An Introduction to Data Mining and Statistical Learning

Gilbert Saporta
Chaire de Statistique Appliquée & CEDRIC, CNAM,
292 rue Saint Martin, F-75003 Paris

gilbert.saporta@cnam.fr
http://cedric.cnam.fr/~saporta

Outline

1. What is data mining?
2. Some unsupervised methods
3. Some supervised methods
4. Statistical modelling
5. Predictive modelling and statistical learning
6. Discussion

1. What is data mining?

- Data mining is a new field at the frontiers of statistics and information technologies (database management, artificial intelligence, machine learning, etc.) which aims at discovering structures and patterns in large data sets.

1.1 Definitions:

- U.M.Fayyad, G.Piatetski-Shapiro : " *Data Mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data "*
- D.J.Hand : " *I shall define Data Mining as the discovery of interesting, unexpected, or valuable structures in large data sets"*

- The metaphor of Data Mining means that there are treasures (or nuggets) hidden under mountains of data, which may be discovered by specific tools.
- Data Mining is concerned with data which were collected for another purpose: it is a *secondary analysis* of data bases that *are collected Not Primarily For Analysis,* but for the management of individual cases (Kardaun, T.Alanko,1998) .
- Data Mining is not concerned with efficient methods for collecting data such as surveys and experimental designs (Hand, 2000)

What is new? Is it a revolution ?

- The idea of discovering facts from data is as old as Statistics which "*is the science of learning from data*" (J.Kettenring, former ASA president).
- In the 60's: Exploratory Data Analysis (Tukey, Benzecri..) « *Data analysis is a tool for extracting the diamond of truth from the mud of data.* » (J.P.Benzécri 1973)

1.2 Data Mining started from:

- an evolution of DBMS towards Decision Support Systems using a Data Warehouse.
- Storage of huge data sets: credit card transactions, phone calls, supermarket bills: giga and terabytes of data are collected automatically.
- Marketing operations: CRM (customer relationship management)
- Research in Artificial Intelligence, machine learning, KDD for Knowledge Discovery in Data Bases

1.3 Goals and tools

- Data Mining is a « secondary analysis » of data collected in an other purpose (management eg)
- Data Mining aims at finding structures of two kinds : models and patterns

- Patterns
  - a characteristic structure exhibited by a few number of points : a small subgroup of customers with a high commercial value, or conversely highly risked.
  - Tools: cluster analysis, visualisation by dimension reduction: PCA, CA etc. association rules.

Models

- Building models is a major activity for statisticians econometricians, and other scientists. A model is a global summary of relationships between variables, which both helps to understand phenomenons and allows predictions.
- DM is not concerned with estimation and tests, of prespecified models, but with discovering models through an algorithmic search process exploring linear and non-linear models, explicit or not: neural networks, decision trees, Support Vector Machines, logistic regression, graphical models etc.
- In DM Models do not come from a theory, but from data exploration.

process or tools?

- DM often appears as a collection of tools presented usually in one package, in such a way that several techniques may be compared on the same data-set.
- But DM is a process, not only tools:

Data → Information → Knowledge

preprocessing   analysis

The challenge of massive data sets: volume explosion (Michel Béra, 2009)

- In the 90s

→

Large in	
Neural Networks	**Statistics**
100,000 Weights	50 parameters
50,000 examples	200 cases

- Today
  - **Web transactions** At Yahoo ! (Fayyad, KDD 2007)

    → ± **16 B events - day,** 425 M visitors - month, **10 Tb data / day**

  - **Radio-frequency identification** (Jiawei, Adma 2006)

    → A retailer with 3,000 stores, selling 10,000 items a day per store

    300 million events per day (after redundancy removal)

  - **Social network** (Kleinberg, KDD 2007)

    → **4.4-million-node network** of declared friendships on blogging community

    240-million-node network of all IM communication over one month on Microsoft Instant Messenger

