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
2. Which of the following cannot be used for dimensionality reduction?
   A) Lasso Regularisation B) PCA
   C) Recursive feature elimination D) Ridge Regularisation
3. Which of the following is not a kernel in Support Vector Machines?
   A) linear B) Radial Basis Function
   C) hyperplane D) polynomial
5. 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
6. 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' will be?
   (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
7. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?
   A) remains same B) increases
   C) decreases D) none of the above
8. 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

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

9. Which of the following are correct about Principal Components?
   A) Principal Components are calculated using supervised learning techniques
   B) B) Principal Components are calculated using unsupervised learning techniques
   C) Principal Components are linear combinations of Linear Variables.
   D) All of the above
10. 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 sugar levels.
11. 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

MACHINE LEARNING

Q10 to Q15 are subjective answer type questions, Answer them briefly.

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

ANS: Outliers are observation that does not follow normal distribution or gaussian distribution curve. Generally, these are occurred due to false sampling, any error in collecting data or during Sampling.

Although outliers are present in data, sometimes they are true values, but disturbs overall performance of Model. So, it is important to handle those outliers.

IQR also known as inter quartile range, is method used to outlie those outliers from data set. In IQR, based on data quartiles are created and points in quartile are plotted in plot, called as box plot. These box plots are easy way of understanding outliers present in data set.

Basically IQR, is divided into four groups, those are Q1, Q2, Q3 and Q4. IQR is difference between Q3 – Q1.

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

Bagging: It is method used to reduce variance in data.

Boosting: It is method used to reduce training errors.

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

It is used to measure accuracy of Linear Regression models. It is calculated by residual mean by total mean and it is always less than and equal to R square.

Formula for the same is,

Adjusted $R^2 = 1 – [((1-R^2) \times (N – 1))/(N – p – 1))]$

Where

R^2 = Sample R Squared

N = Total Sample Size

P = Number of independent Variable

14. What is the difference between standardisation and normalisation?
15. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation.