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:
   C) between -1 and 1

2. Which of the following cannot be used for dimensionality reduction?
   C) Recursive feature elimination

3. Which of the following is not a kernel in Support Vector Machines?
   A) linear

4. Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?
   D) Support Vector Classifier

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' will be?
   (1 kilogram = 2.205 pounds)
   B) same as old coefficient of 'X'

6. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?
   B) increases

7. Which of the following is not an advantage of using random forest instead of decision trees?
   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?
   B. Principal Components are calculated using unsupervised learning techniques
   C. Principal Components are linear combinations of Linear Variables.

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
   D. Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.

10. Which of the following is(are) hyper parameters of a decision tree?
    A) max_depth B) max_features
    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 – Outlier is an extremely high or extremely low data point relative to the nearest data point and the rest of the neighboring co-existing values in a data graph or dataset. Interquartile range method is one of the efficient methods to find outliers. To find outliers we need to first calculate 1st and 3rd quartile and IQR (the difference between these two quartiles) and with these we can find outliers which are not allowed above and below the specific level.

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

    Ans – Bagging and Boosting are two types of Ensemble Learning. Bagging algorithms learns from each other independently in parallel and combines them for determining the model average whereas, boosting algorithms learn sequentially and adaptively to improve model predictions of a learning algorithm.

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

    Ans – Its an ability to make a more accurate view of the correlation between one variable and another. It can be calculated based on value of r-squared, number of independent variables (predictors), total sample size. Formula = $1-[((1-R^2)(n-1))/n-p-1]$

14. What is the difference between standardisation and normalisation?

    Ans – Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation.

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

    Ans – It's a statistical method of evaluating and comparing learning algorithms by dividing data into two segments i.e one used to learn or train a model and the other used to validate the model. The advantage of cross-validation is that it is very accurate among others and the disadvantage is that it required more power to run and it is very expensive in terms of processing power required.