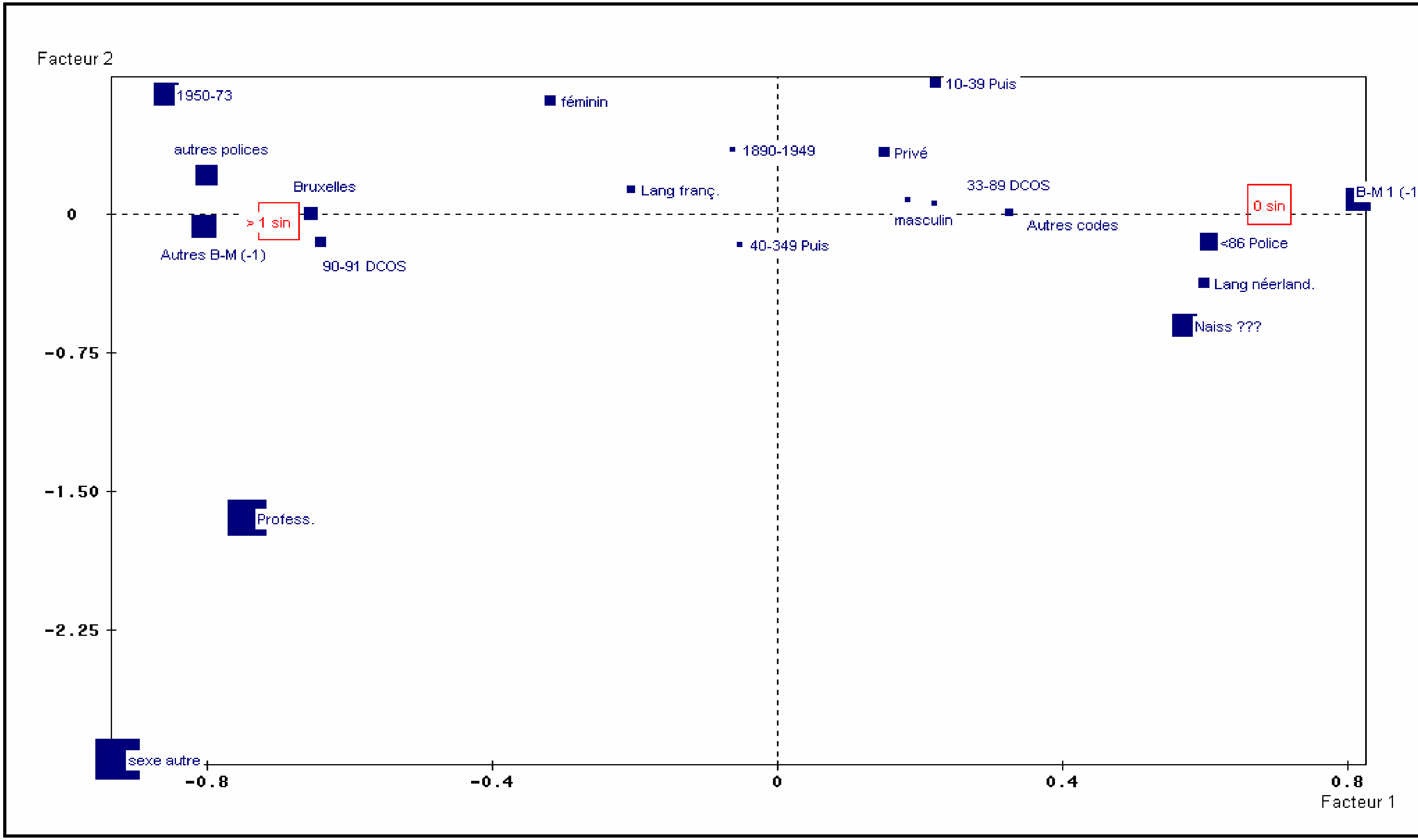
  - **Cellular networks**

    → A telecom carrier generates **hundreds of millions of CDRs / day**

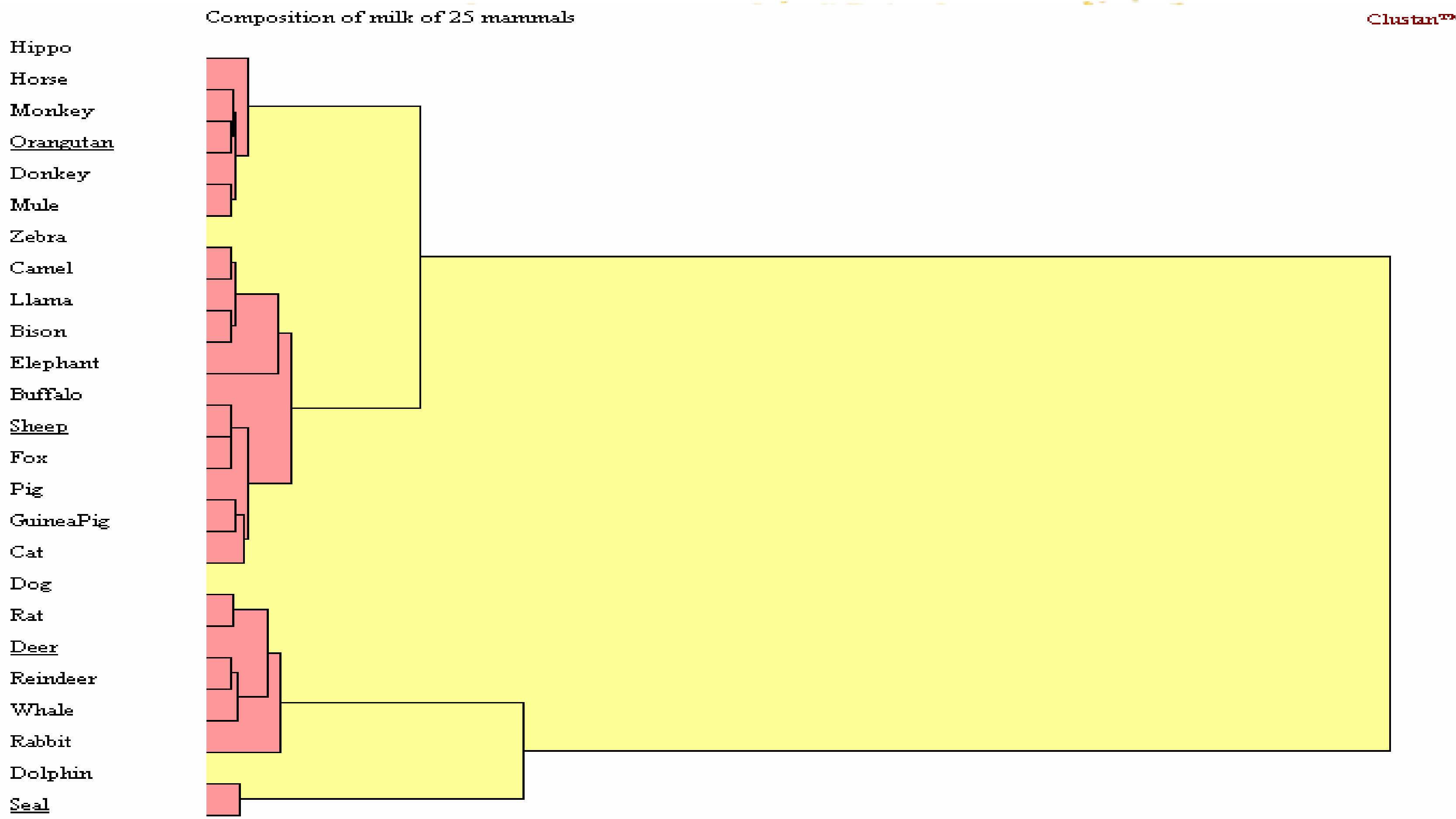
    The network generates technical data : **40 M events / day** in a large city

2. An overview of non supervised methods

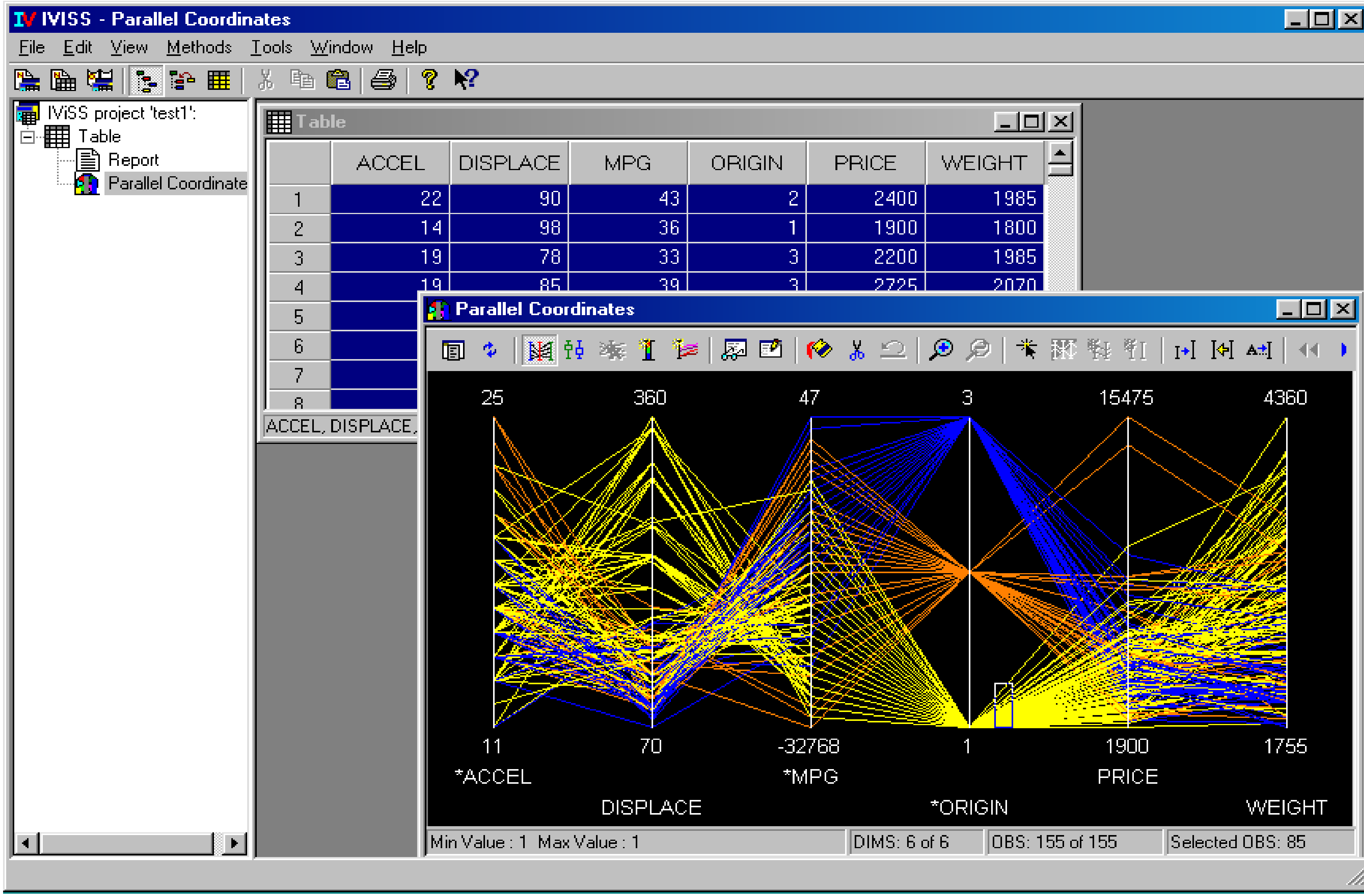
- Dimension reduction
  - Factor analysis
  - Cluster analysis
- Data visualisation
  - parallel coordinates
- Assocation rules discovery


Facteur 2
1950-73
féminin
10-39 Puis
autres polices
1890-1949
Privé
Bruxelles
Lang franç.
33-89 DCOS
0 sin
B-M 1 (-1)
0
> 1 sin
masculin
Autres codes
Autres B-M (-1)
40-349 Puis
<86 Police
90-91 DCOS
Lang néerland.
Naiss ???
-0.75
-1.50
Profess.
-2.25
sexe autre
-0.8
-0.4
0
0.4
0.8
Facteur 1

Hierarchical cluster analysis



Parallel coordinates



2. Association rule discovery, or market basket analysis

- Illustration with a real industrial example at Peugeot-Citroen car manufacturing company.
- (Ph.D of Marie Plasse).

ASSOCIATION RULES MINING

- **Marketing target : basket data analysis**

Basket	Purchases
1	**{bread, butter, milk}**
2	**{bread, meat}**
…	
n	**{fruit juice, fish, strawberries, bread}**

"90% of transactions that purchase bread and butter also purchase milk" (Agrawal et al., 1993)

$$\{\,bread,\ butter\,\} \Rightarrow \{milk\,\}$$

Itemset A	$\Rightarrow$	Itemset C
antecedent		consequent

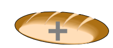
where $A \cap C = \emptyset$

{bread, butter } $\Rightarrow$ *{ milk }*

Itemset A antecedent	$\Rightarrow$	*Itemset C* consequent

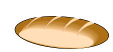
- **Reliability : Support :** % of transactions that contain all items of A and C

$$sup(A \Rightarrow C) = P(A \cap C) = P(C / A) \cdot P(A)$$

  - *Supp = 30 %* ➔ 30% of transactions contain   

- **Strength : Confidence :** % of transactions that contain C among the ones that contain C

$$conf(A \Rightarrow C) = P(C / A) = \frac{P(A \cap C)}{P(A)} = \frac{sup(A \Rightarrow C)}{sup(A)}$$

  - *Conf = 90 %* ➔ 90% of transactions that contain  tain also 

- **Support:** P(A∩C)
- **Confidence:** P(C/A)
- thresholds s0 et c0
- Interesting result only if P(C/A) is much larger than P(C) or P(C/not A) is low.
- **Lift:**

$$\frac{P(C/A)}{P(C)} = \frac{P(C \cap A)}{P(A)P(C)}$$

MOTIVATION

- **Industrial data :**
  - A set of vehicles described by a large set of binary flags

Vehicles	A1	A2	A2	A2	A3	...	AP
	1	0	0	1	0		0
	0	0	1	1	0		0
	0	1	0	0	1		0
	1	0	0	0	1		0
	0	1	0	0	0		1
	0	1	0	0	0		0
	0	0	1	0	0		0

