Machine Learning (ML) MCQs [set-30]

726. _____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.

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 above

Answer: A

727. It's possible to specify if the scaling process must include both mean and standard deviation using the parameters_____.

A. with_mean=True/False

B. with_std=True/False

C. Both A & B

D. None of the Mentioned

Answer: C

728. 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

Answer: C

729. 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)

Answer: A

730. In the mathematical Equation of Linear Regression Y = ?1 + ?2X + ?, (?1, ?2) refers to _____

A. (X-intercept, Slope)

B. (Slope, X-Intercept)

C. (Y-Intercept, Slope)

D. (slope, Y-Intercept)

Answer: C

731. 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?

A. Bias increases and Variance increases

B. Bias decreases and Variance increases

C. Bias decreases and Variance decreases

D. Bias increases and Variance decreases

Answer: D

732. 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

A. 1 and 2

B. 2 and 3

C. 1 and 3

D. 1, 2 and 3

Answer: A

733. The minimum time complexity for training an SVM is O(n2). 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

734. 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

735. 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 variables3. 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

736. 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

Answer: B

Machine Learning (ML) MCQs [set-29]

701. SVM can solve linear and non-linear problems

A. true

B. false

Answer: A

702. 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.

A. true

B. false

Answer: A

703. Hyperplanes are _____boundaries that help classify the data points.

A. usual

B. decision

C. parallel

Answer: B

704. The _____of the hyperplane depends upon the number of features.

A. dimension

B. classification

C. reduction

Answer: A

705. Hyperplanes are decision boundaries that help classify the data points.

A. true

B. false

Answer: A

706. SVM algorithms use a set of mathematical functions that are defined as the kernel.

A. true

B. false

Answer: A

707. In SVR we try to fit the error within a certain threshold.

A. true

B. false

Answer: A

708. 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: C

709. 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

Answer: C

710. Which of the following is not supervised learning?

A. PCA

B. Decision Tree

C. Naive Bayesian

D. Linerar regression

Answer: A

711. _____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.

A. Supervised

B. Semi-supervised

C. Reinforcement

D. Clusters

Answer: B

712. In reinforcement learning, this feedback is usually called as___.

A. Overfitting

B. Overlearning

C. Reward

D. None of above

Answer: C

713. In the last decade, many researchers started training bigger and bigger models, built with several different layers that's why this approach is called_____.

A. Deep learning

B. Machine learning

C. Reinforcement learning

D. Unsupervised learning

Answer: A

714. 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 _____

A. Regression

B. Accuracy

C. Modelfree

D. Scalable

Answer: C

715. _____ showed better performance than other approaches, even without a context-based model

A. Machine learning

B. Deep learning

C. Reinforcement learning

D. Supervised learning

Answer: B

716. If two variables are correlated, is it necessary that they have a linear relationship?

A. Yes

B. No

Answer: B

717. 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

Answer: B

---

718. 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

Answer: C

---

719. In syntax of linear model lm(formula,data,..), data refers to _____

A. Matrix

B. Vector

C. Array

D. List

Answer: B

---

720. 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.

Answer: D

---

721. 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.

Answer: D

722. 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

723. 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

Answer: D

724. _____which can accept a NumPy RandomState generator or an integer seed.

A. make_blobs

B. random_state

C. test_size

D. training_size

Answer: B

725. 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

A. 1

B. 2

C. 3

D. 4

Answer: B

Machine Learning (ML) MCQs [set-28]

676. 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

Answer: C

677. 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.

Answer: C

678. 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

Answer: A

679. 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

Answer: A

680. 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

Answer: B

681. 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

Answer: A

682. What is/are true about kernel in SVM?1. Kernel function map low dimensional data to high dimensional space2. It’s a similarity function

A. 1

B. 2

C. 1 and 2

D. None of these

Answer: C

683. 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~0)?

A. Misclassification would happen

B. Data will be correctly classified

C. Can’t say

D. None of these

Answer: A

684. 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

685. How do you handle missing or corrupted data in a dataset?

A. a. Drop missing rows or columns

B. b. Replace missing values with mean/median/mode

C. c. Assign a unique category to missing values

D. d. All of the above

Answer: D

686. 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

Answer: B

687. 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

688. If there is only a discrete number of possible outcomes called _____.

A. Modelfree

B. Categories

C. Prediction

D. None of above

Answer: B

689. Some people are using the term ___ instead of prediction only to avoid the weird idea that machine learning is a sort of modern magic.

A. Inference

B. Interference

C. Accuracy

D. None of above

Answer: A

690. The term _____ can be freely used, but with the same meaning adopted in physics or system theory.

A. Accuracy

B. Cluster

C. Regression

D. Prediction

Answer: D

691. Common deep learning applications / problems can also be solved using____

A. Real-time visual object identification

B. Classic approaches

C. Automatic labeling

D. Bio-inspired adaptive systems

Answer: B

692. what is the function of 'Unsupervised Learning'?

A. Find clusters of the data and find low-dimensional representations of the data

B. Find interesting directions in data and find novel observations/ database cleaning

C. Interesting coordinates and correlations

D. All

Answer: D

693. What are the two methods used for the calibration in Supervised Learning?

A. Platt Calibration and Isotonic Regression

B. Statistics and Informal Retrieval

Answer: A

694. 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

Answer: B

695. 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

Answer: B

696. 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 decreases and Adjusted R-squared increases

A. 1 and 2

B. 1 and 3

C. 2 and 4

D. None of the above

Answer: A

697. 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.

Answer: D

698. If two variables are correlated, is it necessary that they have a linear relationship?

A. Yes

B. No

Answer: B

699. 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

Answer: A

700. 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 large value of C(C->infinity)?

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

Machine Learning (ML) MCQs [set-27]

651. Any linear combination of the components of a multivariate Gaussian is a univariate Gaussian.

A. True

B. false

Answer: A

652. Solving a non linear separation problem with a hard margin Kernelized SVM (Gaussian RBF Kernel) might lead to overfitting

A. True

B. false

Answer: A

653. SVM is a ---------- algorithm

A. Classification

B. Clustering

C. Regression

D. All

Answer: A

654. SVM is a ---------- learning

A. Supervised

B. Unsupervised

C. Both

D. None

Answer: A

655. The linear SVM classifier works by drawing a straight line between two classes

A. True

B. false

Answer: A

656. What is Model Selection in Machine Learning?

A. The process of selecting models among different mathematical models, which are used to describe the same data set

B. when a statistical model describes random error or noise instead of underlying relationship

C. Find interesting directions in data and find novel observations/ database cleaning

D. All above

Answer: A

657. Which are two techniques of Machine Learning ?

A. Genetic Programming and Inductive Learning

B. Speech recognition and Regression

C. Both A & B

D. None of the Mentioned

Answer: A

658. Even if there are no actual supervisors _____ learning is also based on feedback provided by the environment

A. Supervised

B. Reinforcement

C. Unsupervised

D. None of the above

Answer: B

659. When it is necessary to allow the model to develop a generalization ability and avoid a common problem called_____.

A. Overfitting

B. Overlearning

C. Classification

D. Regression

Answer: A

660. Techniques involve the usage of both labeled and unlabeled data is called___.

A. Supervised

B. Semi-supervised

C. Unsupervised

D. None of the above

Answer: B

661. A supervised scenario is characterized by the concept of a _____.

A. Programmer

B. Teacher

C. Author

D. Farmer

Answer: B

662. overlearning causes due to an excessive _____.

A. Capacity

B. Regression

C. Reinforcement

D. Accuracy

Answer: A

663. Which of the following are several models for feature extraction

A. regression

B. classification

C. None of the above

Answer: C

664. _____ provides some built-in datasets that can be used for testing purposes.

A. scikit-learn

B. classification

C. regression

D. None of the above

Answer: A

665. While using _____ all labels are turned into sequential numbers.

A. LabelEncoder class

B. LabelBinarizer class

C. DictVectorizer

D. FeatureHasher

Answer: A

666. _____produce sparse matrices of real numbers that can be fed into any machine learning model.

A. DictVectorizer

B. FeatureHasher

C. Both A & B

D. None of the Mentioned

Answer: C

667. scikit-learn offers the class_____, which is responsible for filling the holes using a strategy based on the mean, median, or frequency

A. LabelEncoder

B. LabelBinarizer

C. DictVectorizer

D. Imputer

Answer: D

668. Which of the following scale data by removing elements that don't belong to a given range or by considering a maximum absolute value.

A. MinMaxScaler

B. MaxAbsScaler

C. Both A & B

D. None of the Mentioned

Answer: C

669. scikit-learn also provides a class for per-sample normalization,_____

A. Normalizer

B. Imputer

C. Classifier

D. All above

Answer: A

670. _____dataset with many features contains information proportional to the independence of all features and their variance.

A. normalized

B. unnormalized

C. Both A & B

D. None of the Mentioned

Answer: B

671. In order to assess how much information is brought by each component, and the correlation among them, a useful tool is the_____.

A. Concuttent matrix

B. Convergance matrix

C. Supportive matrix

D. Covariance matrix

Answer: D

672. The_____ parameter can assume different values which determine how the data matrix is initially processed.

A. run

B. start

C. init

D. stop

Answer: C

673. _____allows exploiting the natural sparsity of data while extracting principal components.

A. SparsePCA

B. KernelPCA

C. SVD

D. init parameter

Answer: A

674. 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

Answer: A

675. 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

Machine Learning (ML) MCQs [set-26]

626. 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_____.

A. RobustScaler

B. DictVectorizer

C. LabelBinarizer

D. FeatureHasher

Answer: A

627. scikit-learn also provides a class for per-sample normalization, Normalizer. It can apply_____to each element of a dataset

A. max, l0 and l1 norms

B. max, l1 and l2 norms

C. max, l2 and l3 norms

D. max, l3 and l4 norms

Answer: B

628. There are also many univariate methods that can be used in order to select the best features according to specific criteria based on_____.

A. F-tests and p-values

B. chi-square

C. ANOVA

D. All above

Answer: A

629. _____performs a PCA with non-linearly separable data sets.

A. SparsePCA

B. KernelPCA

C. SVD

D. None of the Mentioned

Answer: B

630. 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 these

Answer: B

631. The parameter_____ allows specifying the percentage of elements to put into the test/training set

A. test_size

B. training_size

C. All above

D. None of these

Answer: C

632. In many classification problems, the target _____ is made up of categorical labels which cannot immediately be processed by any algorithm.

A. random_state

B. dataset

C. test_size

D. All above

Answer: B

633. _____adopts a dictionary-oriented approach, associating to each category label a progressive integer number.

A. LabelEncoder class

B. LabelBinarizer class

C. DictVectorizer

D. FeatureHasher

Answer: A

634. 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)

Answer: A

635. In syntax of linear model lm(formula,data,..), data refers to _____

A. Matrix

B. Vector

C. Array

D. List

Answer: B

636. 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

637. 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

Answer: D

638. 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

Answer: A

639. Naive Bayes classifiers are a collection ----------of algorithms

A. Classification

B. Clustering

C. Regression

D. All

Answer: A

640. Naive Bayes classifiers is _____ Learning

A. Supervised

B. Unsupervised

C. Both

D. None

Answer: A

641. Features being classified is independent of each other in Naïve Bayes Classifier

A. False

B. true

Answer: B

642. Features being classified is _____ of each other in Naïve Bayes Classifier

A. Independent

B. Dependent

C. Partial Dependent

D. None

Answer: A

643. Conditional probability is a measure of the probability of an event given that another event has already occurred.

A. True

B. false

Answer: A

644. Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

A. True

B. false

Answer: A

645. Bernoulli Naïve Bayes Classifier is _____distribution

A. Continuous

B. Discrete

C. Binary

Answer: C

646. Multinomial Naïve Bayes Classifier is _____distribution

A. Continuous

B. Discrete

C. Binary

Answer: B

647. Gaussian Naïve Bayes Classifier is _____distribution

A. Continuous

B. Discrete

C. Binary

Answer: A

648. Binarize parameter in BernoulliNB scikit sets threshold for binarizing of sample features.

A. True

B. false

Answer: A

649. Gaussian distribution when plotted, gives a bell shaped curve which is symmetric about the _____ of the feature values.

