- b) Data preprocessing documents
- c) Technical manuals
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: knitr has short technical documents.

2. Point out the correct combination related to output statements.

- a) results: "asis"
- b) echo: true
- c) echo=false
- d) none of the mentioned

[View Answer](#)

Answer: a

Explanation: Global option relating to echo have values TRUE and FALSE.

3. Which of the following is required for not echoing the code?

- a) echo=TRUE
- b) print=TRUE
- c) echo=FALSE
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: Code has to be written to set the global options.

4. Which of the following global options are available for figures in knitr?

- a) fig.height
- b) fig.size
- c) fig.breadth
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: fig.height has numeric value.

5. Which of the following global option has value "hide"?

- a) results
- b) fig.width
- c) echo
- d) none of the mentioned

[View Answer](#)

Answer: a

Explanation: Workflow R Markdown is a format for writing reproducible, dynamic reports with R.

6. Which of the following is the correct order of conversion?

- a) .md->.Rmd->.html
- b) .Rmd->.md->.html
- c) .Rmd->.md->.xml
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: knitr converts markdown document in to html by default.

7. knitr is good for complex time-consuming computations.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: knitr is poor for complex time-consuming computations.

8. Which of the following statement is used for importing knitr library?

- a) library(knitr)
- b) import knitr
- c) lib(knitr)
- d) none of the mentioned

[View Answer](#)

Answer: a

Explanation: knitr is not good for documents that require precise formatting.

9. The document produced by knitr document has which of the following extension?

- a) .md
- b) .rmd
- c) .html
- d) none of the mentioned

[View Answer](#)

Answer: b

Explanation: knitr produces markdown document.

10. Code chunks begin with ``{r} and end with ```.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: Code chunks can have names.

1. What is the role of processing code in the research pipeline?

- a) Transforms the analytical results into figures and tables
- b) Transforms the analytic data into measured data
- c) Transforms the measured data into analytic data
- d) All of the mentioned

[View Answer](#)

Answer: c

Explanation: Data science workflow is a non-linear, iterative process.

2. Which of the following is a goal of literate statistical programming?

- a) Combine explanatory text and data analysis code in a single document
- b) Ensure that data analysis documents are always exported in JPEG format
- c) Require those data analysis summaries are always written in R
- d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: Literate Statistical Practice is a programming methodology.

3. What does it mean to weave a literate statistical program?

- a) Convert a program from S to python
- b) Convert the program into a human readable document
- c) Convert a program to decompress it
- d) All of the mentioned

[View Answer](#)

Answer: b

Explanation: Literate Statistical Programming can be done with knitr.

4. Which of the following is required to implement a literate programming system?

- a) A programming language like Perl

- b) A programming language like Java
- c) A programming language like R
- d) All of the mentioned

[View Answer](#)

Answer: c

Explanation: R is a language and environment for statistical computing and graphics.

5. What is one way in which the knitr system differs from Sweave?

- a) knitr allows for the use of markdown instead of LaTeX
- b) knitr is written in python instead of R
- c) knitr lacks features like caching of code chunks
- d) none of the mentioned

[View Answer](#)

Answer: a

Explanation: knitr is an engine for dynamic report generation with R.

6. Which of the following is useful way to put text, code, data, output all in one document?

- a) Literate statistical programming
- b) Object oriented programming
- c) Descriptive programming
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: Object-oriented programming is a programming language model organized around objects rather than “actions” and data rather than logic.

7. Some chunks have to be re-computed every time you re-knit the file.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: All chunks have to be re-computed every time you re-knit the file.

8. Which of the following tool can be used for integrating text and code in one document?

- a) knitr
- b) ggplot2
- c) NumPy
- d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: knitr is a way to write LaTeX, HTML, and Markdown with R code interlaced.

9. Which of the following should be set on chunk by chunk basis to store results of computation?

- a) cache=TRUE
- b) cache=FALSE
- c) caching=TRUE
- d) none of the mentioned

[View Answer](#)

Answer: a

Explanation: After the first run. The results are loaded from cache.

10. Dependencies are checked explicitly in caching caveats.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: Dependencies are not checked explicitly in caching caveats.

1. Original idea comes of Literate Statistical Practice from _____

- a) Don Knuth
- b) Don Cutting
- c) Douglas Cutting
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: Literate programs are tangled to produce machine readable documents.

2. Point out the correct statement.

- a) An article is stream of code and text
- b) Analysis code is divided in to code chunks only
- c) Literate programs are tangled to produce human readable documents
- d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: Analysis code is divided in to code chunks and text.

3. Which of the following is required for literate programming?

- a) documentation language
- b) mapper language
- c) reducer language
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: Programming language is also required for literate programming.

4. Which of the following is required to implement a literate programming system?

- a) A programming language like Perl
- b) A programming language like Java
- c) A programming language like R
- d) All of the mentioned

[View Answer](#)

Answer: c

Explanation: R is a language and environment for statistical computing and graphics.

5. Which of the following way is required to make work reproducible?

- a) keep track of things
- b) Save output
- c) Save data in proprietary formats
- d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: Save data in NON proprietary formats to make work reproducible.

6. Which of the following disadvantage does literate programming have?

- a) Slow processing of documents
- b) Code is not automatic
- c) No logical order
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: Code and text is in one place.

7. knitr supports only one documentation language.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: knitr supports various documentation languages.

8. Which of the following tool documentation language is supported by knitr?

- a) RMarkdown
- b) LaTeX
- c) HTML
- d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: knitr is available on CRAN.

9. Which of the following package by Yihui is built in to RStudio environment?

- a) rpy2
- b) knitr
- c) ggplot2
- d) none of the mentioned

[View Answer](#)

Answer: b

Explanation: It can be exported to pdf and html.

10. Literate program code is live-automatic “regression test” when building a document.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: Data and results are automatically updated to reflect external changes.

1. Which of the following is the probability calculus of beliefs, given that beliefs follow certain rules?

- a) Bayesian probability
- b) Frequency probability
- c) Frequency inference
- d) Bayesian inference

[View Answer](#)

Answer: a

Explanation: Data scientists tend to fall within shades of gray of these and various other schools of inference.

2. Point out the correct statement.

- a) Bayesian inference is the use of Bayesian probability representation of beliefs to perform inference
- b) NULL is the standard missing data marker used in S
- c) Frequency inference is the use of Bayesian probability representation of beliefs to perform inference
- d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: Frequency probability is the long run proportion of times an event occurs in independent, identically distributed repetitions.

3. Which of the following can be considered as random variable?

- a) The outcome from the roll of a die
- b) The outcome of flip of a coin
- c) The outcome of exam
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: The probability distribution of a discrete random variable is a list of probabilities associated with each of its possible values.

4. Which of the following random variable that take on only a countable number of possibilities?

- a) Discrete
- b) Non Discrete
- c) Continuous
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: Continuous random variable can take any value on some subset of the real line.

5. Point out the wrong statement.

- a) A random variable is a numerical outcome of an experiment
- b) There are three types of random variable
- c) Continuous random variable can take any value on the real line
- d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: There are two types of random variable-continuous and discrete.

6. Which of the following is also referred to as random variable?

- a) stochastic
- b) aleatory
- c) cliete
- d) all of the mentioned

[View Answer](#)

Answer: b

Explanation: Random variable is also known as stochastic variable.

7. Bayesian inference uses frequency interpretations of probabilities to control error rates.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: Frequency inference uses frequency interpretations of probabilities to control error rates.

8. Which of the following condition should be satisfied by function for pmf?

- a) The sum of all of the possible values is 1
- b) The sum of all of the possible values is 0
- c) The sum of all of the possible values is infinite
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: A probability mass function evaluated at a value corresponds to the probability that a random variable takes that value.

9. Which of the following function is associated with a continuous random variable?

- a) pdf
- b) pmv
- c) pmf
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: pdf stands for probability density function.

10. Statistical inference is the process of drawing formal conclusions from data.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: Statistical inference requires navigating the set of assumptions and tools.

1. The expected value or _____ of a random variable is the center of its distribution.

- a) mode
- b) median
- c) mean
- d) bayesian inference

[View Answer](#)

Answer: c

Explanation: A probability model connects the data to the population using assumptions.

2. Point out the correct statement.

- a) Some cumulative distribution function F is non-decreasing and right-continuous
- b) Every cumulative distribution function F is decreasing and right-continuous
- c) Every cumulative distribution function F is increasing and left-continuous
- d) None of the mentioned

[View Answer](#)

Answer: d

Explanation: Every cumulative distribution function F is non-decreasing and right-continuous.

3. Which of the following of a random variable is a measure of spread?

- a) variance
- b) standard deviation
- c) empirical mean
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: Densities with a higher variance are more spread out than densities with a lower variance.

4. The square root of the variance is called the _____ deviation.

- a) empirical
- b) mean
- c) continuous
- d) standard

[View Answer](#)

Answer: d

Explanation: Standard Deviation (SD) is the measure of spread of the numbers in a set of data from its mean value.

5. Point out the wrong statement.

- a) A percentile is simply a quantile with expressed as a percent
- b) There are two types of random variable
- c) R cannot approximate quantiles for you for common distributions
- d) None of the mentioned

[View Answer](#)

Answer: c

Explanation: R can approximate quantiles for you for common distributions.

6. Which of the following inequality is useful for interpreting variances?

- a) Chebyshev
- b) Stautaory
- c) Testory
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: Chebyshev's inequality is also spelled as Tchebyshoff's inequality.

7. For continuous random variables, the CDF is the derivative of the PDF.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: For continuous random variables, the PDF is the derivative of the CDF.

8. Chebyshev's inequality states that the probability of a "Six Sigma" event is less than _____

- a) 10%
- b) 20%
- c) 30%
- d) 3%

[View Answer](#)

Answer: d

Explanation: If a bell curve is assumed, the probability of a "six sigma" event is on the order of one ten millionth of a percent.

9. Which of the following random variables are the default model for random samples?

- a) iid
- b) id
- c) pmd
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: Random variables are said to be iid if they are independent and identically distributed.

10. Cumulative distribution functions are used to specify the distribution of multivariate random variables.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: In the case of a continuous distribution, it gives the area under the probability density function from minus infinity to x.

1. Which of the following goal is incorrectly represented in the below figure?

- a) Relationship between variables
- b) Distribution of variables
- c) Inference about relationships
- d) Causal

[View Answer](#)

Answer: d

Explanation: Causal is not directly related to goal of statistical modelling.

2. Point out the correct statement.

- a) The exponent of a normally distributed random variables follows what is called the log-normal distribution
- b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent
- c) The square of a standard normal random variable follows what is called chi-squared distribution
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: Many random variables, properly normalized, limit to a normal distribution.

3. Which of the following is incorrect with respect to use of Poisson distribution?

- a) Modeling event/time data
- b) Modeling bounded count data
- c) Modeling contingency tables
- d) All of the mentioned

[View Answer](#)

Answer: b

Explanation: Poisson distribution is used for modeling unbounded count data.

4. _____ random variables are used to model rates.

- a) Empirical
- b) Binomial
- c) Poisson
- d) All of the mentioned

[View Answer](#)

Answer: c

Explanation: Poisson distribution is used to model counts.

5. Point out the wrong statement.

- a) The normal distribution is asymmetric and peaked about its mode
- b) A constant times a normally distributed random variable is also normally distributed
- c) Sample means of normally distributed random variables are again normally distributed
- d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: The normal distribution is symmetric and peaked about its mean.

6. Which of the following form the basis for frequency interpretation of probabilities?

- a) Asymptotics
- b) Symptotics
- c) Asymmetry
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: Asymptotics is the term for the behavior of statistics as the sample size.

7. Bernoulli random variables take (only) the values 1 and 0.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: The Bernoulli distribution arises as the result of a binary outcome.

8. The _____ basically states that the sample mean is consistent.

- a) LAN
- b) LLN
- c) LWN
- d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: LLN stands for law of large numbers.

9. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?

- a) Central Limit Theorem
- b) Central Mean Theorem
- c) Centroid Limit Theorem
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: The Central Limit Theorem (CLT) is one of the most important theorems in statistics.

10. The binomial random variables are obtained as the sum of iid Gaussian trials.
- a) True
 - b) False

[View Answer](#)

Answer: a

Explanation: The binomial random variables are obtained as the sum of iid Bernoulli trials.

1. The _____ of the Chi-squared distribution is twice the degrees of freedom.
- a) variance
 - b) standard deviation
 - c) mode
 - d) none of the mentioned

[View Answer](#)

Answer: a

Explanation: The mean of the Chi-squared is its degrees of freedom.

2. Point out the correct statement.
- a) Asymptotics are incredibly useful for simple statistical inference and approximations
 - b) Asymptotics often lead to nice understanding of procedures
 - c) An estimator is consistent if it converges to what you want to estimate
 - d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: Consistency is neither necessary nor sufficient for one estimator to be better than another.

3. Gosset's distribution is invented by which of the following scientist?
- a) William Gosset
 - b) William Gosling
 - c) Gosling Gosset
 - d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: Gosset's distribution is indexed by a degrees of freedom.

4. The _____ of a collection of data is the joint density evaluated as a function of the parameters with the data fixed.
- a) probability
 - b) likelihood
 - c) poisson distribution
 - d) all of the mentioned

[View Answer](#)

Answer: b

Explanation: Likelihood analysis of data uses the likelihood to perform inference regarding the unknown parameter.

5. Point out the wrong statement.
- a) Asymptotics generally give assurances about finite sample performance
 - b) The sample variance and the sample standard deviation are consistent as well
 - c) The sample mean and the sample variance are unbiased as well
 - d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: The kinds of asymptotics that do are orders of magnitude more difficult to work with.

6. Which of the following is a property of likelihood?

- a) Ratios of likelihood values measure the relative evidence of one value of the unknown parameter to another
- b) Given a statistical model and observed data, all of the relevant information contained in the data regarding the unknown parameter is contained in the likelihood
- c) The Resultant likelihood is multiplication of individual likelihood
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: Likelihood is the hypothetical probability that an event that has already occurred would yield a specific outcome.

7. CLT is mostly useful as an approximation.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: The CLT applies in an endless variety of settings.

8. The beta distribution is the default prior for parameters between _____

- a) 0 and 10
- b) 1 and 2
- c) 0 and 1
- d) None of the mentioned

[View Answer](#)

Answer: c

Explanation: Bayesian statistics posits a prior on the parameter of interest.

9. Which of the following mean is a mixture of the MLE and the prior mean?

- a) interior
- b) exterior
- c) posterior
- d) all of the mentioned

[View Answer](#)

Answer: c

Explanation: MLE stands for maximum likelihood.

10. Usually replacing the standard error by its estimated value does change the CLT.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: Usually replacing the standard error by its estimated value doesn't change the CLT.

1. Which of the following testing is concerned with making decisions using data?

- a) Probability
- b) Hypothesis
- c) Causal
- d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: The null hypothesis is assumed true and statistical evidence is required to reject it in favor of a research or alternative hypothesis.

2. Point out the correct statement.

- a) Power of a one sided test is lower than the power of the associated two sided test

- b) Power of a two sided test is greater than the power of the associated one sided test
c) Hypothesis testing is less commonly used
d) None of the mentioned

[View Answer](#)

Answer: d

Explanation: Power of a one sided test is greater than the power of the associated two sided test.

3. Which of the following value is the most common measure of “statistical significance”?
a) P
b) A
c) L
d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: The P-value is the probability under the null hypothesis of obtaining evidence as extreme or more extreme than would be observed by chance alone.

4. What is the purpose of multiple testing in statistical inference?
a) Minimize errors
b) Minimize false positives
c) Minimize false negatives
d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: A false positive is an error in some evaluation process in which a condition tested for is mistakenly found to have been detected.

5. Point out the wrong statement with respect to FDR.
a) FDR is difficult to calculate
b) FDR is relatively less conservative
c) FDR allows for more false positives
d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: FDR stands for false discovery rate.

6. Which of the following is the oldest multiple testing correction?
a) Bonferroni correction
b) Bernoulli correction
c) Likelihood correction
d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: Bonferroni correction is easy to calculate.

7. The pooled estimator is a mixture of the group variances, placing greater weight on whichever has a larger sample size.
a) True
b) False

[View Answer](#)

Answer: a

Explanation: If the sample sizes are the same the pooled variance estimate is the average of the group variances.

8. Which of the following tool is used for constructing confidence intervals and calculating standard errors for difficult statistics?
a) bagger
b) bootstrap
c) jackknife
d) none of the mentioned

[View Answer](#)

Answer: b

Explanation: The bootstrap procedure follows from the so called bootstrap principle.

9. Which of the following tool is used for estimating standard errors and the bias of estimators?

- a) knitr
- b) jackknife
- c) ggplot2
- d) all of the mentioned

[View Answer](#)

Answer: c

Explanation: jackknife involves resampling data.

10. Power is the probability of rejecting the null hypothesis when it is true.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: Power is the probability of rejecting the null hypothesis when it is false.

1. Which of the following function can be replaced with the question mark in the below figure?

- a) boxplot
- b) lplot
- c) levelplot
- d) all of the mentioned

[View Answer](#)

Answer: c

Explanation: levelplot is used plotting “image”.

2. Point out the correct statement.

- a) The mean is a measure of central tendency of the data
- b) Empirical mean is related to “centering” the random variables
- c) The empirical standard deviation is a measure of spread
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: The process of centering and scaling the data is called “normalizing” the data.

3. Which of the following implies no relationship with respect to correlation?

- a) $\text{Cor}(X, Y) = 1$
- b) $\text{Cor}(X, Y) = 0$
- c) $\text{Cor}(X, Y) = 2$
- d) All of the mentioned

[View Answer](#)

Answer: b

Explanation: Correlation is a statistical technique that can show whether and how strongly pairs of variables are related.

4. Normalized data are centered at ___ and have units equal to standard deviations of the original data.

- a) 0
- b) 5

- c) 1
d) 10
[View Answer](#)

Answer: a
Explanation: In statistics and applications of statistics, normalization can have a range of meanings.

5. Point out the wrong statement.
a) Regression through the origin yields an equivalent slope if you center the data first
b) Normalizing variables results in the slope being the correlation
c) Least squares is not an estimation tool
d) None of the mentioned
[View Answer](#)

Answer: c
Explanation: Least squares is an estimation tool.

6. Which of the following is correct with respect to residuals?
a) Positive residuals are above the line, negative residuals are below
b) Positive residuals are below the line, negative residuals are above
c) Positive residuals and negative residuals are below the line
d) All of the mentioned
[View Answer](#)

Answer: a
Explanation: Residuals can be thought of as the outcome with the linear association of the predictor removed.

7. Minimizing the likelihood is the same as maximizing -2 log likelihood.
a) True
b) False
[View Answer](#)

Answer: a
Explanation: Maximizing the likelihood is the same as minimizing 2 log likelihood.

8. Which of the following refers to the circumstance in which the variability of a variable is unequal across the range of values of a second variable that predicts it?
a) Heterogeneity
b) Heteroskedasticity
c) Heteroelasticity
d) None of the mentioned
[View Answer](#)

Answer: b
Explanation: Heteroskedasticity has serious consequences for the OLS estimator.

9. Which of the following outcome is odd man out in the below figure?

- a) R Squared
b) Kappa
c) RMSE
d) All of the mentioned
[View Answer](#)

Answer: b

Explanation: Kappa is categorical outcome.

10. Residuals are useful for investigating best model fit.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: Residuals are useful for investigating poor model fit.

1. Which of the following is the correct formula for total variation?

- a) Total Variation = Residual Variation – Regression Variation
- b) Total Variation = Residual Variation + Regression Variation
- c) Total Variation = Residual Variation * Regression Variation
- d) All of the mentioned

[View Answer](#)

Answer: b

Explanation: The complementary part of the total variation is called unexplained or residual.

2. Point out the correct statement.

- a) A standard error is needed to create a prediction interval
- b) The prediction interval must incorporate the variability in the data around the line
- c) Investors use the residual variance to measure the accuracy of their predictions on the value of an asset
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: In statistics, explained variation measures the proportion to which a mathematical model accounts for the variation of a given data set.

3. Which of the following things can be accomplished with linear model?

- a) Flexibly fit complicated functions
- b) Uncover complex multivariate relationships
- c) Build accurate prediction models
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: Linear models are the single most important applied statistical and machine learning technique.

4. Which of the following statement is incorrect with respect to outliers?

- a) Outliers can have varying degrees of influence
- b) Outliers can be the result of spurious or real processes
- c) Outliers cannot conform to the regression relationship
- d) None of the mentioned

[View Answer](#)

Answer: c

Explanation: Outliers can conform to the regression relationship.

5. Point out the wrong statement.

- a) The fraction of variance unexplained is an established concept in the context of linear regression
- b) “Explained variance” is routinely used in principal component analysis
- c) The general linear model extends simple linear regression (SLR) by adding terms linearly into the model
- d) None of the mentioned

[View Answer](#)

Answer: d

Explanation: Linearity refers to a mathematical relationship or function that can be graphically represented as a straight line.

6. Which of the following can be useful for diagnosing data entry errors?

- a) hat values
- b) dfifit
- c) resid
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: resid returns the ordinary residuals.

7. Multivariate regression estimates are exactly those having removed the linear relationship of the other variables from both the regressor and response.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: Multivariate Data Analysis refers to any statistical technique used to analyze data that arises from more than one variable.

8. Residual _____ plots investigate normality of the errors.

- a) RR
- b) PP
- c) QQ
- d) None of the mentioned

[View Answer](#)

Answer: c

Explanation: Patterns in your residual plots generally indicate some poor aspect of model fit.

9. Which of the following show residuals divided by their standard deviations?

- a) rstudent
- b) cooks.distance
- c) rstandard
- d) all of the mentioned

[View Answer](#)

Answer: c

Explanation: rstandard stands for standardized residuals.

10. The least squares estimate for the coefficient of a multivariate regression model is exactly regression through the origin with the linear relationships.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: Multivariate regression adjusts a coefficient for the linear impact of the other variables.

1. How many components are present in generalized linear models?

- a) 2
- b) 4
- c) 6
- d) None of the mentioned

[View Answer](#)

Answer: d

Explanation: Generalized linear models involve three components.

