MACHINE LEARNING (Assignment 4)

Answer 1:- C) between -1 and 1

Answer 2:- B) PCA

Answer 3:- C) hyperplane

Answer 4:- A) Logistic Regression

Answer 5:- A) 2.205 × old coefficient of 'X

Answer 6 :- B) increases

Answer 7:- C) Random Forests are easy to interpret

Answer 8:- B) Principal Components are calculated using unsupervised learning techniques

C) Principal Components are linear combinations of Linear Variables.

Answer 9:- C) Identifying spam or ham emails

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

Answer10:- A) max_depth

B) max_features

Answer11:- IQR is the range between the first and the third quartiles namely Q1 and Q3.

IQR = Q3 – Q1. The data points which fall below Q1 – 1.5 IQR or above Q3 + 1.5 IQR are outliers.

Once we calculate it, we can use IQR to identify the outliers. We label a point as an outlier if it satisfies one of the following conditions:

- It's greater than 75th percentile + 1.5 IQR
- It's less than 25th percentile – 1.5 IQR

Applying this simple formula, we can easily detect the outliers of our distribution. Boxplot uses the same method to plot the outliers as points outside the whiskers.

The reasons behind that 1.5 coefficient rely upon the normal distribution, but the general idea is to calculate outliers without using some measure that could be affected by them. That's why using, for example, the standard deviation, could lead us to poor results. Quartiles and percentiles are based on counts, so they are less vulnerable to the presence of outliers.

The idea is that if a point is too far from the 75th percentile (or from the 25th percentile), it's a "strange" point that can be labeled as an outlier. The order of magnitude of such a distance is the IQR itself.

Answer12:-

Bagging:-

Various training data subsets are randomly drawn with replacement from the whole training dataset.

If the classifier is unstable (high variance), then we need to apply bagging.

Objective to decrease variance, not bias.

Every model is constructed independently.

Boosting:-

Each new subset contains the components that were misclassified by previous models.
If the classifier is steady and straightforward (high bias), then we need to apply boosting.
Objective to decrease bias, not variance.
New models are affected by the performance of the previously developed model.

Answer13:- Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size. Every time you add a independent variable to a model, the R-squared increases, even if the independent variable is insignificant. It never declines.

Adjusted R Squared = 1 – [((1 – R2) * (n – 1)) / (n – k – 1)]

Answer14:- Standardization is the subtraction of the mean and then dividing by its standard deviation. While Normalization is the process of dividing of a vector by its length and it transforms your data into a range between 0 and 1.

Answer15:- a statistical method of evaluating and comparing learning algorithms by dividing data into two segments.

Advantage

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

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.