Machine Learning Assignment 4

QUES 1. The value of correlation coefficient will always be:

Answer. C) Between -1 and 1

QUES 2. Which of the following cannot be used for dimensionality reduction?

Answer. C) Recursive feature elimination.

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

Answer. A) Linear.

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

Answer. A) Logistic Regression

QUES 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?

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

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

Answer. B) Increases

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

Answer. Random Forests are easy to interpret.

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

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

C) Principal Components are linear combinations of Linear Variables.

QUES 9. Which of the following are applications of clustering?

Answer. A) Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index.

D) Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.

QUES 10. Which of the following is(are) hyper parameters of a decision tree?

Answer. A) max_depth.

B) max_features

D) min_samples_leaf

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

Answer. 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. 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.

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

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

Boosting is a method of merging different types of predictions. Boosting decreases bias, not variance.

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

Answer. Adjusted R^2 is a corrected goodness-of-fit (model accuracy) measure for linear models. It identifies the percentage of variance in the target field that is explained by the input or inputs. R^2 tends to optimistically estimate the fit of the linear regression.

Adjusted R^2 is calculated by dividing the residual mean square error by the total mean square error (which is the sample variance of the target field). The result is then subtracted from 1. Adjusted R^2 is always less than or equal to R^2.

QUES 14. What is the difference between standardization and normalization?

Answer. Normalization: Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

X'= X-Xmin/Xmax-Xmin

Normalization is used when the distribution of data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.

Standardization: Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Standardization is used in cases where the data follows a Gaussian distribution. Unlike Normalization, Standardization does not have a bounding range. So, even if we have outliers in our data, they will not be affected by Standardization.

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

Answer. 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. Cross-validation is a technique that allows us to utilize our training data better for training and evaluating the model.

Advantage: It helps to test the ability of a Machine Learning Model to predict new data.

Disadvantage: The training algorithm has to be rerun from scratch K times, which means it takes K times as much computation to make an evaluation.