A. Mean

B. Variance

C. Discrete

D. Random

Answer: A

650. SVMs directly give us the posterior probabilities P(y = 1jx) and P(y = ?1jx)

A. True

B. false

Answer: B

Machine Learning (ML) MCQs [set-25]

601. 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.

Answer: C

602. 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

Answer: A

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

A. linear

B. quadratic

C. reciprocal

D. inverse

Answer: A

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

A. linear

B. nonlinear

C. categorical

D. symmetrical

Answer: B

605. 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.

Answer: C

606. 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

Answer: D

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

A. linear regression

B. logistic regression

C. simple regression

D. multiple linear regression

Answer: B

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

A. linear regression

B. bayes classifier

C. logistic regression

D. backpropagation learning

Answer: A

609. 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.

Answer: B

610. 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

Answer: D

611. 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

Answer: C

612. 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

Answer: C

613. 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.

Answer: B

614. In reinforcement learning if feedback is negative one it is defined as____.

A. Penalty

B. Overlearning

C. Reward

D. None of above

Answer: A

615. According to____ , it’s a key success factor for the survival and evolution of all species.

A. Claude Shannon\s theory

B. Gini Index

C. Darwin’s theory

D. None of above

Answer: C

616. What is ‘Training set’?

A. Training set is used to test the accuracy of the hypotheses generated by the learner.

B. A set of data is used to discover the potentially predictive relationship.

C. Both A & B

D. None of above

Answer: B

617. Common deep learning applications include____

A. Image classification, Real-time visual tracking

B. Autonomous car driving, Logistic optimization

C. Bioinformatics, Speech recognition

D. All above

Answer: D

618. Reinforcement learning is particularly efficient when_____.

A. the environment is not completely deterministic

B. it\s often very dynamic

C. it\s impossible to have a precise error measure

D. All above

Answer: D

619. if there is only a discrete number of possible outcomes (called categories), the process becomes a_____.

A. Regression

B. Classification.

C. Modelfree

D. Categories

Answer: B

---

620. Which of the following are supervised learning applications

A. Spam detection, Pattern detection, Natural Language Processing

B. Image classification, Real-time visual tracking

C. Autonomous car driving, Logistic optimization

D. Bioinformatics, Speech recognition

Answer: A

---

621. 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.

A. Logical

B. Classical

C. Classification

D. None of above

Answer: D

---

622. What is ‘Overfitting’ in Machine learning?

A. when a statistical model describes random error or noise instead of underlying relationship ‘overfitting’ occurs.

B. Robots are programed so that they can perform the task based on data they gather from sensors.

C. While involving the process of learning ‘overfitting’ occurs.

D. a set of data is used to discover the potentially predictive relationship

Answer: A

---

623. What is ‘Test set’?

A. Test set is used to test the accuracy of the hypotheses generated by the learner.

B. It is a set of data is used to discover the potentially predictive relationship.

C. Both A & B

D. None of above

Answer: A

---

624. _____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

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 above

Answer: B

---

625. How it's possible to use a different placeholder through the parameter_____.

A. regression

B. classification

C. random_state

D. missing_values

Answer: D

Machine Learning (ML) MCQs [set-24]

576. We usually use feature normalization before using the Gaussian k

A. e 1

B. 1 and 2

C. 1 and 3

D. 2 and 3

Answer: B

577. 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

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

A. deduction

B. abduction

C. induction

D. conjunction

Answer: C

579. Computers are best at learning

A. facts.

B. concepts.

C. procedures.

D. principles.

Answer: A

580. Data used to build a data mining model.

A. validation data

B. training data

C. test data

D. hidden data

Answer: B

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

A. hidden attribute.

B. output attribute.

C. input attribute.

D. categorical attribute.

Answer: A

582. 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.

Answer: B

583. 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

Answer: C

584. 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

Answer: C

585. A multiple regression model has the form: y = 2 + 3x1 + 4x2. As x1 increases by 1 unit (holding x2 constant), y will

A. increase by 3 units

B. decrease by 3 units

C. increase by 4 units

D. decrease by 4 units

Answer: C

586. 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

Answer: B

587. 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

Answer: C

588. 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

Answer: D

589. 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

Answer: C

590. 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

Answer: B

591. 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.

Answer: B

592. Another name for an output attribute.

A. predictive variable

B. independent variable

C. estimated variable

D. dependent variable

Answer: B

593. 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.

Answer: C

594. 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.

Answer: D

595. 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

Answer: A

596. 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

Answer: D

597. 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

Answer: B

598. 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.

Answer: A

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

A. training

B. test

C. verification

D. validation

Answer: D

600. 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.

Answer: A

Machine Learning (ML) MCQs [set-23]

551. In a real problem, you should check to see if the SVM is separable and th

A. true

B. false

Answer: B

552. In reinforcement learning, this feedback is usually called as .

A. overfitting

B. overlearning

C. reward

D. none of above

Answer: C

553. In the last decade, many researchers started training bigger and bigger models, built with several different layers that's why this approach is called .

A. deep learning

B. machine learning

C. reinforcement learning

D. unsupervised learning

Answer: A

554. When it is necessary to allow the model to develop a generalization ability and avoid a common problem called .

A. overfitting

B. overlearning

C. classification

D. regression

Answer: A

555. Techniques involve the usage of both labeled and unlabeled data is called .

A. supervised

B. semi- supervised

C. unsupervised

D. none of the above

Answer: B

556. Reinforcement learning is particularly efficient when .

A. the environment is not completely deterministic

B. it\s often very dynamic

C. it\s impossible to have a precise error measure

D. all above

Answer: D

557. 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.

A. logical

B. classical

C. classification

D. none of above

Answer: D

558. if there is only a discrete number of possible outcomes (called categories), the process becomes a .

A. regression

B. classification.

C. modelfree

D. categories

Answer: B

559. 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

Answer: D

560. scikit-learn also provides functions for creating dummy datasets from scratch:

A. make_classifica tion()

B. make_regressio n()

C. make_blobs()

D. all above

Answer: D

561. which can accept a NumPy RandomState generator or an integer seed.

A. make_blobs

B. random_state

C. test_size

D. training_size

Answer: B

562. 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

A. 1

B. 2

C. 3

D. 4

Answer: B

563. It's possible to specify if the scaling process must include both mean and standard deviation using the parameters .

A. with_mean=tru e/false

B. with_std=true/ false

C. both a & b

D. none of the mentioned

Answer: C

564. Which of the following selects the best K high-score features.

A. selectpercentil e

B. featurehasher

C. selectkbest

D. all above

Answer: C

565. What is/are true about ridge regression?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 infinity, we get very, very small coefficients approaching 04. 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

Answer: A

566. 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?

A. increase

B. decrease

C. remain constant

D. can't say

Answer: D

567. 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?

A. bias increases and variance increases

B. bias decreases and variance increases

C. bias decreases and variance decreases

D. bias increases and variance decreases

Answer: D

568. Problem: Players will play if weather is sunny. Is this statement is correct?

A. true

B. false

Answer: A

569. Multinomial Naïve Bayes Classifier is _ distribution

A. continuous

B. discrete

C. binary

Answer: B

570. The minimum time complexity for training an SVM is O(n2). 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

571. 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 variables3. 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

572. Which of the following is not supervised learning?

A. pca

B. decision tree

C. naive bayesian

D. linerar regression

Answer: A

573. Gaussian Naïve Bayes Classifier is _ distribution

A. continuous

B. discrete

C. binary

Answer: A

574. 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

---

575. 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 misclassificati on and simplicity of the model

D. none of the above

Answer: C

Machine Learning (ML) MCQs [set-22]

526. Which of the following is not supervised learning?

A. pca

B. decision tree

C. naive bayesian

D. linerar regression

Answer: A

527. Gaussian Naïve Bayes Classifier is _ distribution

A. continuous

B. discrete

C. binary

Answer: A

528. 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

529. 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 ontest data

C. both a and b

Answer: C

530. 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?

A. yes

B. no

Answer: A

531. Linear SVMs have no hyperparameters that need to be set by cross-validation

A. true

B. false

Answer: B

532. For the given weather data, what is the probability that players will play if weather is sunny

A. 0.5

B. 0.26

C. 0.73

D. 0.6

Answer: D

533. 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

A. 0.4

B. 0.2

C. 0.6

D. 0.45

Answer: B

534. Problem: Players will play if weather is sunny. Is t

A. true

B. false

Answer: A

535. For the given weather data, Calculate probability

A. 0.4

B. 0.64

C. 0.29

D. 0.75

Answer: B

536. For the given weather data, Calculate probability

A. 0.4

B. 0.64

C. 0.36

D. 0.5

Answer: C

537. For the given weather data, what is the probabilit

A. 0.5

B. 0.26

C. 0.73

D. 0.6

Answer: D

538. 100 people are at party. Given data gives informa

A. 0.4

B. 0.2

C. 0.6

D. 0.45

Answer: B

539. 100 people are at party. Given data gives informa

A. true

B. false

Answer: A

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

A. how far the hy

B. how accuratel

C. the threshold amount of error i

Answer: B

541. The effectiveness of an SVM depends upon:

A. selection of ke

B. kernel param

C. soft margin pa

D. all of the abov

Answer: D

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

A. true

B. false

Answer: A

543. The SVM’s are less effective when:

A. the data is line

B. the data is cl

C. the data is noisy and contains

Answer: C

544. Suppose you are using RBF kernel in SVM with high Gamma valu

A. the model wo

B. uthe model wo

C. the model wou

D. none of the ab

Answer: B

545. 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 misclassificati on and simplicity of the model

D. none of the above

Answer: C

546. If I am using all features of my dataset and I achieve 100% accura

A. underfitting

B. nothing, the m

C. overfitting

Answer: C

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

A. text and hype

B. image classifi

C. clustering of n

D. all of the abov

Answer: D

548. 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?

A. you want to in

B. you want to d

C. you will try to c

D. you will try to r

Answer: C

549. We usually use feature normalization before using the Gaussian k

A. e 1

B. 1 and 2

C. 1 and 3

D. 2 and 3

Answer: B

550. Linear SVMs have no hyperparameters that need to be set by cross-valid

A. true

B. false

Answer: B

Machine Learning (ML) MCQs [set-21]

501. 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

Answer: D

502. 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.

Answer: D

503. scikit-learn also provides functions for creating dummy datasets from scratch:

A. make_classifica tion()

B. make_regressio n()

C. make_blobs()

D. all above

Answer: D

504. which can accept a NumPy RandomState generator or an integer seed.

A. make_blobs

B. random_state

C. test_size

D. training_size

Answer: B

505. 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

A. 1

B. 2

C. 3

D. 4

Answer: B

---

506. 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.

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 above

Answer: A

---

507. It's possible to specify if the scaling process must include both mean and standard deviation using the parameters .

