MACHINE LEARNING - 4

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- C) Recursive feature elimination

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

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-

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) Principal Components are calculated using unsupervised learning techniques
C) Principal Components are linear combinations of Linear Variables.

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) 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

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

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.

Ans-Outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In other words they are the unusual values in a dataset.
IQR=Q3-Q1.IQR is the range between first and third quartile range. The datapoints which falls belowQ1-1.5 and above Q3+1.5 are outliers.

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

Ans-Bagging is the method of merging same type of predictions and boosting is a method of merging different types of predictions. In Bagging the result is obtained by averaging the responses of the N learners. Boosting assigns a second set of weights, this time for the N classifiers in order to take a weighted average of their estimates.

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

Ans-Adjusted R squared is a modified version of R-squared that has been adjusted for the number of predictors in the model.Adjusted R-squared value can be calculated based on value of rsquared.Everytime you adda an independent variable to a model ,R squared increases, even if the independent variable is insignificant.

14. What is the difference between standardisation and normalisation?

Ans-Normalization means rescales the value into a range of [0,1] . Standardization means rescales data to have a mean of 0 and a std deviation of 1.

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

Ans- Cross validation is a technique for assessing how statistical analysis generalises to an independent dataset.It is a technique for evaluating machine learning models by training several models on subsets of the available input data and evaluating them on complementary subset of data.
Advantage –It make use of all datapoints and hence it is low bias.
Disadvantage-It leads to higher variation in testing model as we are testing against one data point