Clustering

Nathanaël Carraz Rakotonirina

Mathématiques Informatique et Statistique Appliquées (MISA)
Université d'Antananarivo

Clustering

Clustering is an unsupervised learning approach. The inputs are $\mathcal{D} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$ where $\boldsymbol{x}_i \in \mathbb{R}^D$. Data samples are unlabeled.

Goal

The goal is to assign similar data points to the same cluster.

Applications of clustering:

- resembling genes
- image segmentation
- customer segmentation
- social sub-networks
- user profile

Hierarchical agglomerative clustering

- The input to the algorithm is an $N \times N$ dissimilarity matrix D ($D_{ij} > 0$ is the distance between groups i and j)
- Groups with small dissimilarity are grouped together in a hierarchical fashion.
- The output is a binary tree called **dendogram**.
- By cutting this tree at different heights, we can induce a different number of (nested) clusters.

The algorithm:

1. Initialize each sample as a cluster : $C_i = \{i\}$ for $i = 1$ to N
2. Initialize set of clusters available for merging : $S = \{C_1, ..., C_N\}$
3. Compute the dissimilarity matrix D according to some distance
4. Repeat until all samples are merged in one cluster
   - Pick the two closest clusters: (C_j, C_k)
   - Merge them into a new cluster : $C_l = C_j \cup C_k$
   - Remove C_j and C_k from the available clusters : $S = S \setminus \{C_j, C_k\}$
   - Add C_l to the available clusters : $S = S \cup C_l$
   - Update the dissimilarity matrix by adding the distance between C_l and the other available clusters.

How to measure dissimilarity between groups

- **Single link**: also called nearest neighbor clustering, the distance between two clusters is defined as the distance between the two closest members of each group:

$$d_{\min}(G,H) = \min\{d(i,j), i \in G, j \in H\}$$

- **Complete link**: also called furthest neighbor clustering, the distance between two clusters is defined as the distance between the two most distant pairs:

$$d_{\max}(G,H) = \max\{d(i,j), i \in G, j \in H\}$$

- **Average link**: the distance between two clusters is the average distance between all pairs

$$d_{\text{avg}}(G,H) = \frac{1}{n_G n_H} \sum_{i \in G} \sum_{j \in H} d(i,j)$$

- **Ward's distance**: the distance between two clusters is the distance between centers:

$$d_{\text{avg}}(G,H) = \sqrt{\frac{n_G n_H}{n_G + n_H}} d(\mu_G, \mu_H)$$

K means clustering

It computes similarity in terms of Euclidean distance to learned cluster centers $\mu_k \in \mathbb{R}^D$. Unlike hierachical agglomerative clustering:

- It is faster ($O(N)$ while average link HAC is $O(N^3)$).
- It optimizes a well-defined cost-function (it is a model and not just an algorithm).

The algorithm:

1. Initialize cluster centroids $\boldsymbol{\mu}_1, .., \boldsymbol{\mu}_K \in \mathbb{R}^D$ randomly
2. Repeat until convergence
   - Assign each data point $\boldsymbol{x}_i$ to the closest centeroid: $z_i = \arg\min_k ||\boldsymbol{x}_i - \boldsymbol{\mu}_k||_2^2$ (z_i is the index of the cluster assigned to $\boldsymbol{x}_i$)
   - Update the cluster centroids to the average of the points assigned to them : $\boldsymbol{\mu}_j = \frac{1}{n_j} \sum_{i:z_i=j} \boldsymbol{x}_i$

The loss can be formulated as

$$J(\boldsymbol{M}, \boldsymbol{Z}) = \sum_{i=1}^{N} ||\boldsymbol{x}_i - \boldsymbol{\mu}_{z_i}||^2 = ||\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{M}^\top||_F^2$$

where $\boldsymbol{X} \in \mathbb{R}^{N \times D}$, $\boldsymbol{Z} \in [0,1]^{N \times K}$, $\boldsymbol{M} \in \mathbb{R}^{D \times K}$ (its columns are the clusters $\boldsymbol{\mu}_j$)

Limitations

- The loss is not convex thus it may converge to a local minimum.
- Different initializations can lead to different clusters (different local minima).

Solution

It is run many times using different random initial values for the cluster centroids, and then, pick the best solution. There are other variants.


How are we going to select the number of clusters in K-means clustering ?
Fix it a priori
Elbow trick
Best reconstruction error on a validation set
imgflip.com

How to choose K ?

We cannot use the reconstruction error on a validation set as a way to select the best unsupervised model (like PCA).
It is difficult to asses the number of clusters because of a lack of a ground truth to compare with.

- Manually fixed a priori
- Elbow trick
- Ad-hoc metrics (silhouette score)

Explore further

Other variants of K-means

- K-means++
- K-medoids
- faster K-means

Other clustering methods

- Clustering with mixture models
- Spectral clustering
- DBScan