Introduction to Data Mining

José Hernández-Orallo

Dpto. de Sistemas Informáticos y Computación

Universidad Politécnica de Valencia, Spain

jorallo@dsic.upv.es

Roma, 14-15th May 2009

Outline

- Motivation. BI: Old Needs, New Tools.
- Some DM Examples.
- Data Mining: Definition and Applications
- The KDD Process
- Data Mining Techniques
- Development and Implementation

Taxonomy of DM Techniques

The previous taxonomy is simplified by DM tools:

- **Predictive: (we have one output variable)**
  - ***Classification/categorisation:*** **the output variable is nominal.**
  - ***Regression:*** **the output variable is numerical.**
- **Descriptive: (there is no output variable)**
  - ***Clustering:*** **the goal is to discover groups in the data.**
  - ***Exploratory analysis:***
    - ***Association rules, functional dependencies:*** **the variables are nominal.**
    - ***Factorial/correlation analysis, scatter analysis, multivariate analysis:*** **the variables are numerical.**

Correspondence DM Tasks / Techniques

- *Flexibility: many supervised techniques have been adapted to unsupervised problems (and vice versa).*

TECHNIQUE	PREDICTIVE / SUPERVISED		DESCRIPTIVE / UNSUPERVISED		
	Classification	Regression	Clustering	Association rules	Other (factorial, correl, scatter)
Neural Networks	✓	✓	✓ *		
Decision Trees	✓ (c4.5)	✓ (CART)	✓		
Kohonen			✓		
Linear regression (local, global), exp..		✓			
Logistic Regression	✓				
Kmeans	✓ *		✓		
A Priori (associations)				✓	
factorial analysis, multivariate analysis					✓
CN2	✓				
K-NN	✓		✓		
RBF	✓				
Bayes Classifiers	✓	✓			

Descriptive Methods

Correlation and associations (exploratory analysis, *link analysis*):

- Correlation coefficient (when the attributes are numerical):
  - Example: richness distribution inequality and crime index are positively correlated.
- Associations (when attributes are nominal).
  - Example: tobacco and alcohol are associated.
- Functional dependencies: unidirectional association.
  - Example: the risk level in cardiovascular illnesses depends on tobacco and alcohol (among other things).

Descriptive Methods

Correlations and factorial 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.

Example (Kiel 2000): Visit analysis: 11 patients, 7 factors:

- Health: patient's health (referred to the capability to make a visit). (1-10)
- Need: patient's certainty that the visit is important. (1-10)
- Transportation: transportation availability to the health centre. (1-10)
- Child Care: availability to leave the children on care of another person. (1-10)
- Sick Time: if the patient is working, the ease to get the sick-off time. (1-10)
- Satisfaction: patient satisfaction with their doctor. (1-10)
- Ease: health centre ease to arrange the visit and the efficiency of the visit. (1-10)
- No-Show: indicates if the patient has gone to the doctor's or not during the last year (0-has gone, 1 hasn't)

Descriptive Methods

Correlations and factorial analysis. Example (contd.):

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 do not show in a 64.34%

Descriptive Methods

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 *= rule certainty =* r_c/n

  We can (or not) consider the null values.

Descriptive Methods

Association Rules:

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):**

  T_c = *rule confidence* $= r_c/r_a$: $P(Cons \mid Ante)$

  T_s = *support* $= (r_c$ *or* $r_c/n)$: $P(Cons \wedge Ante)$

Descriptive Methods

Association Rules: Example:

	VINO “EL CABEZÓN”	GASEOSA “CHISPA”	VINO “TÍO PACO”	HORCHATA “XUFER”	BIZCOCHOS “GOLOSO”	GALLETAS “TRIGO”	CHOCOLATE “LA VACA”
T1	1	1	0	0	0	1	0
T2	0	1	1	0	0	0	0
T3	0	0	0	1	1	1	0
T4	1	1	0	1	1	1	1
T5	0	0	0	0	0	1	0
T6	1	0	0	0	0	1	1
T7	0	1	1	1	1	0	0
T8	0	0	0	1	1	1	1
T9	1	1	0	0	1	0	1
T10	0	1	0	0	1	0	0

Descriptive Methods

Association Rules. Example:

- **If we define a minimal support = 2:**
  - **FIRST STAGE: frequent itemsets:**
    - **Seven sets of only one item (seven attributes)**
    - **From the 7!/5!=42 possible cases with two items, we have 15 itemsets with at least the minimal support.**
    - **11 itemsets of three items.**
    - **2 itemsets of four items**
  - **SECOND STAGE: creation of rules from the frequent itemsets:**

IF *bizcochos "Goloso"* AND *horchata "Xufer"* THEN *galletas "Trigo"*	Supp=3, Conf=3/4
IF *bizcochos "Goloso"* AND *galletas "Trigo"* THEN *horchata "Xufer"*	Supp=3, Conf=3/3
IF *galletas "Trigo"* AND *horchata "Xufer"* THEN *bizcochos "Goloso"*	Supp=3, Conf=3/3

Descriptive Methods

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.**

Descriptive Methods

Sequential Association Rules:

- **We can establish associacions such as this:**

  "*if s/he buys X in T s/he 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

Descriptive Methods

Sequential Association Rules:

Example (cont.):

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)

Descriptive Methods

Sequential Association Rules:

Example (cont.):

Mining Results

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

Descriptive Methods

Clustering:

Deals with finding “natural” groups from a dataset such that the instances in the same group have similarities.

