Clustering

Gilles Gasso

INSA Rouen - ASI Departement
Laboratory LITIS

September 29, 2021

Plan

Introduction

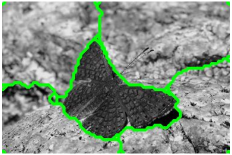
- $\mathcal{D} = \{\boldsymbol{x}_i \in \mathbb{R}^d\}_{i=1}^N$: set of training samples
- Goal : structure the data into homogeneous categories
  Group the samples into clusters so that samples in a cluster are as similar as possible
- Clustering $\equiv$ unsupervised learning

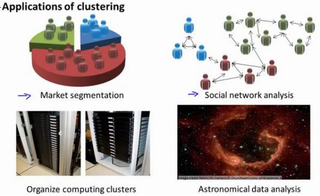
Clustering images

https://courses.cs.washington.edu/courses/cse416/18sp/slides/L11_kmeans.pdf

Applications

Field	Data type	Clusters
Text mining	Texts E-mails	Close texts Automatic folders
Graph mining	Graphs	Social sub-networks
BioInformatics	Genes	Resembling genes
Marketing	Client profile, purchased products	Customer segmentation
Image segmentation	Images	Homogeneous areas in an image
Web log analysis	Clickstream	User profile

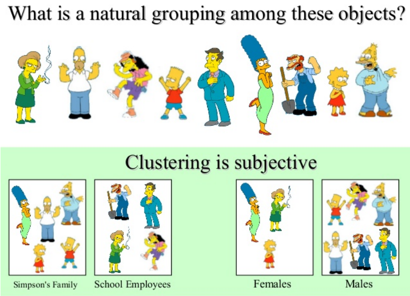




http://images2.programmersought.com/267/fc/fc00092c0966ec1d4b726f60880f9703.png

What is clustering ?

- How to define similarity or dissimilarity between samples
- How to characterize a cluster ?
- Number of clusters
- Which algorithms of clustering?
- How to assess a clustering result



https://image.slidesharecdn.com/k-means-130411020903-phpapp01/95/k-means-clustering-algorithm-4-638.jpg?cb=1365646184

Dissimilarity measure (1)

Concept of dissimilarity

Dissimilarity is a function of the pair $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ with a value in $\mathbb{R}_+$ such that $D(\boldsymbol{x}_1, \boldsymbol{x}_2) = D(\boldsymbol{x}_2, \boldsymbol{x}_1) \geq 0$ and $D(\boldsymbol{x}_1, \boldsymbol{x}_2) = 0 \Rightarrow \boldsymbol{x}_1 = \boldsymbol{x}_2$

Dissimilarity measure: distance $D(\boldsymbol{x}_1, \boldsymbol{x}_2)$ between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2 \in \mathbb{R}^d$

- Minkoswki's distance : $D(\boldsymbol{x}_1, \boldsymbol{x}_2)^q = \sum_{j=1}^{d} |x_{1,j} - x_{2,j}|^q$
  - Euclidean distance corresponds to $q = 2$:
    $D(\boldsymbol{x}_1, \boldsymbol{x}_2)^2 = \sum_{j=1}^{d} (x_{1,j} - x_{2,j})^2 = (\boldsymbol{x}_1 - \boldsymbol{x}_2)^\top (\boldsymbol{x}_1 - \boldsymbol{x}_2)$
  - Manhattan distance $(q = 1)$: $D(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{j=1}^{d} |x_{1,j} - x_{2,j}|$
- Metric linked to the positive definite matrix W :

$$D^2(\boldsymbol{x}_1, \boldsymbol{x}_2) = (\boldsymbol{x}_1 - \boldsymbol{x}_2)^\top \mathrm{W} (\boldsymbol{x}_1 - \boldsymbol{x}_2)$$

Dissimilarity measure (2)

$\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are discrete

- Compute the contingency matrix $A(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{R}^{d \times d}$
  - $\boldsymbol{x}_1 = \begin{pmatrix} 0 & 1 & 2 & 1 & 2 & 1 \end{pmatrix}^\top$ and $\boldsymbol{x}_2 = \begin{pmatrix} 1 & 0 & 2 & 1 & 0 & 1 \end{pmatrix}^\top$
  - $A(\boldsymbol{x}_1, \boldsymbol{x}_2) = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{pmatrix}$
- Hamming's distance: number of indexes where the 2 samples differ

$$D(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{i=1}^{d} \sum_{j=1, j \neq i}^{d} a_{ij}$$

