UNIT I

1. What is classification?

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

Ans: Solution A

2. What is regression?

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

Ans: Solution B

3. What is supervised learning?

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

Ans: Solution B

4. What is Unsupervised learning?

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

Ans: Solution A

5. What is Semi-Supervised learning?

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

Ans: Solution D

6. What is Reinforcement learning?

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

Ans: Solution C

7. Sentiment Analysis is an example of:

Regression,

Classification

Clustering

Reinforcement Learning

Options:

A. 1 Only

B. 1 and 2

C. 1 and 3

D. 1, 2 and 4

Ans : Solution D

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

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

Ans : Solution C

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

Ans : Solution A

10. Data used to build a data mining model.
    a) validation data
    b) training data
    c) test data
    d) hidden data

Ans : Solution B

11. Supervised learning and unsupervised clustering both require at least one
    a) hidden attribute.
    b) output attribute.
    c) input attribute.
    d) categorical attribute.

Ans : Solution A

12. Supervised learning differs from unsupervised clustering in that supervised learning requires
    a) at least one input attribute.
    b) input attributes to be categorical.
    c) at least one output attribute.
    d) output attributes to be categorical.

Ans : Solution B

13. A regression model in which more than one independent variable is used to predict the dependent variable is called
    a) a simple linear regression model
    b) a multiple regression models
    c) an independent model
    d) none of the above

Ans : Solution C

14. A term used to describe the case when the independent variables in a multiple regression model are correlated is
    a) Regression
    b) correlation
    c) multicollinearity
    d) none of the above

Ans : Solution C

15. A multiple regression model has the form: y = 2 + 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

Ans : Solution C

16. A multiple regression model has
    a) only one independent variable
    b) more than one dependent variable
    c) more than one independent variable
    d) none of the above

Ans : Solution B

17. A measure of goodness of fit for the estimated regression equation is the
    a) multiple coefficient of determination
    b) mean square due to error
    c) mean square due to regression
    d) none of the above

Ans : Solution C

18. The adjusted multiple coefficient of determination accounts for
    a) the number of dependent variables in the model
    b) the number of independent variables in the model
    c) unusually large predictors
    d) none of the above

Ans : Solution D

19. The multiple coefficient of determination is computed by
    a) dividing SSR by SST
    b) dividing SST by SSR
    c) dividing SST by SSE
    d) none of the above

Ans : Solution C

20. For a multiple regression model, SST = 200 and SSE = 50. The multiple coefficient of determination is
    a) 0.25

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

Ans : Solution B

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

Ans : Solution B

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

Ans : Solution B

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

Ans : Solution C

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

Ans : Solution D

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

Ans : Solution D

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

Ans : Solution A

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

Ans : Solution A

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

Ans : Solution D

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

Ans : Solution B

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

Ans : Solution A

31. Data used to optimize the parameter settings of a supervised learner model.
    a) Training
    b) Test
    c) Verification
    d) Validation

Ans : Solution D

32. Bootstrapping allows us to
    a) choose the same training instance several times.
    b) choose the same test set instance several times.
    c) build models with alternative subsets of the training data several times.
    d) test a model with alternative subsets of the test data several times.

Ans : Solution A

33. The correlation between the number of years an employee has worked for a company and the salary of the employee is 0.75. What can be said about employee salary and years worked?
    a) There is no relationship between salary and years worked.
    b) Individuals that have worked for the company the longest have higher salaries.
    c) Individuals that have worked for the company the longest have lower salaries.
    d) The majority of employees have been with the company a long time.
    e) The majority of employees have been with the company a short period of time.

Ans : Solution B

34. The correlation coefficient for two real-valued attributes is –0.85. What does this value tell you?
    a) The attributes are not linearly related.
    b) As the value of one attribute increases the value of the second attribute also increases.
    c) As the value of one attribute decreases the value of the second attribute increases.
    d) The attributes show a curvilinear relationship.

Ans : Solution C

35. The average squared difference between classifier predicted output and actual output.
    a) mean squared error
    b) root mean squared error
    c) mean absolute error
    d) mean relative error

Ans : Solution A

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

b) Quadratic
c) reciprocal
d) inverse

Ans : Solution A

37. Regression trees are often used to model _____ data.
a) Linear
b) Nonlinear
c) Categorical
d) Symmetrical

Ans : Solution B

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

Ans : Solution C

39. Logistic regression is a _____ regression technique that is used to model data having a _____outcome.
a) linear, numeric
b) linear, binary
c) nonlinear, numeric
d) nonlinear, binary

Ans : Solution D

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

Ans : Solution B

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

Ans : Solution A

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

Ans : Solution B

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

Ans : Solution D

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

Ans : Solution C

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

Ans : Solution C

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

Ans : Solution B

UNIT –II

1.True- False: Over fitting is more likely when you have huge amount of data to train?
A) TRUE
B) FALSE
Ans **Solution: (B)**
With a small training dataset, it's easier to find a hypothesis to fit the training data exactly i.e. over fitting.

