

Introduction au Data Mining et à l'apprentissage statistique



Gilbert Saporta

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

gilbert.saporta@cnam.fr

<http://cedric.cnam.fr/~saporta>

Plan



1. Qu'est-ce que le data mining?
2. Trois méthodes emblématiques
 - 2.1 Règles d'associations
 - 2.2 Arbres de décision
 - 2.3 Scoring
3. Performance des méthodes de prévision
4. Construction et choix de modèles: théorie de l'apprentissage
5. Le DM, une nouvelle conception de la statistique et du rôle des modèles

1. Qu'est-ce que le data mining?

- Le Data Mining est un nouveau champ situé au croisement de la statistique et des technologies de l'information (bases de données, intelligence artificielle, apprentissage etc.) dont le but est de découvrir des structures dans de vastes ensembles de données.
 - Deux types: **modèles** et « **patterns** » (ou **comportements**)
(*D.Hand*)

1.1 Définitions:



- 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"*

- La métaphore du Data Mining signifie qu'il y a des trésors ou **pépites** cachés sous des montagnes de données que l'on peut découvrir avec des outils spécialisés.
- Le Data Mining analyse des données recueillies à d'autres fins: c'est *une analyse secondaire* de bases de données, souvent conçues pour la gestion de données individuelles (Kardaun, T.Alanko,1998)
- Le Data Mining ne se préoccupe donc pas de collecter des données de manière efficace (sondages, plans d'expériences) (Hand, 2000)

The First International Conference on **Knowledge Discovery and Data Mining** KDD-95

Conference Co-Chairs:
Usama M. Fayyad, Jet Propulsion Laboratory/California Institute of Technology
Ramasamy Uthrusamy, GM Research Laboratories

Program Committee:
R. Agrawal (IBM, USA)
T. Anand (AT&T, USA)
R. Brachman (AT&T Bell Labs, USA)
W. Burdine (NASA Ames, USA)
N. Cercone (University of Regina, Canada)
P. Chesman (NASA Ames, USA)
G. Cooper (University of Pittsburgh, USA)
B. Gaines (University of Calgary, Canada)
C. Glymour (Carnegie-Mellon University, USA)
D. Hand (Open University, UK)
D. Heckerman (Microsoft Research, USA)
S.J. Hong (IBM, USA)
L. Jackel (AT&T Bell Labs, USA)
J. Kerschbatter (Georg Meissner University, USA)
W. Kloosgen (GM, Germany)
D. Madigan (University of Washington, USA)
C. Mathieu (GTE Laboratories, USA)
H. Mannila (University of Helsinki, Finland)
G. Piatetsky-Shapiro (GTE Labs, USA)
D. Pregibon (AT&T Bell Labs, USA)
A. Siebes (CW-Netherlands)
E. Simoudis (IBM, USA)
A. Sison (University of Waterloo, Canada)
P. Smith (Jet Propulsion Laboratory, USA)
A. Steinlin (New York University, USA)
A. Wai (Monash University, Australia)
D. Zeng (University of Regina, Canada)
Z. Zhang (Michigan State University, USA)

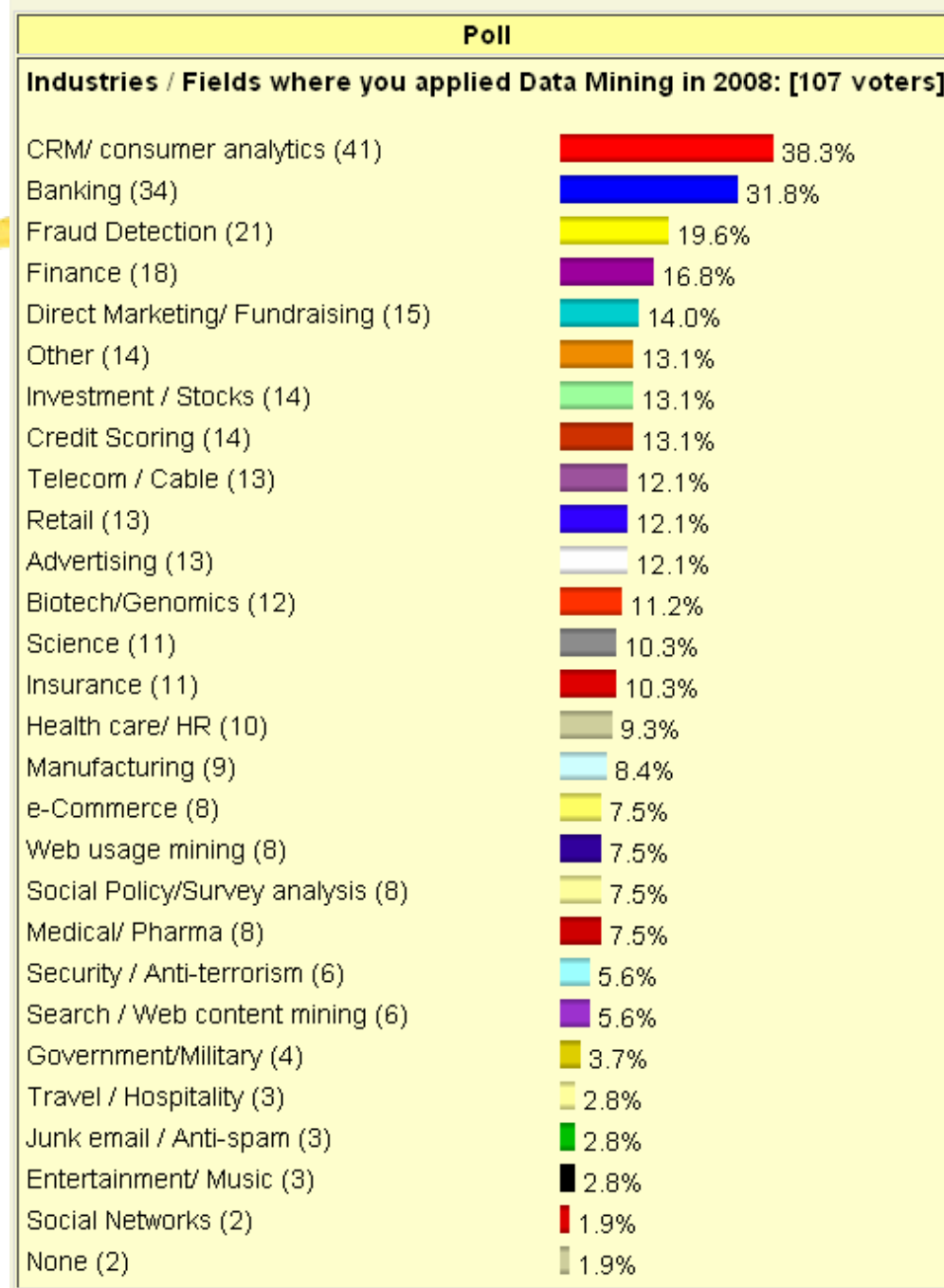
Publication: Proceedings of the Conference on
Knowledge Discovery and Data Mining, KDD-95
Morgan Kaufmann, Chicago, Illinois, USA

Est-ce nouveau? Est-ce une révolution ?

- L'idée de découvrir des faits à partir des données est aussi vieille que la statistique *"Statistics is the science of learning from data. Statistics is essential for the proper running of government, central to decision making in industry, and a core component of modern educational curricula at all levels"* (J.Kettenring, 1997, ancien président de l'ASA).
- Dans les années 60: Analyse Exploratoire (Tukey, Benzécri) « *L'analyse des données est un outil pour dégager de la gangue des données le pur diamant de la véridique nature.* » (J.P.Benzécri 1973)

1.2 le Data Mining est né de :

- L'évolution des SGBD vers l'informatique décisionnelle avec les entrepôts de données (Data Warehouse).
- La constitution de giga bases de données : transactions de cartes de crédit, appels téléphoniques, factures de supermarchés: terabytes de données recueillies automatiquement.
- Développement de la Gestion de la Relation Client (CRM)
 - Marketing client au lieu de marketing produit
 - Attrition, satisfaction, etc.
- Recherches en Intelligence artificielle, apprentissage, extraction de connaissances



1.3 Objectifs et outils

- Le Data Mining cherche des structures de deux types : **modèles** et **patterns**
- Patterns
 - une structure caractéristique possédée par un petit nombre d'observations: niche de clients à forte valeur, ou au contraire des clients à haut risque
 - Outils: classification, visualisation par réduction de dimension (ACP, AFC etc.) règles d'association.

modèles



- Construire des modèles a toujours été une activité des statisticiens. Un modèle est un résumé global des relations entre variables, permettant de **comprendre** des phénomènes, et d'émettre des **prévisions**. « *Tous les modèles sont faux, certains sont utiles* » (G.Box)

Modèles

- Le DM ne traite pas d'estimation et de tests de modèles préspecifiés, mais de la découverte de modèles à l'aide d'un processus de recherche algorithmique d'exploration de modèles:
 - linéaires ou non,
 - explicites ou implicites: réseaux de neurones, arbres de décision, SVM, régression logistique, réseaux bayésiens....
- Les modèles ne sont pas issus d'une théorie mais de l'exploration des données.

- 
- Autre distinction: **prédictif** (supervisé) ou **exploratoire** (non supervisé)

Des outils ou un process?

- Le DM est souvent présenté comme un ensemble intégré d'outils permettant entre autres de comparer plusieurs techniques sur les mêmes données.
- Mais le DM est bien plus qu'une boîte à outils:



Data mining et KDD

- « Le Data Mining est une étape dans le processus d'extraction des connaissances, qui consiste à appliquer des algorithmes d'analyse des données »

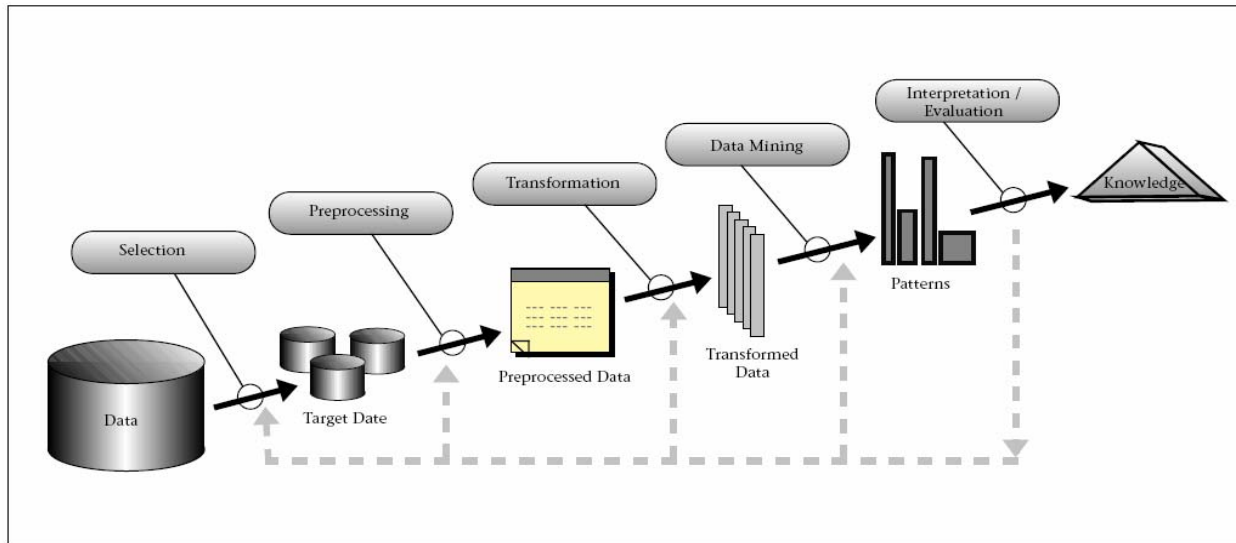



Figure 1. An Overview of the Steps That Compose the KDD Process.

