Material taken from Hastie et al., 2017, section 14.3.

9 CLUSTER ANALYSIS

9.1 INTRODUCTION

☛ In the classification problem, the goal was to classify observations into groups that we knew in advance: we had training data $(\mathbf{X}_1, G_1), \ldots, (\mathbf{X}_n, G_n)$ from each group (supervised learning, we had known class labels G_i).

☛ In cluster analysis, the goal is also to assign individuals to groups but unlike classification, we don't know what these groups are and we have no training data from the groups (unsupervised learning).

☛ We observe only $\mathbf{X}_1, \ldots, \mathbf{X}_n$, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$ (or directly data on dissimilarities, see later). We don't know if there are natural groups but suspect that the individuals may come from several groups and we hope to identify those groups, called clusters.

☛ Example: a new company has some data (some $\mathbf{X}_i$'s) about its customers (for example data on their purchases) and to understand better their behaviour, the company wants to identify clusters of individuals with different consumption behaviour.

It is a new company and so they really don't know what clusters to expect, they do not have training data.

☛ The idea of clustering techniques is to group individuals into clusters, such that individuals within each cluster are more closely related to one another than individuals from different clusters.

☛ Hierarchical clustering: sometimes we may also arrange the clusters into a natural hierarchy. The individuals are grouped into a few large clusters first, then each is further divided into smaller clusters. This sequential division can be done several times.

☛ Cluster analysis is often used as a descriptive tool to see if the $\mathbf{X}_i$'s are likely to come from several groups or not (each group having different properties).

☛ Within a cluster the individuals are similar to each other. This notion depends on the definition of similarity that we use. Different measures of similarity usually lead to different clusters.

☛ When using a clustering technique we have to choose which similarity measure seems to be appropriate for the data at hand. It is not especially easy to determine: we have to think about the data, the problem, and try to identify what seems to be a relevant similarity measure for our problem.

☛ Many clustering algorithms takes as input a dissimilarity matrix.

☛ This is an $n \times n$ matrix D such that D_{ij}, $i, j = 1, \ldots, n$ is the dissimilarity measure between the ith and jth individuals. D_{ij} is the (i, j)th element D. Depending on the case the data could be explanatory vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from which we compute D, or directly in the form of D.

☛ Most algorithms assume that D has nonnegative elements and zero diagonal elements: $D_{ii} = 0$ for $i = 1, 2, ..., n$ because an individual is not dissimilar to themselves.

Larger dissimilarity value: More different
Smaller dissimilarity value: More similar

☛ Most algorithms assume symmetric dissimilarity matrices, so if the original matrix D is not symmetric it must be replaced by $(D + D^T)/2$.

☛ If we are given similarities rather than dissimilarities, unless the algorithm accepts a similarity matrix, we have to first create a dissimilarity matrix. To do this, we usually apply a monotone-decreasing function to the similarities to turn them into dissimilarities.

☛ When D is computed from explanatory vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$, we could also regard dissimilarity as a function, say

$$\mathcal{D} : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^+,$$

which measures the dissimilarity between two individuals. In particular we could write $D_{ij} = \mathcal{D}(\mathbf{X}_i, \mathbf{X}_j)$. Depending on the case, $\mathcal{D}$ may or may not be a distance.

☛ Recall that $\mathcal{D} : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^+$ is a distance iff

$$\begin{aligned}
&\forall a, b \in \mathbb{R}^p : \quad \mathcal{D}(a,b) = \mathcal{D}(b,a) \\
&\forall a, b \in \mathbb{R}^p : \quad \mathcal{D}(a,b) = 0 \iff a = b \\
&\forall a, b, c \in \mathbb{R}^p : \quad \mathcal{D}(a,c) \leq \mathcal{D}(a,b) + \mathcal{D}(b,c).
\end{aligned}$$

☛ If $\mathcal{D}$ is not real distance then we cannot apply, to the matrix D, clustering algorithms based on a real distance.