  - Example: $D(\boldsymbol{x}_1, \boldsymbol{x}_2) = 3$

Dissimilarity between clusters (1)

Distance $D(\mathcal{C}_1, \mathcal{C}_2)$ between 2 clusters $\mathcal{C}_1$ and $\mathcal{C}_2$

- minimum diameter (nearest neighbor) :

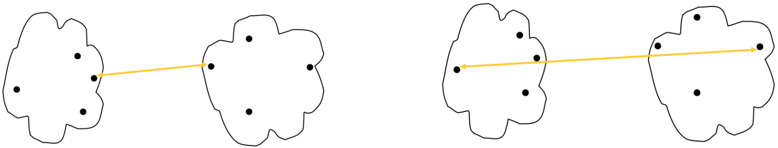
$$D_{\min}(\mathcal{C}_1, \mathcal{C}_2) = \min \left\{ D(\boldsymbol{x}_i, \boldsymbol{x}_j), \boldsymbol{x}_i \in \mathcal{C}_1, \boldsymbol{x}_j \in \mathcal{C}_2 \right\}$$

- maximum diameter :

$$D_{\max}(\mathcal{C}_1, \mathcal{C}_2) = \max \left\{ D(\boldsymbol{x}_i, \boldsymbol{x}_j), \boldsymbol{x}_i \in \mathcal{C}_1, \boldsymbol{x}_j \in \mathcal{C}_2 \right\}$$

Minimum diameter

Maximum diameter



Dissimilarity between clusters (2)

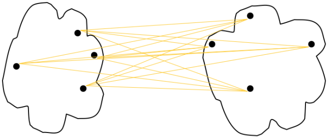
Distance $D(\mathcal{C}_1, \mathcal{C}_2)$ between 2 clusters $\mathcal{C}_1$ and $\mathcal{C}_2$

- average distance :

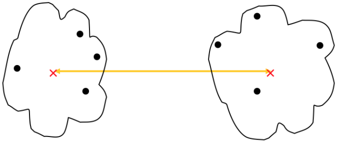
$$D_{\text{moy}}(\mathcal{C}_1, \mathcal{C}_2) = \frac{\sum_{\boldsymbol{x}_i \in \mathcal{C}_1} \sum_{\boldsymbol{x}_j \in \mathcal{C}_2} D(\boldsymbol{x}_i, \boldsymbol{x}_j)}{n_1 n_2}$$

- Ward's distance (between centres) :
$D_{\text{Ward}}(\mathcal{C}_1, \mathcal{C}_2) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} D(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$

Average distance



Distance between centroids



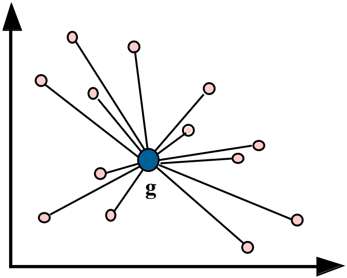
What is a good clustering?

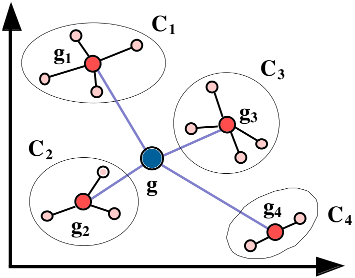
- Every cluster $\mathcal{C}_\ell$ is characterized by:
  - a center: $\boldsymbol{\mu}_\ell = \frac{1}{N_\ell} \sum_{i \in \mathcal{C}_\ell} \boldsymbol{x}_i$ with $N_\ell = \text{card}(\mathcal{C}_\ell)$
  - intra-cluster variation: $J_\ell = \sum_{i \in \mathcal{C}_\ell} D^2(\boldsymbol{x}_i, \boldsymbol{\mu}_\ell)$

    measures how close are the points around $\boldsymbol{\mu}_\ell$. The lower J_ℓ, the smaller is the spread of the samples around $\boldsymbol{\mu}_\ell$
- Within (overall) cluster distance:
  $J_w = \sum_\ell \sum_{i \in \mathcal{C}_\ell} D^2(\boldsymbol{x}_i, \boldsymbol{\mu}_\ell) = \sum_{i \in \mathcal{C}_\ell} J_\ell$
- Let $\boldsymbol{\mu}$ be the centerof the samples: $\boldsymbol{\mu} = \frac{1}{N} \sum_i \boldsymbol{x}_i$
- Inter-cluster distance: $J_b = \sum_\ell N_\ell D^2(\boldsymbol{\mu}_\ell, \boldsymbol{\mu})$

  measures the "distance" between the clusters. The greater the $\boldsymbol{\mu}$, the more the clusters are well separated

