Content

- **Unit 3**: Data analysis
  - Predictive and descriptive tasks
  - Supervised techniques
  - Unsupervised techniques
  - Model evaluation

Correlations and factor analysis:

- Make it possible to establish factor relevance (or irrelevance) and whether the correlation is positive or negative wrt. other factors or the variable on study.

Correlation Matrix

	Health	Need	Transp'tion	Child Care	Sick Time	Satisfaction	Ease	No-Show
Health	1							
Need	-0.7378	1						
Transportation	0.3116	-01041	1					
Child Care	0.3116	-01041	1	1				
Sick Time	0.2771	0.0602	0.6228	0.6228	1			
Satisfaction	0.22008	-0.1337	0.6538	0.6538	0.6257	1		
Ease	0.3887	-0.0334	0.6504	0.6504	0.6588	0.8964	1	
No-Show	0.3955	-0.5416	-0.5031	-0.5031	-0.7249	-0.3988	-0.3278	1

Regression coefficient:

Independent Variable	Coefficient
Health	.6434
Need	.0445
Transportation	-.2391
Child Care	-.0599
Sick Time	-.7584
Satisfaction	.3537
Ease	-.0786

Indicates that an increment of 1 in the Health factor increases the probability that the patient does not show in a 64.34%

Association rules and dependencies

- **Non-directional associations:**

Of the following form:

$$(X_1 = a) \leftrightarrow (X_4 = b)$$

From n rows in the table, we compute the cases in which both parts are simultaneously true or false:

We get confidence T_c :

$$T_c = \text{rule certainty} = r_c/n$$

We can (or not) consider the null values.

Unsupervised Learning

- **Directional associations (also called value dependencies):**

  Of the following form (if *Ante* then *Cons*):

  E.g. if (X1= a, X3=c, X5=d) then (X4=b, X2=a)

  From n rows in the table, the antecedent is true in r_a cases and, from these, in r_c cases so is the consequent, then we have:

  Two parameters T_c (confidence/accuracy) y T_s (support):

  Tc = rule confidence =rc/ra : P(Cons | Ante)

  Ts = support = (ra and rc) /n : P(Cons $\wedge$ Ante)

Unsupervised Learning

- Example

ID	Income	City	Profession	Age	Child	Obese	Married
11251545	5.000.000	Barcelona	Ejecutivo	45	3	S	S
30512526	1.000.000	Melilla	Abogado	25	0	S	N
22451616	3.000.000	León	Ejecutivo	35	2	S	S
25152516	2.000.000	Valencia	Camarero	30	0	S	S
23525251	1.500.000	Benidorm	Animador Parque Temático	30	0	N	N

- Non.Directional associations:

  Married and (Child> 0) are associated (40%, 2 cases from 5 instances).
  Obese and Married are associated (60%, 3 cases from 5 instances)

- Directional associations:

  (Child> 0) → Married (100%, 2 cases).
  Married → Obese (100%, 3 cases)

Association Rules.

- The most common algorithm is "A PRIORI" and derivatives.
- There are many variants for association rules:
  - Associations in hierarchies (e.g. product families and categories).
  - Negative associations: "80% of customers who buy frozen pizzas do not buy lentils".
  - Associations for non-binary attributes.

Sequential Association Rules

"if C buys X in T, C will buy Y in T+P"

- Example:

Transaction Database

Customer	Transaction Time	Purchased Items
John	6/21/97 5:30 pm	Beer
John	6/22/97 10:20 pm	Brandy
Frank	6/20/97 10:15 am	Juice, Coke
Frank	6/20/97 11:50 am	Beer
Frank	6/21/97 9:25 am	Wine, Water, CIder
Mitchell	6/21/97 3:20 pm	Beer, Gin, Cider
Mary	6/20/97 2:30 pm	Beer
Mary	6/21/97 6:17 pm	Wine, Cider
Mary	6/22/97 5:05 pm	Brandy
Robin	6/20/97 11:05 pm	Brandy

Sequential Association Rules

- Example (cntd.):

Customer Sequence

Customer	Customer Sequences
John	(Beer) (Brandy)
Frank	(Juice, Coke) (Beer) (Wine, Water, Cider)
Mitchell	(Beer, Gin, Cider)
Mary	(Beer) (Wine, Cider) (Brandy)
Robin	(Brandy)

Sequential Association Rules

- Example (cntd.):

Mining Results

Sequential Patterns with Support >= 40%	Supporting Customers
(Beer) (Brandy)	John, Frank
(Beer) (Wine, Cider)	Frank, Mary

Clustering:

- Deals with finding "natural" groups from a dataset such that the instances in the same group are similar.
- Clustering methods:
  - Hierarchical: the data is grouped in an tree-like way (e.g. the animal realm).
  - Non-hierarchical: the data is grouped in a one-level partition.

