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.

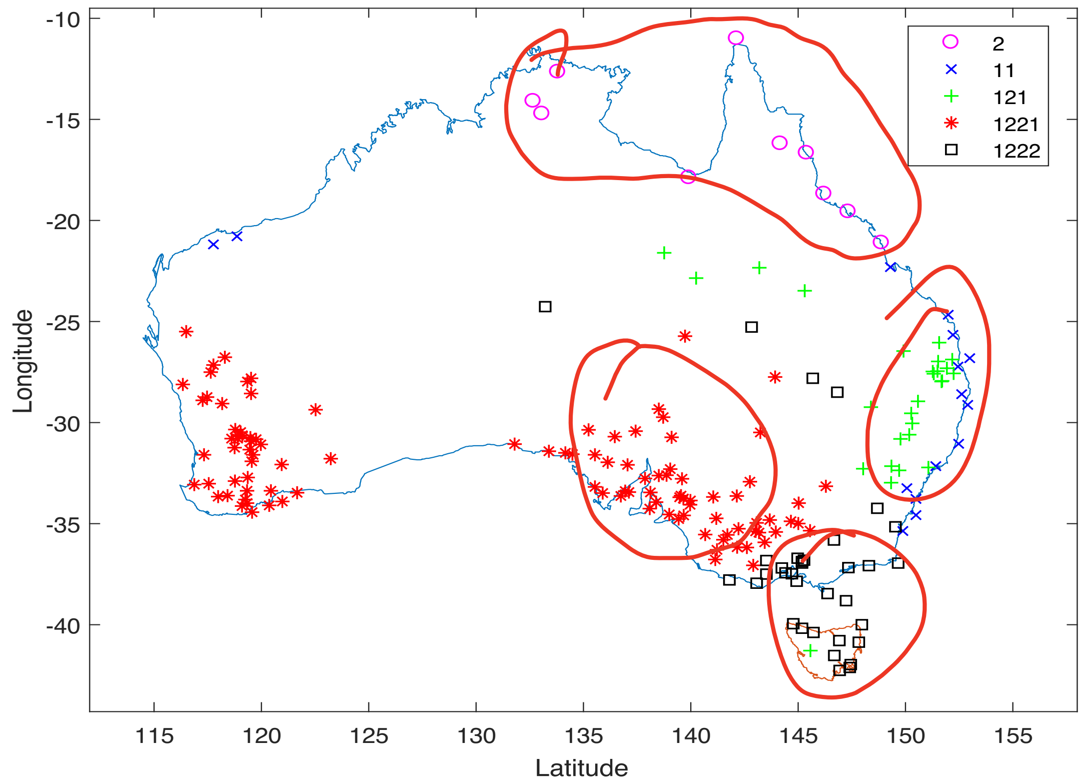
☛ By applying a clustering technique to the data, we hope to identify meaningful groups of individuals.

can I understand some behaviors of my customer from data?

we look at clustering result and identify main groups

Ex: $\mathbf{X}_i$: yearly rainfall measurements at some Australian weather stations. After applying a certain clustering method to the $\mathbf{X}_i$'s, we get the following $K = 5$ clusters:

check rainfall partern is different or not within Australia bounds



The clustering method has roughly clustered the data according to their location in Australia, using only yearly rainfall data $\mathbf{X}_i$.

☛ 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 cluster is further divided into smaller clusters. This sequential division can be done several times. 类似树

Instead of clustering my individuals into K groups: cluster them in a hierachical way

cluster them in 2,3 or four big groups =>

☛ 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), like in the rainfall example.

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

how to define "similarity"

☛ 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 (requires experience). 都试

9.2 DISSIMILARITY MATRICES Dissimilarity matrix → n×n (n个观测值)

- Many clustering algorithms take 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.

对角线是0 →观测和自己的相似度为1，不相似度为0 ∴ $D_{ii}=0$

对角线为0

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

不对称的dissimilarity matrix

→ $D_{ij} = D_{ji}$

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

相似阵转换不相似阵 (用 monotone-decreasing transformation) 不对称要转对称

- 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

function of two vectors in R^p

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

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.

