MACHINE LEARNING

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

Answer: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

Answer: Lasso Regularisation

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

Answer: 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

Answer: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

Answer: 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

Answer: 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
   Answer: D Random Forests provide a reliable feature importance estimate

MACHINE LEARNING

8.

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

9. 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

Answer: D

10. 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.

Answer: B, C , D

11. 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

Answer a,b, c

MACHINE LEARNING

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

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

Answer: Outliers are data points which are beyond the normal range of i.e. beyond +-3 SD. If any value beyond q1 or q3 that is considered as an outlier.

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

Answer: Bagging decreases variance, not bias, and solves over-fitting issues in a model. Boosting decreases bias, not variance.

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

Answer: Adjusted R Square determines the extent of the variance of the dependent variable, which the independent variable can explain. By looking at the adjusted R^2 value, one can judge whether the data in the regression equation is a good fit. The higher the adjusted R^2 the better the regression equation as it implies that the independent variable chosen to determine the dependent variable can explain the variation in the dependent variable.

15. What is the difference between standardisation and normalisation?

Answer: Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance)

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

Answer : Cross Validation is a technique of measuring overfitting in the model. Some times it may not provide good result.