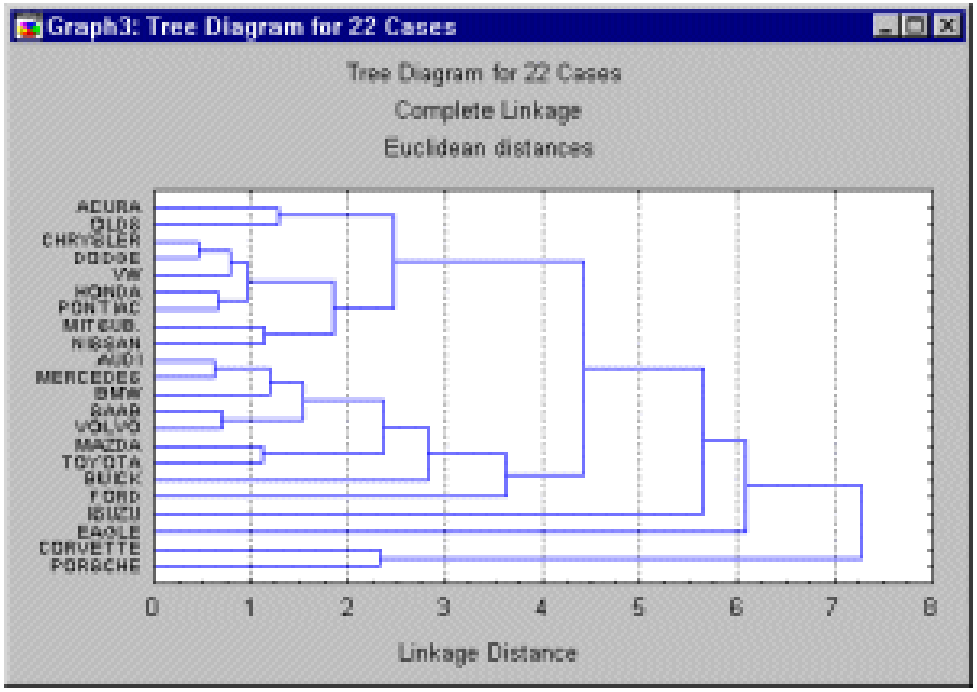
    (a) Parametric: we assume that the conditional densities have some known parametrical form (e.g. Gaussian), and the problem is then reduced to estimating the parameters.

    (b) Non-parametric: we do not assume anything about the way in which the objects are clustered.

Clustering. Hierarchical methods

A simple method consists of separating individuals according to their distance. The limit (linkage distance) is increased in order to make groups.

This gives different clustering at several levels, in a hierarchical way. This is called a Horizontal Hierarchical Tree Plot (or dendrogram)

