Machine Learning

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?

D) Ridge Regularisation

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?

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?

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?

C) decreases

7. Which of the following is not an advantage of using random forest instead of decision trees?

B) Random Forests explains more variance in data then decision trees

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

C) Identifying spam or ham emails

10. Which of the following is(are) hyper parameters of a decision tree?

B) max_features D) min_samples_leaf

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

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to

the analyst to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations

Interquartile Range Definition- The interquartile range defines the difference between the third and the first quartile. Quartiles are the partitioned values that divide the whole series into 4 equal parts. So, there are 3 quartiles. First Quartile is denoted by Q1 known as the lower quartile, the second Quartile is denoted by Q2 and the third Quartile is denoted by Q3 known as the upper quartile. Therefore, the interquartile range is equal to the upper quartile minus lower quartile.

Interquartile Range Formula The difference between the upper and lower quartile is known as the interquartile range. The formula for the interquartile range is

Interquartile range = Upper Quartile – Lower Quartile = Q-3 – Q-1

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

S.NO	Bagging	Boosting
1.	Bagging is a learning approach that aids in enhancing the performance, execution, and precision of machine learning algorithms.	Boosting is an approach that iteratively modifies the weight of observation based on the last classification.
2.	It is the easiest method of merging predictions that belong to the same type.	It is a method of merging predictions that belong to different types.
3.	Here, every model has equal weight.	Here, the weight of the models depends on their performance.
4.	In bagging, each model is assembled independently.	In boosting, the new models are impacted by the implementation of earlier built models.
5.	It helps in solving the over-fitting issue.	It helps in reducing the bias.
6.	In the case of bagging, if the classifier is unstable, then we apply bagging.	In the case of boosting, If the classifier is stable, then we apply boosting.

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

Adjusted R Squared refers to the statistical tool that helps investors measure the extent of the variable's variance, which is dependent and explained with

the independent variable. It considers the impact of only those independent variables that impact the variation of the dependent variable.

Adjusted R Squared or Modified R^2 determines the extent of the variance of the dependent variable, which the independent variable can explain. The specialty of the modified R^2 is that it does not consider the impact of all independent variables but only those which impact the variation of the dependent variable. Therefore, the value of the modified R^2 can also be negative, though it is not always negative.

The formula to calculate the adjusted R square of regression is below:

$R^2 = \{(1 / N) * \Sigma [(xi – x) * (Yi – y)] / (\sigma x * \sigma y)\}^2$

14. What is the difference between standardisation and normalisation

Typically we normalize **data when performing some type of analysis in which we have multiple variables that are measured on different scales and we want each of the variables to have the same range.**

This prevents one variable from being overly influential, especially if it's measured in different units (i.e. if one variable is measured in inches and another is measured in yards).

On the other hand, we typically **standardize** data when we'd like to know how many standard deviations each value in a dataset lies from the mean.

For example, we might have a list of exam scores for 500 students at a particular school and we'd like to know how many standard deviations each exam score lies from the mean score.

In this case, we could standardize the raw data to find out this information. Then, a standardized score of 1.26 would tell us that the exam score of that particular student lies 1.26 standard deviations above the mean exam score.

Whether you decide to normalize or standardize your data, keep the following in mind:

- A **normalized dataset** will always have values that range between 0 and 1.
- A **standardized dataset** will have a mean of 0 and standard deviation of 1, but there is no specific upper or lower bound for the maximum and minimum values.
-

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

Cross Validation in Machine Learning is a great technique to deal with overfitting problem in various algorithms. Instead of training our model on one training dataset, we train our model on many datasets. Below are some of the

advantages and disadvantages of Cross Validation in Machine Learning:

Advantages of Cross Validation

1. Reduces Overfitting: **In Cross Validation, we split the dataset into multiple folds and train the algorithm on different folds. This prevents our model from overfitting the training dataset. So, in this way, the model attains the generalization capabilities which is a good sign of a robust algorithm.**

Note: **Chances of overfitting are less if the dataset is large. So, Cross Validation may not be required at all in the situation where we have sufficient data available.**

2. Hyperparameter Tuning: **Cross Validation helps in finding the optimal value of hyperparameters to increase the efficiency of the algorithm.**

Disadvantages of Cross Validation

1. Increases Training Time: **Cross Validation drastically increases the training time. Earlier you had to train your model only on one training set, but with Cross Validation you have to train your model on multiple training sets.**

For example, if you go with 5 Fold Cross Validation, you need to do 5 rounds of training each on different 4/5 of available data. And this is for only one choice of hyperparameters. If you have multiple choice of parameters, then the training period will shoot too high.

2. Needs Expensive Computation: **Cross Validation is computationally very expensive in terms of processing power required.**