Descriptive methods
Clustering

***K. Gibert*[(1)]**

[(1)]Department of Statistics and Operation Research

Knowledge Engineering and Machine Learning group
Universitat Politècnica de Catalunya, Barcelona

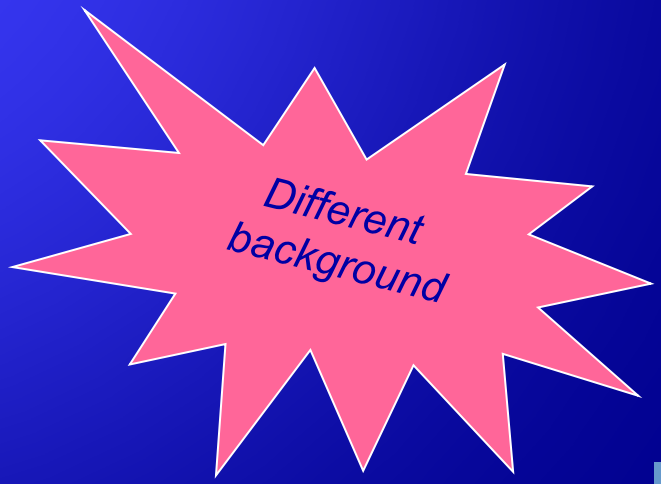
Clustering

Finding homogeneous groups with distinguishable individuals

basic brain activity

First systematic trial: LINNEO (s. XVII)

- Formal solutions
  - Statistics
  - Artificial Intelligence



Clustering
Statistical principles

- Algebraic fundamentals
  - *Only numerical data matrices*
  - Sokal and Sneath 1956 Numerical Taxonomy

Distance required

- Partitioning methods (linear complexity)
  - Number of classes IS AN INPUT
  - K-means, dynamic clouds (nuées dynamiques, Diday)

- Hierarchical methods (quadratic complexity)
  - Number of classes IS AN OUTPUT
  - Ascendents or descendents (for very large n)

Curse of dimensionality

Bad performance if large number of variables (compen[illegible]ct)

A huge "normal " group and many outlier groups (trivial knowledge)

Clustering
Artificial Intelligence principles

- Logic and information theory fundamentals

  Often qualitative data matrices

  - Conceptual clustering (Michalski & Stepp 1983)
    - COBWEB (Fisher 1987)
    - ITERATE (Biswas 1998)
  - Fuzzzy clustering
    - Fuzzy C-Means (Bezdek 1981)

Clustering
Model based approaches

- Probabilistic clustering:
  - *Assume known initial distributions for classes*
  - *EM-algorithm: Two step*
    - ✓ Expectation step: Compute the expetected class of objects (use conditional distributions and posterior probabilities)
    - ✓ Maximization step: Update distributional class parameters to maximize the current class assignments
    - ✓ (use likelihood functiion)
    - ✓ Repeat till no improvement

Distributional assumptions

Convergence not guaranteed

Optimal not guaranteed

- Density Estimation based
  - *Search areas with higher concentration of observations over data cloud*
  - *Assume density homogeneity and some parameters*

Clustering

Hybrid approaches

- Clustering based on rules *[Gibert 1996]*:
  - *Sea sponges [LNStats1994] [Mathware 1997]*
  - *Stellar populations [CyS 1998]*
  - *Thyroid dysfunctions [JAMSDA 1999]*
  - *Characteristic situations in wastewater treatment plants [AIComm2001, 2005]*
  - *Reaction time after electroshock therapy [LNCS2002] [MedicinskaInformatika 2003]*
  - *Response to antidepressants treatment in patients with schizophrenia [ENPP02] [HPP05]*
  - *Functional disability in elderly people [JRR 2004]*
  - *Follow up [MCM 2012]*
  - *Urban planning [NNW05]*
  - *Dependency in severe mental illness [HARPS 2010]*
  - *Comorbidity between severe mental disease and intellectual disability [AIA2007]*
  - *Response to rehabilitation in acquired brain dammage [MedArch2008],*
  - *successfull therapies? (in press)*
  - *Quality of life perceived in patients with spinal cord injury [StudHTI 09][ActaInfMed2009]*
  - *Profile processes in waste water treatment plant [EMS2010]*
  - *Mental Health Systems in under-developed countries (in press)*
  - *Types of Borderline Personality Disorder (in press)*
  - *Characterization of Agitation episodes in severe mental disease (in progress)*