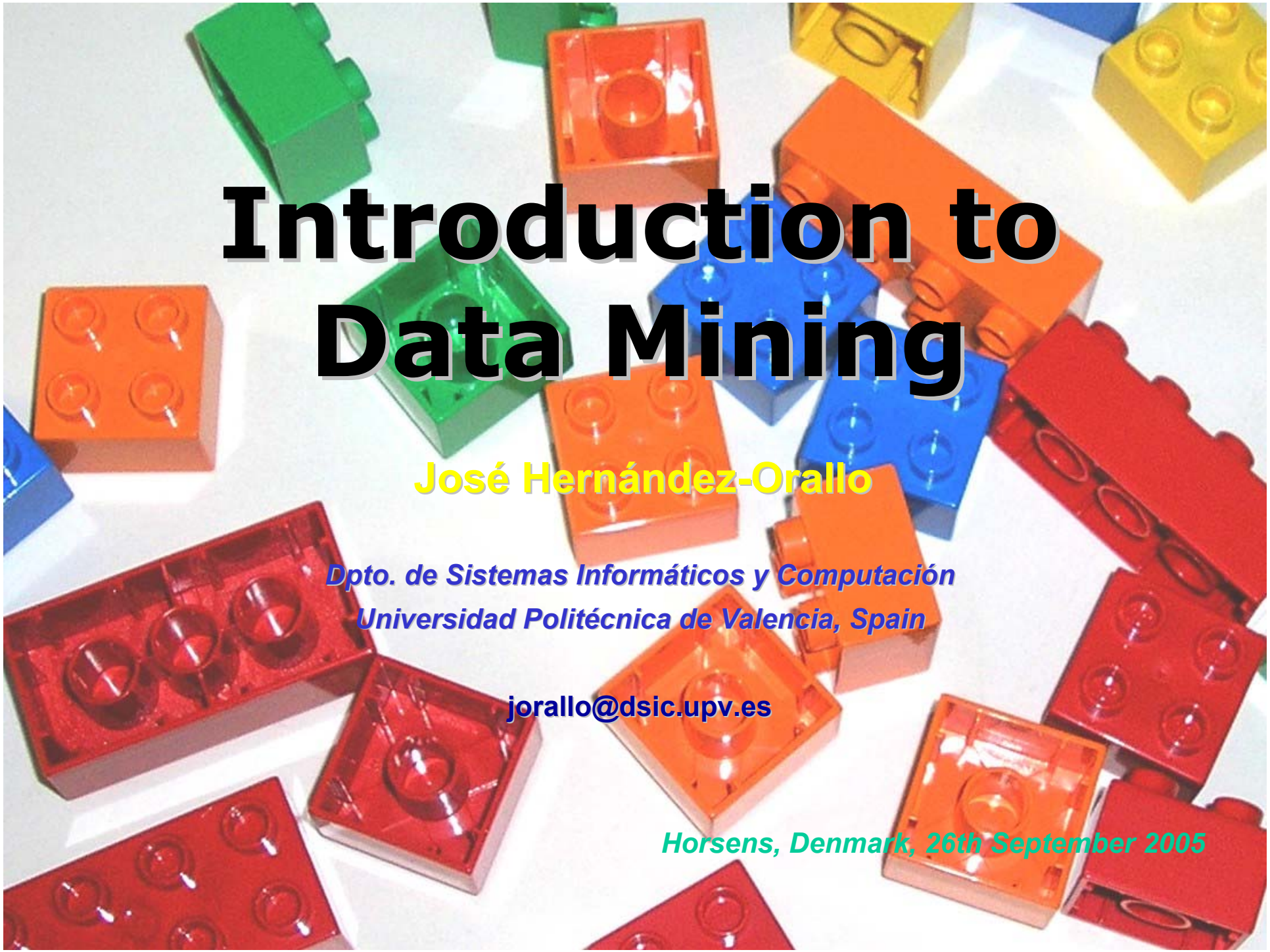


Introduction to Data Mining

José Hernández-Orallo

Dpto. de Sistemas Informáticos y Computación

Universidad Politécnica de Valencia, Spain

jorallo@dsic.upv.es

Horsens, Denmark, 26th September 2005

Outline

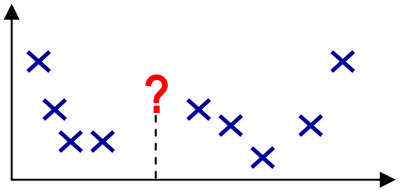
- Motivation. BI: Old Needs, New Tools.
- Some DM Examples.
- Data Mining: Definition and Applications
- The KDD Process
- Data Mining Techniques
- Development and Implementation

Taxonomy of Data Mining Techniques

Examples:

Predictive

- ***Interpolation***:



f(2.2)=?

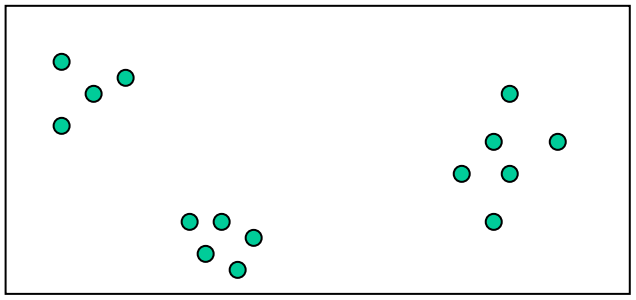
- ***Sequential prediction***: 1, 2, 3, 5, 7, 11, 13, 17, 19, ... ?
- ***Supervised learning***:

1 3 -> even.
3 5 -> even.
7 2 -> odd.

4 2 -> ?

Descriptive

- ***Clustering (unsupervised learning)***:



How many groups are there?
Which groups do I make?

