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
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
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)
   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
6. 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
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

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
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 sugar levels.
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

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.
12. What is the primary difference between bagging and boosting algorithms?
13. What is adjusted R^2 in linear regression. How is it calculated?
14. What is the difference between standardisation and normalisation?
15. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation.

11.A) An outlier is an observation that lies an abnormal distance from other values in a random sample from a population

IQR is the range between the first and the third quartiles namely Q1 and Q3: IQR = Q3 –

The difference between Q3 and Q1 is called the Inter-Quartile Range or IQR. Any data point less than the Lower Bound or more than the Upper Bound is considered as an outlier.

12.A) Bagging:

Bagging attempts to tackle the over-fitting issue.
If the classifier is unstable (high variance), then we need to apply bagging.
Advantages of using Random Forest technique:

It manages a higher dimension data set very well.
It manages missing quantities and keeps accuracy for missing data.

Boosting :

Boosting tries to reduce bias.
If the classifier is steady and straightforward (high bias), then we need to apply boosting.\

Advantages of using Gradient Boosting methods:

It supports different loss functions.
It works well with interactions.
Disadvantages of using a Gradient Boosting method:

It requires cautious tuning of different hyper-parameters.

13.A) Adjusted R2 is a corrected goodness-of-fit (model accuracy) measure for linear models. It identifies the percentage of variance in the target field that is explained by the input or inputs. R2 tends to optimistically estimate the fit of the linear regression.

adjusted R squared formula = I - [(I-R2) X (n-I) / (n-k-I)]

14.A) In Normalisation, the change in values is that they are at a standard scale without distorting the differences in the values. Whereas, Standardisation assumes that the dataset is in Gaussian distribution and measures the variable at different scales, making all the variables equally contribute to the analysis.

15.A) 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.

Cross Validation helps in finding the optimal value of hyperparameters to increase the efficiency of the algorithm. 1. Increases Training Time: Cross Validation drastically increases the training time