2.What is pca.components_ in Sklearn?
Set of all eigen vectors for the projection space
Matrix of principal components
Result of the multiplication matrix
None of the above options
Ans A

3.**Which of the following techniques would perform better for reducing dimensions of a data set?**
A. Removing columns which have too many missing values
B. Removing columns which have high variance in data
C. Removing columns with dissimilar data trends
D. None of these
Ans Solution: **(A)**
If a columns have too many missing values, (say 99%) then we can remove such columns.

4.**It is not necessary to have a target variable for applying dimensionality reduction algorithms.**
A. TRUE
B. FALSE
Ans Solution: **(A)**
LDA is an example of supervised dimensionality reduction algorithm.

5. **PCA can be used for projecting and visualizing data in lower dimensions.**
A. TRUE
B. FALSE
Ans Solution: **(A)**
Sometimes it is very useful to plot the data in lower dimensions. We can take the first 2 principal components and then visualize the data using scatter plot.

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

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

Ans D

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

8. What happens when you get features in lower dimensions using PCA?
The features will still have interpretability
The features will lose interpretability
The features must carry all information present in data
The features may not carry all information present in data
A. 1 and 3
B. 1 and 4
C. 2 and 3
D. 2 and 4
Ans Solution: **(D)**
When you get the features in lower dimensions then you will lose some information of data most of the times and you won't be able to interpret the lower dimension data.

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

Ans Solution: **(D)**
PCA is a deterministic algorithm which doesn't have parameters to initialize and it doesn't have local minima problem like most of the machine learning algorithms has.

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

11. **[True or False] PCA can be used for projecting and visualizing data in lower dimensions.**
A. TRUE
B. FALSE

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

12. **A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college.**
1) Which of the following statement is true in following case?
A) Feature F1 is an example of nominal variable.
B) Feature F1 is an example of ordinal variable.
C) It doesn't belong to any of the above category.
D) Both of these
Solution: (B)
Ordinal variables are the variables which has some order in their categories. For example, grade A should be consider as high grade than grade B.

13. **Which of the following is an example of a deterministic algorithm?**
A) PCA
B) K-Means
C) None of the above
Solution: (A)
A deterministic algorithm is that in which output does not change on different runs. PCA would give the same result if we run again, but not k-means.

UNIT –III

1. Which of the following methods do we use to best fit the data in Logistic Regression?
A) Least Square Error
B) Maximum Likelihood
C) Jaccard distance
D) Both A and B
Ans Solution: B

2. Choose which of the following options is true regarding One-Vs-All method in Logistic Regression.
A) We need to fit n models in n-class classification problem
B) We need to fit n-1 models to classify into n classes
C) We need to fit only 1 model to classify into n classes
D) None of these
Ans **Solution: A**

3. Suppose, You applied a Logistic Regression model on a given data and got a training accuracy X and testing accuracy Y. Now, you want to add a few new features in the same data. Select the option(s) which is/are correct in such a case.
Note: Consider remaining parameters are same.
A) Training accuracy increases
B) Training accuracy increases or remains the same
C) Testing accuracy decreases
D) Testing accuracy increases or remains the same
Ans Solution: A and D
Adding more features to model will increase the training accuracy because model has to consider more data to fit the logistic regression. But testing accuracy increases if feature is found to be significant

4. Which of the following algorithms do we use for Variable Selection?
A) LASSO
B) Ridge
C) Both
D) None of these
Ans Solution: A
In case of lasso we apply a absolute penality, after increasing the penality in lasso some of the coefficient of variables may become zero

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

A) Linear regression is sensitive to outliers
B) Linear regression is not sensitive to outliers
C) Can't say
D) None of these
Ans Solution: (A)
The slope of the regression line will change due to outliers in most of the cases. So Linear Regression is sensitive to outliers.

6. Which of the following methods do we use to find the best fit line for data in Linear Regression?
A) Least Square Error
B) Maximum Likelihood
C) Logarithmic Loss
D) Both A and B
Ans Solution: (A)
In linear regression, we try to minimize the least square errors of the model to identify the line of best fit.

7. Which of the following is true about Residuals?
A) Lower is better
B) Higher is better
C) A or B depend on the situation
D) None of these
Ans Solution: (A)
Residuals refer to the error values of the model. Therefore lower residuals are desired.

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

A) Since the there is a relationship means our model is not good
B) Since the there is a relationship means our model is good
C) Can't say
D) None of these
Ans Solution: (A)
There should not be any relationship between predicted values and residuals. If there exists any relationship between them, it means that the model has not perfectly captured the information in the data.

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

Choose the option which describes bias in best manner.
A) In case of very large x; bias is low
B) In case of very large x; bias is high
C) We can't say about bias
D) None of these
Ans Solution: (B)
If the penalty is very large it means model is less complex, therefore the bias would be high.

