Classification supervisée et credit scoring

Gilbert Saporta

Conservatoire National des Arts et Métiers, Paris

saporta@cnam.fr

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

Plan

1. Introduction
2. Techniques linéaires pour construire une grille de scores
3. Prédicteurs qualitatifs
4. Scoring direct
5. Validation et choix de modèles
6. Le traitement des refusés (reject inference)
7. Scores de durée
8. Conclusion

1.Introduction

Credit scoring is the set of decision models and their underlying techniques that aid lenders in the granting of consumer credit.

Credit scoring is one the most successful applications of statistical modeling in finance and banking. Yet because credit scoring does not have the same glamour as the pricing of exotic financial derivatives or portfolio analysis, the literature on the subject is very limited.

Thomas & al. 2002

Le comité de Bâle sur la supervision bancaire

- Créé en 1974 par le G10

  Banque des Règlements Internationaux (BIS)
  - Réduire la vulnérabilité par la mise en place d'un ratio prudentiel attestant d'un niveau minimal de fonds propres.
- Accords Bâle II

- **Bâle 2**
  - Une « révolution quantitative » (A.L.Rémy Crédit Agricole)

« banks are expected to provide an estimate of the PD and LGD »

  - PD (probability de défaut)
  - LGD (perte en cas de défaut)
  - EAD (exposition en cas de défaut)
  - Calcul du capital nécessaire au niveau de confiance 99.9% à un an

- Impact énorme sur les études statistiques.
  - Exigence de justification statistique et de backtesting imposé par le régulateur (Commission Bancaire)

⇒ Recrutements massifs

- Le « New Basel Capital Accord » régulera les prêts bancaires à partir de 2007

- Contexte statistique du credit scoring:
  - Réponse Y à 2 catégories (« bons » « mauvais »)
  - $X_1,\ldots,X_p$ prédicteurs

- terminologie :
  - classification
  - Apprentissage supervisé
  - discrimination
  - Reconnaissance des formes
  - ...

- Pas seulement un problème de classement
  - Evaluation des risques plus que décision binaire
- Aspects particuliers:
  - Traitement des refusés
  - Prêts à long terme

2. Techniques linéaires de construction de grilles de score

- Analyse discriminante
- Régression logistique
- SVM linéaires
- Régressions régularisées
  - PLS
  - ridge
- Autres (GLM, programmation linéaire...)

2.1 L'analyse discriminante

- **2.1.1 La fonction linéaire de Fisher (1936)**
  - Prédicteurs numériques:

$$\boldsymbol{\beta} = \mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) = \mathbf{W}^{-1}\begin{pmatrix} \overline{x}_1^1 - \overline{x}_2^1 \\ . \\ \overline{x}_1^p - \overline{x}_2^p \end{pmatrix}$$

  - Le « meilleur» prédicteur au sens de la maximisation du T de Student
  - Score de Fisher:

$$S(\mathbf{x}) = (\mathbf{g}_1 - \mathbf{g}_2)'\mathbf{W}^{-1}\mathbf{x} - \frac{1}{2}(\mathbf{g}_1 - \mathbf{g}_2)'\mathbf{W}^{-1}(\mathbf{g}_1 + \mathbf{g}_2)$$

$$= \beta_1\mathbf{x}_1 + \ldots + \beta_p\mathbf{x}_p + \beta_0$$

2.1.2 Une régression « politiquement incorrecte »

- $\mathbf{y}$ à 2 valeurs (-1;+1) ou (0;1) ou (a;b)
- a=n/n_1 b=-n/n_2

$$\boldsymbol{\beta} = \mathbf{V}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$$

$$D_p^2 = \frac{n(n-2)}{n_1 n_2} \frac{R^2}{1-R^2}$$

- D_p distance de Mahalanobis
- Une source d'incompréhension et de controverses !