是D(a,b)的含义：

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

$$
\begin{aligned}
&① \ \forall a, b \in \mathbb{R}^p : \quad \mathcal{D}(a, b) = \mathcal{D}(b, a) \\
&② \ \forall a, b \in \mathbb{R}^p : \quad \mathcal{D}(a, b) = 0 \iff a = b \\
&③ \ \forall a, b, c \in \mathbb{R}^p : \quad \mathcal{D}(a, c) \leq \mathcal{D}(a, b) + \mathcal{D}(b, c).
\end{aligned}
$$

triangular. rule.

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

real distance: 满足上述条件①②③

9.3 DISSIMILARITIES BASED ON ATTRIBUTES 不同的度量.

☛ 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 $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$. 一行

☛ 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. 两个度量.

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

数值型变量:

☛ 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 $x, y \in \mathbb{R}$,

绝对值差

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

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

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

或平方差.

Can also take

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

平方项: 让小的更小, 大的更大 => put more emphasis on larger differences

less emphasis on small differences

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"

两个观测间的相关系数 即两个向量之间的

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

where on this occasion

的观测 所有值相加 求均值

average of all components of vector

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

变量个数

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.,

Dissimilarity

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

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

☛ Categorical/nominal variables: 类别

☛ 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.

☛ Ordinal variables: (有顺序的类别)

☛ 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

ordinal变量的转换:
$$\frac{i - 1/2}{M}, i = 1, \ldots, M$$

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.

☛ For dissimilarities based on a notion $d(x, y)$ of dissimilarity between two values x and y, to create our dissimilarity matrix:

☛ the simplest way is to take

对某些变量加重权(着重考虑某些特征的相似/差异)

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

☛ However we can also take a weighted version of this, i.e. take

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

(权重越大，特征影响越大)

是 $w_j \le 1$

where the $w_j's$ are positive weights which depend on the context.

☛ w_j regulates the relative influence of X_j on dissimilarity: we can put more weight on some components if we believe they are more important for clustering but it is often not easy to know in advance which components are important since we often have no idea of the clusters that will be created.

☛ Putting the same weight to each component ($w_j = 1$) doesn't always mean that all variables are given the same importance in clustering.

不假 权重都不意味着你的特征都同样重要，类似PCA（有的X的 Variance很大，他们会导致 dominate entire PCA (composition) 在PCA中

☛ For example if we use $d(x, y) = (x - y)^2$, so that

甚至有时候是因为一些变量的 scale大，比如有的特征 variable 数值非常大（km，千等）

$$D_{ik} = \mathcal{D}(\mathbf{X}_i, \mathbf{X}_k) = \sum_{j=1}^{p} (X_{ij} - X_{kj})^2,$$

then components which have a larger variance contribute more to the dissimilarity measure than others, not because they are more important for clustering but just because their scale is larger than that of other variables and so they artificially have more weight on D_{ik}.

一种 scaling 的方式　scale before clustering.

☛ There to make all variables have equal importance we could take

通过乘以，you put the all the variables into the same varibility. ⟹ $w_j = 1/\hat{s}_{j,j}$, variance of jth component.

with $\hat{s}_{j,j}$ the empirical variance of X_j computed from $\mathbf{X}_1, \dots, \mathbf{X}_n$.

所以即使不加 weight，有方差的变量会 dominate dissimilarity，不是因为他们重要，只是因为他们有更大的 scale

⟹ ~~rescal~~ rescale data before doing clustering

☛ For a general d used to compute D_{ik}, putting a weight

$$w_j = 1/\bar{d}_j,$$

where

$$\bar{d}_j = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{k=1}^{n} d(X_{ij}, X_{kj}),$$

will usually result in giving each component equal influence in the computation of the dissimilarity.

☛ Note: in the case where $d(x, y) = (x - y)^2$, this gives

$$\bar{d}_j = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{k=1}^{n} (X_{ij} - X_{kj})^2 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{k=1}^{n} (X_{ij}^2 - 2X_{ij}X_{kj} + X_{kj}^2)$$
$$= \frac{2}{n} \sum_{i=1}^{n} X_{ij}^2 - 2\bar{X}_j^2 = 2\hat{s}_{j,j}.$$

variance of j th component

$2[E(X^2) - E(X)^2]$

☛ But: maybe putting equal weight to each component is not always a good idea! Maybe some of the components should be assigned more weight because they are more relevant for clustering.