A. with_mean=tru e/false

B. with_std=true/ false

C. both a & b

D. none of the mentioned

Answer: C

---

508. Which of the following selects the best K high-score features.

A. selectpercentil e

B. featurehasher

C. selectkbest

D. all above

Answer: C

---

509. 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

Answer: C

510. What is/are true about ridge regression?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 infinity, we get very, very small coefficients approaching 04. 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

Answer: A

511. 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

512. 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)

Answer: A

513. In the mathematical Equation of Linear Regression Y = ?1 + ?2X + ?, (?1, ?2) refers to

A. (x-intercept, slope)

B. (slope, x- intercept)

C. (y-intercept, slope)

D. (slope, y- intercept)

Answer: C

514. Suppose that we have N independent variables (X1,X2… Xn) 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 X1) with Y is -0.95.Which of the following is true for X1?

A. relation between the x1 and y is weak

B. relation between the x1 and y is strong

C. relation between the x1 and y is neutral

D. correlation can't judge the relationship

Answer: B

515. 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?

A. increase

B. decrease

C. remain constant

D. can't say

Answer: D

516. 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?

A. bias increases and variance increases

B. bias decreases and variance increases

C. bias decreases and variance decreases

D. bias increases and variance decreases

Answer: D

517. 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

A. 1 and 2

B. 2 and 3

C. 1 and 3

D. 1, 2 and 3

Answer: A

518. Problem: Players will play if weather is sunny. Is this statement is correct?

A. true

B. false

Answer: A

519. Multinomial Naïve Bayes Classifier is _ distribution

A. continuous

B. discrete

C. binary

Answer: B

520. For the given weather data, Calculate probability of not playing

A. 0.4

B. 0.64

C. 0.36

D. 0.5

Answer: C

521. The minimum time complexity for training an SVM is O(n2). 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

522. 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

523. 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

524. 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 variables3. 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

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

A. true

B. false

Answer: A

Machine Learning (ML) MCQs [set-20]

476. 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

477. 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

Answer: C

478. 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?

A. yes

B. no

Answer: A

479. Linear SVMs have no hyperparameters that need to be set by cross-validation

A. true

B. false

Answer: B

480. For the given weather data, what is the probability that players will play if weather is sunny

A. 0.5

B. 0.26

C. 0.73

D. 0.6

Answer: D

481. 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

A. 0.4

B. 0.2

C. 0.6

D. 0.45

Answer: B

482. Linear SVMs have no hyperparameters

A. true

B. false

Answer: B

483. What are the different Algorithm techniques in Machine Learning?

A. supervised learning and semi-

B. unsupervised learning and transduction

C. both a & b

D. none of the mentioned

Answer: C

484. 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

A. supervised

B. semi- supervised

C. reinforcement

D. clusters

Answer: B

485. In reinforcement learning, this feedback is usually called as .

A. overfitting

B. overlearning

C. reward

D. none of above

Answer: C

486. In the last decade, many researchers started training bigger and bigger models, built with several different layers that's why this approach is called .

A. deep learning

B. machine learning

C. reinforcement learning

D. unsupervised learning

Answer: A

487. What does learning exactly mean?

A. robots are programed so that they can

B. a set of data is used to discover the

C. learning is the ability to change

D. it is a set of data is used to discover the

Answer: C

488. When it is necessary to allow the model to develop a generalization ability and avoid a common problem called .

A. overfitting

B. overlearning

C. classification

D. regression

Answer: A

489. Techniques involve the usage of both labeled and unlabeled data is called .

A. supervised

B. semi- supervised

C. unsupervised

D. none of the above

Answer: B

490. 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

A. regression

B. accuracy

C. modelfree

D. scalable

Answer: C

491. showed better performance than other approaches, even without a context-based model

A. machine learning

B. deep learning

C. reinforcement learning

D. supervised learning

Answer: B

492. Which of the following sentence is correct?

A. machine learning relates with the study,

B. data mining can be defined as the process

C. both a & b

D. none of the above

Answer: C

493. What is 'Overfitting' in Machine learning?

A. when a statistical model describes random error or noise instead of

B. robots are programed so that they can perform the task based on data they gather from

C. while involving the process of learning 'overfitting' occurs.

D. a set of data is used to discover the potentially predictive relationship

Answer: A

494. What is 'Test set'?

A. test set is used to test the accuracy of the hypotheses generated by the learner.

B. it is a set of data is used to discover the potentially predictive relationship.

C. both a & b

D. none of above

Answer: A

495. what is the function of 'Supervised Learning'?

A. classifications, predict time series, annotate strings

B. speech recognition, regression

C. both a & b

D. none of above

Answer: C

496. Commons unsupervised applications include

A. object segmentation

B. similarity detection

C. automatic labeling

D. all above

Answer: D

497. Reinforcement learning is particularly efficient when .

A. the environment is not completely deterministic

B. it\s often very dynamic

C. it\s impossible to have a precise error measure

D. all above

Answer: D

498. 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.

A. logical

B. classical

C. classification

D. none of above

Answer: D

499. Common deep learning applications include

A. image classification, real-time visual tracking

B. autonomous car driving, logistic optimization

C. bioinformatics, speech recognition

D. all above

Answer: D

500. if there is only a discrete number of possible outcomes (called categories), the process becomes a .

A. regression

B. classification.

C. modelfree

D. categories

Answer: B

Machine Learning (ML) MCQs [set-19]

451. Which of the following method is used to find the optimal features for cluster analysis

A. k-means

B. density-based spatial clustering

C. spectral clustering find clusters

D. all above

Answer: D

452. scikit-learn also provides functions for creating dummy datasets from scratch:

A. make_classification()

B. make_regression()

C. make_blobs()

D. all above

Answer: D

453. which can accept a NumPy RandomState generator or an integer seed.

A. make_blobs

B. random_state

C. test_size

D. training_size

Answer: B

454. 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

A. 1

B. 2

C. 3

D. 4

Answer: B

455. 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.

A. labelencoder class

B. dictvectorizer

C. labelbinarizer class

D. featurehasher

Answer: C

456. 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.

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 above

Answer: A

457. It's possible to specify if the scaling process must include both mean and standard deviation using the parameters .

A. with_mean=true/false

B. with_std=true/false

C. both a & b

D. none of the mentioned

Answer: C

458. Which of the following selects the best K high-score features.

A. selectpercentile

B. featurehasher

C. selectkbest

D. all above

Answer: C

459. 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.

A. 1 and 4

B. 2 and 3

C. 1 and 3

D. none of theses

Answer: A

---

460. 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

Answer: C

---

461. What is/are true about ridge regression?1. When lambda is 0, model works like linear regression model2. When lambda is 0, model doesnt work like linear regression model3. When lambda goes to infinity, we get very, very small coefficients approaching 04. 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

Answer: A

---

462. 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

---

463. Function used for linear regression in R

A. lm(formula, data)

B. lr(formula, data)

C. lrm(formula, data)

D. regression.linear(formula,

Answer: A

464. In the mathematical Equation of Linear Regression Y?=??1 + ?2X + ?, (?1, ?2) refers to

A. (x-intercept, slope)

B. (slope, x-intercept)

C. (y-intercept, slope)

D. (slope, y-intercept)

Answer: C

465. Suppose that we have N independent variables (X1,X2 Xn) 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 X1) with Y is -0.95.Which of the following is true for X1?

A. relation between the x1 and y is weak

B. relation between the x1 and y is strong

C. relation between the x1 and y is neutral

D. correlation can•t judge the relationship

Answer: B

466. 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?

A. increase

B. decrease

C. remain constant

D. can't say

Answer: D

467. 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?

A. bias increases and variance increases

B. bias decreases and variance increases

C. bias decreases and variance decreases

D. bias increases and variance decreases

Answer: D

---

468. 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 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

Answer: A

---

469. Problem:Players will play if weather is sunny. Is this statement is correct?

A. true

B. false

Answer: A

---

470. For the given weather data, Calculate probability of not playing

A. 0.4

B. 0.64

C. 0.36

D. 0.5

Answer: C

---

471. 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?

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

---

472. The minimum time complexity for training an SVM is O(n2). According to this fact, what sizes of datasets are not best suited for SVMs?

A. large datasets

B. small datasets

C. medium sized datasets

D. size does not matter

Answer: A

---

473. 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

---

474. 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

---

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

A. true

B. false

Answer: A

Machine Learning (ML) MCQs [set-18]

426. If two variables are correlated, is it necessary that they have a linear relationship?

A. yes

B. no

Answer: B

427. Correlated variables can have zero correlation coeffficient. True or False?

A. true

B. false

Answer: A

428. Suppose we fit Lasso Regression to a data set, which has 100 features (X1,X2X100). 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

Answer: B

429. 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

Answer: C

430. 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

Answer: D

---

431. In syntax of linear model lm(formula,data,..), data refers to

A. matrix

B. vector

C. array

D. list

Answer: B

---

432. Linear Regression is a supervised machine learning algorithm.

A. true

B. false

Answer: A

---

433. It is possible to design a Linear regression algorithm using a neural network?

A. true

B. false

Answer: A

---

434. 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

---

435. 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

Answer: C

436. 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 dont 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.

Answer: D

437. 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.

Answer: D

438. 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

Answer: D

439. 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.

Answer: B

440. 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

Answer: B

441. In a real problem, you should check to see if the SVM is separable and then include slack variables if it is not separable.

A. true

B. false

Answer: B

442. 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

443. 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?

A. true

B. false

Answer: A

444. For the given weather data, Calculate probability of playing

A. 0.4

B. 0.64

C. 0.29

D. 0.75

Answer: B

445. In SVR we try to fit the error within a certain threshold.

A. true

B. false

Answer: A

446. In reinforcement learning, this feedback is usually called as .

A. overfitting

B. overlearning

C. reward

D. none of above

Answer: C

447. Which of the following sentence is correct?

A. machine learning relates with the study, design and development of the algorithms that give computers the capability to learn without being explicitly programmed.

B. data mining can be defined as the process in which the unstructured data tries to extract knowledge or unknown interesting patterns.

C. both a & b

D. none of the above

Answer: C

448. Reinforcement learning is particularly

A. the environment is not

B. it\s often very dynamic

C. it\s impossible to have a

D. all above

Answer: D

449. Lets 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

Answer: D

450. 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.

Answer: D

Machine Learning (ML) MCQs [set-17]

401. 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

Answer: B

402. Which of the following statement(s) can

A. 1 and 2

B. 1 and 3

C. 2 and 4

D. none of the above

Answer: A

403. 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.

Answer: D

404. If two variables are correlated, is it necessary that they have a linear relationship?

A. yes

B. no

Answer: B

405. 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.

Answer: D

406. Suppose you are using a Linear SVM classifier with 2 class classification

A. yes

B. no

Answer: A

407. If you remove the non-red circled points from the data, the decision boundary will change?

A. true

B. false

Answer: B

408. 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

Answer: A

409. 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)?

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

410. SVM can solvelinearand non- linearproblems

A. true

B. false

Answer: A

411. 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.

A. true

B. false

Answer: A

412. Hyperplanes are boundaries that help classify the data points.

A. usual

B. decision

C. parallel

Answer: B

413. The of the hyperplane depends upon the number of features.

A. dimension

B. classification

C. reduction

Answer: A

414. Hyperplanes are decision boundaries that help classify the data points.

A. true

B. false

Answer: A

415. SVMalgorithmsusea set of mathematical functions that are defined as thekernel.

A. true

B. false

Answer: A

416. In SVM, Kernel function is used to map a lower dimensional data into a higher dimensional data.

A. true

B. false

Answer: A

417. In SVR we try to fit the error within a

A. true

B. false

Answer: A

418. 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: C

419. 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

Answer: C

420. Which of the following isnotsupervised learning?

A. ••pca

B. ••decision tree

C. ••naive bayesian

D. linerar regression

Answer: A

421. 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.

