MACHINE LEARNING WORKSHEET 4

Q1 TO 10 OBJECTIVE ANSWERS

1] c) between -1 and 1

2] c) Recursive feature elimination

3] a) linear

4] a) logistic regression

5] d) cannot be determine

6] b) increases

7] b) random forests explains more variance in data then decision trees

8] b) principal components are calculated using unsupervised learning techniques

c) Principal components are linear combinations of linear variables

9] a) Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index

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

10] a) max_depth

b) max_features

Q11 TO 15 SUBJECTIVE ANSWERS

11] Outliers are observations that lie abnormally far away from other values in a dataset. There are values uncommonly far from the middle. It is an extremely high or low data point relative to the nearest data point and the rest of the neighboring co-existing values. They are a statistical method used to detect extreme data points from a given distribution of data. The IQR is the difference between the 25th percentile (Q1) and the 75th percentile (Q3) in a dataset. It measures the spread of the middle 50% of values.

12] Bagging tries to tackle the over-fitting problem. It is ensemble learning method that is generally used to reduce variance within a noisy dataset that a random sample of data in a training set is selected with replacement meaning that the single data points can be selected more than once.

Boosting tries to reduce bias. It is another ensemble process to create a set of predictors that can fit consecutive trees, generally random samples and at every phase the objective is to solve net error from the previous trees.

13] Adjusted R-squared determines the extent of the variance of the dependent variable, which the independent variable show the specialty of the R^2 is that it does not consider the impact of all independent variables but only those which impact the variation of the dependent variable. Thus, the value of R^2 can also be negative, though it is not always negative. The formula which is to calculate the adjusted R square of regression is below:

$$R^2 = \{(1 / N) * \Sigma [(x_i - x) * (Y_i - y)] / (\sigma_x * \sigma_y)\}^2$$

14] 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 to scale a variable to have a values between a desired range while standardization transforms data to have a mean of zero and a standard deviation of 1.

15] Cross validation is a resampling method that uses different portions of the data to test and train a model on different iterations. One advantage: They are more accurate estimate of out-of-sample accuracy and more efficient use of data.

One disadvantage: The training algorithm has to be rerun from scratch k times.

THANK YOU