2. Point out the wrong statement.

- a) Additive response models don't make much sense if the response is discrete, or strictly positive
- b) Transformations are often easy to interpret in linear model
- c) Regression models are used to predict one variable from one or more other variables
- d) All of the mentioned

[View Answer](#)

Answer: b

Explanation: Transformations are often hard to interpret in linear model.

3. Which of the following component is involved in generalized linear models?

- a) An exponential family model for the response
- b) A systematic component via a linear predictor
- c) A link function that connects the means of the response to the linear predictor
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: GLM is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution.

4. Collection of exchangeable binary outcomes for the same covariate data are called _____ outcomes.

- a) random
- b) direct
- c) binomial
- d) none of the mentioned

[View Answer](#)

Answer: c

Explanation: The multivariate regression model for binary outcomes gives odds ratios, not risk ratios.

5. Point out the wrong statement.

- a) Asymptotics are used for inference usually
- b) Adding squared terms makes it continuously differentiable at the knot points
- c) Adding squared terms makes it twice continuously differentiable at the knot points
- d) None of the mentioned

[View Answer](#)

Answer: c

Explanation: Adding cubic terms makes it twice continuously differentiable at the knot points.

6. Which of the following is example use of Poisson distribution?

- a) Analyzing contingency table data
- b) Modeling web traffic hits
- c) Incidence rates
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: The Poisson distribution is a useful model for counts and rates.

7. Principal components or factor analytic models on covariates are often useful for reducing complex covariate spaces.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: The space of models explodes quickly as you add interactions and polynomial terms.

8. How many outcomes are possible with bernoulli trial?

- a) 2
- b) 3
- c) 4
- d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: Bernoulli trial is a random experiment with exactly two possible outcomes.

9. Which of the following analysis is a statistical process for estimating the relationships among variables?

- a) Causal
- b) Regression
- c) Multivariate
- d) All of the mentioned

[View Answer](#)

Answer: b

Explanation: Regression models provide the scientist with a powerful tool, allowing predictions about past, present, or future events to be made with information about past or present events.

10. Linear models are the most useful applied statistical technique.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: Linear model do have limitations.

1. Which of the following can be used to generate balanced cross-validation groupings from a set of data?

- a) createFolds
- b) createSample
- c) createResample
- d) none of the mentioned

[View Answer](#)

Answer: a

Explanation: createResample can be used to make simple bootstrap samples.

2. Point out the wrong statement.

- a) Simple random sampling of time series is probably the best way to resample times series data.
- b) Three parameters are used for time series splitting
- c) Horizon parameter is the number of consecutive values in test set sample
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: Simple random sampling of time series is probably not the best way to resample times series data.

3. Which of the following function can be used to maximize the minimum dissimilarities?

- a) sumDiss
- b) minDiss
- c) avgDiss
- d) all of the mentioned

[View Answer](#)

Answer: d

Explanation: sumDiss can be used to maximize the total dissimilarities.

4. Which of the following function can create the indices for time series type of splitting?

- a) newTimeSlices
- b) createTimeSlices
- c) binTimeSlices
- d) none of the mentioned

[View Answer](#)

Answer: b

Explanation: Rolling forecasting origin techniques are associated with time series type of splitting.

5. Point out the correct statement.

- a) Asymptotics are used for inference usually
- b) Caret includes several functions to pre-process the predictor data
- c) The function dummyVars can be used to generate a complete set of dummy variables from one or more factors
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: The function dummyVars takes a formula and a data set and outputs an object that can be used to create the dummy variables using the predict method.

6. Which of the following can be used to create sub-samples using a maximum dissimilarity approach?

- a) minDissim
- b) maxDissim
- c) inmaxDissim
- d) all of the mentioned

[View Answer](#)

Answer: b

Explanation: Splitting is based on the predictors.

7. caret does not use the proxy package.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: caret uses the proxy package.

8. Which of the following function can be used to create balanced splits of the data?

- a) newDataPartition
- b) createDataPartition
- c) renameDataPartition
- d) none of the mentioned

[View Answer](#)

Answer: b

Explanation: If the y argument to this function is a factor, the random sampling occurs within each class and should preserve the overall class distribution of the data.

9. Which of the following package tools are present in caret?

- a) pre-processing
- b) feature selection
- c) model tuning
- d) all of the mentioned

[View Answer](#)

Answer: d

Explanation: There are many different modeling functions in R.

10. caret stands for classification and regression training.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: The caret package is a set of functions that attempt to streamline the process for creating predictive models.

1. Which of the following function is a wrapper for different lattice plots to visualize the data?

- a) levelplot
- b) featurePlot
- c) plotsample
- d) none of the mentioned

[View Answer](#)

Answer: b

Explanation: featurePlot is used for data visualization in caret.

2. Point out the wrong statement.

- a) In every situation, the data generating mechanism can create predictors that only have a single unique value

- b) Predictors might have only a handful of unique values that occur with very low frequencies
c) The function `findLinearCombos` uses the QR decomposition of a matrix to enumerate sets of linear combinations
d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: In some situations, the data generating mechanism can create predictors that only have a single unique value.

3. Which of the following function can be used to identify near zero-variance variables?

- a) `zeroVar`
b) `nearVar`
c) `nearZeroVar`
d) all of the mentioned

[View Answer](#)

Answer: c

Explanation: The `saveMetrics` argument can be used to show the details and usually defaults to FALSE.

4. Which of the following function can be used to flag predictors for removal?

- a) `searchCorrelation`
b) `findCausation`
c) `findCorrelation`
d) none of the mentioned

[View Answer](#)

Answer: c

Explanation: Some models thrive on correlated predictors.

5. Point out the correct statement.

- a) `findLinearColumns` will also return a vector of column positions can be removed to eliminate the linear dependencies
b) `findLinearCombos` will return a list that enumerates dependencies
c) the function `findLinearRows` can be used to generate a complete set of row variables from one factor
d) none of the mentioned

[View Answer](#)

Answer: b

Explanation: For each linear combination, it will incrementally remove columns from the matrix and test to see if the dependencies have been resolved.

6. Which of the following can be used to impute data sets based only on information in the training set?

- a) `postProcess`
b) `preProcess`
c) `process`
d) all of the mentioned

[View Answer](#)

Answer: b

Explanation: This can be done with K-nearest neighbors.

7. The function `preProcess` estimates the required parameters for each operation.

- a) True
b) False

[View Answer](#)

Answer: a

Explanation: `predict.preProcess` is used to apply them to specific data sets.

8. Which of the following can also be used to find new variables that are linear combinations of the original set with independent components?

- a) ICA
b) SCA
c) PCA
d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: ICA stands for independent component analysis.

9. Which of the following function is used to generate the class distances?

- a) preprocess.classDist
- b) predict.classDist
- c) predict.classDistance
- d) all of the mentioned

[View Answer](#)

Answer: b

Explanation: By default, the distances are logged.

10. The preProcess class can be used for many operations on predictors.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: Operations include centering and scaling.

1. varImp is a wrapper around the evimp function in the _____ package.

- a) numpy
- b) earth
- c) plot
- d) none of the mentioned

[View Answer](#)

Answer: b

Explanation: The earth package is an implementation of Jerome Friedman's Multivariate Adaptive Regression Splines.

2. Point out the wrong statement.

- a) The trapezoidal rule is used to compute the area under the ROC curve
- b) For regression, the relationship between each predictor and the outcome is evaluated
- c) An argument, para, is used to pick the model fitting technique
- d) All of the mentioned

[View Answer](#)

Answer: c

Explanation: An argument, nonpara, is used to pick the model fitting technique.

3. Which of the following curve analysis is conducted on each predictor for classification?

- a) NOC
- b) ROC
- c) COC
- d) All of the mentioned

[View Answer](#)

Answer: b

Explanation: For two class problems, a series of cutoffs is applied to the predictor data to predict the class.

4. Which of the following function tracks the changes in model statistics?

- a) varImp
- b) varImpTrack
- c) findTrack
- d) none of the mentioned

[View Answer](#)

Answer: a

Explanation: GCV change value can also be tracked.

5. Point out the correct statement.

- a) The difference between the class centroids and the overall centroid is used to measure the variable influence

- b) The Bagged Trees output contains variable usage statistics
c) Boosted Trees uses different approach as a single tree
d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: The larger the difference between the class centroid and the overall center of the data, the larger the separation between the classes.

6. Which of the following model include a backwards elimination feature selection routine?

- a) MCV
b) MARS
c) MCRS
d) All of the mentioned

[View Answer](#)

Answer: b

Explanation: MARS stands for Multivariate Adaptive Regression Splines.

7. The advantage of using a model-based approach is that is more closely tied to the model performance.

- a) True
b) False

[View Answer](#)

Answer: a

Explanation: Model-based approach is able to incorporate the correlation structure between the predictors into the importance calculation.

8. Which of the following model sums the importance over each boosting iteration?

- a) Boosted trees
b) Bagged trees
c) Partial least squares
d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: gbm package can be used here.

9. Which of the following argument is used to set importance values?

- a) scale
b) set
c) value
d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: All measures of importance are scaled to have a maximum value of 100.

10. For most classification models, each predictor will have a separate variable importance for each class.

- a) True
b) False

[View Answer](#)

Answer: a

Explanation: The exceptions are classification trees, bagged trees and boosted trees.

1. Which of the following is the valid component of the predictor?

- a) data
b) question
c) algorithm
d) all of the mentioned

[View Answer](#)

Answer: d

Explanation: A prediction is a statement about the future.

2. Point out the wrong statement.

- a) In Sample Error is also called generalization error
- b) Out of Sample Error is the error rate you get on the new dataset
- c) In Sample Error is also called resubstitution error
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: Out of Sample Error is also called generalization error.

3. Which of the following is correct order of working?

- a) questions->input data ->algorithms
- b) questions->evaluation ->algorithms
- c) evaluation->input data ->algorithms
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: Evaluation is done in the last.

4. Which of the following shows correct relative order of importance?

- a) question->features->data->algorithms
- b) question->data->features->algorithms
- c) algorithms->data->features->question
- d) none of the mentioned

[View Answer](#)

Answer: b

Explanation: Garbage in should be equal to garbage out.

5. Point out the correct statement.

- a) In Sample Error is the error rate you get on the same dataset used to model a predictor
- b) Data have two parts-signal and noise
- c) The goal of predictor is to find signal
- d) None of the mentioned

[View Answer](#)

Answer: d

Explanation: Perfect in sample prediction can be built.

6. Which of the following is characteristic of best machine learning method?

- a) Fast
- b) Accuracy
- c) Scalable
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: There is always a trade-off in prediction accuracy.

7. True positive means correctly rejected.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: True positive means correctly identified.

8. Which of the following trade-off occurs during prediction?

- a) Speed vs Accuracy

- b) Simplicity vs Accuracy
- c) Scalability vs Accuracy
- d) None of the mentioned

[View Answer](#)

Answer: d

Explanation: Interpretability also matters during prediction.

9. Which of the following expression is true?

- a) In sample error < out sample error
- b) In sample error > out sample error
- c) In sample error = out sample error
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: Out of sample error is given more importance.

10. Backtesting is a key component of effective trading-system development.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: Backtesting is the process of applying a trading strategy or analytical method to historical data to see how accurately the strategy or method would have predicted actual results.

1. Which of the following is correct use of cross validation?

- a) Selecting variables to include in a model
- b) Comparing predictors
- c) Selecting parameters in prediction function
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: Cross-validation is also used to pick type of prediction function to be used.

2. Point out the wrong combination.

- a) True negative=correctly rejected
- b) False negative=correctly rejected
- c) False positive=correctly identified
- d) All of the mentioned

[View Answer](#)

Answer: c

Explanation: False positive means incorrectly identified.

3. Which of the following is a common error measure?

- a) Sensitivity
- b) Median absolute deviation
- c) Specificity
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: Sensitivity and specificity are statistical measures of the performance of a binary classification test, also known in statistics as classification function.

4. Which of the following is not a machine learning algorithm?

- a) SVG
- b) SVM
- c) Random forest
- d) None of the mentioned

[View Answer](#)

Answer: a
Explanation: SVM stands for scalable vector machine.

5. Point out the wrong statement.
a) ROC curve stands for receiver operating characteristic
b) Foretime series, data must be in chunks
c) Random sampling must be done with replacement
d) None of the mentioned

[View Answer](#)

Answer: d
Explanation: Random sampling with replacement is the bootstrap.

6. Which of the following is a categorical outcome?
a) RMSE
b) RSquared
c) Accuracy
d) All of the mentioned

[View Answer](#)

Answer: c
Explanation: RMSE stands for Root Mean Squared Error.

7. For k cross-validation, larger k value implies more bias.
a) True
b) False

[View Answer](#)

Answer: b
Explanation: For k cross-validation, larger k value implies less bias.

8. Which of the following method is used for trainControl resampling?
a) repeatedcv
b) svm
c) bag32
d) none of the mentioned

[View Answer](#)

Answer: a
Explanation: repeatedcv stands for repeated cross-validation.

9. Which of the following can be used to create the most common graph types?
a) qplot
b) quickplot
c) plot
d) all of the mentioned

[View Answer](#)

Answer: a
Explanation: qplot() is short for a quick plot.

10. For k cross-validation, smaller k value implies less variance.
a) True
b) False

[View Answer](#)

Answer: a
Explanation: Larger k value implies more variance.

1. Predicting with trees evaluate _____ within each group of data.
a) equality
b) homogeneity
c) heterogeneity

- d) all of the mentioned
[View Answer](#)

Answer: b

Explanation: Predicting with trees is easy to interpret.

2. Point out the wrong statement.

- a) Training and testing data must be processed in different way
b) Test transformation would mostly be imperfect
c) The first goal is statistical and second is data compression in PCA
d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: Training and testing data must be processed in same way.

3. Which of the following method options is provided by train function for bagging?

- a) bagEarth
b) treebag
c) bagFDA
d) all of the mentioned

[View Answer](#)

Answer: d

Explanation: Bagging can be done using bag function as well.

4. Which of the following is correct with respect to random forest?

- a) Random forest are difficult to interpret but often very accurate
b) Random forest are easy to interpret but often very accurate
c) Random forest are difficult to interpret but very less accurate
d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: Random forest is top performing algorithm in prediction.

5. Point out the correct statement.

- a) Prediction with regression is easy to implement
b) Prediction with regression is easy to interpret
c) Prediction with regression performs well when linear model is correct
d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: Prediction with regression gives poor performance in non linear settings.

6. Which of the following library is used for boosting generalized additive models?

- a) gamBoost
b) gbm
c) ada
d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: Boosting can be used with any subset of classifier.

7. The principal components are equal to left singular values if you first scale the variables.

- a) True
b) False

[View Answer](#)

Answer: b

Explanation: The principal components are equal to left singular values if you first scale the variables.

8. Which of the following is statistical boosting based on additive logistic regression?

- a) gamBoost
- b) gbm
- c) ada
- d) mboost

[View Answer](#)

Answer: a

Explanation: mboost is used for model based boosting.

9. Which of the following is one of the largest boost subclass in boosting?

- a) variance boosting
- b) gradient boosting
- c) mean boosting
- d) all of the mentioned

[View Answer](#)

Answer: b

Explanation: R has multiple boosting libraries.

10. PCA is most useful for non linear type models.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: PCA is most useful for linear type models.

1. Which of the following is correct about regularized regression?

- a) Can help with bias trade-off
- b) Cannot help with model selection
- c) Cannot help with variance trade-off
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: Regularized regression does not perform as well as random forest.

2. Point out the wrong statement.

- a) Model based approach may be computationally convenient
- b) Model based approach use Bayes theorem
- c) Model based approach are reasonably inaccurate on real problems
- d) All of the mentioned

[View Answer](#)

Answer: c

Explanation: Model based approach are reasonably accurate on real problems.

3. Which of the following methods are present in caret for regularized regression?

- a) ridge
- b) lasso
- c) relaxo
- d) all of the mentioned

[View Answer](#)

Answer: d

Explanation: In caret one can tune over the no of predictors to retain instead of defined values for penalty.

4. Which of the following method can be used to combine different classifiers?

- a) Model stacking
- b) Model combining
- c) Model structuring
- d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: Model ensembling is also used for combining different classifiers.

5. Point out the correct statement.

- a) Combining classifiers improves interpretability
- b) Combining classifiers reduces accuracy
- c) Combining classifiers improves accuracy
- d) All of the mentioned

[View Answer](#)

Answer: c

Explanation: You can combine classifier by averaging.

6. Which of the following function provides unsupervised prediction?

- a) cl_forecast
- b) cl_nowcast
- c) cl_precast
- d) none of the mentioned

[View Answer](#)

Answer: d

Explanation: cl_predict function is clue package provides unsupervised prediction.

7. Model based prediction considers relatively easy version for covariance matrix.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: Model based prediction considers relatively easy version for covariance matrix.

8. Which of the following is used to assist the quantitative trader in the development?

- a) quantmod
- b) quantile
- c) quantity
- d) mboost

[View Answer](#)

Answer: a

Explanation: Quandl package is similar to quantmod.

9. Which of the following function can be used for forecasting?

- a) predict
- b) forecast
- c) ets
- d) all of the mentioned

[View Answer](#)

Answer: b

Explanation: Forecasting is the process of making predictions of the future based on past and present data and analysis of trends.

10. Predictive analytics is same as forecasting.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: Predictive analytics goes beyond forecasting.

1. Which of the following project is used for calling R products from web?

- a) OpenCPU
- b) OpenDisk
- c) OpenMem

- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: OpenCPU is complementary to OpenCPU.

2. Point out the wrong statement.

- a) Shiny is platform for creating interactive programs embedded in to web page
- b) Shiny is invented by R folks
- c) Time required to create data products using shiny is more
- d) All of the mentioned

[View Answer](#)

Answer: c

Explanation: Time to create data products is less using shiny.

3. Which of the following statement will install shiny?

- a) install.packages("shiny")
- b) install.library("shiny")
- c) install.lib("shiny")
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: Shiny applications are automatically "live" in the same way that spreadsheets are live.

4. Which of the following can be done by shiny?

- a) Tabbed main panels
- b) Editable data tables
- c) Dynamic UI
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: shiny allows users to upload files.

5. Point out the correct statement.

- a) shiny project is a directory containing at least three parts
- b) shiny project is a file containing at least three parts
- c) shiny project consist is a directory containing only one part
- d) none of the mentioned

[View Answer](#)

Answer: d

Explanation: shiny project consist is a directory containing at least two parts.

6. Which of the following function can interrupt execution and can be called continuously?

- a) browser()
- b) browse()
- c) search()
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: Debugging shiny apps can be difficult.

7. runApp() will run the shiny and open the browser window.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: The chart is rendered within the browser using Flash.

8. Which of the following function is for single checkbox widget?

- a) checkboxInput
- b) dateInput
- c) singleboxInput
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: Shiny comes with a family of pre-built widgets, each created with a transparently named R function.

9. How many components are involved in shiny?

- a) 3
- b) 4
- c) 5
- d) none of the mentioned

[View Answer](#)

Answer: d

Explanation: Shiny apps have two components:user-interface script and server script.

10. All of the styled elements are handled through server.R.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: All of the styled elements are handled through ui.R.

1. Which of the following framework is compatible with slidify?

- a) io2015
- b) io2012
- c) d3
- d) all of the mentioned

[View Answer](#)

Answer: b

Explanation: D3 is a JavaScript library for visualizing data with HTML, SVG, and CSS.

2. Point out the wrong statement.

- a) Slidify is created by Ramnath Vaidyanathan
- b) Slidify is non customizable
- c) Slidify presentation are just HTML files
- d) All of the mentioned

[View Answer](#)

Answer: b

Explanation: Slidify is customizable and extendable.

3. Which of the following statement will load slidify?

- a) library(slidify)
- b) install.library(slidify)
- c) install.load(slidify)
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: Devtools should be installed in advance.

4. Which of the following will be used to compose the content of the presentation?

- a) ui.RMD
- b) index.RMD
- c) server.RMD
- d) all of the mentioned

[View Answer](#)

Answer: b

Explanation: index.RMD is an R markdown document.

5. Point out the correct statement.

- a) Slidify allows embedded code chunks
- b) Slidify presentation cannot be shared easily
- c) Slidify is difficult to use
- d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: Slidify allows mathematical formulas as well.

6. Which of the following statement generates a html slide deck from index.Rmd?

- a) slidify("index.Rmd")
- b) lib.slidify("index.Rmd")
- c) slidifylib("index.Rmd")
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: It is a static file, which means that you can open it in your browser locally and it should display fine.

7. The first part of index.Rmd is XML code.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: The first part of index.Rmd is YAML code.

8. Which of the following statement will install slidify from github?

- a) install_github('slidify', 'ramnathv')
- b) install_github('slidify', 'r')
- c) install('slidify', 'ramnathv')
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: Slidify is not on CRAN.

9. Which of the following element can be added to slidify?

- a) Quiz
- b) RCharts
- c) Shiny apps
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: Many interactive elements can be added to slidify.

10. MathJax is a cross-browser JavaScript library that displays mathematical notation in web browsers.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: MathJax uses MathML.

1. Which of the following is R interface to google charts?

- a) googleVis
- b) googleChart
- c) googleDataVis

- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: googleVis allow users to create interactive charts based on data frames.

2. Point out the wrong statement.

- a) The plot command does open a graphics device in the modern way
- b) Motion Chart is only displayed when hosted on a web server
- c) gvisMotionChart is used to create motion chart
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: The plot command does not open a graphics device in the traditional way.

3. Which of the following create a Google Gadget based on a Google Visualization Object?

- a) createGadget
- b) createGoogleGadget
- c) newGoogleGadget
- d) all of the mentioned

[View Answer](#)

Answer: b

Explanation: createGoogleGadget returns a Google Gadget XML string.

4. Which of the following reads a data.frame and creates text output referring to the Google Visualization API?

- a) gvisAnnotatedLine
- b) gvisTimeLine
- c) gvisAnnotatedTimeLine
- d) none of the mentioned

[View Answer](#)

Answer: c

Explanation: An annotated time line is an interactive time series line chart with optional annotations.

5. Point out the correct statement.

- a) gvisAnnotationChart returns list of class "gvis" and "list"
- b) The gvisAreaChart function reads a data.frame and creates text output referring to the Google Visualization API
- c) gvisAreaChart returns list of class "gvis" and "list"
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: This can be included into a web page, or as a stand-alone page.

