Unsupervised clustering

$$X = \begin{bmatrix} x_1^1 & \cdots & x_1^p & \ell_1 \\ \vdots & & \vdots & \vdots \\ x_n^1 & \cdots & x_n^p & \ell_n \end{bmatrix}$$

$x_i \in \mathbb{R}^p$

$\ell_i \in \{1, \dots, g\}$

But: ℓ_i are hidden, g isn't known

Goal:
- estimate g
- estimate $\ell_1, \dots, \ell_n$

A general idea: units belonging to the same cluster (sharing the same label) are more similar than units belonging to different.

To implement this idea $\Longrightarrow$ specify a dissimilarity

$d(x, y)$ dissimilarity between x & $y \in \mathbb{R}^p$

$d: \mathbb{R}^p \times \mathbb{R}^p \longrightarrow [0, +\infty)$

Properties:
1. $d(x, x) = 0 \quad \forall x \in \mathbb{R}^p$
1'. $d(x, y) = 0 \iff x = y \quad \forall x, y \in \mathbb{R}^p$
2. $d(x, y) = d(y, x) \quad \forall x, y \in \mathbb{R}^p$
3. $d(x, y) \le d(x, z) + d(z, y) \quad \forall x, y, z \in \mathbb{R}^p$

If d satisfies 1', 2, 3 — metric

4. $d(x, y) \le \max\{d(x, z), d(z, y)\}$

If d satisfies 1', 2, 4 — ultra metric

$x = \{\alpha, \beta, \gamma, \delta, \varepsilon, \zeta, \theta\}$

$y = \{\alpha, \beta, \gamma, \omega, \nu, \mu\}$

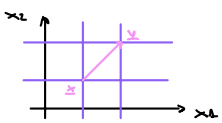
↳ $q = n$ first symbol where the 2 seq x and y are diff

$d(x, y) = \frac{1}{2^n}$ ULTRAMETRIC

SHORT LIST of TRIVIAL distances

$x, y \in \mathbb{R}^p$

- $d(x, y) = \sqrt{(x-y)'(x-y)}$ Euclidean distance
  often x, y have been std.
- $d(x, y) = \sqrt{(x-y)' \Sigma (x-y)}$
  Σ covariance $p \times p$ estimated by S pooled
- $d^r(x, y) = \sum_{i=1}^{p} |x_i - y_i|^2$ Minkowsky's dist (ℓ^r)
  $r = 2 \longrightarrow$ Euclidean dist
  $r = 1 \longrightarrow$ Manhattan dist



$x, y \in (\mathbb{R}^+)^p$

- $d(x, y) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{x_i + y_i}$ Canberra dist

$x, y \in \{0, 1\}^p \qquad x = (0, 1, 1, 0, 0)$

- $d^2_{exc}(x, y) = \sum_{i=1}^{p} (x_i - y_i)^2 = \#$ discordances

EXAMPLE

Raw data = {Blue, Brown, Green}

$x = (x_1, x_2)$

$(0, 1)$ – Blue

$(1, 0)$ – Brown

$(0, 0)$ – Green

$x, y \in \{0, 1\}^p$ — Contingency Matrix

x \ y	1	0	
1	a	b	
0	c	d	
			p

$d^2_{exbd}(x, y) = c + b$

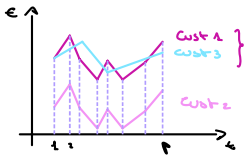
$d(x, y) = 1 - \frac{a}{p}$

$d_Q: \mathbb{R}^p \times \mathbb{R}^p \longrightarrow [0, +\infty)$

$d_C: \{0, 1\}^q \times \{0, 1\}^r \longrightarrow [0, +\infty)$

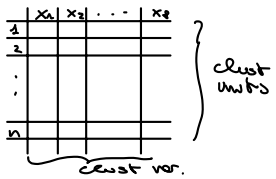
$d = \lambda d_Q + (1 - \lambda) d_C \qquad \lambda \in [0, 1]$

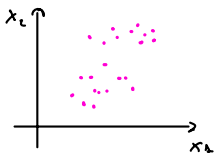
EXAMPLE

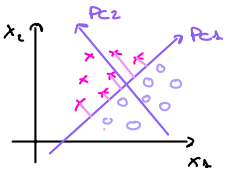


→ Smaller euclidean distance, we might think that they are similar, but if we want to know how they respond to a sale, cust 1 and 2 are similar instead. Depends on what we want to know.

$x = (x_1, \dots, x_p)$ $\quad d(x,y) = \sqrt{2(1-\text{corr}(x,y))}$ ← Clust variables?

