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. D) Ridge 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.A) linear

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. D) Cannot be determined

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

MACHINE LEARNING

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. 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

   Answer.
   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.

   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

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

    Answer. C) n_estimators

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.

Answer.

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. IQR is the range between the first and the third quartiles namely Q1 and Q3. IQR=Q3-Q1.The data points which you 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?

Answer.
Bagging is the simplest way of combining predictions that belongs to the same type, while boosting is a way of combining predictions that belong to the different types. Bagging aims to decrease variance no bias, while boosting aims to decrease bias not variance.

13. What is adjusted R^2 in linear regression. How is it calculated?

Answer.
Adjusted R-squared,a modified version of R-Squared. adds precision and he reliability by considering the impact of additional independent variables that tend to skew their result of R-squared measurements. Adjusted R squared is calculated by dividing the residual mean square error by the total mean square error.

14. What is the difference between standardisation and normalisation?

Answer.
Normalization

Minimum and maximum value of features are used for scaling. It is used when features are of different scales.scales values between [0,1] or [-1,1].It is really affected by outliers.

Standarization

Mean and standard deviation is used for scaling. It is used when we want to ensure zero mean and unit standard deviation. It is not bounded to a certain range. It is much less affected by outliers.

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

Answer.
Cross validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments.One used to learn or train a model and the other used to validate the model.

Advantage

Reduces overfitting, Hyperparameter Tuning

Disadvantage

Increases Training Time, Needs Expensive Computation