10. Which of the following option is true?
A) Linear Regression errors values has to be normally distributed but in case of Logistic Regression it is not the case
B) Logistic Regression errors values has to be normally distributed but in case of Linear Regression it is not the case
C) Both Linear Regression and Logistic Regression error values have to be normally distributed
D) Both Linear Regression and Logistic Regression error values have not to be normally distributed
Ans Solution: A

11. Suppose you have trained a logistic regression classifier and it outputs a new example x with a prediction ho(x) = 0.2. This means
Our estimate for P(y=1 | x)
Our estimate for P(y=0 | x)
Our estimate for P(y=1 | x)
Our estimate for P(y=0 | x)
Ans Solution: B

12. **True-False: Linear Regression is a supervised machine learning algorithm.**
A) TRUE
B) FALSE
Solution: (A)
Yes, Linear regression is a supervised learning algorithm because it uses true labels for training. Supervised learning algorithm should have input variable (x) and an output variable (Y) for each example.

13. **True-False: Linear Regression is mainly used for Regression.**
A) TRUE
B) FALSE
Solution: (A)
Linear Regression has dependent variables that have continuous values.

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

A) TRUE
B) FALSE

Solution: (A)

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

15. **Which of the following methods do we use to find the best fit line for data in Linear Regression?**
A) Least Square Error
B) Maximum Likelihood
C) Logarithmic Loss
D) Both A and B
Solution: (A)
In linear regression, we try to minimize the least square errors of the model to identify the line of best fit.

16. **Which of the following evaluation metrics can be used to evaluate a model while modeling a continuous output variable?**
A) AUC-ROC
B) Accuracy
C) Logloss
D) Mean-Squared-Error
Solution: (D)
Since linear regression gives output as continuous values, so in such case we use mean squared error metric to evaluate the model performance. Remaining options are use in case of a classification problem.

17. **True-False: Lasso Regularization can be used for variable selection in Linear Regression.**
A) TRUE
B) FALSE
Solution: (A)
True, In case of lasso regression we apply absolute penalty which makes some of the coefficients zero.

18. **Which of the following is true about Residuals ?**
A) Lower is better
B) Higher is better

C) A or B depend on the situation
D) None of these
Solution: (A)
Residuals refer to the error values of the model. Therefore lower residuals are desired.

19. Suppose that we have N independent variables (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
Solution: (B)
The absolute value of the correlation coefficient denotes the strength of the relationship. Since absolute correlation is very high it means that the relationship is strong between X1 and Y.

20. **Looking at above two characteristics, which of the following option is the correct for Pearson correlation between V1 and V2?**
If you are given the two variables V1 and V2 and they are following below two characteristics.
1. If V1 increases then V2 also increases
2. If V1 decreases then V2 behavior is unknown
A) Pearson correlation will be close to 1
B) Pearson correlation will be close to -1
C) Pearson correlation will be close to 0
D) None of these

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

21. **Suppose Pearson correlation between V1 and V2 is zero. In such case, is it right to conclude that V1 and V2 do not have any relation between them?**
A) TRUE
B) FALSE
Solution: (B)
Pearson correlation coefficient between 2 variables might be zero even when they have a relationship between them. If the correlation coefficient is zero, it just means that that they don't move together. We can take examples like y=|x| or y=x^2.

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

A) TRUE

B) FALSE

Solution: (B)

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

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

1. We don't have to choose the learning rate
2. It becomes slow when number of features is very large
3. Thers is no need to iterate

A) 1 and 2

B) 1 and 3

C) 2 and 3

D) 1,2 and 3

Solution: (D)

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

Question Context 24-26:

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

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

A) In case of very large x; bias is low

B) In case of very large x; bias is high

C) We can't say about bias

D) None of these

Solution: (B)

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

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

A) Some of the coefficient will become absolute zero

B) Some of the coefficient will approach zero but not absolute zero

C) Both A and B depending on the situation

D) None of these

Solution: (B)

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

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

A) Some of the coefficient will become zero

B) Some of the coefficient will be approaching to zero but not absolute zero
C) Both A and B depending on the situation
D) None of these
Solution: (A)
As already discussed, lasso applies absolute penalty, so some of the coefficients will become zero.

27. **Which of the following statement is true about outliers in Linear regression?**
A) Linear regression is sensitive to outliers
B) Linear regression is not sensitive to outliers
C) Can't say
D) None of these
Solution: (A)
The slope of the regression line will change due to outliers in most of the cases. So Linear Regression is sensitive to outliers.

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

A) Since the there is a relationship means our model is not good
B) Since the there is a relationship means our model is good
C) Can't say
D) None of these
Solution: (A)
There should not be any relationship between predicted values and residuals. If there exists any relationship between them,it means that the model has not perfectly captured the information in the data.