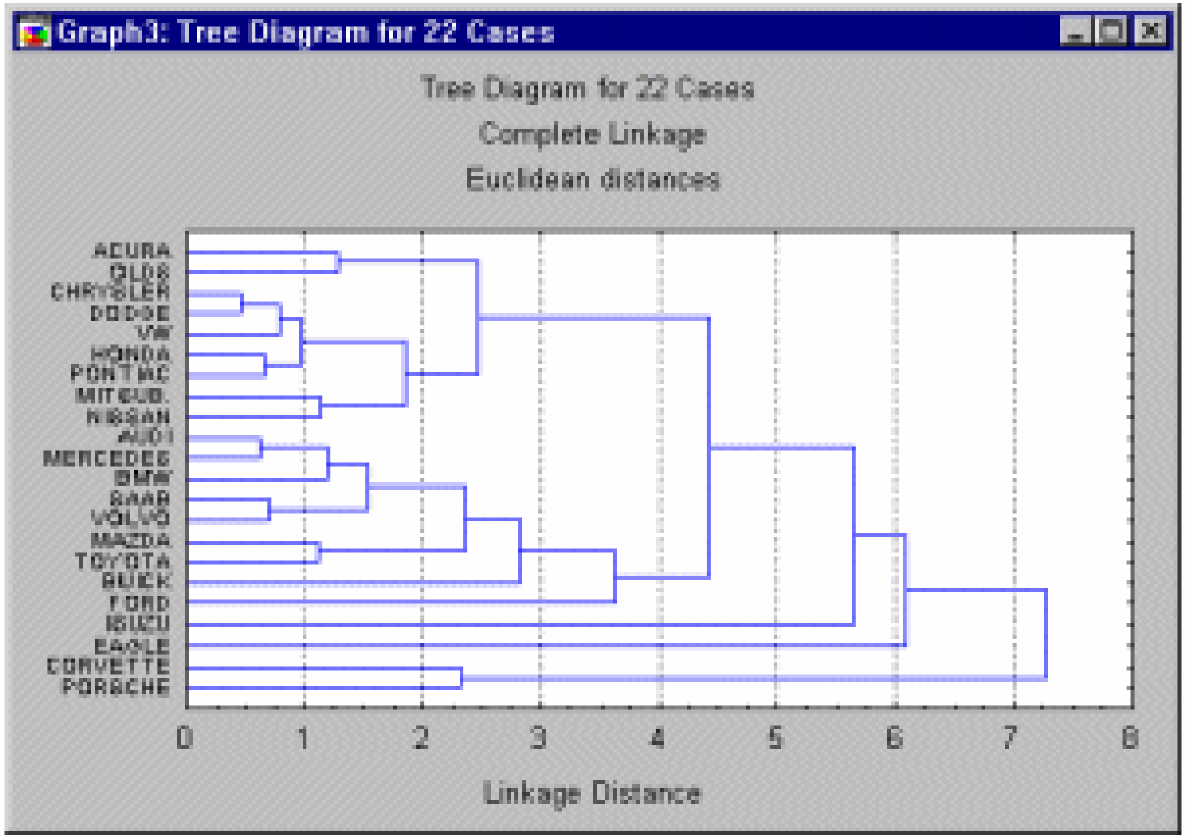
- **Clustering method:**
  - **Hierarchical: the data is grouped in an tree-like way (e.g. the animal realm).**
  - **Non-hierarchical: the data is grouped in a one-level partition.**
    - **(a) Parametrical: we assume that the conditional densities have some known parametrical form (e.g. Gaussian), and the problem is then reduced to estimate the parameters.**
    - **(b) Non-parametrical: do not assume anything about the way in which the objects are grouped.**

Descriptive Methods

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.