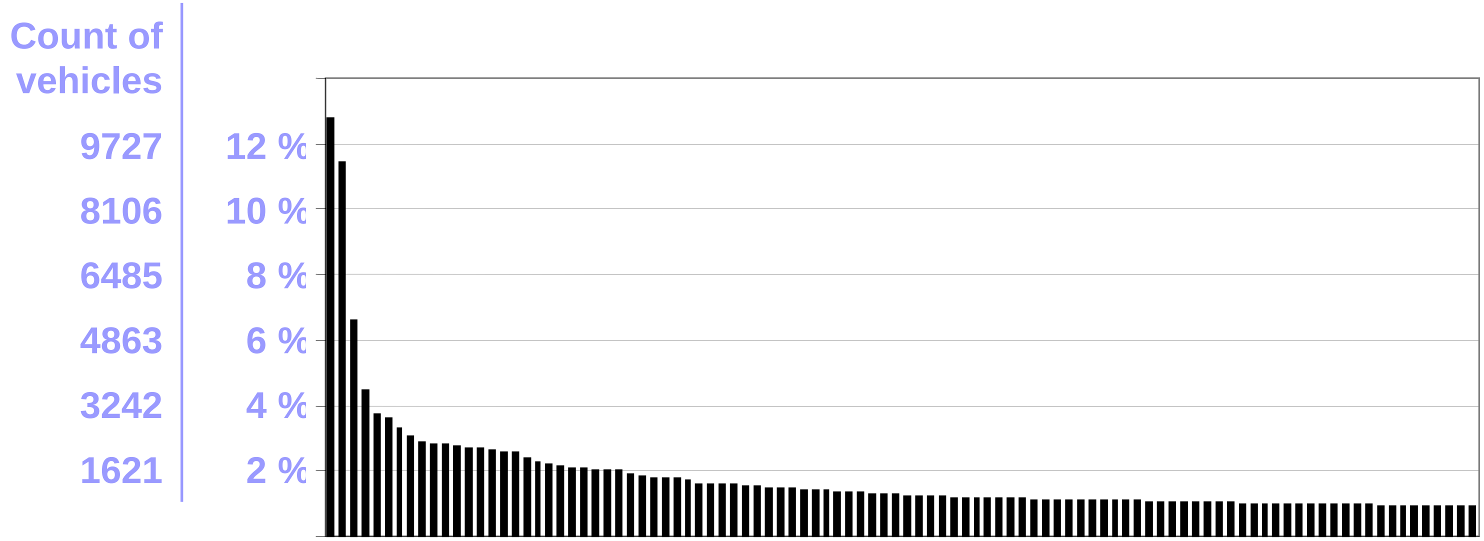
- **Motivation : decision-making aid**
  - Always searching for a greater quality level, the car manufacturer can take advantage of knowledge of associations between attributes.

- **Our work :**
  - We are looking for patterns in data : Associations discovery

DATA FEATURE

- **Data size :**
  - **More than 80 000 vehicles (≈transactions) ➔ 4 months of manufacturing**
  - **More than 3000 attributes (≈items)**

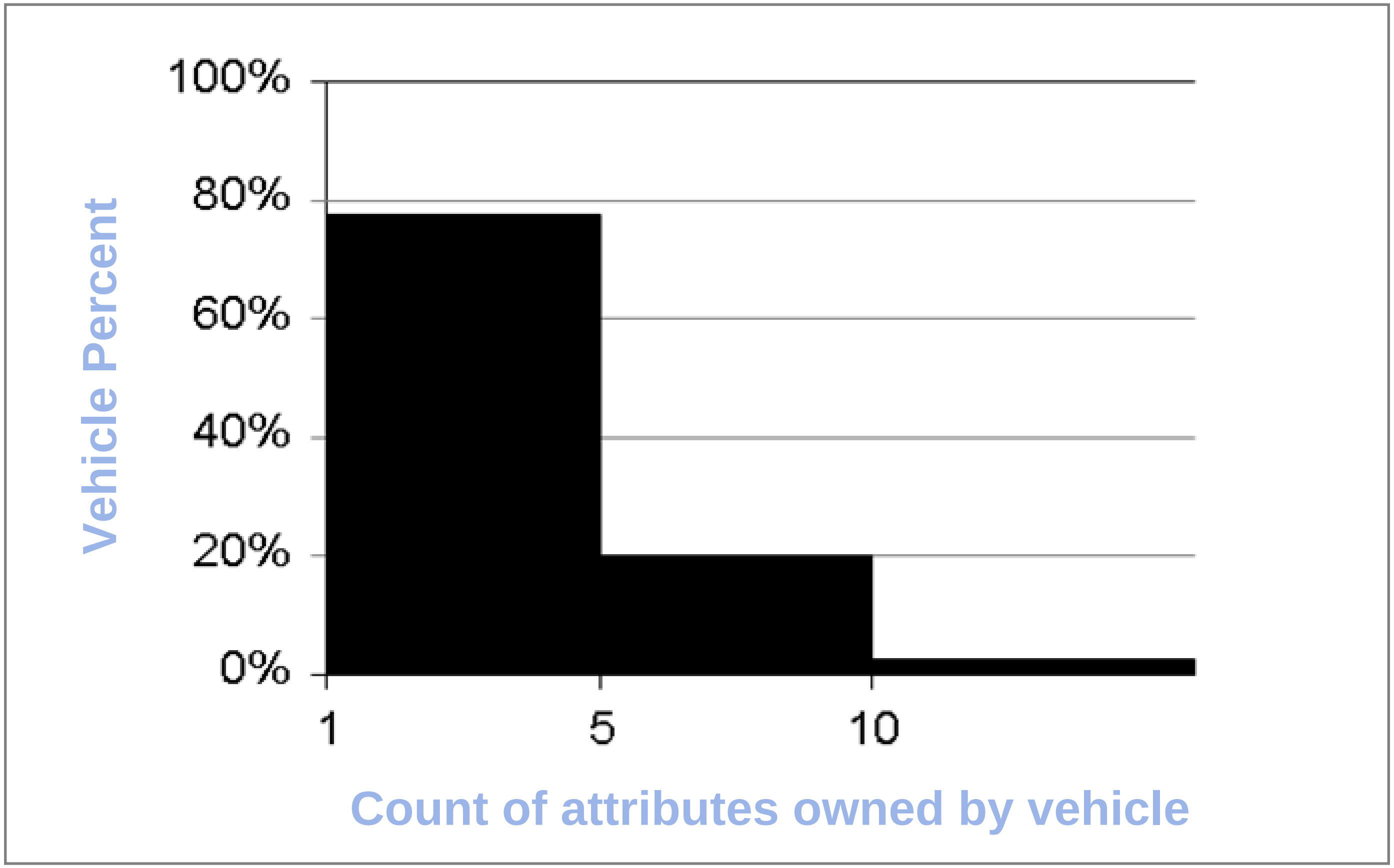
- **Sparse data :**



Count & percent of the 100 more frequent attributes

DATA FEATURE

- **Count of co-occurrences per vehicle :**



OUPUT : ASSOCIATION RULES

Minimum support (minimum count of vehicles that support the rule)	Minimum confidence	Count of rules	Maximum size of rules
500	**50 %**	**16**	**3**
400	**50 %**	**29**	**3**
300	**50 %**	**194**	**5**
250	**50 %**	**1299**	**6**
200	**50 %**	**102 981**	**10**
100	**50 %**	**1 623 555**	**13**

- **Aims :**
  - **Reduce count of rules**
  - **Reduce size of rules**

- **A first reduction is obtained by manual grouping :**

Minimum support	Minimum confidence	Count of rules	Maximum size of rules
100	**50 %**	**600636**	**12**

COMBINING CLUSTER ANALYSIS AND ASSOCIATION RULES

- **10-clusters partition with hierarchical clustering and Russel Rao coefficient**

Cluster	Number of variables in the cluster	Number of rules found in the cluster	Maximum size of rules
1	2	0	0
2	12	481170	12
3	2	0	0
4	5	24	4
5	117	55	4
6	4	22	4
7	10	33	4
8	5	22	4
9	16	1	2
10	2928	61	4

- **Cluster 2 is atypical and produces many complex rules**

- **Mining association rules inside each cluster except atypical cluster :**

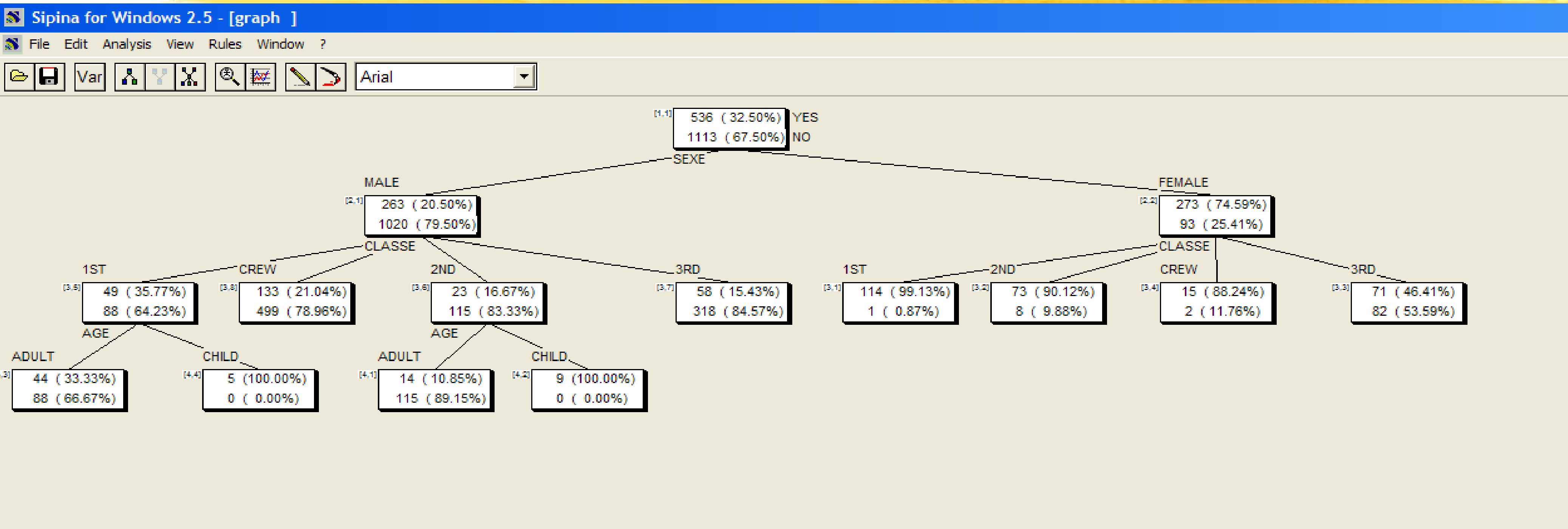
	Count of rules	Maximum size of rules	Reduction of the count of rules
Without clustering	600636	12	.
Ward - Russel & Rao	218	4	More than 99%