2. Trois techniques emblématiques du Data Mining



- Une méthode non supervisée:
 - Règles d'association

- Deux méthodes supervisées
 - Arbres de décision
 - Scores








2.1 La recherche de règles d'association ou l'analyse du panier de la ménagère

- Illustré avec un exemple industriel provenant de PSA Peugeot-Citroen .
- (Thèse CIFRE de Marie Plasse).

PROBLEMATIQUE INDUSTRIELLE








Les données

→ Plus de 80000 véhicules décrits par plus de 3000 attributs binaires

Véhicules	A1	A2	A3	A4	A5	...	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	1		1
	0	1	0	0	1		0
	0	0	1	0	0		0

Matrice de données binaires

=

Véhicules	Attributs présents
	{A1, A4}
	{A3, A4}
	{A2, A5}
	{A1, A5}
	{A2, A5, Ap}
	{A2, A5}
	{A3}

Données de transaction

- Trouver des corrélations entre les attributs...
- ... grâce à la recherche de règles d'association

LA RECHERCHE DE REGLES D'ASSOCIATION

Rappel de la méthode

→ Origine marketing : analyser les ventes des supermarchés

*"lorsqu'un client achète du pain et du beurre,
il achète 9 fois sur 10 du lait en même temps"*

→ Formalisation : $A \rightarrow C$ où $A \cap C = \emptyset$

→ Fiabilité : **Support** : % de transactions contenant A et C

$$\text{sup}(A \rightarrow C) = P(A \cap C) = P(C / A) \cdot P(A)$$

→ Précision : **Confiance** : % de transactions contenant C sachant qu'elles ont A

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

→ Algorithmes :

- Recherche des **sous-ensembles fréquents** (avec minsup)
- Extraction des **règles d'association** (avec minconf)



$$s(A \rightarrow C) = 30\%$$

⇒ 30% des transactions contiennent à la fois



$$c(A \rightarrow C) = 90\%$$

⇒ 90% des transactions qui contiennent
contiennent aussi



- Apriori (Agrawal & Srikant, 1994)
- Partition (Saverese et al., 1995)
- Sampling (Brin & Motwani, 1997)
- Eclat (Zaki, 2000)
- FP-Growth (Han & Pei, 2003)

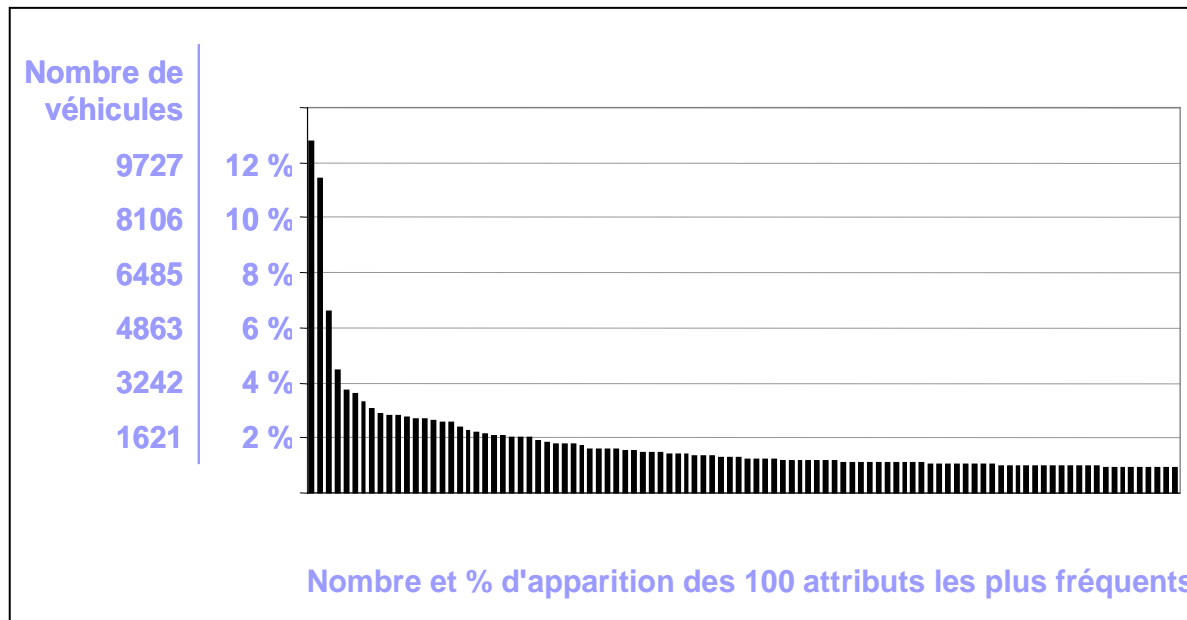
LA RECHERCHE DE REGLES D'ASSOCIATION

Spécificités des données

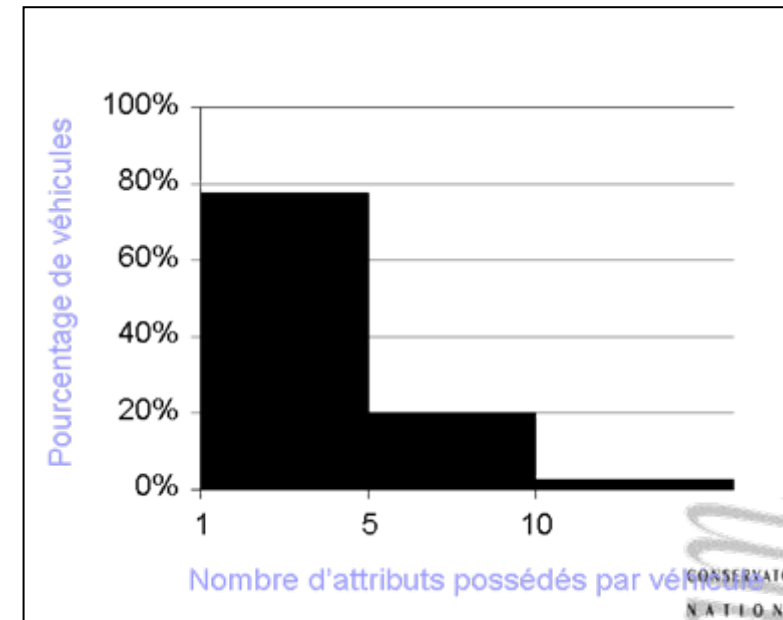
→ 80000 véhicules décrits par 3000 attributs binaires : environ 4 mois de production

→ Des données clairsemées :

→ Répartition des 100 attributs les plus fréquents :



→ Nombre d'attributs présents par véhicule :



→ 4 attributs en moyenne

LA RECHERCHE DE REGLES D'ASSOCIATION

Extraction des règles

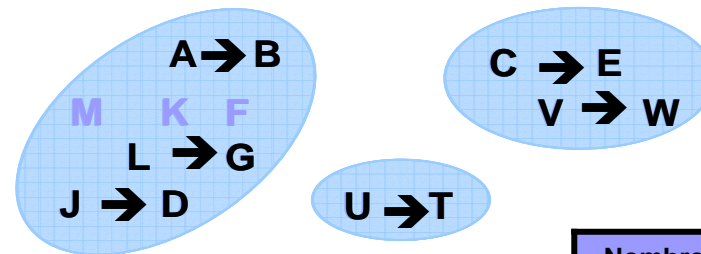
→ Règles extraites :

Support minimum (nb de véhicules vérifiant la règle)	Confiance minimum	Nombre de règles	Taille maximum des règles obtenues
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

→ Réduire le nombre et la complexité des règles tout en gardant une valeur faible pour le support minimum

→ Réalisation d'une classification de variables préalable (Plasse et al., 2005)

→ Recherche des règles à l'intérieur de chaque groupe :



→ Résultats :







	Nombre de règles	Complexité maximum	Réduction du nombre de règles
Sans classification : Rappel premier résultat	1 623 555	13	.
Sans classification : regroupement manuel	600636	12	60%
Avec classification préalable	218	4	99%

LES INDICES DE PERTINENCE

Sélection des "meilleures" règles

- Pour faire valider les règles par un expert du terrain, il faut sélectionner les "meilleures" règles
- On peut les classer par ordre décroissant de leur intérêt statistique
- Il existe plusieurs indices pour évaluer la pertinence des règles
- Un des plus connus et utilisés : **le lift** (Brin et al., 1997)

$$\text{lift}(A \Rightarrow C) = \frac{P(A \cap C)}{P(A) \cdot P(C)}$$

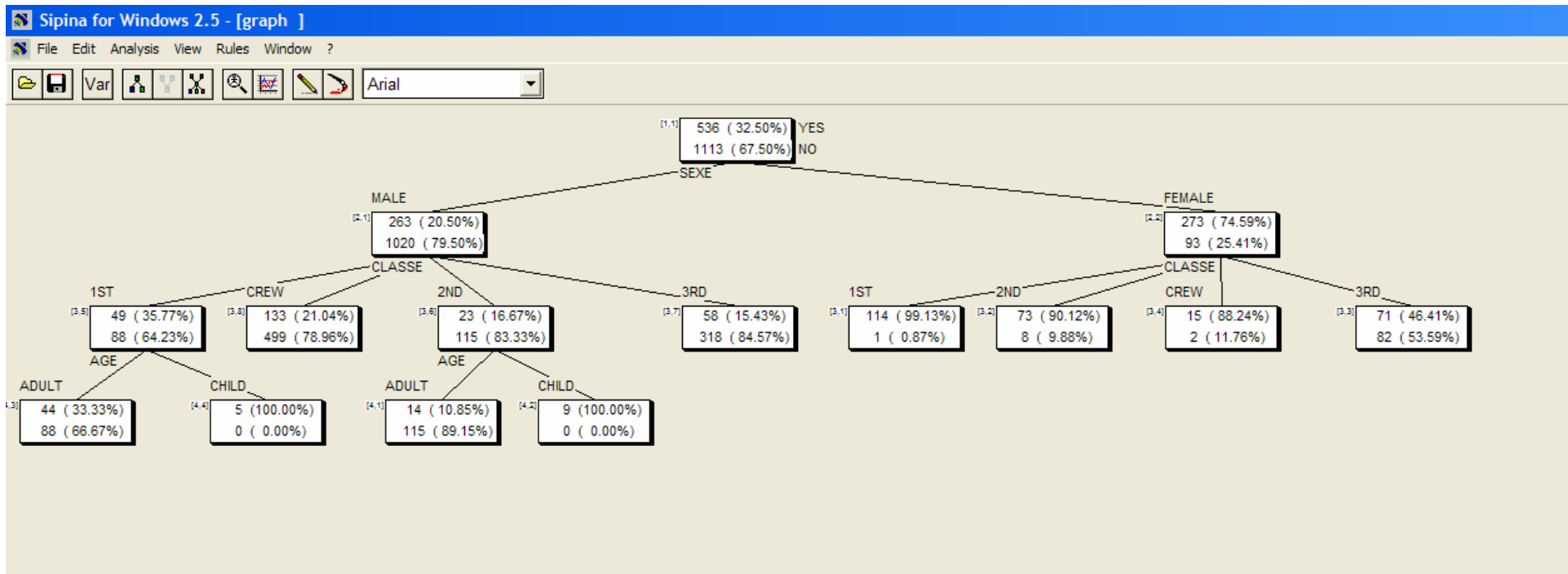
Exemple : $\text{lift} = 2$ → les transactions contenant  +  +  sont 2 fois plus nombreuses que si l'achat de  +  et l'achat de  étaient indépendants

2.2 Arbres de décision

- Développées autour de 1960 (AID de Morgan & Sonquist) et très utilisées en marketing, ces méthodes délaissées par les statisticiens ont connu un regain d'intérêt avec les travaux de Breiman & al. (1984) qui en ont renouvelé la problématique: elles sont devenues un des outils les plus populaires du **data mining** en raison de la lisibilité des résultats. On peut les utiliser pour prédire une variable Y quantitative (arbres de régression) ou qualitative (arbres de décision, de classification, de segmentation) à l'aide de prédicteurs quantitatifs ou qualitatifs. Les termes de **partitionnement récursif** ou de **segmentation** sont parfois utilisés

logiciel gratuit SIPINA

<http://eric.univ-lyon2.fr>



- Résolution des problèmes de discrimination et de régression en divisant successivement l'échantillon en sous-groupes.
- Il s'agit de sélectionner parmi les variables explicatives celle qui est la plus liée à la variable à expliquer. Cette variable fournit une première division de l'échantillon en plusieurs sous-ensembles appelés segments. Puis on réitère cette procédure à l'intérieur de chaque segment en recherchant la deuxième meilleure variable, et ainsi de suite ...
- Il s'agit donc d'une **classification descendante** à but prédictif opérant par sélection de variables : chaque classe doit être la plus homogène possible vis à vis de Y

Arbres binaires ou non?