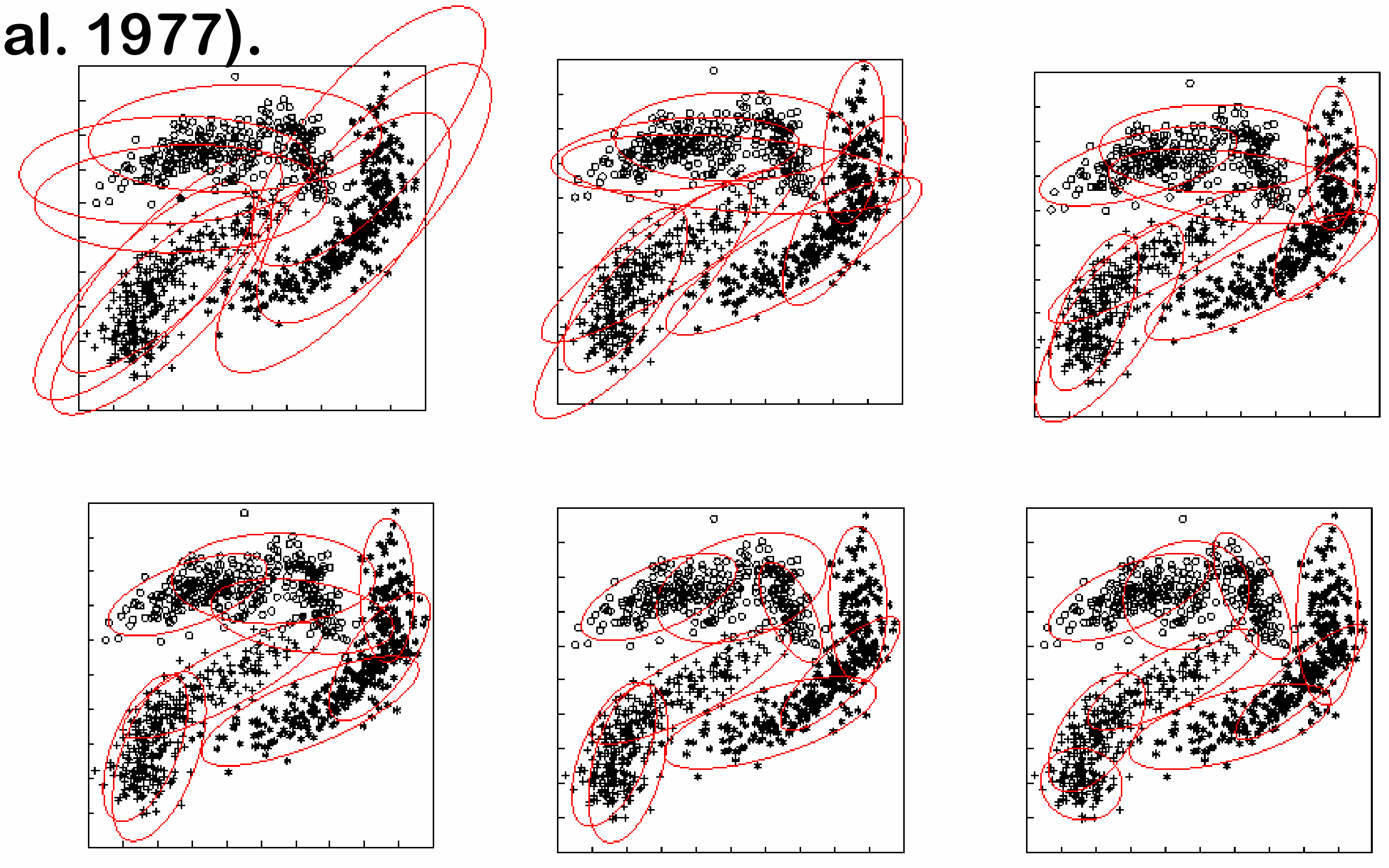
This gives different clustering at several levels, in a hierarchical way.
This is called a *Horizontal Hierarchical Tree Plot* (or dendrogram)



Descriptive Methods

Clustering. Parametrical Methods:

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



Charts: Enrique Vidal

Descriptive Methods

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:

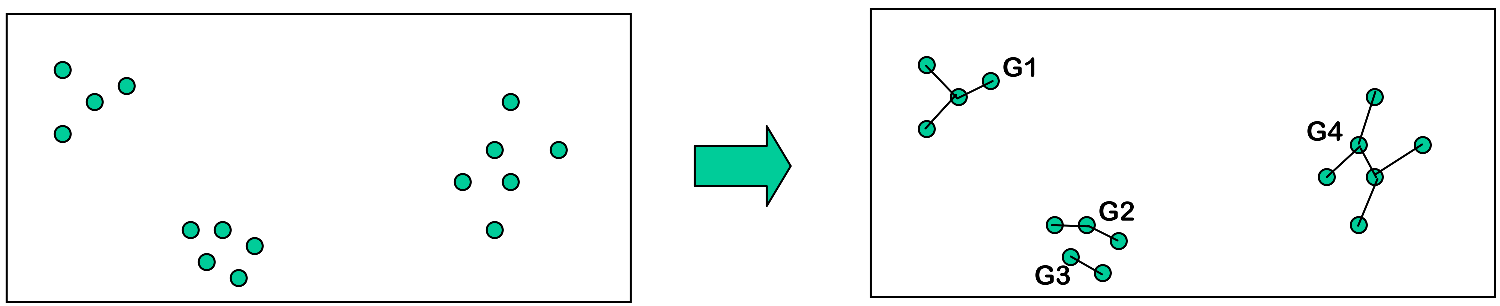
- **Cobweb (Fisher 1987).**
- **AUTOCLASS (Cheeseman & Stutz 1996)**

Descriptive Methods

Clustering. Non-Parametrical Methods

1-NN (Nearest Neighbour):

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



The connectivity between points generates the clusters.

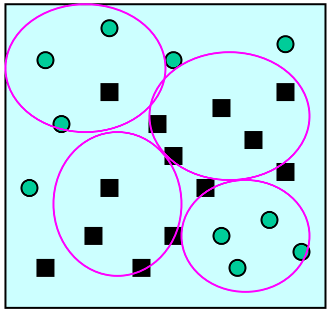
- In some cases, the clusters are too slow.
  - Variants: k-NN.

