Clustering Methods and PCA

Jiaqi Bi, University of Toronto

August 5, 2021

Contents

1 Why do we use Clustering?

Clustering is used for finding subgroups or clusters in a data set. It is a technique for data mining.

Example 1.1. We have a set of n observations, each with p features. The n observations could correspond to tissue samples for patients with breast cancer; these could be clinical measurements. We may have a reason to believe that there is some heterogeneity among the n tissue samples; perhaps there are a few different unknown subtypes of breast cancer. Clustering could be used to find these subgroups.

There are two best known clustering methods:

- K means clustering: We seek to partition the observations into a pre-specified number of clusters.
- Hierarchical clustering: We do not know in advance how many clusters we want

Definition 1.2. Dendrogram

A tree-like visual representation of the observations that allows us to view at once the clusterings obtained for each possible number of clusters from 1 to n.

1.1 K means clustering

$C_1, ..., C_K$ denote sets containing the indices i.e., $1, ..., K$ of the observations in each cluster. These sets have the following properties:

- $C_1 \cup C_2 \cup ... \cup C_K = \{1, ..., n\}$ Each observation belongs to at least one of the K clusters.
- $C_K \cap C_{K'} = \emptyset$ for all $k \neq k'$ The clusters are non-overlapping: no observation belongs to more than one cluster.

If the ith observation is in the kth cluster, call it $i \in C_k$. Note that K-means clustering is a good clustering as within-cluster variation is as small as possible.

Definition 1.3. Within-cluster variation

For cluster C_k is a measure $W(C_k)$ of the amount by which the observations within a cluster differ from each other. We need to solve:

$$\underset{C_1,...,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

Formula:

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

where $|C_k|$ is the number of observations in the kth cluster. Combining these two formulas will get a optimization problem that defines K means clustering:

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

The algorithm of K Means Clustering:

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
   - For each of the K clusters, compute the cluster centroid. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
   - Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance).

Following formulae explains why the algorithm is to decrease the value of the objective (optimization problem) at each step:

$$\frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = 2 \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$$

where $\bar{x}_{kj} = \frac{1}{|C_k|} \sum_{i \in C_k} x_{ij}$ is the mean for feature j in cluster C_k.

To perform K means clustering, we must decide how many clusters we expect in the data. The problem of selecting K is far from simple.

1.2 Hierarchical clustering

There is a con of K means clustering such that it needs us to pre specify the number of clusters K. The method of Hierarchical clustering does not require that we need a particular choice of K.

The most common type of hierarchical clustering:

- Bottom-up
- Agglomerative