2.1.3 ADL et modèle probabiliste

- ADL optimale (règle de Bayes) pour des prédicteurs gaussiens avec matrices de covariance identiques
- Si *a priori* différents :

$$S(\mathbf{x}) = (\mathbf{g}_1 - \mathbf{g}_2)'\mathbf{W}^{-1}\mathbf{x} - \ln(\frac{p_2}{p_1}) - \frac{1}{2}(\mathbf{g}_1 - \mathbf{g}_2)'\mathbf{W}^{-1}(\mathbf{g}_1 + \mathbf{g}_2)$$

  - Probabilité *a posteriori* $P(G_1 / \mathbf{x}) = \dfrac{\exp(S(\mathbf{x}))}{1+\exp(S(\mathbf{x}))}$

fonction logistique

- Peut quand même être appliquée si ces hypothèses ne sont pas satisfaites

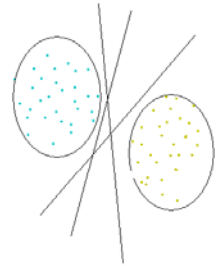
2.2 Régression logistique

$$\pi(\mathbf{x}) = P(Y = 1 / \mathbf{X} = \mathbf{x}) = \frac{e^{\beta_0 + \beta_1 \mathbf{x}_1 + ... + \beta_p \mathbf{x}_p}}{1 + e^{\beta_0 + \beta_1 \mathbf{x}_1 + ... + \beta_p \mathbf{x}_p}}$$

- Berkson (1944), Cox (1958): biostatistique, épidémiologie
- Plus tard en économétrie avec McFadden (1973), prix Nobel
- Facteurs de risque plutôt que prévision individuelle

$$score = \beta_0 + \beta_1 \mathbf{x}_1 + ... + \beta_p \mathbf{x}_p$$

- La méthode préférée des économètres, devenu le « standard » de l'industrie bancaire
  - Plus « scientifique »?: prédiction des probabilités, estimation par le maximum de vraisemblance, erreurs standard calculables, coefficients uniques interprétés comme odds-ratios
  - Facilité des logiciels: prise en compte aisée de prédicteurs qualitatifs, sans manipulation d'indicatrices



- Mais:
  - Pas de solution si séparation parfaite
  - MV conditionnel, propriétés asymptotiques
  - Le bootstrap permet de calculer les erreur standard en ADL
  - Peu de différence en pratique:

    « It is generally felt that logistic regression is a safer, more robust bet than the LDA model, relying on fewer assumptions . It is our experience that the models give very similar results , even when LDA is used in inappropriately, such as with qualitative variables. » Hastie and al.(2001)
  - Le choix d'une méthode ne doit pas relever de l'idéologie mais de sa performance

2.3 Probabilités a posteriori et stratification

- Estimer P demande de connaître les vraies probabilités a priori
- Les modifier change seulement β_0 en ADL et en logistique:
  - Important pour les probabilités , pas pour un score

2.4 Autres méthodes issues de la régression linéaire

En cas de multicolinéarité.

- **2.4.1 Régression ridge**

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$$

$$\min \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right\|^2 \quad \text{with} \quad \left\| \boldsymbol{\beta} \right\|^2 < d^2$$

- k optimisé par validation croisée (KXEN)

2.4.2 Analyse discriminante PLS

- Des composantes expliquant à la fois Y et les X
- Critère de Tucker:

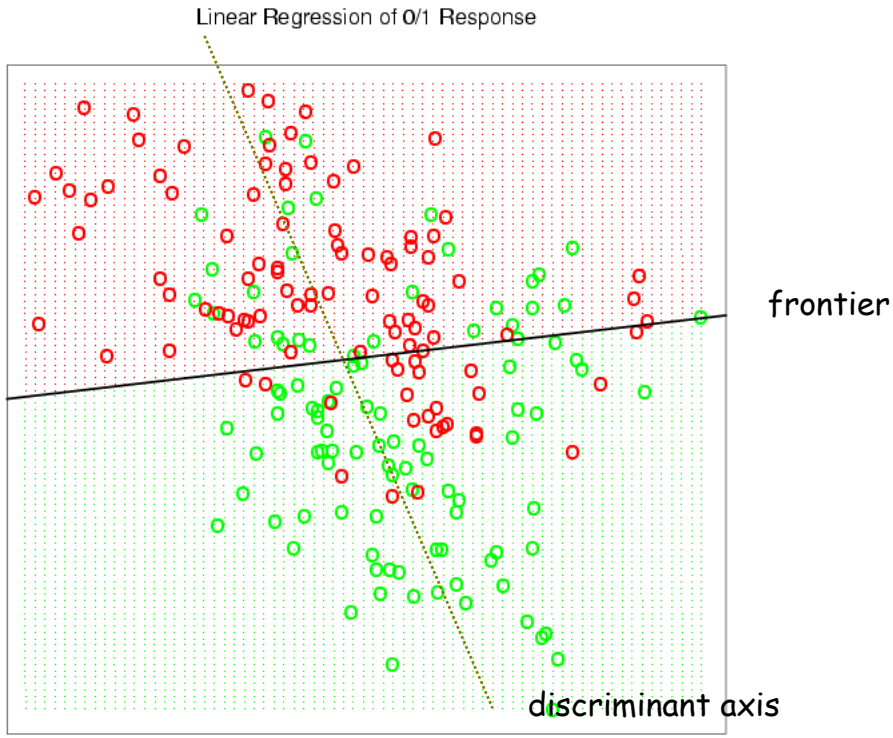
$$\max \quad (\mathrm{cov}(\mathbf{y};\mathbf{Xw}))^2$$