Descriptive Methods

Clustering. Non-Parametrical Methods

***k*-means clustering:**

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



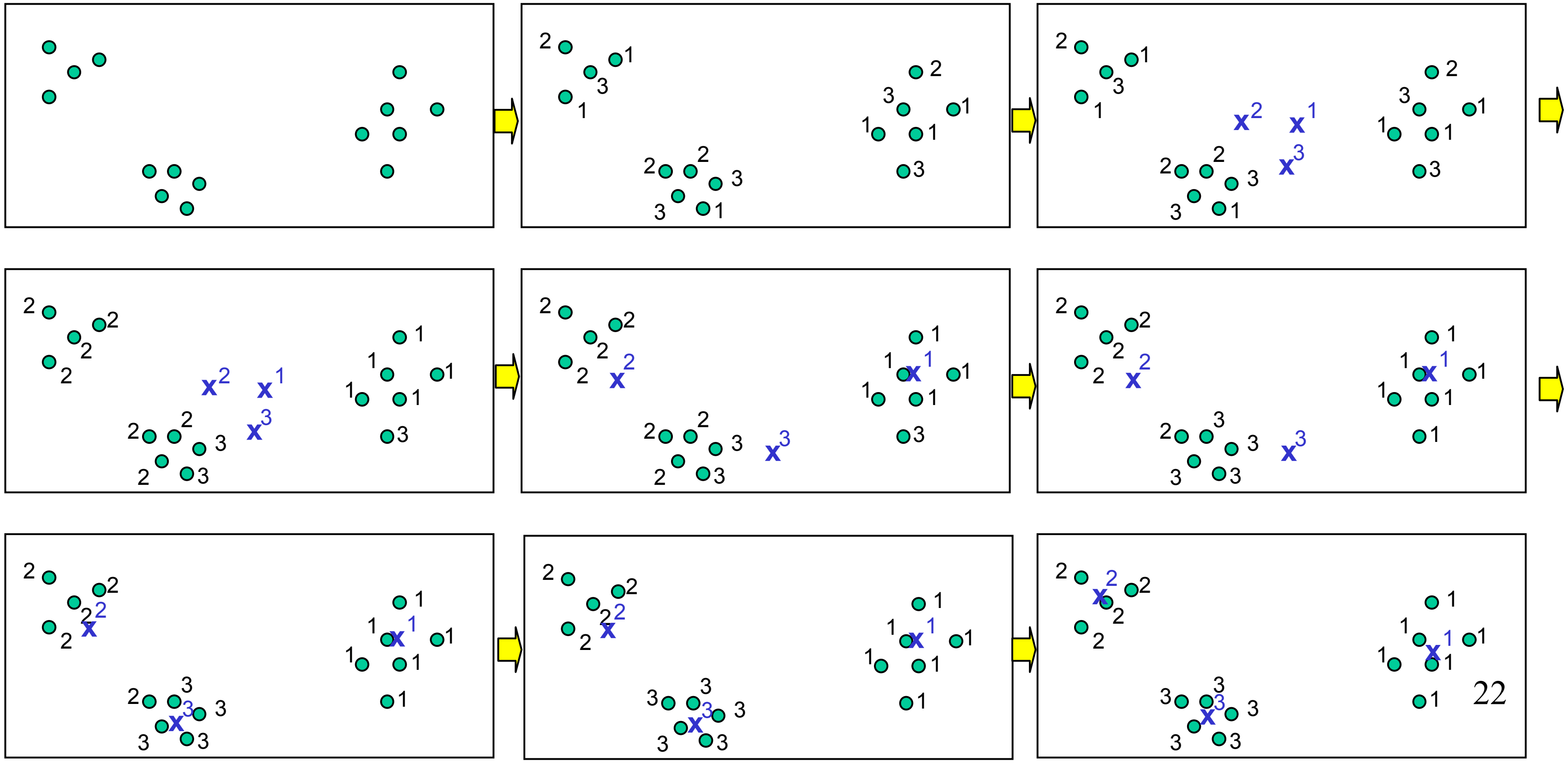
***On-line k*-means clustering (competitive learning):**

- **Incremental refinement.**

Descriptive Methods

Clustering. Non-Parametrical Methods

k-means clustering:

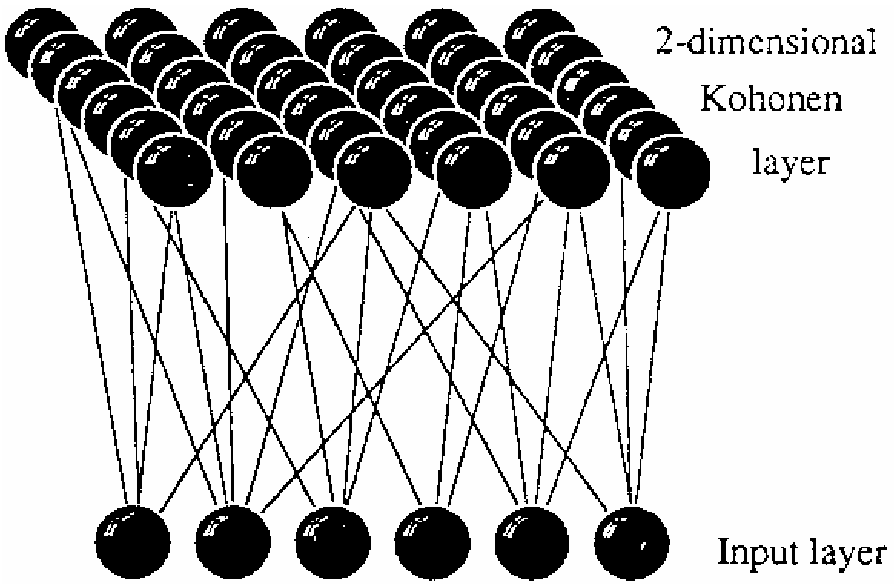


Descriptive Methods

Clustering. Non-Parametrical Methods

SOM (Self-Organizing Maps) or Kohonen Networks

- *Also known as LVQ (linear-vector quantization) or associative memory networks (Kohonen 1984).*

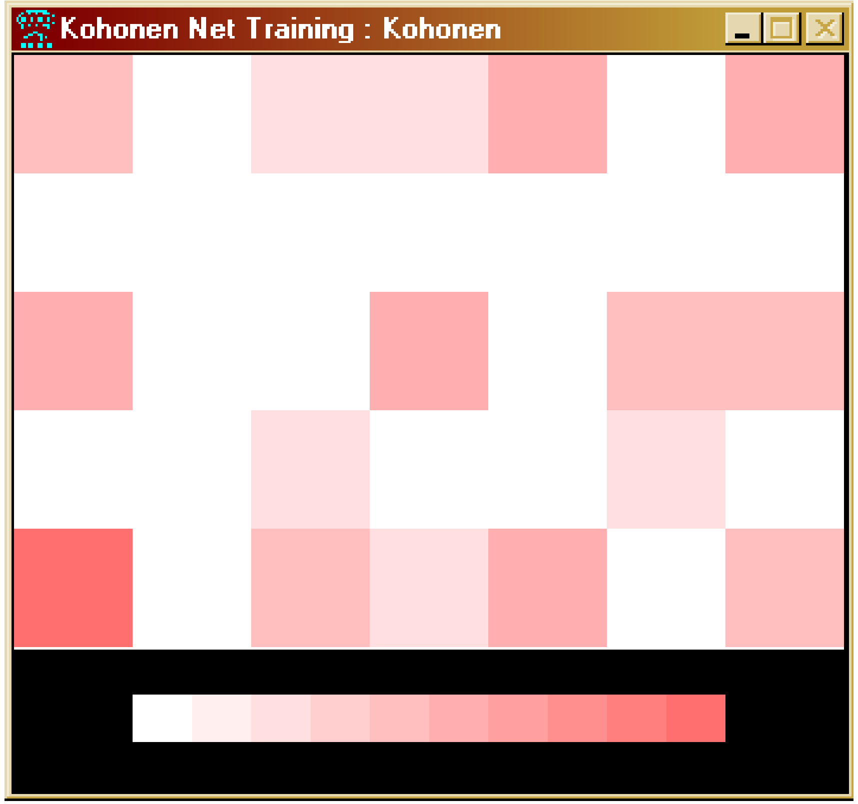


The neuron matrix is the last layer in a bidimensional grid.

Descriptive Methods

Clustering. Non-Parametrical Methods

SOM (Self-Organizing Maps) or Kohonen Networks



It can also be seen as a network which reduces the dimensionality to 2.

Because of this, it is usual to make a bidimensional representation with the result of the network in order to find clusters visually.

Descriptive Methods

Other Descriptive Methods
Statistical Analysis:

- **Data distribution analysis.**
- **Anomalous data detection.**
- **Scatter analysis.**

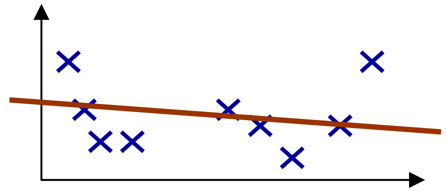
  - ***Frequently, these analyses are used previously to determine the most appropriate method for a supervised (predictive) task.***
  - ***They are also used regularly for data cleansing and preparation.***

Predictive Methods

Global Linear Regression.

- **The coefficients of a linear function *f* are estimated**

For more than two dimensions it can be solved through *gradient descent*



Non-linear Regression.

- **Logarithmic Estimation (the function to obtain is substituted by $y=\ln(f)$). Then, we use linear regression to calculate the coefficients. Next, when we want to predict, we just compute $f = e^y$.**