A. supervised

B. semi-supervised

C. reinforcement

D. clusters

Answer: B

422. In reinforcement learning, this feedback is

A. overfitting

B. overlearning

C. reward

D. none of above

Answer: C

423. In the last decade, many researchers started training bigger and bigger models, built with several different layers that's why this approach is called .

A. deep learning

B. machine learning

C. reinforcement learning

D. unsupervised learning

Answer: A

424. 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

A. regression

B. accuracy

C. modelfree

D. scalable

Answer: C

425. showed better performance than other approaches, even without a context- based model

A. machine learning

B. deep learning

C. reinforcement learning

D. supervised learning

Answer: B

Machine Learning (ML) MCQs [set-16]

376. 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.

Answer: D

377. 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

Answer: A

378. 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

Answer: A

379. Can we calculate the skewness of variables based on mean and median?

A. true

B. false

Answer: B

380. 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

Answer: B

381. 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-

A. 1 and 2

B. 1 and 3

C. 2 and 4

D. none of the above

Answer: A

382. 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

Answer: B

383. Conditional probability is a measure of the probability of an event given that another event has already occurred.

A. true

B. false

Answer: A

384. What is/are true about kernel in SVM?1. Kernel function map low dimensional data to high dimensional space2. Its a similarity function

A. 1

B. 2

C. 1 and 2

D. none of these

Answer: C

385. 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 small C (C~0)?

A. misclassification would happen

B. data will be correctly classified

C. can't say

D. none of these

Answer: A

386. 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

387. If you remove the non-red circled points from the data, the decision boundary will

A. true

B. false

Answer: B

388. 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

389. The SVMs 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

390. If there is only a discrete number of possible outcomes called .

A. modelfree

B. categories

C. prediction

D. none of above

Answer: B

391. Some people are using the term instead of prediction only to avoid the weird idea that machine learning is a sort of modern magic.

A. inference

B. interference

C. accuracy

D. none of above

Answer: A

392. The term can be freely used, but with the same meaning adopted in physics or system theory.

A. accuracy

B. cluster

C. regression

D. prediction

Answer: D

393. Common deep learning applications / problems can also be solved using

A. real-time visual object identification

B. classic approaches

C. automatic labeling

D. bio-inspired adaptive systems

Answer: B

394. Identify the various approaches for machine learning.

A. concept vs classification learning

B. symbolic vs statistical learning

C. inductive vs analytical learning

D. all above

Answer: D

395. what is the function of Unsupervised Learning?

A. find clusters of the data and find low-dimensional representations of the data

B. find interesting directions in data and find novel observations/ database cleaning

C. interesting coordinates and correlations

D. all

Answer: D

396. What are the two methods used for the calibration in Supervised Learning?

A. platt calibration and isotonic regression

B. statistics and informal retrieval

Answer: A

397. Which of the following are several models for feature extraction

A. regression

B. classification

C. none of the above

Answer: C

398. Let's say, a Linear regression model perfectly fits the training data (train error

A. you will always have test error zero

B. you can not have test error zero

C. none of the above

Answer: C

399. Which of the following assumptions do we make while deriving linear regression parameters?

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

A. 1,2 and 3.

B. 1,3 and 4.

C. 1 and 3.

D. all of above.

Answer: D

400. Suppose we fit Lasso Regression to a data set, which has 100 features (X1,X2X100). 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

Answer: B

Machine Learning (ML) MCQs [set-15]

351. In reinforcement learning if feedback is negative one it is defined as .

A. penalty

B. overlearning

C. reward

D. none of above

Answer: A

352. According to , it's a key success factor for the survival and evolution of all species.

A. claude shannon\s theory

B. gini index

C. darwin's theory

D. none of above

Answer: C

353. A supervised scenario is characterized by the concept of a .

A. programmer

B. teacher

C. author

D. farmer

Answer: B

354. overlearning causes due to an excessive .

A. capacity

B. regression

C. reinforcement

D. accuracy

Answer: A

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

A. pca

B. k-means

C. none of the above

Answer: A

356. Which of the following model model include a backwards elimination feature selection routine?

A. mcv

B. mars

C. mcrs

D. all above

Answer: B

357. Can we extract knowledge without apply feature selection

A. yes

B. no

Answer: A

358. While using feature selection on the data, is the number of features decreases.

A. no

B. yes

Answer: B

359. Which of the following are several models

A. regression

B. classification

C. none of the above

Answer: C

360. provides some built-in datasets that can be used for testing purposes.

A. scikit-learn

B. classification

C. regression

D. none of the above

Answer: A

361. While using all labels are turned into sequential numbers.

A. labelencoder class

B. labelbinarizer class

C. dictvectorizer

D. featurehasher

Answer: A

362. produce sparse matrices of real numbers that can be fed into any machine learning model.

A. dictvectorizer

B. featurehasher

C. both a & b

D. none of the mentioned

Answer: C

363. scikit-learn offers the class , which is responsible for filling the holes using a strategy based on the mean, median, or frequency

A. labelencoder

B. labelbinarizer

C. dictvectorizer

D. imputer

Answer: D

364. Which of the following scale data by removing elements that don't belong to a given range or by considering a maximum absolute value.

A. minmaxscaler

B. maxabsscaler

C. both a & b

D. none of the mentioned

Answer: C

365. scikit-learn also provides a class for per- sample normalization,

A. normalizer

B. imputer

C. classifier

D. all above

Answer: A

366. dataset with many features contains information proportional to the independence of all features and their variance.

A. normalized

B. unnormalized

C. both a & b

D. none of the mentioned

Answer: B

367. In order to assess how much information is brought by each component, and the correlation among them, a useful tool is the .

A. concuttent matrix

B. convergance matrix

C. supportive matrix

D. covariance matrix

Answer: D

368. The parameter can assume different values which determine how the data matrix is initially processed.

A. run

B. start

C. init

D. stop

Answer: C

369. allows exploiting the natural sparsity of data while extracting principal components.

A. sparsepca

B. kernelpca

C. svd

D. init parameter

Answer: A

370. 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

Answer: A

371. Overfitting is more likely when you have huge amount of data to train?

A. true

B. false

Answer: B

372. 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

373. Lets 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

Answer: C

374. 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.

Answer: C

375. 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

Answer: A

Machine Learning (ML) MCQs [set-14]

326. In given image, P(H) is probability.

A. posterior

B. prior

Answer: B

327. Conditional probability is a measure of the probability of an event given that another

A. true

B. false

Answer: A

328. Bayes theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

A. true

B. false

Answer: A

329. Bernoulli Nave Bayes Classifier is distribution

A. continuous

B. discrete

C. binary

Answer: C

330. Multinomial Nave Bayes Classifier is distribution

A. continuous

B. discrete

C. binary

Answer: B

331. Gaussian Nave Bayes Classifier is distribution

A. continuous

B. discrete

C. binary

Answer: A

332. Binarize parameter in BernoulliNB scikit sets threshold for binarizing of sample features.

A. true

B. false

Answer: A

333. Gaussian distribution when plotted, gives a bell shaped curve which is symmetric about the of the feature values.

A. mean

B. variance

C. discrete

D. random

Answer: A

334. SVMs directly give us the posterior probabilities P(y = 1jx) and P(y = ??1jx)

A. true

B. false

Answer: B

335. Any linear combination of the components of a multivariate Gaussian is a univariate Gaussian.

A. true

B. false

Answer: A

336. Solving a non linear separation problem with a hard margin Kernelized SVM (Gaussian RBF Kernel) might lead to overfitting

A. true

B. false

Answer: A

337. SVM is a algorithm

A. classification

B. clustering

C. regression

D. all

Answer: A

338. SVM is a learning

A. supervised

B. unsupervised

C. both

D. none

Answer: A

339. The linearSVMclassifier works by drawing a straight line between two classes

A. true

B. false

Answer: A

340. Which of the following function provides unsupervised prediction ?

A. cl_forecastb

B. cl_nowcastc

C. cl_precastd

D. none of the mentioned

Answer: D

341. Which of the following is characteristic of best machine learning method ?

A. fast

B. accuracy

C. scalable

D. all above

Answer: D

342. What are the different Algorithm techniques in Machine Learning?

A. supervised learning and semi-supervised learning

B. unsupervised learning and transduction

C. both a & b

D. none of the mentioned

Answer: C

343. What is the standard approach to supervised learning?

A. split the set of example into the training set and the test

B. group the set of example into the training set and the test

C. a set of observed instances tries to induce a general rule

D. learns programs from data

Answer: A

344. Which of the following is not Machine Learning?

A. artificial intelligence

B. rule based inference

C. both a & b

D. none of the mentioned

Answer: B

345. What is Model Selection in Machine Learning?

A. the process of selecting models among different mathematical models, which are used to describe the same data set

B. when a statistical model describes random error or noise instead of underlying relationship

C. find interesting directions in data and find novel observations/ database cleaning

D. all above

Answer: A

346. Which are two techniques of Machine Learning ?

A. genetic programming and inductive learning

B. speech recognition and regression

C. both a & b

D. none of the mentioned

Answer: A

347. Even if there are no actual supervisors learning is also based on feedback provided by the environment

A. supervised

B. reinforcement

C. unsupervised

D. none of the above

Answer: B

348. What does learning exactly mean?

A. robots are programed so that they can perform the task based on data they gather from sensors.

B. a set of data is used to discover the potentially predictive relationship.

C. learning is the ability to change according to external stimuli and remembering most of all previous experiences.

D. it is a set of data is used to discover the potentially predictive relationship.

Answer: C

349. When it is necessary to allow the model to develop a generalization ability and avoid a common problem called .

A. overfitting

B. overlearning

C. classification

D. regression

Answer: A

350. Techniques involve the usage of both labeled and unlabeled data is called .

A. supervised

B. semi-supervised

C. unsupervised

D. none of the above

Answer: B

Machine Learning (ML) MCQs [set-13]

301. 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 these

Answer: B

302. 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

Answer: A

303. What is PCA, KPCA and ICA used for?

A. principal components analysis

B. kernel based principal component analysis

C. independent component analysis

D. all above

Answer: D

304. Can a model trained for item based similarity also choose from a given set of items?

A. yes

B. no

Answer: A

305. What are common feature selection methods in regression task?

A. correlation coefficient

B. greedy algorithms

C. all above

D. none of these

Answer: C

306. The parameter allows specifying the percentage of elements to put into the test/training set

A. test_size

B. training_size

C. all above

D. none of these

Answer: C

307. In many classification problems, the target is made up of categorical labels which cannot immediately be processed by any algorithm.

A. random_state

B. dataset

C. test_size

D. all above

Answer: B

308. adopts a dictionary-oriented approach, associating to each category label a progressive integer number.