- ***Exploratory Analysis: Correlations, Associations and Dependencies***

Taxonomy of DM Techniques

PREDICTIVE: Interpolation and Sequential Prediction.

- Generally the same techniques for both, but very different depending on whether the output variable is numerical or nominal:
  - ***Continuous output data (numerical/real) :***
    - ***Linear regression***:
      - Global (classical) linear regression.
      - Locally weighted linear regression.
    - ***Non-linear regression***: logarithmic, pick & mix, ...
  - ***Discrete output data (nominal):***
    - Specific techniques
    - Different from classification in the way that there is a temporal character of the data.

Taxonomy of DM Techniques

PREDICTIVE: Supervised learning.

Depending on the number and kind of "classes":

- *discrete*: "classification" or "categorisation".
  - Depending on whether we estimate a function (exclusive) or a correspondence (non-exclusive) :
    - classification: disjoint classes.
      - *example: determine the blood group from the parents' blood groups.*
    - categorisation: overlapping classes.
      - *example: determine the topics covered by a webpage.*
- Continuous or orderly discrete: known as "estimation".
  - *example: estimate the number of children for a given family from the features of this and other families.*

Taxonomy of DM Techniques

PREDICTIVE: Supervised learning (classification).

- Techniques:
  - k-NN (Nearest Neighbor).
  - k-means (competitive learning).
  - Perceptron Learning.
  - Multilayer ANN methods (e.g. backpropagation).
  - Radial Basis Functions.
  - Decision Tree Learning (e.g. ID3, C4.5, CART).
  - Bayes Classifiers.
  - Center Splitting Methods.
  - Rules (CN2)
  - Pseudo-relational: Supercharging, Pick-and-Mix.
  - Relational: ILP, IFLP, *SCIL*.

Similarity-Based

Fence and Fill

Taxonomy of DM Techniques

DESCRIPTIVE: Exploratory Analysis

- Techniques:
  - Correlation analysis
  - Associations.
  - Dependencies.
  - Anomalous data detection.
  - Scatter analysis.

Taxonomy of DM Techniques

DESCRIPTIVE: Clustering (Unsupervised learning)

- *Clustering* techniques:
  - k-means (competitive learning).
  - Kohonen neural networks
  - EM (Estimated Means) (Dempster et al. 1977).
  - Cobweb (Fisher 1987).
  - AUTOCLASS
  - ...

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	✓	✓			

Hypothesis Evaluation

Which hypothesis to choose?

- APPROXIMATIONS:
  - To assume a priori distributions.
  - Simplicity criteria (or minimum message/description, MML/MDL).
  - Partition: Training Set and Test Set.
    - Cross-validation.
  - Based on reinforcement.

Hypothesis Evaluation

SAMPLE PARTITION

- The evaluation of a hypothesis (model) with the same data which have been used to generate the hypothesis always gives much too optimistic results.

Solution: split into Training Set and Test Set.

- If the available data is large:
  - *Training Set*: set with which the algorithm learns one or more hypotheses.
  - *Test Set*: set with which the best hypothesis is chosen and its validity estimated.
- If the available data is small:
  - In order to take the most from the data, we can use **cross-validation**.

Hypothesis Evaluation

Which measure is used for evaluation?

- For problems with a discrete class, we estimate the "accuracy" (percentage of right guesses) or other measures (AUC, logloss, f-measure, …).

- For problems with a *continuous class*, we can use the squared error, the absolute error, the relative error or others.

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 :

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

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* $= (r_c \text{ or } r_c / n)$: $P(Cons \wedge Ante)$

Descriptive Methods

Association Rules. Example:

Id	*Family Income*	*City*	*Profession*	*Age*	*Chldrn*	*Fat*	*Married*
11251545	5.000.000	Barcelona	Executive	45	3	Y	Y
30512526	1.000.000	Melilla	Lawyer	25	0	Y	N
22451616	3.000.000	León	Executive	35	2	Y	Y
25152516	2.000.000	Valencia	Waiter	30	0	Y	Y
23525251	1.500.000	Benidorm	Thematic Park Animator	30	0	N	N

- Non-directional Associations:

  Married and (Chldrn > 0) are associated (80%, 4 cases).
  Fat and married are associated (80%, 4 cases)

- Directional Associations:

  (Chldrn > 0) → Married (100%, 2 cases).
  Married → Fat (100%, 3 cases)
  Fat → Married (75%, 3 cases)

Descriptive Methods
Unsupervised Learning

Clustering:

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

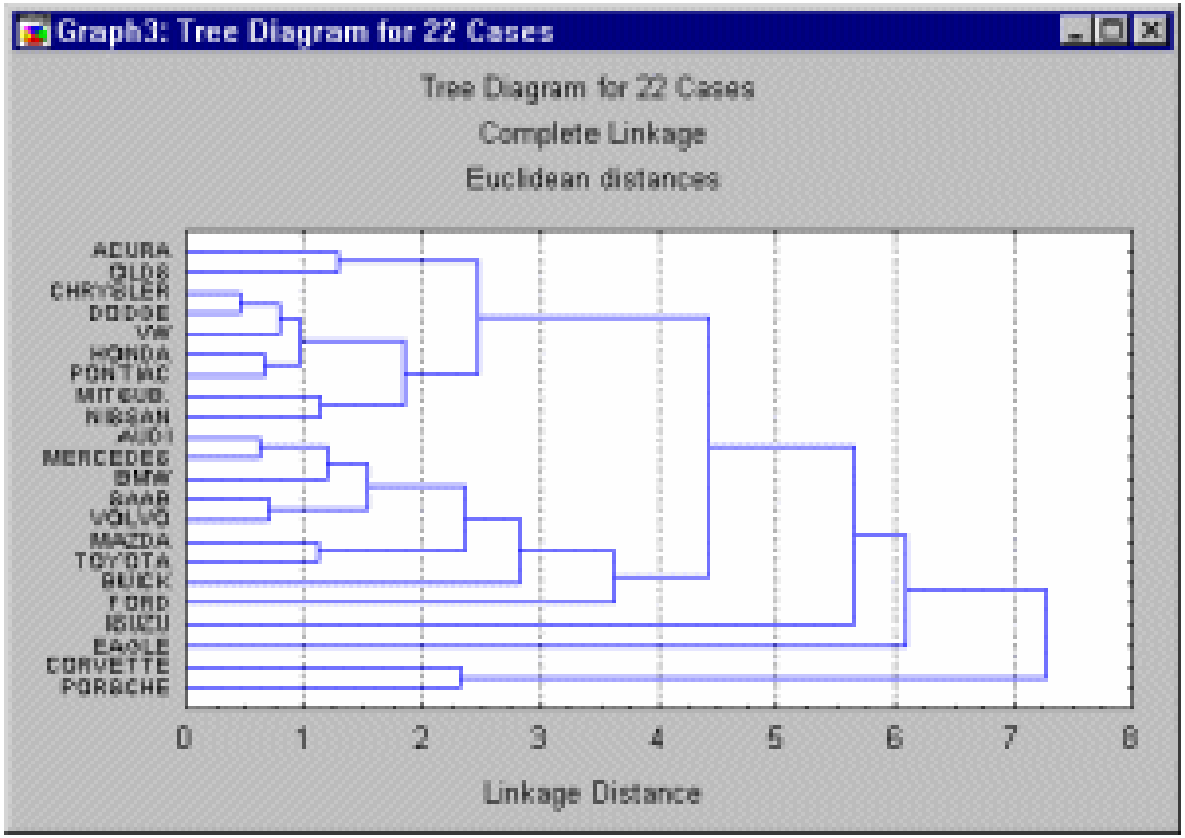
- Clustering method:
  - Hierarchical: the data is grouped in an tree-like way (e.g. the animal realm).
  - Non-hierarchical: the data is grouped in a one-level partition.
    - (a) Parametrical: we assume that the conditional densities have some known parametrical form (e.g. Gaussian), and the problem 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
Unsupervised Learning

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.

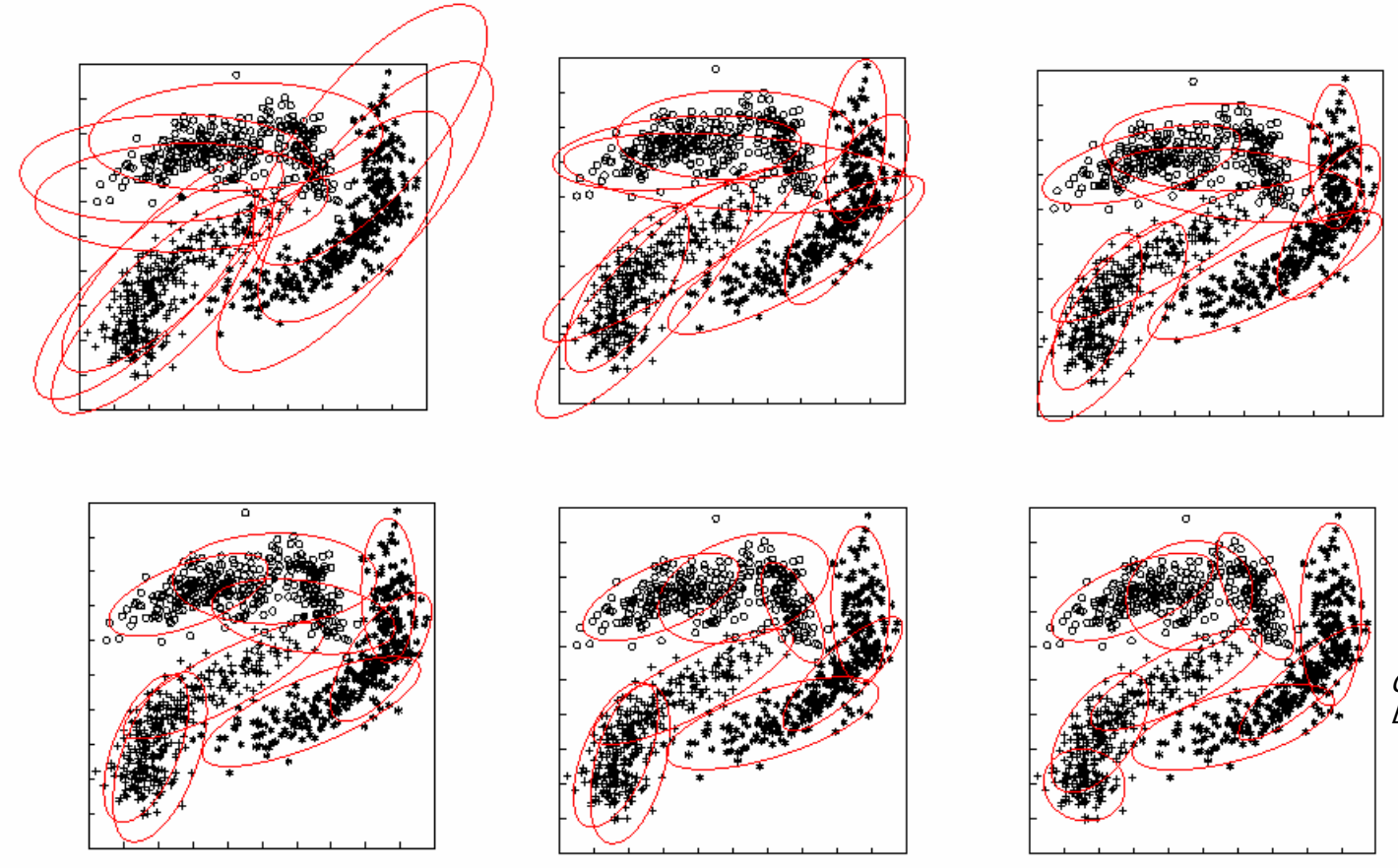
This gives different clustering at several levels, in a hierarchical way. This is called a *Horizontal Hierarchical Tree Plot* (or dendrogram)