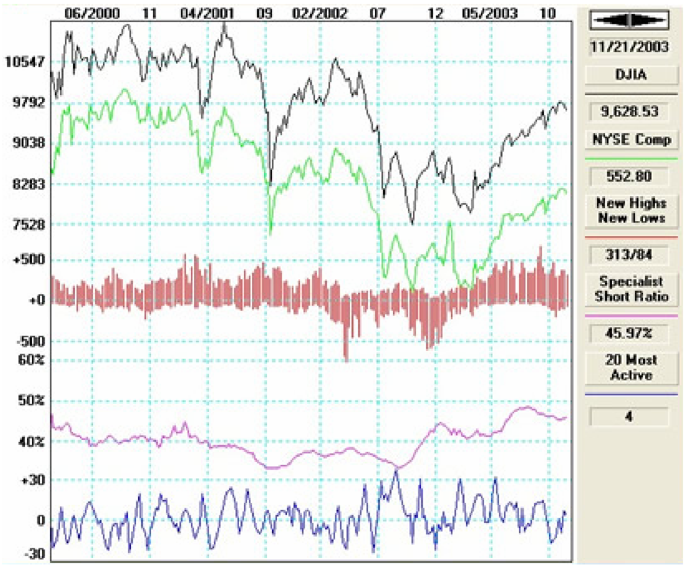
Pick and Mix - Supercharging

- **New dimensions are added, combining the given dimensions. E.g. $x_4 = x_1 \cdot x_2$, $x_5 = x_3^2$, $x_6 = x_1^{x_2}$ and next we get a linear function for x_1, x_2, x_3, x_4, x_5, x_6**

Predictive Methods

General non-linear regression.

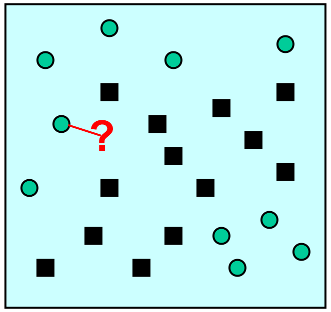
- Adaptative regression and time series. In this case, we usually assume a time order for one of the variables:



- Markov chains.
- Vector Quantization
- MARS (Multiple Adaptive Regression Splines) Algorithm.
- …

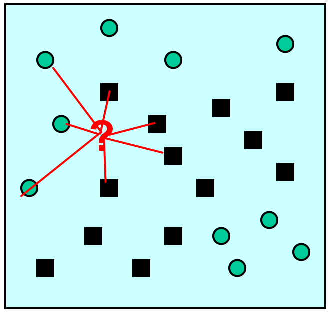
Predictive Methods

k-NN (Nearest Neighbour): can be used for classification



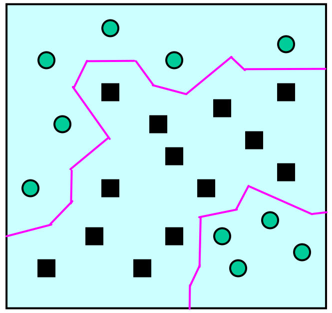
Classifies circle

1-nearest neighbor



Classifies square

7-nearest neighbor

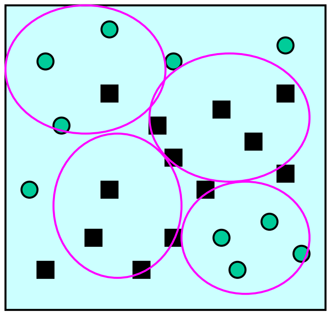


1-nearest neighbor PARTITION

(Poliedric or Voronoi)

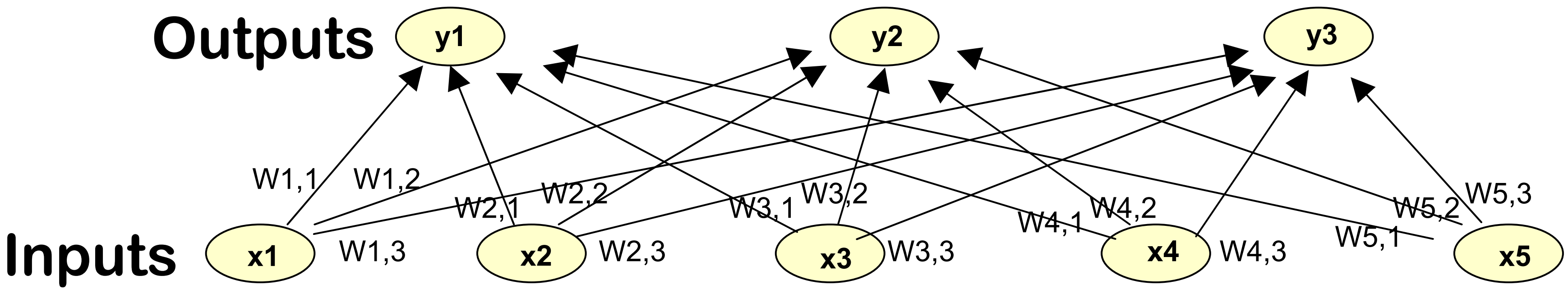
***k*-means clustering:**

- **Can also be adapted to Supervised Learning, if used conveniently.**



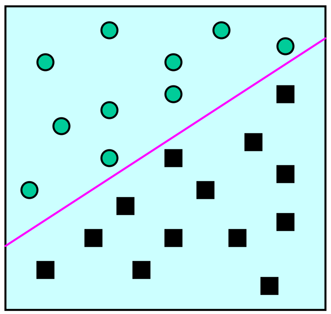
Predictive Methods

Perceptron Learning.