A. labelencoder class

B. labelbinarizer class

C. dictvectorizer

D. featurehasher

Answer: A

309. 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

Answer: C

310. 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

Answer: D

---

311. 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

Answer: D

---

312. 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)

Answer: A

---

313. In syntax of linear model lm(formula,data,..), data refers to

A. matrix

B. vector

C. array

D. list

Answer: B

---

314. In the mathematical Equation of Linear Regression Y = ?1 + ?2X + ?, (?1, ?2) refers to

A. (x-intercept, slope)

B. (slope, x-intercept)

C. (y-intercept, slope)

D. (slope, y-intercept)

Answer: C

---

315. Linear Regression is a supervised machine learning algorithm.

A. true

B. false

Answer: A

316. It is possible to design a Linear regression algorithm using a neural network?

A. true

B. false

Answer: A

317. Overfitting is more likely when you have huge amount of data to train?

A. true

B. false

Answer: B

318. 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

Answer: A

319. 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

320. Naive Bayes classifiers are a collection ----------of algorithms

A. classification

B. clustering

C. regression

D. all

Answer: A

321. Naive Bayes classifiers is Learning

A. supervised

B. unsupervised

C. both

D. none

Answer: A

322. Features being classified is independent of each other in Nave Bayes Classifier

A. false

B. true

Answer: B

323. Features being classified is of each other in Nave Bayes Classifier

A. independent

B. dependent

C. partial dependent

D. none

Answer: A

324. 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.

A. true

B. false

Answer: A

325. In given image, P(H|E) is probability.

A. posterior

B. prior

Answer: A

Machine Learning (ML) MCQs [set-12]

276. 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

A. 1, 3 and 4

B. 1,3 and 6

C. 1,3, 4 and 6

D. all of above

Answer: D

277. 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.

A. bagging

B. 1,3 and 6

C. a or b

D. none of these

Answer: A

278. 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

B. no

C. can’t say

Answer: B

279. Which of the following is NOT supervised learning?

A. pca

B. decision tree

C. linear regression

D. naive bayesian

Answer: A

280. According to , it's a key success factor for the survival and evolution of all species.

A. claude shannon\s theory

B. gini index

C. darwin's theory

D. none of above

Answer: C

281. How can you avoid overfitting ?

A. by using a lot of data

B. by using inductive machine learning

C. by using validation only

D. none of above

Answer: A

282. What are the popular algorithms of Machine Learning?

A. decision trees and neural networks (back propagation)

B. probabilistic networks and nearest neighbor

C. support vector machines

D. all

Answer: D

283. What is Training set?

A. training set is used to test the accuracy of the hypotheses generated by the learner.

B. a set of data is used to discover the potentially predictive relationship.

C. both a & b

D. none of above

Answer: B

284. Common deep learning applications include

A. image classification, real-time visual tracking

B. autonomous car driving, logistic optimization

C. bioinformatics, speech recognition

D. all above

Answer: D

285. what is the function of Supervised Learning?

A. classifications, predict time series, annotate strings

B. speech recognition, regression

C. both a & b

D. none of above

Answer: C

286. Commons unsupervised applications include

A. object segmentation

B. similarity detection

C. automatic labeling

D. all above

Answer: D

287. Reinforcement learning is particularly efficient when .

A. the environment is not completely deterministic

B. it\s often very dynamic

C. it\s impossible to have a precise error measure

D. all above

Answer: D

288. if there is only a discrete number of possible outcomes (called categories), the process becomes a .

A. regression

B. classification.

C. modelfree

D. categories

Answer: B

289. Which of the following are supervised learning applications

A. spam detection, pattern detection, natural language processing

B. image classification, real-time visual tracking

C. autonomous car driving, logistic optimization

D. bioinformatics, speech recognition

Answer: A

290. 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.

A. logical

B. classical

C. classification

D. none of above

Answer: D

291. Which of the following sentence is correct?

A. machine learning relates with the study, design and

B. data mining can be defined as the process in which the

C. both a & b

D. none of the above

Answer: C

292. What is Overfitting in Machine learning?

A. when a statistical model describes random error or noise instead of underlying relationship overfitting occurs.

B. robots are programed so that they can perform the task based on data they gather from sensors.

C. while involving the process of learning overfitting occurs.

D. a set of data is used to discover the potentially predictive relationship

Answer: A

293. What is Test set?

A. test set is used to test the accuracy of the hypotheses generated by the learner.

B. it is a set of data is used to discover the potentially predictive relationship.

C. both a & b

D. none of above

Answer: A

294. 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

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 above

Answer: B

---

295. How it's possible to use a different placeholder through the parameter .

A. regression

B. classification

C. random_state

D. missing_values

Answer: D

---

296. 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 .

A. robustscaler

B. dictvectorizer

C. labelbinarizer

D. featurehasher

Answer: A

---

297. scikit-learn also provides a class for per- sample normalization, Normalizer. It can apply to each element of a dataset

A. max, l0 and l1 norms

B. max, l1 and l2 norms

C. max, l2 and l3 norms

D. max, l3 and l4 norms

Answer: B

---

298. There are also many univariate methods that can be used in order to select the best features according to specific criteria based on .

A. f-tests and p-values

B. chi-square

C. anova

D. all above

Answer: A

---

299. Which of the following selects only a subset of features belonging to a certain percentile

A. selectpercentile

B. featurehasher

C. selectkbest

D. all above

Answer: A

---

300. performs a PCA with non-linearly separable data sets.

A. sparsepca

B. kernelpca

C. svd

D. none of the mentioned

Answer: B

Machine Learning (ML) MCQs [set-11]

251. True or False: Ensemble learning can only be applied to supervised learning methods.

A. true

B. false

Answer: B

252. 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.

A. true

B. false

Answer: B

253. 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

Answer: A

254. True or False: Ensemble of classifiers may or may not be more accurate than any of its individual model.

A. true

B. false

Answer: A

255. 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

B. no

C. can’t say

Answer: B

256. Generally, an ensemble method works better, if the individual base models have _____? Note: Suppose each individual base models have accuracy greater than 50%.

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

Answer: A

257. 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.

A. bagging

B. boosting

C. a or b

D. none of these

Answer: A

258. 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

A. 0.05

B. 0.06

C. 0.07

D. 0.09

Answer: B

259. 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

A. true

B. false

Answer: A

260. 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

A. 1 and 3

B. 2 and 3

C. 1 and 2

D. all of above

Answer: D

261. How is the model capacity affected with dropout rate (where model capacity means the ability of a neural network to approximate complex functions)?

A. model capacity increases in increase in dropout rate

B. model capacity decreases in increase in dropout rate

C. model capacity is not affected on increase in dropout rate

D. none of these

Answer: B

262. True or False: Dropout is computationally expensive technique w.r.t. bagging

A. true

B. false

Answer: B

263. 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

A. 1-2-3

B. 1-3-4

C. 2-1-3

D. none of above

Answer: D

---

264. 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. both

D. none of above

Answer: C

---

265. Below are the two ensemble models:
1. E1(M1, M2, M3) and
2. E2(M4, M5, M6)
Above, Mx 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

A. e1

B. e2

C. any of e1 and e2

D. none of these

Answer: B

---

266. True or False: In boosting, individual base learners can be parallel.

A. true

B. false

Answer: B

267. 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

A. 1 and 2

B. 2 and 3

C. 1 and 3

D. all of these

Answer: C

268. 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

A. you will have only k features after the first stage

B. you will have only m features after the first stage

C. you will have k+m features after the first stage

D. you will have k*n features after the first stage

Answer: B

269. Which of the following is the difference between stacking and blending?

A. stacking has less stable cv compared to blending

B. in blending, you create out of fold prediction

C. stacking is simpler than blending

D. none of these

Answer: D

270. 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

A. 1 and 2

B. 2 and 3

C. 1 and 3

D. all of above

Answer: A

---

271. Q25. Which of the following are advantages of stacking?
1) More robust model
2) better prediction
3) Lower time of execution

A. 1 and 2

B. 2 and 3

C. 1 and 3

D. all of the above

Answer: A

---

272. 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

A. 1 and 2

B. 2 and 3

C. 1 and 3

D. all of above

Answer: C

---

273. 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

A. 1 and 3

B. 2 and 3

C. 1 and 2

D. 1, 2 and 3

Answer: D

---

274. 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

Answer: C

---

275. 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

Answer: D

Machine Learning (ML) MCQs [set-10]

226. Skewness of Normal distribution is _____

A. negative

B. positive

C. 0

D. undefined

Answer: C

227. 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 linear relationship

Answer: C

228. 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?

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. c1: (4,4), c2: (3,3), c3: (7,7)

Answer: D

229. In Naive Bayes equation P(C / X)= (P(X / C) *P(C)) / P(X) which part considers "likelihood"?

A. p(x/c)

B. p(c/x)

C. p(c)

D. p(x)

Answer: A

230. 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, 2 and 3

D. 1 and 2

Answer: D

231. 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

Answer: A

232. Which of the following is an application of NN (Neural Network)?

A. sales forecasting

B. data validation

C. risk management

D. all of the mentioned

Answer: D

233. Neural Networks are complex _____ with many parameters.

A. linear functions

B. nonlinear functions

C. discrete functions

D. exponential functions

Answer: A

234. 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

Answer: C

---

235. Which one of the following is not a major strength of the neural network approach?

A. neural network learning algorithms are guaranteed to converge to an optimal solution

B. neural networks work well with datasets containing noisy data

C. neural networks can be used for both supervised learning and unsupervised clustering

D. neural networks can be used for applications that require a time element to be included in the data

Answer: A

---

236. The network that involves backward links from output to the input and hidden layers is called

A. self organizing maps

B. perceptrons

C. recurrent neural network

D. multi layered perceptron

Answer: C

---

237. 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

A. 1

B. 2

C. 3&4

D. 1,2,3&4

Answer: D

238. 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. a

B. b

C. c

D. b&c

Answer: C

239. 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?

A. ??bagging

B. boosting

C. stacking

D. randomization

Answer: A

240. 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

A. 1, 4, 3, 2

B. 3, 1, 2, 4

C. 4, 3, 2, 1

D. 1, 2, 3, 4

Answer: A

241. In which neural net architecture, does weight sharing occur?

A. recurrent neural network

B. convolutional neural network

C. . fully connected neural network

D. both a and b

Answer: D

242. 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

A. 1 and 2

B. 2 and 3

C. 1 and 3

D. 1,2 and 3

Answer: C

243. Given above is a description of a neural network. When does a neural network model become a deep learning model?

A. when you add more hidden layers and increase depth of neural network

B. when there is higher dimensionality of data

C. when the problem is an image recognition problem

D. when there is lower dimensionality of data

Answer: A

244. 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

A. 1, 2, 3, 4, 5

B. 4, 3, 1, 5, 2

C. 3, 2, 1, 5, 4

D. 5, 4, 3, 2, 1

Answer: B

245. 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?

A. 238

B. 76

C. 248

D. 348

Answer: D

246. Increase in size of a convolutional kernel would necessarily increase the performance of a convolutional network.

A. true

B. false

Answer: B

247. The F-test

A. an omnibus test

B. considers the reduction in error when moving from the complete model to the reduced model

C. considers the reduction in error when moving from the reduced model to the complete model

D. can only be conceptualized as a reduction in error

Answer: C

248. 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

Answer: D

249. 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

Answer: C

---

250. 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

Answer: D

Machine Learning (ML) MCQs [set-9]

201. 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:

A. k-means clustering

B. conceptual clustering

C. expectation maximization

D. agglomerative clustering

Answer: A

202. Which one of the following is the main reason for pruning a Decision Tree?

A. to save computing time during testing

B. to save space for storing the decision tree

C. to make the training set error smaller

D. to avoid overfitting the training set

Answer: D

203. 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?