Descriptive Methods
Unsupervised Learning

Clustering. Parametrical Methods:

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



Charts:
Enrique Vidal

Descriptive Methods
Unsupervised Learning

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:

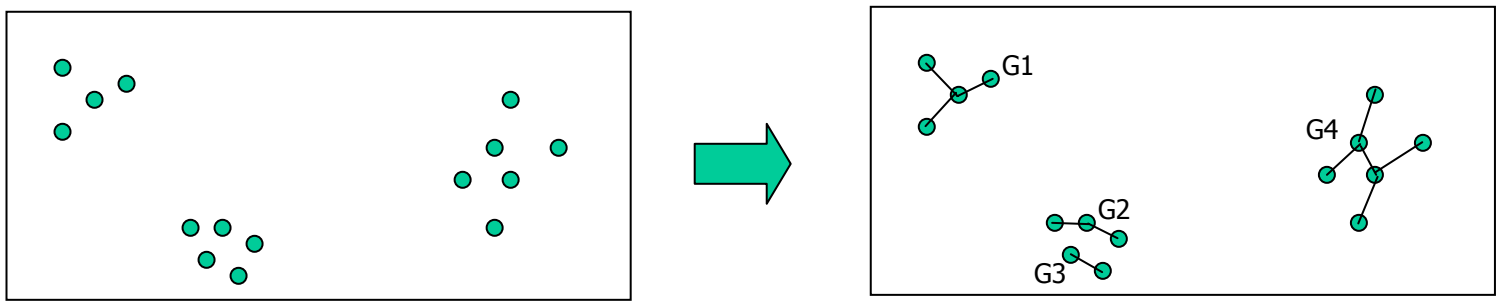
- Cobweb (Fisher 1987).
- AUTOCLASS (Cheeseman & Stutz 1996)

Descriptive Methods
Unsupervised Learning

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.

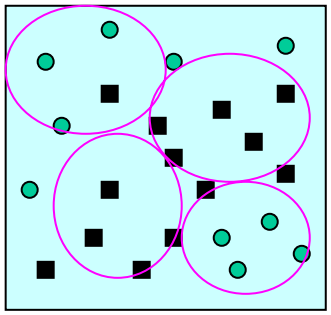
- In some cases, the clusters are too slow.
  - Variants: k-NN.

Descriptive Methods
Unsupervised Learning

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):

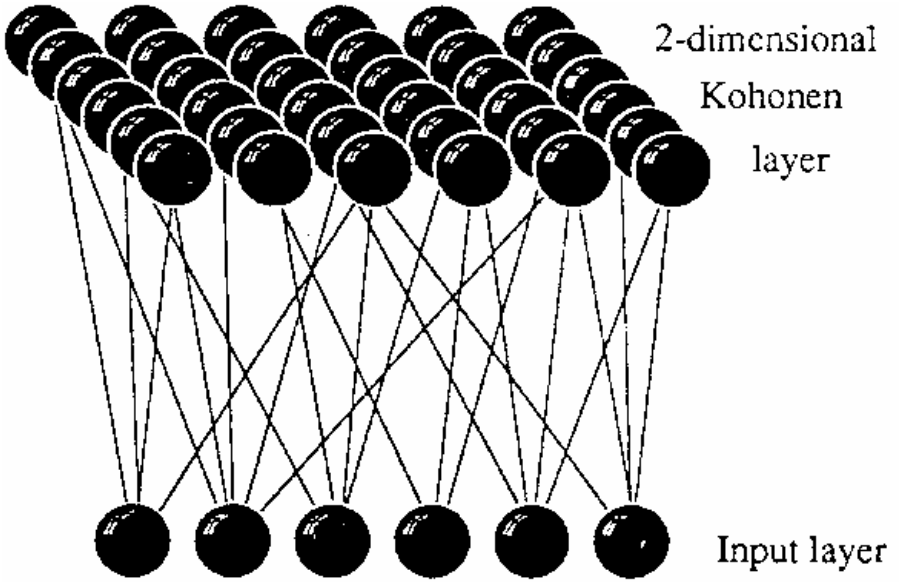
- Incremental refinement.