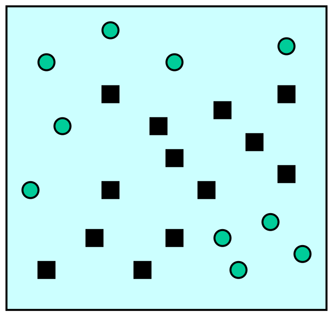


- **Computes a linear function.**

$$y'_j = \sum_{i=1}^{n} w_{i,j} \cdot x_i$$



LINEAR PARTITION POSSIBLE

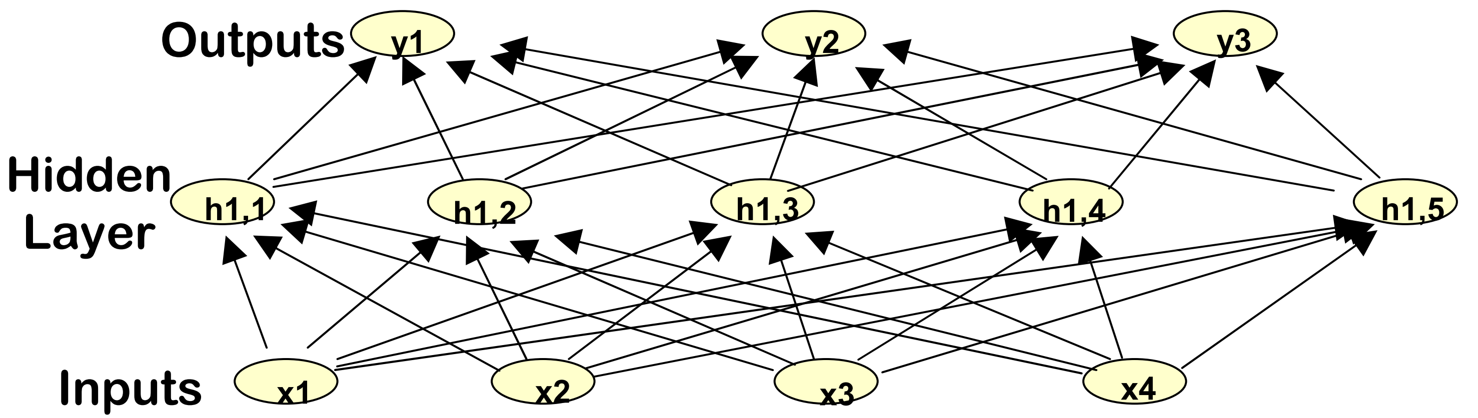


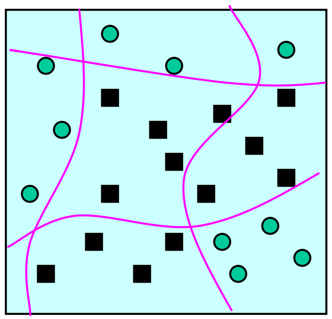
LINEAR PARTITION IMPOSSIBLE

Predictive Methods

Multilayer Perceptron (Artificial Neural Networks, ANN).

- The one-layer perceptron is not able to learn even the most simplest functions.
- We add new internal layers.



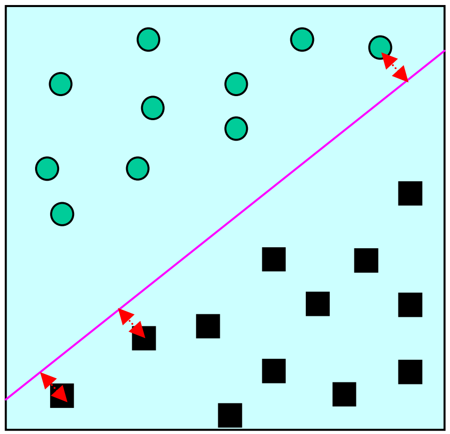


NON-LINEAR MULTIPLE PARTITION IS POSSIBLE WITH 4 INTERNAL UNITS

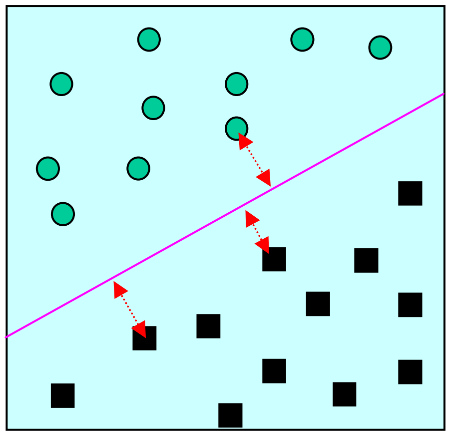
Predictive Methods

Support Vector Machines (SVM) / Kernel methods

- The basis is a very simple classifier.
  - The typical classifier is just the line (in more dimensions, a hyperplane) which splits the two classes more neatly in such a way that the three nearest examples to the borderline (the three support vectors) are as far as possible.



Data is perfectly split, but the three nearest examples are very near to the border line.



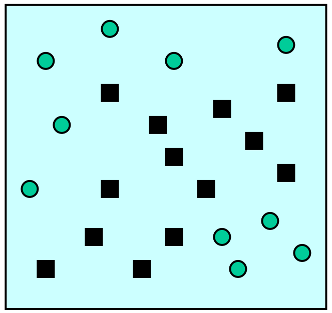
Data is perfectly split, but now the three nearest examples are much farther.

Predictive Methods

Support Vector Machines (SVM) / Kernel methods

- This linear discriminant is very efficient (even for hundreds of dimensions/attributes), since only a few examples are considered (many of them far away are just not considered).