Question Context 29-31:
Suppose that you have a dataset D1 and you design a linear regression model of degree 3 polynomial and you found that the training and testing error is "0" or in another terms it perfectly fits the data.
29. What will happen when you fit degree 4 polynomial in linear regression?
A) There are high chances that degree 4 polynomial will over fit the data
B) There are high chances that degree 4 polynomial will under fit the data
C) Can't say
D) None of these
Solution: (A)
Since is more degree 4 will be more complex(overfit the data) than the degree 3 model so it will again perfectly fit the data. In such case training error will be zero but test error may not be zero.

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

A) It is high chances that degree 2 polynomial will over fit the data

B) It is high chances that degree 2 polynomial will under fit the data

C) Can't say

D) None of these

Solution: (B)

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

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

A) Bias will be high, variance will be high

B) Bias will be low, variance will be high

C) Bias will be high, variance will be low

D) Bias will be low, variance will be low

Solution: (C)

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

Question Context 32-33:

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

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

A) Increase

B) Decrease

C) Remain constant

D) Can't Say

Solution: (D)

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

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

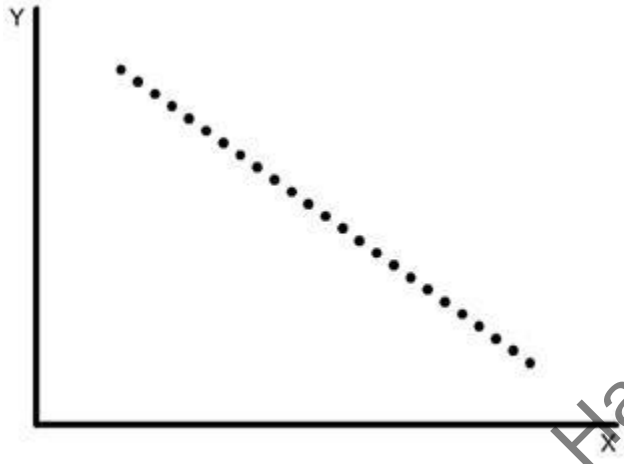
A) Bias increases and Variance increases

B) Bias decreases and Variance increases

C) Bias decreases and Variance decreases

D) Bias increases and Variance decreases
E) Can't Say False
Solution: (D)
As we increase the size of the training data, the bias would increase while the variance would decrease.

Question Context 34:
Consider the following data where one input(X) and one output(Y) is given.



34. What would be the root mean square training error for this data if you run a Linear Regression model of the form (Y = A0+A1X)?

A) Less than 0
B) Greater than zero
C) Equal to 0
D) None of these
Solution: (C)
We can perfectly fit the line on the following data so mean error will be zero.

Question Context 35-36:
Suppose you have been given the following scenario for training and validation error for Linear Regression.

Scenario	Learning Rate	Number of iterations	Training Error	Validation Error
1	0.1	1000	100	110
2	0.2	600	90	105

3	0.3	400	110	110
4	0.4	300	120	130
5	0.4	250	130	150

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

A) 1

B) 2

C) 3

D) 4

Solution: (B)

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

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

A) Training Error will decrease and Validation error will increase

B) Training Error will increase and Validation error will increase

C) Training Error will increase and Validation error will decrease

D) Training Error will decrease and Validation error will decrease

E) None of the above

Solution: (D)

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

Question Context 37-38:

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

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

1. I will add more variables
2. I will start introducing polynomial degree variables
3. I will remove some variables

A) 1 and 2

B) 2 and 3

C) 1 and 3

D) 1, 2 and 3

Solution: (A)

In case of under fitting, you need to induce more variables in variable space or you can add some polynomial degree variables to make the model more complex to be able to fir the data better.

38. Now situation is same as written in previous question(under fitting).Which of following regularization algorithm would you prefer?

A) L1
B) L2
C) Any
D) None of these
Solution: (D)
I won't use any regularization methods because regularization is used in case of overfitting.

39. **True-False: Is Logistic regression a supervised machine learning algorithm?**
A) TRUE
B) FALSE
Solution: A
True, Logistic regression is a supervised learning algorithm because it uses true labels for training. Supervised learning algorithm should have input variables (x) and an target variable (Y) when you train the model .

40. **True-False: Is Logistic regression mainly used for Regression?**
A) TRUE
B) FALSE
Solution: B
Logistic regression is a classification algorithm, don't confuse with the name regression.

41. **True-False: Is it possible to design a logistic regression algorithm using a Neural Network Algorithm?**
A) TRUE
B) FALSE
Solution: A
True, Neural network is a is a *universal* approximator so it can implement linear regression algorithm.

42. **True-False: Is it possible to apply a logistic regression algorithm on a 3-class Classification problem?**
A) TRUE
B) FALSE
Solution: A
Yes, we can apply logistic regression on 3 classification problem, We can use One Vs all method for 3 class classification in logistic regression.

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

A) Least Square Error
B) Maximum Likelihood
C) Jaccard distance
D) Both A and B
Solution: B
Logistic regression uses maximum likely hood estimate for training a logistic regression.

44. **Which of the following evaluation metrics can not be applied in case of logistic regression output to compare with target?**
A) AUC-ROC
B) Accuracy
C) Logloss
D) Mean-Squared-Error
Solution: D
Since, Logistic Regression is a classification algorithm so it's output can not be real time value so mean squared error can not use for evaluating it

45. **One of the very good methods to analyze the performance of Logistic Regression is AIC, which is similar to R-Squared in Linear Regression. Which of the following is true about AIC?**
A) We prefer a model with minimum AIC value
B) We prefer a model with maximum AIC value
C) Both but depend on the situation
D) None of these
Solution: A
We select the best model in logistic regression which can least AIC.

46. **[True-False] Standardisation of features is required before training a Logistic Regression.**
A) TRUE
B) FALSE
Solution: B
Standardization isn't required for logistic regression. The main goal of standardizing features is to help convergence of the technique used for optimization.

47. Which of the following algorithms do we use for Variable Selection?
A) LASSO
B) Ridge
C) Both
D) None of these

Solution: A
In case of lasso we apply a absolute penality, after increasing the penality in lasso some of the coefficient of variables may become zero.

Context: 48-49

Consider a following model for logistic regression: P (y =1|x, w)= g(w0 + w1x)
where g(z) is the logistic function.

In the above equation the P (y =1|x; w) , viewed as a function of x, that we can get by changing the parameters w.

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

A) (0, inf)
B) (-inf, 0)
C) (0, 1)
D) (-inf, inf)

Solution: C

For values of *x* in the range of real number from −∞ to +∞ Logistic function will give the output between (0,1)

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

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

Solution: A

Explanation is same as question number 10

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

A) odds will be 0
B) odds will be 0.5
C) odds will be 1
D) None of these

Solution: C

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

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

A) (– ∞ , ∞)
B) (0,1)
C) (0, ∞)
D) (- ∞, 0)

Solution: A

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

52. **Which of the following option is true?**

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

Solution:A

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

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

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

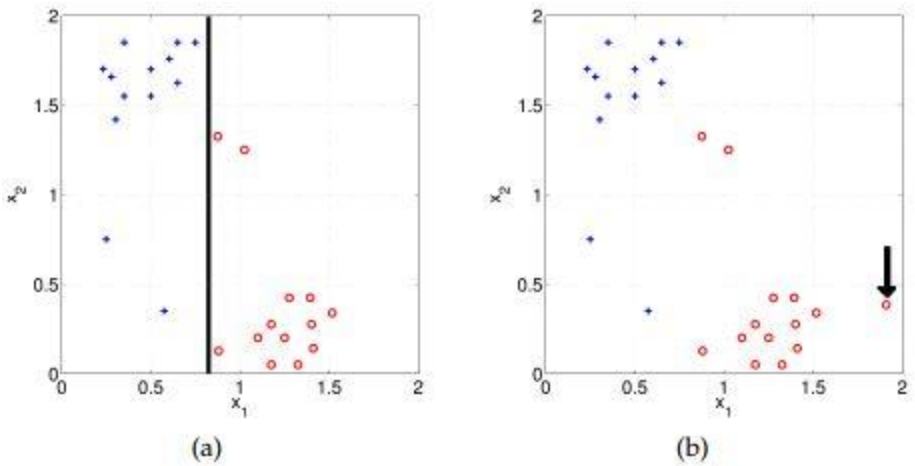
Logit_inv(x): is a inverse logit function of any number "x"

A) Logistic(x) = Logit(x)
B) Logistic(x) = Logit_inv(x)
C) Logit_inv(x) = Logit(x)
D) None of these

Solution: B

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

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



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

Solution: A

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

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

Note: Consider remaining parameters are same.

