Data mining
Data preparation

Elena Baralis
Politecnico di Torino

Main steps in data preparation

- Data cleaning
  - Incomplete data
  - Noisy data
  - Identification of outliers and exceptions
  - Management of inconsistencies
- Data integration
- Data transformation
  - normalization
  - aggregation
- Data reduction
  - To obtain a reduced representation of the dataset that is smaller in volume, but it can provide similar analytical results
    - discretization
    - sampling

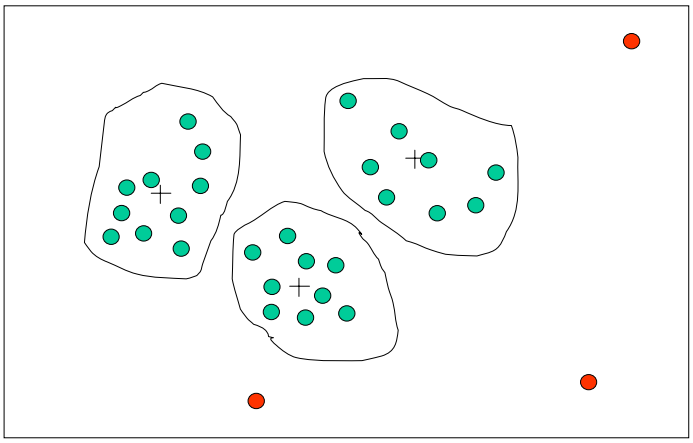
Incomplete data

- Missing values are due to different reasons, typically related to the data entry process
  - Malfunctions of the tools
  - Information was not collected because non relevant
- Solutions
  - Ignore the tuple with missing information
  - Use a special value (NULL, N/A)
  - Use the average value of the attribute
  - Use the average value of the attribute inside the class

Noisy data

- Random error or significant variance of a measure
  - Malfunctions of the tools/network transmission
  - Technological limitations
  - Data entry problems
- Solutions
  - clustering or regression to identify and eliminate outliers
  - discretization

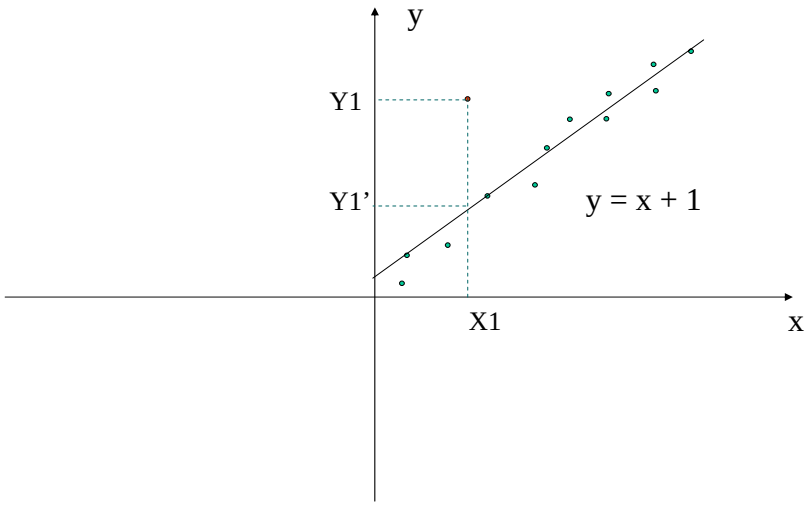
Clustering



From Han, Kamber,"Data mining; Concepts and Techniques", Morgan Kaufmann 2006

Elena Baralis
Politecnico di Torino

Regression



From Han, Kamber,"Data mining; Concepts and Techniques", Morgan Kaufmann 2006

Elena Baralis
Politecnico di Torino

Normalization

- It is a type of data transformation
  - The values of an attribute are scaled so as to fall within a small specified range, typically (-1,+1) or (0,+1)
- Techniques
  - min-max normalization

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - z-score normalization

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

  - decimal scaling

$$v' = \frac{v}{10^j}$$

j is the smallest integer such that $\max(|v'|) < 1$

Data reduction

- It generates a reduced representation of the dataset. This representation is smaller in volume, but it can provide similar analytical results
  - sampling
    - It reduces the cardinality of the set
  - feature selection
    - It reduces the number of attributes
  - discretization
    - It reduces the cardinality of the attribute domain

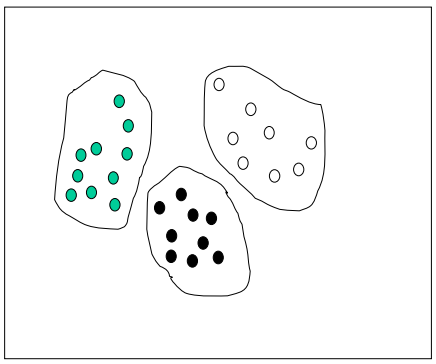
Sampling

- Selection of a subset (sample) of the initial data
  - Representative sample
  - It reduces the complexity of data mining algorithms
  - It doesn't always actually reduce the number of I/O pages
- Techniques
  - Random sampling
    - Without replacement
    - With replacement
  - Stratified sampling
    - For each class of interest, the sample contains a fraction that is proportional to class population in the initial database
    - Suitable for non uniform data distribution

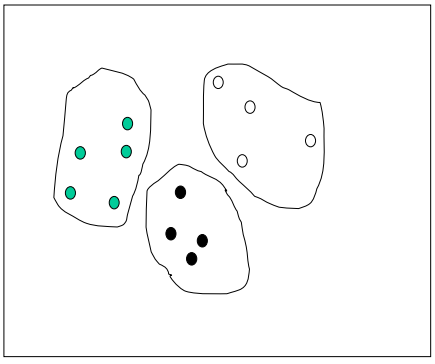
Elena Baralis
Politecnico di Torino

Sampling

Initial database



Stratified sampling



From Han, Kamber,"Data mining; Concepts and Techniques", Morgan Kaufmann 2006

Elena Baralis
Politecnico di Torino

Feature selection

- Selection of a subset of attributes, belonging to the initial schema
  - Selection of a minimal set of attributes preserving the initial data distribution
  - Especially useful for clustering algorithms
- It causes a reduction in the number of extracted patterns
  - It increases the interpretability of the result
- Heuristics techniques
  - The exhaustive exploration is impossible because of the wide number of possible choices
  - One attribute at a time is added/removed, and the result is observed

Elena Baralis
Politecnico di Torino

Discretization

- It splits the domain of a continuous attribute in a set of intervals
  - It reduces the cardinality of the attribute domain
- Techniques
  - N intervals with the same width $W=(v_{max} – v_{min})/N$
    - Easy to implement
    - It can be badly affected by outliers and sparse data
    - Incremental approach
  - N intervals with (approximately) the same cardinality
    - It better fits sparse data and outliers
    - Non incremental approach
  - clustering
    - It well fits sparse data and outliers

Elena Baralis
Politecnico di Torino

Discretization

Price	Width (W=10)	Cardinality (N=2)	Clustering
7	[0,10]	[7,20]	[7,7]
20	[11,20]	[22,50]	[20,22]
22	[21,30]	[51,53]	[50,53]
50	[31,40]		
51	[41,50]		
53	[51,60]		

Elena Baralis
Politecnico di Torino