Illustration





Total inertia of the points = Inertia Intra-cluster + Inertia Inter-cluster

A good clustering ...

is the one which minimizes the within distance and maximizes the inter-cluster distance

Approaches of clustering

- Hierarchical clustering
- K-means clustering

Hierarchical clustering: principle

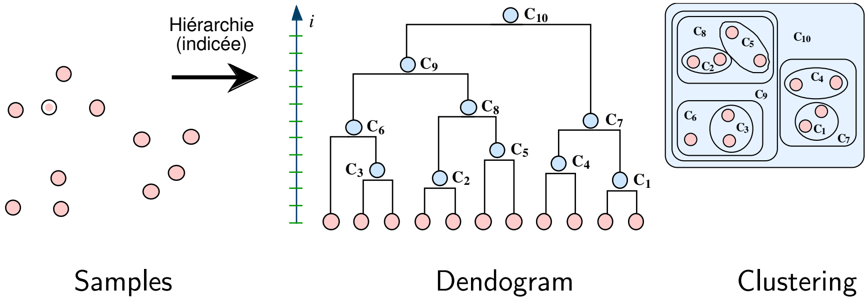
Bottom up approach

The clusters are iteratively "merged" with their nearest clusters.

Algorithm

- Initialization:
  - Each sample is a cluster,
  - Compute the pairwise distance matrix $\boldsymbol{M}$ with $M_{ij} = D(\boldsymbol{x}_i, \boldsymbol{x}_j)$
- Repeat
  - Select from $\boldsymbol{M}$ the two closest clusters $\mathcal{C}_I$ and $\mathcal{C}_J$
  - Merge $\mathcal{C}_I$ and $\mathcal{C}_J$ into the cluster $\mathcal{C}_G$
  - Update $\boldsymbol{M}$ by computing the distance between $\mathcal{C}_G$ and the remaining clusters
- Until all samples are merged into one cluster

Hierarchical clustering: illustration



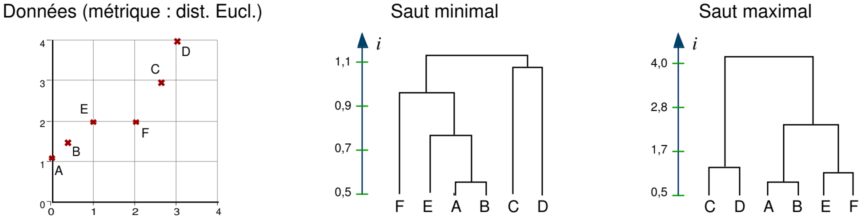
- Dendrogram: represents the successive mergings
- Height of a cluster in the dendrogram = distance between the 2 clusters before their merging

Merging two clusters

Common metrics

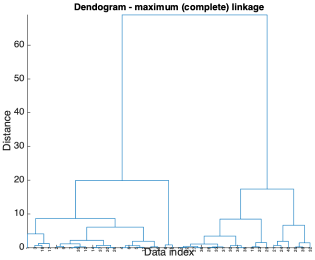
- Single linkage (minimum) based on $D_{\text{min}}(\mathcal{C}_1, \mathcal{C}_2)$
  - produces large clusters (by chaining effect)
  - sensitivity to noised data
- Complete linkage (maximum) based on $D_{\text{max}}(\mathcal{C}_1, \mathcal{C}_2)$
  - produces specific clusters (only very close clusters are combined)
  - sensitivity to noised data
- Average linkage based on $D_{\text{moy}}(\mathcal{C}_1, \mathcal{C}_2)$
  - produces classes with close variance
- Ward distance $D_{\text{Ward}}(\mathcal{C}_1, \mathcal{C}_2)$
  - tends to minimize within variance of clusters being merged

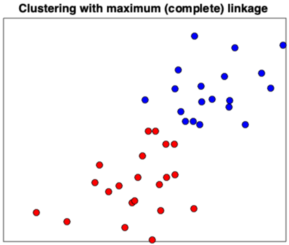
Influence of linkage criterion (1)

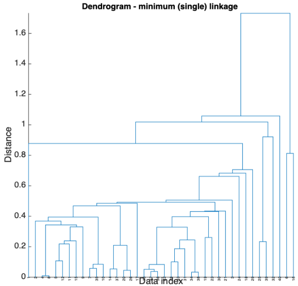


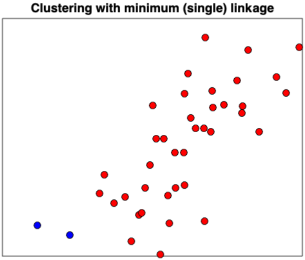
- Clustering result may change w.r.t the selected linkage measure