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

Solution: A and D

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

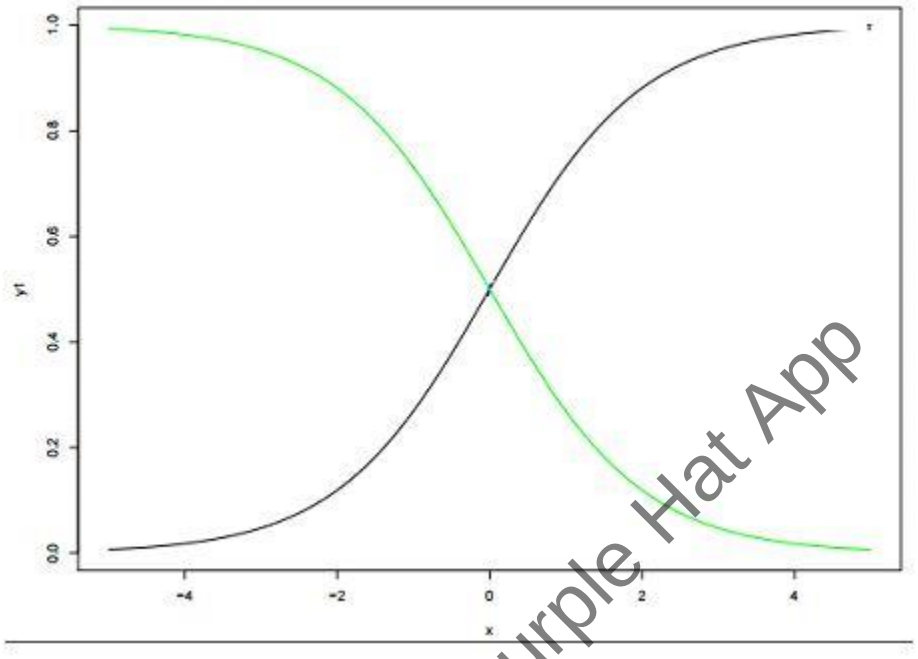
56. **Choose which of the following options is true regarding One-Vs-All method in Logistic Regression.**

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

Solution: A

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

57. **Below are two different logistic models with different values for β0 and β1.**



Which of the following statement(s) is true about β0 and β1 values of two logistics models (Green, Black)?

Note: consider Y = β0 + β1*X. Here, β0 is intercept and β1 is coefficient.

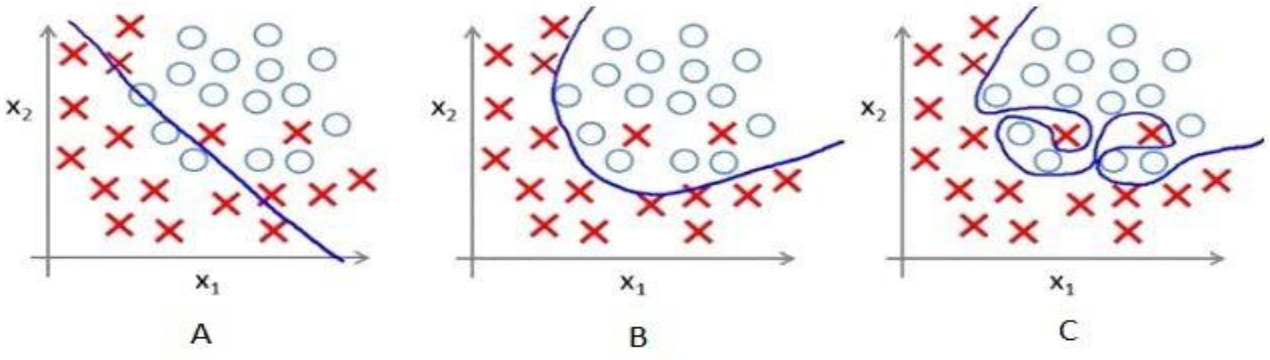
A) β1 for Green is greater than Black
B) β1 for Green is lower than Black
C) β1 for both models is same
D) Can't Say

Solution: B

β0 and β1: β0 = 0, β1 = 1 is in X1 color(black) and β0 = 0, β1 = −1 is in X4 color (green)

Context 58-60

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



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

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

Solution: C

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

59. What do you conclude after seeing this visualization?

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

A) 1 and 3
B) 1 and 3
C) 1, 3 and 4
D) 5

Solution: C

The trend in the graphs looks like a quadratic trend over independent variable X. A higher degree(Right graph) polynomial might have a very high accuracy on the train population but is expected to fail badly

on test dataset. But if you see in left graph we will have training error maximum because it underfits the training data

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

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

Solution: A

Since, more regularization means more penality means less complex decision boundry that shows in first figure A.