6. Which of the following is used for creating interacting tables?

- a) gvisGeoChart
- b) gvisTable
- c) gvisLineChart
- d) all of the mentioned

[View Answer](#)

Answer: b

Explanation: gvisLineChart is used for creating line charts.

7. gvisAnnotatedTimeLine returns list of class "gvis" and "list".

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: The chart is rendered within the browser using flash.

8. The actual chart of gvisBarChart is rendered by the web browser using _____ or VML.

- a) JPEG
- b) SVG
- c) PDF
- d) All of the mentioned

[View Answer](#)

Answer: b

Explanation: gvisBarChart reads data frame.

9. Which of the following is used for creating tree maps?

- a) gvisGeoChart
- b) gvisTable
- c) gvisTreeMap
- d) all of the mentioned

[View Answer](#)

Answer: c

Explanation: gvisGeoChart is used for interactive maps.

10. gvisAnnotationChart charts are interactive time series line charts that support annotations.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: Unlike the gvisAnnotatedTimeLine, which uses flash, annotation charts are SVG/VML and should be preferred whenever possible.

1. Which of the following is contained in NumPy library?

- a) n-dimensional array object
- b) tools for integrating C/C++ and Fortran code
- c) fourier transform
- d) all of the mentioned

[View Answer](#)

Answer: d

Explanation: NumPy is the fundamental package for scientific computing with Python.

2. Point out the wrong statement.

- a) ipython is an enhanced interactive Python shell
- b) matplotlib will enable you to plot graphics
- c) rPy provides a lot of scientific routines that work on top of NumPy
- d) all of the mentioned

[View Answer](#)

Answer: c

Explanation: SciPy provides a lot of scientific routines that work on top of NumPy.

3. The _____ function returns its argument with a modified shape, whereas the _____ method modifies the array itself.

- a) reshape, resize
- b) resize, reshape
- c) reshape2, resize
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: If a dimension is given as -1 in a reshaping operation, the other dimensions are automatically calculated.

4. To create sequences of numbers, NumPy provides a function _____ analogous to range that returns arrays instead of lists.

- a) arange
- b) aspace
- c) aline
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: When arange is used with floating point arguments, it is generally not possible to predict the number of elements obtained.

5. Point out the correct statement.

- a) NumPy main object is the homogeneous multidimensional array
- b) In Numpy, dimensions are called axes
- c) Numpy array class is called ndarray
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: The number of axes is called rank.

6. Which of the following function stacks 1D arrays as columns into a 2D array?

- a) row_stack
- b) column_stack
- c) com_stack
- d) all of the mentioned

[View Answer](#)

Answer: b

Explanation: column_stack is equivalent to vstack only for 1D arrays.

7. ndarray is also known as the alias array.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: numpy.array is not the same as the Standard Python Library class array.array.

8. Which of the following method creates a new array object that looks at the same data?

- a) view
- b) copy
- c) paste
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: The copy method makes a complete copy of the array and its data.

9. Which of the following function can be used to combine different vectors so as to obtain the result for each n-uplet?

- a) iid_
- b) ix_
- c) ixd_
- d) all of the mentioned

[View Answer](#)

Answer: b

Explanation: Length of the 1D boolean array must coincide with the length of the dimension (or axis) you want to slice.

10. ndarray.dataitemSize is the buffer containing the actual elements of the array.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: ndarray.data is the buffer containing the actual elements of the array.

1. Which of the following sets the size of the buffer used in ufuncs?

- a) bufsize(size)
- b) setszie(size)

- c) setbufsize(size)
- d) all of the mentioned

[View Answer](#)

Answer: c

Explanation: Adjusting the size of the buffer may therefore alter the speed at which ufunc calculations of various sorts are completed.

2. Point out the wrong statement.

- a) A universal function is a function that operates on ndarrays in an element-by-element fashion
- b) In NumPy, universal functions are instances of the numpy.ufunction class
- c) Many of the built-in functions are implemented in compiled C code
- d) All of the mentioned

[View Answer](#)

Answer: b

Explanation: ufunc instances can also be produced using the frompyfunc factory function.

3. Which of the following attribute should be used while checking for type combination input and output?

- a).types
- b).type
- c).class
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: Universal functions in NumPy are flexible enough to have mixed type signatures.

4. Which of the following returns an array of ones with the same shape and type as a given array?

- a) all_like
- b) ones_like
- c) one_alike
- d) all of the mentioned

[View Answer](#)

Answer: b

Explanation: The optional output arguments of the function can be used to help you save memory for large calculations.

5. Point out the wrong statement.

- a) Each universal function takes array inputs and produces array outputs
- b) Broadcasting is used throughout NumPy to decide how to handle disparately shaped arrays
- c) The output of the ufunc is necessarily an ndarray, if all input arguments are ndarrays
- d) All of the mentioned

[View Answer](#)

Answer: c

Explanation: The output of the ufunc is not necessarily an ndarray, if all input arguments are not ndarrays.

6. Which of the following set the floating-point error callback function or log object?

- a) setter
- b) settercall
- c) setterstack
- d) all of the mentioned

[View Answer](#)

Answer: b

Explanation: seterr sets how floating-point errors are handled.

7. Some ufuncs can take output arguments.

- a) True
- b) False

[View Answer](#)

Answer: b

Explanation: All ufuncs can take output arguments. If necessary, output will be cast to the data-type of the provided output array.

8. _____ decompose the elements of x into mantissa and two's exponent.

- a) trunc
- b) fmod
- c) frexp
- d) ldexp

[View Answer](#)

Answer: c

Explanation: fmod function return the element-wise remainder of division.

9. Which of the following function take only single value as input?

- a) iscomplex
- b) minimum
- c) fmin
- d) all of the mentioned

[View Answer](#)

Answer: a

Explanation: iscomplex function returns a bool array, where true if input element is complex.

10. The array object returned by `__array_prepare__` is passed to the ufunc for computation.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: If the class has an `__array_wrap__` method, the returned ndarray result will be passed to that method just before passing control back to the caller.

1. When talking to a speech recognition program, the program divides each second of your speech into 100 separate _____

- a) Codes
- b) Phonemes
- c) Samples
- d) Words

[View Answer](#)

Answer: c

Explanation: None.

2. Which term is used for describing the judgmental or commonsense part of problem solving?

- a) Heuristic
- b) Critical
- c) Value based
- d) Analytical

[View Answer](#)

Answer: a

Explanation: None.

3. Which stage of the manufacturing process has been described as "the mapping of function onto form"?

- a) Design
- b) Distribution
- c) Project management
- d) Field service

[View Answer](#)

Answer: a

Explanation: None.

4. Which kind of planning consists of successive representations of different levels of a plan?

- a) hierarchical planning
- b) non-hierarchical planning
- c) project planning
- d) all of the mentioned

[View Answer](#)

Answer: a
Explanation: None.

5. What was originally called the “imitation game” by its creator?

- a) The Turing Test
- b) LISP
- c) The Logic Theorist
- d) Cybernetics

[View Answer](#)

Answer: a
Explanation: None.

6. Decision support programs are designed to help managers make _____

- a) budget projections
- b) visual presentations
- c) business decisions
- d) vacation schedules

[View Answer](#)

Answer: c
Explanation: None.

7. PROLOG is an AI programming language, which solves problems with a form of symbolic logic known as predicate calculus. It was developed in 1972 at the University of Marseilles by a team of specialists. Can you name the person who headed this team?

- a) Alain Colmerauer
- b) Niklaus Wirth
- c) Seymour Papert
- d) John McCarthy

[View Answer](#)

Answer: a
Explanation: None.

8. Programming a robot by physically moving it through the trajectory you want it to follow be called _____

- a) contact sensing control
- b) continuous-path control
- c) robot vision control
- d) pick-and-place control

[View Answer](#)

Answer: b
Explanation: None.

9. To invoke the LISP system, you must enter _____

- a) AI
- b) LISP
- c) CL (Common Lisp)
- d) Both LISP and CL

[View Answer](#)

Answer: b
Explanation: None.

10. In LISP, what is the function (list-length <list>)?

- a) returns a new list that is equal to <list> by copying the top-level element of <list>
- b) returns the length of <list>
- c) returns t if <list> is empty
- d) all of the mentioned

[View Answer](#)

Answer: b
Explanation: None.

11. ART (Automatic Reasoning Tool) is designed to be used on _____

- a) LISP machines
- b) Personal computers
- c) Microcomputers
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: None.

12. Which particular generation of computers is associated with artificial intelligence?

- a) Second
- b) Fourth
- c) Fifth
- d) Third

[View Answer](#)

Answer: c

Explanation: None.

13. Shaping teaching techniques to fit the learning patterns of individual students is the goal of _____

- a) decision support
- b) automatic programming
- c) intelligent computer-assisted instruction
- d) expert systems

[View Answer](#)

Answer: c

Explanation: None.

14. Which of the following function returns t if the object is a symbol in LISP?

- a) (* <object>)
- b) (symbolp <object>)
- c) (nonnumeric <object>)
- d) (constantp <object>)

[View Answer](#)

Answer: b

Explanation: None.

15. The symbols used in describing the syntax of a programming language are _____

- a) 0
- b) {}
- c) ''''
- d) <>

[View Answer](#)

Answer: d

Explanation: None.

1. Ambiguity may be caused by _____

- a) syntactic ambiguity
- b) multiple word meanings
- c) unclear antecedents
- d) all of the mentioned

[View Answer](#)

Answer: d

Explanation: None.

2. Which company offers the LISP machine considered "the most powerful symbolic processor available"?

- a) LMI
- b) Symbolics
- c) Xerox

- d) Texas Instruments
[View Answer](#)

Answer: b
Explanation: None.

3. What of the following is considered a pivotal event in the history of Artificial Intelligence?

- a) 1949, Donald O, The organization of Behavior
- b) 1950, Computing Machinery and Intelligence
- c) 1956, Dartmouth University Conference Organized by John McCarthy
- d) 1961, Computer and Computer Sense

[View Answer](#)

Answer: c
Explanation: None.

4. What are the two subfields of Natural language processing?

- a) symbolic and numeric
- b) time and motion
- c) algorithmic and heuristic
- d) understanding and generation

[View Answer](#)

Answer: c
Explanation: None.

5. High-resolution, bit-mapped displays are useful for displaying _____

- a) clearer characters
- b) graphics
- c) more characters
- d) all of the mentioned

[View Answer](#)

Answer: c
Explanation: None.

6. A bidirectional feedback loop links computer modeling with _____

- a) artificial science
- b) heuristic processing
- c) human intelligence
- d) cognitive science

[View Answer](#)

Answer: c
Explanation: None.

7. Which of the following have people traditionally done better than computers?

- a) recognizing relative importance
- b) finding similarities
- c) resolving ambiguity
- d) all of the mentioned

[View Answer](#)

Answer: c
Explanation: None.

8. In LISP, the function evaluates both and is _____

- a) set
- b) setq
- c) add
- d) eva

[View Answer](#)

Answer: a
Explanation: None.

9. Which type of actuator generates a good deal of power but tends to be messy?
- a) electric
 - b) hydraulic
 - c) pneumatic
 - d) both hydraulic & pneumatic

[View Answer](#)

Answer: b
Explanation: None.

10. Research scientists all over the world are taking steps towards building computers with circuits patterned after the complex interconnections existing among the human brain's nerve cells. What name is given to such type of computers?
- a) Intelligent computers
 - b) Supercomputers
 - c) Neural network computers
 - d) Smart computers

[View Answer](#)

Answer: c
Explanation: None.

11. The integrated circuit was invented by Jack Kilby of _____
- a) MIT
 - b) Texas Instruments
 - c) Xerox
 - d) All of the mentioned

[View Answer](#)

Answer: b
Explanation: None.

12. People overcome natural language problems by _____
- a) grouping attributes into frames
 - b) understanding ideas in context
 - c) identifying with familiar situations
 - d) both understanding ideas in context & identifying with familiar situations

[View Answer](#)

Answer: d
Explanation: None.

13. The Cedar, BBN Butterfly, Cosmic Cube and Hypercube machine can be characterized as _____
- a) SISD
 - b) MIMD
 - c) SIMD
 - d) MISD

[View Answer](#)

Answer: b
Explanation: None.

14. A series of AI systems, developed by Pat Langley to explore the role of heuristics in scientific discovery is _____
- a) RAMD
 - b) BACON
 - c) MIT
 - d) DU

[View Answer](#)

Answer: b
Explanation: None.

1. Nils Nilsson headed a team at SRI that created a mobile robot named _____

- a) Robotics
- b) Dedalus
- c) Shakey
- d) Vax

[View Answer](#)

Answer: c

Explanation: None.

2. An Artificial Intelligence technique that allows computers to understand associations and relationships between objects and events is called _____

- a) heuristic processing
- b) cognitive science
- c) relative symbolism
- d) pattern matching

[View Answer](#)

Answer: c

Explanation: None.

3. The new organization established to implement the Fifth Generation Project is called _____

- a) ICOT (Institute for New Generation Computer Technology)
- b) MITI (Ministry of International Trade and Industry)
- c) MCC (Microelectronics and Computer Technology Corporation)
- d) SCP (Strategic Computing Program)

[View Answer](#)

Answer: a

Explanation: None.

4. What is the field that investigates the mechanics of human intelligence?

- a) history
- b) cognitive science
- c) psychology
- d) sociology

[View Answer](#)

Answer: b

Explanation: None.

5. What is the name of the computer program that simulates the thought processes of human beings?

- a) Human logic
- b) Expert reason
- c) Expert system
- d) Personal information

[View Answer](#)

Answer: c

Explanation: None.

6. What is the name of the computer program that contains the distilled knowledge of an expert?

- a) Database management system
- b) Management information System
- c) Expert system
- d) Artificial intelligence

[View Answer](#)

Answer: c

Explanation: None.

7. Claude Shannon described the operation of electronic switching circuits with a system of mathematical logic called _____

- a) LISP
- b) XLISP
- c) Neural networking

- d) Boolean algebra
[View Answer](#)

Answer: c
Explanation: None.

8. A computer program that contains expertise in a particular domain is called?

- a) intelligent planner
b) automatic processor
c) expert system
d) operational symbolizer

[View Answer](#)

Answer: c
Explanation: None.

9. What is the term used for describing the judgmental or commonsense part of problem solving?

- a) Heuristic
b) Critical
c) Value based
d) Analytical

[View Answer](#)

Answer: a
Explanation: None.

10. What was originally called the “imitation game” by its creator?

- a) The Turing Test
b) LISP
c) The Logic Theorist
d) Cybernetics

[View Answer](#)

Answer: a
Explanation: None.

11. Decision support programs are designed to help managers make _____

- a) budget projections
b) visual presentations
c) business decisions
d) vacation schedules

[View Answer](#)

Answer: c
Explanation: None.

12. Programming a robot by physically moving it through the trajectory you want it to follow is called _____

- a) contact sensing control
b) continuous-path control
c) robot vision control
d) pick-and-place control

[View Answer](#)

Answer: b
Explanation: None

1. What is the primary interactive method of communication used by humans?

- a) reading
b) writing
c) speaking
d) all of the mentioned

[View Answer](#)

Answer: c
Explanation: None.

2. Elementary linguistic units that are smaller than words are?
- a) allophones
 - b) phonemes
 - c) syllables
 - d) all of the mentioned

[View Answer](#)

Answer: d
Explanation: None.

3. In LISP, the atom that stands for "true" is _____
- a) t
 - b) ml
 - c) y
 - d) time

[View Answer](#)

Answer: a
Explanation: None.

4. A mouse device may be _____
- a) electro-chemical
 - b) mechanical
 - c) optical
 - d) both mechanical and optical

[View Answer](#)

Answer: d
Explanation: None.

5. An expert system differs from a database program in that only an expert system _____
- a) contains declarative knowledge
 - b) contains procedural knowledge
 - c) features the retrieval of stored information
 - d) expects users to draw their own conclusions

[View Answer](#)

Answer: b
Explanation: None.

6. Arthur Samuel is linked inextricably with a program that played _____
- a) checkers
 - b) chess
 - c) cricket
 - d) football

[View Answer](#)

Answer: a
Explanation: None.

7. Natural language understanding is used in _____
- a) natural language interfaces
 - b) natural language front ends
 - c) text understanding systems
 - d) all of the mentioned

[View Answer](#)

Answer: d
Explanation: None.

8. Which of the following are examples of software development tools?

- a) debuggers
- b) editors
- c) assemblers, compilers and interpreters
- d) all of the mentioned

[View Answer](#)

Answer: d

Explanation: None.

9. Which is the first AI programming language?

- a) BASIC
- b) FORTRAN
- c) IPL(Inductive logic programming)
- d) LISP

[View Answer](#)

Answer: d

Explanation: None.

10. The Personal Consultant is based on?

- a) EMYCIN
- b) OPS5+
- c) XCON
- d) All of the mentioned

[View Answer](#)

Answer: d

Explanation: None.

1. What is Machine learning?

- a) The autonomous acquisition of knowledge through the use of computer programs
- b) The autonomous acquisition of knowledge through the use of manual programs
- c) The selective acquisition of knowledge through the use of computer programs
- d) The selective acquisition of knowledge through the use of manual programs

[View Answer](#)

Answer: a

Explanation: Machine learning is the autonomous acquisition of knowledge through the use of computer programs.

2. Which of the factors affect the performance of learner system does not include?

- a) Representation scheme used
- b) Training scenario
- c) Type of feedback
- d) Good data structures

[View Answer](#)

Answer: d

Explanation: Factors that affect the performance of learner system does not include good data structures.

3. Different learning methods does not include?

- a) Memorization
- b) Analogy
- c) Deduction
- d) Introduction

[View Answer](#)

Answer: d

Explanation: Different learning methods does not include the introduction.

4. In language understanding, the levels of knowledge that does not include?

- a) Phonological
- b) Syntactic
- c) Empirical

- d) Logical
[View Answer](#)

Answer: c
Explanation: In language understanding, the levels of knowledge that does not include empirical knowledge.

5. A model of language consists of the categories which does not include?

- a) Language units
- b) Role structure of units
- c) System constraints
- d) Structural units

[View Answer](#)

Answer: d
Explanation: A model of language consists of the categories which does not include structural units.

6. What is a top-down parser?

- a) Begins by hypothesizing a sentence (the symbol S) and successively predicting lower level constituents until individual preterminal symbols are written
- b) Begins by hypothesizing a sentence (the symbol S) and successively predicting upper level constituents until individual preterminal symbols are written
- c) Begins by hypothesizing lower level constituents and successively predicting a sentence (the symbol S)
- d) Begins by hypothesizing upper level constituents and successively predicting a sentence (the symbol S)

[View Answer](#)

Answer: a
Explanation: A top-down parser begins by hypothesizing a sentence (the symbol S) and successively predicting lower level constituents until individual preterminal symbols are written.

7. Among the following which is not a horn clause?

- a) p
- b) $\emptyset p \vee q$
- c) $p \rightarrow q$
- d) $p \rightarrow \emptyset q$

[View Answer](#)

Answer: d
Explanation: $p \rightarrow \emptyset q$ is not a horn clause.

8. The action 'STACK(A, B)' of a robot arm specify to _____

- a) Place block B on Block A
- b) Place blocks A, B on the table in that order
- c) Place blocks B, A on the table in that order
- d) Place block A on block B

[View Answer](#)

Answer: d
Explanation: The action 'STACK(A,B)' of a robot arm specify to Place block A on block B.

1. How many terms are required for building a bayes model?

- a) 1
- b) 2
- c) 3
- d) 4

[View Answer](#)

Answer: c
Explanation: The three required terms are a conditional probability and two unconditional probability.

2. What is needed to make probabilistic systems feasible in the world?

- a) Reliability
- b) Crucial robustness
- c) Feasibility
- d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: On a model-based knowledge provides the crucial robustness needed to make probabilistic system feasible in the real world.

3. Where does the bayes rule can be used?

- a) Solving queries
- b) Increasing complexity
- c) Decreasing complexity
- d) Answering probabilistic query

[View Answer](#)

Answer: d

Explanation: Bayes rule can be used to answer the probabilistic queries conditioned on one piece of evidence.

4. What does the bayesian network provides?

- a) Complete description of the domain
- b) Partial description of the domain
- c) Complete description of the problem
- d) None of the mentioned

[View Answer](#)

Answer: a

Explanation: A Bayesian network provides a complete description of the domain.

5. How the entries in the full joint probability distribution can be calculated?

- a) Using variables
- b) Using information
- c) Both Using variables & information
- d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: Every entry in the full joint probability distribution can be calculated from the information in the network.

6. How the bayesian network can be used to answer any query?

- a) Full distribution
- b) Joint distribution
- c) Partial distribution
- d) All of the mentioned

[View Answer](#)

Answer: b

Explanation: If a bayesian network is a representation of the joint distribution, then it can solve any query, by summing all the relevant joint entries.

7. How the compactness of the bayesian network can be described?

- a) Locally structured
- b) Fully structured
- c) Partial structure
- d) All of the mentioned

[View Answer](#)

Answer: a

Explanation: The compactness of the bayesian network is an example of a very general property of a locally structured system.

8. To which does the local structure is associated?

- a) Hybrid
- b) Dependant
- c) Linear
- d) None of the mentioned

[View Answer](#)

Answer: c

Explanation: Local structure is usually associated with linear rather than exponential growth in complexity.

9. Which condition is used to influence a variable directly by all the others?

- a) Partially connected
- b) Fully connected
- c) Local connected
- d) None of the mentioned

[View Answer](#)

Answer: b

Explanation: None.

10. What is the consequence between a node and its predecessors while creating bayesian network?

- a) Functionally dependent
- b) Dependant
- c) Conditionally independent
- d) Both Conditionally dependant & Dependant

[View Answer](#)

Answer: c

Explanation: The semantics to derive a method for constructing bayesian networks were led to the consequence that a node can be conditionally independent of its predecessors.

1. A _____ is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.

- a) Decision tree
- b) Graphs
- c) Trees
- d) Neural Networks

[View Answer](#)

Answer: a

Explanation: Refer the definition of Decision tree.