- **The number of rules to analyse has significantly decreased**
- **The output rules are more simple to analyse**
- **Clustering has detected an atypical cluster of attributes to treat separately**

3. Some supervised methods

- Trees
- Scores

Decision trees

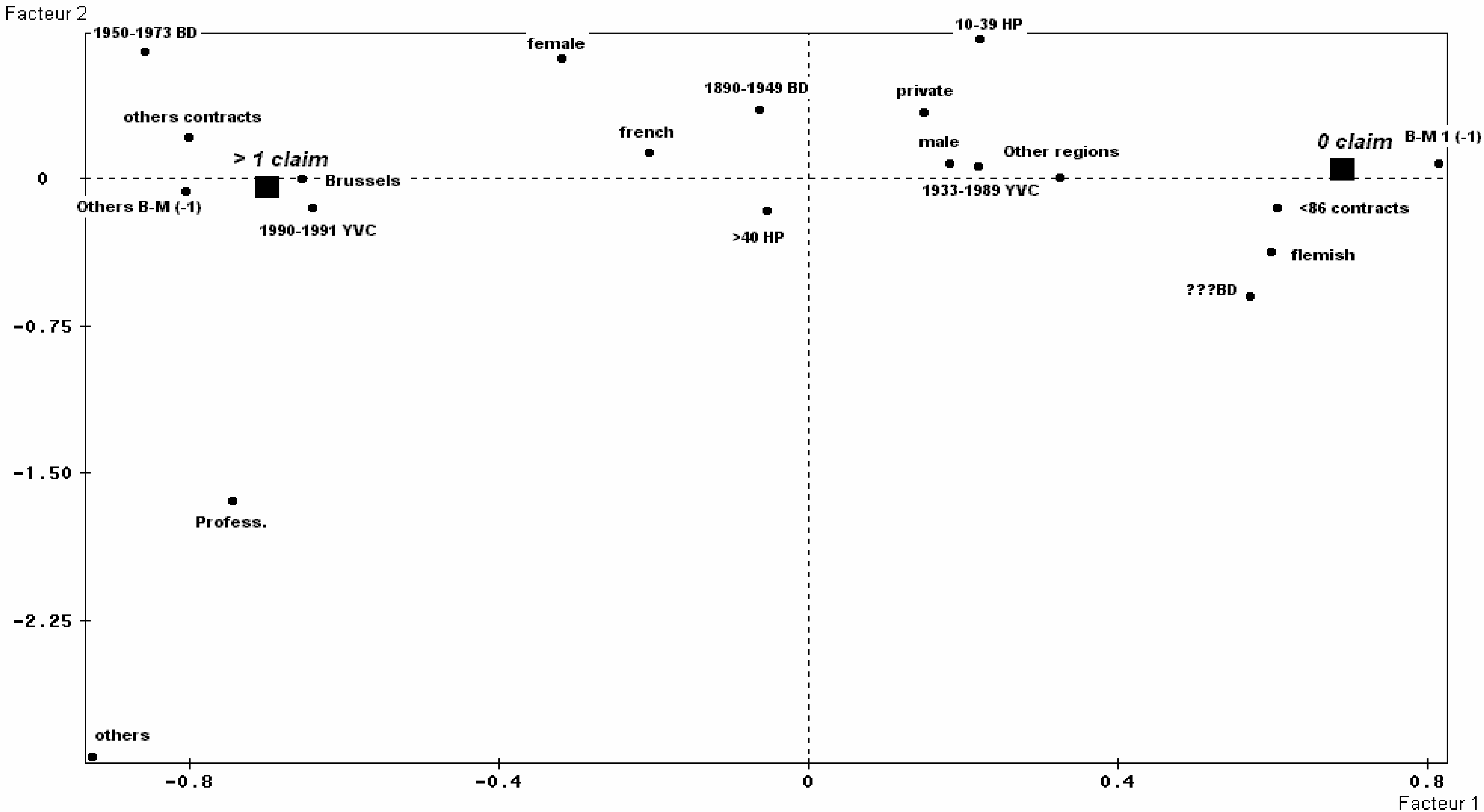


A scoring case study

An insurance example

- 1106 belgian automobile insurance contracts :
- 2 groups: « 1 good », « 2 bad »
- 9 predictors: 20 categories
  - Use type(2), gender(3), language (2), agegroup (3), region (2), bonus-malus (2), horsepower (2), duration (2), age of vehicle (2)

Principal plane MCA



Fisher's LDA

FACTORS	CORRELATIONS	LOADINGS
1 F 1	0.719	6.9064
2 F 2	*0.055*	*0.7149*
3 F 3	-0.078	-0.8211
4 F 4	*-0.030*	*-0.4615*
5 F 5	0.083	1.2581
6 F 6	*0.064*	*1.0274*
7 F 7	*-0.001*	*0.2169*
8 F 8	0.090	1.3133
9 F 9	-0.074	-1.1383
10 F 10	-0.150	-3.3193
11 F 11	*-0.056*	*-1.4830*
INTERCEPT		0.093575

R2 =	0.57923	F =	91.35686
D2 =	5.49176	T2 =	1018.69159

Score= 6.90 F1 - 0.82 F3 + 1.25 F5 + 1.31 F8 - 1.13 F9 - 3.31 F10

- **Transforming scores**
  - Standardisation between 0 and 1000 is often convenient
  - Linear transformation of score implies the same transformation for the cut-off point

Scorecard

CATEGORIES	COEFFICIENTS DISCRIMINANT FUNCTION	TRANSFORMED COEFFICIENTS (SCORE)
2 . Use type		
USE1 - Profess.	-4.577	0.00
USE2 - private	0.919	53.93
4 . Gender		
MALE - male	0.220	24.10
FEMA - female	-0.065	21.30
OTHE - companies	-2.236	0.00
5 . Language		
FREN – French	-0.955	0.00
FLEM - flemish	2.789	36.73
24 . Birth date		
BD1 - 1890-1949 BD	0.285	116.78
BD2 - 1950-1973 BD	-11.616	0.00
BD? - ???BD	7.064	183.30
25 . Region		
REG1 - Brussels	-6.785	0.00
REG2 – Other regions	3.369	99.64
26 . Level of bonus-malus		
BM01 - B-M 1 (-1)	17.522	341.41
BM02 - Others B-M (-1)	-17.271	0.00
27 . Duration of contract		
C<86 - <86 contracts	2.209	50.27
C>87 - others contracts	-2.913	0.00
28 . Horsepower		
HP1 - 10-39 HP	6.211	75.83
HP2 - >40 HP	-1.516	0.00
29 . year of vehicle construction		
YVC1 - 1933-1989 YVC	3.515	134.80
YVC2 - 1990-1991 YVC	-10.222	0.00

logistic regression

$$P(G_1|\mathbf{x}) = \frac{\exp(S(\mathbf{x}))}{1+\exp(S(\mathbf{x}))} = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}$$