$$(\mathrm{cov}(\mathbf{y};\mathbf{Xw}))^2 = r^2(\mathbf{y};\mathbf{Xw}).V(\mathbf{Xw}).V(\mathbf{y})$$

- Composantes suivantes sur les résidus. Arrêt par validation croisée.
- Suite de régressions simples (ni inversion, ni diagonalisation)

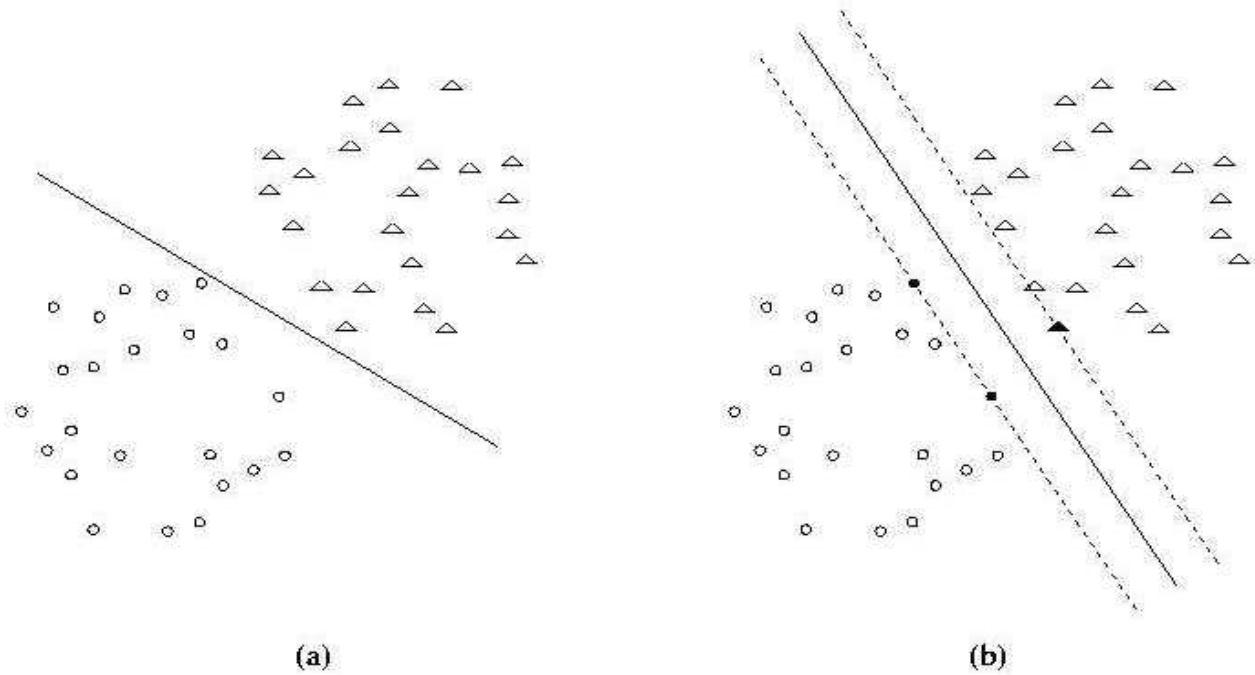
2.5 SVM linéaires

- Score linéaire= frontière linéaire

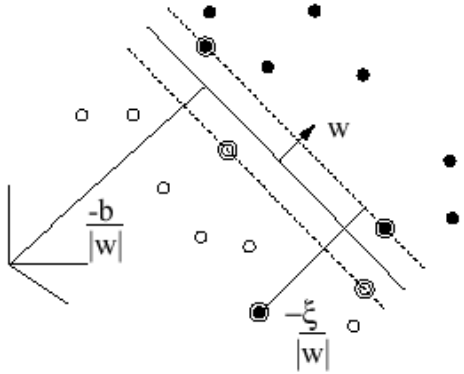


Adapted from Hastie & al. 2001

- L’hyperplan optimal de Vapnik maximise la marge (distance du point le plus proche à la frontière)



- Cas non séparable :



- Trade-off entre marge et taux d'erreur
- Programme quadratique

$$y_i = \pm 1$$

$$\min \|\mathbf{w}\|$$

$$\text{subject to : } \begin{cases} y_i(\mathbf{x_i'}\mathbf{w} + b) \geq 1 - \xi_i \\ \sum \xi_i < \gamma \end{cases}$$

