

Q1. The value of correlation coefficient will always be:

Ans. (A) between 0 and 1

Q2. Which of the following cannot be used for dimensionality reduction?

Ans. (C) Recursive feature elimination

Q3. Which of the following is not a kernel in Support Vector Machines?

Ans. (A) linear

Q4. Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?

Ans. (D) Support Vector Classifier

Q5. 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?

Ans. (C) old coefficient of 'X' $\div 2.205$

Q6. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?

Ans. (B) increases

Q7. Which of the following is not an advantage of using random forest instead of decision trees?

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

Q8. Which of the following are correct about Principal Components?

Ans. (D) All of the above

Q9. Which of the following are applications of clustering?

Ans. (A), (C), (D)

Q10. (A), (C), (D)

Q11. What are outliers? Explain the Inter Quartile Range (IQR) method for outlier detection.

Ans. Outliers are those value in data which are significant different from rest data in the datasets. Inter Quartile is the difference of first quartile (Q1) and third quartile(Q3). After find IQR, Q1 and Q3, then calculate lower fence and upper fence with below formulas.

Lower fence = $Q1 - 1.5(IQR)$

Upper Fence = $Q3 + 1.5(IQR)$

After find lower fence and Upper fence the all data are treated as outliers which li lower than lower fence and higher than upper fence (outliers <lower fence and outliers > upper fence)

Q12. What is the primary difference between bagging and boosting algorithms?

Ans. In bagging, training data subset are drawn randomly with replacement with entire training dataset. But in boosting, each new training data subset has a component which is misclassified by the previous model.

In bagging, tackle the overfitting issue, but in Boosting, reduced the bias.

In Bagging, Objective to decrease variance, no effect in bias, but in boosting, objective to decrease in bias, no effect in variance.

Q13. What is adjusted R² in linear regression. How is it calculated?

Ans. Adjusted R² is an upgraded version of R². It has been adjusted number of predictions in model.

Formula: - Adjusted R² = $1 - \frac{(1 - R^2)(N - 1)}{(N - p - 1)}$

Where- R² → simple R²

N → Total sample Size

p → Number of independent variables

Q14. What is the difference between standardisation and normalisation?

Ans. In Normalisation, value changed a standard scale without distorting its difference values.

In Standardisation, change its value with data mean and standard deviation.

Q15. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation.

Ans. Cross-validation is a technique which allows us to utilize our training data for better training and evaluate the model.

Advantage: Cross validation gives an idea, how to model generalize for unknown data set.

Disadvantage: It takes a long time for train model