Descriptive Methods
Unsupervised Learning

Clustering. Non-Parametrical Methods

SOM (Self-Organizing Maps) or Kohonen Networks

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

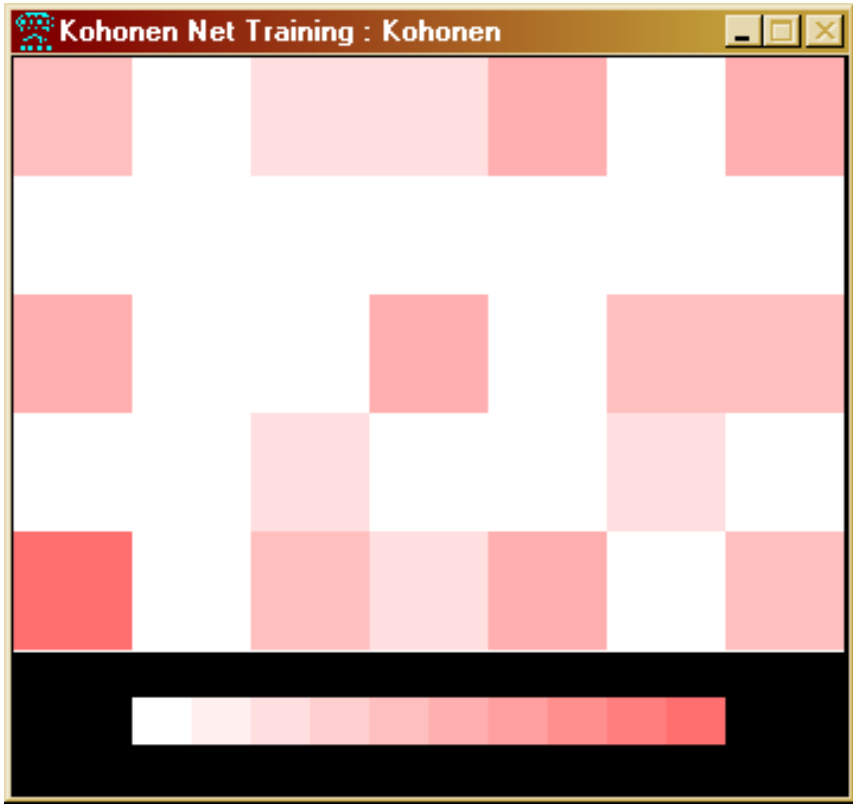


The neuron matrix is the last layer in a bidimensional grid.

Descriptive Methods
Unsupervised Learning

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.

Other Descriptive Methods

Statistical Analysis:

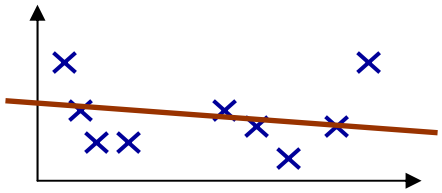
- Data distribution analysis.
- Anomalous data detection.
- Scatter analysis.

  - *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.
Interpolation and Sequential Prediction

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

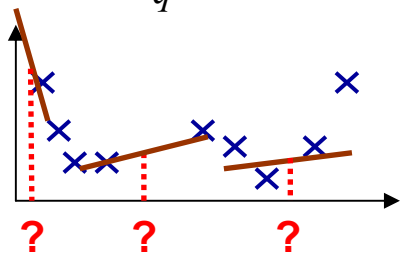
- New dimensions are added, combining the given dimensions. E.g. $x_4 = x_1 \cdot x_2$, $x_5 = x_3^2$, $x_6 = x_1^{x_2}$ and next we get a linear function for x_1, x_2, x_3, x_4, x_5, x_6

Predictive Methods. Interpolation and Sequential Prediction

Locally weighted linear regression.

- The linear function is approximated for each point x_q that needs to be interpolated:

$$\hat{f}(x) = w_0 + w_1 x_1 + \ldots + w_m x_m$$



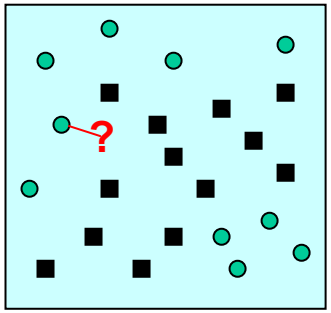
Adaptive Regression.

- *Specialised in sequential prediction. It is regularly used in sound and video compression, in networks (the new frame), etc.*

- Algorithms much more sophisticated (Markov chains, VQ)

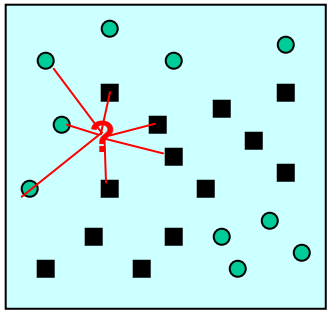
Predictive Methods.
Supervised Learning

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



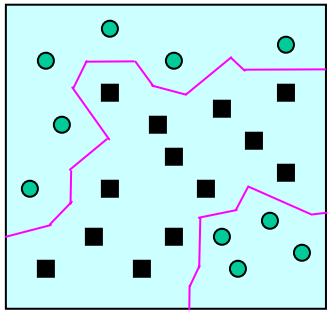
Classifies circle

1-nearest neighbor



Classifies square

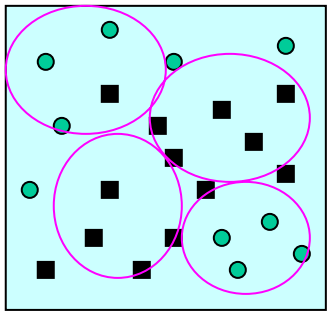
7-nearest neighbor



1-nearest neighbor PARTITION (Poliedric or Voronoi)

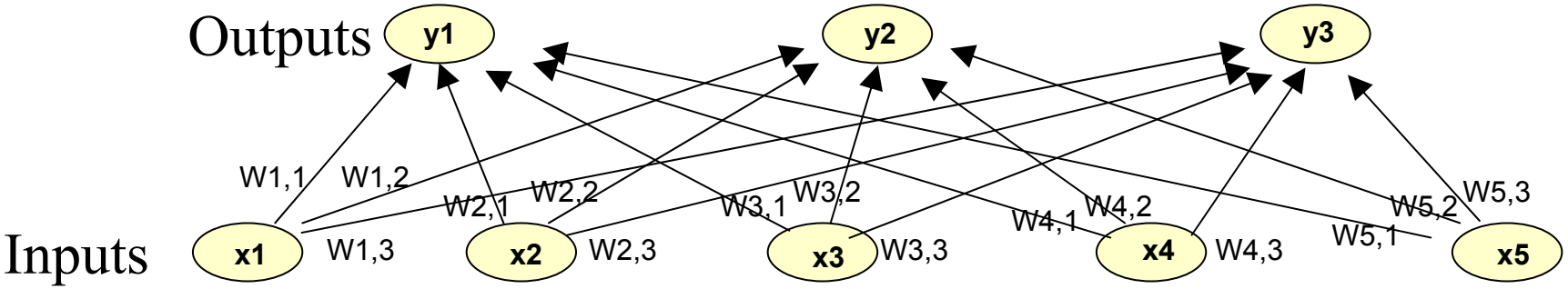
***k*-means clustering**:

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



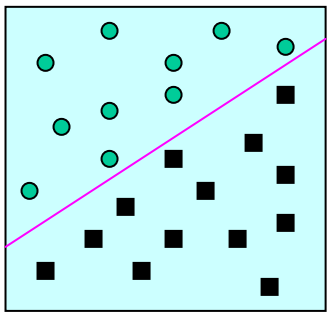
Supervised Learning

Perceptron Learning.