- En présence d'un prédicteur qualitatif, on pourrait utiliser des arbres non binaires en découpant en m sous ensembles : cette idée n'est en général pas bonne car elle conduit à des subdivisions avec trop peu d'observations et souvent non pertinentes.
- L'intérêt des arbres binaires est de pouvoir regrouper les modalités qui ne se distinguent pas vis à vis de y .

La méthode CART

- La méthode CART permet de construire un arbre de décision binaire par divisions successives de l'échantillon en deux sous-ensembles.
- Il n'y a pas de règle d'arrêt du processus de division des segments : à l'obtention de l'arbre complet, une procédure d'élagage permet de supprimer les branches les moins informatives.
- Au cours de cette phase d'élagage, la méthode sélectionne un sous arbre 'optimal' en se fondant sur un critère d'erreur calculé sur un échantillon test

Divisions d'un nœud (arbres binaires)

- Les divisions possibles dépendent de la nature statistique de la variable :
 - variable binaire $B(0,1)$: une division possible
 - variable nominale N (k modalités) : $2^{k-1} - 1$ divisions possibles
 - variable ordinale O (k modalités) : $k-1$ divisions possibles
 - variable quantitative Q (q valeurs distinctes) : $q-1$ divisions possibles

Discrimination : arrêt des divisions, affectation



- Nœud terminal :
 - s'il est pur ou s'il contient des observations toutes identiques
 - s'il contient trop peu d'observations
- Un segment terminal est affecté à la classe qui est la mieux représentée

Discrimination : T.E.A.



- Représente la proportion d'individus mal classés dans l'ensemble des segments terminaux

Discrimination : Sélection du meilleur sous-arbre

■ Échantillon d'apprentissage :

- Construction de l'arbre complet A_{max} puis élagage : à partir de l'arbre complet, on détermine la séquence optimale de sous-arbres emboîtés $\{A_{max-1}, \dots, A_h, \dots, A_1\}$ avec $1 \leq h < max$
- Le taux d'erreur en apprentissage (TEA) de A_h vérifie :

$$TEA(A_h) = \min_{A \in S_h} \{TEA(A)\}$$

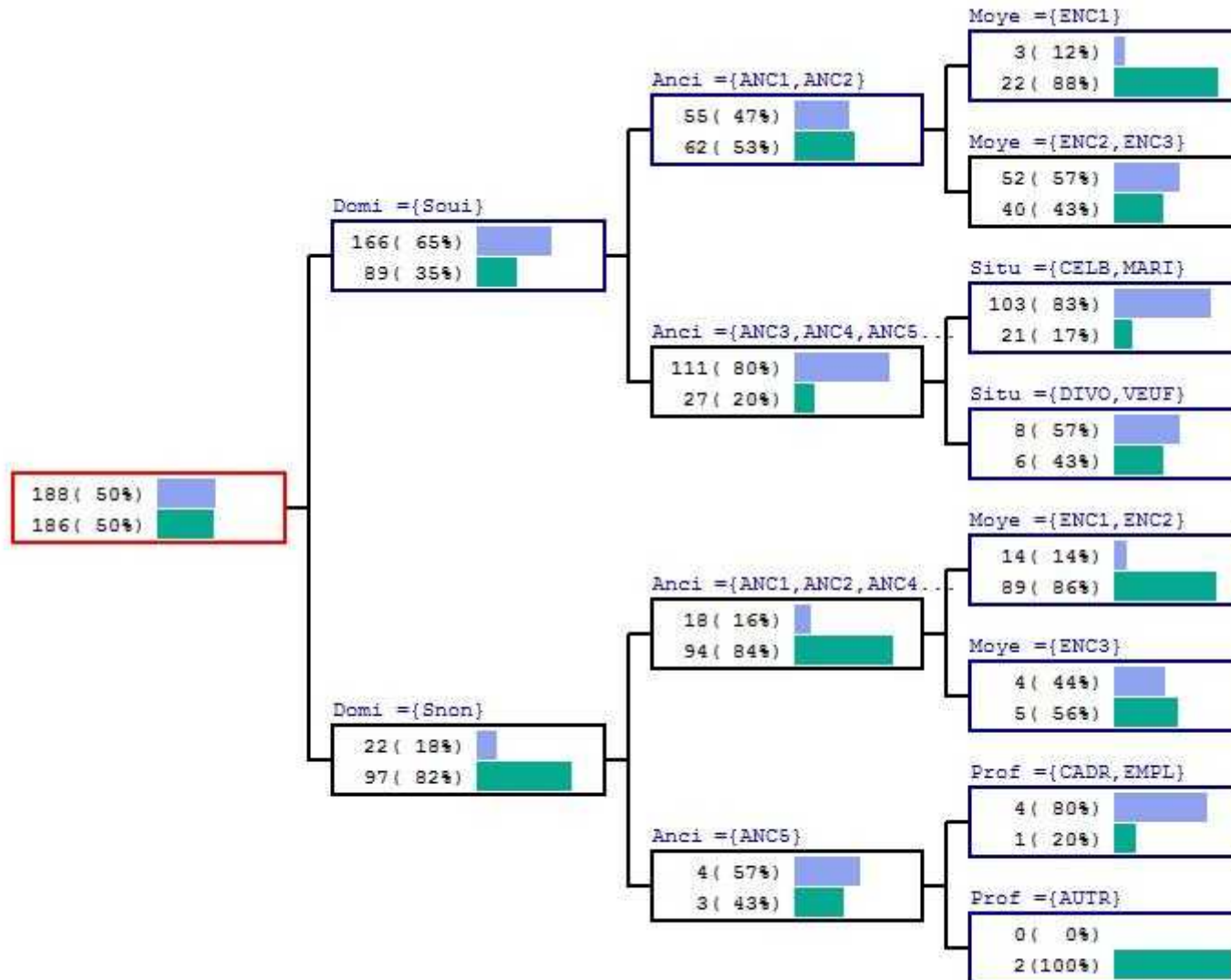
- Où S_h est l'ensemble des sous-arbres de A_{max} ayant h segments terminaux

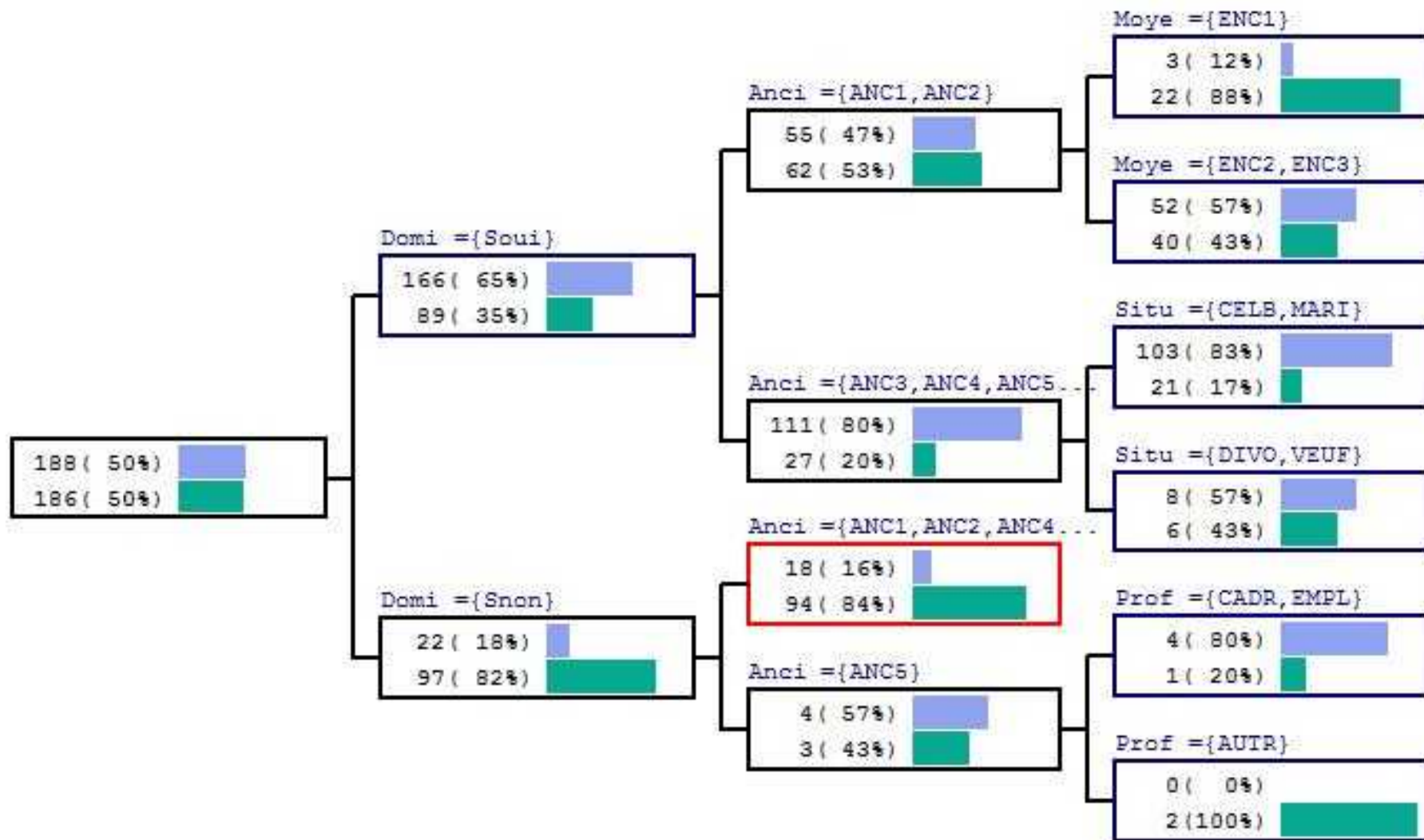
■ Échantillon-test :