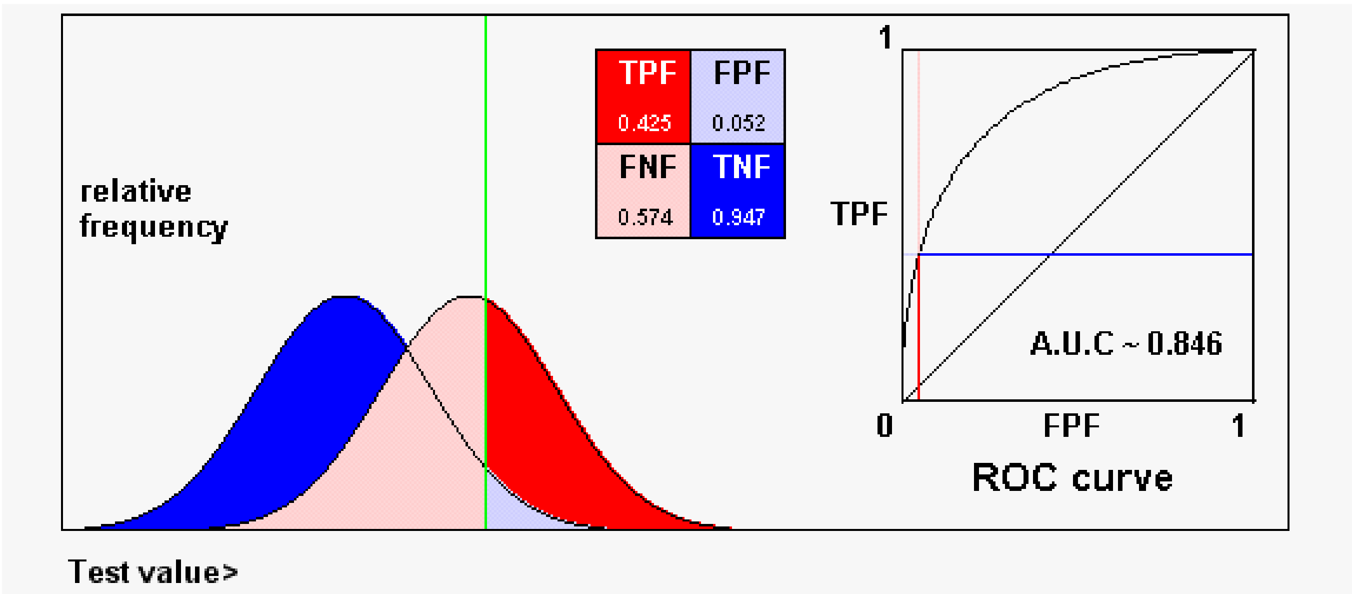
- A direct estimation of the posterior probability
- Estimation techniques differ: least squares in LDA , conditional maximum likelihood in logistic regression.

Performance measures for supervised binary classification

- Misclassification rate or score performance?
  - Error rate implies a strict decision rule.
- Scores
  - A score is a rating: the threshold is chosen by the end-user
  - Probability P(G1/x): also a score ranging from 0 to 1. Almost any technique gives a score.

ROC curve and AUC

- A synthesis of score performance for any threshold *s*. **x** is classified in group 1 if $S(\mathbf{x}) > s$
- Using s as a parameter, the ROC curve links the true positive rate 1-β to the false positive rate α .

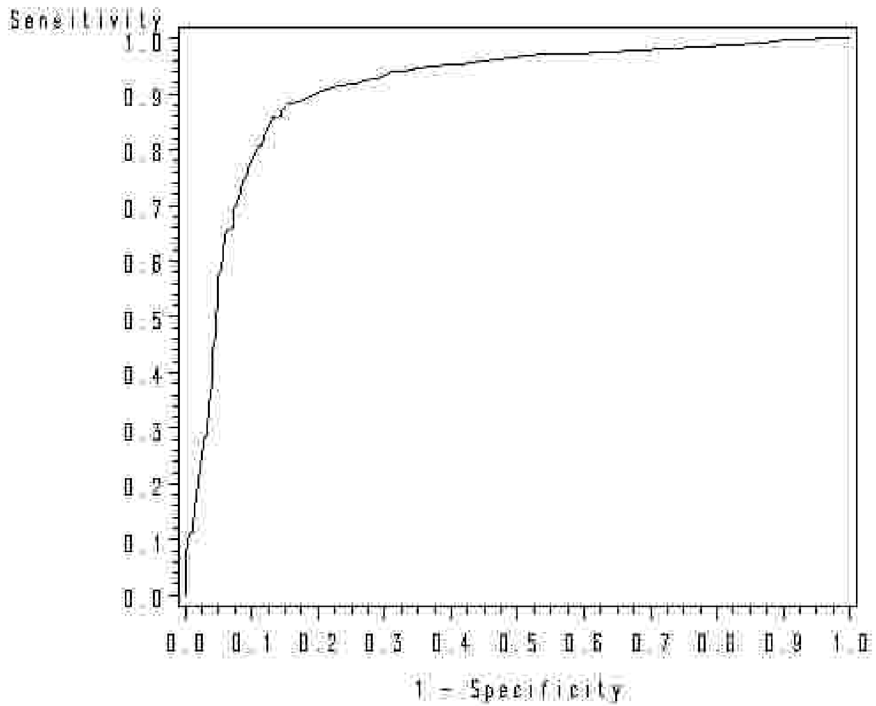


ROC curve and AUC

- AUC : area under curve
  - Probability of concordance P(X1>X2)

$$AUC = \int_{s=+\infty}^{s=-\infty} (1-\beta(s))d\alpha(s)$$

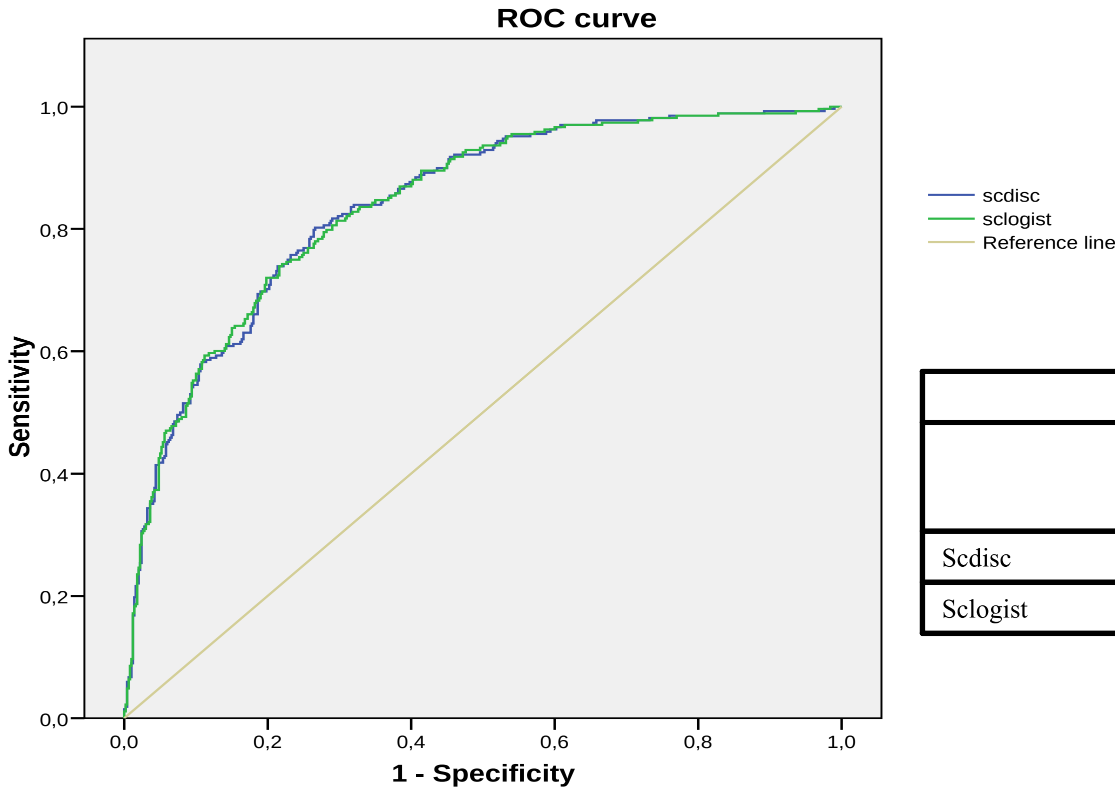
  - Estimated by the proportion of concordant pairs among $n_1 n_2$
  - Related to Mann-Whitney's U statistic : AUC = U/n1n2



Model choice through AUC

- As long as there is no crossing: the best model is the one with the largest AUC or G.
  - No need of nested models
- But comparing models on the basis of the learning sample may be misleading since the comparison will be generally in favour of the more complex model.
- Comparison should be done on hold-out (independent) data to prevent overfitting

Performance comparisons



	AUC	Std Err.	Asymptotic confidence Interval 95%	
			Lower bound	Upper bound
Scdisc	0.839	0.015	0.810	0.868
Sclogist	0.839	0.015	0.811	0.868

4.Statistical models

- About statistical models
  - Unsupervised case: a representation of a probabilisable real world: X r.v. $\in$ parametric family $f(x;\theta)$
  - Supervised case: response $Y=\Phi(X)+\varepsilon$
- Different goals
  - Unsupervised: good fit with parsimony
  - Supervised: accurate predictions

4.1. Model choice and penalized likelihood

- The likelihood principle (Fisher, 1920)

sample of n iid observations:

$$L(x_1, .., x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

The best model is the one which maximizes the likelihood, ie the probability of having observed the data. ML estimation etc.

Overfitting risk

- Likelihood increases with the number of parameters..
  - Variable selection: a particular case of model selection

Need for parsimony

- Occam's razor

William of Occham
(1285-1348)

from wikipedia

An English Franciscan friar and scholastic philosopher. He was summoned to Avignon in 1324 by Pope John XXII on accusation of heresy, and spent four years there in effect under house arrest.

William of Ockham has inspired in U.Eco's The Name of the Rose, the monastic detective William of Baskerville, who uses logic in a similar manner.

