Machine Learning Worksheet-4

Ques1:

c) between -1 and 1

Ques2:

b) PCA

Ques3:

a) linear

Ques4:

a) Logistic Regression

Ques5:

a) 2.205 × old coefficient of 'X'

Ques6:

b) increase

Ques7:

c) Random Forests are easy to interpret

Ques8:

b) Principal Components are calculated using unsupervised learning techniques

c) Principal Components are linear combinations of Linear Variables.

Ques9:

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

Ques10:

a) max_depth

b) max_features

d) min_samples_leaf

Ques11:

Outlier is an observation that lies an abnormal distance from other values in random samples from a population.

IQR is the range between first and third quartile (IQR = Q3-Q1).

To detect above outliers we calculate by Q3 + 1.5 IQR and below outliers are calculated by Q1 – 1.5 IQR.

Ques12:

In Bagging, individual trees are independent of each other whereas in Boosting, individual trees are not independent of each other.

Bagging reduces variance whereas Boosting reduces both variance as well as biasness.

Ques13:

If there are more than one independent variable in our equation then we use adjusted R^2 to evaluate the fit of a linear model.

It is calculated by:

$$R^2\ adjusted = 1 - \frac{(1-R^2)(N-1)}{N-p-1}$$

Where,

R = sample R squared

p = no. of predictors

N = Total sample size

Ques14:

The transformer of Standardisation is StandardScaler whereas the transformer of Normalisation is MinMaxScaler.

Standardisation is not bounded to a certain range but for normalization the scales values between [0,1] or [-1,1].

Ques15:

Cross validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments one is for train and another one is test.

Advantage of cross validation is to protect against overfitting while predicting the model.

Disadvantage of the method is that it randomly divide the data into train and test set by k folds.