2. Decision Tree is a display of an algorithm.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: None.

3. What is Decision Tree?

- a) Flow-Chart
- b) Structure in which internal node represents test on an attribute, each branch represents outcome of test and each leaf node represents class label
- c) Flow-Chart & Structure in which internal node represents test on an attribute, each branch represents outcome of test and each leaf node represents class label
- d) None of the mentioned

[View Answer](#)

Answer: c

Explanation: Refer the definition of Decision tree.

4. Decision Trees can be used for Classification Tasks.

- a) True
- b) False

[View Answer](#)

Answer: a

Explanation: None.

5. Choose from the following that are Decision Tree nodes?

- a) Decision Nodes
- b) End Nodes
- c) Chance Nodes

- d) All of the mentioned
[View Answer](#)

Answer: d
Explanation: None.

6. Decision Nodes are represented by _____
a) Disks
b) Squares
c) Circles
d) Triangles
[View Answer](#)

Answer: b
Explanation: None.

7. Chance Nodes are represented by _____
a) Disks
b) Squares
c) Circles
d) Triangles
[View Answer](#)

Answer: c
Explanation: None.

8. End Nodes are represented by _____
a) Disks
b) Squares
c) Circles
d) Triangles
[View Answer](#)

Answer: d
Explanation: None.

9. Which of the following are the advantage/s of Decision Trees?
a) Possible Scenarios can be added
b) Use a white box model, If given result is provided by a model
c) Worst, best and expected values can be determined for different scenarios
d) All of the mentioned
[View Answer](#)

Answer: d
Explanation: None

1. What is true about Machine Learning?

- A. Machine Learning (ML) is that field of computer science
B. ML is a type of artificial intelligence that extract patterns out of raw data by using an algorithm or method.
C. The main focus of ML is to allow computer systems learn from experience without being explicitly programmed or human intervention.
D. All of the above

[View Answer](#)

Ans : D

Explanation: All statement are true about Machine Learning.

2. ML is a field of AI consisting of learning algorithms that?

- A. Improve their performance
B. At executing some task

- C. Over time with experience
- D. All of the above

[View Answer](#)

Ans : D

Explanation: ML is a field of AI consisting of learning algorithms that : Improve their performance (P), At executing some task (T), Over time with experience (E).

3. $p \rightarrow 0q$ is not a?

- A. hack clause
- B. horn clause
- C. structural clause
- D. system clause

[View Answer](#)

Ans : B

Explanation: $p \rightarrow 0q$ is not a horn clause.

4. The action _____ of a robot arm specify to Place block A on block B.

- A. STACK(A,B)
- B. LIST(A,B)
- C. QUEUE(A,B)
- D. ARRAY(A,B)

[View Answer](#)

Ans : A

Explanation: The action 'STACK(A,B)' of a robot arm specify to Place block A on block B.

5. A_____ begins by hypothesizing a sentence (the symbol S) and successively predicting lower level constituents until individual preterminal symbols are written.

- A. bottow-up parser
- B. top parser
- C. top-down parser
- D. bottom parser

[View Answer](#)

Ans : C

Explanation: A top-down parser begins by hypothesizing a sentence (the symbol S) and successively predicting lower level constituents until individual preterminal symbols are written.

6. A model of language consists of the categories which does not include _____.

- A. System Unit
- B. structural units.
- C. data units
- D. empirical units

[View Answer](#)

Ans : B

Explanation: A model of language consists of the categories which does not include structural units.

7. Different learning methods does not include?

- A. Introduction
- B. Analogy
- C. Deduction
- D. Memorization

[View Answer](#)

Ans : A

Explanation: Different learning methods does not include the introduction.

8. The model will be trained with data in one single batch is known as ?

- A. Batch learning
- B. Offline learning
- C. Both A and B
- D. None of the above

[View Answer](#)

Ans : C

Explanation: we have end-to-end Machine Learning systems in which we need to train the model in one go by using whole available training data. Such kind of learning method or algorithm is called Batch or Offline learning.

9. Which of the following are ML methods?

- A. based on human supervision
- B. supervised Learning
- C. semi-reinforcement Learning
- D. All of the above

[View Answer](#)

Ans : A

Explanation: The following are various ML methods based on some broad categories : Based on human supervision, Unsupervised Learning, Semi-supervised Learning and Reinforcement Learning

10. In Model based learning methods, an iterative process takes place on the ML models that are built based on various model parameters, called ?

- A. mini-batches
- B. optimizedparameters
- C. hyperparameters
- D. superparameters

[View Answer](#)

Ans : C

Explanation: In Model based learning methods, an iterative process takes place on the ML models that are built based on various model parameters, called hyperparameters.

Clustering VS Classification

MCQ

- What is the relation between the distance between clusters and the corresponding class discriminability?
 - proportional
 - inversely-proportional
 - no-relation

Ans: (a)

- To measure the density at a point, consider
 - sphere of any size
 - sphere of unit volume
 - hyper-cube of unit volume
 - both (b) and (c)

Ans: (d)

- Agglomerative clustering falls under which type of clustering method?
 - partition
 - hierarchical
 - none of the above

Ans: (b)

- Indicate which is/are a method of clustering
 - linkage method
 - split and merge
 - both a and b
 - neither a nor b

Ans: (c)

- K means and K-medoids are example of which type of clustering method?
 - Hierarchical
 - partition
 - probabilistic
 - None of the above.

Ans: (b)

6. Unsupervised classification can be termed as
- a. distance measurement
 - b. dimensionality reduction
 - c. clustering
 - d. none of the above

Ans: (d)

7. Indicate which one is a method of density estimation
- a. Histogram based
 - b. Branch and bound procedure
 - c. Neighborhood distance
 - d. all of the above

Ans: (c)

Linear Algebra

MCQ

1. Which of the properties are true for matrix multiplication
 - a. Distributive
 - b. Commutative
 - c. both a and b
 - d. neither a nor b

Ans: (a)

2. Which of the operations can be valid with two matrices of different sizes?
 - a. addition
 - b. subtraction
 - c. multiplication
 - d. Division

Ans: (c)

3. Which of the following statements are true?
 - a. $\text{trace}(A)=\text{trace}(A')$
 - b. $\det(A)=\det(A')$
 - c. both a and b
 - d. neither a nor b

Ans: (c)

4. Which property ensures that inverse of a matrix exists?
 - a. determinant is non-zero
 - b. determinant is zero
 - c. matrix is square
 - d. trace of matrix is positive value.

Ans: (a)

5. Identify the correct order of general to specific matrix?
 - a. square->identity->symmetric->diagonal
 - b. symmetric->diagonal->square->Identity
 - c. square->diagonal->Identity->symmetric

- d. square->symmetric->diagonal->identity

Ans: (d)

6. Which of the statements are true?

- a. If A is a symmetric matrix, $\text{inv}(A)$ is also symmetric
- b. $\det(\text{inv}(A)) = 1/\det(A)$
- c. If A and B are invertible matrices, AB is an invertible matrix too.
- d. all of the above

Ans: (d)

7. Which of the following options hold true?

- a. $\text{inv}(\text{inv}(A)) = A$
- b. $\text{inv}(kA) = \text{inv}(A)/k$
- c. $\text{inv}(A') = \text{inv}(A)'$
- d. all of the above

Ans: (d)

Eigenvalues and Eigenvectors

MCQ

1. The Eigenvalues of a matrix $\begin{bmatrix} 2 & 7 \\ -1 & -6 \end{bmatrix}$ are

- a. 3 and 0
- b. -2 and 7
- c. -5 and 1
- d. 3 and -5

Ans: (c)

2. The Eigenvalues of $\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$ are

- a. -1, 1 and 2
- b. 1, 1 and -2
- c. -1, -1 and 2
- d. 1, 1 and 2

Ans: (c)

3. The Eigenvectors of $\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$ are

- a. (1 1 1), (1 0 1) and (1 1 0)
- b. (1 1 -1), (1 0 -1) and (1 1 0)
- c. (-1 1 -1), (1 0 1) and (1 1 0)
- d. (1 1 1), (-1 0 1) and (-1 1 0)

Ans: (d)

4. Indicate which of the statements are true?

- a. A and A*A have same Eigenvectors
- b. If m is an Eigenvalue of A, then m^2 is an Eigenvalue of A*A.
- c. both a and b
- d. neither a nor b

Ans: (c)

5. Indicate which of the statements are true?

- a. If m is an Eigenvalue of A, then m is an Eigenvalue of A'

- b. If m is an Eigenvalue of A , then $1/m$ is the Eigenvalue of $\text{inv}(A)$
- c. both a and b
- d. neither a nor b

Ans: (c)

6. Indicate which of the statements are true?

- a. A singular matrix must have a zero Eigenvalue
- b. A singular matrix must have a negative Eigenvalue
- c. A singular matrix must have a complex Eigenvalue
- d. (d) All of the above

Ans: (a)

Vector Spaces

MCQ

1. Which of these is a vector space?

- a. $\{(x \ y \ z \ w)' \in R^4 | x + y - z + w = 0\}$
- b. $\{(x \ y \ z)' \in R^3 | x + y + z = 0\}$
- c. $\{(x \ y \ z)' \in R^3 | x^2 + y^2 + z^2 = 1\}$
- d. $\left\{ \begin{pmatrix} a & 1 \\ b & c \end{pmatrix} | a, b, c \in R \right\}$

Ans: (a)

2. Under which of the following operations $\{(x, y) | x, y \in R\}$ is a vector space?

- a. $(x_1, y_1) + (x_2, y_2) = (x_1 + x_2, y_1 + y_2)$ and $r.(x, y) = (rx, y)$
- b. $(x_1, y_1) + (x_2, y_2) = (x_1 + x_2, y_1 + y_2)$ and $r.(x, y) = (rx, 0)$
- c. both a and b
- d. neither a nor b

Ans: (d)

3. Which of the following statements are true?

- a. $r \cdot \vec{v} = \vec{0}$, if and only if $r=0$
- b. $r_1 \cdot \vec{v} = r_2 \cdot \vec{v}$, if and only if $r_1 = r_2$
- c. set of all matrices under usual operations is not a vector space
- d. all of the above

Ans: (d)

4. What is the dimension of the subspace $H = \left\{ \begin{bmatrix} a - 3b + 6c \\ 5a + 4d \\ b - 2c - d \\ 5d \end{bmatrix} : a, b, c, d \in R \right\}$

- a. 1
- b. 2
- c. 3
- d. 4

Ans: (c)

5. What is the rank of the matrix $\begin{bmatrix} 2 & -1 & 1 & -6 & 8 \\ 1 & -2 & -4 & 3 & -2 \\ -7 & 8 & 10 & 3 & -10 \\ 4 & -5 & -7 & 0 & 4 \end{bmatrix}$

- a. 2
- b. 3
- c. 4
- d. 5

Ans: (a)

6. If v_1, v_2, v_3, v_4 are in R^4 and v_3 is not a linear combination of v_1, v_2, v_4 , then $\{v_1, v_2, v_3, v_4\}$ must be linearly independent.

- a. True
- b. False

Ans: (b). For example, if $v_4 = v_1 + v_2$, then $1v_1 + 1v_2 + 0v_3 - 1v_4 = 0$.

7. The vectors $x_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$, $x_2 = \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}$, $x_3 = \begin{pmatrix} 3 \\ 1 \\ 4 \end{pmatrix}$ are :
- a. Linearly dependent
 - b. Linearly independent

Ans: (a). Because $2x_1 + x_2 - x_3 = 0$.

8. The vectors $x_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, $x_2 = \begin{pmatrix} -5 \\ 3 \end{pmatrix}$ are :
- a. Linearly dependent
 - b. Linearly independent

Ans: (b).

Rank and SVD

MCQ

1. The number of non-zero rows in an echelon form is called?

- a. reduced echelon form
- b. rank of a matrix
- c. conjugate of the matrix
- d. cofactor of the matrix

Ans: (b)

2. Let A and B be arbitrary $m \times n$ matrices. Then which one of the following statement is true

- a. $\text{rank}(A + B) \leq \text{rank}(A) + \text{rank}(B)$
- b. $\text{rank}(A + B) < \text{rank}(A) + \text{rank}(B)$
- c. $\text{rank}(A + B) \geq \text{rank}(A) + \text{rank}(B)$
- d. $\text{rank}(A + B) > \text{rank}(A) + \text{rank}(B)$

Ans: (a)

3. The rank of the matrix $\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ is

- a. 0
- b. 2
- c. 1
- d. 3

Ans: (a)

4. The rank of $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ is

- a. 3
- b. 2
- c. 1
- d. 0

Ans: (c)

5. Consider the following two statements:

- I. The maximum number of linearly independent column vectors of a matrix A is called the rank of A.
- II. If A is an $n \times n$ square matrix, it will be nonsingular if $\text{rank } A = n$.

With reference to the above statements, which of the following applies?

- a. Both the statements are false
- b. Both the statements are true
- c. I is true but II is false.
- d. I is false but II is true

Ans: (b)

6. The rank of a 3×3 matrix C ($= AB$), found by multiplying a non-zero column matrix A of size 3×1 and a non-zero row matrix B of size 1×3 , is

- a. 0
- b. 1
- c. 2
- d. 3

Ans: (b)

7. Find the singular values of the matrix $B = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$

- a. 2 and 4
- b. 3 and 4
- c. 2 and 3
- d. 3 and 1

Ans: (d)

8. “Grahm-Schmidt” Process involves factorizing a matrix as a multiplication of two matrices

- a. One is Orthogonal and the other one is upper-triangular
- b. Both are symmetric
- c. One is symmetric and the other one is anti-symmetric
- d. One is diagonal and the other one is symmetric

Ans: (a)

9. SVD is defined as $A = U \Sigma V^T$ where U consists of Eigenvectors of

- a. AA^T
- b. $A^T A$
- c. AA^{-1}
- d. A^*A

Ans: (a)

10. SVD is defined as $A = U \Sigma V^T$, where Σ is :

- a. diagonal matrix having singular values
- b. diagonal matrix having arbitrary values
- c. identity matrix
- d. non diagonal matrix

Ans: (a)

Normal Distribution and Decision Boundary I

MCQ

1. Three components of Bayes decision rule are class prior, likelihood and ...
 - a. Evidence
 - b. Instance
 - c. Confidence
 - d. Salience

Ans: (a)

2. Gaussian function is also called ... function
 - a. Bell
 - b. Signum
 - c. Fixed Point
 - d. Quintic

Ans: (a)

3. The span of the Gaussian curve is determined by the of the distribution
 - a. Mean
 - b. Mode
 - c. Median
 - d. Variance

Ans: (d)

4. When the value of the data is equal to the mean of the distribution in which it belongs to, the Gaussian function attains ... value
 - a. Minimum
 - b. Maximum
 - c. Zero
 - d. None of the above

Ans: (b)

5. The full width of the Gaussian function at half the maximum is
 - a. 2.35σ
 - b. 1.5σ
 - c. 0.5σ
 - d. 0.355σ

Ans: (a)

6. Property of correlation coefficient is

- a. $-1 \leq \rho_{xy} \leq 1$
- b. $-0.5 \leq \rho_{xy} \leq 1$
- c. $-1 \leq \rho_{xy} \leq 1.5$
- d. $-0.5 \leq \rho_{xy} \leq 0.5$

Ans: (a)

7. The correlation coefficient can be viewed as ... angle between two vectors in \mathbb{R}^D

- a. Sin
- b. Cos
- c. Tan
- d. Sec

Ans: (b)

8. For a n-dimensional data, number of correlation coefficient is equal to

- a. nC_2
- b. $n-1$
- c. n^2
- d. $\log(n)$

Ans: (a)

9. Iso-contour lines of smaller radius depicts value of the density function

- a. Higher
- b. Lower
- c. Equal
- d. None of the above

Ans: (a)

Normal Distribution and Decision Boundary II

MCQ

1. If the covariance matrix is strictly diagonal with equal variance then the iso-contour lines (data scatter) of the data resembles
 - a. Concentric circle
 - b. Ellipse
 - c. Oriented Ellipse
 - d. None of the above

Ans: (a)

2. Nature of the decision boundary is determined by
 - a. Decision Rule
 - b. Decision boundary
 - c. Discriminant function
 - d. None of the above

Ans: (c)

3. In Supervised learning, class labels of the training samples are
 - a. Known
 - b. Unknown
 - c. Doesn't matter
 - d. Partially known

Ans: (a)

4. In learning is online then it is called
 - a. Supervised
 - b. Unsupervised
 - c. Semi-supervised
 - d. None of the above

Ans: (b)

5. In supervised learning, the process of learning is
 - a. Online
 - b. Offline
 - c. Partially online and offline
 - d. Doesn't matter

Ans: (b)

6. For spiral data the decision boundary will be
- Linear
 - Non-linear
 - Does not exist

Ans: (b)

7. In a 2-class problem, if the discriminant function satisfies $g_1(x) = g_2(x)$ then, the data point lies
- On the DB
 - Class 1's side
 - Class 2's side
 - None of the above

Ans: (a)

Bayes Theorem

MCQ

1. $P(\vec{X})P(w_i|\vec{X}) =$
 - a. $P(1 - \vec{X})P(w_i|\vec{X})$
 - b. $P(\vec{X})P(1 - w_i|\vec{X})$
 - c. $P(\vec{X}|w_i)P(w_i)$
 - d. $P(\vec{X} - w_i)P(w_i|\vec{X})$

Ans: (c)

2. In Bayes Theorem, unconditional probability is called as
 - a. Evidence
 - b. Likelihood
 - c. Prior
 - d. Posterior

Ans: (a)

3. In Bayes Theorem, Class conditional probability is called as
 - a. Evidence
 - b. Likelihood
 - c. Prior
 - d. Posterior

Ans: (b)

4. When the covariance term in Mahalobian distance becomes Identity then the distance is similar to
 - a. Euclidean distance
 - b. Manhattan distance
 - c. City block distance
 - d. Geodesic distance

Ans: (a)

5. The decision boundary for an N-dimensional ($N > 3$) data will be a
 - a. Point
 - b. Line
 - c. Plane
 - d. Hyperplane

Ans: (d)

6. Bayes error is the bound of probability of classification error.
- Lower
 - Upper

Ans: (a)

7. Bayes decision rule is the theoretically classifier that minimize probability of classification error.
- Best
 - Worst
 - Average

Ans: (a)

Linear Discriminant Function and Perceptron Learning

MCQ

1. A perceptron is:
 - a. a single McCulloch-Pitts neuron
 - b. an autoassociative neural network
 - c. a double layer autoassociative neural network
 - d. All the above

Ans: (a)

2. Perceptron is used as a classifier for
 - a. Linearly separable data
 - b. Non-linearly separable data
 - c. Linearly non-separable data
 - d. Any data

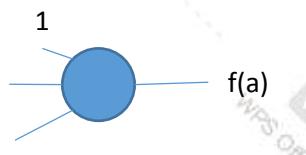
Ans: (a)

3. A 4-input neuron has weights 1, 2, 3 and 4. The transfer function is linear with the constant of proportionality being equal to 2. The inputs are 4, 10, 5 and 20 respectively. The output will be:

- a. 238
- b. 76
- c. 119
- d. 178

Ans: (a)

4. Consider a perceptron for which training sample, $u \in R^2$ and

$$f(a) = \begin{cases} 1 & \text{for } a > 0 \\ 0 & \text{for } a = 0 \\ -1 & \text{for } a < 0 \end{cases}$$


Let the desired output (y) be 1 when elements of class A = {(1,2),(2,4),(3,3),(4,4)} is applied as input and let it be -1 for the class B = {(0,0),(2,3),(3,0),(4,2)}. Let the initial connection weights $w_0(0) = +1$, $w_1(0) = -2$, $w_2(0) = +1$ and learning rate be $\eta = 0.5$.

This perceptron is to be trained by perceptron convergence procedure, for which the weight update formula is $(t + 1) = w(t) + \eta(y - f(a))u$, where $f(a)$ is the actual output.

A. If $u = (4,4)$ is applied as input, then $w(1) = ?$

- a. $[2,2,5]^T$
- b. $[2,1,5]^T$
- c. $[2,1,1]^T$
- d. $[2,0,5]^T$

Ans: (a)

B. If $(4,2)$ is then applied, what will be $w(2)$

- a. $[1,-2,3]^T$
- b. $[-1,-2,3]^T$
- c. $[1,-2,-3]^T$
- d. $[1,2,3]^T$

Ans: (a)

5. Perceptron training rule converges, if data is

- a. Linearly separable
- b. Non-linearly separable
- c. Linearly non-separable data
- d. Any data

Ans: (a)

6. Is XOR problem solvable using a single perceptron

- a. Yes
- b. No
- c. Can't say

Ans: (b)

7. Consider a perceptron for which training sample, $u \in R^2$ and actual output, $x \in \{0,1\}$, let the desired output be 0 when elements of class A= $\{(2,4),(3,2),(3,4)\}$ is applied as input and let it be 1 for the class B= $\{(1,0),(1,2),(2,1)\}$. Let the learning rate η be 0.5 and initial connection weights are $w_0=0$, $w_1=1$, $w_2=1$. Answer the following questions:

A. Shall the perceptron convergence procedure terminate if the input patterns from class A and B are repeatedly applied by choosing a very small learning rate?

- a. Yes
- b. No
- c. Can't say

Ans: (a). Since Classes are linearly separable.

- B. Now add sample (5,2) to class B, what is your answer now, i.e. will it converge or not?
 - a. Yes
 - b. No
 - c. Can't say

Ans: (b). After adding above sample, classes become non linear separable.