weight

* Scaling : ruin everything *

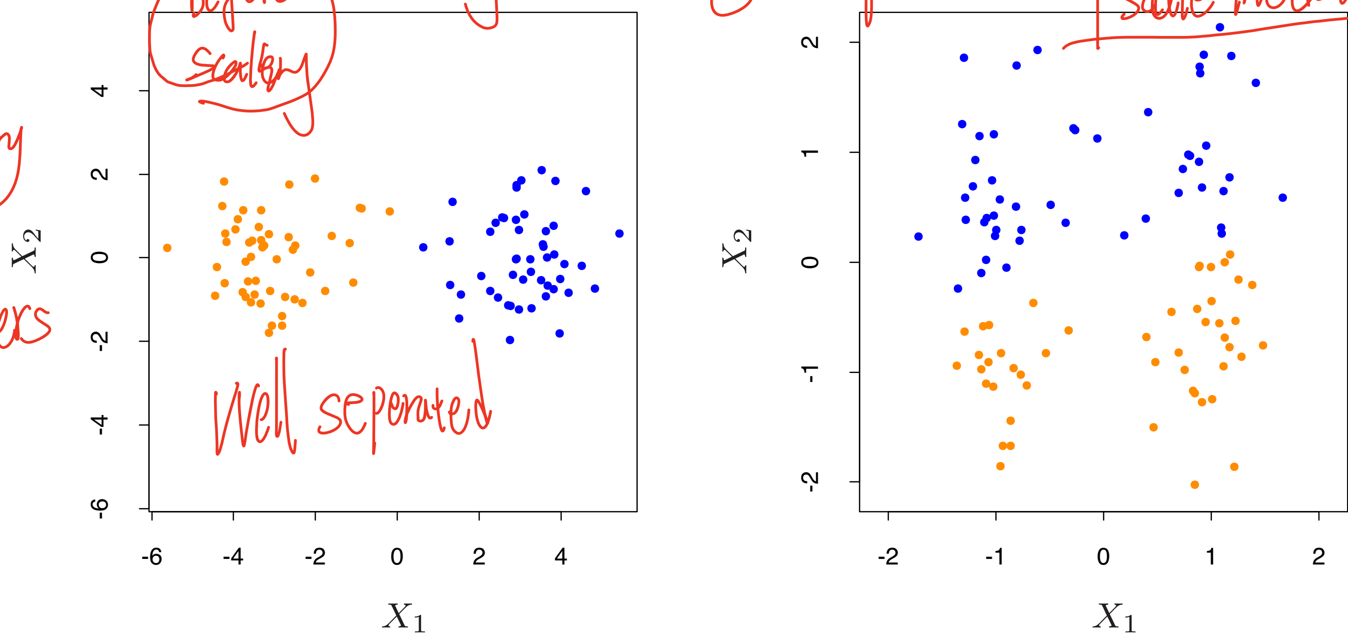
Ex, page 506 of Hastie et al. (2017): clustering some data in 2 groups using the K-means algorithm applied to non standardised (left) and standardised (right – equivalent to putting weight $w_j = 1/(2\hat{s}_{j,j})$) data: in this example, standardising makes the clusters less distinguishable.

K-means cluster that data by comparing the mean of each clusters

before scaling

scaling is not always a good idea.

scale the data



C

FIGURE 14.5. *Simulated data: on the left, K-means clustering (with $K=2$) has been applied to the raw data. The two colors indicate the cluster memberships. On the right, the features were first standardized before clustering. This is equivalent to using feature weights $1/[2 \cdot \mathrm{var}(X_j)]$. The standardization has obscured the two well-separated groups. Note that each plot uses the same units in the horizontal and vertical axes.*

☛ The problem is that there is no recipe for guessing which components are the most important for clustering as we are in an unsupervised problem: we don't really know what we're looking for and we have no training data to guide us. unspervised method

☛ In some problems the user may have some idea about the type of data that seem the most important for clustering in their particular problem.

☛ Each problem is different and users need to think carefully about their own problem to decide how to weigh the components. This is crucial to the success of any clustering algorithm.

☆ there is no clear target on clustering. ☆ there is no sigle trueth (there are different ways of clustering individuals there is a one which) ~~better than~~

There is no wrong and right.