**MACHINE LEARNING**

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

**ANS :- 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

**ANS :- 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

**ANS :- B) Radial Basis Function**

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

**ANS:- B) Naïve Bayes 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

**ANS :- A) 2.205 × old coefficient of ‘X’**

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

**ANS :- C) decreases**

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

**ANS :- 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

**ANS :- 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.

**ANS :- A) Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index**

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

**ANS :- A) max\_depth**

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.**

An outlier is a value or point that differs substantially from the rest of the data. The **quartiles** of a ranked set of data values are three points which divide the data into exactly four equal parts, each part comprising of quarter data.   
 Q1 is defined as the middle number between the smallest number and the median of the data set.Q2 is the [median](https://www.geeksforgeeks.org/mean-median-unsorted-array/) of the data.Q3 is the middle value between the median and the highest value of the data set.

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

while they are built independently for Bagging, Boosting tries to add new models that do well where previous models fail. only Boosting determines weights for the data to tip the scales in favor of the most difficult cases. it is an equally weighted average for Bagging and a weighted average for Boosting, more weight to those with better performance on training data. only Boosting tries to reduce bias. On the other hand, Bagging may solve the over-fitting problem, while Boosting can increase it.

**13. What is adjusted R2 in linear regression. How is it calculated?**

The adjusted R-squared is a modified version of R-squared that adjusts for the number of predictors in a regression model. It is calculated as:**Adjusted R2 = 1 – [ (1-R2)\* (n-1)/ (n-k-1)]** where: R2: The R2 of the model

**14. What is the difference between standardisation and normalisation?**

The two most discussed scaling methods are Normalization and Standardization. Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

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

Cross-validation is a**statistical method used to estimate the performance (or accuracy) of machine learning models**. It is used to protect against overfitting in a predictive model, particularly in a case where the amount of data may be limited.

**Advantages** :- It helps evaluate the quality of your model.

It helps to reduce/avoid problems of overfitting and underfitting.

It lets you select the model that will deliver the best performance on unseen data.

**Disadvantages**:- 1. Cross Validation significantly lengthens the training period. ...

2.In most cases, the structure you're studying develops over time in predictive modelling. As a result, you may notice variations in the training and validation sets.

3. Cross Validation requires a lot of computing power.