Linear and Non-Linear Decision Boundaries

MCQ

1. Decision Boundary in case of same covariance matrix, with identical diagonal elements is :
 - a. Linear
 - b. Non-Linear
 - c. None of the above

Ans: (a)

2. Decision Boundary in case of diagonal covariance matrix, with identical diagonal elements is given by $W^T(X - X_0) = 0$, where W is given by:
 - a. $(\mu_k - \mu_l)/\sigma^2$
 - b. $(\mu_k + \mu_l)/\sigma^2$
 - c. $(\mu_k^2 + \mu_l^2)/\sigma^2$
 - d. $(\mu_k + \mu_l)/\sigma$

Ans: (a)

3. Decision Boundary in case of arbitrary covariance matrix but identical for all class is :
 - a. Linear
 - b. Non-Linear
 - c. None of the above

Ans: (a)

4. Decision Boundary in case of arbitrary covariance matrix but identical for all class is given by $W^T(X - X_0) = 0$, where W is given by:
 - a. $(\mu_k - \mu_l)/\sigma^2$
 - b. $\Sigma^{-1}(\mu_k - \mu_l)$
 - c. $(\mu_k^2 + \mu_l^2)/\sigma^2$
 - d. $\Sigma^{-1}(\mu_k^2 - \mu_l^2)$

Ans: (b)

5. Decision Boundary in case of arbitrary covariance matrix and also unequal is :
 - a. Linear
 - b. Non-Linear
 - c. None of the above

Ans: (b)

6. Discriminant function in case of arbitrary covariance matrix and all parameters are class dependent is given by $(X^T W_i X + w_i^T X + w_{io}) = 0$, where W is given by:

- a. $-\frac{1}{2} \Sigma_i^{-1}$
- b. $\Sigma_i^{-1} \mu_i$
- c. $-\frac{1}{2} \Sigma_i^{-1} \mu_i$
- d. $-\frac{1}{4} \Sigma_i^{-1}$

Ans: (a)

PCA

MCQ

1. The tool used to obtain a PCA is
 - a. LU Decomposition
 - b. QR Decomposition
 - c. SVD
 - d. Cholesky Decomposition

Ans: (c)

2. PCA is used for
 - a. Dimensionality Enhancement
 - b. Dimensionality Reduction
 - c. Both
 - d. None

Ans: (b)

3. The scatter matrix of the transformed feature vector is given by
 - a. $\sum_{k=1}^N (x_k - \mu)(x_k - \mu)^T$
 - b. $\sum_{k=1}^N (x_k - \mu)^T (x_k - \mu)$
 - c. $\sum_{k=1}^N (\mu - x_k)(\mu - x_k)^T$
 - d. $\sum_{k=1}^N (\mu - x_k)^T (\mu - x_k)$

Ans: (a)

4. PCA is used for
 - a. Supervised Classification
 - b. Unsupervised Classification
 - c. Semi-supervised Classification
 - d. Cannot be used for classification

Ans: (b)

5. The vectors which correspond to the vanishing singular values of a matrix that span the null space of the matrix are:
- Right singular vectors
 - Left singular vectors
 - All the singular vectors
 - None

Ans: (a)

6. If S is the scatter of the data in the original domain, then the scatter of the transformed feature vectors is given by
- S^T
 - S
 - WSW^T
 - $W^T SW$

Ans: (d)

7. The largest Eigen vector gives the direction of the
- Maximum scatter of the data
 - Minimum scatter of the data
 - No such information can be interpreted
 - Second largest Eigen vector which is in the same direction.

Ans: (a)

8. The following linear transform does not have a fixed set of basis vectors:
- DCT
 - DFT
 - DWT
 - PCA

Ans: (d)

9. The Within Class scatter matrix is given by:
- $\sum_{i=1}^c \sum_{k=1}^N (x_k - \mu_i)(x_k - \mu_i)^T$
 - $\sum_{i=1}^c \sum_{k=1}^N (x_k - \mu_i)^T (x_k - \mu_i)$
 - $\sum_{i=1}^c \sum_{k=1}^N (x_i - \mu_k)(x_i - \mu_k)^T$
 - $\sum_{i=1}^c \sum_{k=1}^N (x_i - \mu_k)^T (x_i - \mu_k)$

Ans: (a)

10. The Between Class scatter matrix is given by:

- a. $\sum_{i=1}^c N_i(\mu_i - \mu)(\mu_i - \mu)^T$
- b. $\sum_{i=1}^c N_i(\mu_i - \mu)^T(\mu_i - \mu)$
- c. $\sum_{i=1}^c N_i(\mu - \mu_i)(\mu - \mu_i)^T$
- d. $\sum_{i=1}^c N_i(\mu - \mu_i)^T(\mu - \mu_i)$

Ans: (a)

11. Which of the following is unsupervised technique?

- a. PCA
- b. LDA
- c. Bayes
- d. None of the above

Ans: (a)

Linear Discriminant Analysis

MCQ

1. Linear Discriminant Analysis is
 - a. Unsupervised Learning
 - b. Supervised Learning
 - c. Semi-supervised Learning
 - d. None of the above

Ans: (b)

2. The following property of a within-class scatter matrix is a must for LDA:
 - a. Singular
 - b. Non-singular
 - c. Does not matter
 - d. Problem-specific

Ans: (b)

3. In Supervised learning, class labels of the training samples are
 - a. Known
 - b. Unknown
 - c. Doesn't matter
 - d. Partially known

Ans: (a)

4. The upper bound of the number of non-zero Eigenvalues of $S_w^{-1}S_B$ ($C = \text{No. of Classes}$)
 - a. $C - 1$
 - b. $C + 1$
 - c. C
 - d. None of the above

Ans: (a)

5. If S_w is singular and $N < D$, its rank is at most (N is total number of samples, D dimension of data, C is number of classes)
 - a. $N + C$
 - b. N
 - c. C
 - d. $N - C$

Ans: (d)

6. If S_w is singular and $N < D$ the alternative solution is to use (N is total number of samples, D dimension of data)

- a. EM
- b. PCA
- c. ML
- d. Any one of the above

Ans: (b)

GMM

MCQ

1. A method to estimate the parameters of a distribution is
 - a. Maximum Likelihood
 - b. Linear Programming
 - c. Dynamic Programming
 - d. Convex Optimization

Ans: (a)

2. Gaussian mixtures are also known as
 - a. Gaussian multiplication
 - b. Non-linear super-position of Gaussians
 - c. Linear super-position of Gaussians
 - d. None of the above

Ans: (c)

3. The mixture coefficients of the GMM add upto
 - a. 1
 - b. 0
 - c. Any value greater than 0
 - d. Any value less than 0

Ans: (a)

4. The mixture coefficients are
 - a. Strictly positive
 - b. Positive
 - c. Strictly negative
 - d. Negative

Ans: (b)

5. The mixture coefficients can take a value
 - a. Greater than zero
 - b. Greater than 1
 - c. Less than zero
 - d. Between zero and 1

Ans: (d)

6. For Gaussian mixture models, parameters are estimated using a closed form solution by

- a. Expectation Minimization
- b. Expectation Maximization
- c. Maximum Likelihood
- d. None of the above

Ans: (b)

7. Latent Variable in GMM is also known as:

- a. Prior Probability
- b. Posterior Probability
- c. Responsibility
- d. None of the above

Ans: (b,c)

8. A GMM with K Gaussian mixture has K covariance matrices, with dimension:

- a. Arbitrary
- b. K X K
- c. D X D (Dimension of data)
- d. N X N (No of samples in the dataset)

Ans: c

Complete Machine Learning MCQs Unit Wise | SPPU Final Year

October 13, 2020 by Shivam



Machine Learning MCQs UNIT Wise, Final Year SPPU.

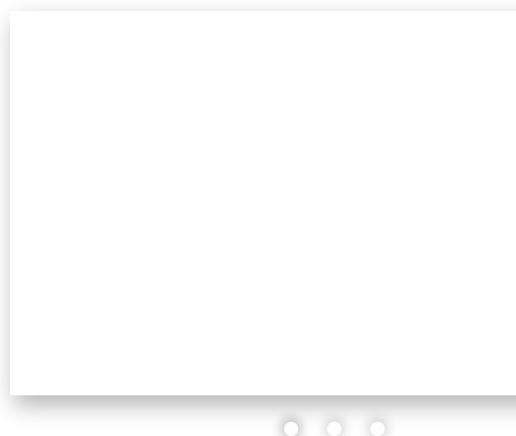
Machine Learning being the most prominent areas of the era finds its place in the curriculum of many universities or institutes, among which is **Savitribai Phule Pune University(SPPU)**.

Machine Learning subject, having subject no.: **410250**, the first compulsory subject of **8th semester** and has 3 credits in the course, according to the new credit system. This subject is the first compulsory subject that includes all the basics of this topic to its

efficient algorithms. If any student develops interest in this subject, going through this course will be a good start.

This subject gives knowledge from the introduction of Machine Learning terminologies and types like supervised, unsupervised, etc. to its various techniques like clustering, classification, etc.

As we know, the syllabus of the upcoming final exams contains only the first four units of this course, so, the below-given MCQs cover the first 4 units of ML subject as:-



Unit 1. Introduction to Machine Learning

Unit 2. Feature Selection

Unit 3. Regression

Unit 4. Naïve Bayes and Support Vector Machine

So, here are the **MCQs on the subject Machine Learning from the course of Computer branch, SPPU**, which will clearly help you out on the upcoming exams.

Machine Learning MCQs UNIT I

1. What is classification?

a) when the output variable is a category, such as “red” or “blue” or “disease” and “no

disease".

- b) when the output variable is a real value, such as "dollars" or "weight".

Ans: Solution A

2. What is regression?

- a) When the output variable is a category, such as "red" or "blue" or "disease" and "no disease".
- b) When the output variable is a real value, such as "dollars" or "weight".

Ans: Solution B

3. What is supervised learning?

- a) All data is unlabelled and the algorithms learn to inherent structure from the input data
- b) All data is labelled and the algorithms learn to predict the output from the input data
- c) It is a framework for learning where an agent interacts with an environment and receives a reward for each interaction
- d) Some data is labelled but most of it is unlabelled and a mixture of supervised and unsupervised techniques can be used.

Ans: Solution B

4. What is Unsupervised learning?

- a) All data is unlabelled and the algorithms learn to inherent structure from the input data
- b) All data is labelled and the algorithms learn to predict the output from the input data
- c) It is a framework for learning where an agent interacts with an environment and

receives a reward for each interaction

- d) Some data is labelled but most of it is unlabelled and a mixture of supervised and unsupervised techniques can be used.

Ans: Solution A

5. What is Semi-Supervised learning?

- a) All data is unlabelled and the algorithms learn to inherent structure from the input data
- b) All data is labelled and the algorithms learn to predict the output from the input data
- c) It is a framework for learning where an agent interacts with an environment and receives a reward for each interaction
- d) Some data is labelled but most of it is unlabelled and a mixture of supervised and unsupervised techniques can be used.

Ans: Solution D

6. What is Reinforcement learning?

- a) All data is unlabelled and the algorithms learn to inherent structure from the input data
- b) All data is labelled and the algorithms learn to predict the output from the input data
- c) It is a framework for learning where an agent interacts with an environment and receives a reward for each interaction
- d) Some data is labelled but most of it is unlabelled and a mixture of supervised and unsupervised techniques can be used.

Ans: Solution C

7. Sentiment Analysis is an example of:

- a)Regression,
- b)Classification
- c)Clustering
- d)Reinforcement Learning

Options:

- A. 1 Only
- B. 1 and 2
- C. 1 and 3
- D. 1, 2 and 4

Ans : Solution D

8. The process of forming general concept definitions from examples of concepts to be learned.

- a) Deduction
- b) abduction
- c) induction
- d) conjunction

Ans : Solution C

9. Computers are best at learning

- a) facts.
- b) concepts.
- c) procedures.
- d) principles.

Ans : Solution A

10. Data used to build a data mining model.

- a) validation data
- b) training data
- c) test data
- d) hidden data

Ans : Solution B

11. Supervised learning and unsupervised clustering both require at least one

- a) hidden attribute.
- b) output attribute.
- c) input attribute.
- d) categorical attribute.

Ans : Solution A

12. Supervised learning differs from unsupervised clustering in that supervised learning requires

- a) at least one input attribute.
- b) input attributes to be categorical.
- c) at least one output attribute.
- d) output attributes to be categorical.

Ans : Solution B

13. A regression model in which more than one independent variable is used to predict the

dependent variable is called

- a) a simple linear regression model
- b) a multiple regression models
- c) an independent model
- d) none of the above

Ans : Solution C

14. A term used to describe the case when the independent variables in a multiple regression model

are correlated is

- a) Regression
- b) correlation
- c) multicollinearity
- d) none of the above

Ans : Solution C

15. A multiple regression model has the form: $y = 2 + 3x_1 + 4x_2$. As x_1 increases by 1 unit (holding x_2 constant), y will
- a) increase by 3 units
 - b) decrease by 3 units
 - c) increase by 4 units
 - d) decrease by 4 units

Ans : Solution C

16. A multiple regression model has
- a) only one independent variable
 - b) more than one dependent variable
 - c) more than one independent variable
 - d) none of the above

Ans : Solution B

17. A measure of goodness of fit for the estimated regression equation is the
- a) multiple coefficient of determination
 - b) mean square due to error
 - c) mean square due to regression
 - d) none of the above

Ans : Solution C

18. The adjusted multiple coefficient of determination accounts for
- a) the number of dependent variables in the model
 - b) the number of independent variables in the model
 - c) unusually large predictors
 - d) none of the above

Ans : Solution D

19. The multiple coefficient of determination is computed by
- a) dividing SSR by SST
 - b) dividing SST by SSR
 - c) dividing SST by SSE
 - d) none of the above

Ans : Solution C

20. For a multiple regression model, $SST = 200$ and $SSE = 50$. The multiple coefficient of determination is

- a) 0.25
- b) 4.00
- c) 0.75
- d) none of the above

Ans : Solution B

21. A nearest neighbor approach is best used

- a) with large-sized datasets.
- b) when irrelevant attributes have been removed from the data.
- c) when a generalized model of the data is desirable.
- d) when an explanation of what has been found is of primary importance.

Ans : Solution B

22. Another name for an output attribute.

- a) predictive variable
- b) independent variable
- c) estimated variable
- d) dependent variable

Ans : Solution B

23. Classification problems are distinguished from estimation problems in that

- a) classification problems require the output attribute to be numeric.

- b) classification problems require the output attribute to be categorical.
- c) classification problems do not allow an output attribute.
- d) classification problems are designed to predict future outcome.

Ans : Solution C

24. Which statement is true about prediction problems?
- a) The output attribute must be categorical.
 - b) The output attribute must be numeric.
 - c) The resultant model is designed to determine future outcomes.
 - d) The resultant model is designed to classify current behavior.

Ans : Solution D

25. Which statement about outliers is true?
- a) Outliers should be identified and removed from a dataset.
 - b) Outliers should be part of the training dataset but should not be present in the test data.
 - c) Outliers should be part of the test dataset but should not be present in the training data.
 - d) The nature of the problem determines how outliers are used.

Ans : Solution D

26. Which statement is true about neural network and linear regression models?
- a) Both models require input attributes to be numeric.
 - b) Both models require numeric attributes to range between 0 and 1.
 - c) The output of both models is a categorical attribute value.
 - d) Both techniques build models whose output is determined by a linear sum of weighted input attribute values.

Ans : Solution A

27. Which of the following is a common use of unsupervised clustering?
- a) detect outliers
 - b) determine a best set of input attributes for supervised learning
 - c) evaluate the likely performance of a supervised learner model
 - d) determine if meaningful relationships can be found in a dataset

Ans : Solution A

28. The average positive difference between computed and desired outcome values.

- a) root mean squared error
- b) mean squared error
- c) mean absolute error
- d) mean positive error

Ans : Solution D

29. Selecting data so as to assure that each class is properly represented in both the training and test set.

- a) cross validation
- b) stratification
- c) verification
- d) bootstrapping

Ans : Solution B

30. The standard error is defined as the square root of this computation.

- a) The sample variance divided by the total number of sample instances.
- b) The population variance divided by the total number of sample instances.
- c) The sample variance divided by the sample mean.
- d) The population variance divided by the sample mean.

Ans : Solution A

31. Data used to optimize the parameter settings of a supervised learner model.

- a) Training

- b) Test
- c) Verification
- d) Validation

Ans : Solution D

32. Bootstrapping allows us to

- a) choose the same training instance several times.
- b) choose the same test set instance several times.
- c) build models with alternative subsets of the training data several times.
- d) test a model with alternative subsets of the test data several times.

Ans : Solution A

33. The correlation between the number of years an employee has worked for a company and the salary of the employee is 0.75. What can be said about employee salary and years worked?

- a) There is no relationship between salary and years worked.
- b) Individuals that have worked for the company the longest have higher salaries.
- c) Individuals that have worked for the company the longest have lower salaries.
- d) The majority of employees have been with the company a long time.
- e) The majority of employees have been with the company a short period of time.

Ans : Solution B

34. The correlation coefficient for two real-valued attributes is -0.85 . What does this value tell you?

- a) The attributes are not linearly related.
- b) As the value of one attribute increases the value of the second attribute also increases.
- c) As the value of one attribute decreases the value of the second attribute increases.
- d) The attributes show a curvilinear relationship.

Ans : Solution C

35. The average squared difference between classifier predicted output and actual output.

- a) mean squared error
- b) root mean squared error
- c) mean absolute error

d) mean relative error

Ans : Solution A

36. Simple regression assumes a _____ relationship between the input attribute and output

attribute.

- a) Linear
- b) Quadratic
- c) reciprocal
- d) inverse

Ans : Solution A

37. Regression trees are often used to model _____ data.

- a) Linear
- b) Nonlinear
- c) Categorical
- d) Symmetrical

Ans : Solution B

38. The leaf nodes of a model tree are

- a) averages of numeric output attribute values.
- b) nonlinear regression equations.
- c) linear regression equations.
- d) sums of numeric output attribute values.

Ans : Solution C

39. Logistic regression is a _____ regression technique that is used to model data having a _____ outcome.

- a) linear, numeric
- b) linear, binary
- c) nonlinear, numeric
- d) nonlinear, binary

Ans : Solution D

40. This technique associates a conditional probability value with each data instance.

- a) linear regression
- b) logistic regression
- c) simple regression
- d) multiple linear regression

Ans : Solution B

41. This supervised learning technique can process both numeric and categorical input attributes.

- a) linear regression
- b) Bayes classifier
- c) logistic regression
- d) backpropagation learning

Ans : Solution A

42. With Bayes classifier, missing data items are

- a) treated as equal compares.
- b) treated as unequal compares.
- c) replaced with a default value.
- d) ignored.

Ans : Solution B

43. This clustering algorithm merges and splits nodes to help modify nonoptimal partitions.

- a) agglomerative clustering
- b) expectation maximization
- c) conceptual clustering

d) K-Means clustering

Ans : Solution D

44. This clustering algorithm initially assumes that each data instance represents a single cluster.

- a) agglomerative clustering
- b) conceptual clustering
- c) K-Means clustering
- d) expectation maximization

Ans : Solution C

45. This unsupervised clustering algorithm terminates when mean values computed for the current iteration of the algorithm are identical to the computed mean values for the previous iteration.

- a) agglomerative clustering
- b) conceptual clustering
- c) K-Means clustering
- d) expectation maximization

Ans : Solution C

46. Machine learning techniques differ from statistical techniques in that machine learning methods

- a) typically assume an underlying distribution for the data.
- b) are better able to deal with missing and noisy data.
- c) are not able to explain their behavior.

d) have trouble with large-sized datasets.

Ans : Solution B

Machine Learning MCQs UNIT -II

1. True- False: Over fitting is more likely when you have huge amount of data to train?

A) TRUE

B) FALSE

Ans Solution: (B)

With a small training dataset, it's easier to find a hypothesis to fit the training data exactly i.e. over fitting.

2. What is pca.components_ in Sklearn?

A) Set of all eigen vectors for the projection space

B) Matrix of principal components

C) Result of the multiplication matrix

D) None of the above options

Ans A

3. Which of the following techniques would perform better for reducing dimensions of a data set?

A. Removing columns which have too many missing values

B. Removing columns which have high variance in data

C. Removing columns with dissimilar data trends

D. None of these

Ans Solution: (A) If a columns have too many missing values, (say 99%) then we can remove such columns.

4. It is not necessary to have a target variable for applying dimensionality reduction algorithms.

A. TRUE

B. FALSE

Ans Solution: (A)

LDA is an example of supervised dimensionality reduction algorithm

5. PCA can be used for projecting and visualizing data in lower dimensions.

- A. TRUE
- B. FALSE

Ans Solution: (A)

Sometimes it is very useful to plot the data in lower dimensions. We can take the first 2 principal components and then visualize the data using scatter plot.

6. The most popularly used dimensionality reduction algorithm is Principal Component Analysis (PCA). Which of the following is/are true about PCA?

- 1.PCA is an unsupervised method
 - 2.It searches for the directions that data have the largest variance
 - 3.Maximum number of principal components \leq number of features
 - 4.All principal components are orthogonal to each other
- A. 1 and 2
 - B. 1 and 3
 - C. 2 and 3
 - D. All of the above

Ans D

7. PCA works better if there is?

- 1.A linear structure in the data
 - 2.If the data lies on a curved surface and not on a flat surface
 - 3.If variables are scaled in the same unit
- A. 1 and 2
 - B. 2 and 3

C. 1 and 3

D. 1,2 and 3

Ans Solution: (C)

8. What happens when you get features in lower dimensions using PCA?

1.The features will still have interpretability

2.The features will lose interpretability

3.The features must carry all information present in data

4.The features may not carry all information present in data

A. 1 and 3

B. 1 and 4

C. 2 and 3

D. 2 and 4

Ans Solution: (D)

When you get the features in lower dimensions then you will lose some information of data most of the times and you won't be able to interpret the lower dimension data.

9. Which of the following option(s) is / are true?

1.You need to initialize parameters in PCA

2.You don't need to initialize parameters in PCA

3.PCA can be trapped into local minima problem

4.PCA can't be trapped into local minima problem

A. 1 and 3

B. 1 and 4

C. 2 and 3

D. 2 and 4

Ans Solution: (D)

PCA is a deterministic algorithm which doesn't have parameters to initialize and it doesn't have local minima problem like most of the machine learning algorithms has.

10. What is of the following statement is true about t-SNE in comparison to PCA?

A. When the data is huge (in size), t-SNE may fail to produce better results.

B. T-SNE always produces better result regardless of the size of the data

C. PCA always performs better than t-SNE for smaller size data.

D. None of these

Ans Solution: (A)

Option A is correct

11. [True or False] PCA can be used for projecting and visualizing data in lower dimensions.

A. TRUE

B. FALSE

Solution: (A)

Sometimes it is very useful to plot the data in lower dimensions. We can take the first 2 principal components and then visualize the data using scatter plot.