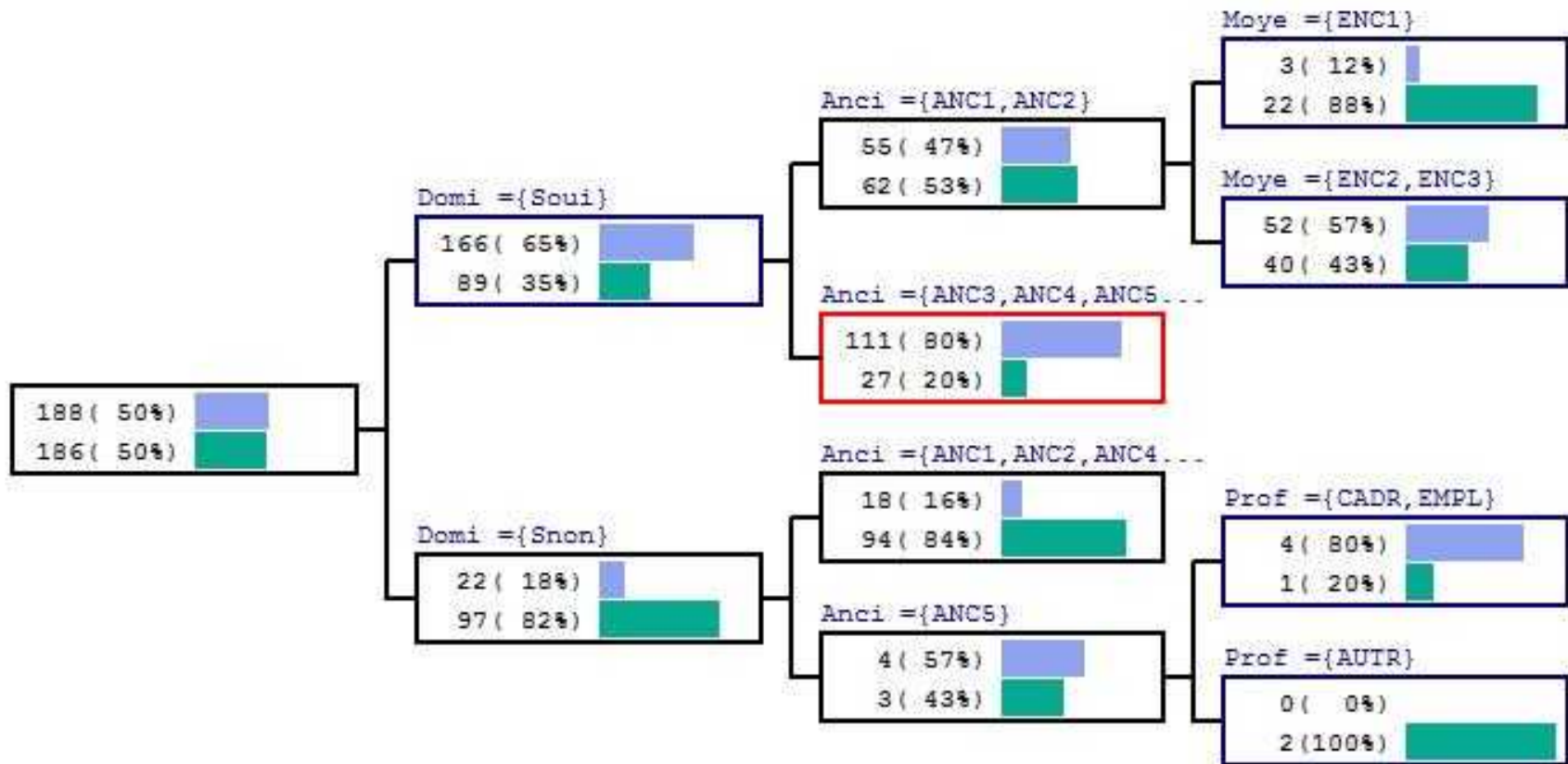
- Choix de A^* tel que l'erreur de classement en test (ETC) vérifie :

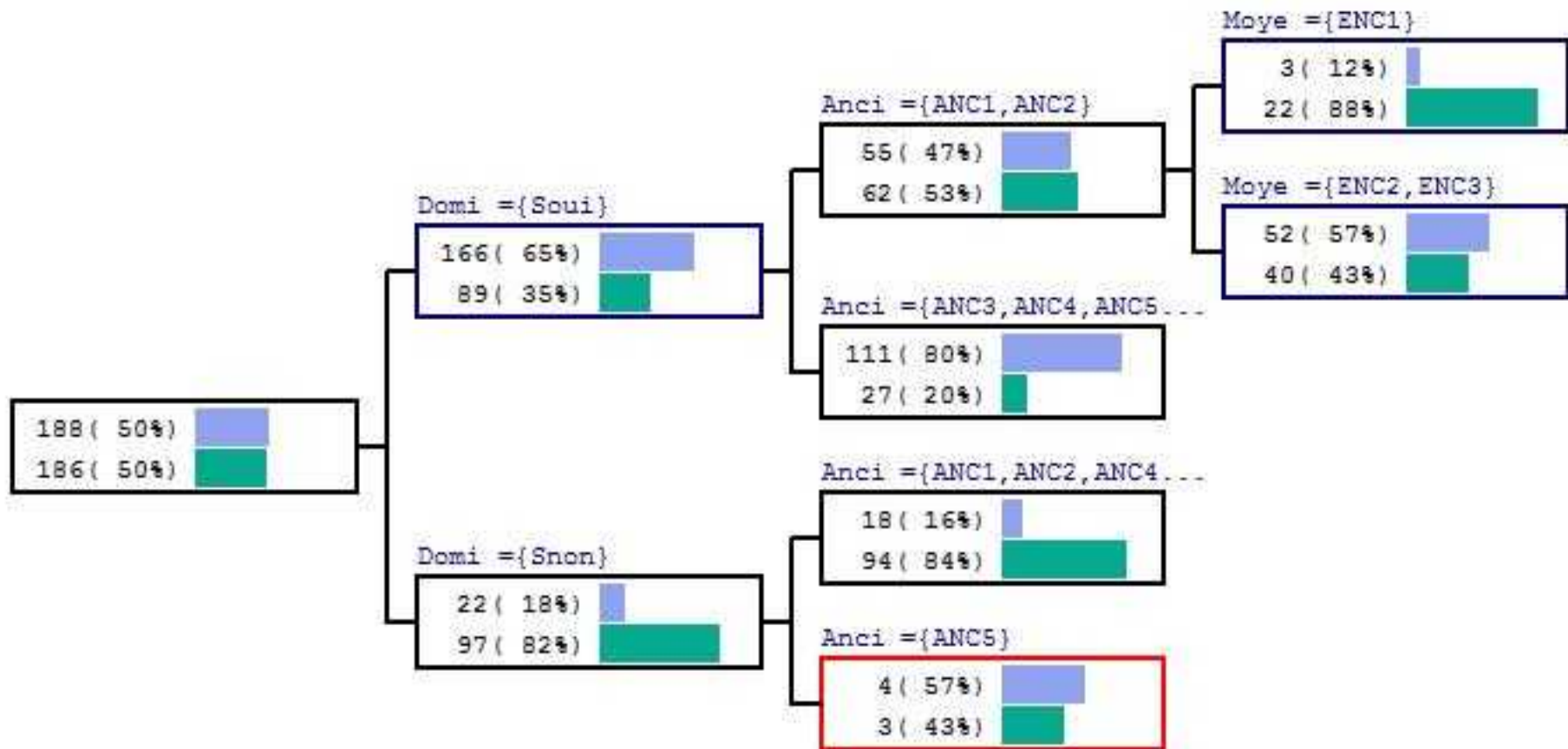
$$ETC(A^*) = \min_{1 \leq h \leq max} \{ETC(A_h)\}$$

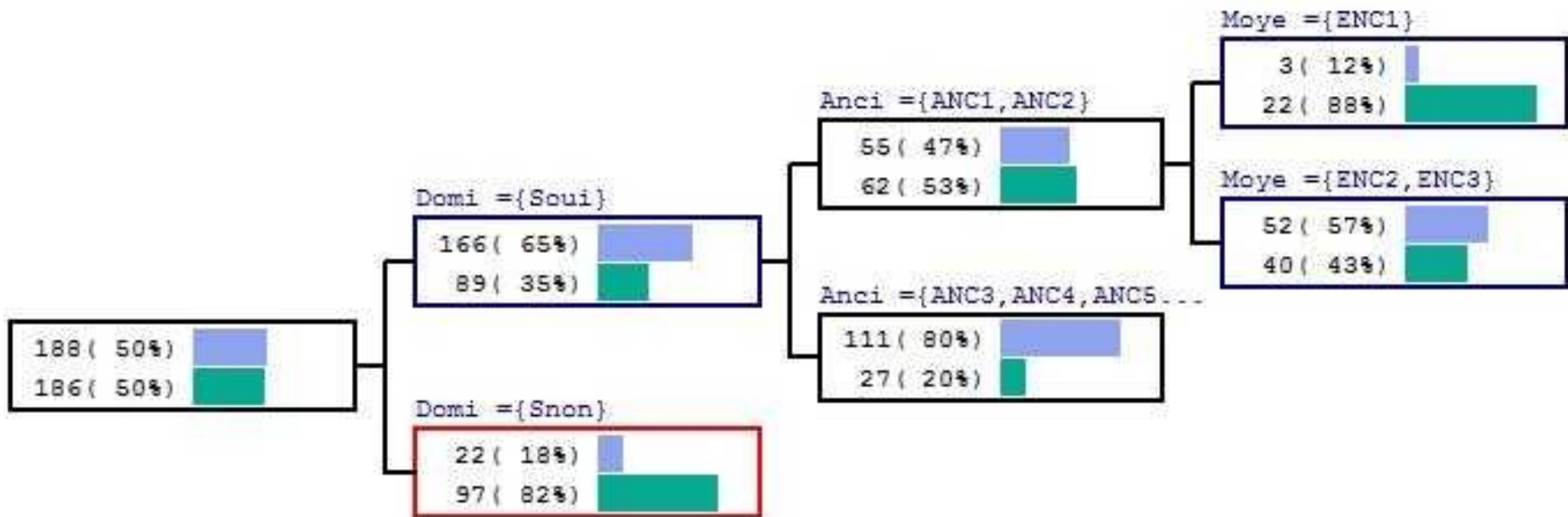
Exemple: bons et mauvais clients d'une banque (SPAD)











Matrice de confusion

OBSERVE	PREDIT	
	BON	MAUV
BON	163	25
MAUV	67	119

Avantages et inconvénients



- Les méthodes de segmentation fournissent une alternative intéressante aux méthodes paramétriques usuelles : elles ne nécessitent pas d'hypothèse sur les données, et les résultats sont plus simples à exploiter
- MAIS : elles fournissent souvent des arbres instables (une division conditionne les suivantes, les branches coupées ne repoussent pas...).

