BIG DATA STATISTICS FOR BUSINESS

AY 2023-24

Giovanni De Luca
Parthenope University of Naples

1

Unsupervised learning

- Unsupervised learning: for each statistical unit we have *p* variables, but no response variable is observed.
- We are not interested in prediction, because we do not have a response variable *Y*.
- In this case, the goal is the study of the relationship between the variables or between the observations, or grouping the observations into distinct groups.

2

Unsupervised learning

- We will focus on a particular type of unsupervised learning:

1. clustering, a broad class of methods for discovering unknown subgroups in data.

3

Unsupervised learning

- In general, unsupervised learning is often much more challenging.
- The exercise tends to be more subjective, and there is no simple goal for the analysis, such as prediction of a response.

4

Unsupervised learning

- If we fit a predictive model using a supervised learning technique, then it is possible to check our work by comparing fitted (or predicted) values of the response *Y* for observations used in fitting the model.
- On the other hand, in unsupervised learning, there is no way to check our work because we don't know the true answer: the problem is unsupervised.

5

Cluster analysis

- The cluster analysis is an important method belonging to the unsupervised learning techniques family.

6

Cluster analysis

- The goal of cluster analysis is to ascertain, on the basis of the variables X_1, X_2 . . . , X_p, whether the *n* observations fall into relatively distinct groups.
- For example, in a market segmentation study, we might observe multiple characteristics (variables) for potential customers, such as zip code, family income, and shopping habits.
- We might believe that the customers fall into different groups (e.g.: big spenders versus low spenders).

7

Cluster analysis

- When we cluster the observations of a dataset, we seek to partition them into distinct groups so that the observations within each group are quite similar to each other, while observations in different groups are quite different from each other.
- This is an unsupervised problem because we are trying to discover structure — in this case, distinct clusters — on the basis of a dataset. But we do not know the «true» structure.
- Of course, to make this concrete, we must define what it means for two or more observations to be similar or different.

8

Clustering methods

- There exist a great number of clustering methods.
- We focus on the two most popular clustering methods:

1. hierarchical clustering
2. K-means clustering

9

Hierarchical clustering

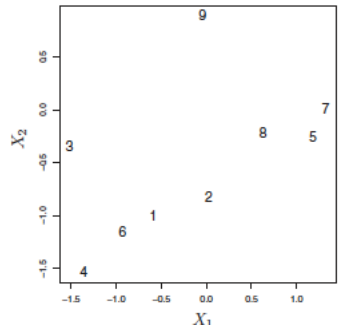
- Hierarchical clustering does not require to pre-specify the number of clusters *K*.
- Hierarchical clustering results in an attractive tree-based representation of the observations, called a dendrogram.

10

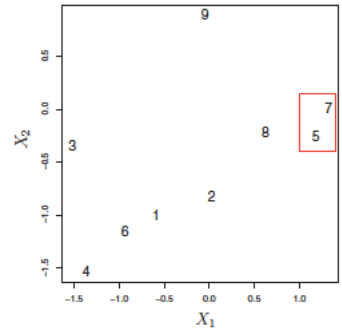
The hierarchical clustering algorithm

- The most common type of hierarchical clustering is bottom-up or agglomerative clustering.
- At the beginning, each of the *n* observations is treated as its own cluster (*n* observations = *n* clusters).
- Use the distance between each pair of observations.
- The two observations that are most similar to each other are then fused so that there now are *n* − 1 clusters.

11



12

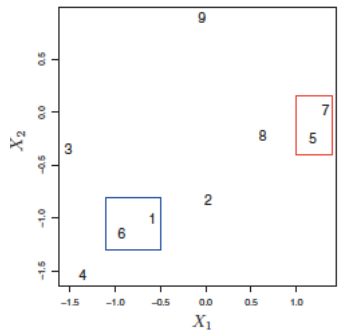


13

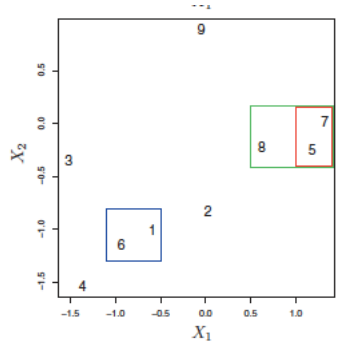
The hierarchical clustering algorithm

- Next the two clusters that are most similar to each other are fused again, so that there now are $n - 2$ clusters.

