MACHINE LEARNING

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: A) 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: 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: C) Decision Tree 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

ANSWER: a) 2.205 × 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: 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

ANSWER: c) Random Forests 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), D), B), C)

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), B) AND D)

Q11 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: Outlier detection is an important part of many machine learning problems. The quality and performance of machine learning models depend on the quality of data. However, datasets often contain poor quality samples, noisy points, or outliers. Outliers are points that do not fit well with the rest of the data.

    INTER QUARTILE RANGE (IQR).

    This is a very simple technique using statistical measures. If you've ever studied a boxplot, you know what the terms median, percentile, and interquartile range mean. Box plots show the distribution of the data. Quartiles are measured every 25 percent of the total data points. The first quartile represents the 25th percentile of values, the second represents the median or 50th percentile, the third and his fourth quartile represent the 75th and 100th percentiles respectively (maximum value). The distance between the 1st and 3rd quartiles therefore represents the range of the central 50% values, the so-called interquartile range.

    Finding outliers is easy. Find the interquartile range and choose a multiplier k, usually equal to 1.5. Ranges of values above Q3 + K*IQR and below Q1 – K*IQR are considered outliers.

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

S.NO	Bagging	Boosting
1.	Bagging is a learning approach that aids in enhancing the performance, execution, and precision of machine learning algorithms.	Boosting is an approach that iteratively modifies the weight of observation based on the last classification.
2.	It is the easiest method of merging predictions that belong to the same type.	It is a method of merging predictions that belong to different types.
3.	Here, every model has equal weight.	Here, the weight of the models depends on their performance.
4.	In bagging, each model is assembled independently.	In boosting, the new models are impacted by the implementation of earlier built models.
5.	It helps in solving the over-fitting issue.	It helps in reducing the bias.
6.	In the case of bagging, if the classifier is unstable, then we apply bagging.	In the case of boosting, If the classifier is stable, then we apply boosting.

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

ANSWER: Adjusted R-squared is a modified version of R-squared that accounts for predictors that are not significant in the regression model. In other words, the fitted coefficient of determination indicates whether adding additional predictors improves the regression model. To understand the fitted R-squared, we need to understand R-squared.

Adjusted R-squared

Where k is the number of regressors and n is the sample size.

Overwhelms the reduction by k when the newly added variables are sufficient to improve the model's performance. Otherwise, increasing k decreases the adjusted r-squared value.

$$R^2_{adj} = 1 - [\frac{(1 - R^2)(n-1)}{n-k-1}]$$

The adjusted coefficient of determination determines the amount of variance in the dependent variable that the independent variable can explain. You can use the fitted R^2 values to assess whether the data fit the regression equation well. The higher the adjusted R^2, the better the regression equation, because it means that the independent variables chosen to determine the dependent variable can explain the variation in the dependent variable.

The corrected R^2 value can be negative, but it is not always negative. In the fitted R-squared, the variation of the independent variables is

Affects the variation of the dependent variable. Not applicable for R^2, only relevant for adjusted R^2 values.

14. What is the difference between standardisation and normalisation?
ANSWER:

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.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

15. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation.
ANSWER: Cross-validation in machine learning is a great technique for dealing with overfitting problems in various algorithms. Instead of training the model on one training dataset, train the model on many datasets. Below are some of the pros and cons of cross-validation in machine learning.

1. Reduce overfitting:
Cross-validation divides the dataset into multiple folds and trains the algorithm on the different folds. This prevents the model from overfitting the training data set. This is how the model gains generalization ability and is a good sign of a robust algorithm.

2. Hyper parameter tuning:
Cross-validation helps find the optimal values of hyper parameters and increases the efficiency of algorithms.