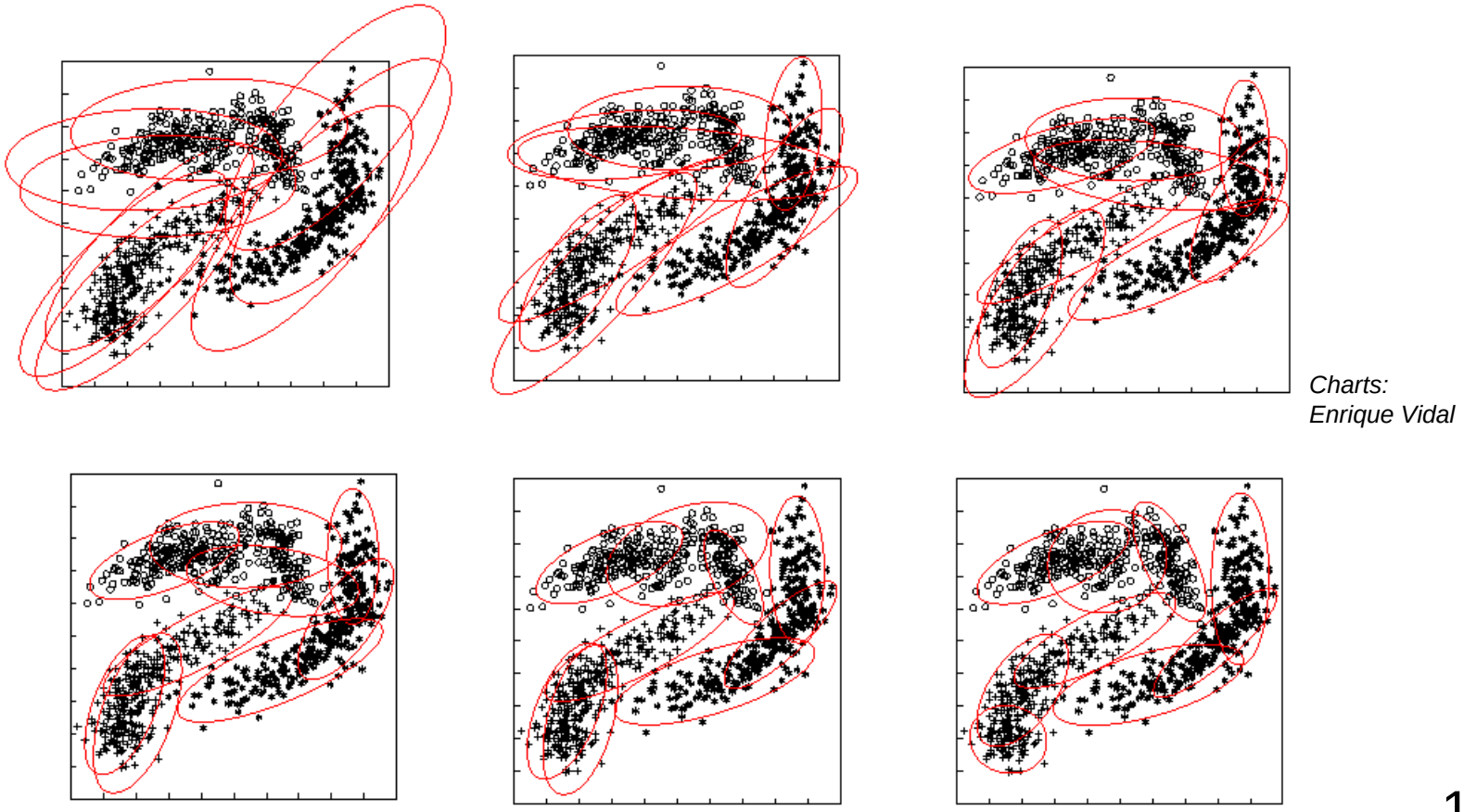


Clustering. Hierarchical methods

- How the tree is constructed
  - Agglomerative: bottom-up approach
  - Divisive: top-down approach
- Linkage criterion: given two clusters A and B, its distance can be calculated as:
  - Complete linkage: $\max \{ d(a,b) : a \in A, b \in B \}$
  - Single linkage: $\min \{ d(a,b) : a \in A, b \in B \}$
  - Unweighted average linkage: $\text{avg} \{ d(a,b) : a \in A, b \in B \}$
  - Centroid linkage $\|c_A - c_B\|$ where c_A and c_B are the centroids of A and B

Clustering. Parametrical Methods:

(e.g., the algorithm EM, Estimated Means) (Dempster et al. 1977).



Charts:
Enrique Vidal

Clustering. Non-Parametrical Methods

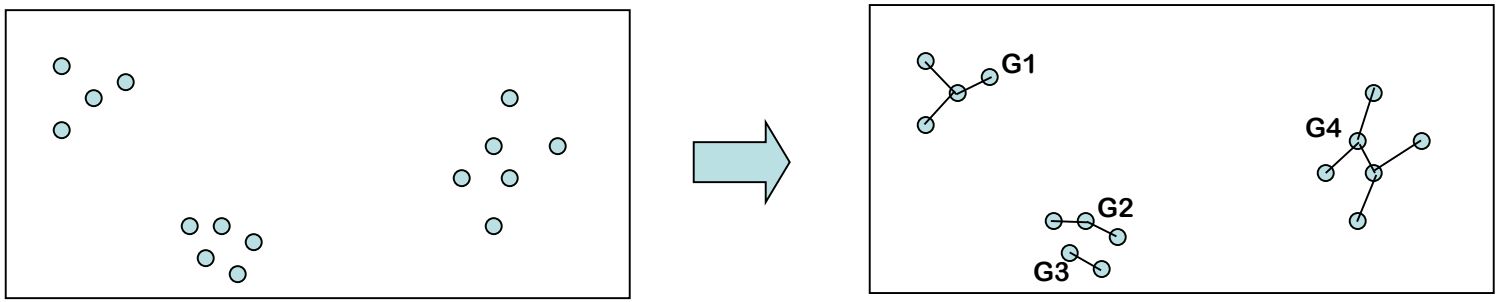
- Methods:
  - *k*-NN
  - *k*-means clustering,
  - online *k*-means clustering,
  - centroids
  - SOM (Self-Organizing Maps) or Kohonen networks.

- Other more specific algorithms:
  - Cobweb (Fisher 1987).
  - AUTOCLASS (Cheeseman & Stutz 1996)

Clustering. Non-Parametric Methods

- **1-NN (Nearest Neighbour):**

Given several examples in the variable space, each point is connected to its nearest point:



(it's like a dendrogram at the lowest level)

The connectivity between points generates the clusters.

In some cases, the clusters are too small

Variants: k-NN.

Clustering. Non-Parametrical Methods

- ***k*-means clustering**:

Is used to find the *k* most dense points in an arbitrarily set of points.

- Algorithm:
  1. randomly split the examples into k (non-empty) sets and calculate the midpoint (or prototype) of each one.
  2. each example is reassigned to the nearest prototype giving new sets
  3. calculate the midpoint of the new groups .
  4. repeat steps 2 and 3 while the prototypes change.

Clustering. Non-Parametrical Methods

k-means clustering: