WORKSHOP 6 – Clustering

CMP3744M/CMP9732M – Algorithms for Data Mining

Overview

In this exercise, you are expected to write a Python code to implement the k-means algorithm. You can write you own code to implement the algorithm detailed in the lecture; you are also allowed to use existing software packages to implement the algorithm.

Tasks:

1) Download the file kmeans.csv from blackboard. There are two attributes in the data.
2) Show the scatter plots of the data; you will see that there are three clusters.
3) Write a Python code to implement the k-means algorithm on the data. You can use the Python module sklearn to import the k-means aglorithm, for example, `from sklearn.cluster import KMeans'.
4) Visualise the scatter plot clusters with different colours and their centroids.
5) Optimal number of clusters: The number of clusters has to be provided before applying k-means algorithm to the data. However, we do not aware the optimal number of clusters in the data. To find the optimal number of clusters, we can optimize the aggregate distance over all the data. Therefore, to find the optimal number, your task is to do the following steps: 1) vary the number of clusters K from 1 to 10, and apply k-means to the data for each K; 2) calculate the aggregate distance for each K; 3) visualise the aggregate distance against K; 4) select the optimal K where `mountain' ends and `rubbles' begins.