2.3 *Le scoring*

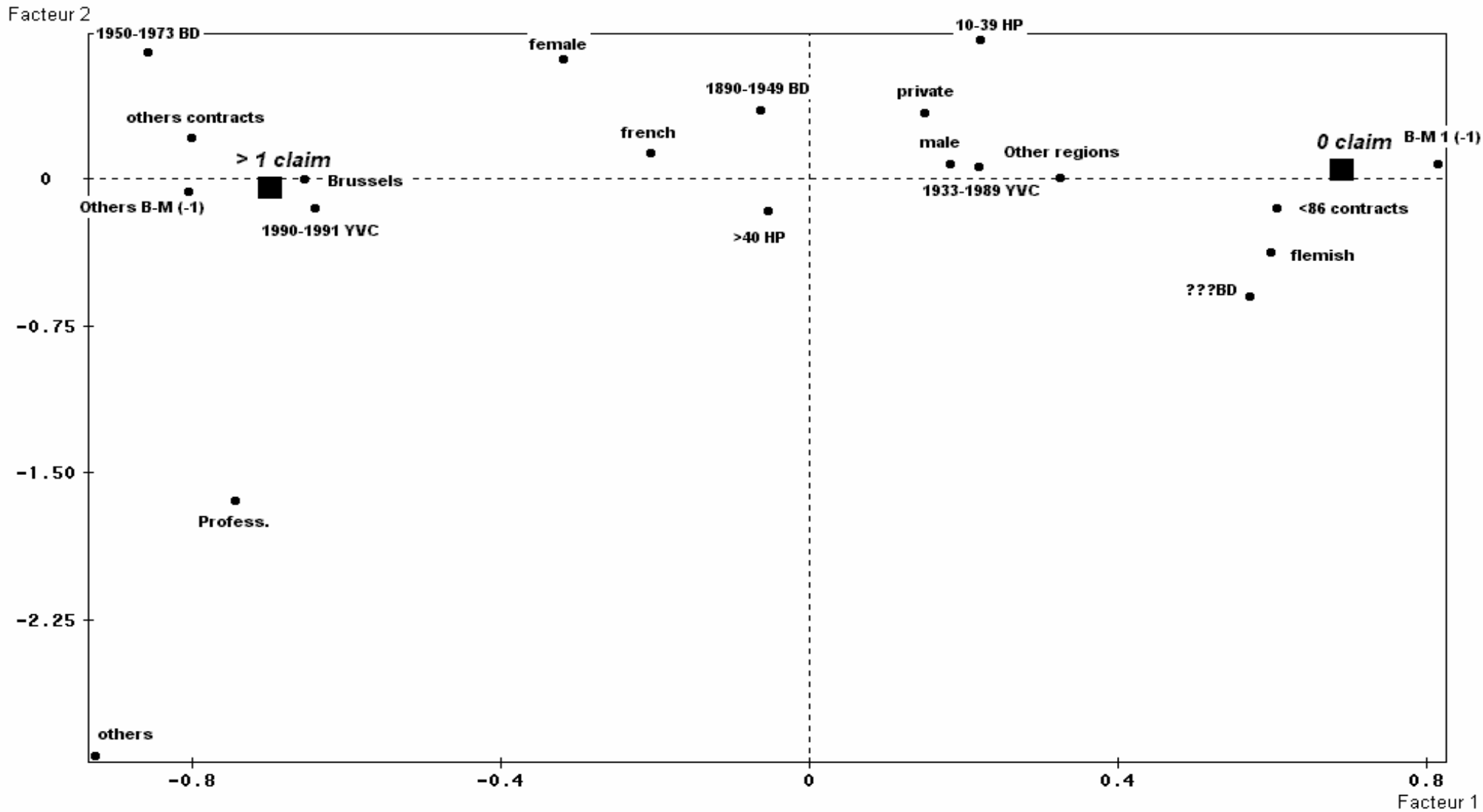


- Prédire une variable à deux modalités :
ex.: qualité d'un client, survie d'un malade etc.
- Construction d'une note de risque (score S)
combinaison des prédicteurs
- Fixation d'un seuil de décision
 - Si $S > s$ on classe dans une modalité, sinon dans l'autre

Exemple assurance (SPAD)

- 1106 contrats automobile belges:
- 2 groupes: « 1 bons », « 2 mauvais »
- 9 prédicteurs: 20 catégories
 - Usage (2), sexe (3), langue (2), age (3), région (2), bonus-malus (2), puissance (2), durée (2), age du véhicule (2)

ACM



ADL de Fisher sur les composantes

FACTEURS	CORRELATIONS	COEFFICIENTS
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
CONSTANTE		0.093575

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

$$\text{Score} = 6.90 F1 - 0.82 F3 + 1.25 F5 + 1.31 F8 - 1.13 F9 - 3.31 F10$$



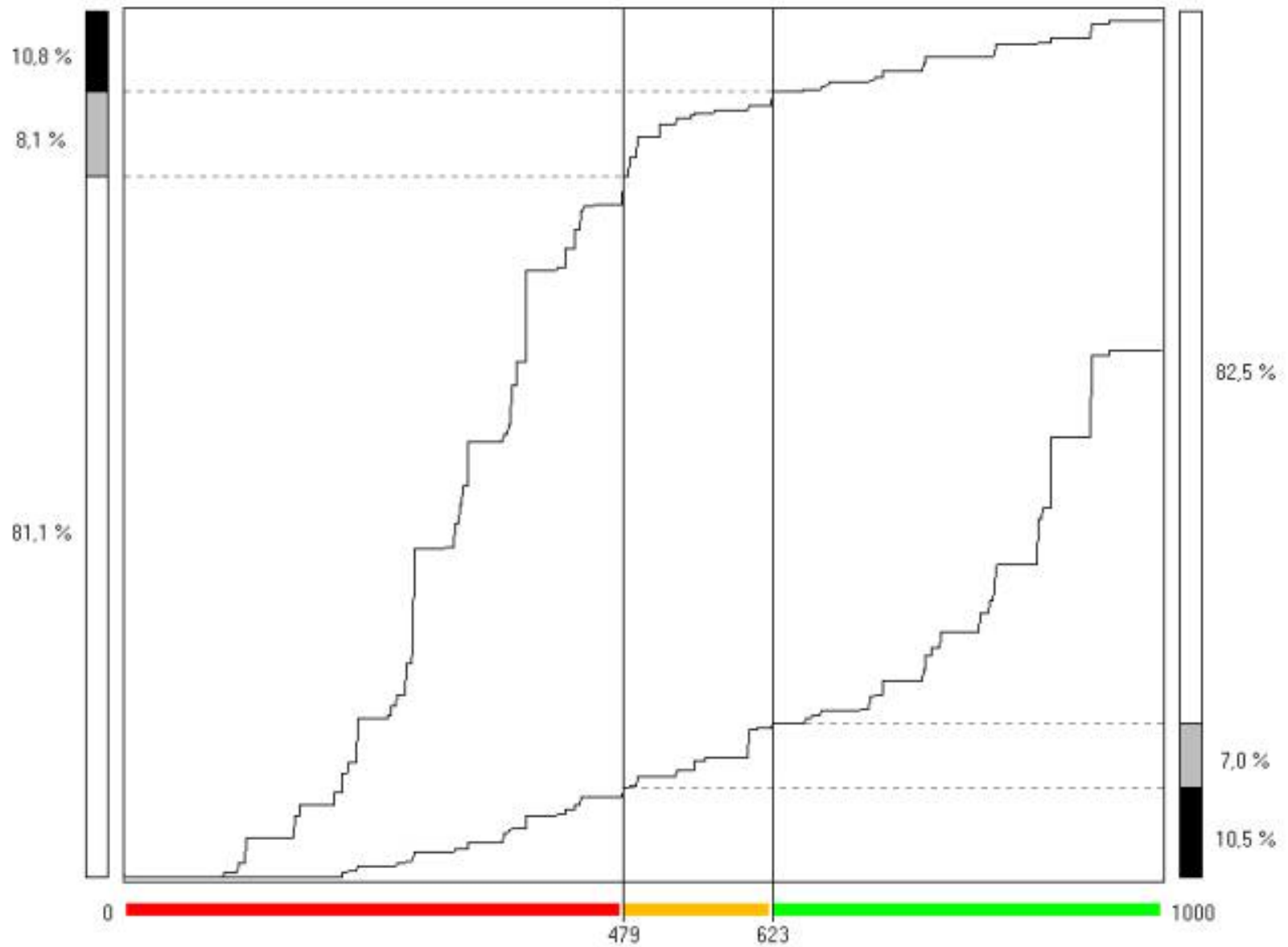
■ scores normalisés

- Echelle de 0 à 1000

- Transformation linéaire du score et du seuil

Grille de score (« 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



Cas des prédicteurs numériques



- Si prédicteurs numériques (taux d'endettement, revenu...)
- Découpage en classes
 - Avantages, détection des liaisons non linéaires

Une autre méthode : régression logistique

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

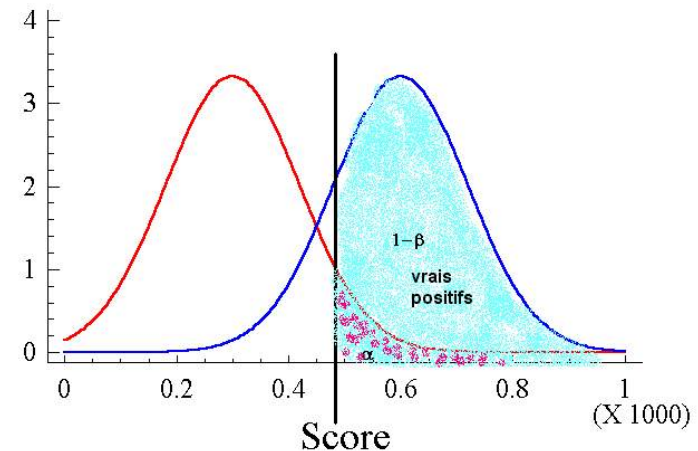
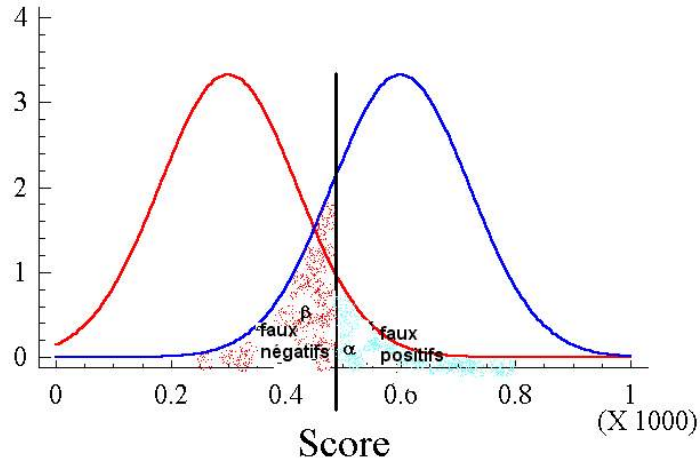
- Estimation directe de la probabilité a posteriori
- Maximum de vraisemblance conditionnel au lieu des moindres carrés.