- **Classifieur ou fonction de score**

$$f(\mathbf{x}) = \mathbf{w}'\mathbf{x} + b = \sum_{\alpha_i > 0} \alpha_i y_i \mathbf{x}_i' \mathbf{x} + b$$

- $f(\mathbf{x})$ ne dépend que des points supports
- combinaison linéaire
- Décision selon le signe de $f(\mathbf{x})$
- Plus robuste aux outliers que l'ADL

3. Prédicteurs qualitatifs

- Fréquent en crédit à la consommation, mais pas dans les publications..
  - Profession
  - Emploi
  - Statut matrimonial
  - ...

Catégorisation de prédicteurs numériques

- Ex: classes d'âge au lieu de l'âge
- Une perte de précision?
- Traitement des non-linéarités

$$S = \sum_{j=1}^{p} \varphi_j(X_j) \qquad \varphi_j \text{ fonctions en escalier}$$

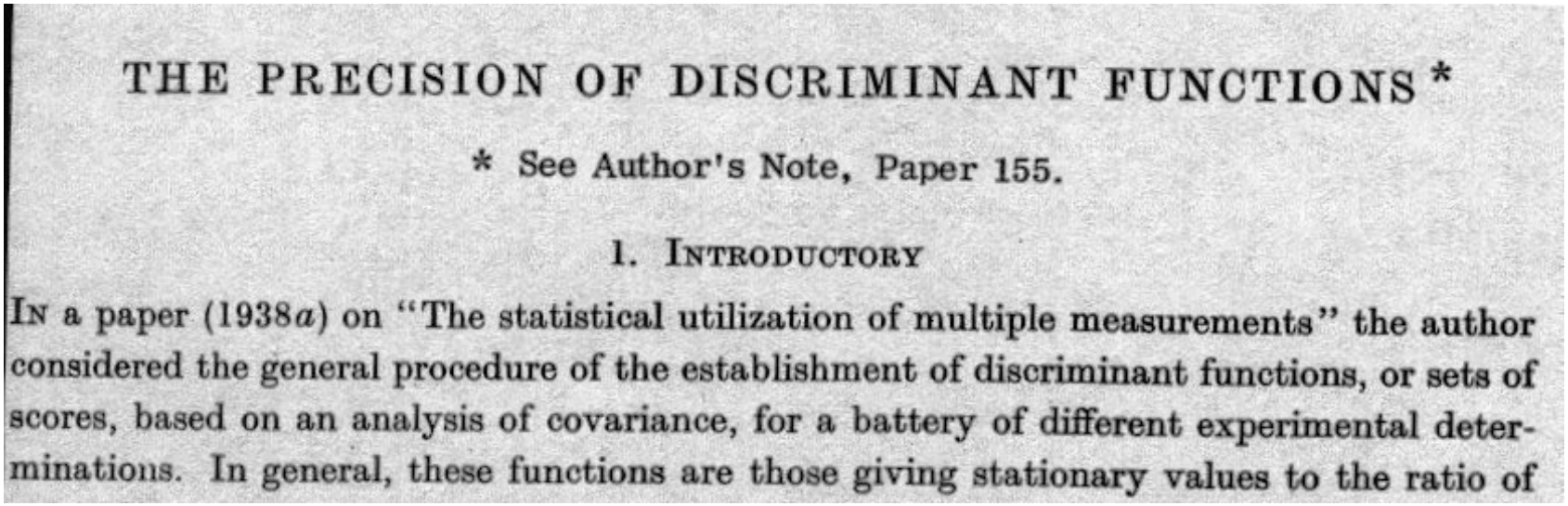
- Utile pour les outliers: robustesse
- Traitement des valeurs manquantes