14



15



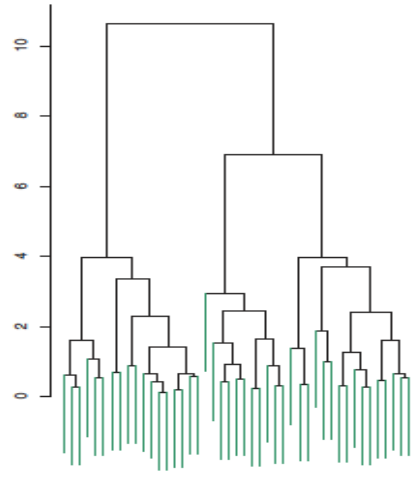
16

The hierarchical clustering algorithm

- The algorithm proceeds in this way until all of the observations belong to one single cluster.
- A dendrogram can now be built.

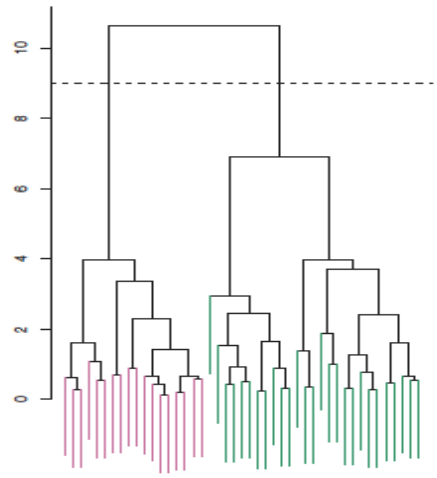
17

Dendrogram



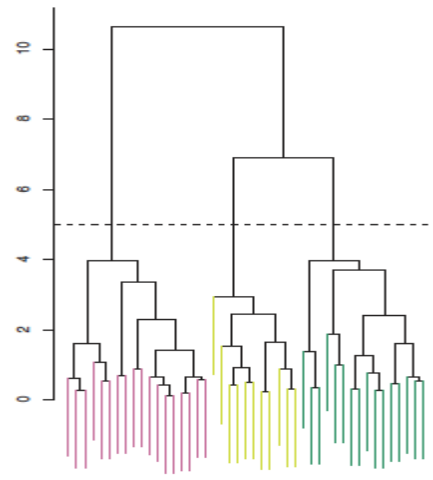
18

Dendrogram



19

Dendrogram



20

How to compute distances?

- Euclidean distance can be computed when the variables are numerical.
- Euclidean distance between unit *i* and unit *j* according to the variable X is given by

$$d_{ij} = \sqrt{(X_i - X_j)^2}$$

21

How to compute distances?

- Euclidean distance between unit *i* and unit *j* according to the (standardized) variables X_1 and X_2 is given by

$$d_{ij} = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2}$$

22

How to compute distances?

- Euclidean distance between unit *i* and unit *j* according to the (standardized) variables X_1, X_2 ,…, X_p is given by

$$d_{ij} = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \cdots + (X_{pi} - X_{pj})^2}$$

$$d_{ij} = \sqrt{\sum_{h=1}^{p} (X_{hi} - X_{hj})^2}$$

23

How to compute distances?

- Gower distance is usually adopted in presence of both numerical and categorical variables. It is given by

$$d_{ij} = \frac{\sum_{h=1}^{p} d_{ij,h}}{\sum_{h=1}^{p} \delta_{ij,h}}$$

- $\delta_{ij,h}$ takes value 1 if the units *i* and *j* can be compared with respect to the *h*-th variable and 0 otherwise (if the phenomenon is absent in both units, e.g. no-no).
- For numerical variables

$$d_{ij,h} = \frac{|X_{hi} - X_{hj}|}{range(X_h)}$$

24

Properties of the distance

- Generic distance d_{ij} between unit *i* and unit *j* has the following properties:

1. Non-negativity: $d_{ij} \geq 0$
2. Identity: $d_{ij} = 0$ if $i = j$
3. Simmetry: $d_{ij} = d_{ji}$
4. Triangular inequality: $d_{ij} \leq d_{is} + d_{sj}$

