MACHINE LEARNING (worksheet-4)

Question-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

Answer- C) between -1 and 1

Question-2 Which of the following cannot be used for dimensionality reduction?

A) Lasso Regularisation

B) PCA

C) Recursive feature elimination

D) Ridge Regularisation

Answer- B) PCA

Question-3 Which of the following is not a kernel in Support Vector Machines?

A) linear

B) Radial Basis Function

C) hyperplane

D) polynomial

Answer- A) linear

Question-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

Answer- A) Logistic Regression

Question-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

Answer- D) Cannot be determined

Question- 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

Answer- B) increases

Question-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

Answer- A) Random Forests reduce overfitting

Question- 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

Answer- B) and C)

Question-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.

Answer- A) and C)

Question-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

Answer- C) and D)

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

Answer- The difference between Q3 and Q1 is called the Inter Quartile Range (IQR). Any data point less than the Lower Bound or more than the Upper Bound is considered as an outlier.

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

Answer- Bagging is a method of merging the same type of predictions. Boosting is a method of merging different types of predictions. Bagging decreases variance, not bias, and solves over-fitting issues in a model. Boosting decreases bias, not variance.

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

Answer- Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases when the new term improves the model more than would be expected by chance.

Adjusted R-squared is calculated by dividing the residual mean square error by the total mean square error. The result is then subtracted from 1. Adjusted R2 is always less than or equal to R2.

Question-14 What is the difference between standardisation and normalisation?

Answer- In Normalisation, the change in values is that they at a standard scale without distorting the difference in the values. Whereas, Standardisation assumes that the dataset is in Gaussian distribution and measures the variable at different scales, making all the variables equally contribute to the analysis.

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

Answer- 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, particular in a case where the amount of data may be limited.

An advantage of using this method is that we make use of all data points and hence it is low bias. The major drawback of this method is that it leads to higher variation in the testing model as we are testing against one data points.