

MACHINE LEARNING

1Q. The value of correlation coefficient will always be:

Ans. C) between -1 and 1

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

Ans. B) PCA

3Q. Which of the following is not a kernel in Support Vector Machines?

Ans. A) linear

4Q. Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?

Ans. A) Logistic Regression

5Q. 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)

Ans. B) same as old coefficient of 'X'

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

Ans. B) increases

7Q. Which of the following is not an advantage of using random forest instead of decision trees?

Ans. C) Random Forests are easy to interpret

8Q. Which of the following are correct about Principal Components?

Ans. B) Principal Components are calculated using unsupervised learning techniques

C) Principal Components are linear combinations of Linear Variables.

9Q. Which of the following are applications of clustering?

Ans. 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.

D) Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.

10Q. Which of the following is(are) hyper parameters of a decision tree?

Ans. A) max_depth
D) min_samples_leaf

11Q. What are outliers? Explain the Inter Quartile Range (IQR) method for outlier detection.

Ans. An outlier is a value or point that differs substantially from the rest of the data. Inter Quartile Range (IQR) is the range between the first and third quartile namely Q1 and Q3. (IQR = Q3 – Q1) the data point fall below Q1 – 1.5 or above Q3 + 1.5 IQR are outlier.

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

Ans. Bagging is a method of merging the same type of predictions. Boosting is a method of merging different types of predictions. Bagging decreases variance, not bias, but. Boosting decreases bias, not variance.

13Q. What is adjusted R² in linear regression. How is it calculated?

Ans. Adjusted R² is a corrected goodness-of-fit or model accuracy measure for linear models. It identifies the percentage of variance in the target field that is explained by the inputs.

Adjusted R Squared = $1 - [((1 - R^2) * (n - 1)) / (n - k - 1)]$

14Q. What is the difference between standardization and normalization?

Ans.

- 1) In case of standardization mean and standard deviation is used for scaling while in case of normalization Minimum and maximum value of features are used for scaling
- 2) standardization is used when we want to ensure zero mean and unit standard deviation while normalization is used when features are of different scales.
- 3) It is not bounded to a certain range while Scales values between [0, 1] or [-1, 1].
- 4) standardization is much less effected by outliers while normalization is really effected by outliers.

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

Ans. Cross-validation is a technique which is used statistical method for training data better for training and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model.

As an 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. In complex machine learning model

As a disadvantage: Cross-validation is computationally very expensive as we need to train on multiple training sets