FELLOWSHIP PROGRAM IN AI/ML

USE CASE - DAY 14: 23.08.2018

Use case 1:

Description:

We have given a dataset for you to understand Kmeans clustering algorithm and to visualise how kmeans works.

Exercise 1 : Import the data and plot it using matplotlib and visualise the data. (The data is very much cleaned and you wont get such data in real life examples)

Exercise 2: Initialise three clusters to start k - means and plot the three cluster points along with the previous data set.

Exercise 3: K-means algorithm iteratively performs these two steps:

- The first step assigns clusters to points by assigning them to the cluster with nearest centriod.
- The second step calculates the new mean from the points belonging to the cluster.
- Repeat these two steps till convergence.
- Visualise each step

Exercise 4 : Change number of clusters and see how plot is changing.

Use case 2:

Description:

In this experiment, we discuss hierarchical clustering methods. Heirarchial clustering algorithms build nested clusters by repeatedly merging two clusters in the bottom up approach and successively splitting a cluster into two in the top-down approach. This hierarchy of clusters is represented as a dendrogram. Dendrogram is a tree diagram popularly used to illustrate the arrangement of the heirarchy of clusters produced by hierarchical clustering algorithms.

MNIST is a classic dataset of handwritten images. It is a popular dataset used for benchmarking classification algorithms.

Hierarchical clustering generally fall into two types:

Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

Exercise 1 : Import the data set and visualise the clusters formed. Use from sklearn import datasets to import mnist data set and visualise the clusters formed

Exercise 2 : Use SpectralEmbedding to Embed in 2D and plot using agglomerative clustering

Exercise 3 : The linkage criteria determines the distance metric used for the merge strategy:

> Use ward which minimizes the sum of squared differences within all clusters. It is a variance-minimizing approach and in this sense is similar to the k-means objective function but addressed with an agglomerative hierarchical approach.
>
> Also check with other linkage criteria.

Exercise 4 : Plot dendogram

Exercise 5 : If we define the clusters to be the set of points with distance between each other equal to 20, visualise how many number of clusters will be obtained?