MACHINE LEARNING

1 **In Q1 to Q7, only one option is correct, Choose the correct option:**

1. The value of correlation coefficient will always be:
   A) between 0 and 1 B) greater than -1
   C) between -1 and 1 D) between 0 and -1

Answer – C) between -1 and 1

2. Which of the following cannot be used for dimensionality reduction?
   A) Lasso Regularisation B) PCA
   C) Recursive feature elimination D) Ridge Regularisation

Answer – C) Recursive feature elimination

3. Which of the following is not a kernel in Support Vector Machines?
   A) linear B) Radial Basis Function
   C) hyperplane D) polynomial

Answer – C) hyperplane

4. Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?
   A) Logistic Regression B) Naïve Bayes Classifier
   C) Decision Tree Classifier D) Support Vector Classifier

Answer – A) Logistic Regression

5. In a Linear Regression problem, 'X' is independent variable and 'Y' is dependent variable, where 'X' represents weight in pounds. If you convert the unit of 'X' to kilograms, then new coefficient of 'X' willbe?
   (1 kilogram = 2.205 pounds)
   A) 2.205 × old coefficient of 'X' B) same as old coefficient of 'X'
   C) old coefficient of 'X' ÷ 2.205 D) Cannot be determined

Answer – A) 2.205 x old coefficient of 'X'

6. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy ofthe model?
   A) remains same B) increases
   C) decreases D) none of the above

Answer – B) increases

MACHINE LEARNING

7. Which of the following is not an advantage of using random forest instead of decision trees?
   A) Random Forests reduce overfitting
   B) Random Forests explains more variance in data then decision trees
   C) Random Forests are easy to interpret
   D) Random Forests provide a reliable feature importance estimate

Answer – C) Random Forests are easy to interpret

In Q8 to Q10, more than one options are correct, Choose all the correct options:

8. Which of the following are correct about Principal Components?
   A) Principal Components are calculated using supervised learning techniques
   B) Principal Components are calculated using unsupervised learning techniques
   C) Principal Components are linear combinations of Linear Variables.
   D) All of the above

Answer – B) & C)

9. Which of the following are applications of clustering?
   A) Identifying developed, developing and under-developed countries on the basis of factors like GDP,poverty index, employment rate, population and living index
   B) Identifying loan defaulters in a bank on the basis of previous years' data of loan accounts.
   C) Identifying spam or ham emails
   D) Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugarlevels.

Answer – B) & C)

10. Which of the following is(are) hyper parameters of a decision tree?
    A) max_depth
    B) max_features
    C) n_estimators
    D) min_samples_leaf

Answer – B) & D)

Q11 to Q15 are subjective answer type questions, Answer them briefly.

11. What are outliers? Explain the Inter Quartile Range (IQR) method for outlier detection.

Answer –

Outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In other words, they are the unusual values in a dataset.

Inter Quartile Range (IQR) has the formula

IQR=Q3-Q1,

MACHINE LEARNING

Where IQR is the range between first and third quartile range. The datapoints which falls below Q1-1.5 and above Q3+1.5 are outliers.

IQR is used to measure variability by dividing a data set into quartiles. The data is sorted in ascending order and split into 4 equal parts. Q1, Q2, Q3 called first, second and third quartiles are the values which separate the 4 equal parts.

- Q1 represents the 25th percentile of the data.
- Q2 represents the 50th percentile of the data.
- Q3 represents the 75th percentile of the data.

If a dataset has 2n / 2n+1 data points, then
Q1 = median of the dataset &
Q2 = median of n smallest data points.
Q3 = median of n highest data points.

IQR is the range between the first and the third quartiles namely Q1 and Q3: IQR = Q3 – Q1. The data points which fall below Q1 – 1.5 IQR or above Q3 + 1.5 IQR are outliers.

12. What is the primary difference between bagging and boosting algorithms?

Answer –

Difference between **Bagging** and **Boosting** algorithms are:

a. Bagging is a way to decrease the variance in the prediction by generating additional data for training from dataset using combinations with repetitions to produce multi-sets of the original data. Boosting is an iterative technique which adjusts the weight of an observation based on the last classification.

b. Bagging is the method of merging same type of predictions and Boosting is a method of merging
different types of predictions.

c. In Bagging the result is obtained by averaging the responses of the N learners. Boosting assigns a second set of weights, this time for the N classifiers in order to take a weighted average of their estimates.

13. What is adjusted R^2 in linear regression. How is it calculated?

Answer -

Adjusted R^2 is a modified version of **R^2** that has been adjusted for the number of predictors in the model. Its value can be calculated based on value of **R^2**.

Every time you add an independent variable to a model, R squared increases, even if the independent variable is insignificant.

MACHINE LEARNING

Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size.

It is calculated as:

$$\textbf{Adjusted } R^2 = 1 - [(1-R^2)*(n-1)/(n-k-1)]$$

Where, ***n*** is the number of points in data sample, and ***k*** is the number of independent regressors, i.e. the number of variables in model, excluding the constant.

14. What is the difference between standardization and normalization?

Answer –

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as **Min-Max scaling**.

It rescales the values into a range of [0,1]. This might be useful in some cases where all parameters need to have the same positive scale. However, the outliers from the data set are lost.

$$X_{changed} = X - X_{min}\, X_{max} - X_{min}\, X_{changed}$$

$$= X - X_{min}\, X_{max} - X_{min}$$

Standardization is another scaling technique where the values are centred around the mean with a unit standard deviation. It means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

It rescales data to have a mean (μμ) of 0 and standard deviation (σσ) of 1 (unit variance).

$$X_{changed} = X - \mu\, \sigma$$

15. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation.

Answer –

Cross-Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model. It is a technique for assessing how statistical analysis generalizes to an independent dataset. It is a technique for evaluating machine learning models by training several models on subsets of the available input data and evaluating them on complementary subset of data.

Advantage: More accurate estimate of out-of-sample accuracy.

Disadvantage: It is computationally very expensive as we need to train on multiple training sets.