Machine Learning WORKSHEET-4¶

1. ANS: C) between -1 and 1
2. ANS: C) Recursive feature elimination
3. ANS: C) hyperplane
4. ANS: D) Support Vector Classifier
5. ANS: C) old coefficient of 'X' ÷ 2.205
6. ANS: B) increases
7. ANS: B) Random Forests explains more variance in data then decision trees
8. ANS B) Principal Components are calculated using unsupervised learning techniques

   C) Principal Components are linear combinations of Linear Variables.

9. 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.

   C) Identifying spam or ham emails

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

10. ANS: A) max_depth B) max_features

    D) min_samples_leaf

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

ANS: An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations.

IQR is used to measure variability by dividing a data set into quartiles. The data is sorted in ascending order and split into 4 equal parts. Q1, Q2, Q3 called first, second and third quartiles are the values which separate the 4 equal parts.

Q1 represents the 25th percentile of the data. Q2 represents the 50th percentile of the data. Q3 represents the 75th percentile of the data. If a dataset has 2n / 2n+1 data points, then Q1 = median of the dataset. Q2 = median of n smallest data points. Q3 = median of n highest data points.

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.

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

Machine Learning WORKSHEET-4¶

ANS: Bagging is a homogeneous weak learners' model that learns from each other independently in parallel and combines them for determining the model average.

Boosting is also a homogeneous weak learners' model but works differently from Bagging. In this model, learners learn sequentially and adaptively to improve model predictions of a learning algorithm.

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

ANS: The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance.

Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size.

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

Where:

n – Number of points in your data set. k – Number of independent variables in the model, excluding the constant

14. What is the difference between standardisation and normalisation?

--> In normalisation Minimum and maximum value of features are used for scaling, In standarisation Mean and standard deviation is used for scaling.

Normalisation is used when features are of different scales. Standarization is used when we want to ensure zero mean and unit standard deviation.

In normalisation Scales values between [0, 1] or [-1, 1]. standarisation is not bounded to a certain range.

Normalisation is really affected by outliers. Standarisation is much less affected by outliers.

Scikit-Learn provides a transformer called MinMaxScaler for Normalization. Scikit-Learn provides a transformer called StandardScaler for standardization.

Normalisation transformation squishes the n-dimensional data into an n-dimensional unit hypercube. Standarisation translates the data to the mean vector of original data to the origin and squishes or expands.

Normalisation is useful when we don't know about the distribution. Standarisation is useful when the feature distribution is Normal or Gaussian.

Normalisation is a often called as Scaling Normalization. Standarisation is a often called as Z-Score Normalization.

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

--> Cross validation is a technique for assessing how the statistical analysis generalises to an independent data set.It is a technique for evaluating machine learning models by training several

models on subsets of the available input data and evaluating them on the complementary subset of the data.

Advantage:-

An advantage of using this method is that we make use of all data points and hence it is low bias.

Disadvatage:-

The major drawback of this method is that it leads to higher variation in the testing model as we are testing against one data point. If the data point is an outlier it can lead to higher variation.