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

2. Which of the following cannot be used for dimensionality reduction?

   A) Lasso Regularisation B) PCA

   C) Recursive feature elimination D) Ridge Regularisation

Ans.) D

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

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.) D

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’ willbe?

   (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.)

6. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy ofthe model?

   A) remains same B) increases

   C) decreases D) none of the above

Ans.) B

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

MACHINE LEARNING

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

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 sugarlevels.

   Ans.) ALL

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, B & C

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.

Ans.) An outlier is an observation that lies at an abnormal distance from the other values in a values in a random sample from a population.

The IQR describes the middle 50% of values when ordered from lowest to highest. To find the interquartile range (IQR), first find the median (middle value) of the lower and upper half of the data. These values are quartile 1 (Q1) and quartile 3 (Q3). The range is calculated by subtracting the lowest value from the highest value. While a large range means high variability, a small range means low variability in a distribution. The IQR is the difference between Q3 and Q1.

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

Ans.)

Bagging	**Boosting**
Here various training data subsets are drawn randomly with replacement from the whole training dataset in independently created	Here each new subset contains the components that were misclassified by previous models. Thus, new models are

models.	affected by previous models.
We can apply this when the classifier is unstable.	We can apply this when the classifier has a high bias or is steady and straightforward.
It can tackle over-fitting.	It tries to reduce bias.
Each model is given equal performance.	Models are weighted as per their performance.
Decreases variance.	Decreases bias.
Can easily connect to predictions of a same group.	Can easily connect to predictions of different types.

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

Ans.) Adjusted R-squared (R2) helps measure linear models with model accuracy. The value is calculated on the basis of the value of r-squared, number of independent variables and total sample size. Thus, everytime an independent variable is added to a model, the r-squared increases. It never declines even if the independent variables are insignificant.

14. What is the difference between standardisation and normalisation?

Ans.)

Standardisation	**Normalisation**
Used with Gaussian distribution	Used when there is no Gaussian distribution
Considered when algorithms make assumptions about data distribution	Considered when algorithms do not make assumptions
Not bounded by range	Scales in a range of [0,1] or [1,1]
Slightly affected by outliers.	Highly affected by outliers.

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

Ans.) Cross validation or CV is one of the key aspects of testing learning models. It is a statistical method of evaluating and comparing learning algorithms by dividing data into training (to learn) and testing (to validate).

Advantage of Cross Validation:

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.

Disadvantage of Cross Validation:

Increases Training Time: Cross Validation drastically increases the training time as one has to train the model not only on one training set, but on multiple training sets. It also is very expensive at the same time.

****