REPRESENTATIVE-BASED CLUSTERING

Given the dataset with m points in a d-dimensional space, $D = \{x_i\}_{i=1}^{m}$, and given the number of desired clusters k, the goal of representative-based clustering is to partition the dataset into k groups or clusters, which is called a clustering and is denoted as $C = \{C_1, C_2, .., C_k\}$. In addition for each cluster C_i there exist a representative point that summarizes the cluster, a common choice being the mean (also called the centroid) μ_i of all points in the cluster, that is:

$$\mu_i = \frac{1}{m_i} \sum_{x_j = C_i} x_j$$

where $m_i = |C_i|$ is the number of points in cluster C_i.

In this class we describe one approach for representative-based clustering, namely the K-means algorithm.

K-MEANS CLUSTERING ALGORITHM

Given a clustering $C = \{C_1, C_2, .., C_k\}$ we need some scoring function that evaluates its quality or goodness. The sum of squared errors scoring function is defined as:

$$SSE(C) = \sum_{i=1}^{k} \sum_{x_j \in C_i} \left\| x_j - \mu_i \right\|^2$$

The goal is to find a clustering that minimizes the SSE score:

$$C^* = argmin_c(SSE)$$

- K-means initializes the cluster means by randomly generating k points in the data space. This is done by generating a value uniformly at random within the range for each dimension.
- Each iteration of K-Means consists of two steps:
    - Cluster assignment
    - Centroid update

Given the k cluster mean, in the cluster assignment step, each point $x_i \in D$ is assigned to the closest mean, which induces a clustering, with each cluster C_i comprising points that are closer to μ_i than any other cluster mean.

This means each point x_i is assigned to cluster C_{j^*}, where

$$J^* = argmin_{i=1}^{k}\left\{\left\|x_j - \mu_i\right\|^2\right\}$$

Given a set of cluster $C_i, i = 1, \dots, k$, in the centroid update step, new mean values are computed for each cluster from the points in C_i. The

The cluster assignment and centroid update steps are carried out iteratively until we reach a fixed point or local minima. Though practically speaking, one can assume that K-means has converged if the centroids do not change from one iteration to the next.

===

K-MEANS CLUSTERING ALGORITHM PSEUDO CODE

$t = 0$

Randomly initialize k centroids: $\mu_1^t, \mu_2^t, \dots, \mu_k^t \in \mathbb{R}^d$

Repeat:

$t \leftarrow t + 1$

$C_j \leftarrow \emptyset$ for all $j = 1, \dots, k$

// Cluster Assignment Step

foreach $x_j \in D$ **do**

$j^* \leftarrow argmin_i\left\{\left\|x_j - \mu_i^t\right\|^2\right\}$ // Assign x_j to the closest centroid

$C_{j^*} \leftarrow C_{j^*} \cup \{x_j\}$

// Centroid Update Step

foreach $i = 1 \ to\ k$ **do**

$\mu_i^t \leftarrow \frac{1}{|C_i|}\sum_{x_j \in C_i} x_j$

Until $\sum_{i=1}^{k}\left\|\mu_i^t - \mu_i^{t-1}\right\|^2 \leq \varepsilon$

TEXTBOOK REFERENCE

Tan, P., Steinbach, M., Karpatne, A. & Kumar, V., 2019. *Introduction to Data Mining*. 2nd ed. New York: Pearson Education.

Zaki, M. J. & Wagner, M., 2014. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. New York: Cambridge University Press.