$\mathcal{X} =$ 





d is specified $\Longrightarrow$ $D = [d_{ij}]$ $\quad n \times n$ $\quad$ dust matrix

$d_{ij} = d(x_i, x_j)$ $\quad i,j = 1, \dots, n$

$\mathcal{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ $\quad x_i \in \mathbb{R}^p$ $\qquad D = \begin{bmatrix} 0 & & d_{1i} \\ & \diagdown & \\ d_{ij} & & 0 \end{bmatrix}$ $\quad$ if d is reflexive, e symmetric

Distance between finite subsets of $\mathbb{R}^p$

U, V finite subsets of $\mathbb{R}^p$, $\quad ? d(U,V)$

- **single linkage**

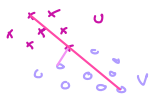
  $d(U,V) = \min \{ d(x_i, y_i) : x_i \in U, y_i \in V \}$

- **complete linkage**

  $d(U,V) = \max \{ d(x_i, y_i) : x_i \in U, y_i \in V \}$

- **average linkage**

  $d(U,V) = \frac{1}{\#U \cdot \#V} \sum_{\substack{x_i \in U \\ y_i \in V}} d(x_i, y_i)$



Hierarchical Agglomeration (Clust. Agglomerativa)

- D distance
- linkage

Initialization:

- every unit is a cluster

Until convergence repeat

Step 1 Cluster together the two closest clusters

Step 2 Update D

Example

$n = 5$

$D =$

	1	2	3	4	5
1	0				
2	2	0			
3	6	5	0		
4	10	9	4	0	
5	9	8	5	3	0

Round 1: Cluster $\{1\}$ & $\{2\}$ in $\{1,2\}$, $\quad d(\{1\}, \{2\}) = 2$

$D_1 =$

	{1,2}	3	4	5
{1,2}	0			
3	5	0		
4	9	4	0	
5	8	5	3	0

Round 2: Cluster $\{4\}$ & $\{5\}$, $\quad d(\{5\}, \{4\}) = 3$

$D_2 =$

	{1,2}	3	{4,5}
{1,2}	0		
3	5	0	
{4,5}	8	4	0

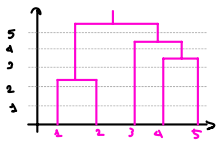
Round 3: Cluster $\{3\}$ & $\{4,5\}$, $d(\{3\},\{4,5\}) = 4$

$D_3 =$

	$\{1,2\}$	$\{3,4,5\}$
$\{1,2\}$	0	
$\{3,4,5\}$	5	0

Round 4: Cluster $\{1,2\}$ & $\{3,4,5\}$ $d(\{1,2\},\{3,4,5\}) = 5$

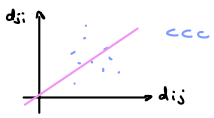
A dendrogram to summarize the entire process:



Obs: jitter the data first ie. introduce some small noise

Cophenetic dist: $d_c(x,y)$ = dist d when x & y merge in the cluster

↳ ultra metric



Ward's method for hierchical clustering

$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ traing set $x_i \in \mathbb{R}^p$

$d: \mathbb{R}^p \times \mathbb{R}^p \to [0,+\infty)$

$d^2(x,y) = (x-y)'(x-y)$

C cluster of points in $\mathbb{R}^p$ (finite)

$\bar{x} = \underset{z \in \mathbb{R}^p}{\arg\min} \sum_{x_i \in C} d^2(x, x_i)$ · $\frac{1}{|C|} \sum_{x_i \in C} x_i$ (barycentre)

for C_i $i = 1, \dots, k$:
- C_i cluster of points in X
- $\bar{x}_i$ barycentre of C_i
- $ESS_i = \sum_{x \in C_i} d^2(x, \bar{x}_i) = \sum_{x \in C_i} (x - \bar{x}_i)'(x - \bar{x}_i)$
- If X has been split in clusters $C_1, \dots, C_k$: $ESS = ESS_1 + ESS_2 + \dots + ESS_k$

Algorithm

1. Initialize: Every point in the cluster $ESS = 0$
2. Until convergence, aggregate the two clusters that produce the minimum increase in ESS

$\Longrightarrow$ dendrogram

Non hierarch. clustering – K-means

Q: Cluster the training set in K (given) subsets $C_1, \dots, C_k$.

1. $\bigcup_{i=1}^{k} C_i = X$
2. $C_i \cap C_j = \emptyset$ $i, j = 1, \dots, k$ $i \neq j$

$d: \mathbb{R}^p \times \mathbb{R}^p \to [0,+\infty)$ metric or pseudo metric

C cluster of points in $\mathbb{R}^p$ (finite)

$\bar{x} = \underset{z \in \mathbb{R}^p}{\arg\min} \sum_{x_i \in C} d^2(x, x_i)$ centroid

Goal: given K find $C_1, \dots, C_k$ satisfying 1 & 2 and s.t. $\sum_{i=1}^{k} \sum_{x \in C_i} d^2(x, \bar{x}_i)$ is minimum OPTIMIZATION PROBLEM

Algorithm (solution proposed by K-means)