12. A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college.

1) Which of the following statement is true in following case?

A) Feature F1 is an example of nominal variable.

B) Feature F1 is an example of ordinal variable.

C) It doesn't belong to any of the above category.

D) Both of these

Solution: (B)

Ordinal variables are the variables which has some order in their categories. For example, grade A should be consider as high grade than grade B.

13. Which of the following is an example of a deterministic algorithm?

A) PCA

B) K-Means

C) None of the above

Solution: (A)

A deterministic algorithm is that in which output does not change on different runs.

PCA would give the same result if we run again, but not k-means

Machine Learning MCQs UNIT –III

1. Which of the following methods do we use to best fit the data in Logistic Regression?

- A) Least Square Error
- B) Maximum Likelihood
- C) Jaccard distance
- D) Both A and B

Ans Solution: B

2. Choose which of the following options is true regarding One-Vs-All method in Logistic

Regression.

- A) We need to fit n models in n-class classification problem
- B) We need to fit n-1 models to classify into n classes
- C) We need to fit only 1 model to classify into n classes
- D) None of these

Ans Solution: A

3. Suppose, You applied a Logistic Regression model on a given data and got a training accuracy X and testing accuracy Y. Now, you want to add a few new features in the same data. Select the option(s) which is/are correct in such a case.

Note: Consider remaining parameters are same.

- A) Training accuracy increases
- B) Training accuracy increases or remains the same
- C) Testing accuracy decreases
- D) Testing accuracy increases or remains the same

Ans Solution: A and D

Adding more features to model will increase the training accuracy because model has to

consider more data to fit the logistic regression. But testing accuracy increases if feature is found to be significant

4. Which of the following algorithms do we use for Variable Selection?

- A) LASSO
- B) Ridge
- C) Both
- D) None of these

Ans Solution: A

In case of lasso we apply a absolute penalty, after increasing the penalty in lasso some of the coefficient of variables may become zero

5. Which of the following statement is true about outliers in Linear regression?

- A) Linear regression is sensitive to outliers
- B) Linear regression is not sensitive to outliers
- C) Can't say
- D) None of these

Ans Solution: (A)

The slope of the regression line will change due to outliers in most of the cases. So

Linear

Regression is sensitive to outliers.

6. Which of the following methods do we use to find the best fit line for data in Linear Regression?

- A) Least Square Error
- B) Maximum Likelihood
- C) Logarithmic Loss
- D) Both A and B

Ans Solution: (A)

In linear regression, we try to minimize the least square errors of the model to identify the line of best fit.

7. Which of the following is true about Residuals?

- A) Lower is better
- B) Higher is better
- C) A or B depend on the situation
- D) None of these

Ans Solution: (A)

Residuals refer to the error values of the model. Therefore lower residuals are desired.

8. Suppose you plotted a scatter plot between the residuals and predicted values in linear

regression and you found that there is a relationship between them. Which of the following

conclusion do you make about this situation?

- A) Since the there is a relationship means our model is not good
- B) Since the there is a relationship means our model is good
- C) Can't say
- D) None of these

Ans Solution: (A)

There should not be any relationship between predicted values and residuals. If there exists any relationship between them, it means that the model has not perfectly captured the information in the data.

9. Suppose you have fitted a complex regression model on a dataset. Now, you are using Ridge regression with penalty x .

Choose the option which describes bias in best manner.

- A) In case of very large x ; bias is low
- B) In case of very large x ; bias is high
- C) We can't say about bias
- D) None of these

Ans Solution: (B)

If the penalty is very large it means model is less complex, therefore the bias would be high.

10. Which of the following option is true?

- A) Linear Regression errors values has to be normally distributed but in case of Logistic Regression it is not the case
- B) Logistic Regression errors values has to be normally distributed but in case of Linear Regression it is not the case
- C) Both Linear Regression and Logistic Regression error values have to be normally distributed
- D) Both Linear Regression and Logistic Regression error values have not to be normally distributed

Ans Solution: A

11. Suppose you have trained a logistic regression classifier and it outputs a new example x with

a prediction $h_0(x) = 0.2$. This means

Our estimate for $P(y=1 | x)$

Our estimate for $P(y=0 | x)$

Our estimate for $P(y=1 | x)$

Our estimate for $P(y=0 | x)$

Ans Solution: B

12. True-False: Linear Regression is a supervised machine learning algorithm.

A) TRUE

B) FALSE

Solution: (A)

Yes, Linear regression is a supervised learning algorithm because it uses true labels for training. Supervised learning algorithm should have input variable (x) and an output variable (y) for each example

13. True-False: Linear Regression is mainly used for Regression.

A) TRUE

B) FALSE

Solution: (A)

Linear Regression has dependent variables that have continuous values.

14. True-False: It is possible to design a Linear regression algorithm using a neural network?

A) TRUE

B) FALSE

Solution: (A)

True. A Neural network can be used as a universal approximator, so it can definitely implement a linear regression algorithm.

15. Which of the following methods do we use to find the best fit line for data in Linear Regression?

- A) Least Square Error
- B) Maximum Likelihood
- C) Logarithmic Loss
- D) Both A and B

Solution: (A)

In linear regression, we try to minimize the least square errors of the model to identify the line of best fit.

16. Which of the following evaluation metrics can be used to evaluate a model while modeling a continuous output variable?

- A) AUC-ROC
- B) Accuracy
- C) Logloss
- D) Mean-Squared-Error

Solution: (D)

Since linear regression gives output as continuous values, so in such case we use mean squared error metric to evaluate the model performance. Remaining options are use in case of a classification problem.

17. True-False: Lasso Regularization can be used for variable selection in Linear Regression.

- A) TRUE
- B) FALSE

Solution: (A)

True, In case of lasso regression we apply absolute penalty which makes some of the coefficients zero.

18. Which of the following is true about Residuals ?

- A) Lower is better
- B) Higher is better
- C) A or B depend on the situation
- D) None of these

Solution: (A)

Residuals refer to the error values of the model. Therefore lower residuals are desired.

19. Suppose that we have N independent variables (X_1, X_2, \dots, X_n) and dependent variable is Y. Now Imagine that you are applying linear regression by fitting the best fit line using least square error on this data. You found that correlation coefficient for one of its variable (Say X_1) with Y is -0.95.

Which of the following is true for X_1 ?

- A) Relation between the X_1 and Y is weak
- B) Relation between the X_1 and Y is strong
- C) Relation between the X_1 and Y is neutral
- D) Correlation can't judge the relationship

Solution: (B)

The absolute value of the correlation coefficient denotes the strength of the relationship.

Since absolute correlation is very high it means that the relationship is strong between X_1 and Y.

20. Looking at above two characteristics, which of the following option is the correct for

Pearson correlation between V1 and V2? If you are given the two variables V1 and V2 and they are following below two characteristics.

- 1. If V1 increases then V2 also increases
- 2. If V1 decreases then V2 behavior is unknown

- A) Pearson correlation will be close to 1
- B) Pearson correlation will be close to -1
- C) Pearson correlation will be close to 0
- D) None of these

Solution: (D)

We cannot comment on the correlation coefficient by using only statement 1. We need to consider the both of these two statements. Consider V1 as x and V2 as $|x|$. The correlation coefficient would not be close to 1 in such a case.

21. Suppose Pearson correlation between V1 and V2 is zero. In such case, is it right to conclude that V1 and V2 do not have any relation between them?

- A) TRUE
- B) FALSE

Solution: (B)

Pearson correlation coefficient between 2 variables might be zero even when they have a relationship between them. If the correlation coefficient is zero, it just means that that they don't move together. We can take examples like $y=|x|$ or $y=x^2$.

22. True- False: Overfitting is more likely when you have huge amount of data to train?

- A) TRUE
- B) FALSE

Solution: (B)

With a small training dataset, it's easier to find a hypothesis to fit the training data exactly i.e. overfitting.

23. We can also compute the coefficient of linear regression with the help of an analytical method called "Normal Equation". Which of the following is/are true about Normal Equation?

- 1. We don't have to choose the learning rate
 - 2. It becomes slow when number of features is very large
 - 3. There is no need to iterate
- A) 1 and 2
 - B) 1 and 3

- C) 2 and 3
- D) 1,2 and 3

Solution: (D)

Instead of gradient descent, Normal Equation can also be used to find coefficients.

Question Context 24-26:

Suppose you have fitted a complex regression model on a dataset. Now, you are using Ridge regression with penalty x .

24. Choose the option which describes bias in best manner.

- A) In case of very large x ; bias is low
- B) In case of very large x ; bias is high
- C) We can't say about bias
- D) None of these

Solution: (B)

If the penalty is very large it means model is less complex, therefore the bias would be high.

25. What will happen when you apply very large penalty?

- A) Some of the coefficient will become absolute zero
- B) Some of the coefficient will approach zero but not absolute zero
- C) Both A and B depending on the situation
- D) None of these

Solution: (B)

In lasso some of the coefficient value become zero, but in case of Ridge, the coefficients become close to zero but not zero.

26. What will happen when you apply very large penalty in case of Lasso?

A) Some of the coefficient will become zero

B) Some of the coefficient will be approaching to zero but not absolute zero

C) Both A and B depending on the situation

D) None of these

Solution: (A)

As already discussed, lasso applies absolute penalty, so some of the coefficients will become zero.

27. Which of the following statement is true about outliers in Linear regression?

A) Linear regression is sensitive to outliers

B) Linear regression is not sensitive to outliers

C) Can't say

D) None of these

Solution: (A)

The slope of the regression line will change due to outliers in most of the cases. So

Linear

Regression is sensitive to outliers.

28. Suppose you plotted a scatter plot between the residuals and predicted values in linear

regression and you found that there is a relationship between them. Which of the following

conclusion do you make about this situation?

A) Since the there is a relationship means our model is not good

B) Since the there is a relationship means our model is good

C) Can't say

D) None of these

Solution: (A)

There should not be any relationship between predicted values and residuals. If there exists any relationship between them,it means that the model has not perfectly captured the information in the data.

Question Context 29-31:

Suppose that you have a dataset D1 and you design a linear regression model of degree 3

polynomial and you found that the training and testing error is “0” or in another terms it perfectly fits the data.

29. What will happen when you fit degree 4 polynomial in linear regression?

- A) There are high chances that degree 4 polynomial will over fit the data
- B) There are high chances that degree 4 polynomial will under fit the data
- C) Can't say
- D) None of these

Solution: (A)

Since is more degree 4 will be more complex(overfit the data) than the degree 3 model so it will again perfectly fit the data. In such case training error will be zero but test error may not be zero.

30. What will happen when you fit degree 2 polynomial in linear regression?

- A) It is high chances that degree 2 polynomial will over fit the data
- B) It is high chances that degree 2 polynomial will under fit the data
- C) Can't say
- D) None of these

Solution: (B)

If a degree 3 polynomial fits the data perfectly, it's highly likely that a simpler model(degree 2 polynomial) might under fit the data.

31. In terms of bias and variance. Which of the following is true when you fit degree 2 polynomial?

- A) Bias will be high, variance will be high
- B) Bias will be low, variance will be high
- C) Bias will be high, variance will be low
- D) Bias will be low, variance will be low

Solution: (C)

Since a degree 2 polynomial will be less complex as compared to degree 3, the bias will be high and variance will be low

Question Context 32-33:

We have been given a dataset with n records in which we have input attribute as x and output attribute as y. Suppose we use a linear regression method to model this data. To test our linear regressor, we split the data in training set and test set randomly.

32. Now we increase the training set size gradually. As the training set size increases, what do you expect will happen with the mean training error?

- A) Increase

- B) Decrease
- C) Remain constant
- D) Can't Say

Solution: (D)

Training error may increase or decrease depending on the values that are used to fit the model. If the values used to train contain more outliers gradually, then the error might just increase.

33. What do you expect will happen with bias and variance as you increase the size of training data?

- A) Bias increases and Variance increases
- B) Bias decreases and Variance increases
- C) Bias decreases and Variance decreases
- D) Bias increases and Variance decreases
- E) Can't Say False

Solution: (D)

As we increase the size of the training data, the bias would increase while the variance would decrease.

Question Context 34:

Consider the following data where one input(X) and one output(Y) is given

34. What would be the root mean square training error for this data if you run a Linear Regression model of the form ($Y = A_0 + A_1X$)?

- A) Less than 0
- B) Greater than zero
- C) Equal to 0

D) None of these

Solution: (C)

We can perfectly fit the line on the following data so mean error will be zero.

Question Context 35-36:

Suppose you have been given the following scenario for training and validation error for Linear Regression.

35. Which of the following scenario would give you the right hyper parameter?

A) 1

B) 2

C) 3

D) 4

Solution: (B)

Option B would be the better option because it leads to less training as well as validation error.

36. Suppose you got the tuned hyper parameters from the previous question. Now, Imagine

you want to add a variable in variable space such that this added feature is important. Which of the following thing would you observe in such case?

A) Training Error will decrease and Validation error will increase

B) Training Error will increase and Validation error will increase

- C) Training Error will increase and Validation error will decrease
- D) Training Error will decrease and Validation error will decrease
- E) None of the above

Solution: (D)

If the added feature is important, the training and validation error would decrease.

Question Context 37-38:

Suppose, you got a situation where you find that your linear regression model is under fitting the data.

37. In such situation which of the following options would you consider?

- 1. I will add more variables
 - 2. I will start introducing polynomial degree variables
 - 3. I will remove some variables
- A) 1 and 2
 - B) 2 and 3
 - C) 1 and 3
 - D) 1, 2 and 3

Solution: (A)

In case of under fitting, you need to induce more variables in variable space or you can add

some polynomial degree variables to make the model more complex to be able to fit the data better.

38. Now situation is same as written in previous question(under fitting). Which of following

regularization algorithm would you prefer?

- A) L1
- B) L2
- C) Any
- D) None of these

Solution: (D)

I won't use any regularization methods because regularization is used in case of overfitting.

39. True-False: Is Logistic regression a supervised machine learning algorithm?

- A) TRUE

B) FALSE

Solution: A

True, Logistic regression is a supervised learning algorithm because it uses true labels for

training. Supervised learning algorithm should have input variables (x) and an target variable (Y) when you train the model .

40. True-False: Is Logistic regression mainly used for Regression?

A) TRUE

B) FALSE

Solution: B

Logistic regression is a classification algorithm, don't confuse with the name regression.

41. True-False: Is it possible to design a logistic regression algorithm using a Neural Network Algorithm?

A) TRUE

B) FALSE

Solution: A

True, Neural network is a universal approximator so it can implement linear regression algorithm.

42. True-False: Is it possible to apply a logistic regression algorithm on a 3-class Classification

problem?

A) TRUE

B) FALSE

Solution: A

Yes, we can apply logistic regression on 3 classification problem, We can use One Vs all method for 3 class classification in logistic regression.

43. Which of the following methods do we use to best fit the data in Logistic Regression?

A) Least Square Error

B) Maximum Likelihood

C) Jaccard distance

D) Both A and B

Solution: B

Logistic regression uses maximum likelihood estimate for training a logistic regression.

44. Which of the following evaluation metrics can not be applied in case of logistic regression output to compare with target?

A) AUC-ROC

B) Accuracy

C) Logloss

D) Mean-Squared-Error

Solution: D

Since, Logistic Regression is a classification algorithm so its output can not be real time value so mean squared error can not be used for evaluating it

45. One of the very good methods to analyze the performance of Logistic Regression is AIC,

which is similar to R-Squared in Linear Regression. Which of the following is true about AIC?

A) We prefer a model with minimum AIC value

B) We prefer a model with maximum AIC value

C) Both but depend on the situation

D) None of these

Solution: A

We select the best model in logistic regression which has least AIC.

46. [True-False] Standardisation of features is required before training a Logistic Regression.

A) TRUE

B) FALSE

Solution: B

Standardization isn't required for logistic regression. The main goal of standardizing features is to help convergence of the technique used for optimization.

47. Which of the following algorithms do we use for Variable Selection?

A) LASSO

- B) Ridge
- C) Both
- D) None of these

Solution: A

In case of lasso we apply a absolute penalty, after increasing the penalty in lasso some of the coefficient of variables may become zero

Context: 48-49

Consider a following model for logistic regression: $P(y=1|x, w) = g(w_0 + w_1x)$ where $g(z)$ is the logistic function. In the above equation the $P(y=1|x; w)$, viewed as a function of x , that we can get by changing the parameters w .

48 What would be the range of p in such case?

- A) $(0, \infty)$
- B) $(-\infty, 0)$
- C) $(0, 1)$
- D) $(-\infty, \infty)$

Solution: C

For values of x in the range of real number from $-\infty$ to $+\infty$ Logistic function will give the output between $(0,1)$

49 In above question what do you think which function would make p between $(0,1)$?

- A) logistic function
- B) Log likelihood function
- C) Mixture of both
- D) None of them

Solution: A

Explanation is same as question number 10

50. Suppose you have been given a fair coin and you want to find out the odds of getting heads. Which of the following option is true for such a case?

- A) odds will be 0
- B) odds will be 0.5
- C) odds will be 1
- D) None of these

Solution: C

Odds are defined as the ratio of the probability of success and the probability of failure. So in case of fair coin probability of success is $1/2$ and the probability of failure is $1/2$ so odd would be 1

51. The logit function(given as $l(x)$) is the log of odds function. What could be the range of logit function in the domain $x=[0,1]$?

- A) $(-\infty, \infty)$
- B) $(0,1)$
- C) $(0, \infty)$
- D) $(-\infty, 0)$

Solution: A

For our purposes, the odds function has the advantage of transforming the probability function, which has values from 0 to 1, into an equivalent function with values between 0 and ∞ . When we take the natural log of the odds function, we get a range of values from $-\infty$ to ∞ .

52. Which of the following option is true?

- A) Linear Regression errors values has to be normally distributed but in case of Logistic Regression it is
not the case
- B) Logistic Regression errors values has to be normally distributed but in case of Linear Regression it is
not the case
- C) Both Linear Regression and Logistic Regression error values have to be normally distributed
- D) Both Linear Regression and Logistic Regression error values have not to be normally distributed

Solution:A

53. Which of the following is true regarding the logistic function for any value "x"?

Note:

Logistic(x): is a logistic function of any number " x "

Logit(x): is a logit function of any number " x "

Logit_inv(x): is a inverse logit function of any number " x "

- A) $\text{Logistic}(x) = \text{Logit}(x)$
- B) $\text{Logistic}(x) = \text{Logit_inv}(x)$
- C) $\text{Logit_inv}(x) = \text{Logit}(x)$
- D) None of these

Solution: B

54. How will the bias change on using high(infinite) regularisation?

Suppose you have given the two scatter plot “a” and “b” for two classes(blue for positive and red for negative class). In scatter plot “a”, you correctly classified all data points using logistic regression (black line is a decision boundary).

- A) Bias will be high
- B) Bias will be low
- C) Can't say
- D) None of these

Solution: A

Model will become very simple so bias will be very high.

55. Suppose, You applied a Logistic Regression model on a given data and got a training accuracy X and testing accuracy Y. Now, you want to add a few new features in the same data. Select the option(s) which is/are correct in such a case.

Note: Consider remaining parameters are same.

- A) Training accuracy increases
- B) Training accuracy increases or remains the same
- C) Testing accuracy decreases
- D) Testing accuracy increases or remains the same

Solution: A and D

Adding more features to model will increase the training accuracy because model has to consider more data to fit the logistic regression. But testing accuracy increases if feature is found to be significant

56. Choose which of the following options is true regarding One-Vs-All method in

Logistic Regression.

- A) We need to fit n models in n-class classification problem
- B) We need to fit n-1 models to classify into n classes
- C) We need to fit only 1 model to classify into n classes
- D) None of these

Solution: A

If there are n classes, then n separate logistic regression has to fit, where the probability of each category is predicted over the rest of the categories combined.

57. Below are two different logistic models with different values for β_0 and β_1

Which of the following statement(s) is true about β_0 and β_1 values of two logistics models (Green, Black)?

Note: consider $Y = \beta_0 + \beta_1 * X$. Here, β_0 is intercept and β_1 is coefficient.

- A) β_1 for Green is greater than Black
- B) β_1 for Green is lower than Black
- C) β_1 for both models is same
- D) Can't Say

Solution: B

β_0 and β_1 : $\beta_0 = 0, \beta_1 = 1$ is in X1 color(black) and $\beta_0 = 0, \beta_1 = -1$ is in X4 color (green)

Context 58-60 Below are the three scatter plot(A,B,C left to right) and hand drawn decision boundaries for logistic regression.

58. Which of the following above figure shows that the decision boundary is overfitting the training data?

- A) A
- B) B
- C) C
- D) None of these

Solution: C

Since in figure 3, Decision boundary is not smooth that means it will over-fitting the data.

59. What do you conclude after seeing this visualization?

- 1. The training error in first plot is maximum as compare to second and third plot.
- 2. The best model for this regression problem is the last (third) plot because it has minimum training error (zero).
- 3. The second model is more robust than first and third because it will perform best on unseen data.
- 4. The third model is overfitting more as compare to first and second.
- 5. All will perform same because we have not seen the testing data.

- A) 1 and 3
- B) 1 and 3
- C) 1, 3 and 4
- D) 5

Solution: C

The trend in the graphs looks like a quadratic trend over independent variable X. A higher degree(Right graph) polynomial might have a very high accuracy on the train population but is expected to fail badly on test dataset. But if you see in left graph we will have training error maximum because it underfits the training data

60. Suppose, above decision boundaries were generated for the different value of regularization. Which of the above decision boundary shows the maximum regularization?

- A) A
- B) B
- C) C
- D) All have equal regularization

Solution: A

Since, more regularization means more penalty means less complex decision boundary that shows in first figure A.

61. What would do if you want to train logistic regression on same data that will take less time as well as give the comparatively similar accuracy(may not be same)?

Suppose you are using a Logistic Regression model on a huge dataset. One of the problem you may face on such huge data is that Logistic regression will take very long time to train.

- A) Decrease the learning rate and decrease the number of iteration
- B) Decrease the learning rate and increase the number of iteration
- C) Increase the learning rate and increase the number of iteration
- D) Increase the learning rate and decrease the number of iteration

Solution: D

If you decrease the number of iteration while training it will take less time for surely but will not give the same accuracy for getting the similar accuracy but not exact you need to increase the learning rate.

