MACHINE LEARNING

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
   Ans: between -1 to 1

2. Which of the following cannot be used for dimensionality reduction?
   A) Lasso Regularisation B) PCA
   C) Recursive feature elimination D) Ridge Regularisation
   Ans: PCA

3. Which of the following is not a kernel in Support Vector Machines?
   A) linear B) Radial Basis Function
   C) hyperplane D) polynomial
   Ans: hyperplane

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
   Ans:

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
   Ans: 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?
   A) remains same B) increases
   C) decreases D) none of the above
   Ans: increases

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
   Ans: Random Forests explains more variance in data then decision trees

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.

MACHINE LEARNING

D) All of the above
Ans: All of the above

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.
Ans: A, B & D

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

Ans: A, B & D

11. What are outliers? Explain the Inter Quartile Range (IQR) method for outlier detection.
Ans: The outliers may suggest experimental errors, variability in a measurement, or an anomaly. 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 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.

13. What is adjusted R2 in linear regression. How is it calculated?
Ans: Adjusted R2 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. Adjusted R squared 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.

MACHINE LEARNING

14. What is the difference between standardisation and normalisation?
    Ans: In Normalisation, the change in values is that they are at a standard scale without distorting the differences in the values. Whereas, Standardisation measures the variable at different scales, making all the variables equally contribute to the analysis.

15. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation.
    Ans: Cross-validation is a technique that allows us to utilize our training data better for training and evaluating the model.

    Advantage : Cross-Validation is a very powerful tool. It helps us better use our data, and it gives us much more information about our algorithm performance.

    Disadvantage: 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.