What happens if the data is not linearly separable?



- A kernel function is applied in order to increase the number of dimensions, which usually implies that now the data becomes linearly separable..

Predictive Methods

Decision Trees (ID3 (Quinlan), C4.5 (Quinlan), CART).

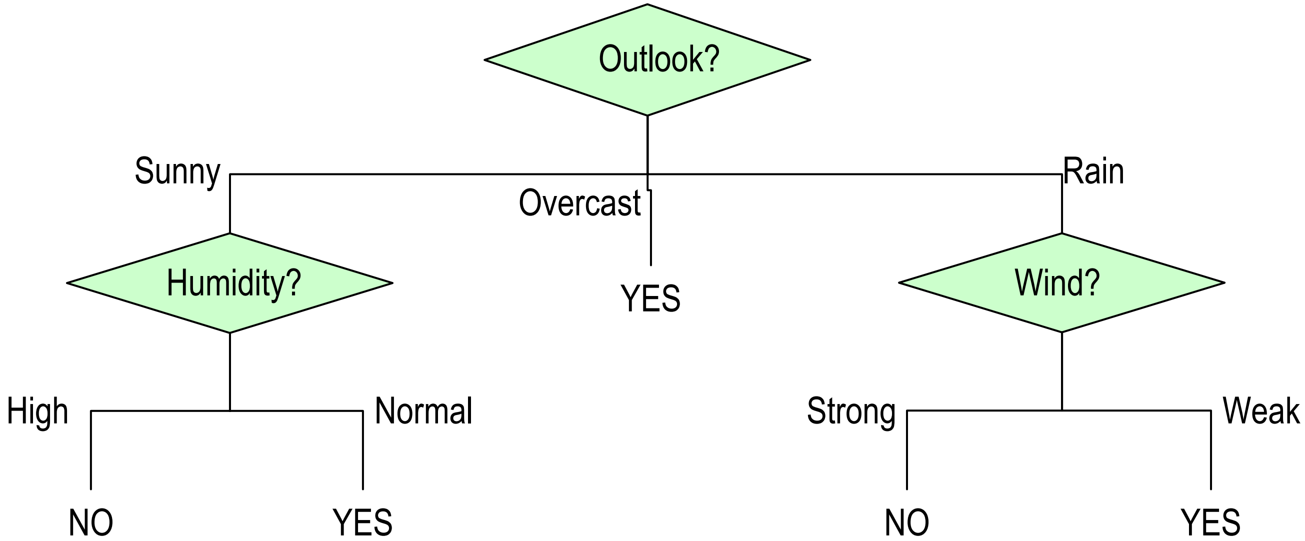
- **Example C4.5 with nominal data:**

Example	Sky	Temperature	Humidity	Wind	PlayTennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Predictive Methods

Decision Trees.

- **Example C4.5 with nominal data:**



E.g. the instance:

(Outlook = sunny, Temperature = cool, Humidity = high, Wind = strong

is NO.

Predictive Methods

Naive Bayes Classifiers.

- More frequently used with nominal/discrete variables. E.g. playtennis:
- We want to classify a new instance:

(Outlook = sunny, Temperature = cool, Humidity = high, Wind = strong)

$$V_{NB} = \arg\max_{c_i \in \{yes,no\}} P(c_i) \prod_j P(x_j \mid c_i) =$$

$$= \arg\max_{c_i \in \{yes,no\}} P(c_i) \cdot P(Outlook = sunny \mid c_i) \cdot P(Temperature = cool \mid c_i)$$

$$\cdot P(Humidity = high \mid c_i) \cdot P(Wind = strong \mid c_i)$$

- Estimating the 10 necessary probabilities:

  P(Playtennis=yes)=9/14=.64,   P(Playtennis=no)=5/14=.36

  P(Wind=strong|Playtennis=yes)=3/9=.33

  P(Wind=strong|Playtennis=no)=3/5=.60

  ...

- We have that:

  P(yes)P(sunny|yes)P(cool|yes)P(high|yes)P(strong|yes)=0.0053

  P(no)P(sunny|no)P(cool|no)P(high|no)P(strong|no)=0.206

Predictive Methods

Method comparison:

- **k-NN:**
  - Easy to use.
  - Efficient if the number of examples is not very high.
  - The value *k* can fixed for many applications.
  - The partition is very expressive (complex borders).
  - Only intelligible visually (2D or 3D).
  - Robust to noise but not to non-relevant attributes (distances increases, known as the "the curse of dimensionality")

- **Neural Networks**
  - **(multilayer):**
    - The number of layers and elements for each layer are difficult to adjust.
    - Appropriate for discrete or *continuous* outputs.
    - Low intelligibility.
    - Very sensitive to outliers (anomalous data).
    - Many examples needed.

Predictive Methods

Method comparison (contd.):

- **Naive Bayes:**
  - Very easy to use.
  - Very efficient (even with many variables).
  - THERE IS NO MODEL.
  - Robust to noise.

- **Decision Trees:**
  - **(C4.5):**
    - Very easy to use.
    - Admit discrete and continuous attributes.
    - The output must be finite and discrete (although there are regression decision trees)
    - Noise tolerant, to non-relevant attributes and *missing attribute values*.
    - High intelligibility.