62. Which of the following image is showing the cost function for $y = 1$.

Following is the loss function in logistic regression(Y-axis loss function and x axis log probability) for two class classification problem.

Note: Y is the target class

- A) A
- B) B
- C) Both
- D) None of these

Solution: A

A is the true answer as loss function decreases as the log probability increases

63. Suppose, Following graph is a cost function for logistic regression.

Now, How many local minimas are present in the graph?

- A) 1
- B) 2
- C) 3
- D) 4

Solution: C

There are three local minima present in the graph

64. Can a Logistic Regression classifier do a perfect classification on the below data?

Note: You can use only X1 and X2 variables where X1 and X2 can take only two binary values(0,1).

- A) TRUE
- B) FALSE
- C) Can't say
- D) None of these

Solution: B

No, logistic regression only forms linear decision surface, but the examples in the figure are not linearly separable

Machine Learning MCQs UNIT IV

1. The SVM's are less effective when:

- A) The data is linearly separable
- B) The data is clean and ready to use
- C) The data is noisy and contains overlapping points

Ans Solution: C

When the data has noise and overlapping points, there is a problem in drawing a clear hyperplane without misclassifying.

2. The cost parameter in the SVM means:

- A) The number of cross-validations to be made
- B) The kernel to be used
- C) The tradeoff between misclassification and simplicity of the model
- D) None of the above

Ans Solution: C

The cost parameter decides how much an SVM should be allowed to "bend" with the data. For a low cost, you aim for a smooth decision surface and for a higher cost, you

aim to classify more points correctly. It is also simply referred to as the cost of misclassification.

3. Which of the following are real world applications of the SVM?

- A) Text and Hypertext Categorization
- B) Image Classification
- C) Clustering of News Articles
- D) All of the above

Ans Solution: D

SVM's are highly versatile models that can be used for practically all real world problems ranging from regression to clustering and handwriting recognitions.

4. Which of the following is true about Naive Bayes ?

- A) Assumes that all the features in a dataset are equally important
- B) Assumes that all the features in a dataset are independent
- C) Both A and B – answer
- D) None of the above options

Ans Solution: C

5 What do you mean by generalization error in terms of the SVM?

- A) How far the hyperplane is from the support vectors
- B) How accurately the SVM can predict outcomes for unseen data
- C) The threshold amount of error in an SVM

Ans Solution: B

Generalisation error in statistics is generally the out-of-sample error which is the measure of how accurately a model can predict values for previously unseen data.

6 The SVM's are less effective when:

- A) The data is linearly separable
- B) The data is clean and ready to use
- C) The data is noisy and contains overlapping points

Ans Solution: C

When the data has noise and overlapping points, there is a problem in drawing a clear hyperplane without misclassifying.

7 What is/are true about kernel in SVM?

- 1. Kernel function map low dimensional data to high dimensional space

2. It's a similarity function

- A) 1
- B) 2
- C) 1 and 2
- D) None of these

Ans Solution: C

Both the given statements are correct

Question Context: 8– 9

Suppose you are using a Linear SVM classifier with 2 class classification problem. Now you have been given the following data in which some points are circled red that are representing support vectors.

8. If you remove the following any one red points from the data. Does the decision boundary will change?

- A) Yes
- B) No

Solution: A

These three examples are positioned such that removing any one of them introduces slack in the constraints. So the decision boundary would completely change.

9. [True or False] If you remove the non-red circled points from the data, the decision boundary will change?

- A) True
- B) False

Solution: B

On the other hand, rest of the points in the data won't affect the decision boundary much.

10. What do you mean by generalization error in terms of the SVM?

- A) How far the hyperplane is from the support vectors
- B) How accurately the SVM can predict outcomes for unseen data
- C) The threshold amount of error in an SVM

Solution: B

Generalization error in statistics is generally the out-of-sample error which is the measure of how accurately a model can predict values for previously unseen data.

11. When the C parameter is set to infinite, which of the following holds true?

- A) The optimal hyperplane if exists, will be the one that completely separates the data
- B) The soft-margin classifier will separate the data
- C) None of the above

Solution: A

At such a high level of misclassification penalty, soft margin will not hold existence as there will be no room for error.

12. What do you mean by a hard margin?

- A) The SVM allows very low error in classification
- B) The SVM allows high amount of error in classification
- C) None of the above

Solution: A

A hard margin means that an SVM is very rigid in classification and tries to work extremely well in the training set, causing overfitting.

13. The minimum time complexity for training an SVM is $O(n^2)$. According to this fact, what sizes of datasets are not best suited for SVM's?

13. The minimum time complexity for training an SVM is $O(n^2)$. According to this fact, what sizes of datasets are not best suited for SVM's?

- A) Large datasets
- B) Small datasets
- C) Medium sized datasets
- D) Size does not matter

Solution: A

Datasets which have a clear classification boundary will function best with SVM's.

14. The effectiveness of an SVM depends upon:

- A) Selection of Kernel
- B) Kernel Parameters
- C) Soft Margin Parameter C
- D) All of the above

Solution: D

The SVM effectiveness depends upon how you choose the basic 3 requirements mentioned above in such a way that it maximises your efficiency, reduces error and overfitting.

15. Support vectors are the data points that lie closest to the decision surface.

- A) TRUE
- B) FALSE

Solution: A

They are the points closest to the hyperplane and the hardest ones to classify. They also have a direct bearing on the location of the decision surface.

16. The SVM's are less effective when:

- A) The data is linearly separable
- B) The data is clean and ready to use
- C) The data is noisy and contains overlapping points

Solution: C

When the data has noise and overlapping points, there is a problem in drawing a clear hyperplane without misclassifying.

17. Suppose you are using RBF kernel in SVM with high Gamma value. What does this signify?

- A) The model would consider even far away points from hyperplane for modeling
- B) The model would consider only the points close to the hyperplane for modeling
- C) The model would not be affected by distance of points from hyperplane for modeling
- D) None of the above

Solution: B

The gamma parameter in SVM tuning signifies the influence of points either near or far away from the hyperplane. For a low gamma, the model will be too constrained and include all points of the training dataset, without really capturing the shape. For a higher gamma, the model will capture the shape of the dataset well.

18. The cost parameter in the SVM means:

- A) The number of cross-validations to be made
- B) The kernel to be used
- C) The tradeoff between misclassification and simplicity of the model
- D) None of the above

Solution: C

The cost parameter decides how much an SVM should be allowed to “bend” with the data. For a low cost, you aim for a smooth decision surface and for a higher cost, you aim to classify more points correctly. It is also simply referred to as the cost of misclassification.

19. Suppose you are building a SVM model on data X. The data X can be error prone which means that you should not trust any specific data point too much. Now think that you want to build a SVM model which has quadratic kernel function of polynomial degree 2 that uses Slack variable C as one of its hyper parameter. Based upon that give the answer for following question.

What would happen when you use very large value of C(C->infinity)?

Note: For small C was also classifying all data points correctly

- A) We can still classify data correctly for given setting of hyper parameter C
- B) We can not classify data correctly for given setting of hyper parameter C
- C) Can't Say
- D) None of these

Solution: A

For large values of C, the penalty for misclassifying points is very high, so the decision boundary will perfectly separate the data if possible.

20. What would happen when you use very small C (C~0)?

- A) Misclassification would happen
- B) Data will be correctly classified
- C) Can't say
- D) None of these

Solution: A

The classifier can maximize the margin between most of the points, while misclassifying a few points, because the penalty is so low.

21. If I am using all features of my dataset and I achieve 100% accuracy on my training set, but ~70% on validation set, what should I look out for?

- A) We can still classify data correctly for given setting of hyper parameter C
- B) We can not classify data correctly for given setting of hyper parameter C
- C) Can't Say
- D) None of these

Solution: A

For large values of C, the penalty for misclassifying points is very high, so the decision boundary will perfectly separate the data if possible.

20. What would happen when you use very small C ($C \sim 0$)?

- A) Misclassification would happen
- B) Data will be correctly classified
- C) Can't say
- D) None of these

Solution: A

The classifier can maximize the margin between most of the points, while misclassifying a few points, because the penalty is so low.

21. If I am using all features of my dataset and I achieve 100% accuracy on my training set, but ~70% on validation set, what should I look out for?

- A) Underfitting
- B) Nothing, the model is perfect
- C) Overfitting

Solution: C

If we're achieving 100% training accuracy very easily, we need to check to verify if we're overfitting our data.

22. Which of the following are real world applications of the SVM?

- A) Text and Hypertext Categorization

- B) Image Classification
- C) Clustering of News Articles
- D) All of the above

Solution: D

SVM's are highly versatile models that can be used for practically all real world problems ranging from regression to clustering and handwriting recognitions.

Question Context: 23 – 25

Suppose you have trained an SVM with linear decision boundary after training SVM, you correctly infer that your SVM model is under fitting.

23. Which of the following option would you more likely to consider iterating SVM next time?

- A) You want to increase your data points
- B) You want to decrease your data points
- C) You will try to calculate more variables
- D) You will try to reduce the features

Solution: C

The best option here would be to create more features for the model.

24. Suppose you gave the correct answer in previous question. What do you think that is actually happening?

- 1. We are lowering the bias
 - 2. We are lowering the variance
 - 3. We are increasing the bias
 - 4. We are increasing the variance
- A) 1 and 2
 - B) 2 and 3
 - C) 1 and 4
 - D) 2 and 4

Solution: C

Better model will lower the bias and increase the variance

25. In above question suppose you want to change one of it's(SVM) hyperparameter so that effect would be same as previous questions i.e model will not under fit?

- A) We will increase the parameter C
- B) We will decrease the parameter C
- C) Changing in C don't effect

D) None of these

Solution: A

Increasing C parameter would be the right thing to do here, as it will ensure regularized model

26. We usually use feature normalization before using the Gaussian kernel in SVM.

What is true about feature normalization?

1. We do feature normalization so that new feature will dominate other
2. Some times, feature normalization is not feasible in case of categorical variables
3. Feature normalization always helps when we use Gaussian kernel in SVM

A) 1

B) 1 and 2

C) 1 and 3

D) 2 and 3

Solution: B

Statements one and two are correct.

Question Context: 27-29

Suppose you are dealing with 4 class classification problem and you want to train a SVM model on the data for that you are using One-vs-all method. Now answer the below questions?

27. How many times we need to train our SVM model in such case?

A) 1

B) 2

C) 3

D) 4

Solution: D

For a 4 class problem, you would have to train the SVM at least 4 times if you are using a one-vs-all method.

28. Suppose you have same distribution of classes in the data. Now, say for training 1 time in one vs all setting the SVM is taking 10 second. How many seconds would it require to train one-vs-all method end to end?

A) 20

B) 40

C) 60

D) 80

Solution: B

It would take $10 \times 4 = 40$ seconds

29 Suppose your problem has changed now. Now, data has only 2 classes. What would you think how many times we need to train SVM in such case?

- A) 1
- B) 2
- C) 3
- D) 4

Solution: A

Training the SVM only one time would give you appropriate results

Question context: 30 –31

Suppose you are using SVM with linear kernel of polynomial degree 2, Now think that you have applied this on data and found that it perfectly fit the data that means, Training and testing accuracy is 100%.

30. Now, think that you increase the complexity (or degree of polynomial of this kernel). What would you think will happen?

- A) Increasing the complexity will over fit the data
- B) Increasing the complexity will under fit the data
- C) Nothing will happen since your model was already 100% accurate
- D) None of these

Solution: A

Increasing the complexity of the data would make the algorithm overfit the data.

31. In the previous question after increasing the complexity you found that training accuracy was still 100%. According to you what is the reason behind that?

- 1. Since data is fixed and we are fitting more polynomial term or parameters so the algorithm starts memorizing everything in the data
- 2. Since data is fixed and SVM doesn't need to search in big hypothesis space

- A) 1
- B) 2
- C) 1 and 2
- D) None of these

Solution: C

Both the given statements are correct.

32. What is/are true about kernel in SVM?

1. Kernel function map low dimensional data to high dimensional space
 2. It's a similarity function
- A) 1
B) 2
C) 1 and 2
D) None of these

Solution: C

Both the given statements are correct.

Machine Learning MCQs UNIT V

1. Which of the following is a widely used and effective machine learning algorithm based on the idea of bagging?

- a) Decision Tree
- b) Regression
- c) Classification
- d) Random Forest

Ans D

2. Which of the following is a disadvantage of decision trees?

- a) Factor analysis
- b) Decision trees are robust to outliers
- c) Decision trees are prone to be overfit
- d) None of the above

Ans C

3. Can decision trees be used for performing clustering?

- a. True
- b. False

Ans Solution: (A)

Decision trees can also be used to find clusters in the data but clustering often generates natural clusters and is not dependent on any objective function.

4. Which of the following algorithm is most sensitive to outliers?

- a. K-means clustering algorithm

- b. K-medians clustering algorithm
- c. K-modes clustering algorithm
- d. K-medoids clustering algorithm

Ans Solution: (A)

5 Sentiment Analysis is an example of:

- a. Regression
- b. Classification
- c. Clustering
- d. Reinforcement Learning

Options:

- a. 1 Only
- b. 1 and 2
- c. 1 and 3
- d. 1, 2 and 4

Ans D

6 Which of the following is the most appropriate strategy for data cleaning before performing clustering analysis, given less than desirable number of data points:

Capping and flooring of variables Removal of outliers

Options:

- a. 1 only
- b. 2 only
- c. 1 and 2
- d. None of the above

Ans A

7 Which of the following is/are true about bagging trees?

- 1. In bagging trees, individual trees are independent of each other
 - 2. Bagging is the method for improving the performance by aggregating the results of weak learners
- A) 1
 - B) 2
 - C) 1 and 2
 - D) None of these

Ans Solution: C

Both options are true. In Bagging, each individual trees are independent of each other because they consider different subset of features and samples.

8. Which of the following is/are true about boosting trees?

1. In boosting trees, individual weak learners are independent of each other
 2. It is the method for improving the performance by aggregating the results of weak learners
- A) 1
B) 2
C) 1 and 2
D) None of these

Ans Solution: B

In boosting tree individual weak learners are not independent of each other because each tree correct the results of previous tree. Bagging and boosting both can be consider as improving the base learners results.

9. In Random forest you can generate hundreds of trees (say T1, T2Tn) and then aggregate the results of these tree. Which of the following is true about individual (Tk) tree in Random Forest?

1. Individual tree is built on a subset of the features
 2. Individual tree is built on all the features
 3. Individual tree is built on a subset of observations
 4. Individual tree is built on full set of observations
- A) 1 and 3
B) 1 and 4
C) 2 and 3
D) 2 and 4

Ans Solution: A

Random forest is based on bagging concept, that consider fraction of sample and fraction of feature for building the individual trees.

10. Suppose you are using a bagging based algorithm say a RandomForest in model building.Which of the following can be true?

1. Number of tree should be as large as possible
 2. You will have interpretability after using Random Forest
- A) 1
B) 2

- C) 1 and 2
- D) None of these

Ans Solution: A

Since Random Forest aggregate the result of different weak learners, If It is possible we would want more number of trees in model building. Random Forest is a black box model you will lose interpretability after using it.

11. Which of the following is/are true about Random Forest and Gradient Boosting ensemble methods?

- 1. Both methods can be used for classification task
 - 2. Random Forest is use for classification whereas Gradient Boosting is use for regression task
 - 3. Random Forest is use for regression whereas Gradient Boosting is use for Classification task
 - 4. Both methods can be used for regression task
- A) 1
 - B) 2
 - C) 3
 - D) 4
 - E) 1 and 4

Solution: E

Both algorithms are design for classification as well as regression task

12. In Random forest you can generate hundreds of trees (say T₁, T₂T_n) and then aggregate the results of these tree. Which of the following is true about individual(T_k) tree in Random Forest?

- 1. Individual tree is built on a subset of the features
 - 2. Individual tree is built on all the features
 - 3. Individual tree is built on a subset of observations
 - 4. Individual tree is built on full set of observations
- A) 1 and 3
 - B) 1 and 4
 - C) 2 and 3
 - D) 2 and 4

Solution: A

Random forest is based on bagging concept, that consider fraction of sample and fraction of feature for building the individual trees.

13. Which of the following algorithm doesn't uses learning Rate as of one of its hyperparameter?

1. Gradient Boosting

2. Extra Trees

3. AdaBoost

4. Random Forest

A) 1 and 3

B) 1 and 4

C) 2 and 3

D) 2 and 4

Solution: D

Random Forest and Extra Trees don't have learning rate as a hyperparameter.

14. Which of the following algorithm are not an example of ensemble learning algorithm?

A) Random Forest

B) Adaboost

C) Extra Trees

D) Gradient Boosting

E) Decision Trees

Solution: E

Decision trees doesn't aggregate the results of multiple trees so it is not an ensemble algorithm.

15. Suppose you are using a bagging based algorithm say a RandomForest in model building. Which of the following can be true?

1. Number of tree should be as large as possible

2. You will have interpretability after using RandomForest

A) 1

B) 2

C) 1 and 2

D) None of these

Solution: A

Since Random Forest aggregate the result of different weak learners, If It is possible we would want more number of trees in model building. Random Forest is a black box model you will lose interpretability after using it.

16. True-False: The bagging is suitable for high variance low bias models?

- A) TRUE
- B) FALSE

Solution: A

The bagging is suitable for high variance low bias models or you can say for complex models.

17. To apply bagging to regression trees which of the following is/are true in such case?

- 1. We build the N regression with N bootstrap sample
 - 2. We take the average the of N regression tree
 - 3. Each tree has a high variance with low bias
- A) 1 and 2
 - B) 2 and 3
 - C) 1 and 3
 - D) 1,2 and 3

Solution: D

All of the options are correct and self-explanatory

18. How to select best hyper parameters in tree based models?

- A) Measure performance over training data
- B) Measure performance over validation data
- C) Both of these
- D) None of these

Solution: B

We always consider the validation results to compare with the test result.

19. In which of the following scenario a gain ratio is preferred over Information Gain?

- A) When a categorical variable has very large number of category
- B) When a categorical variable has very small number of category
- C) Number of categories is the not the reason
- D) None of these

Solution: A

When high cardinality problems, gain ratio is preferred over Information Gain technique.

20. Suppose you have given the following scenario for training and validation error for Gradient Boosting. Which of the following hyper parameter would you choose in such case?

- A) 1
- B) 2
- C) 3
- D) 4

Solution: B

Scenario 2 and 4 has same validation accuracies but we would select 2 because depth is lower is better hyper parameter.

21. Which of the following is/are not true about DBSCAN clustering algorithm:

- 1. For data points to be in a cluster, they must be in a distance threshold to a core point
- 2. It has strong assumptions for the distribution of data points in dataspace
- 3. It has substantially high time complexity of order $O(n^3)$
- 4. It does not require prior knowledge of the no. of desired clusters
- 5. It is robust to outliers

Options:

- A. 1 only
- B. 2 only
- C. 4 only
- D. 2 and 3

Solution: D

✗ DBSCAN can form a cluster of any arbitrary shape and does not have strong

assumptions for the distribution of data points in the data space.

✗ DBSCAN has a low time complexity of order $O(n \log n)$ only

22. Point out the correct statement.

- a) The choice of an appropriate metric will influence the shape of the clusters
- b) Hierarchical clustering is also called HCA
- c) In general, the merges and splits are determined in a greedy manner
- d) All of the mentioned

Answer: d

Explanation: Some elements may be close to one another according to one distance and farther away according to another.

23. Which of the following is required by K-means clustering?

- a) defined distance metric
- b) number of clusters
- c) initial guess as to cluster centroids
- d) all of the mentioned

Answer: d

Explanation: K-means clustering follows partitioning approach.

24. Point out the wrong statement.

- a) k-means clustering is a method of vector quantization
- b) k-means clustering aims to partition n observations into k clusters
- c) k-nearest neighbor is same as k-means
- d) none of the mentioned

Answer: c

Explanation: k-nearest neighbour has nothing to do with k-means.

25. Which of the following function is used for k-means clustering?

- a) k-means
- b) k-mean
- c) heat map
- d) none of the mentioned

Answer: a

Explanation: K-means requires a number of clusters.

26. K-means is not deterministic and it also consists of number of iterations.

- a) True
- b) False

Answer: a

Explanation: K-means clustering produces the final estimate of cluster centroids.

Note: we are not connected with SPPU in any way.

This content is just for practice purpose, actual questions asked in exam may vary.

We do not claim any copyright of the above content

For any Suggestions / Queries / Copyright Claim / Content Removal Request contact us
at

READ MORE: 10 Best Machine Learning Institutes in Pune 2020

READ MORE: The Complete Guide To Become A Machine Learning Engineer

Yogita Sahu
Content Writer

 Trending

- < Top Money Earning Apps In India
- > Complete Information and Cyber Security MCQs | SPPU Final Year

1 thought on “Complete Machine Learning MCQs Unit Wise | SPPU Final Year”

Seat No -

Total number of questions : 60

11930_MACHINE LEARNING

Time : 1hr

Max Marks : 50

N.B

- 1) All questions are Multiple Choice Questions having single correct option.
 - 2) Attempt any 50 questions out of 60.
 - 3) Use of calculator is allowed.
 - 4) Each question carries 1 Mark.
 - 5) Specially abled students are allowed 20 minutes extra for examination.
 - 6) Do not use pencils to darken answer.
 - 7) Use only black/blue ball point pen to darken the appropriate circle.
 - 8) No change will be allowed once the answer is marked on OMR Sheet.
 - 9) Rough work shall not be done on OMR sheet or on question paper.
 - 10) Darken ONLY ONE CIRCLE for each answer.
-

Q.no 1. Which of these is not a frequent pattern mining algorithm?

A : Apriori

B : FP growth

C : Decision trees

D : Eclat

Q.no 2. A table with all possible value of a random variable and its corresponding probabilities is called

A : Probability Mass Function

B : Probability Density Function

C : Cumulative distribution function

D : Probability Distribution**Q.no 3. If you use an ensemble of different base models, is it necessary to tune the hyper parameters of all base models to improve the ensemble performance?**

A : Yes

D : No

C : Can't Say

D : May be

Q.no 4. Following are the descriptive models

A : Clustering

B : Classification

C : Association rule

D : Classification and Association Rule

Q.no 5. What is the minimum no. of variables/ features required to perform clustering?