☛ In most cases where we want to cluster data, we observe p variables (aka attributes) $X_1, \ldots, X_p$ for each of n individuals. For $i = 1, \ldots, n$, we observe a vector $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$.

☛ Many clustering algorithms take as input a dissimilarity matrix $\Rightarrow$ we use those observations to construct it.

☛ A simple way of doing this is to take the (i, k)th element of the dissimilarity matrix D to be

$$D_{ik} = \mathcal{D}(\mathbf{X}_i, \mathbf{X}_k) = \sum_{j=1}^{p} d(X_{ij}, X_{kj}) \,,$$

where $d(X_{ij}, X_{kj})$ is a measure of dissimilarity between individuals i and k for the variable X_j.

☛ However there are different ways to define dissimilarity, depending on the nature of the data.

☛ Quantitative variables: $X_1, \ldots, X_p$ in the form of continuous real-valued numbers.

☛ Often use

$$D_{ik} = \mathcal{D}(\mathbf{X}_i, \mathbf{X}_k) = \sum_{j=1}^{p} d(X_{ij}, X_{kj}) \,,$$

where, for $x, y \in \mathbb{R}$,

$$d(x, y) = \ell(|x - y|)$$

with ℓ an increasing function and $|\cdot|$ the absolute value. Most often:

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

Can also take

$$d(x, y) = |x - y|.$$

The absolute difference gives the same importance to small and large differences whereas the squared difference make small differences smaller and large differences larger $\Rightarrow$ puts more emphasis on larger differences.

☛ Another possibility is to measure similarity between the ith and the kth individuals through a "correlation"

$$\rho(\mathbf{X}_i, \mathbf{X}_k) = \frac{\sum_{j=1}^{p}(X_{ij} - \bar{\mathbf{X}}_i)(X_{kj} - \bar{X}_k)}{\sqrt{\sum_{j=1}^{p}(X_{ij} - \bar{\mathbf{X}}_i)^2 \sum_{j=1}^{p}(X_{kj} - \bar{X}_k)^2}},$$

where on this occasion

$$\bar{\mathbf{X}}_i = \sum_{j=1}^{p} X_{ij}/p.$$

This is not the usual correlation of a random variable, as the latter would be summed over the individuals, not over the components!

☛Instead, here $\rho(\mathbf{X}_i, \mathbf{X}_k)$ it is some sort of notion of correlation between two individuals rather than between two variables.

☛ From the similarity $\rho(\mathbf{X}_i, \mathbf{X}_k)$ we can define dissimilarity by, e.g.,

$$D_{ik} = \mathcal{D}(\mathbf{X}_i, \mathbf{X}_k) = 1 - \rho(\mathbf{X}_i, \mathbf{X}_k)$$

(this will always be between 0 and 2 and $D_{ii} = 0$).

- Categorical/nominal variables:

  - Variables which have several categories (take several values), but there is no notion of ordering (or preference) between those values.

  - Example: a variable that would take the values black, orange, blue, green.

  - In that case the user has to define a way to measure the degree of difference between any two pairs of values. Since there is no number coming from the variables themselves, we have to come up with such a measure ourselves.

  - There is a literature of techniques especially designed for categorical variables. See literature if interested.

invent numbers from scratch

☛ Ordinal variables:

☛ These can be quantitative or categorical but even if they are categorical, there is an order between them. If they are quantitative, then only the order of the numbers matters.

☛ Examples: academic grades (A, B, C, D, F – fail), degree of preference (can't stand, dislike, OK, like, terrific), rank data (when data are ranked according to preference, they are given rank 1, 2, 3, etc).

☛ Suppose the ordinal data take M distinct values. To compute dissimilarity measures, the M values are usually replaced by

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

where $i = 1, \ldots, M$ correspond to the order of the original M values (order as in 1=preferred, 2=2nd preferred etc).

☛ Then we just work with these recoded variables as if they were quantitative variables.