Prétraitement

- Sélection de variables, discrétisation, détection des interactions $X_j * X_k$: consommatrices de temps
- De nouveaux outils :
  - K2C, Khiops, Datalab..

3.1 ADL pour prédicteurs qualitatifs: un peu de (pré)histoire

- Fisher (1940)
  - Un seul prédicteur
  - Equations de l'analyse des correspondances
  - Introduction du vocable « Scores »



THE PRECISION OF DISCRIMINANT FUNCTIONS *

* See Author's Note, Paper 155.

1. INTRODUCTORY

IN a paper (1938*a*) on "The statistical utilization of multiple measurements" the author considered the general procedure of the establishment of discriminant functions, or sets of scores, based on an analysis of covariance, for a battery of different experimental determinations. In general, these functions are those giving stationary values to the ratio of

For example, in a contingency table individuals are cross classified in two categories, such as eye colour and hair colour, as in the following example (Tocher's data for Caithness compiled by K. Maung of the Galton Laboratory).

Eye colour	Hair colour					
	Fair	Red	Medium	Dark	Black	Total
Blue	326	38	241	110	3	718
Light	688	116	584	188	4	1580
Medium	343	84	909	412	26	1774
Dark	98	48	403	681	85	1315
Total	1455	286	2137	1391	118	5387

Variation among the four eye colours may be regarded as due to variations in three variates defined conveniently in some such way as the following:

Eye colour	x_1	x_2	x_3
Blue	0	0	0
Light	1	0	0
Medium	0	1	0
Dark	0	0	1

We may then ask for what eye colour scores, i.e. for what linear function of x_1, x_2, x_3, are the five hair colour classes most distinct. The answer may be found in a variety of ways. For example, by starting with arbitrarily chosen scores for eye colour, determining from these average scores for hair colour, and using these latter to find new scores for eye colour.

Apart from a contraction of scale by a factor R^2 for each completed cycle, this form tends to a limit, and yields scores such as the following:

Eye colour	x	Hair colour	y
Light	−0·9873	Fair	−1·2187
Blue	−0·8968	Red	−0·5226
Medium	0·0753	Medium	−0·0941
Dark	1·5743	Dark	1·3189
		Black	2·4518

The particular values given above have been standardized so as to have mean values zero, and mean square deviations unity. In the sample from which they are derived each score has a linear regression on the other, the regression coefficient being 0·44627; this is, of course, equal to the correlation coefficient between the two scores regarded as variates. Hotelling has called pairs of functions of this kind canonical components. It may be noticed that no assumption is introducéd as to the order of the classes of each category. In Tocher's schedule Light eyes come between Blue and Medium, but the discriminant function puts Blue between Medium and Light, though near the latter.

3.2 Cas général : p prédicteurs

- Quantification optimale:
  - Attribuer des notes partielles aux catégories des prédicteurs pour maximiser la distance de Mahalanobis dans $\mathbb{R}^p$
- Une analyse discriminante où les variables qualitatives sont remplacées par les indicatrices des modalités