Occam's razor states that the explanation of any phenomenon should make as few assumptions as possible, eliminating, or "shaving off", those that make no difference in the observable predictions of the explanatory hypothesis or theory.

lex parsimoniae :

entia non sunt multiplicanda praeter necessitatem,

or:

entities should not be multiplied beyond necessity.

penalized likelihood

Nested (?) family of parametric models, with k parameters: trade-off between the fit and the complexity

Akaïke :

- $AIC = -2 \ln(L) + 2k$

Schwartz :

- $BIC = -2 \ln(L) + k \ln(n)$

- Choose the model which minimizes AIC or BIC

4.2 AIC and BIC: different theories

- AIC : approximation of Kullback-Leibler divergence between the true model and the best choice inside the family

$$I(f;g)=\int f(t)\ln\frac{f(t)}{g(t)}dt = E_f(\ln(f(t))-E_f(\ln(g(t))$$

$$E_{\hat{\theta}}E_f(\ln(g(t;\hat{\theta}))\sim \ln(L(\hat{\theta}))-k$$

AIC and BIC: different theories

- BIC : bayesian choice between m models M_i . For each model $P(\theta_i / M_i)$. The posterior probability of M_i knowing the data **x** is proportional to P(Mi) P(**x**/Mi). With equal priors $P(M_i)$:

$$\ln(P(\mathbf{x}/M_i) \sim \ln(P(\mathbf{x}/\hat{\theta}_i, M_i) - \frac{k}{2}\ln(n)$$

- The most probable model *Mi a posteriori* is the one with minimal *BIC*.

AIC and BIC: different uses

- BIC favourises more parsimonious models than AIC due to its penalization
- AIC (not BIC) is biased : if the true model belongs to the family *Mi* , the probability that AIC chooses the true model does not tend to one when the number of observations goes to infinity.
- It is inconsistent to use AIC and BIC simultaneously
- Other penalisations such as $AIC3 = -2\ln\left(L(\hat{\theta})\right) + 3k$ theory?

4.3 Limitations

- Refers to a “true” which generally does not exist, especially if n tends to infinity. “Essentially, all models are wrong, but some are useful ” G.Box (1987)
- Penalized likelihood cannot be computed for many models:
  - Decision trees, neural networks, ridge and PLS regression etc.
  - No likelihood, which number of parameters?

5. Predictive modelling

- In Data Mining applications (CRM, credit scoring etc.) models are used to make predictions.
- Model efficiency: capacity to make good predictions and not only to fit to the data (forecasting instead of backforecasting: in other words it is the future and not the past which has to be predicted).

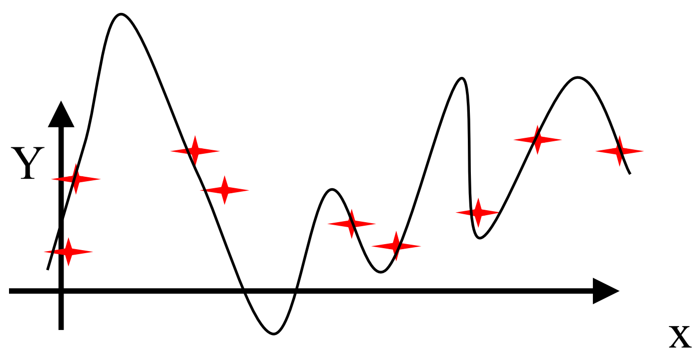
- **Classical framework**
  - Underlying theory
  - Narrow set of models
  - Focus on parameter estimation and goodness of fit
  - Error: white noise
- **Data mining context**
  - Models come from data
  - Algorithmic models
  - Focus on control of generalization error
  - Error: minimal

The black-box problem and supervised learning (N.Wiener, V.Vapnik)

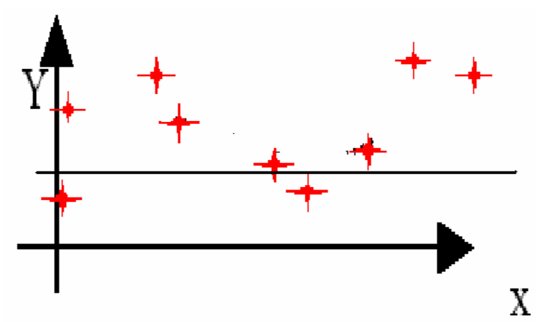
- Given an input x, a non-deterministic system gives a variable $y = f(x)+e$. From n pairs (x_i,y_i) one looks for a function which approximates the unknown function f.
- Two conceptions:
  - A good approximation is a function close to f
  - A good approximation is a function which has an error rate close to the black box, ie which performs as well

5.1 Model choice and Statistical Learning Theory

- How to choose a model in a family of models (eg: degree of a polynomial regression)?