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

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

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

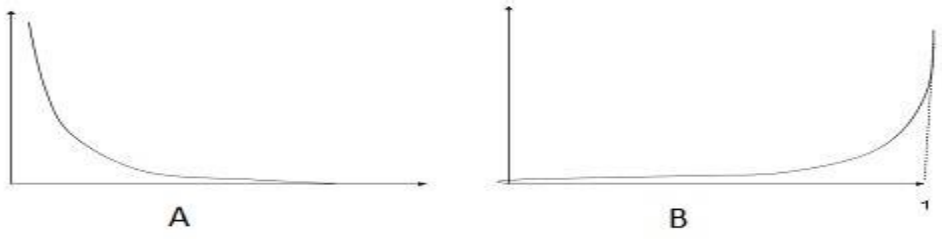
Solution: D

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

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

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

Note: Y is the target class

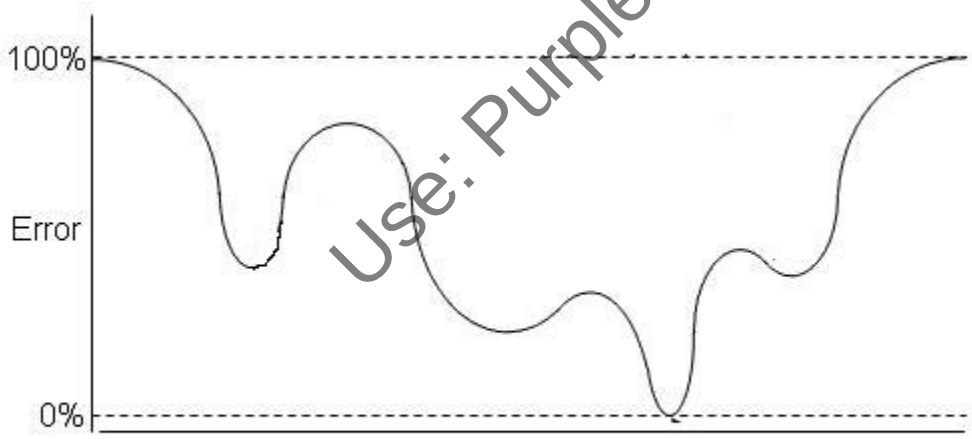


A) A
B) B
C) Both
D) None of these

Solution: A

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

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



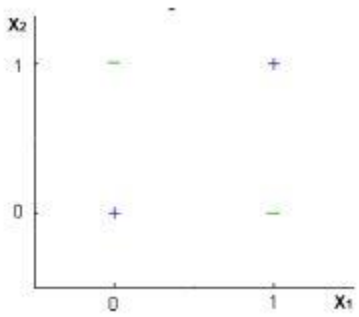
Now, How many local minimas are present in the graph?

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

Solution: C

There are three local minima present in the graph

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



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

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

Solution: B

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

UNIT IV

1. The SVM's are less effective when:

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

Ans Solution: C

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

2. The cost parameter in the SVM means:

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

Ans Solution: C

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

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

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

Ans Solution: D

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

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

Assumes that all the features in a dataset are equally important

Assumes that all the features in a dataset are independent

Both A and B - answer

None of the above options

Ans Solution: C

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

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

Ans Solution: B

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

6 The SVM's are less effective when:

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

Ans Solution: C

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

7 What is/are true about kernel in SVM?

1. Kernel function map low dimensional data to high dimensional space
2. It's a similarity function

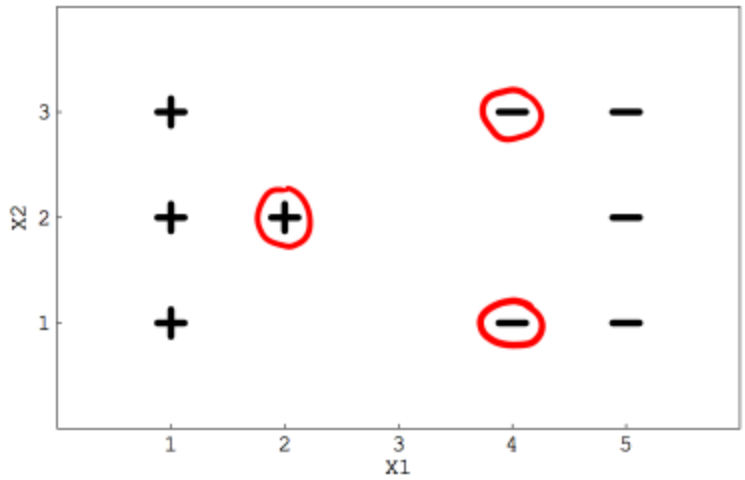
A) 1
B) 2
C) 1 and 2
D) None of these

Ans Solution: C

Both the given statements are correct**.**

Question Context:8– 9

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



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

A) Yes
B) No

Solution: A

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

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

A) True
B) False

Solution: B

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

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

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

Solution: B

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

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

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

Solution: A

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

12. What do you mean by a hard margin?

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

Solution: A

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

13. The minimum time complexity for training an SVM is O(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

Solution: A

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

14. The effectiveness of an SVM depends upon:

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

Solution: D

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

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

A) TRUE
B) FALSE

Solution: A

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

16. The SVM's are less effective when:

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

Solution: C

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

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

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

Solution: B

The gamma parameter in SVM tuning signifies the influence of points either near or far away from the hyperplane.

For a low gamma, the model will be too constrained and include all points of the training dataset, without really capturing the shape.

For a higher gamma, the model will capture the shape of the dataset well.

