THE SPARKS FOUNDATION

DATA SCIENCE AND BUSINESS ANALYTICS INTERNSHIP

SUBMITTED BY – UMANG AGGARWAL

Task Name : Prediction using Unsupervised ML

1. From the given 'Iris' dataset, predict the optimum number of clusters and represent it visually.
2. Use R or Python or perform this task.

Solution

Step 1: Import Libraries

We'll import necessary libraries like pandas for making data , sklearn for loading dataset and for implementing K-Means algorithm, and matplotlib for plotting scatter plots.

Step 2: Load Iris Dataset

We'll load iris dataset from sklearn default datasets.

Step 3: Define Target and Predictors

Step 4: Explore Dataset

We'll explore dataset by just plotting a scatter plot of dataset and will see all the data points.

Step 5: Finding Optimum No. of Clusters

We'll use elbow method to determine optimal number of clusters. The basic idea behind k-means clustering, is to define clusters such that the total within-cluster sum of square (WSS) is minimized. The total WSS measures the compactness of the clustering and we want it to be as small as possible.

The Elbow method looks at the total WSS as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't improve much better the total WSS.

The optimal number of clusters can be defined as follow:

1. Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
2. For each k, calculate the total within-cluster sum of square (wss).
3. Plot the curve of wss according to the number of clusters k.
4. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

Step 6: Applying K-Means

After finding optimal no. of clusters we'll apply K-Means algorithm on dataset.

K means works through the following iterative process:

1. Pick a value for k (the number of clusters to create).
2. Initialize k 'centroids' (starting points) in your data.
3. Create your clusters. Assign each point to the nearest centroid.
4. Make your clusters better. Move each centroid to the center of its cluster.
5. Repeat steps 3–4 until your centroids converge.

Step 7: Visualizing Clusters

In the end we'll visualize the clusters.