A : 0

B : 1

C : 2

D : 3

Q.no 6. The shape of the Normal Curve is ----

A : Spiked

B : Flat

C : Circular

D : Bell shaped

Q.no 7. In random experiment, observations of random variable are classified as ----

A : Events

B : Composition

C : Trials

D : Functions

Q.no 8. Movie Recommendation systems are an example of:

A : Classification and Clustering

B : Clustering and Reinforcement Learning

C : Reinforcement Learning and Regression

D : Regression and classification

Q.no 9. In a Binomial Distribution, the mean and variance are equal.

A : Yes

B : No

C : Can't Say

D : May be

Q.no 10. The weight of persons in a state is a -----

A : Continuous random variable

B : Discrete random variable

C : Irregular random variable

D : Uncertain random variable

Q.no 11. It is not necessary to have a target variable for applying dimensionality reduction algorithms

A : True

B : false

C : Can't Say

D : May be

Q.no 12. The difference between the actual Y value and the predicted Y value found using a regression equation is called the

A : slope

B : residual

C : outlier

D : scatter plot

Q.no 13. Which of the following is the types of supervised learning

A : Classification

B : Clustering

C : Reinforcement Learning

D : k-means

Q.no 14. A perceptron adds up all the weighted inputs it receives, and if it exceeds a certain value, it outputs a 1, otherwise it just outputs a 0.

A : true

B : False

C : Sometimes – it can also output intermediate values as well

D : Can't say

Q.no 15. Mean and variance of Poisson's distribution is the same.

A : Yes

B : No

C : Can't Say

D : May be

Q.no 16. Variance of a constant 'a' is

A : 0

B : a

C : a/2

D : 1

Q.no 17. K means and K-medoids are example of which type of clustering method?

A : partition

B : Hierarchical

C : probabilistic

D : they are not clustering methods

Q.no 18. How many coefficients do you need to estimate in a simple linear regression model

X A : 1

B : 2

C : 3

D : 4

Q.no 19. What are closed itemsets?

A : An itemset for which at least one proper super-itemset has same support

B : An itemset whose no proper super-itemset has same support

C : An itemset for which at least super-itemset has same confidence

D : An itemset whose no proper super-itemset has same confidence

Q.no 20. The classification is considered to be

A : Supervised Learning

B : Unsupervised Learning

C : semi-Supervised Learning

D : Deep Learning

Q.no 21. In the mathematical Equation of Linear Regression $Y = \Theta_0 + \Theta_1 * x + \epsilon$, (Θ_0, Θ_1) refers to _

A : (X-intercept, Slope)

B : (Slope, X-Intercept)

C : (Y-Intercept, Slope)

D : (slope, Y-Intercept)

Q.no 22. What is gini index?

A : It is a type of index structure

B : It is a measure of purity

C : It is an index as well as measure of purity

D : Nothing related to ML

Q.no 23. Which is example of Multi-Class Classification?

A : Plant species classification

B : Social network analysis

C : Medical imaging

D : Email spam detection

Q.no 24. Application of machine learning methods to large databases is called

A : Data Mining.

B : Artificial Intelligence

C : Big Data Computing

D : Internet of Things

Q.no 25. Like the probabilistic view, the _____ view allows us to associate a probability of membership with each classification.

A : exemplar

B : deductive

C : classical

D : inductive

Q.no 26. In terms of the bias-variance decomposition, a 1-nearest neighbor classifier has than a 3-nearest neighbor classifier.

A : higher variance

B : higher bias

C : lower variance

D : Higher Bias & Lower variance

Q.no 27. What is the purpose of performing cross-validation?

A : To assess the predictive performance of the models

B : To judge how the trained model performs inside the sample on test data

C : To find the maxima or minima at the local point

D : Normalize the data

Q.no 28. The VC dimension of hypothesis space H1 is larger than the VC dimension of hypothesis space H2. Which of the following can be inferred from this?

A : The number of examples required for learning a hypothesis in H1 is larger than the number of examples required for H2.

B : The number of examples required for learning a hypothesis in H1 is smaller than the number of examples required for H2.

C : No relation to number of samples required for PAC learning

D : The number of examples required for learning a hypothesis in H1 is smaller than the number of examples required for H1.

Q.no 29. Support Vector Machine is -----

A : Logical Model

B : Probabilistic Model

C : Geometric Model

D : Neural Network model

Q.no 30. Lasso can be interpreted as least-squares linear regression where

A : weights are regularized with the L1 norm

B : the weights have a Gaussian prior

C : weights are regularized with the L2 norm

D : the solution algorithm is simpler

Q.no 31. Having multiple perceptrons can actually solve the XOR problem satisfactorily: this is because each perceptron can partition off a linear part of the space itself, and they can then combine their results.

A : True – this works always, and these multiple perceptrons learn to classify even complex problems

B : False – perceptrons are mathematically incapable of solving linearly inseparable functions, no matter what you do

C : True – perceptrons can do this but are unable to learn to do it – they have to be explicitly hand-coded

D : False – just having a single perceptron is enough

Q.no 32. The Classification predictions can be evaluated using _____

A : accuracy

B : Error rate

C : root mean squared error

D : Value of K

Q.no 33. Which of the following is a good test dataset characteristic? (A) Large enough to yield meaningful results (B) Is representative of the dataset as a whole

A : A only

B : B only

C : Both A and B

D : None of A and B

Q.no 34. The cost parameter in the SVM means:

A : The number of cross-validations to be made

B : The kernel to be used

C : The tradeoff between misclassification and simplicity of the model

D : None of the above

Q.no 35. The Both PCA and Lasso can be used for feature selection. Which of the following statements are true?

- A : Lasso selects a subset (not necessarily a strict subset) of the original features
- B : PCA and Lasso both allow you to specify how many features are chosen
- C : PCA produces features that are non linear combinations of the original features
- D : PCA and Lasso are the same if you use the kernel trick

Q.no 36. Supervised learning and unsupervised clustering both require which is correct according to the statement.

- A : output attribute.
- B : hidden attribute.
- C : input attribute.
- D : categorical attribute

Q.no 37. A person trained to interact with a human expert in order to capture their knowledge.

- A : knowledge programmer
- B : knowledge developer
- C : knowledge engineer
- D : knowledge extractor

Q.no 38. $E(X) = n*p*q$ is for which distribution?

- A : Bernoulli's
- B : Binomial
- C : Poisson's
- D : Normal

Q.no 39. What is back propagation?

- A : It is another name given to the curvy function in the perceptron
- B : It is the transmission of error back through the network to adjust the inputs

C : It is the transmission of error back through the network to allow weights to be adjusted so that the network can learn

D : None of the mentioned

Q.no 40. Linear Regression has dependent variables that have _____

A : K-Values

B continuous values.

C : N-Values

D : Square Values

Q.no 41. If the probability that a bomb dropped from a place will strike the target is 60% and if 10 bombs are dropped, find mean and variance?

A : 0.4, 0.25

B : 6,24

C : 0.4, 0.17

D : 0.6, 0.17

Q.no 42. Which of the following, specifies the prior probability of each utterance?

A : Sound Model

B : Language Model

C : Visual Model

D : System Model

Q.no 43. Among the following, is viewed as problem of probabilistic inference?

A : Speech recognition

B : Speaking

C : Hearing

D : Utterance

Q.no 44. What is Morphological Segmentation?

A : Does Discourse Analysis

B : Separate words into individual morphemes and identify the class of the morphemes

C : Is an extension of propositional logic

D : Automatic Text Summarization

Q.no 45. Wrapper methods are hyper-parameter selection methods that

A : Should be used whenever possible because they are computationally efficient

B : Should be avoided unless there are no other options because they are always prone to overfitting.

C : Are useful mainly when the learning machines are “black boxes”

D : Should be avoided altogether.

Q.no 46. Find the mean and variance of X? Where $x=\{0, 1, 2, 3, 4\}$ and $f(x) = \{ 1/9, 2/9, 3/9, 2/9, 1/9\}$

A . 2 ,4 / 3

B : 3,4/3

C : 2,2/3

D : 3,3/3

Q.no 47. Which of the following option is / are correct regarding benefits of ensemble model?

1. Better performance
2. Generalized models
3. Better interpretability

A : 1 and 2

B : 1 and 3

C : 2 and 3

D : 1,2 and 3

Q.no 48. What is the expectation of X? Where $x=\{0, 1, 2, 3\}$ and $f(x) = \{ 1/6, 2/6, 2/6, 1/6\}$

A : 0.5

B . 1.5

C : 2.5

D : 3.5

Q.no 49. What is a heuristic function?

A : A function to solve mathematical problems

B : A function which takes parameters of type string and returns an integer value

C : A function whose return type is nothing

~~D.~~ A function that maps from problem state descriptions to measures of desirability

Q.no 50. If $P(x) = 0.5$ and $x = 4$, then $E(x) = ?$

A : 1

B : 0.5

C : 4

~~D.~~ 2

Q.no 51. Different learning methods does not include?

A : Memorization

B : Analogy

C : Deduction

~~D.~~ Introduction

Q.no 52. Knowledge may be

~~A.~~ Declarative and procedural

B : Declarative and non-procedural

C : Procedural and non-procedural

D : Declarative, procedural and non-procedural

Q.no 53. Suppose you are using RBF kernel in SVM with high Gamma value. What does this signify?

A : The model would consider even far away points from hyperplane for modeling

~~B.~~ The model would consider only the points close to the hyperplane for modeling

C : The model would not be affected by distance of points from hyperplane for modeling

D : None of the above

Q.no 54. Which of the following is a reasonable way to select the number of principal components "k"?

~~A.~~ Choose k to be the smallest value so that at least 99% of the variance is retained. - answer

B : Choose k to be 99% of m ($k = 0.99*m$, rounded to the nearest integer).

C : Choose k to be the largest value so that 99% of the variance is retained.

D : Use the elbow method

Q.no 55. In a Binomial Distribution, if 'n' is the number of trials and 'p' is the probability of success, then the mean value is given by

- A : $n \cdot p$
- B : n
- C : p
- D : $n \cdot p \cdot (1-p)$

Q.no 56. Suppose your model is demonstrating high variance across the different training sets. Which of the following is NOT valid way to try and reduce the variance?

- A : Increase the amount of training data in each training set
- B : Improve the optimization algorithm being used for error minimization.**
- C : Decrease the model complexity
- D : Reduce the noise in the training data

Q.no 57. Which of the following is direct application of frequent itemset mining?

- A : Social Network Analysis**
- B : Market Basket Analysis
- C : Outlier Detection
- D : Intrusion Detection

Q.no 58. You are given seismic data and you want to predict next earthquake , this is an example of

- A : Supervised learning**
- B : Reinforcement learning
- C : Unsupervised learning
- D : Dimensionality reduction

Q.no 59. How can we assign the weights to output of different models in an ensemble?

1. Use an algorithm to return the optimal weights
2. Choose the weights using cross validation
3. Give high weights to more accurate models

- A : 1 and 2
- B : 1 and 3
- C : 2 and 3
- D : 1,2 and 3**

Q.no 60. What does FP growth algorithm do?

- A : It mines all frequent patterns through pruning rules with lesser support
- B : It mines all frequent patterns through pruning rules with higher support
- C : It mines all frequent patterns by constructing a FP tree**
- D : It mines all frequent patterns without any pruning rules with higher support

Q.no 1. Which of the following is true?

- A : Both apriori and FP-Growth uses horizontal data format**
- B : Both apriori and FP-Growth uses vertical data format
- C : Apriori uses horizontal and FP-Growth uses vertical data format
- D : Apriori uses vertical and FP-Growth uses horizontal data format

Q.no 2. In the regression equation $Y = 75.65 + 0.50*X$, the intercept is

- A : 0.5
- B : 75.65**
- C : 1
- D : indeterminable

Q.no 3. Which of the following statements about Naive Bayes is incorrect?

- A : Attributes are equally important.
- B : Attributes are statistically dependent of one another given the class value.**
- C : Attributes are statistically independent of one another given the class value.
- D : Attributes can be nominal or numeric

Q.no 4. In Standard normal distribution, the value of median is ----

- A : 2
- B : 1
- C : 0**
- D : Not Fixed

Q.no 5. In boosting, individual base learners can be parallel.

- A : Yes
- B : No**

C : Can't Say

D : May be

Q.no 6. The apriori algorithm works in a ----- and ----- fashion?

A : top-down and depth-first

B : top-down and breath-first

C : bottom-up and depth-first

D : bottom-up and breath-first

Q.no 7. If machine learning model output involves target variable then that model is called as

A : Descriptive model

B : Predictive Model

C : Reinforcement Learning

D : All of the above

Q.no 8. VC dimension stands for

A : Vepin Chervo

B : vapnik chervonenkis

C : Vebinic cheronic

D : Velocory Cherumal

Q.no 9. Impact of high variance on the training set ?

A : overfitting

B : underfitting

C : both underfitting & overfitting

D : Depents upon the dataset

Q.no 10. Which of the following algorithm is not an example of an ensemble method?

A : Extra Tree Regressor

B : Random Forest

C : Gradient Boosting

D : Decision Tree

Q.no 11. A random variable that assumes a finite or a countably infinite number of values is called -----

A : Continuous random variable

B : Discrete random variable

C : Irregular random variable

D : Uncertain random variable

Q.no 12. Ensembles will yield bad results when there is significant diversity among the models.

A : Yes

B : No

C : Can't Say

D : May be

Q.no 13. Mutually Exclusive events _____

A : Contain all sample points

B : Contain all common sample points

C : Does not contain any sample point

D : Does not contain any common sample point

Q.no 14. What are closed frequent itemsets?

A : A closed itemset

B : A frequent itemset

C : An itemset which is both closed and frequent

D : An empty set

Q.no 15. In simple term, machine learning is

(A) Training based on historical data

(B) Prediction to answer a query

A : A only

B : B only

C : Both A and B

D : None of A and B

Q.no 16. Normal Distribution is symmetric is about ----

A : Variance

B : Standard Deviation

C : Mean

D : Covariance

Q.no 17. Linear Regression is a _____ machine learning algorithm.

A : Supervised

B : Unsupervised

C : Semi-Supervised

D : Can't say

Q.no 18. If machine learning model output doesnot involves target variable then that model is called as

A : Descriptive model

B : Predictive Model

C : Reinforcement Learning

D : All of the above

Q.no 19. Out of the following values, which one is not possible in probability?

A : 1

B : 0

C : 0.5

D : -0.5

Q.no 20. What are maximal frequent itemsets?

A : A frequent itemset whose no super-itemset is frequent

B : A frequent itemset whose super-itemset is also frequent

C : A non-frequent itemset whose super-itemset is frequent

D : A non-frequent itemset whose super-itemset is also non-frequent

Q.no 21. Which of the following indicates the fundamental of least squares?

A : arithmetic mean should be maximized

B : arithmetic mean should be zero

C : arithmetic mean should be neutralized

D : arithmetic mean should be minimized

Q.no 22. Which of the following is true about averaging ensemble?

A : It can only be used in classification problem

B : It can only be used in regression problem

C : It can be used in both classification as well as regression

D : It is not used anywhere

Q.no 23. You trained a binary classifier model which gives very high accuracy on the training data, but much lower accuracy on validation data. Which is false.

A : This is an instance of overfitting

B : This is an instance of underfitting

C : The training was not well regularized

D : The training and testing examples are sampled from different distributions

Q.no 24. Which is example of Binary classification?

A : Social network analysis

B : Medical imaging

C : Email spam detection

D : Image segmentation

Q.no 25. Which of the following sentence is FALSE regarding regression?

A : It relates inputs to outputs.

B : It is used for prediction.

C : It may be used for interpretation.

D : It discovers causal relationships

Q.no 26. Which of the following methods do we use to best fit the data in Logistic Regression?

A : Least Square Error

B : Maximum Likelihood

C : Jaccard distance

D : Both A and B

Q.no 27. How do we know if we are underfitting?

- A : If by increasing capacity we decrease generalization error
- B : If the error representing the training set is relatively small
- C : If by increasing capacity we increase generalization error
- D : generalization error is large

Q.no 28. Chance Nodes are represented by _____

- A : disk
- B : square
- C : circle
- D : triangle

Q.no 29. For a given data having 100 examples,if squared SE1,SE2, and SE3 are 13.33,3.33, and 4.00 respectively, calculate Mean Squared Error(MSE)?

- A : 0.663
- B : 0.2066
- C : 0.345
- D : 0.567

Q.no 30. Some telecommunication company wants to segment their customers into distinct groups ,this is an example of

- A : Supervised learning
- B : Reinforcement learning
- C : Unsupervised learning
- D : Data extraction

Q.no 31. The error function most suited for gradient descent using logistic regression is

- A : The entropy function.
- B : The squared error.
- C : The cross-entropy function.
- D : The number of mistakes.

Q.no 32. Which is the Decision tree creation algorithm?

- A : K-Means

B : ID3

C : K--medoid

D : Naive Bayes

Q.no 33. What are support vectors?

A : All the examples that have a non-zero weight α_k in a SVM

B : The only examples necessary to compute $f(x)$ in an SVM.

C : All of the above

D : None of the above

Q.no 34. Which algorithm that can be used for binary classification?

A : K-Means

B : Naive Bayes

C : K--medoid

D : Hierarchical

Q.no 35. Neural Networks are complex _____ with many parameters.

A : Linear Functions

B : Nonlinear Functions

C : Discrete Functions

D : Exponential Functions

Q.no 36. What are the reasons for overfitting?

A : Small number of features

B : Training set is too large

C : Noisy data & large number of features

D : Testing data is to small

Q.no 37. Which of the following classifications would best suit the student performance classification systems?

A : if then analysis

B : Market Basket Analysis

C : Regression analysis

D : Cluster analysis

Q.no 38. Decision Nodes are represented by _____

A : disk

B : square

C : circle

D : triangle

Q.no 39. What are two steps of tree pruning work?

A : Pessimistic pruning and post pruning

B : Postpruning and Prepruning

C : Cost complexity pruning and time complexity pruning

D : Pessimistic pruning and Optimistic pruning

Q.no 40. If X and Y in a regression model are totally unrelated,

A : the correlation coefficient would be -1

B : the coefficient of determination would be 0

C : the coefficient of determination would be 1

D : the SSE would be 0

Q.no 41. Type of matrix decomposition model is

A : Descriptive model

B : Predictive Model

C : Logical model

D : Geometrical model

Q.no 42. In a Binomial Distribution, if p, q and n are probability of success, failure and number of trials respectively then variance is given by

A : $n*p$

B : $n*p*q$

C : p

D : $n*p*(1-p)$

Q.no 43. What do you mean by support(A)?

A : Total number of transactions containing A

B : Total number of transactions not containing A

C : Number of transactions containing A / Total number of transactions

D : Number of transactions not containing A / Total number of transactions

Q.no 44. What is the relation between candidate and frequent itemsets?

A : A candidate itemset is always a frequent itemset

B : A frequent itemset must be a candidate itemset

C : No relation between candidate and frequent

D : candidate and frequent are same

Q.no 45. Point out the wrong statement.

A : Regression through the origin yields an equivalent slope if you center the data first

B : Normalizing variables results in the slope being the correlation

C : Least squares is not an estimation tool

D : None of the mentioned

Q.no 46. Find lambda in Poisson's distribution if the probabilities of getting a head in biased coin toss as 3/4 and 6 coins are tossed.

A : 3.5

B : 4.5

C : 5.5

D : 6.5

Q.no 47. In the example of predicting number of babies based on stork's population ,Number of babies is

A : outcome

B : feature

C : observation

D : attribute

Q.no 48. What does Apriori algorithm do?

A : It mines all frequent patterns through pruning rules with lesser support

B : It mines all frequent patterns through pruning rules with higher support

C : It mines all frequent patterns through pruning rules without support

D : It mines all frequent patterns without any pruning rules with higher support

Q.no 49. A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college. Here feature type is -----

A : nominal

B : ordinal

C : categorical

D : boolean

Q.no 50. Which of the following is an example of feature extraction?

A : Construction bag of words from an email

B : Applying PCA to project high dimensional data

C : Removing stop words

D : Forward selection

Q.no 51. Let S1 and S2 be the set of support vectors and w1 and w2 be the learnt weight vectors for a linearly separable problem using hard and soft margin linear SVMs respectively. Which of the following are correct?

A : S1 is subset of S2

B : S1 may not be a subset of S2

C : w1 = w2

D : All of the above

Q.no 52. Indicate which one is a method of density estimation

A : Histogram based

B : Branch and bound procedure

C : Neighborhood distance

D : Elbow method

Q.no 53. Which of the following techniques would perform better for reducing dimensions of a data set?

A : Removing columns which have too many missing values

B : Removing columns which have high variance in data

C : Removing columns with dissimilar data trends

D : None of these

Q.no 54. Which of the following methods can not achieve zero training error on any linearly separable dataset?

A : Decision tree

B : 15-nearest neighbors

C : Hard-margin SVM

D : Perceptron

Q.no 55. For a Poisson Distribution, if mean(m) = 1, then $P(1)$ is?

A : Indeterminate

B : e

C : e/2

D : 1/e

Q.no 56. Suppose, you have 2000 different models with their predictions and want to ensemble predictions of best x models. Now, which of the following can be a possible method to select the best x models for an ensemble?

A : Step wise forward selection

B : Step wise backward elimination

C : Step wise forward and backward elimination

D : Gradiant Descent

Q.no 57. When do you consider an association rule interesting?

A : If it only satisfies min_support

B : If it only satisfies min_confidence

C : If it satisfies both min_support and min_confidence

D : There are no measures to check so

Q.no 58. Previous probabilities in Bayes Theorem that are changed with help of new available information are classified as ----

A : independent probabilities

B : posterior probabilities

C : interior probabilities

D : dependent probabilities

Q.no 59. If $E(x) = 2$ and $E(z) = 4$, then $E(z - x) = ?$

A : 2

B : 6

C : 0

D : can't say

Q.no 60. What is not true about FP growth algorithms?

A : It mines frequent itemsets without candidate generation

B : There are chances that FP trees may not fit in the memory

C : FP trees are very expensive to build

D : It expands the original database to build FP trees.