Influence of linkage criterion (2)









Approaches of clustering

- Hierarchical clustering
- K-means clustering

Clustering by data partitioning

Goal

- $\mathcal{D} = \{\boldsymbol{x}_i \in \mathbb{R}^d\}_{i=1}^N$ a set of training samples
- Search of a partition in K clusters (with $K < N$)

Direct approach

- Build all possible partitions
- Retain the best partition among them

NP-hard problem

The number of possible partitions increases exponentially:
$\#\text{Clusters} = \frac{1}{K!}\sum_{k=1}^{K}(-1)^{K-k} C_k^K k^N$.
For $N = 10$ and $K = 4$, we have 34105 possible partitions !

Data partitioning

Workaround solution

- Determine the K clusters $\{\mathcal{C}_\ell\}_{\ell=1}^K$ and their centers $\{\boldsymbol{\mu}_\ell\}_{\ell=1}^K$ that minimize the cluster within-distance J_w

$$\min_{\{\mathcal{C}_\ell\}_{\ell=1}^K, \{\boldsymbol{\mu}_\ell\}_{\ell=1}^K} \sum_{\ell=1}^{K} \sum_{i \in \mathcal{C}_\ell} \|\boldsymbol{x}_i - \boldsymbol{\mu}_\ell\|^2$$

- Global solution: NP-hard problem
- A local solution (not necessarily the optimal partition) can be attained using a simple algorithm: K-means

K-means clustering

A well-known clustering algorithm

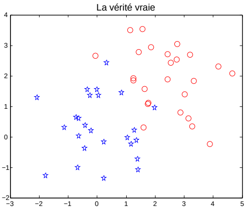
Principle

- Assume the centroids $\boldsymbol{\mu}_\ell, \ell = 1, \cdots, K$ are fixed
  - assign any point $\boldsymbol{x}_i$ to only one cluster
  - $\boldsymbol{x}_i$ is assigned to the closest cluster $\mathcal{C}_k$ (according to the distance between $\boldsymbol{x}_i$ and the clusters' center $\boldsymbol{\mu}_1\ell$)
- Given the clusters $\mathcal{C}_\ell, \ell = 1, \cdots, K$,
  - estimate their centers $\boldsymbol{\mu}_\ell, \ell = 1, \cdots, K$
- Repeat the previous steps until convergence

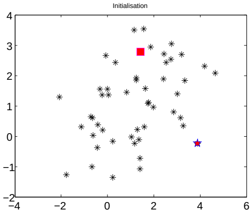
K-Means: illustration

Clustering in $K = 2$ classes

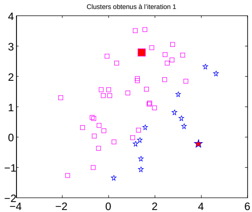
Data



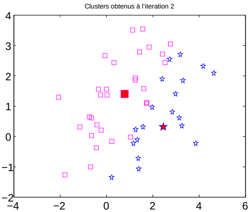
Initialization



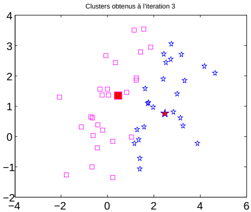
Iteration 1



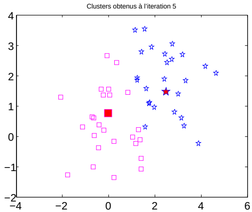
Iteration 2



Iteration 3



Iteration 5



K-Means: Llyod's algorithm

- Initialize the centers $\boldsymbol{\mu}_1, \cdots \boldsymbol{\mu}_K$
- Repeat
  - Assign each point $\boldsymbol{x}_i$ to the closest cluster

$$\forall i \in \{1, \cdots, N\} \quad s_i \leftarrow \arg\min_{\ell} \|\boldsymbol{x}_i - \boldsymbol{\mu}_\ell\|^2 \quad \text{and} \quad \mathcal{C}_k = \{i : s_i = k\}$$

