MACHINE LEARNING

1. 1. The value of correlation coefficient will always be:
   C) between -1 and 1

2. 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?
   C) hyperplane

4. Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?
   A) Logistic Regression

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?
   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

8. Which of the following are correct about Principal Components?
   A) Principal Components are calculated using supervised learning techniques,
   C) Principal Components are linear combinations of Linear Variables.

9. Which of the following are applications of clustering?
   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, D) min_samples_leaf

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

    In a dataset the outliers are the datapoint which are significantly different from other datapoints. They have unexpected high values as compared to the other data in the dataset.These are the result of various kind of errors for example error in data collection and data entry, experimental error and instrumental error, error due to natural factors. Presence of outliers adversly affect the statistical parameter values like mean,median & standard deviation.It also increases bias.

IQR :

Any values which doesn't falls in the range of [Q1-1.5 x IQR to Q3+1.5 x IQR] is called as an outlier. Here Q1 and Q3 are the first and third quartile respectively and IQR is called as Inter Quartile Range. There are 3 different quartile namely Q1,Q2,Q3. Q2 represents the median of the entire dataset and it also divide the dataset in two equal parts. Q1 or the first quartile represents the median of upper part while Q3 or the third quartile represents the median of lower part of data. So the interquartile range is given as IQR= Q3-Q1 .(1.5 x IQR) in the above formula represents the upper and lower whisker limit which are the highest and lowest occuring value in the dataset. The outliers can be visualized by boxplot. In the boxplot the box represents the IQR and the two bars represent the whisker.

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

Bagging:

It decreases the variance .Weight in all model remain same .

Boosting:

It decreases the bias .Weight for the model is determined by performane .

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

Just like R2 adjusted R2 also shows the best fit line but it adjust the number of variable. Adding more variable will decreases the value of adjusted R2. Its value is always less than or equal to R2. Both R2 and the adjusted R2 gives an idea abut the best fit line.

R2_adjusted= 1-[(1-R2)(n-1)/(n-k-1)]

14. What is the difference between standardisation and normalisation?

Standardisation :

It is the process of converging all datapoint in normal distributon range.where the mean =0 and standard deviation =1.

called by: from sklearn.preprocessing import StandardScaler

Normalization :

It is the process of converging all the datapoint within a range of [0,1]. Also termed as min-max scaling.

called by: from sklearn.preprocessing import MinMaxScaler

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

Cross validation used to determine the performance of machine learning models.

After completion of training we test the model on the trained data it may not execute at optimum performance. Here comes the Cross validation.

It create a sample of unseen data from the dataset. After training this unseen sample can be used to test to determine the performance of the model.

Various kind of cross-vall technique are there,

- LOOCV
- KFOLD CV
- HOLDOUT CV

HOLDOUT CV is the basic from of cross validation. Here it simply shuffel and split the

dataset into two part at a given ratio.
KFOLD CV is an improvement over HOLDOUT CV. It runs K times and each time it generate a different set for testing.
Cross validation is more accurate and efficient as it uses all data both for testing and training. It reduces the bias
As it uses all data for training and testing the execution time increases.