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 : D) 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

   Answer:

FLIP ROBO

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 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 : D) All of the above

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 : all of the above

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 B) max_features D) min_samples_leaf

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 a mathematical value that stands out from the other values in a set of data. Outliers are values that are unusually distant from the middle, to put it simply.

By dividing a data set into quartiles, the IQR is used to measure variability. The information is divided into 4 equal portions and arranged in ascending order. The values that divide the four equal halves are known as the first, second, and third quartiles, or Q1, Q2, and Q3, respectively.

The data's 25th percentile is represented by Q1.

The data's 50th percentile is represented by Q2.

The 75th percentile of the data is represented by Q3.

If a dataset contains 2n/2n+1 data points, then Q1 is the dataset's median.

Q2 is the average of the n smallest data points.

Q3 is the average of the n highest data points.

The interquartile range, or IQR, is the space between the first and third quartiles, or Q1 and Q3: IQR = Q3 - Q1. Outliers are data points that are either below or above the median (Q1 - 1.5 IQR or Q3 + 1.5 IQR).

Consider the following data: 6, 2, 1, 5, 4, 3, 50. 50 is obviously an aberration if these statistics indicate the number of donut consumed during lunch.

Capping and flooring to remove such Outliers.

FLIP ROBO

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

Answer:

Bagging	Boosting
Various training data subsets are randomly drawn with replacement from the whole training dataset.	Each new subset contains the components that were misclassified by previous models.
Bagging attempts to tackle the over-fitting issue.	Boosting tries to reduce bias.
If the classifier is unstable (high variance), then we need to apply bagging.	If the classifier is steady and straightforward (high bias), then we need to apply boosting.
Every model receives an equal weight.	Models are weighted by their performance.
Objective to decrease variance, not bias.	Objective to decrease bias, not variance.
It is the easiest way of connecting predictions that belong to the same type.	It is a way of connecting predictions that belong to the different types.

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

Answer: A variant of R-squared that has been changed to account for the number of predictors in the model is known as adjusted R-squared. When the additional term enhances the model more than would be predicted by chance, the adjusted R-squared rises. When a predictor enhances the model by less than anticipated, it falls.

R-Squared (R^2 or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model

MACHINE LEARNING

$$R^2 = 1 - \frac{RSS}{TSS}$$

R^2 = coefficient of determination

RSS = sum of squares of residuals

TSS = total sum of squares

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.	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.
6.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
7.	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 is a statistical technique for assessing and contrasting learning algorithms that involves splitting the data into two sections: one for learning or training a model and the other for model validation.

Cross-validation benefits include a more precise assessment of out-of-sample accuracy. Data is used more "efficiently" because every observation is used for both testing and training.

The drawback of this approach is that the evaluation process requires k times as much computing because the training algorithm must be run k times from begin. This method can be modified by randomly dividing the data k times into a test and training set.