K-means Clustering Example

Xin Tang

Bellevue University

Abstract

This is a make-up example of how K-means clustering algorithm will help to solve a real-life problem. Even the data set and approach are makeup, the scenario and business need are real

## Few words on K-means clustering

Clustering is dividing data points into homogeneous classes or clusters:

* + Points in the same group are as similar as possible
  + Points in different group are as dissimilar as possible

The K-means clustering is used for unsupervised clustering, in which the groups/clusters have no pre-defined labels. When a collection of objects is given, this algorithm will help us to put objects into groups based on similarity.

## Business Scenario:

In a semiconductor equipment manufacturing site, with the high turn rate of work force, the assembly defect and safety incident increased. One major factor is that the new hires lack experience and there are not enough experienced workers assigned as their mentor/supervisor.

There is a plan to set aside some experienced workers as mentors and supervise the new hires during daily operation. We assume that the location of all the defect-prone work centers are known. We want to know the numbers of mentors that need to be assigned, as well as how many need to be assigned to each work center.

The challenge is that the experienced technicians are limited, taking additional supervise duty will reduce their productivity and reduce the overall manufacturing output. As such, we cannot afford to assign them based on a fixed number of technicians or build centers. We want to assign a minimum quantity of them to this duty, and cover as many new hires as possible, if one can cover multiple build centers, it will be even better.

K-means Clustering will group these defects into clusters and define a cluster center for each cluster, which will be the experienced technicians assigned. These clusters centers are the centroids of each cluster and are at a minimum distance from all the points of a particular cluster, which means a group of mistake prone new hires. Henceforth, the experienced technicians assigned will be close to all the new hires who need help and have all the skillset needed to help the technicians in the same cluster. At the same time, each cluster will share least similarity between each other, which will maximum the purpose of each mentor assigned.

The data available will be the number of build mistakes cases, the number of new hires, the numbers of work centers. After running the algorithm, the K will be defined, which is the number of mentors needed.