A too complex model: too good fit

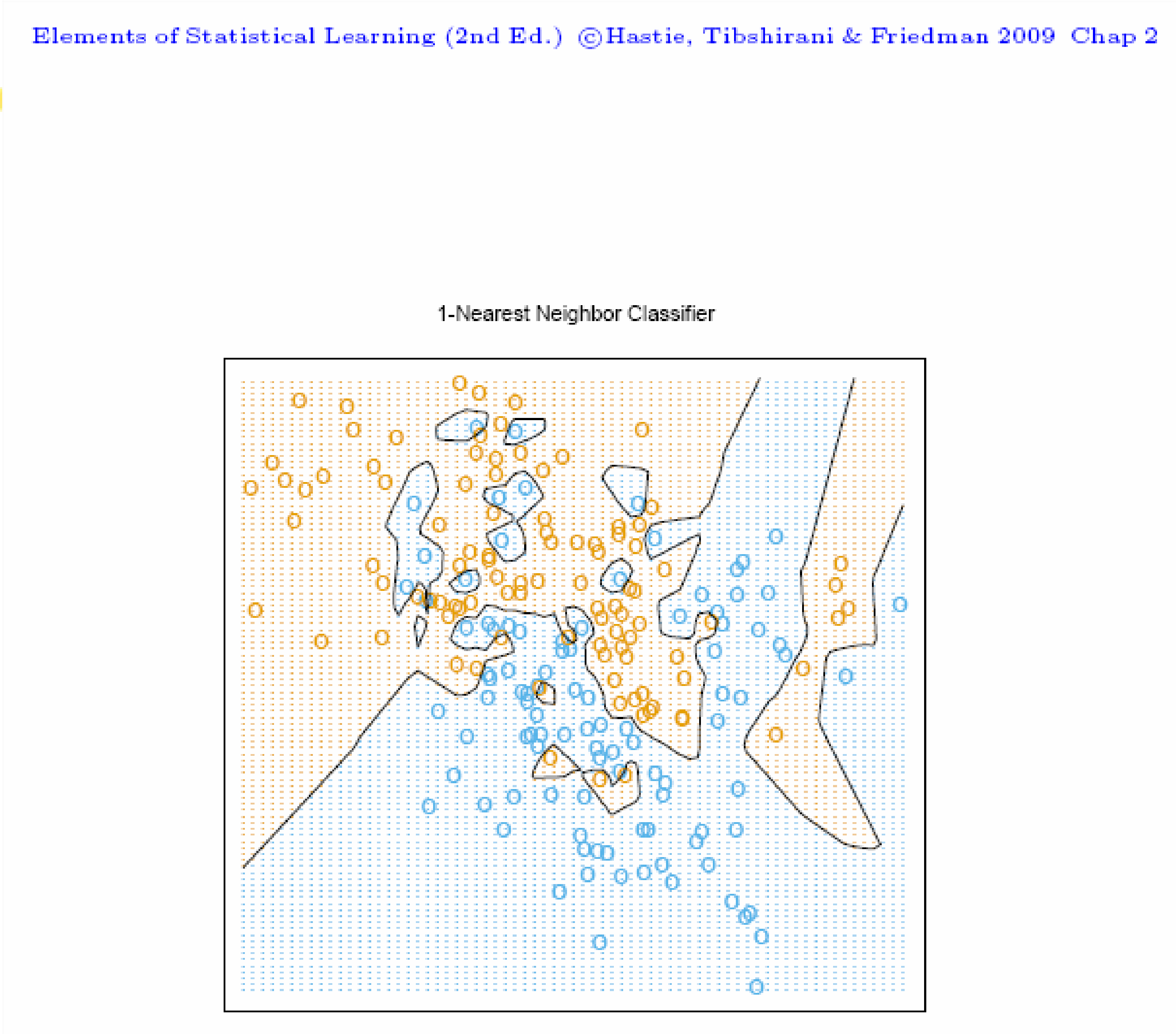


A too simple (but robust) model: bad fit

K-nearest neighbours

- Infarctus data set

K-nearest neighbours



5.2 Model complexity and prediction error

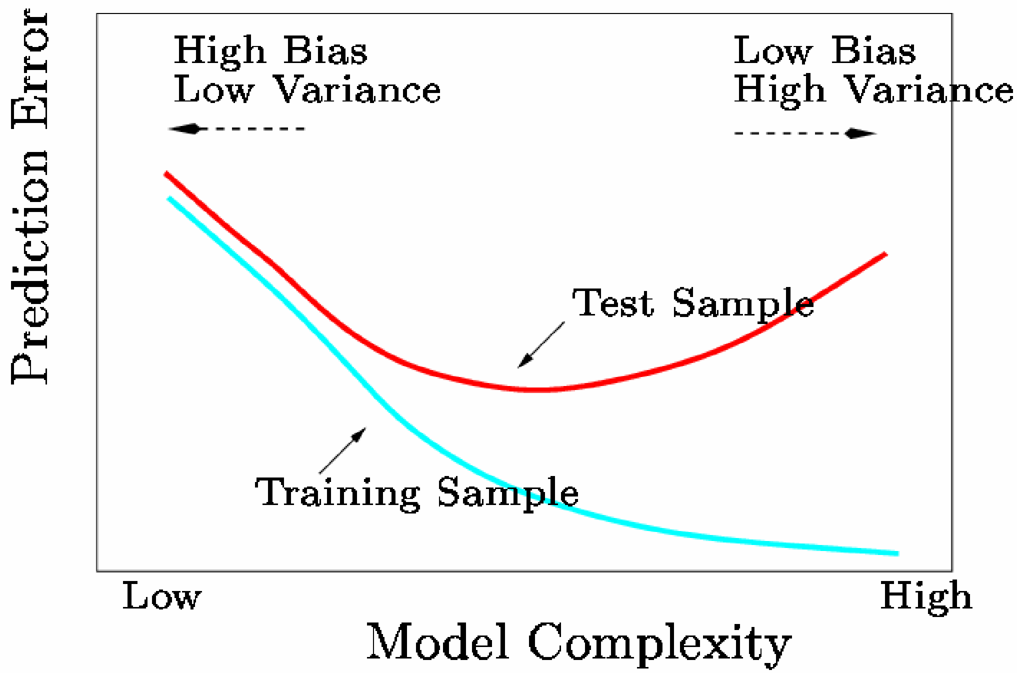


Figure 2.11: *Test and training error as a function of model complexity.*

Model complexity

- The more complex a model, the better the fit but with a high prediction variance.
- Optimal choice: trade-off
- But how can we measure the complexity of a model?

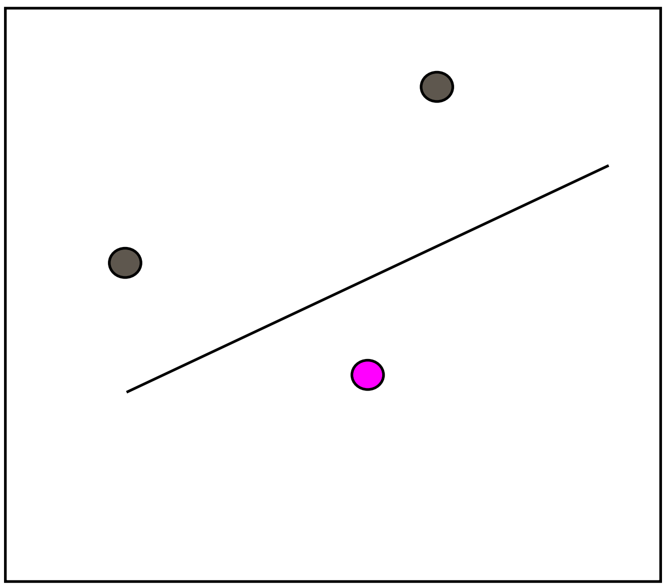
5.3 Vapnik-Cervonenkis dimension for binary supervised classification

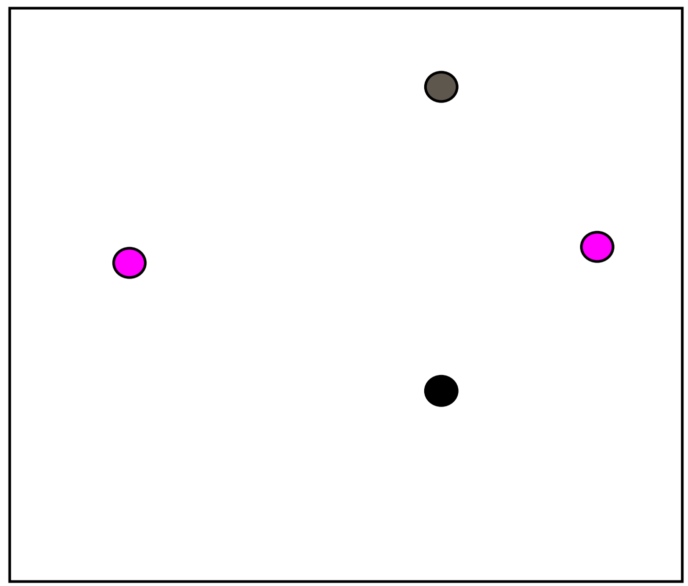
- A measure of complexity related to the separating capacity of a family of classifiers.

- Maximum number of points which can be separated by the family of functions whatever are their labels ± 1

Example

- In 2-D, the VC dimension of "free" linear classifiers is 3 (in p-D VCdim=p+1)

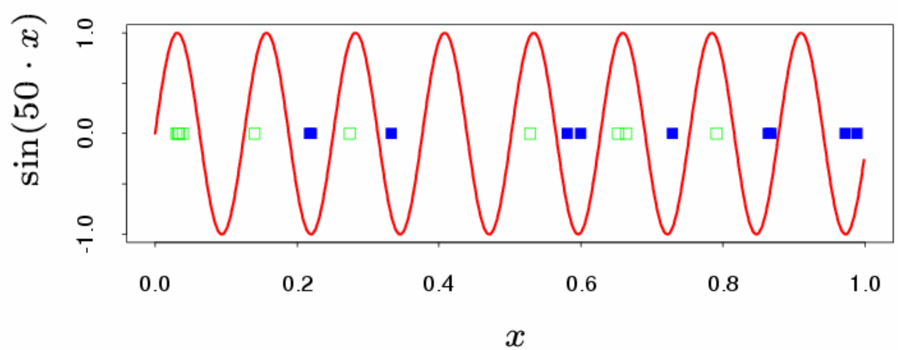




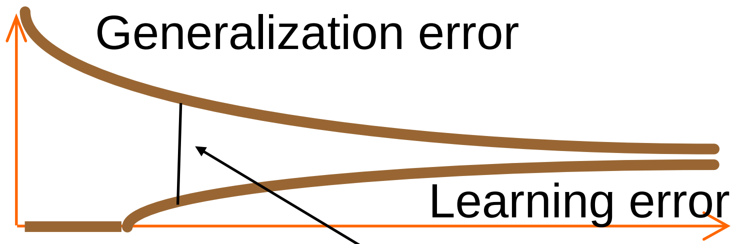
- **But VC dimension is NOT equal to the number of free parameters: can be more or less**

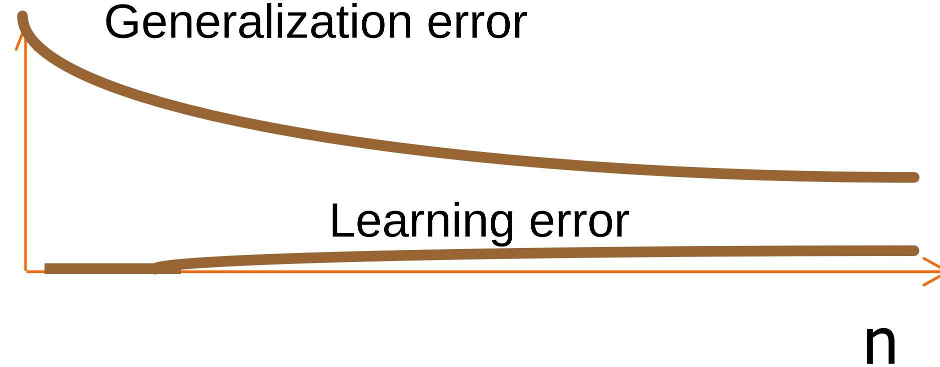
  - The VC dimension of $f(x,w) = sign\ (sin\ (w.x)\ )$
    $c < x < 1,\ c>0,$
    with only one parameter w is infinite.

Hastie et al. 2001



- Consistent learning
- Non consistent learning





h must be finite

Vapnik's inequality $$R < R_{\text{emp}} + \sqrt{\frac{h\left(\ln\left(2n/h\right)+1\right)-\ln\left(\alpha/4\right)}{n}}$$

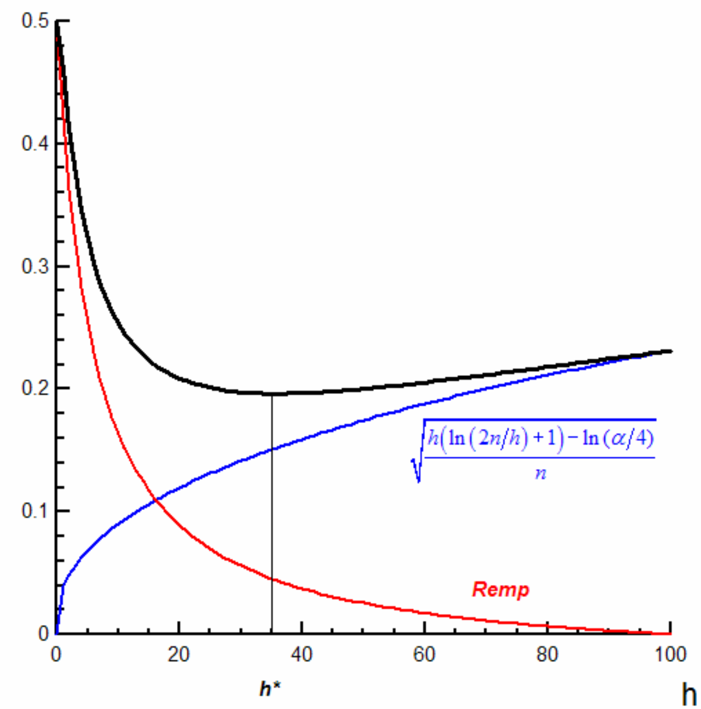
5.4 Model choice by Structural Risk Minimization (SRM)

- Vapnik's inequality:

$$R < R_{\text{emp}} + \sqrt{\frac{h\left(\ln\left(2n/h\right)+1\right)-\ln\left(\alpha/4\right)}{n}}$$

- Comments:
  - the complexity of a family of models may increase when n increases, provided *h* is finite
  - Small values of *h* gives a small difference between *R* and *Remp* . It explains why regularized (ridge) regression, as well as dimension reduction techniques, provide better results in generalisation than ordinary least squares.