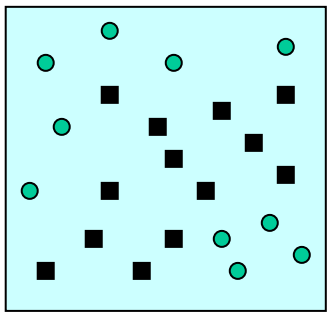


- Computes a linear function.

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



LINEAR PARTITION POSSIBLE

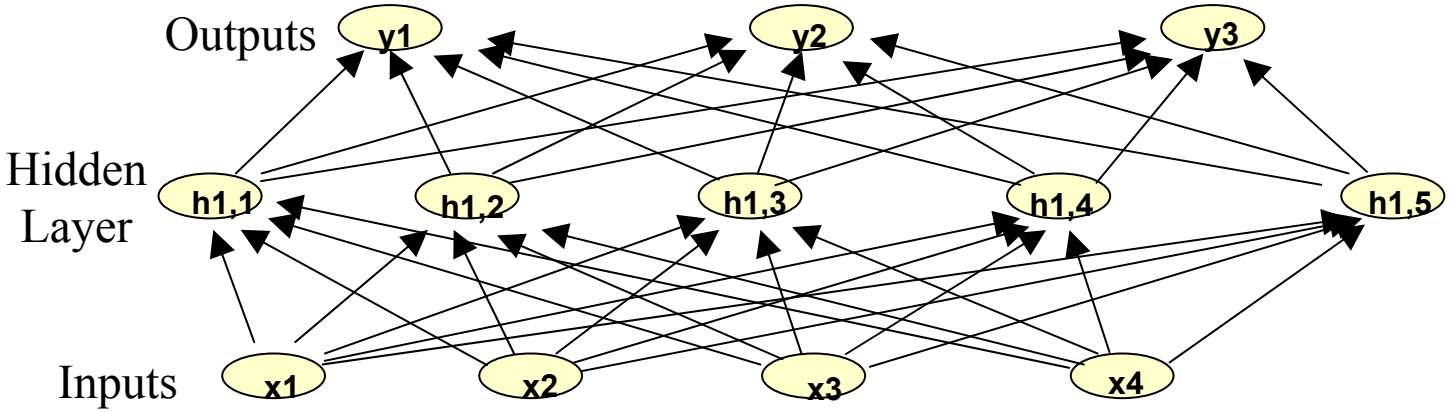


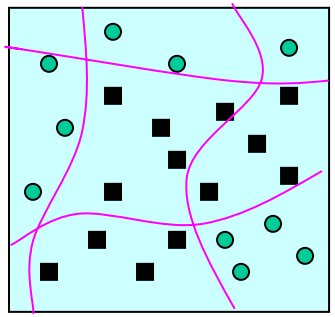
LINEAR PARTITION IMPOSSIBLE

Supervised Learning

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

Supervised Learning

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

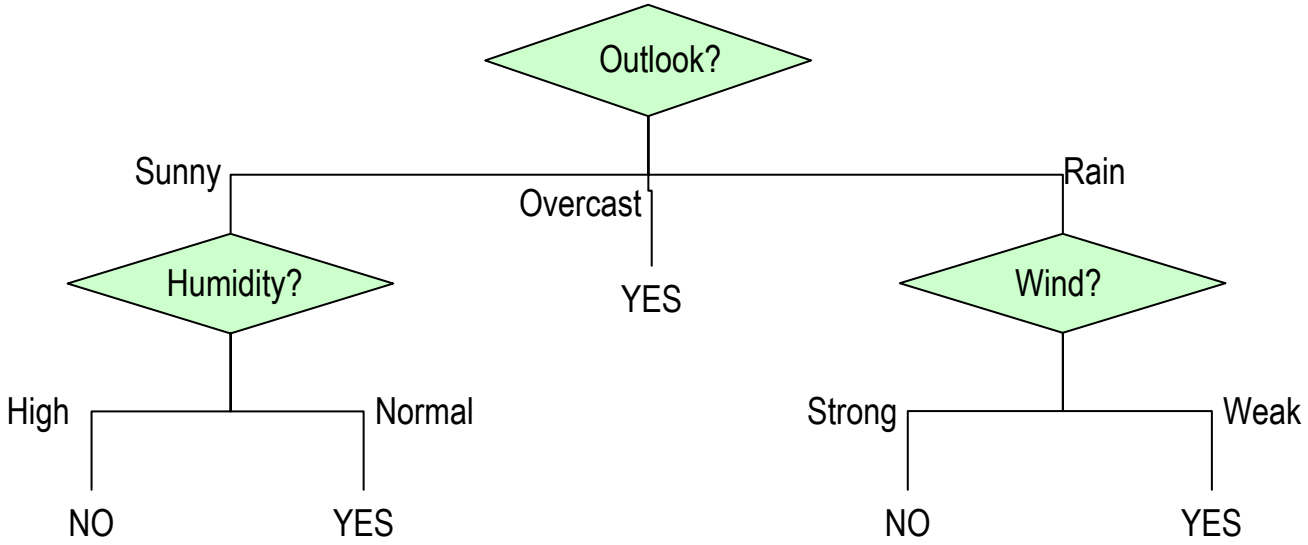
- Example C4.5 with nominal data:

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

Supervised Learning

Decision Trees.

