Erasmus University Rotterdam

Tatev Karen Aslanyan

K-Means Algorithm

K-means is one of the most well-known algorithms for cluster analysis, which is originally known as Forgy's method, (Forgy, 1965). (Sohrab, 2007) apply this clustering technique to model Customer Lifetime Value based on RFM attributes. (Cheng, 2009) combine RFM measures and K-means algorithm to improve classification accuracy and derive some classification rules. One of the main benefits of K-means clustering is the ease of implementation, its efficiency and the short running time compared to other clustering methods, for instance, the hierarchical clustering approach, especially if k is small. The simplicity of this approach makes it easy to explain the results in contrast to support vector machines or artificial neural networks while the flexibility of this method allows for easy adjust if problems occur. Moreover, K-means becomes a good solution for pre-clustering, reducing the space into disjoint smaller subspaces, where other clustering approaches can be applied. However, there are also some disadvantages of this clustering method. One of them is the requirement to pre-specify the number of clusters K. Moreover, the K-means approach is sensitive to outliers and determines the local optimum rather than global.

Hierarchical clustering is an alternative approach that does not require a particular choice of K which results in an attractive tree-based representation of the observations. However, with a large number of variables, K-means clustering may be computationally faster than hierarchical method especially if k is small. Additionally, hierarchical clustering is very sensitive to outliers, and it is not suitable for large data sets. The main idea behind the method is to assign points that are close to each other into the same cluster, checking the distance of each point from the center of the cluster that it belongs to and add all these distances. K-means clustering requires the number of clusters K to be pre-specified, and then the algorithm will assign each observation to precisely one of the K clusters. Specifically, the K-means clustering procedure results from an intuitive mathematical problem. Let $C_1, ..., C_K$ denote the sets that contain the indices of the observations in each cluster. These sets should satisfy the following properties:

1. $C_1 \cup C_2 \cup ... \cup C_K = 1, ..., n$.
2. $C_k \cap C_{k'} = \emptyset$, for all k $\neq k'$.

In other words, the clusters should be non-overlapping, and each observation should belong to at least one of the K clusters. Moreover, the clustering is optimal which has the smallest possible within-cluster variation. The within variation of C_k cluster is a measure $W(C_k)$ of the amount by which the observations in a cluster differ from each other. Therefore, the following optimization problem should be solved:

$$\min_{C_1,..,C_K} \sum_{k=1}^{K} W(C_k) \tag{0.1}$$

The within-cluster variation is defined using the squared Euclidean distance as follows:

$$W(C_k) = \frac{1}{\mid C_k \mid} \sum_{i,i \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \tag{0.2}$$

The number of observations in kth cluster is denoted by $\mid C_k \mid$. Thus, the optimization problem for K-means can be described as follows:

$$\min_{C_1,..,C_K} \sum_{k=1}^{K} \frac{1}{\mid C_k \mid} \sum_{i,i \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \tag{0.3}$$

The pseudocode of the K-means is presented in Algorithm 2.

Algorithm 1 K-means algorithm

```
Step 1: Assign each data point to a random cluster
$z^0_{i\bar{k}} = 1$ for $\bar{k} \in \{1, 2, ...., K\}$ and $z^0_{ik} = 0$ for $k \neq \bar{k}$ and k $\in \{1, 2, ...., K\}$
t=1

Step 2: while clusters change do
solve $\mu^t = \arg\min_\mu \sum_{k=1}^{K} \sum_{i=1}^{n} \mid\mid x_i - \mu_k \mid\mid^2 \underbrace{z^{t-1}_{ik}}_{\text{fixed}}$
where $\mu^t_k$ is the centroid of cluster k at iteration t-1.
solve $z^t = \arg\min_z \sum_{k=1}^{K} \sum_{i=1}^{n} \mid\mid x_i - \underbrace{\mu^t_k}_{\text{fixed}} \mid\mid^2 z_{ik}$
subject to $\sum_{k=1}^{K} z_{ik} = 1, i = 1, 2, ...., n$

where $z_{ik} \in \{0, 1\}, i = 1, 2, ...., n; k = 1, 2, ...., K$. Point i is assigned to the cluster with the
closest $\mu$.
t$\leftarrow t + 1$
```