Methods of Internal Analysis
→ unsupervised learning

So far, we have seen methods that relate a response variable to explanatory variables, in order to use this relationship to explain/predict a quantity of interest. → supervised learning

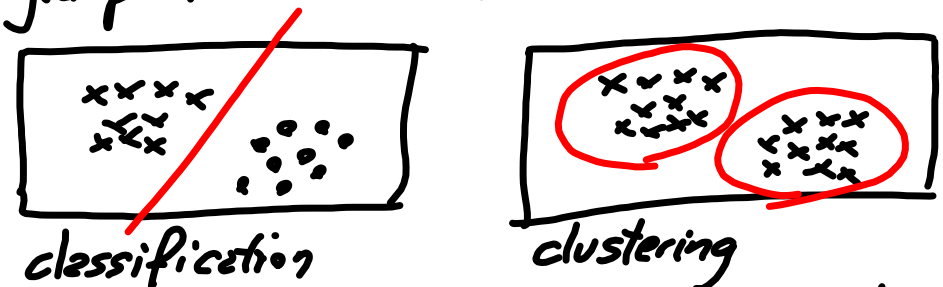
For the rest of the course, we will treat the case of no response variable → all the available variables are on the same level → unsupervised learning

GOAL: understand the relationships among observations, mainly for exploratory analyses

EXAMPLES: Principal Component Analysis (PCA)
clustering ← today

Cluster Analysis

GOAL: group n observations into K clusters



- no information about the characteristics of the groups
- some interpretation can be done a posteriori

EXAMPLES: 1) we want to separate customers in homogeneous groups, e.g. to address them with specific marketing campaigns. (note that we do not know a priori the number of clusters)

2) we want to split the customers in K groups, e.g. if we can perform only K different marketing campaigns (we know the number of cluster, we only need to find them)

In both cases, we want to allocate similar observations to the same group, and dissimilar to different groups

KEY CONCEPT of cluster analysis: define what is "similar" and what is "dissimilar" → dissimilarity distance

Define $d(i,i')$ the dissimilarity between observations i and i'
This dissimilarity $d(i,i')$ is based on the dissimilarities for each of the p observed variables, $d_j(x_{ij}, x_{i'j})$, $j=1,\dots,p$, where $x_i = (x_{i1},\dots,x_{ip})$ is the i-th observation

In order to be a dissimilarity, the function $d_j(x_{ij}, x_{i'j})$ must satisfy the following conditions:

1) $d_j(x,x) = 0$
2) $d_j(x,x') \geq 0$
3) (not necessary but recommended) $d_j(x,x') = d_j(x',x)$ simmetry

If $d_j(x_{ij}, x_{i'j})$ also satisfies the triangle inequality

4) $d_j(x,y) + d_j(y,z) \geq d_j(x,z)$

then $d_j(x_{ij}, x_{i'j})$ is called distance.

For continuous (quantitative) variables, the most used distance is by far the square of Euclidean distance

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

For qualitative variables, often

$$d(x, x') = 1 - \mathbf{1}(x = x') \quad \text{where } \mathbf{1}(x = x') = \begin{cases} 1 \text{ if } x = x' \\ 0 \text{ if } x \neq x' \end{cases}$$

$$\mathbf{1}(x \neq x')$$

For ordinal variables, a conventional score is usually assigned, and then the variables are treated like quantitative
(e.g. c('bad', 'sufficient', 'good') → c(-1, 0, 1))

For both quantitative and qualitative (ordinal and not ordinal) it may be useful to first proceed with a normalization step because the scale of the variables is important to determine the influence of each dimensions

e.g. cm 

It is also relevant for the number of classes

e.g. A-B-C-D-E-F vs pass/not pass

NB: standardization/normalization is not good in every case, there are situations in which it harms, making the results worse

but it is nevertheless recommended.