3. Performance des méthodes de prévision



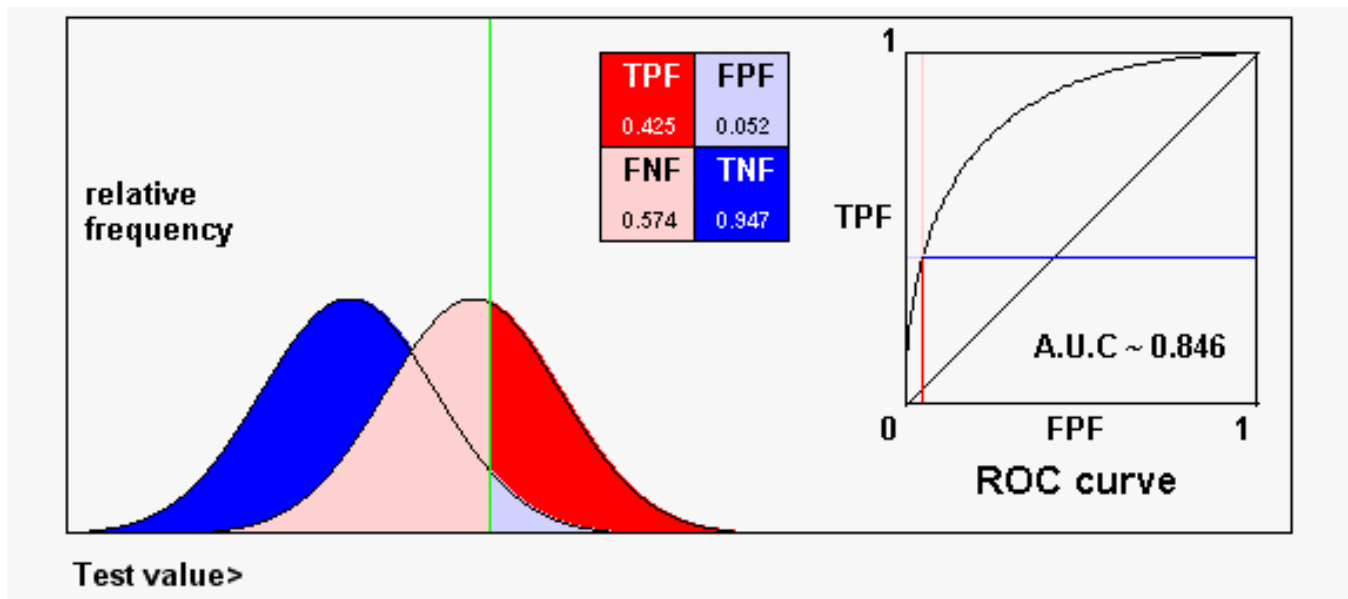
- Mesures de performances en classification binaire supervisée
 - le taux d'erreur suppose le choix d'un seuil
 - le score est plus riche. L'utilisateur choisit son seuil
 - une proba d'appartenance $P(G1/x)$ est aussi un score mais compris entre 0 et 1: à peu près toutes les méthodes fournissent un score

- Groupe à détecter G_1 : scores élevés
- Sensibilité $1-\beta = P(S > s / G_1)$: % de vrais positifs
- Spécificité $1-\alpha = P(S < s / G_2)$: % de vrais négatifs



courbe ROC

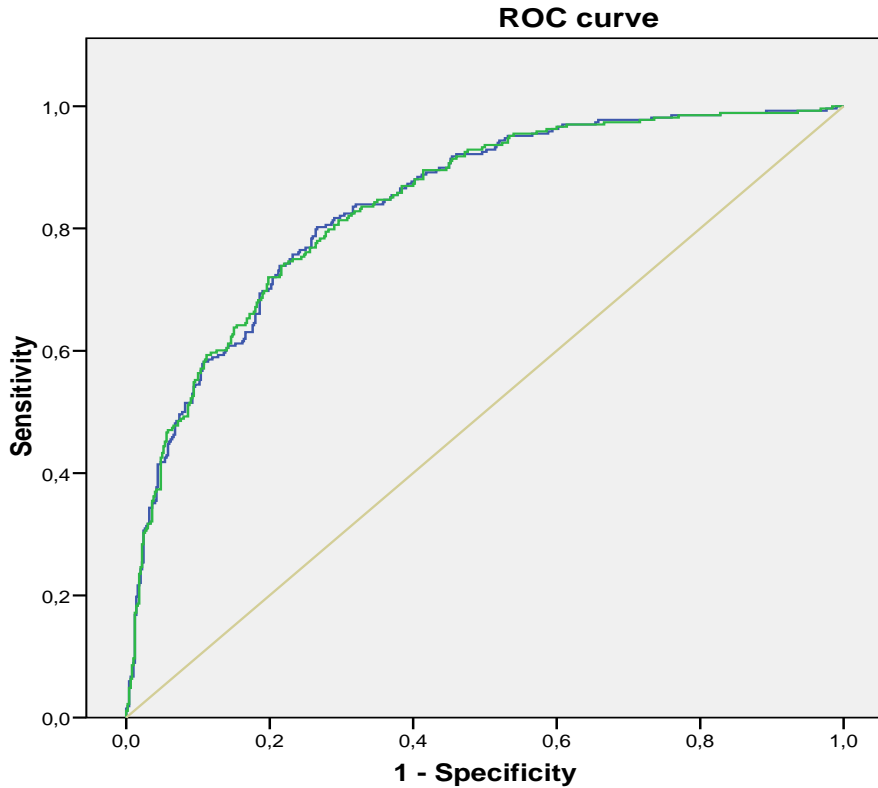
- Une synthèse de la performance d'un score quand le seuil s varie. x est classé en G1 si $S(x) > s$
- La courbe ROC curve relie le taux de vrais positifs $1 - \beta$ au taux de faux négatifs α .



L' AUC



- La surface sous la courbe ROC est un indice global de performance variant de 0.5 à 1
- Indice de Gini: deux fois la surface entre le courbe et la diagonale $G=2AUC-1$
- AUC et G permettent de choisir entre plusieurs modèles si les courbes ne se croisent pas
- Mais attention à ne pas comparer sur l'échantillon d'apprentissage un modèle simple avec un modèle complexe.

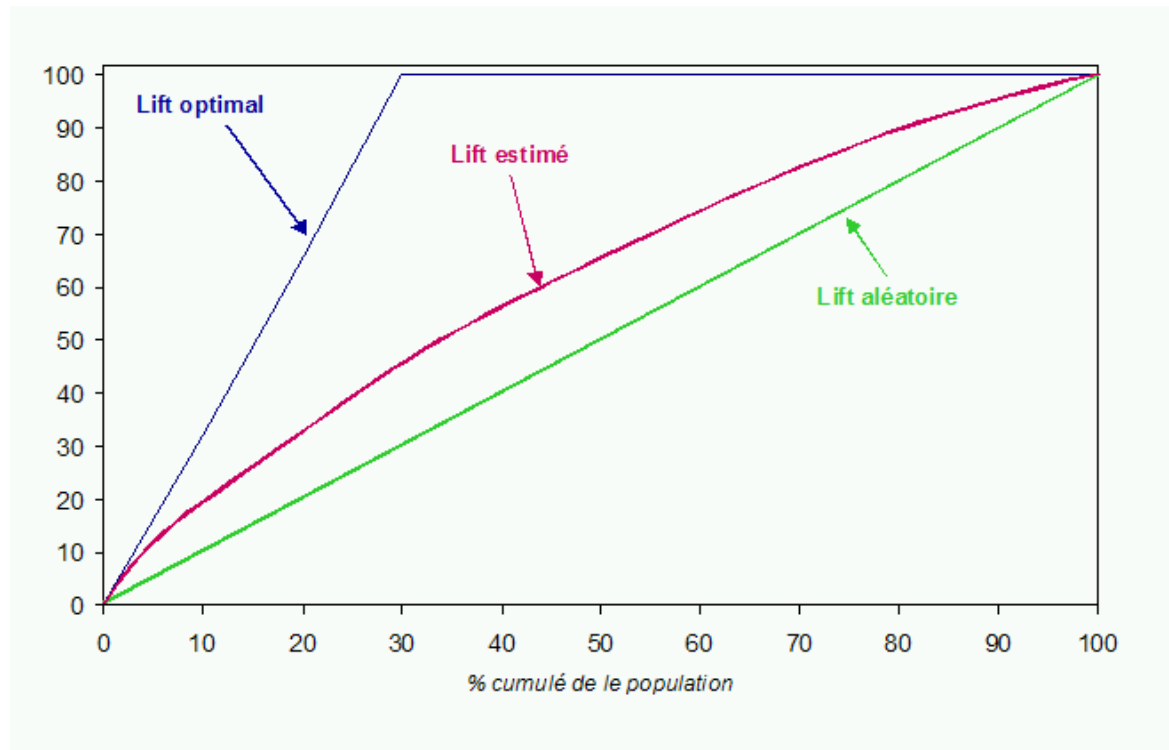


— scdisc
 — sclogist
 — Reference line

	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

Courbe de lift

% de la cible



Coefficient K_i (K_{xen})

- $K_i = (\text{surface entre lift estimé et aléatoire}) / (\text{surface entre lift idéal et aléatoire})$
- $K_i = 2AUC - 1 = G$

5. Construction et choix de modèles: théorie de l'apprentissage

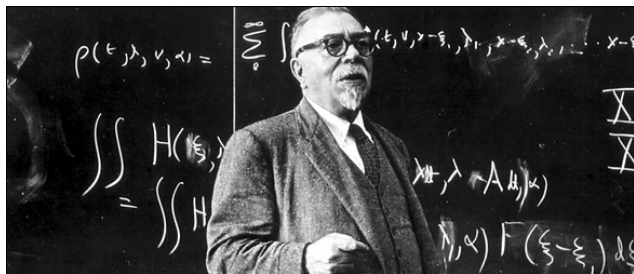
- A la recherche de modèles simples et efficaces
 - ne pas confondre ajustement (prédire le passé) et capacité de généralisation (prédire l'avenir)

De Guillaume d'Ockham à Vladimir Vapnik

- Guillaume d'Ockham 1319



- Norbert Wiener 1948



- Vladimir Vapnik 1982





Guillaume d'Occam (1285 - 3 avril 1349), dit le « docteur invincible » franciscain philosophe logicien et théologien scolastique.


Etudes à Oxford, puis Paris. Enseigne quelques années à Oxford.

Accusé d'hérésie, convoqué pour s'expliquer à Avignon, se réfugie à Munich, excommunié, à la cour de Louis de Bavière, lui-même excommunié. Meurt de l'épidémie de peste noire. Réhabilité par Innocent VI en 1359.

A inspiré le personnage du moine franciscain Guillaume de Baskerville dans le « Nom de la rose » d'Umberto Eco.

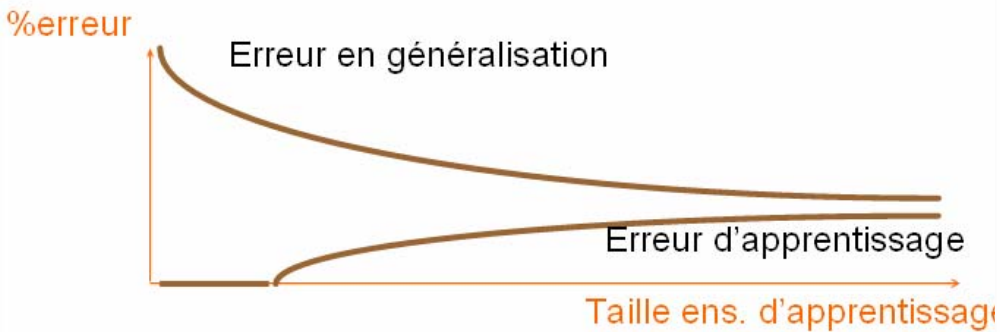
Premier jour, vêpres : « il ne faut pas multiplier les explications et les causes sans qu'on en ait une stricte nécessité. »

Apprentissage, généralisation et complexité

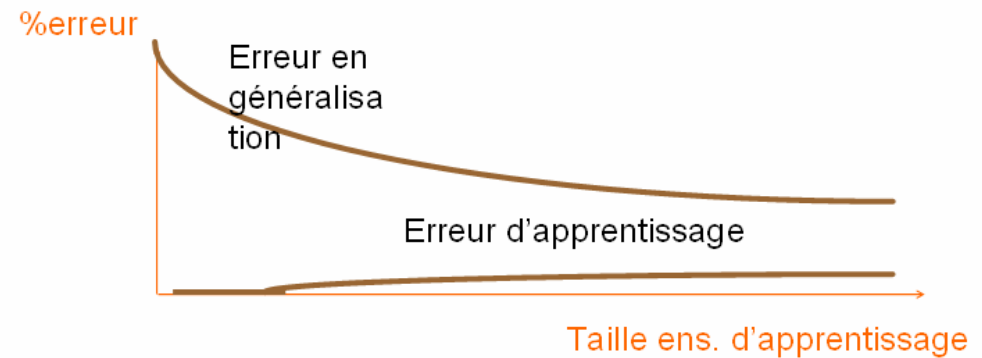


- Mesures d'erreur
 - Risque empirique sur les données utilisées
 - Risque R sur de futures données «généralisation»
- Comportement
 - selon le nombre de données disponibles
 - selon la complexité du modèle