- With SRM, instead of minimizing R, one minimizes the upper bound: R_{emp} + confidence interval.
- For any distribution , SRM provides the best solution with probability 1 (universally strong consistency) Devroye (1996) Vapnik (2006).



5.5 High dimensional problems and regularization

- Many ill-posed problems in applications (eg genomics) where p>>n
- In statistics (LS estimation) Tikhonov regularization = ridge regression; a constrained solution of Af= F under Ω(f)$\leq$c (convex and compact set)

$$\min \left(\|Af - F\|^2 + \gamma\Omega(f) \right)$$

- Other techniques: projection onto a low dimensional subspace: principal components (PCR), partial least squares regression (PLS), support vector machines (SVM)

Ridge regression

- the VC dimension of $f(X,w) = sign\left(\sum_{i=1}^{p}(w_i x_i) + 1\right)$

subject to: $\|W\|^2 = \sum_{i=1}^{p} w_i^2 \leq \dfrac{1}{C}$

may be far lower than *p*+1:

$$h \leq \min\left[int\left(\frac{R^2}{C^2}\right); p\right] + 1 \qquad \|X\| \leq R$$

- Since Vapnik's inequality is an universal one, the upper bound may be too large.
- Exact VC-dimension are very difficult to obtain, and in the best case, one only knows bounds
- But even if the previous inequality is not directly applicable, SRM theory proved that the complexity differs from the number of parameters, and gives a way to handle methods where penalized likelihood is not applicable.

5.6 Empirical model choice

- **The 3 samples procedure** (Hastie & al., 2001)
  - Learning set: estimates model parameters
  - Test : selection of the best model
  - Validation : estimates the performance for future data
- **Resample** (eg: ‘bootstrap, 10-fold CV, …)
- **Final model : with all available data**
  - **Estimating model performance is different from estimating the model**

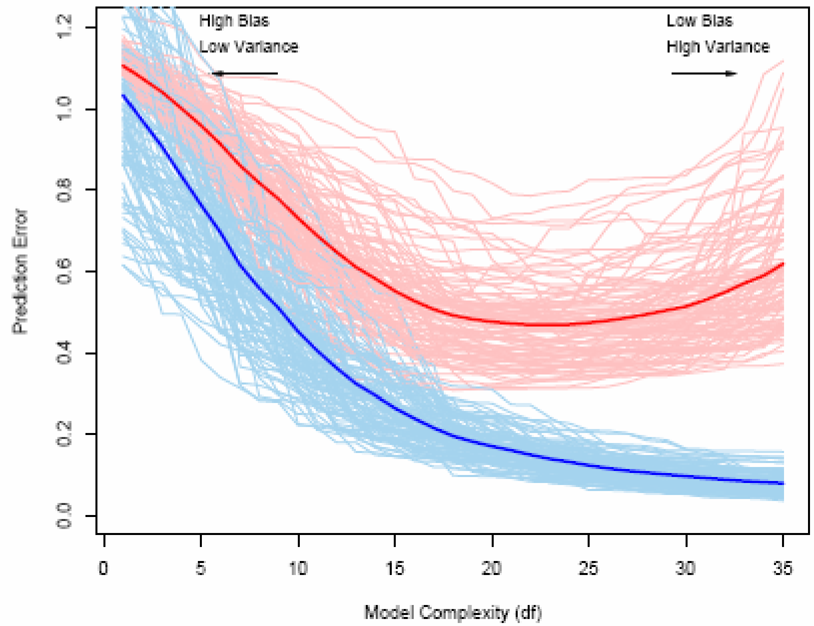
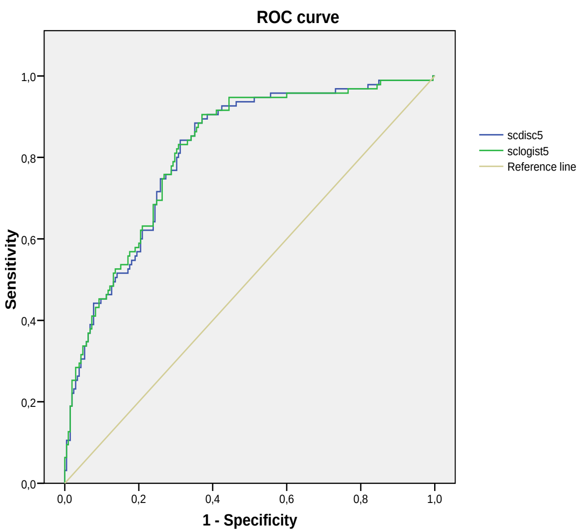
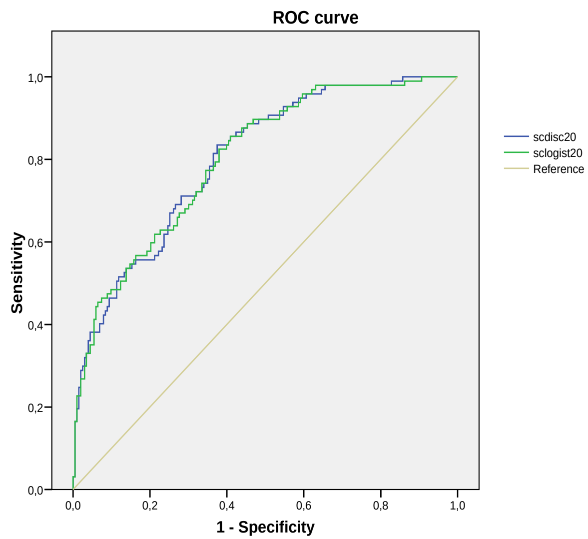


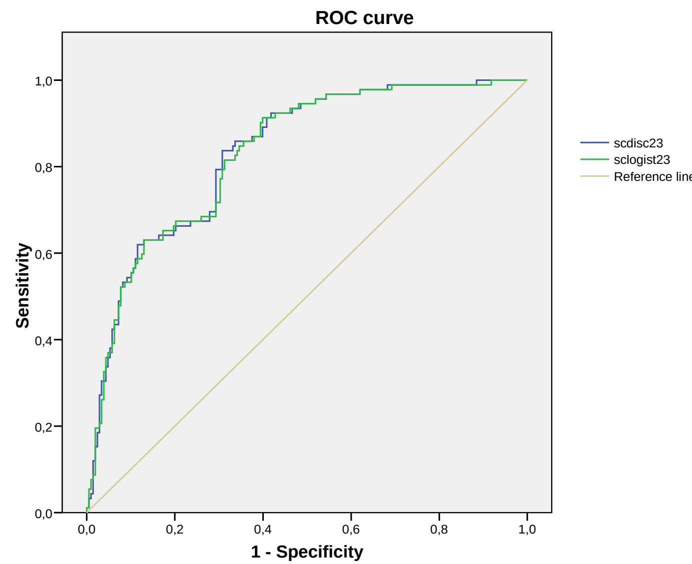
FIGURE 7.1. *Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error* $\overline{\mathrm{err}}$, *while the light red curves show the conditional test error* $\mathrm{Err}_{\mathcal{T}}$ *for* 100 *training sets of size* 50 *each, as the model complexity is increased. The solid curves show the expected test error* Err *and the expected training error* $\mathrm{E}[\overline{\mathrm{err}}]$.

Variability

Training set 70%, validation set 30%, 30 times







- Linear discriminant analysis performs as well as logistic regression
- AUC has a small (due to a large sample) but non neglectable variability
- Large variability in subset selection (Saporta, Niang, 2006)

6 . Discussion

- **Models of data $\neq$ models for prediction**
- Models in Data Mining: no longer a (parsimonious) representation of real world coming from a scientific theory but merely a «blind» prediction technique.
- Penalized likelihood is intellectually appealing but of no help for complex models where parameters are constrained.
- **Statistical Learning Theory provides the concepts for supervised learning in a DM context: avoids overfitting and false discovery risk.**

- One should use adequate and objective performance measures and not “ideology” to choose between models: eg AUC for binary classification
- Empirical comparisons need resampling but assume that future data will be drawn from the same distribution: uncorrect when there are changes in the population
- New challenges:
  - Data streams
  - Complex data

References

- Académie des Sciences (2000): Rapport sur la science et la technologie n°8, *La statistique*,
- J.Friedman (1997) : *Data Mining and statistics, what's the connection?* http://www-stat.stanford.edu/~jhf/ftp/dm-stat.ps
- Giudici, P. (2003) *Applied Data Mining*, Wiley
- Hastie, T., Tibshirani, F., Friedman J. (2009) *Elements of Statistical Learning*, 2nd edition, Springer
- Tufféry, S. (2007) *Data Mining et Statistique Décisionnelle*, Technip