

Cluster analysis

- no response, working on x
- 1) We want to divide customers based on their characteristics (segment of the market) to give them suitable offers
- 2) The company can afford k ^{different} marketing campaigns, we want to split the customers in k groups

In both examples, we want groups with observations as similar as possible within the groups, ~~with~~ groups as different as possible

KEY CONCEPTS in cluster analysis: similarity between observations

The dissimilarity $d(i, i')$ between two observations i and i' is based on the dissimilarities on each of their variables $d_j(x_{ij}, x_{i'j})$, $j=1, \dots, p$

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

$$1) d_j(x, x) = 0$$

$$2) d_j(x, x') \geq 0$$

$$3) (\text{not always, desired}) \quad d_j(x, x') = d_j(x', x) \quad \text{symmetry}$$

In order to be also a distance, $d_j(x_{ij}, x_{i'j})$ must also satisfy the triangular inequality,

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

- For continuous variables j , by far the most common distance is the Euclidean distance,

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

- For categorical variables,

$$d(x, x') = 1 (x \neq x') = \begin{cases} 1 & \text{if } x \neq x' \\ 0 & \text{if } x = x' \end{cases}$$

- For ordinal variables, we often assign a score to each level, and then treat it as a continuous variable

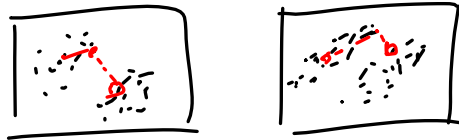
$$\text{E.g. } x_i = \{\text{low, medium, high}\} \rightarrow \{-1, 0, 1\}$$

$$\frac{i - 1/2}{M}, \quad i = 1, \dots, M$$

It is important to normalize the different variables, in order to have dissimilarities in the different directions (variables) that are comparable:

- same scale (eg., only meters, not km vs m)
- ordinal/categorical (take into account how fine is the grid)

Normalization/standardization in this case is tricky, because it is not always good. In the context of cluster analysis, it is not always positive,



Once we find a $d_j(x, x')$ for each j , we need to aggregate them to obtain the final dissimilarity between i and i' . The simplest solution is to add them

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

This assures:

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

Every way of combine $d_j(x_{ij}, x_{i'j})$ that satisfy 1), 2), and 3) can work

- Boon's suggestion $\sum_{j=1}^p w_j d_j(x_{ij}, x_{i'j})$ for suitable choices of w_j
e.g. $\frac{1}{d_j}$

Other choices for dissimilarities

- (weighted) Euclidean distance $d(i, i') = \left(\sum_{j=1}^p w_j (x_{ij} - x_{i'j})^2 \right)^{1/2}$
- Minkowski distance $= \left(\sum_{j=1}^p w_j (x_{ij} - x_{i'j})^\lambda \right)^{1/\lambda} \quad \lambda \geq 1$
- Manhattan distance $= \sum_{j=1}^p w_j |x_{ij} - x_{i'j}|$
- L_∞ norm $\max_j |x_{ij} - x_{i'j}|$

- if we have different kinds of variables (continuous, categorical, ordinal) it is often better to first aggregate those of the same kind, then all together

$$d(i, i') = \frac{w_1 d_1(i, i') + w_2 d_2(i, i') + w_3 d_3(i, i')}{w_1 + w_2 + w_3}$$

with suitable choice of w_1, w_2 , and w_3

All the dissimilarities $d(i, i')$ $i, i' = 1, \dots, n$ are then organised in a dissimilarity matrix D

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

The matrix D is the basis for most of the clustering techniques, whose job is to put together observations with low $d(i, i')$

Usually D is symmetric, if it's not we use

$$\underline{D} = (D + D^T) / 2$$