Apprentissage consistant



Apprentissage non consistant



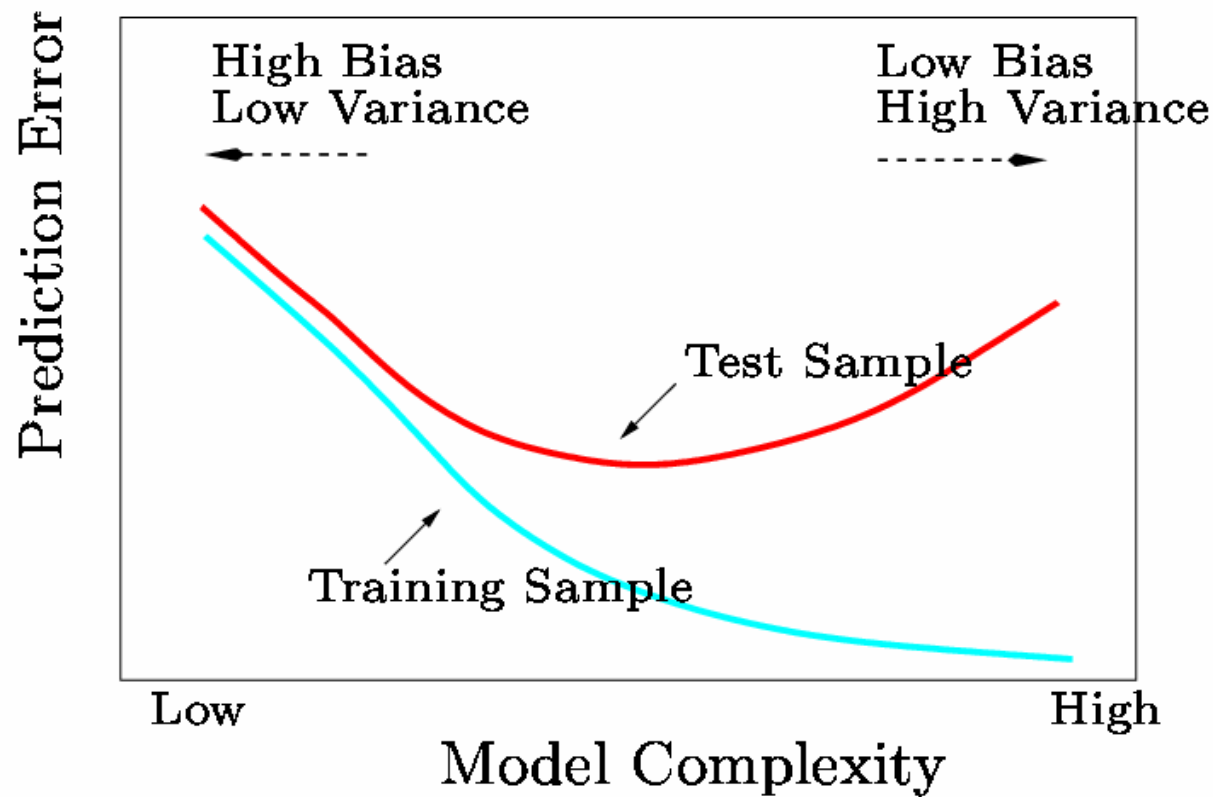


Figure 2.11: *Test and training error as a function of model complexity.*


Hastie et al., 2009

complexité d'un modèle



- Plus un modèle est complexe, mieux il s'ajuste en apprentissage mais avec de grands risques en test.
- \exists compromis optimal
- Comment mesurer la complexité d'un modèle?
 - V.Vapnik a montré que ce n'est pas le nombre de paramètres

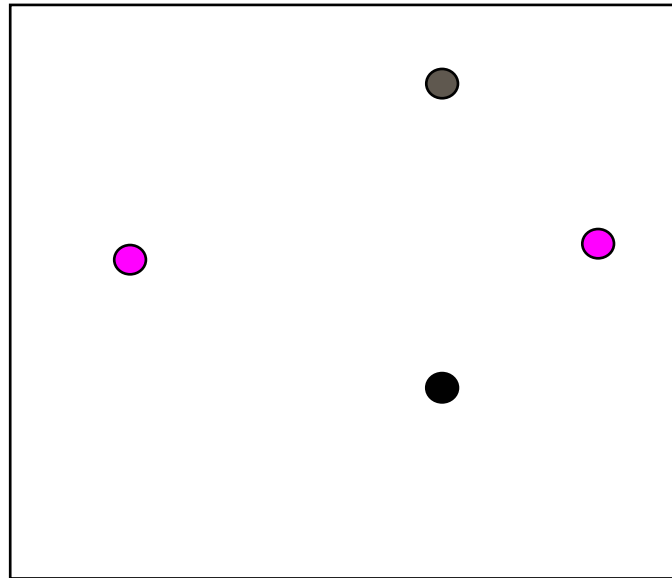
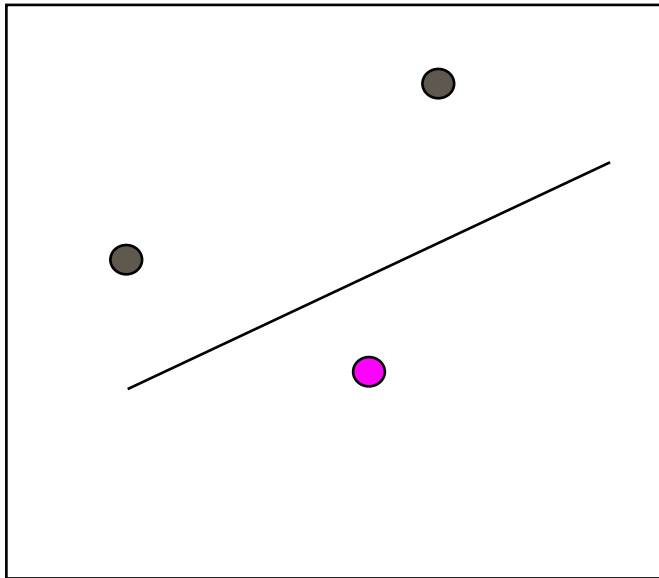
VC-dimension d'une famille de classifieurs



- Une mesure du pouvoir séparateur liée au nombre maximal de points séparables parfaitement. Notée h

Exemple

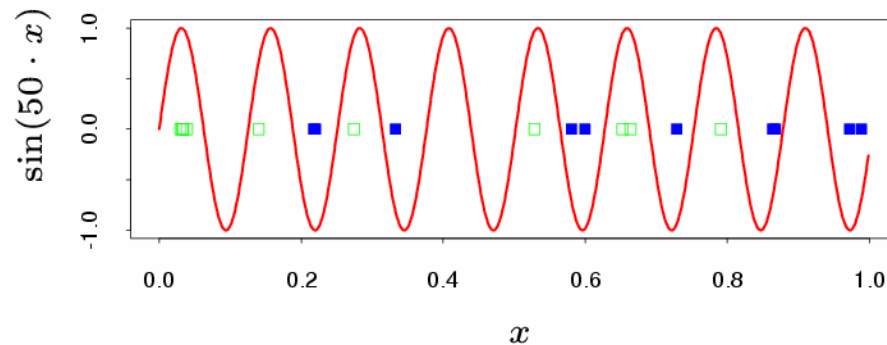
- En 2-D, la VC dimension des classifieurs linéaires non contraints est 3 (en p-D $VCdim=p+1$)



■ La VC dimension n'est pas égale au nombre de paramètres libres: elle peut être plus grande ou plus petite

■ La VC dimension de $f(x, w) = \text{sign}(\sin(w \cdot x))$
 $c < x < 1, c > 0,$
est infinie alors qu'il n'y a qu'un paramètre .

Hastie et al. 2001



La régression ridge

- La VC dimension de l'ensemble des indicatrices linéaires


$$f(X, \mathbf{w}) = \text{sign}\left(\sum_{i=1}^p (w_i x_i) + 1\right)$$

$$\|X\| \leq R$$

satisfaisant à la condition : $\|W\|^2 = \sum_{i=1}^p w_i^2 \leq \frac{1}{C}$

dépend de C et peut prendre toute valeur de 0 à $p+1$.

$$h \leq \min \left[\text{ent} \left(\frac{R^2}{C^2} \right); p \right] + 1$$

- 
- La ridge comme technique de régularisation
 - utile si le nombre de variables est grand
 - fournit des résultats plus robustes que les moindres carrés: coefficients plus stables
 - léger biais mais meilleur pouvoir prédictif

Inégalité de Vapnik

- Avec la probabilité $1 - \alpha$:

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

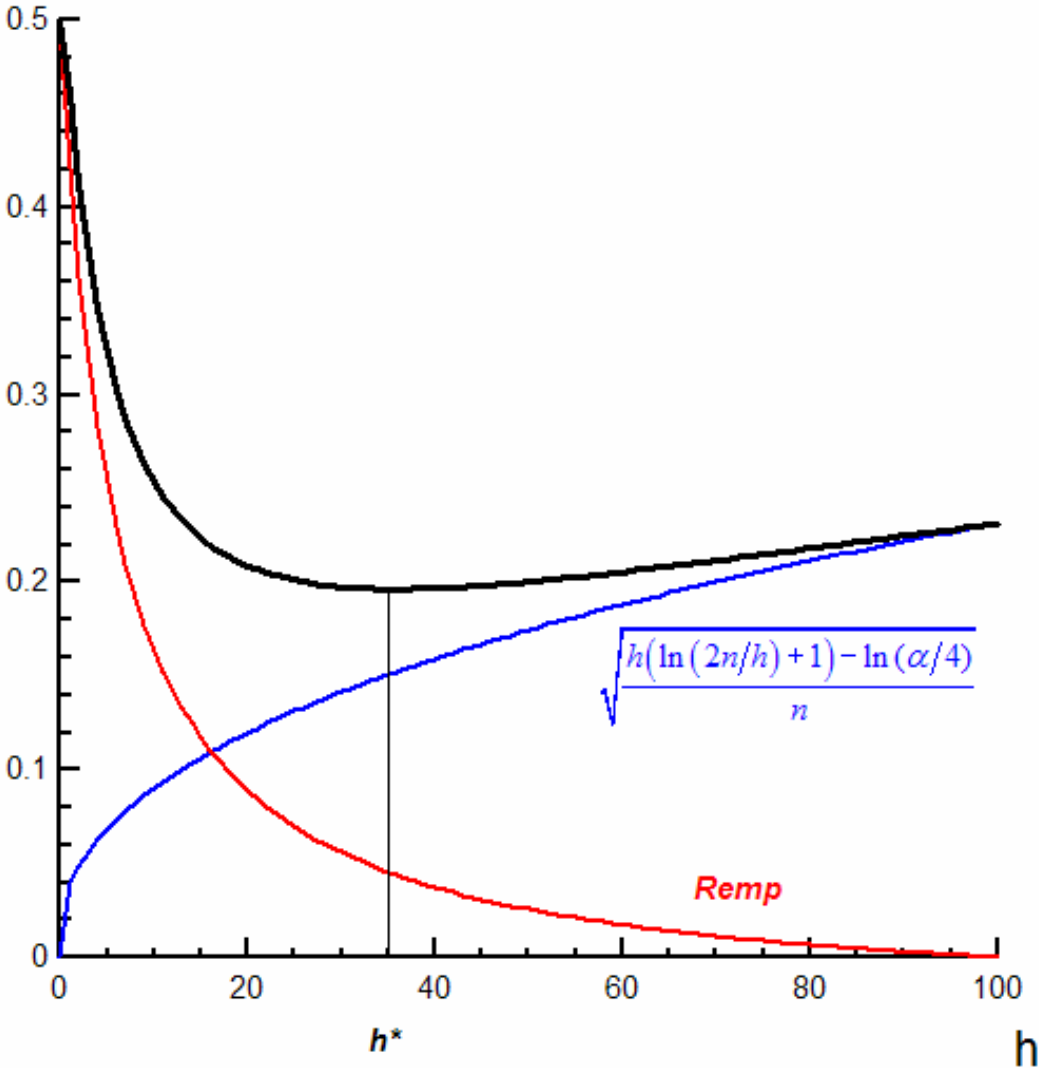
- ne fait pas intervenir p mais la VC dimension h
- Ne fait pas intervenir la distribution de probabilité P