A. your decision trees are too shallow.

B. you need to increase the learning rate.

C. you are overfitting.

D. incorrect data

Answer: A

204. The K-means algorithm:

A. requires the dimension of the feature space to be no bigger than the number of samples

B. has the smallest value of the objective function when k = 1

C. minimizes the within class variance for a given number of clusters

D. converges to the global optimum if and only if the initial means are chosen as some of the samples themselves

Answer: C

205. 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

A. 1 and 2

B. 1 and 3

C. 2 and 3

D. 1, 2 and 3

Answer: D

206. 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

A. 1 and 2

B. 2 and 3

C. 2 and 4

D. 1, 2 and 4

Answer: D

207. Hierarchical clustering is slower than non-hierarchical clustering?

A. true

B. false

C. depends on data

D. cannot say

Answer: A

208. High entropy means that the partitions in classification are

A. pure

B. not pure

C. useful

D. useless

Answer: B

209. 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?

A. decision trees

B. density-based clustering

C. model-based clustering

D. k-means clustering

Answer: B

210. The main disadvantage of maximum likelihood methods is that they are _____

A. mathematically less folded

B. mathematically less complex

C. mathematically less complex

D. computationally intense

Answer: D

211. The maximum likelihood method can be used to explore relationships among more diverse sequences, conditions that are not well handled by maximum parsimony methods.

A. true

B. false

C. -

D. -

Answer: A

212. Which Statement is not true statement.

A. k-means clustering is a linear clustering algorithm.

B. k-means clustering aims to partition n observations into k clusters

C. k-nearest neighbor is same as k-means

D. k-means is sensitive to outlier

Answer: C

213. what is Feature scaling done before applying K-Mean algorithm?

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

Answer: A

214. With Bayes theorem the probability of hypothesis HÂ¾ specified by P(H) Â¾ is referred to as

A. a conditional probability

B. an a priori probability

C. a bidirectional probability

D. a posterior probability

Answer: B

215. 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

A. 0.0398

B. 0.0389

C. 0.0368

D. 0.0396

Answer: D

216. What is the naïve assumption in a Naïve Bayes Classifier.

A. all the classes are independent of each other

B. all the features of a class are independent of each other

C. the most probable feature for a class is the most important feature to be cinsidered for classification

D. all the features of a class are conditionally dependent on each other

Answer: D

217. 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 sereis is 0.12. what is the probability that a person doesn't like to watch either?

A. 0.32

B. 0.2

C. 0.44

D. 0.56

Answer: A

218. What is the actual number of independent parameters which need to be estimated in P dimensional Gaussian distribution model?

A. p

B. 2p

C. p(p+1)/2

D. p(p+3)/2

Answer: D

219. 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.

A. 1 is true, 2 is false

B. 1 is false, 2 is true

C. 1 is true, 2 is true

D. 1 is false, 2 is false

Answer: C

220. Which of the following quantities are minimized directly or indirectly during parameter estimation in Gaussian distribution Model?

A. negative log-likelihood

B. log-liklihood

C. cross entropy

D. residual sum of square

Answer: A

221. 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.

A. 0

B. 1

C. 0.1

D. 0.9

Answer: B

222. Given a rule of the form IF X THEN Y, rule confidence is defined as the conditional probability that Select one:

A. y is false when x is known to be false.

B. y is true when x is known to be true.

C. x is true when y is known to be true

D. x is false when y is known to be false.

Answer: B

223. 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

Answer: B

224. 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

225. How many terms are required for building a bayes model?

A. 1

B. 2

C. 3

D. 4

Answer: C

Machine Learning (ML) MCQs [set-8]

176. In Apriori algorithm, if 1 item-sets are 100, then the number of candidate 2 item-sets are

A. 100

B. 200

C. 4950

D. 5000

Answer: C

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

A. are better able to deal with missing and noisy data

B. typically assume an underlying distribution for the data

C. have trouble with large-sized datasets

D. are not able to explain their behavior

Answer: A

178. 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 donâ€™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

A. 0.0368

B. 0.0396

C. 0.0389

D. 0.0398

Answer: B

179. What is the final resultant cluster size in Divisive algorithm, which is one of the hierarchical clustering approaches?

A. zero

B. three

C. singleton

D. two

Answer: C

180. Given a frequent itemset L, If |L| = k, then there are

A. 2k – 1 candidate association rules

B. 2k candidate association rules

C. 2k – 2 candidate association rules

D. 2k -2 candidate association rules

Answer: C

181. Which Statement is not true statement.

A. k-means clustering is a linear clustering algorithm.

B. k-means clustering aims to partition n observations into k clusters

C. k-nearest neighbor is same as k-means

D. k-means is sensitive to outlier

Answer: B

182. 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

A. 1 and 2

B. 2 and 3

C. 2 and 4

D. 1, 2 and 4

Answer: C

183. 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 like 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 above

Answer: D

184. What are two steps of tree pruning work?

A. pessimistic pruning and optimistic pruning

B. postpruning and prepruning

C. cost complexity pruning and time complexity pruning

D. none of the options

Answer: B

185. 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".

A. 0.4

B. 0.6

C. 0.8

D. 0.42

Answer: D

186. 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??

D. not useful in ml algorithm

Answer: C

187. How to select best hyperparameters in tree based models?

A. measure performance over training data

B. measure performance over validation data

C. both of these

D. random selection of hyper parameters

Answer: B

188. 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

A. 1 and 3

B. 1 and 2

C. 2 and 3

D. 1, 2 and 3

Answer: D

189. What are tree based classifiers?

A. classifiers which form a tree with each attribute at one level

B. classifiers which perform series of condition checking with one attribute at a time

C. both options except none

D. not possible

Answer: C

190. What is gini index?

A. gini index??operates on the categorical target variables

B. it is a measure of purity

C. gini index performs only binary split

D. all (1,2 and 3)

Answer: D

191. Tree/Rule based classification algorithms generate ... rule to perform the classification.

A. if-then.

B. while.

C. do while

D. switch.

Answer: A

192. Decision Tree is

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. both a & b

D. class of instance

Answer: C

193. 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. it can be used for constants

Answer: A

194. 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. wrong client data

Answer: A

195. hich 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

Answer: A

196. Which statement is true about the K-Means algorithm? Select one:

A. the output attribute must be cateogrical.

B. all attribute values must be categorical.

C. all attributes must be numeric

D. attribute values may be either categorical or numeric

Answer: C

197. 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.

A. 1, 3 and 4

B. 1, 2 and 3

C. 1, 2 and 4

D. 1,2,3,4

Answer: D

198. 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. 1,2 and 3

Answer: D

199. 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

A. 1 and 2

B. 2 and 3

C. 1, 2, and 3??

D. 1 and 3

Answer: C

200. How will you counter over-fitting in decision tree?

A. by pruning the longer rules

B. by creating new rules

C. both by pruning the longer rules' and ' by creating new rules'

D. over-fitting is not possible

Answer: A

Machine Learning (ML) MCQs [set-7]

151. The number of iterations in apriori _____ Select one: a. b. c. d.

A. increases with the size of the data

B. decreases with the increase in size of the data

C. increases with the size of the maximum frequent set

D. decreases with increase in size of the maximum frequent set

Answer: C

152. Frequent item sets is

A. superset of only closed frequent item sets

B. superset of only maximal frequent item sets

C. subset of maximal frequent item sets

D. superset of both closed frequent item sets and maximal frequent item sets

Answer: D

153. A good clustering method will produce high quality clusters with

A. high inter class similarity

B. low intra class similarity

C. high intra class similarity

D. no inter class similarity

Answer: C

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

A. both techniques build models whose output is determined by a linear sum of weighted input attribute values

B. the output of both models is a categorical attribute value

C. both models require numeric attributes to range between 0 and 1

D. both models require input attributes to be numeric

Answer: D

155. Which Association Rule would you prefer

A. high support and medium confidence

B. high support and low confidence

C. low support and high confidence

D. low support and low confidence

Answer: C

156. In a Rule based classifier, If there is a rule for each combination of attribute values, what do you called that rule set R

A. exhaustive

B. inclusive

C. comprehensive

D. mutually exclusive

Answer: A

157. The apriori property means

A. if a set cannot pass a test, its supersets will also fail the same test

B. to decrease the efficiency, do level-wise generation of frequent item sets

C. to improve the efficiency, do level-wise generation of frequent item sets d.

D. if a set can pass a test, its supersets will fail the same test

Answer: A

158. If an item set 'XYZ' is a frequent item set, then all subsets of that frequent item set are

A. undefined

B. not frequent

C. frequent

D. can not say

Answer: C

159. Clustering is _____ and is example of _____learning

A. predictive and supervised

B. predictive and unsupervised

C. descriptive and supervised

D. descriptive and unsupervised

Answer: D

160. To determine association rules from frequent item sets

A. only minimum confidence needed

B. neither support not confidence needed

C. both minimum support and confidence are needed

D. minimum support is needed

Answer: C

161. If {A,B,C,D} is a frequent itemset, candidate rules which is not possible is

A. c –> a

B. d –>abcd

C. a –> bc

D. b –> adc

Answer: B

162. Which Association Rule would you prefer

A. high support and low confidence

B. low support and high confidence

C. low support and low confidence

D. high support and medium confidence

Answer: B

163. 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

A. conceptual clustering

B. k-means clustering

C. expectation maximization

D. agglomerative clustering

Answer: B

164. Classification rules are extracted from _____

A. decision tree

B. root node

C. branches

D. siblings

Answer: A

165. What does K refers in the K-Means algorithm which is a non-hierarchical clustering approach?

A. complexity

B. fixed value

C. no of iterations

D. number of clusters

Answer: D

---

166. How will you counter over-fitting in decision tree?

A. by pruning the longer rules

B. by creating new rules

C. both by pruning the longer rules' and ' by creating new rules'

D. none of the options

Answer: A

---

167. What are two steps of tree pruning work?

A. pessimistic pruning and optimistic pruning

B. postpruning and prepruning

C. cost complexity pruning and time complexity pruning

D. none of the options

Answer: B

---

168. Which of the following sentences are true?

A. in pre-pruning a tree is \pruned\ by halting its construction early

B. a pruning set of class labelled tuples is used to estimate cost complexity

C. the best pruned tree is the one that minimizes the number of encoding bits

D. all of the above

Answer: D

---

169. 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?

A. accuracy of the decision tree model on the given data set

B. f1 measure of the decision tree model on the given data set

C. 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

D. comprehensibility of the decision tree model, measured in terms of the size of the corresponding rule set

Answer: C

170. 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. a and b

B. a and d

C. b, c and d

D. all of the above

Answer: C

171. 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. a and b

B. a and d

C. b, c and d

D. all of the above

Answer: A

172. Which among the following statements best describes our approach to learning decision trees

A. identify the best partition of the input space and response per partition to minimise sum of squares error

B. identify the best approximation of the above by the greedy approach (to identifying the partitions)

C. identify the model which gives the best performance using the greedy approximation (option (b)) with the smallest partition scheme

D. identify the model which gives performance close to the best greedy approximation performance (option (b)) with the smallest partition scheme

Answer: D

173. 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?

A. 10.8, 13.33, 14.48

B. 10.8, 13.33, 12.06

C. 7.2, 10, 8.8

D. 7.2, 10, 8.6

Answer: C

174. 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. a and b

B. a and d

C. b, c and d

D. all of the above

Answer: D

175. Time Complexity of k-means is given by

A. o(mn)

B. o(tkn)

C. o(kn)

D. o(t2kn)

Answer: B

Machine Learning (ML) MCQs [set-6]

126. How can we best represent ‘support’ for the following association rule: “If X and Y, then Z”.

A. {x,y}/(total number of transactions)

B. {z}/(total number of transactions)

C. {z}/{x,y}

D. {x,y,z}/(total number of transactions)

Answer: C

127. Choose the correct statement with respect to ‘confidence’ metric in association rules

A. 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.

B. a high value of confidence suggests a weak association rule

C. 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.

D. confidence is not measured in terms of (estimated) conditional probability.

Answer: A

128. What are tree based classifiers?

A. classifiers which form a tree with each attribute at one level

B. classifiers which perform series of condition checking with one attribute at a time

C. both options except none

D. none of the options

Answer: C

129. What is gini index?

A. it is a type of index structure

B. it is a measure of purity

C. both options except none

D. none of the options

Answer: B

130. 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 approact used by ID3 is greedy

A. a and b

B. a and d

C. b, c and d

D. all of the above

Answer: C

---

131. Multivariate split is where the partitioning of tuples is based on a combination of attributes rather than on a single attribute.