1. Initialization step: – Split at random X in K subset (run different initialization because we need to know if it's a local min)
   – or assign at random K-centroid
2. Until convergence repeat: S1. Compute $\bar{x}_1, \dots, \bar{x}_k$ s.t. $\bar{x}_i = \underset{x \in \mathbb{R}^p}{\arg\min} \sum_{x_i \in C_i} d^2(x_i, x)$ $i = 1, \dots, k$

   S2. $\forall x \in X$, assign x to cluster $i = 1, \dots, k$ if $d^2(x, \bar{x}_i) = \min\{d^2(x, \bar{x}_1), \dots, d^2(x, \bar{x}_k)\}$

   STOP when $\bar{x}_1, \dots, \bar{x}_k$ do not change

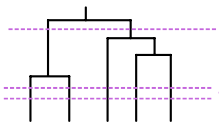
Obs: Step 1 is the difficult one

Simplification: for $i = 1, \dots, k$ $\bar{x}_i = \underset{x \in X}{\arg\min} \sum_{x_i \in C_i} d(x_i, x)$ (K-medoids)

Pros and Cons

- Complete link, aver. link, k-means, Ward's, ...
  ⟹ Create ellipsoidal clusters (wrt d)
- Single link
  ⟹ chain effect, but explores structures that are not ellipsoids

How to choose k?

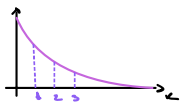


← we are choosing a cut-point on the dendrogram

← Like that we have consistent cut

⟶ DENDROGRAM

→ k-means, Ward, ... : $\sum_{i=1}^{k} \sum_{x \in C_i} d^2(x, \bar{x}_i) = \omega(k)$



Non parametric density based Clustering – DBSCAN

Idea: Clusters are regions of high density

Low density regions are boundaries between clusters, or noise, or outliers.

$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ training set $x_i \in \mathbb{R}^p$

$d: \mathbb{R}^p \times \mathbb{R}^p \to [0, +\infty)$ metric

For $\varepsilon > 0$, $x \in \mathbb{R}^p$ $N_\varepsilon(x) = \{y \in \mathbb{R}^p : d(x, y) < \varepsilon\}$ ε-neighb of x

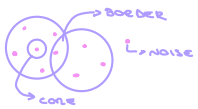
For $i = 1, \dots, n$ $N_\varepsilon(x_i)$ $|N_\varepsilon(x_i)| =$ n° of points in X falling in $N_\varepsilon(x_i)$

Two parameters to be fixed in advance: 1. $\varepsilon > 0$ the radius

2. min Pts ⩾ 1 (an integer)

Point classes in X:

- x_i is a core point if $|N_\varepsilon(x_i)| \geq$ min Pts
- x_i is a BORDER point if it's not a core point but it belongs to an ε-neigh of a core point
- x_i is NOISE if it's not core or border



Definitions

To identify dense contiguous regions:

- x_j is DIRECTLY DENSITY REACHABLE for x_i if:
  a. x_i is core
  b. $x_j \in N_\varepsilon(x_i)$
- x_j is DENSITY REACHABLE from $x_i \in X$ if there is $y_1, \dots, y_k$ $k \geq 1$, s.t.:
  a. $y_1, \dots, y_{k-1}$ are core
  b. $y_1 = x_i$, $y_k = x_j$
  c. $y_j \in N_\varepsilon(y_{j-1})$
- $x_i \in X$ and $x_j \in X$ are DENSITY CONNECTED if there is an $x \in X$ s.t. both x_i and x_j are density reach from x.

DBScan identifies the $C \subseteq X$ s.t.:

1. If x_j is dens reach from x_i and $x_i \in C \implies x_j \in C$
2. For all couples x_i and x_j in C, x_i and x_j must be density connected

Multidimensional scaling (MDS)

$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ training set, $x_i \in \mathbb{R}^p$ $\forall i = 1, \dots, n$

$d: \mathbb{R}^p \times \mathbb{R}^p \to [0, +\infty)$, $d_{ij} = d(x_i, x_j)$ for $i, j = 1, \dots, n$

$D(n \times n)$ dissimilarities matrix

Question: $X = \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix}$ $z_i \in \mathbb{R}^q$ $(q \ll p)$ $i = 1, \dots, n$

$d: \mathbb{R}^q \times \mathbb{R}^q \to [0, +\infty)$, $d(z_i, z_j) = \delta_{ij}$ for $i, j = 1, \dots, n$ s.t. $\delta_{ij} = d_{ij}$

(low-dimensional space with same distance as before)

Solutions not unique, rotation, translation, etc, so we talk about family of configurations.

Classical MDS: find $z_1, \dots, z_n$ s.t. $\sum_{i \neq j} (d_{ij} - \delta_{ij})^2$ (if d is euclidean = PCA)

Kruskal: find $z_1, \dots, z_n$ s.t. $\text{stress} = \frac{\sum_{i \neq j} (\theta(d_{ij}) - \delta_{ij})^2}{\sum \delta_{ij}^2}$ is minimized (also w.r.t. $\theta: \mathbb{R} \to \mathbb{R}$ monotone ↗)