Once $d_j(x, x')$ is chosen for each dimension, the results must be combined in a single $d(i,i')$. The simplest option is

$$d(i,i') = \sum_{j=1}^{p} d_j(x_{ij}, x_{i'j})$$

There are several other ways, but they must satisfy:

1) $d(i,i) = 0$
2) $d(i,i') \geq 0 \iff d(i,i') = 0 \iff d_j(x_{ij}, x_{i'j}) = 0 \ \forall j$
3) $d(i,i') = d(i',i)$ (ideally)

When <u>all</u> variables (dimensions) are <u>quantitative</u>, we can use

- (weighted) Euclidean distance $\left(\sum_{j=1}^{p} w_j (x_{ij} - x_{i'j})^2\right)^{1/2}$

  where $w_j = \begin{cases} 1 \\ 1/s_j^2 \\ 1/\text{range}^2 \end{cases}$

- Mahalanobis distance $(x_{ij} - x_{i'j})^T \Sigma^{-1} (x_{ij} - x_{i'j})$ with Σ semi definitive positive
- Minkowski distance $\left(\sum_{j=1}^{p} w_j |x_{ij} - x_{i'j}|^\lambda\right)^{1/\lambda}$ with $\lambda \geq 1$
- Manhattan distance $\sum_{j=1}^{p} w_j |x_{ij} - x_{i'j}|$
- Canberra distance $\sum_{j=1}^{p} \frac{|x_{ij} - x_{i'j}|}{|x_{ij}| + |x_{i'j}|}$
- L_∞ norm $\max_j |x_{ij} - x_{i'j}|$

When (as often in real data problem) there is a mixture of quantitative, qualitative and ordinal variables, it is reasonable to first aggregate the variables of the same kind together:

$d^{(1)}(i,i')$ using only quantitative variables
$d^{(2)}(i,i')$ " " qualitative "
$d^{(3)}(i,i')$ " " ordinal "

and then combine them in a further step

$$d(i,i') = \frac{w_1 d^{(1)}(i,i') + w_2 d^{(2)}(i,i') + w_3 d^{(3)}(i,i')}{w_1 + w_2 + w_3}$$

for a suitable choice of the weights.

Once we have the $d(i,i')$ for each pairs (i,i'), we can organise them in a $n \times n$ dissimilarity matrix D, with 0 on the diagonal and only non-negative entries

$$D = \begin{pmatrix} d(1,1)=0 & d(1,2) & \dots & d(1,n) \\ d(2,1) & 0 & & d(2,n) \\ \vdots & & \ddots & \vdots \\ d(n,1) & d(n,2) & \dots & 0 \end{pmatrix}$$

Usually it is symmetric, otherwise we can always make it symmetric by

$$D = (D + D^T)/2$$

D is the basis for most of the clustering algorithms (which requires symmetric D)

↓

use D to find a way to group similar observations into the same cluster, and dissimilar into different clusters

clustering

- non-hierarchical
- hierarchical
  - agglomerative
  - divisive

We said that the goal of clustering is to separate observations in K groups. Once this is done, we can decompose the total dissimilarity (sum of all elements of D)

$$\sum_{i=1}^{n}\sum_{i'=1}^{n} d(i,i') = \sum_{k=1}^{K}\sum_{G(i)=k}\left(\sum_{G(i')=k} d(i,i') + \sum_{G(i')\neq k} d(i,i')\right)$$

$$= \underbrace{\sum_{k=1}^{K}\sum_{G(i)=k}\sum_{G(i')=k} d(i,i')}_{D_{within}} + \underbrace{\sum_{k=1}^{K}\sum_{G(i)=k}\sum_{G(i')\neq k} d(i,i')}_{D_{between}}$$

where D_{within} is the dissimilarity within the groups and $D_{between}$ the dissimilarity between the groups

The goal is to minimize the dissimilarity within the groups, that is actually the same as maximizing the dissimilarity between the groups (total dissimilarity is fixed w.r.t. to the clustering)

We want to allocate n observations in K clusters. Since there is a finite number of ways to do that, in theory we could try all the possibilities and select that that provides the smallest Dunn's.
Obviously this is not possible due to the large number of observations. The number of possibilities is indeed

$$S(n,K) = \frac{1}{K!}\sum_{k=1}^{K}(-1)^{K-k}\binom{K}{k}k^n$$

E.g. $S(10,4) = 34'105$ feasible

$S(19,4) = 10^{10}$ clearly unfeasible

$\Rightarrow$ we need more clever algorithms $\rightarrow$ the most famous being K-means