25

Distance matrix

- Starting from *n* units whse distances have been computed, we can define the distance matrix of order $n \times n$ (*n* rows, *n* columns)

$$D = \begin{bmatrix} 0 & d_{12} & d_{13} & \dots & d_{1n} \\ d_{21} & 0 & d_{23} & \dots & d_{2n} \\ d_{31} & d_{32} & 0 & \dots & \dots \\ \dots & \dots & \dots & 0 & \dots \\ d_{n1} & d_{n2} & \dots & \dots & 0 \end{bmatrix}$$

26

Distance matrix

- Given the property of simmetry, we have

$$D = \begin{bmatrix} 0 & d_{12} & d_{13} & \dots & d_{1n} \\ d_{12} & 0 & d_{23} & \dots & d_{2n} \\ d_{13} & d_{23} & 0 & \dots & \dots \\ \dots & \dots & \dots & 0 & \dots \\ d_{1n} & d_{2n} & \dots & \dots & 0 \end{bmatrix}$$

27

Gower distance: example

UNIT	PROPERTY	COMPANY	INCOME
1	YES	A	50000
2	NO	A	40000
3	YES	B	30000
4	YES	C	10000

Distance between units 1 and 2.

28

Gower distance: example

UNIT	PROPERTY	COMPANY	INCOME
1	YES	A	50000
2	NO	A	40000
3	YES	B	30000
4	YES	C	10000

Distance between units 1 and 3.

29

Problem

- We have an idea of the distance between pairs of observations, but how do we define the distance between two clusters or one observation and a cluster?
- Actually, the idea of distance between a pair of observations needs to be extended to a pair of groups of observations.
- This extension is achieved by developing the notion of linkage, which defines the distance between two groups of observations.

30

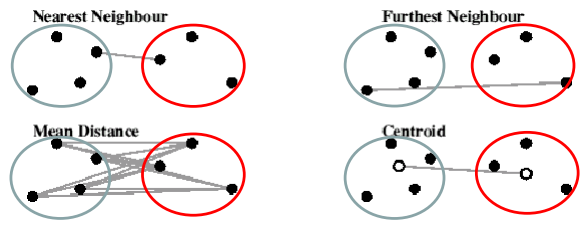
Linkage

- There are four possible linkages:

1. Single linkage (or nearest neighbour)
2. Complete linkage (or furthest neighbour)
3. Average linkage (mean distance)
4. Centroid linkage

31

Linkage



32

Silhouette analysis

- Silhouette analysis can be used to study the separation distance between the resulting clusters.
- For a unit, the silhouette value is a measure of how similar an object is to its own cluster compared to other clusters.
- It is a measure of the quality of a clustering.
- A high average silhouette indicates a good clustering.

33

Silhouette index

- For each unit i the silhouette is a number $s(i)$ between -1 and 1,

$$-1 \leq s(i) \leq 1$$

- $s(i)$ close to 1 means that the observation is appropriately clustered
- $s(i)$ close to -1 means that the observation would be more appropriate if it was clustered in its neighbouring cluster
- $s(i)$ close to 0 means that the observation is on the border of two natural clusters.

34

Average Silhouette Index

- For a dataset we can compute the Average Silhouette Index for a specific number of clusters.
- The average of $s(i)$ over all data of the entire dataset is a measure of how appropriately the data have been clustered.
- Thus silhouette plot may be used to determine the natural number of clusters within a dataset.
- The optimal number of clusters *k* is the one that maximize the average silhouette over a range of possible values for *k*, from 1 to a maximum value.

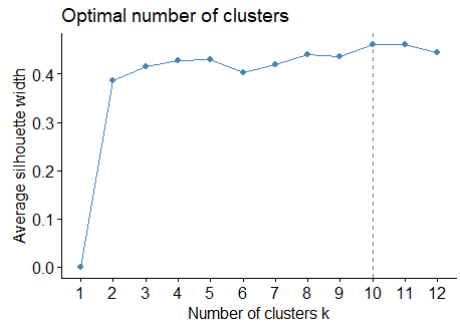
35

Average Silhouette Index

- The algorithm is the following:

1. Perform clustering algorithm for different values of *k* (for instance, by varying *k* from 1 to 10 clusters).
2. For each *k*, calculate the average silhouette of observations.
3. Plot the curve of the average silhouette according to the number of clusters *k*.
4. The location of the maximum is considered as the appropriate number of clusters.

36


Optimal number of clusters
Average silhouette width
0.0
0.1
0.2
0.3
0.4
1 2 3 4 5 6 7 8 9 10 11 12
Number of clusters k

37