Principe de minimisation structurée du risque (SRM)

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

- lorsque n/h est faible (h trop grand), le deuxième terme est grand
- L'idée générale du SRM est de minimiser la somme des deux termes à la droite de l'inéquation.

n fixé



Contrôle de h

- h doit être fini
- h/n doit être petit: si n augmente, on peut augmenter la complexité du modèle
- h décroît avec:
 - Réduction de dimension (cf. Disqual)
 - La marge (SVM)
 - k en régression ridge
- Mais h difficile à obtenir

Les 3 échantillons:

- Apprentissage: pour estimer les paramètres des modèles
- Test : pour choisir le meilleur modèle
- Validation : pour estimer la performance sur des données futures

- Rééchantillonner: validation croisée, bootstrap

Modèle final: avec toutes les données disponibles

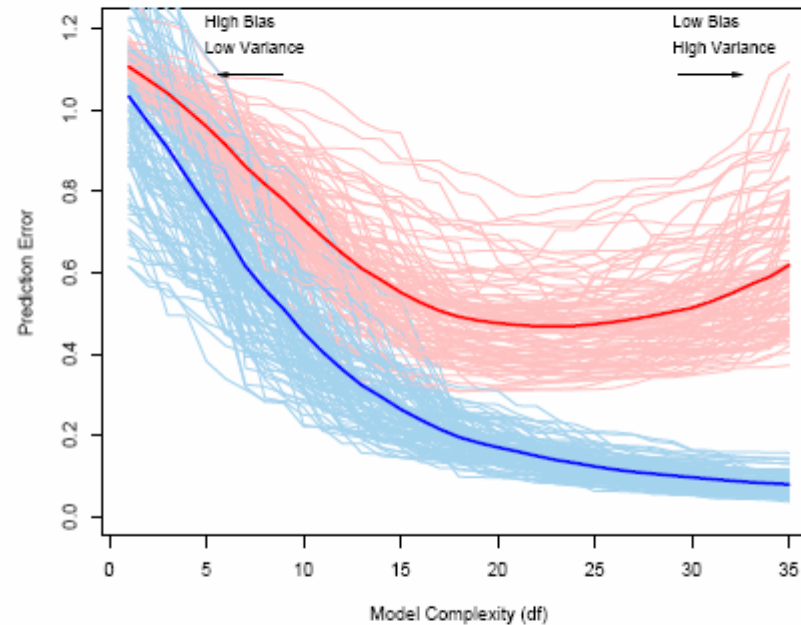
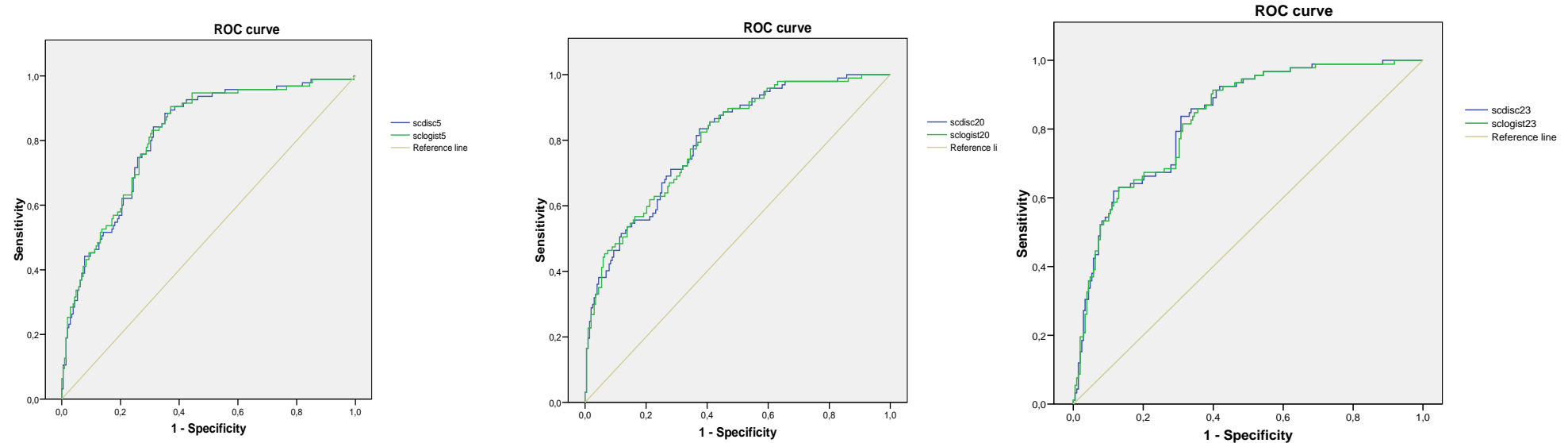


FIGURE 7.1. Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error $\overline{\text{err}}$, while the light red curves show the conditional test error $\text{Err}_{\mathcal{T}}$ for 100 training sets of size 50 each, as the model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

Variabilité

Apprentissage 70%, validation
30%, 30 retirages



6. Data mining : une nouvelle conception de la statistique et du rôle des modèles

- Modèles pour **comprendre** ou modèles pour **prévoir**?
 - Compréhension des données et de leur mécanisme générateur à travers une représentation simple (parcimonieuse)
 - Prédire de nouvelles observations avec une bonne précision

■ Paradoxe n° 1

- Un « bon » modèle statistique ne donne pas nécessairement des prédictions précises au niveau individuel. Exemple facteurs de risque en épidémiologie

■ Paradoxe n°2

- On peut **prévoir sans comprendre**:
 - pas besoin d'une théorie du consommateur pour faire du ciblage
 - un modèle n'est qu'un algorithme

- En data mining, un bon modèle est celui qui donne de bonnes prévisions
 - capacité prédictive sur de nouvelles observations «généralisation »
 - différent de l'ajustement aux données (prédire le passé)
 - Un modèle trop précis sur les données se comporte de manière instable sur de nouvelles données : phénomène de **surapprentissage**
 - Un modèle trop robuste (rigide) ne donnera pas un bon ajustement sur les données
 - **modèles issus des données**

Le défi de l'explosion du volume de données (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

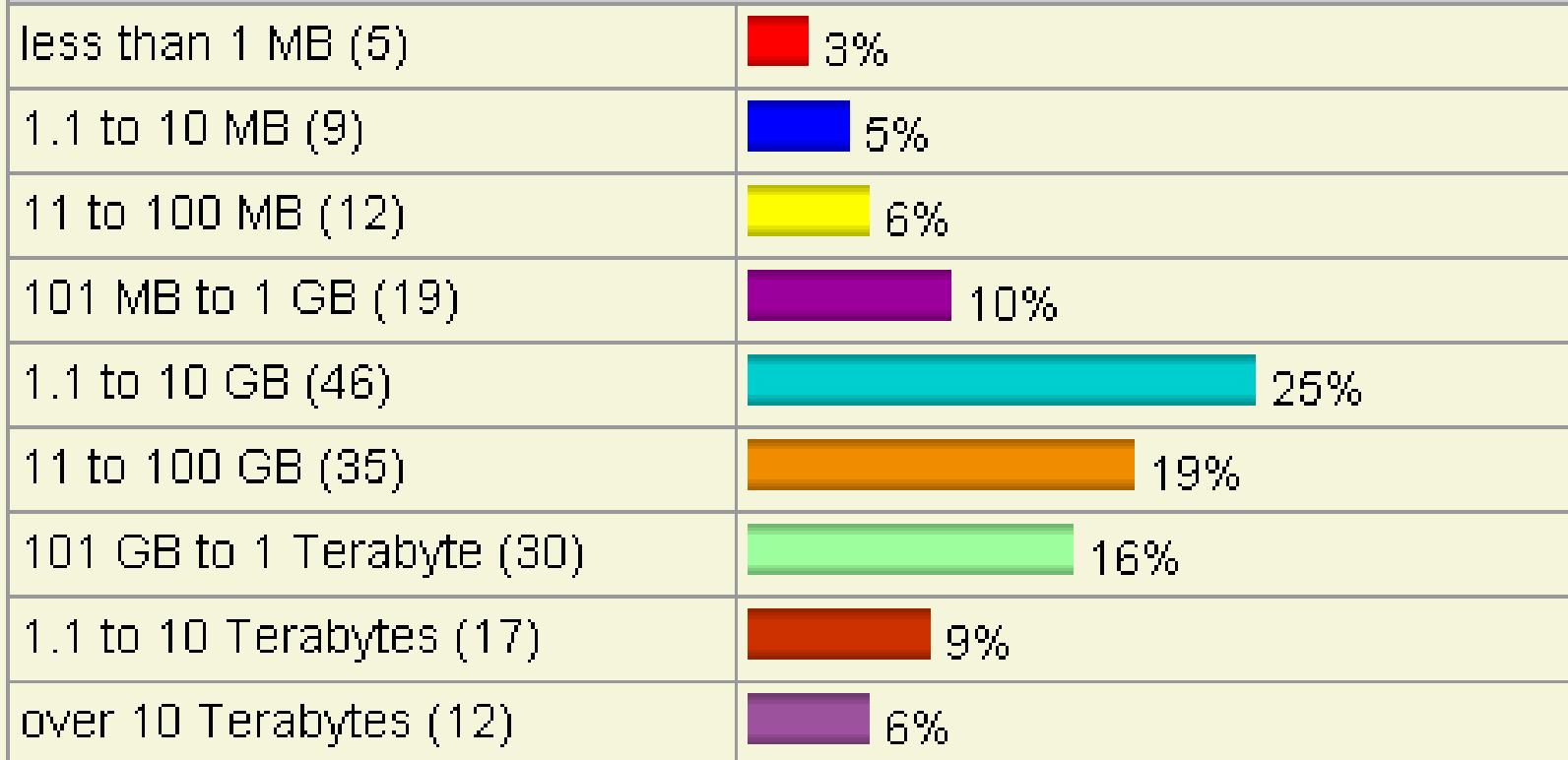
- **Cellular networks**




A telecom carrier generates **hundreds of millions of CDRs / day**
The network generates technical data : **40 M events / day** in a large city

Largest database data-mined Poll

What was the largest database or dataset you data-mined? [185 votes total]



<http://www.kdnuggets.com>

- 
- Tirer le meilleur des deux approches: **data driven** et **hypothesis driven** en les combinant
 - Plus que jamais un **métier d'avenir** pour ceux qui savent combiner les compétences statistiques et informatique
 - **éthique** et traitements de données personnelles

Références

- 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>
- Hastie, Tibshirani, Friedman (2009): *The Elements of Statistical Learning*, 2nd edition, Springer-Verlag, <http://www-stat.stanford.edu/~hastie/Papers/ESLII.pdf>
- Nisbet R., Elder J., Miner G. (2009): *Handbook of Statistical Analysis and Data Mining Applications*, Academic Press
- Tufféry, S. (2009) *Data Mining et Statistique Décisionnelle*, Technip



Merci pour votre attention