A. true

B. false

Answer: A

---

132. Gain ratio tends to prefer unbalanced splits in which one partition is much smaller than the other

A. true

B. false

Answer: A

---

133. The gini index is not biased towards multivalued attributed.

A. true

B. false

Answer: B

---

134. Gini index does not favour equal sized partitions.

A. true

B. false

Answer: B

---

135. When the number of classes is large Gini index is not a good choice.

A. true

B. false

Answer: A

---

136. Attribute selection measures are also known as splitting rules.

A. true

B. false

Answer: A

137. his clustering approach initially assumes that each data instance represents a single cluster.

A. expectation maximization

B. k-means clustering

C. agglomerative clustering

D. conceptual clustering

Answer: C

138. Which statement is true about the K-Means algorithm?

A. the output attribute must be cateogrical

B. all attribute values must be categorical

C. all attributes must be numeric

D. attribute values may be either categorical or numeric

Answer: C

139. KDD represents extraction of

A. data

B. knowledge

C. rules

D. model

Answer: B

140. The most general form of distance is

A. manhattan

B. eucledian

C. mean

D. minkowski

Answer: B

141. Which of the following algorithm comes under the classification

A. apriori

B. brute force

C. dbscan

D. k-nearest neighbor

Answer: D

142. Hierarchical agglomerative clustering is typically visualized as?

A. dendrogram

B. binary trees

C. block diagram

D. graph

Answer: A

143. The _____ step eliminates the extensions of (k-1)-itemsets which are not found to be frequent,from being considered for counting support

A. partitioning

B. candidate generation

C. itemset eliminations

D. pruning

Answer: D

144. The distance between two points calculated using Pythagoras theorem is

A. supremum distance

B. eucledian distance

C. linear distance

D. manhattan distance

Answer: B

145. Which one of these is not a tree based learner?

A. cart

B. id3

C. bayesian classifier

D. random forest

Answer: C

146. Which one of these is a tree based learner?

A. rule based

B. bayesian belief network

C. bayesian classifier

D. random forest

Answer: D

147. What is the approach of basic algorithm for decision tree induction?

A. greedy

B. top down

C. procedural

D. step by step

Answer: A

148. 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

Answer: A

149. 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

150. 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

A. k-means clustering

B. conceptual clustering

C. expectation maximization

D. agglomerative clustering

Answer: A

40 Questions to test a data scientist on Machine Learning [Solution: SkillPower – Machine Learning, DataFest 2017]

CAREER | INTERMEDIATE | INTERVIEWS | MACHINE LEARNING | SKILLTEST

Introduction

Machine Learning is one of the most sought after skills these days. If you are a data scientist, then you need to be good at Machine Learning – no two ways about it. As part of DataFest 2017, we organized various skill tests so that data scientists can assess themselves on these critical skills. These tests included Machine Learning, Deep Learning, Time Series problems and Probability. This article will lay out the solutions to the machine learning skill test. If you missed out on any of the above skill tests, you can still check out the questions and answers through the articles linked above.

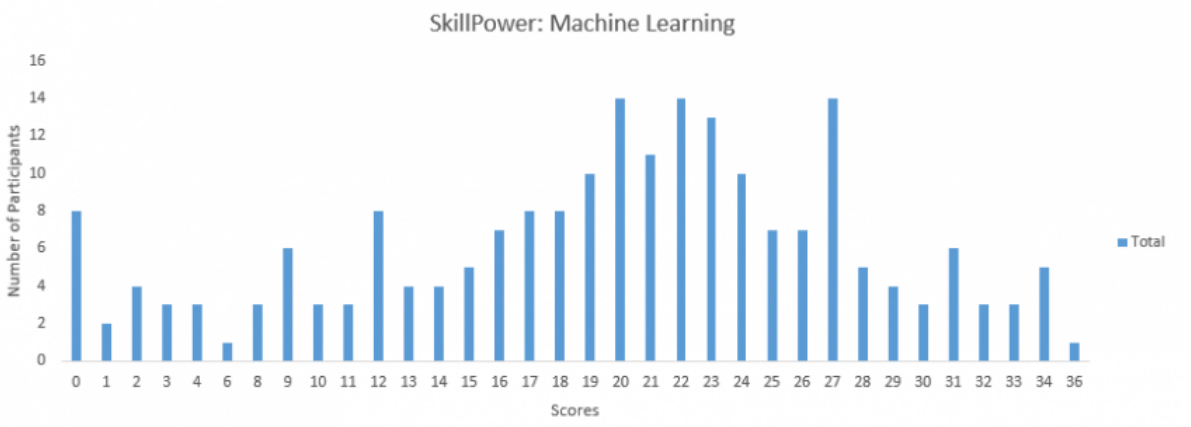
In Machine Learning skill test, more than 1350 people registered for the test. The test was designed to test your conceptual knowledge in machine learning and make you industry ready. If you missed on the real time test, you can still read this article to find out how you could have answered correctly.

Here are the leaderboard rankings for all the participants.

These questions, along with hundreds of others, are part of our 'Ace Data Science Interviews' course. It's a comprehensive guide, with tons of resources, to crack data science interviews and land your dream role! And if you're just starting your data science journey, then check out our most popular course – 'Introduction to Data Science'!

Overall Scores

Below are the distribution scores, they will help you evaluate your performance.



You can access the final scores here. More than 210 people participated in the skill test and the highest score obtained was 36. Here are a few statistics about the distribution.

Mean Score: 19.36

Median Score: 21

Mode Score: 27

Useful Resources

Machine Learning basics for a newbie

Essentials of Machine Learning Algorithms (with Python and R Codes)

Deep Learning vs. Machine Learning – the essential differences you need to know!

Introduction to Data Science Course

Ace Data Science Interviews Course

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)

$Y=X2$. 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)

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?

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 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.enew 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 (t1, t2, t3, …….. tn) and a new observation (q1). The black box outputs the nearest neighbor of q1 (say ti) and its corresponding class label ci.

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 (q1) in the black box algorithm so this algorithm would return a nearest observation and its class.

In second step, you through it out nearest observation from train data and again input the observation (q1). 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 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?

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.

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 x2, we can efficiently perform classification. So at first w1 will become 0. As regularization parameter increases more, w2 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.

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

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) C1 = C2 = C3

B) C1 > C2 > C3

C) C1 < C2 < C3

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. Read more here.

End Notes

I hope you enjoyed the questions and were able to test your knowledge about machine learning. If you have any questions or doubts, feel free to post them below.

Check out all the upcoming events **here**.

Article Url - https://www.analyticsvidhya.com/blog/2017/04/40-questions-test-data-scientist-machine-learning-solution-skillpower-machine-learning-datafest-2017/

Ankit Gupta

Ankit is currently working as a data scientist at UBS who has solved complex data mining problems in many domains. He is eager to learn more about data science and machine learning algorithms.

Machine Learning (ML) MCQs [set-5]

101. MLE estimates are often undesirable because

A. they are biased

B. they have high variance

C. they are not consistent estimators

D. none of the above

Answer: B

102. 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

Answer: A

103. Neural networks

A. optimize a convex cost function

B. always output values between 0 and 1

C. can be used for regression as well as classification

D. all of the above

Answer: C

104. Linear Regression is a _____ machine learning algorithm.

A. supervised

B. unsupervised

C. semi-supervised

D. can\t say

Answer: A

105. Which of the following methods/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

106. 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

Answer: B

107. 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

Answer: A

108. 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

Answer: D

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

A. quadratic

B. inverse

C. linear

D. reciprocal

Answer: C

110. In the regression equation Y = 75.65 + 0.50X, the intercept is

A. 0.5

B. 75.65

C. 1

D. indeterminable

Answer: B

111. 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.

A. binary classification

B. multilabel classification

C. simple linear regression

D. multiple linear regression

Answer: D

112. 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?

A. you will add more features

B. you will remove some features

C. all of the above

D. none of the above

Answer: A

113. 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?

A. increase

B. decrease

C. remain constant

D. can't say

Answer: D

114. 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?

A. bias increases and variance increases

B. bias decreases and variance increases

C. bias decreases and variance decreases

D. bias increases and variance decreases

Answer: D

---

115. 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

A. (i) and (ii)

B. (ii) and (iii)

C. (iii) and (iv)

D. none of these

Answer: B

---

116. 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

Answer: D

---

117. Suppose that we have N independent variables (X1,X2… Xn) 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 X1) with Y is 0.95.

A. relation between the x1 and y is weak

B. relation between the x1 and y is strong

C. relation between the x1 and y is neutral

D. correlation can't judge the relationship

Answer: B

---

118. 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

Answer: C

---

119. Which of the following statements are true for a design matrix X ? Rn×d with d > n? (The rows are n sample points and the columns represent d features.)

A. least-squares linear regression computes the weights w = (xtx)?1 xty

B. the sample points are linearly separable

C. x has exactly d ? n eigenvectors with eigenvalue zero

D. at least one principal component direction is orthogonal to a hyperplane that contains all the sample points

Answer: D

---

120. 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

Answer: C

---

121. 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?

A. you will add more features

B. you will remove some features

C. all of the above

D. none of the above

Answer: A

122. 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

Answer: B

123. 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

A. (i) and (ii)

B. (ii) and (iii)

C. (iii) and (iv)

D. none of these

Answer: B

124. Which of the following statements are true for a design matrix X ? Rn×d with d > n? (The rows are n sample points and the columns represent d features.)

A. least-squares linear regression computes the weights w = (xtx)?1 xty

B. the sample points are linearly separable

C. x has exactly d ? n eigenvectors with eigenvalue zero

D. at least one principal component direction is orthogonal to a hyperplane that contains all the sample points

Answer: D

125. Problem in multi regression is ?

A. multicollinearity

B. overfitting

C. both multicollinearity & overfitting

D. underfitting

Answer: C

Machine Learning (ML) MCQs [set-4]

76. Suppose we train a hard-margin linear SVM on n > 100 data points in R2, 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)?

A. 2

B. 3

C. n

D. n+1

Answer: D

77. 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 ? s2

B. s1 may not be a subset of s2

C. w1 = w2

D. all of the above

Answer: B

78. Which statement about outliers is true?

A. outliers should be part of the training dataset but should not be present in the test data

B. outliers should be identified and removed from a dataset

C. the nature of the problem determines how outliers are used

D. outliers should be part of the test dataset but should not be present in the training data

Answer: C

79. If TP=9 FP=6 FN=26 TN=70 then Error rate will be

A. 45 percentage

B. 99 percentage

C. 28 percentage

D. 20 perentage

Answer: C

80. 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

81. he minimum time complexity for training an SVM is O(n2). 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

82. Perceptron Classifier is

A. unsupervised learning algorithm

B. semi-supervised learning algorithm

C. supervised learning algorithm

D. soft margin classifier

Answer: C

83. Type of dataset available in Supervised Learning is

A. unlabeled dataset

B. labeled dataset

C. csv file

D. excel file

Answer: B

84. which among the following is the most appropriate kernel that can be used with SVM to separate the classes.