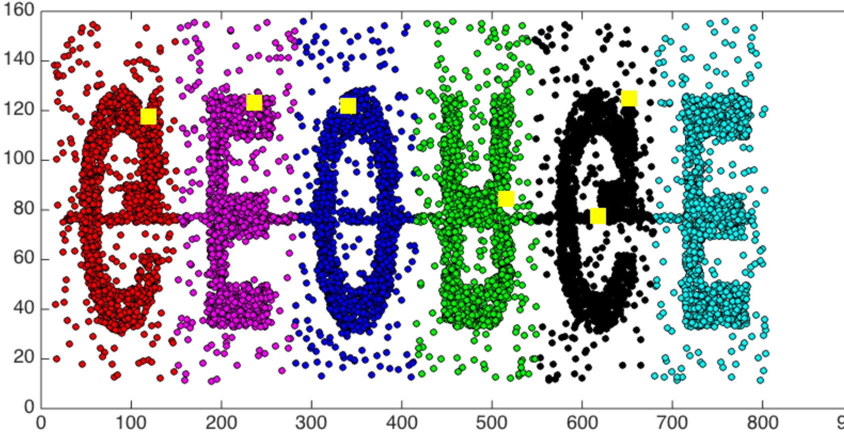
  - Compute the center $\boldsymbol{\mu}_k$ of each cluster

$$\boldsymbol{\mu}_\ell = \frac{1}{N_\ell} \sum_{i \in \mathcal{C}_\ell} \boldsymbol{x}_i \quad \text{with} \quad N_\ell = \text{card}(\mathcal{C}_\ell)$$

- Until convergence

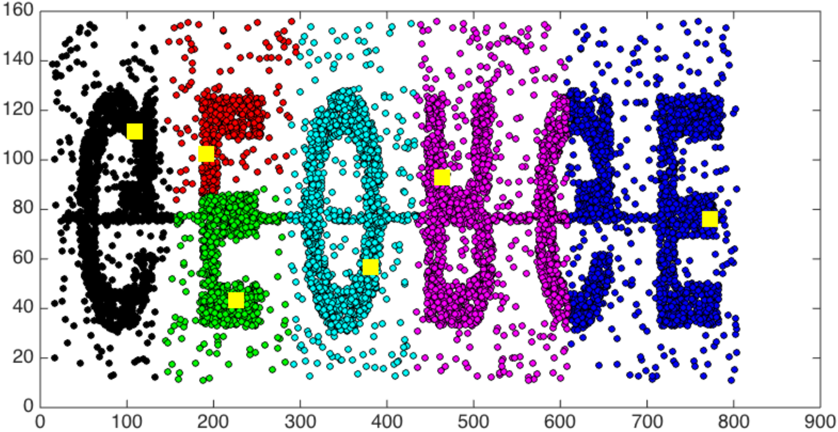
K-Means: example (1)

Initial centers: plain yellow squares



K-Means: example (2)

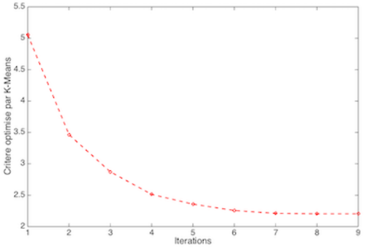
Initial centers: plain yellow squares



$\Longrightarrow$ Different initializations lead to different partitions !

K-Means: remarks and limitations

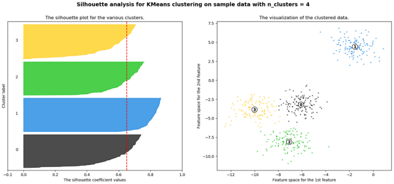
- The criterion J_w decreases at each iteration.

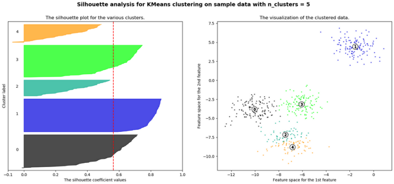


- The algorithm converges to (at least) a local minimum of J_w
- Initialization of $\boldsymbol{\mu}_k$:
  - select randomly within the range of definition of $\boldsymbol{x}_i$
  - select randomly among $\boldsymbol{x}_i$
- Different initializations can lead to different clusters (convergence to local minimum)

K-Means: some issues

- Number of clusters
  - Hard to assess the number of clusters
  - Fixed a priori (e.g.: we want to split customers into K groups)
  - Use the "elbow trick" on the variation of $J_w(K)$ w.r.t K
  - Use ad-hoc metrics such as silhouette score





Conclusion

- Clustering: unsupervised learning
- Group data into homogeneous clusters
- The number of clusters is application-dependent; can be selected based on ad-hoc metrics such as silhouette score
- Several algorithm: hierarchical clustering, K-means, but also DBScan, Spectral clustering, . . .