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
   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
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.(A) 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.(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
   Ans.(B) Random Forests explains more variance in data then decision trees

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)(C).

MACHINE LEARNING

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.
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).(D)

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

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

Ans. An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal.

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.

Formula:-

higher side ==> Q3 +(1.5* IQR)

Lower side ==> Q1 -(1.5*IQR)

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

Ans. **Bagging :**

- Bagging is used when our objective is to reduce the variance of a decision tree.
- Various training data subsets are randomly drawn with replacement from the whole training dataset.
- Every model receives an equal weight.

Boosting :

- Boosting is another ensemble procedure to make a collection of predictors.
- Each new subset contains the components that were misclassified by previous models. Boosting tries to reduce bias.
- Models are weighted by their performance.

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

Ans. The adjusted R-squared is a modified version of R-squared that adjusts for predictors that are not significant in a regression model.

Compared to a model with additional input variables, a lower adjusted R-squared indicates that the additional input variables are not adding value to the model.

$$R^2_{adjusted} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

where

R^2 = sample R-square

p = Number of predictors

N = Total sample size.

MACHINE LEARNING

4. What is the difference between standardisation and normalisation?

Ans.

S.NO.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

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

Ans. 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.

Advantages of Cross Validation

1. Reduces Overfitting: In Cross Validation, we split the dataset into multiple folds and train the algorithm on different folds. This prevents our model from overfitting the training dataset. So, in this way, the model attains the generalization capabilities which is a good sign of a robust algorithm.

2. Hyperparameter Tuning: Cross Validation helps in finding the optimal value of hyperparameters to increase the efficiency of the algorithm.

Disadvantages of Cross Validation

1. Increases Training Time: 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.

2. Needs Expensive Computation: Cross Validation is computationally very expensive in terms of processing power required.

MACHINE LEARNING