1 In Q1 to Q7, only one option is correct, Choose the correct option:

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 C) ) between -1 and 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 B) PCA

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

A) linear B) Radial Basis Function

C) hyperplane D) polynomial

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 A) Logistic Regression

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 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?

A) remains same B) increases

C) decreases D) none of the above

B) 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

C) Random Forest are easy to interpret

In Q8 to Q10, more than one options are correct, Choose all the correct options:

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

Ans B) and C)

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 and 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) max\_depth. B) max\_features D) min\_samples\_leaf

MACHINE LEARNING

Q10 to Q15 are subjective answer type questions, Answer them briefly.

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

Outliers are datapoints that differs significantly from the rest of the datapoints, Inter Quartile Range is a statistical method for understanding the spread of the data and is the difference between 75th and 25th percentiles. It is one of the methods for detection of outliers. Any datapoints 1.5 times the IQR beyond 75th or 25th percentile is considered as an outlier.

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

While both are ensemble learning technique bagging utilises weak learners that work independently in parallel whereas Boosting utilises weak learners but sequentially to get model prediction.

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

Ans In linear regression, R2 is used in estimating the goodness of fit of linear regression. In R2 it considers every variable are able to explain the dependent variable. Adjusted R2 penalizes attributes that are uncorrelated.

Calculation:

R2 Adjusted =

Where R2 = sample R- square

p = Number of predictors

N = Total sample size

14. What is the difference between standardisation and normalisation?

Standardization and normalisation are both feature scaling techniques which is used for bringing the values of independent variables to fixed range which is especially important in algorithms whose prediction is computed using distance such as knn, k-means clustering etc.

We use standardisation for scaling normally distributed data. In standardisation, the values are rescaled such that the mean is equal to zero and standard deviation is 1 hence obtaining values that have a Gaussian distribution.

Normalisation or Min-Max Scaling rescales such that the values are bounded between 0 and 1.

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

Ans. Cross validation is a technique of re- evaluating our predicted values by training it on several subsets of the data and ensure that our model is performing well and has generalized a pattern if they had similar performances. The advantage of cross validation is it avoids over-fitting. One dis advantage of cross validation is it can get costly especially if we use better cross validation techniques.