Machine Learning Assignment 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. 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?

Ans. C) Old coefficient of 'X' ÷ 2.205

6. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?

Ans. C) decreases

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

Ans. A) Random Forests reduce overfitting

8. Which of the following are correct about Principal Components?

Ans. B) Principal Components are calculated using unsupervised learning techniques.

9. Which of the following are applications of clustering?

Ans. ALL ARE CORRECT

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?

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. outlier is an observation of data that does not fit the rest of the data. It is sometimes called an extreme value. When you graph an outlier, it will appear not to fit the pattern of the graph. Some outliers are due to mistakes (for example, writing down 50 instead of 500) while others may indicate that something unusual is happening. IQR is used to measure variability by dividing a data set into quartiles. The data is sorted in ascending order and split into 4 equal parts. 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. If a dataset has 2n / 2n+1 data points, then Q1 = median of the dataset. Q2 = median of n smallest data points. Q3 = median of n highest data points. 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 only controls high variance in a model while boosting controls both bias and variance. So, boosting is considered to be more effective.
- In bagging, each weak learner has equal say in the final decision while in boosting the weak learner which generates high accuracy has more say in the final decision.

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

Ans. 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 R2 = 1 – [(1-R2)*(n-1)/(n-k-1)]

where: R 2 : The R 2 of the model n: The number of observations k: The number of predictor variables

14. What is the difference between standardisation and normalisation?

Ans. Normalization is a part of data processing and cleansing techniques. The main goal of normalization is to make the data homogenous over all records and fields. It helps in creating a linkage between the entry data which in turn helps in cleaning and improving data quality. Whereas data standardization is the process of placing dissimilar features on the same scale. Standardized data in other words can be defined as rescaling the attributes in such a way that their mean is 0 and standard deviation becomes 1.

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

Ans. Cross-validation is a technique for evaluating a machine learning model and testing its performance. CV is commonly used in applied ML tasks. It helps to compare and select an appropriate model for the specific predictive modelling problem. CV is easy to understand, easy to implement, and it tends to have a lower bias than other methods used to count the model's efficiency scores. All this makes cross-validation a powerful tool for selecting the best model for the specific task.
- Advantage - 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 generalisation capabilities which is a good sign of a robust algorithm.
- Disadvantage - 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.