Data Science and Business Analytics
The Sparks Foundation

Task 2 : Prediction using Unsupervised ML

Problem : From the given ‘iris’ data set, predict the optimum number of clusters and represent it visually.

Shivam Deshpande

Clustering

- Clustering is a process of dividing the dataset into the groups that Consists of similar data points
- Points in the same group are as similar as possible and points in different Groups are as different as possible.
- Unsupervised ML technique

Real life Applications

- Amazon
  Recommendation system uses the clustering to show you the recommended Products based on the past purchase history.
- Netflix
  It recommends you the movies based on previous watch history.

K-Means Clustering

- Exclusive Clustering method
- The data points or items exclusively belongs to one cluster
- It’s goal is to group similar elements or data points into a cluster
- Applied to numeric or continuous data
- K = Number of clusters
- Steps : 1) Select the number of clusters i.e k
  2) Select k data points as initial cluster centroids
  3) Take new data point and measure its distance from each cluster centre
  4) Add the selected data point to the nearest cluster based on the distance Measured.

  Repeat the steps for each data point

Elbow Method

- Define clusters such that the total intra cluster variation or the total Within Cluster Sum of Squares (WCSS) is minimized.
- This measures the compactness of clustering
- Elbow method looks at WCSS as the function of number of clusters
- Steps : 1) Compute K-Means algorithm by using different values of k
  Vary k from 1 to 11
  2) For each k calculate the total WCSS
  3) Plot the graph of WCSS with the number of clusters
  4) The location of a bend or a knee in the plot gives us the optimum number of clusters for a given data set.

Thank You.