A. linear kernel

B. gaussian rbf kernel

C. polynomial kernel

D. option 1 and option 3

Answer: B

85. The SVMs 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. option 1 and option 2

Answer: C

86. 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. opton 1 and option 2

Answer: B

87. What is the precision value for following confusion matrix of binary classification?

A. 0.91

B. 0.09

C. 0.9

D. 0.95

Answer: B

88. Which of the following are components of generalization Error?

A. bias

B. vaiance

C. both of them

D. none of them

Answer: C

89. Which of the following is not a kernel method in SVM?

A. linear kernel

B. polynomial kernel

C. rbf kernel

D. nonlinear kernel

Answer: A

90. During the treatement 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 given chemotherapy?

A. precision

B. recall

C. call

D. score

Answer: A

91. Which one of the following is suitable? 1. When the hypothsis space is richer, overfitting is more likely. 2. when the feature space is larger , overfitting is more likely.

A. true, false

B. false, true

C. true,true

D. false,false

Answer: C

92. Which of the following is a categorical data?

A. branch of bank

B. expenditure in rupees

C. prize of house

D. weight of a person

Answer: A

93. The soft margin SVM is more preferred than the hard-margin SVM when-

A. the data is linearly seperable

B. the data is noisy and contains overlapping points

C. the data is not noisy and linearly seperable

D. the data is noisy and linearly seperable

Answer: B

94. In SVM which has quadratic kernel function of polynomial degree 2 that has slack variable C as one hyper paramenter. What would happen if we use very large value for C

A. we can still classify the data correctly for given setting of hyper parameter c

B. we can not classify the data correctly for given setting of hyper parameter c

C. we can not classify the data at all

D. data can be classified correctly without any impact of c

Answer: A

95. In SVM, RBF kernel with appropriate parameters to perform binary classification where the data is non-linearly seperable. In this scenario

A. the decision boundry in the transformed feature space in non-linear

B. the decision boundry in the transformed feature space in linear

C. the decision boundry in the original feature space in not considered

D. the decision boundry in the original feature space in linear

Answer: B

96. 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

A. 1 is true, 2 is false

B. 1 is false, 2 is true

C. 1 is true, 2 is true

D. 1 is false, 2 is false

Answer: C

97. 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]

A. 0.75

B. 0.97

C. 0.95

D. 0.85

Answer: B

98. Which of the following method is used for multiclass classification?

A. one vs rest

B. loocv

C. all vs one

D. one vs another

Answer: A

---

99. 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 sereis is 0.12. what is the probability that a person doesn't like to watch either?

A. 0.32

B. 0.2

C. 0.44

D. 0.56

Answer: C

---

100. 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?

A. 12

B. 24

C. 48

D. 72

Answer: D

Machine Learning (ML) MCQs [set-3]

51. In multiclass classification number of classes must be

A. less than two

B. equals to two

C. greater than two

D. option 1 and option 2

Answer: C

52. Which of the following can only be used when training data are linearlyseparable?

A. linear hard-margin svm

B. linear logistic regression

C. linear soft margin svm

D. the centroid method

Answer: A

53. Impact of high variance on the training set ?

A. overfitting

B. underfitting

C. both underfitting & overfitting

D. depents upon the dataset

Answer: A

54. 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. both 1 & 2

D. none of the above

Answer: A

55. 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: A

56. What are support vectors?

A. all the examples that have a non-zero weight ??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

Answer: C

57. 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

Answer: A

58. What is the purpose of the Kernel Trick?

A. to transform the data from nonlinearly separable to linearly separable

B. to transform the problem from regression to classification

C. to transform the problem from supervised to unsupervised learning.

D. all of the above

Answer: A

59. Which of the following can only be used when training data are linearlyseparable?

A. linear hard-margin svm

B. linear logistic regression

C. linear soft margin svm

D. parzen windows

Answer: A

60. The firing rate of a neuron

A. determines how strongly the dendrites of the neuron stimulate axons of neighboring neurons

B. is more analogous to the output of a unit in a neural net than the output voltage of the neuron

C. only changes very slowly, taking a period of several seconds to make large adjustments

D. can sometimes exceed 30,000 action potentials per second

Answer: B

61. 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

Answer: D

62. 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

63. The kernel trick

A. can be applied to every classification algorithm

B. is commonly used for dimensionality reduction

C. changes ridge regression so we solve a d ?? d linear system instead of an n ?? n system, given n sample points with d features

D. exploits the fact that in many learning algorithms, the weights can be written as a linear combination of input points

Answer: D

64. How does the bias-variance decomposition of a ridge regression estimator compare with that of ordinary least squares regression?

A. ridge has larger bias, larger variance

B. ridge has smaller bias, larger variance

C. ridge has larger bias, smaller variance

D. ridge has smaller bias, smaller variance

Answer: C

65. 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

66. 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. t is a model trained using unsupervised learning. it can be used for classification but not for regression.

Answer: C

67. Which of the following can help to reduce overfitting in an SVM classifier?

A. use of slack variables

B. high-degree polynomial features

C. normalizing the data

D. setting a very low learning rate

Answer: A

68. 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?

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

69. 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

70. 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

Answer: B

71. 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 traning data in each traning set

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

72. 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

Answer: B

73. 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

74. 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.

Answer: C

75. 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

Answer: B

Machine Learning (ML) MCQs [set-2]

26. Like the probabilistic view, the _____ view allows us to associate a probability of membership with each classification.

A. exampler

B. deductive

C. classical

D. inductive

Answer: D

27. Database query is used to uncover this type of knowledge.

A. deep

B. hidden

C. shallow

D. multidimensional

Answer: D

28. A person trained to interact with a human expert in order to capture their knowledge.

A. knowledge programmer

B. knowledge developer r

C. knowledge engineer

D. knowledge extractor

Answer: D

29. 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

Answer: C

30. 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

Answer: A

31. Which learning Requires Self Assessment to identify patterns within data?

A. unsupervised learning

B. supervised learning

C. semisupervised learning

D. reinforced learning

Answer: A

32. 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.

A. both are true

B. 1 is true and 2 is false

C. both are false

D. 1 is false and 2 is true

Answer: B

33. The "curse of dimensionality" referes

A. all the problems that arise when working with data in the higher dimensions, that did not exist in the lower dimensions.

B. all the problems that arise when working with data in the lower dimensions, that did not exist in the higher dimensions.

C. all the problems that arise when working with data in the lower dimensions, that did not exist in the lower dimensions.

D. all the problems that arise when working with data in the higher dimensions, that did not exist in the higher dimensions.

Answer: A

34. In simple term, machine learning is

A. training based on historical data

B. prediction to answer a query

C. both a and b??

D. automization of complex tasks

Answer: C

35. 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

Answer: A

36. Following are the descriptive models

A. clustering

B. classification

C. association rule

D. both a and c

Answer: D

37. Different learning methods does not include?

A. memorization

B. analogy

C. deduction

D. introduction

Answer: D

38. A measurable property or parameter of the data-set is

A. training data

B. feature

C. test data

D. validation data

Answer: B

39. Feature can be used as a

A. binary split

B. predictor

C. both a and b??

D. none of the above

Answer: C

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

A. true

B. false

Answer: A

41. 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 method2. It searches for the directions that data have the largest variance3. Maximum number of principal components <= number of features4. All principal components are orthogonal to each other

A. 1 & 2

B. 2 & 3

C. 3 & 4

D. all of the above

Answer: D

42. 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 varinace 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

Answer: A

43. Which of the folllowing 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

Answer: B

44. Prediction is

A. the result of application of specific theory or rule in a specific case

B. discipline in statistics used to find projections in multidimensional data

C. value entered in database by expert

D. independent of data

Answer: A

45. You are given sesimic 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

Answer: A

46. 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

Answer: C

47. 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?

A. variable f1 is an example of nominal variable

B. variable f1 is an example of ordinal variable

C. it doesn\t belong to any of the mentioned categories

D. it belongs to both ordinal and nominal category

Answer: B

48. What can be major issue in Leave-One-Out-Cross-Validation(LOOCV)?

A. low variance

B. high variance

C. faster runtime compared to k-fold cross validation

D. slower runtime compared to normal validation

Answer: B

49. 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?

A. classification

B. regression

C. kmeans algorithm

D. reinforcement learning

Answer: D

50. Support Vector Machine is

A. logical model

B. proabilistic model

C. geometric model

D. none of the above

Answer: C

Machine Learning (ML) MCQs [set-1]

1. Application of machine learning methods to large databases is called

A. data mining.

B. artificial intelligence

C. big data computing

D. internet of things

Answer: A

2. 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

Answer: B

3. In what type of learning labelled training data is used

A. unsupervised learning

B. supervised learning

C. reinforcement learning

D. active learning

Answer: B

4. In following type of feature selection method we start with empty feature set

A. forward feature selection

B. backword feature selection

C. both a and b??

D. none of the above

Answer: A

5. In PCA the number of input dimensiona are equal to principal components

A. true

B. false

Answer: A

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

A. true

B. false

Answer: A

7. Which of the following is the best machine learning method?

A. scalable

B. accuracy

C. fast

D. all of the above

Answer: D

8. What characterize unlabeled examples in machine learning

A. there is no prior knowledge

B. there is no confusing knowledge

C. there is prior knowledge

D. there is plenty of confusing knowledge

Answer: D

9. What does dimensionality reduction reduce?

A. stochastics

B. collinerity

C. performance

D. entropy

Answer: B

10. Data used to build a data mining model.

A. training data

B. validation data

C. test data

D. hidden data

Answer: A

11. The problem of finding hidden structure in unlabeled data is called…

A. supervised learning

B. unsupervised learning

C. reinforcement learning

D. none of the above

Answer: B

12. Of the Following Examples, Which would you address using an supervised learning Algorithm?

A. given email labeled as spam or not spam, learn a spam filter

B. given a set of news articles found on the web, group them into set of articles about the same story.

C. given a database of customer data, automatically discover market segments and group customers into different market segments.

D. find the patterns in market basket analysis

Answer: A

13. Dimensionality Reduction Algorithms are one of the possible ways to reduce the computation time required to build a model

A. true

B. false

Answer: A

14. 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

A. supervised learning

B. unsupervised learning

C. semisupervised learning

D. reinforcement learning

Answer: A

15. 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 the above

Answer: C

16. Following are the types of supervised learning

A. classification

B. regression

C. subgroup discovery

D. all of the above

Answer: D

17. Type of matrix decomposition model is

A. descriptive model

B. predictive model

C. logical model

D. none of the above

Answer: A

18. Following is powerful distance metrics used by Geometric model

A. euclidean distance

B. manhattan distance

C. both a and b??

D. square distance

Answer: C

19. The output of training process in machine learning is

A. machine learning model

B. machine learning algorithm

C. null

D. accuracy

Answer: A

20. 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

Answer: B

21. PCA is

A. forward feature selection

B. backword feature selection

C. feature extraction

D. all of the above

Answer: C

22. Dimensionality reduction algorithms are one of the possible ways to reduce the computation time required to build a model.

A. true

B. false

Answer: A

23. 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

Answer: A

24. 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

Answer: C

25. What characterize is hyperplance in geometrical model of machine learning?

A. a plane with 1 dimensional fewer than number of input attributes

B. a plane with 2 dimensional fewer than number of input attributes

C. a plane with 1 dimensional more than number of input attributes

D. a plane with 2 dimensional more than number of input attributes

Answer: B