Machine Learning Worksheet 4

1) The value of correlation coefficient will always be

Ans : between 1 and 1

2) Which of the following 1 cannot be used for dimensionality reduction?

Ans : Recursive feature elimination

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

Ans : linear

4) Amongst the following, which one is least suitable for a dataset having non boundaries?

Ans : 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?

Ans : 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?

Ans : increases

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

Ans : Random Forests provide a reliable feature importance estimate

8) Which of the following are correct about Principal Components

Ans : All of the above

9) Which of the following are applications of cl ustering

Ans : 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 : 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. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations.

The outliers may suggest experimental errors, variability in a measurement, or an anomaly. The age of a person may wrongly be recorded as 200 rather than 20 Years. Such an outlier should definitely be discarded from the dataset.However, not all outliers are bad. Some outliers signify that data is significantly different from others. For example, it may indicate an anomaly like bank fraud or a rare disease.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 is used when our objective is to reduce the variance of a decision tree. Here the concept is to create a few subsets of data from the training sample, which is chosen randomly with replacement. Now each collection of subset data is used to prepare their decision trees thus, we end up with an ensemble of various models. The average of all the assumptions from numerous tress is used, which is more powerful than a single decision tree.

Boosting is another ensemble procedure to make a collection of predictors. In other words, we fit consecutive trees, usually random samples, and at each step, the objective is to solve net error from the prior trees.If a given input is misclassified by theory, then its weight is increased so that the upcoming hypothesis is more likely to classify it correctly by consolidating the entire set at last converts weak learners into better performing models.Gradient Boosting is an expansion of the boosting procedure.

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

Ans : 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.Adjusted R-Squared can be calculated mathematically in terms of sum of squares. The only difference between R-square and Adjusted R-square equation is degree of freedom.

Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size. Every time you add a independent variable to a model, the R-squared increases, even if the independent variable is insignificant. It never declines. Whereas

Adjusted R-squared increases only when independent variable is significant and affects dependent variable.

14. What is the difference between standardisation and normalisation?

Ans : The process of arranging the data in a database is known as Normalization. It is a scaling technique used to reduce redundancy in which the values are shifted and scaled in a range of 0 and 1. Normalization is used to remove the unwanted characteristics from the dataset, and it is useful when there are no outliers as it can not handle them.

Data standardization is a process in which the data is restructured in a uniform format. In statistics, standardization compares the variables by putting all the variables on the same scale. It is done by transforming the features by subtracting from the mean and dividing by the standard deviation. This process is also known as the Z-score.

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

Ans : Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.The three steps involved in cross-validation are as follows :

1.Reserveortion of sample data-set.

2.Using the rest data-set train the model.

3.Test the model using the reserve portion of the data-set.

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.

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.