MACHINE LEARNING – 4

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

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

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

4. Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?
   Ans. 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?
   Ans. 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?
   Ans. D) none of the above

7. Which of the following is not an advantage of using random forest instead of decision trees?
   Ans. C) Random Forests are easy to interpret

8. Which of the following are correct about Principal Components?
   Ans. 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?
   Ans. 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.
   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?
    Ans. A) max_depth
    B) max_features
    D) min_samples_leaf

11. What are outliers? Explain the Inter Quartile Range (IQR) method for outlier detection.
    Ans. An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. Outliers are extreme values that stand out greatly from the overall pattern of values in a dataset or graph. Outliers is an extremely high or extremely low data point relative to the nearest data point.

    Q1, Q2, Q3 called first, second and third quartiles are the values which separate the 4 equal parts.
    - Q1 represents the 25th percentile of the data.
    - Q2 represents the 50th percentile of the data.
    - Q3 represents the 75th percentile of the data.

    IQR is the range between the first and the third quartiles namely Q1 and Q3: IQR = Q3 – Q1.
    The data points which fall below Q1 – 1.5 IQR or above Q3 + 1.5 IQR are outliers.

12. What is the primary difference between bagging and boosting algorithms?
    Ans. Bagging: It is a homogeneous weak learners' model that learns from each other independently in parallel and combines them for determining the model average. It aims to decrease variance, not bias.
    Boosting: It is also a homogeneous weak learners' model but works differently from Bagging. In this model, learners learn sequentially and adaptively to improve model predictions of a learning algorithm. It aims to decrease bias, not variance.

13. What is adjusted R2 in linear regression. How is it calculated?
    Ans. The adjusted R-squared is a modified version of R-squared that accounts for predictors that are not significant in a regression model. In other words, the adjusted R-squared shows whether adding additional predictors improve a regression model or not. It measures the proportion of variation explained by only those independent variables that really help in explaining the dependent variable. It penalizes you for adding independent variable that do not help in predicting the dependent variable. If you add more and more useless variables to a model, adjusted r-squared will decrease. If you add more useful variables, adjusted r-squared will increase. Adjusted R2 is always less than or equal to R2.

$$R^2_{adj} = 1 - \left[\frac{(1 - R^2)(n-1)}{n-k-1} \right]$$

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

S.No.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling.	Mean and standard deviation is used for scaling. i.e. Xnew = (X - mean)/Std
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

15. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation.
    Ans. Cross-validation is a technique for validating the model efficiency by training it on the subset of input data and testing on previously unseen subset of the input data. We can also say that it is a technique to check how a statistical model generalizes to an independent dataset. Cross-Validation is a resampling technique with the fundamental idea of splitting the dataset into 2 parts- training data and test data. Train data is used to train the model and the unseen test data is used for prediction. If the model performs well over the test data and gives good accuracy, it means the model hasn't overfitted the training data and can be used for prediction.

    Advantage of Cross Validation: 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. By using this method, we make use of all data points and hence it is low bias.

Disadvantage of Cross Validation: 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.