MACHINE_LEARNING IV

The value of correlation coefficient will always be:

between -1 and 1

Which of the following cannot be used for dimensionality reduction?

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

C Recursive feature elimination

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

ANS (C) Decision Tree Classifier

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

ANS) increases

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

A) Random Forests reduce overfitting

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) All of the above

9. Which of the following are applications of clustering?

ANS) Identifying developed, developing and under-developed countries on the basis of factors like GDP,

poverty index, employment rate, population and living index

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

ANS) max_features

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

ANS Outliers are data points that lie outside the overall pattern in a distribution. Thus, a data point that is distant from the remaining data points in the sample is NOT necessarily an outlier. Instead, a data point deviating from the model fit (the pattern of underlying population) is an outlier.

One common technique to detect outliers is using IQR (interquartile range). In specific, IQR is the middle 50% of data, which is Q3-Q1. Q1 is the first quartile, Q3 is the third quartile, and quartile divides an ordered dataset into 4 equal-sized groups. In Python, we can use percentile function in NumPy package to find Q1 and Q3.

The interquartile range method defines outliers as values larger than Q3 + 1.5 * IQR or the values smaller than Q1 – 1.5 * IQR.

Say we have collected the midterm grade of 500 students and stored the data in an array called grades. We want to know if there are students getting extremely high or extremely low score. In other words, we want to find the outliers in terms of midterm grade.

First, we use percentile function to find Q1 and Q3. The first argument is the data, and the second argument is the percentiles to compute. We can either pass 1 percentile at a time (method 1) or store the percentiles we want to get in a list (method 2)

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

ANS -- In Bagging the result is obtained by averaging the responses of the N learners (or majority vote). However, Boosting assigns a second set of weights, this time for the N classifiers, in order to take a weighted average of their estimates. In the Boosting training stage, the algorithm allocates weights to each resulting model.

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

ANS --The adjusted R-squared is a modified version of R-squared that adjusts for the number of predictors in a regression model. It is calculated as: Adjusted R2 = 1 – [(1-R2)* (n-1)/ (n-k-1)] where: R2: The R2 of the model

14. What is the difference between standardisation and normalisation?

As nouns the difference between standardization and normalization is that standardization is the process of complying (or evaluate by comparing) with a standard while normalization is any process that makes something more normal or regular, which typically means conforming to some regularity or rule, or returning from some state of abnormality.

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

Advantages and Disadvantages of Cross Validation in Machine Learning

Cross Validation in Machine Learning is a great technique to deal with overfitting problem in various algorithms. Instead of training our model on one training dataset, we train our model on many datasets. Below are some of the advantages and disadvantages of Cross Validation in Machine Learning:

Advantages of Cross Validation

1. 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.

Note: Chances of overfitting are less if the dataset is large. So, Cross Validation may not be required at all in the situation where we have sufficient data available.

2. Hyperparameter Tuning: Cross Validation helps in finding the optimal value of hyperparameters to increase the efficiency of the algorithm.

Disadvantages of Cross Validation

1. Increases Training Time: Cross Validation drastically increases the training time. Earlier you had to train your model only on one training set, but with Cross Validation you have to train your model on multiple training sets.

For example, if you go with 5 Fold Cross Validation, you need to do 5 rounds of training each on different 4/5 of available data. And this is for only one choice of hyperparameters. If you have multiple choice of parameters, then the training period will shoot too high.

2. Needs Expensive Computation: Cross Validation is computationally very expensive in terms of processing power required.