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)
2. Which of the following cannot be used for dimensionality reduction?
   A) Lasso Regularisation B) PCA
   C) Recursive feature elimination D) Ridge Regularisation
   Answer:- 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
   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:- B) same as old coefficient of 'X'
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:- A) remains same
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:- 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
   Answer:- B) Principal Components are calculated using unsupervised learning techniques
   C) Principal Components are linear combinations of Linear Variables

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.

   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:- A) max_depth
    D) min_samples_leaf

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.

   Answer:- Any observations that are more than 1.5 IQR below Q1 or more than 1.5 IQR above Q3 are considered outliers.

   IQR (Interquartile Range) is the difference between the third and the first quartile of a distribution (or the 75th percentile minus the 25th percentile). It is a measure of how wide our distribution is since this range contains half of the points of the dataset. It's very useful to make an idea of the shape of the distribution. For example, it is the width of the boxes in the boxplot.

   - **Find the IQR(inter quantile range) to identify outliers**

   ```
   # 1st Quantile
   q1=data.quantile(0.25)
   # 3rd Quantile
   q3=data.quantile(0.75)
   #IQR
   iqr=q3-q1
   # Outlier Detection Formula
   # Higher Side= Q3+(1.5*iqr)
   # Lower Side= q1-(1.5*iqr)
   ```

MACHINE LEARNING

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

Answer:-The difference between bagging & boosting algorithms are listed below:-

S.no	Bagging	Boosting
1	Training data subsets are drawn randomly with replacement from the entire training data set	Each new subset contains the components that were misclassified by previous models
2	Bagging attempts to tackle the over fitting issue	Boosting tries to reduce bias
3	Every Model is built independently	New models are affected by the performance of the previously developed model
4	Every model receives an equal weight	Models are weighted by their performance
5	Objective to decrease variance, not bias	Objective to decrease bias, not variance
6	**Model is built sequentially**	**Model is built parallely**

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

Answer:-

Adjusted R2 is a corrected goodness-of-fit (model accuracy) measure for linear models. It identifies the percentage of variance in the target field that is explained by the input or inputs. R2 tends to optimistically estimate the fit of the linear regression. Selecting the model with the highest value of R squared is not correct approach as the value of R squared shall always increase whenever a new feature is been taken into consideration, so the alternative is to use adjusted R2 which penalizes the model complexity.

Adjusted R squared is calculated by dividing the residual mean square error by the total mean square error . The result is then subtracted from 1. Adjusted R2 is always less than or equal to R2.

4. What is the difference between standardisation and normalisation?

Answer:-

The difference is: in **scaling**, we are changing the range of your data, while in **normalization**; we are changing the shape of the distribution of your data

MACHINE LEARNING

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

Answer:-

There are chances that the model is over fitted and under fitted because your model has trained itself on given data. It has seen the data before and thus it fails to generalize well over it.

So to avoid over fitting and under fitting of data we use cross validation.

Advantage: Cross-Validation is a very powerful tool. It helps us better use our data, and it gives us much more information about our algorithm performance

Disadvantage: Higher Training Time: with cross-validation, we need to train the model on multiple training sets. Expensive Computation: Cross validation is computationally very expensive as we need to train on multiple training sets

MACHINE LEARNING