18. The cost parameter in the SVM means:

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

Solution: C

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

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

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

Note: For small C was also classifying all data points correctly

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

Solution: A

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

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

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

Solution: A

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

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

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

Solution: C

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

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

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

Solution: D

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

Question Context: 23 – 25

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

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

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

Solution: C

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

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

1. We are lowering the bias
2. We are lowering the variance
3. We are increasing the bias
4. We are increasing the variance

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

Solution: C

Better model will lower the bias and increase the variance

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

A) We will increase the parameter C
B) We will decrease the parameter C
C) Changing in C don't effect
D) None of these

Solution: A

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

26. **We usually use feature normalization before using the Gaussian kernel in SVM. What is true about feature normalization?**

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

A) 1
B) 1 and 2
C) 1 and 3
D) 2 and 3

Solution: B

Statements one and two are correct.

Question Context: 27-29

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

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

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

Solution: D

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

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

A) 20
B) 40
C) 60
D) 80

Solution: B

It would take 10×4 = 40 seconds

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

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

Solution: A

Training the SVM only one time would give you appropriate results

Question context: 30 –31

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

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

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

Solution: A

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

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

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

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

Solution: C

Both the given statements are correct.

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

1. Kernel function map low dimensional data to high dimensional space
2. It's a similarity function

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

Solution: C

Both the given statements are correct.

UNIT V

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

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

Ans D

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

a) Factor analysis
b) Decision trees are robust to outliers
c) Decision trees are prone to be overfit

d) None of the above

Ans C

3. Can decision trees be used for performing clustering?

a. True
b. False

Ans Solution: (A)

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

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

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

Ans Solution: (A)

5 **Sentiment Analysis is an example of:**

Regression

Classification

Clustering

Reinforcement Learning

Options:

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

Ans D

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

Capping and flouring of variables

Removal of outliers

Options:

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

Ans A

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

1. In bagging trees, individual trees are independent of each other
2. Bagging is the method for improving the performance by aggregating the results of weak learners

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

Ans **Solution: C**

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

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

1. In boosting trees, individual weak learners are independent of each other
2. It is the method for improving the performance by aggregating the results of weak learners

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

Ans **Solution: B**

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

9. In Random forest you can generate hundreds of trees (say T1, T2 …..Tn) and then aggregate the results of these tree. Which of the following is true about individual (Tk) tree in Random Forest?

1. Individual tree is built on a subset of the features
2. Individual tree is built on all the features
3. Individual tree is built on a subset of observations
4. Individual tree is built on full set of observations

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

Ans Solution: A

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

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

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

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

Ans Solution: A

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

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

1. Both methods can be used for classification task
2. Random Forest is use for classification whereas Gradient Boosting is use for regression task
3. Random Forest is use for regression whereas Gradient Boosting is use for Classification task
4. Both methods can be used for regression task

A) 1
B) 2
C) 3
D) 4
E) 1 and 4

Solution: E

Both algorithms are design for classification as well as regression task.

12. In Random forest you can generate hundreds of trees (say T1, T2 …..Tn) and then aggregate the results of these tree. Which of the following is true about individual(Tk) tree in Random Forest?

1. Individual tree is built on a subset of the features
2. Individual tree is built on all the features
3. Individual tree is built on a subset of observations
4. Individual tree is built on full set of observations

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

Solution: A

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

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

1. Gradient Boosting
2. Extra Trees
3. AdaBoost
4. Random Forest

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

Solution: D

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

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

A) Random Forest
B) Adaboost
C) Extra Trees
D) Gradient Boosting
E) Decision Trees

Solution: E

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

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

1. Number of tree should be as large as possible
2. You will have interpretability after using RandomForest

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

Solution: A

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

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

A) TRUE
B) FALSE

Solution: A

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

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

1. We build the N regression with N bootstrap sample
2. We take the average the of N regression tree
3. Each tree has a high variance with low bias

A) 1 and 2
B) 2 and 3
C) 1 and 3
D) 1,2 and 3

Solution: D

All of the options are correct and self-explanatory

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

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

Solution: B

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

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

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

Solution: A

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

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

Scenario	Depth	Training Error	Validation Error
1	2	100	110
2	4	90	105
3	6	50	100
4	8	45	105

5	10	30	150

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

Solution: B

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

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

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

Options:

A. 1 only

B. 2 only

C. 4 only

D. 2 and 3

Solution: D

- DBSCAN can form a cluster of any arbitrary shape and does not have strong assumptions for the distribution of data points in the data space.
- DBSCAN has a low time complexity of order O (n log n) only.

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

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

23. Which of the following is required by K-means clustering?
a) defined distance metric
b) number of clusters
c) initial guess as to cluster centroids
d) all of the mentioned

Answer: d
Explanation: K-means clustering follows partitioning approach.

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

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

25. Which of the following function is used for k-means clustering?
a) k-means
b) k-mean
c) heat map
d) none of the mentioned

Answer: a
Explanation: K-means requires a number of clusters.

26. K-means is not deterministic and it also consists of number of iterations.
a) True
b) False

Answer: a
Explanation: K-means clustering produces the final estimate of cluster centroids.

27.