- Example C4.5 with nominal data:



E.g. the instance:

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

is NO.

Supervised Learning

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

Supervised Learning

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.

Supervised Learning

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.

Supervised Learning. Oversampling

Oversampling:

- In many classification problems on databases, there may be a much higher proportion of one class over the rest. There are minority classes.
- Problem: the algorithm can take the minority class as noise and can be ignored by the theory. Example:
  - If a binary problem (*yes* / *no*) there are only 1% examples from class *no*, the model "everything is from class yes" would have 99% accuracy.

Solutions:

- Use oversampling...
- ROC Analysis

Supervised Learning. Oversampling

Oversampling / balancing:

- Oversampling consists in repeat/filter the examples of the classes with lower/higher proportion, so maintaining the classes with higher/lower proportion.
  - Changes the class distribution, but allows a better use of the examples of the rare classes.

When should we use oversampling?

- When a class is very rare: e.g. predict machine failures, anomalies, exceptions, etc.
- When all the classes (especially the minority classes) must be validated. E.g. if a rare class is the one of the unlawful customers.

Caveat: we must be very careful when evaluating the models.

Supervised Learning. Macro-average

Macro-average:

- An alternative to oversampling consists in calculating the average in a different way.
- Typically, accuracy is calculated as follows:

$$acc(h) = hits / total$$

(known as *micro-averaged accuracy*)

- The alternative is to calculate accuracy as follows:

$$acc(h) = \frac{hits_{class1} / total_{class1} + hits_{class2} / total_{class2} + ... + hits_{class-n} / total_{class-n}}{no.classes}$$

(known as *macro-averaged accuracy*)

This way we obtain a much more balanced result.

Supervised Learning.
Cost and Confusion Matrices.

Classification errors (class confusion) :

- In many data mining cases, the classification error over one class does not have the same economical, ethical or human consequences than the opposite error.

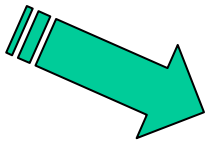
    - Example: classify a wheel tyre delivery as flawless or as defective.

Supervised Learning.
Cost and Confusion Matrices.

Cost and Confusion Matrices:

- There are techniques to weigh the classes → we can combine the "confusion matrices" with the "cost matrices" :

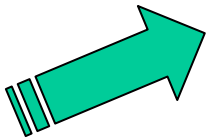
COST		*actual*		
		low	medium	high
predicted	low	0€	5€	2€
	medium	200€	-2000€	10€
	high	10€	1€	-15€



Overall cost:

-29787€

ERROR		*actual*		
		low	medium	high
predicted	low	20	0	13
	medium	5	15	4
	high	4	7	60



Supervised Learning.
Cost and Confusion Matrices.

Classification Errors and Mailings:

- Even more... There are specific techniques to evaluate the convenience of mailing campaigns (selective advertising through mail) :
  - Example: A company plans to do a mailing in order to foster the purchase of one of its products. In case of a positive response, the customers buy products with an average value of 100€. If 55% are production costs (fix and variable), we have that each positive response implies an average profit of 45€.
  - Each mailing costs 1€ (courier, brochures) and the whole of the campaign design (independently of the number) would have a cost of 20,000€.
  - With 1,000,000 customers, where 1% responds positively, how can we evaluate and apply a model which tells (ranks) which are the best customers for the campaign?

Supervised Learning.
Cost and Confusion Matrices.

Classification Errors and Mailings. Example:

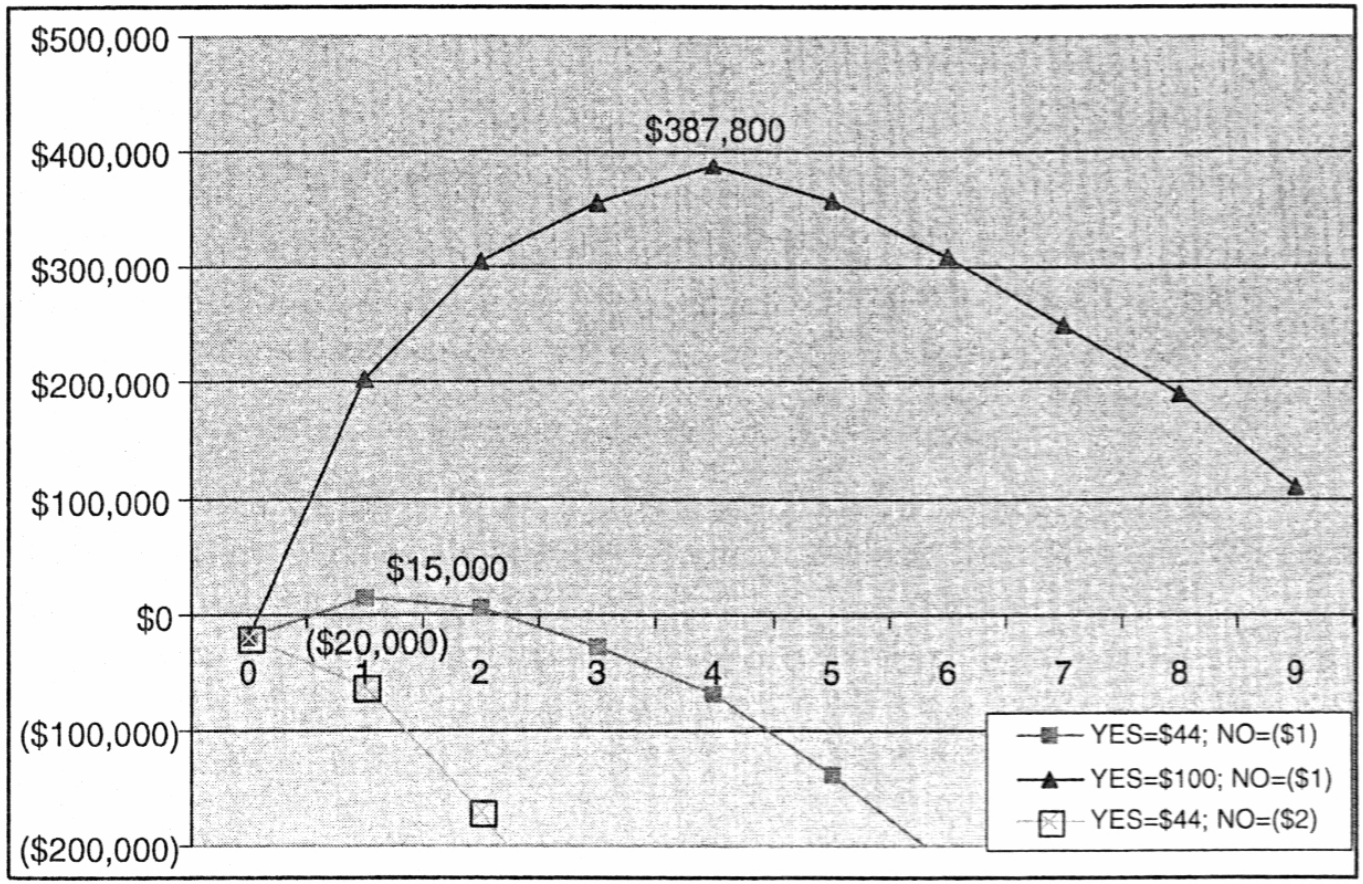
Table which shows the benefits for each decile:

DECILE	GAINS	CUM	LIFT	SIZE	SIZE(YES)	SIZE(NO)	PROFIT
0%	0.0%	0%	0.000	0	0	0	-($20,000)
10%	30.0%	30%	3.000	100,000	3,000	97,000	+ $15,000
20%	20.0%	50%	2.500	200,000	5,000	195,000	+ $5,000
30%	15.0%	65%	2.167	300,000	6,500	293,500	-($27,500)
40%	13.0%	78%	1.950	400,000	7,800	392,200	-($69,000)
50%	7.0%	85%	1.700	500,000	8,500	491,500	-($137,500)
60%	5.0%	90%	1.500	600,000	9,000	591,000	-($215,000)
70%	4.0%	94%	1.343	700,000	9,400	690,600	-($297,000)
80%	4.0%	98%	1.225	800,000	9,800	790,200	-($379,000)
90%	2.0%	100%	1.111	900,000	10,000	890,000	-($470,000)
100%	0.0%	100%	1.000	1,000,000	10,000	990,000	-($570,000)

Supervised Learning.
Cost and Confusion Matrices.

Classification Errors and Mailings. Example (contd.):

Graph showing the benefit for three different campaigns:



Supervised Learning.
Cost and Confusion Matrices.

Classification Errors:

- In this kind of problems (binary or ordered), it is preferable to construct models which can make probabilistic predictions (rankings), because these can be combined with costs in a more effective way.

  - E.g.: It is preferable to have a model which assess (in a 0 a 10 scale) how good a customer is, instead of a model which only tells whether the customer is bad or good.

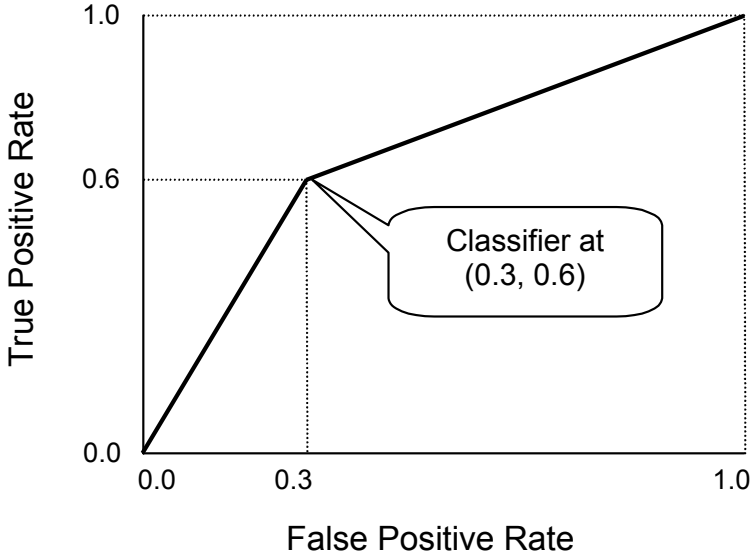
Supervised Learning.
ROC Analysis.

ROC Analysis (Receiver Operating Characteristic):

- Based on drawing the "true-positive rate" on the *y-axis* and the "false-positive rate" on the *x-axis*. E.g, given the following confusion matrix:

		Actual T	Actual F
Predicted	T	30	30
	F	20	70

- We would have TPR= 0.6 and FPR= 0.3.



Predictive Methods
Model Combination

- *BOOSTING/STACKING/RANDOMISING:*
  - The same algorithm is used to get different models from the same dataset.
  - Next, the models are combined.
- *VOTING/ARBITER/COMBINER:*
  - Different algorithms are used to learn different models from the same dataset.
  - Next, the models are combined.

- Different ways of combining the models:
  - WEIGHTING MAJORITY: the value is obtained by computing the mean (continuous output) or the median (discrete case).
  - STACKING/CASCADE: each model is use as an input variable for a second layer model.