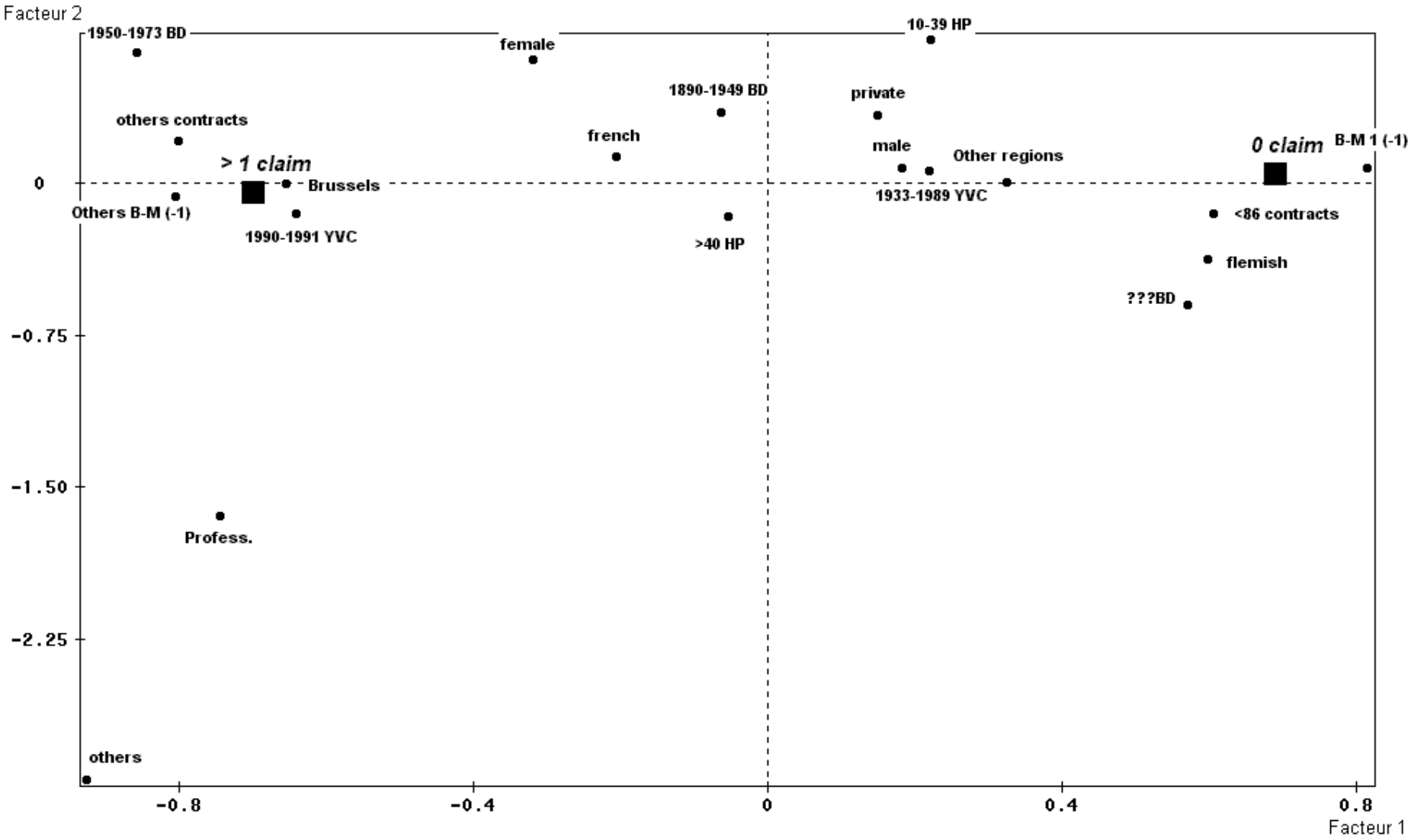
$$X = \left(\begin{array}{ccc|cc} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{array}\right)$$

- X de rang insuffisant: $\text{rank}(X) = \Sigma m_i - p$
  - Solution classique: éliminer une indicatrice par prédicteur (note nulle)
  - Disqual (Saporta, 1975):
    - ADL sur une sélection de composantes de l'Analyse des Correspondances Multiples de X. Semblable à la régression sur composantes principales
    - Sélection experte selon deux critères: inertie et corrélation avec la réponse

Exemple assurance (jeu d'essai SPAD)

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

ACM


Facteur 2
1950-1973 BD
female
10-39 HP
1890-1949 BD
private
others contracts
french
male
Other regions
0 claim
B-M 1 (-1)
> 1 claim
0
Brussels
1933-1989 YVC
Others B-M (-1)
<86 contracts
1990-1991 YVC
>40 HP
flemish
???BD
-0.75
-1.50
Profess.
-2.25
others
-0.8
-0.4
0
0.4
0.8
Facteur 1

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

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

- **3.3 scores normalisés**
  - Echelle de 0 à 1000
  - Transformation linéaire du score et du seuil

Grille de score

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

3.4 Discriminante PLS et discrimination barycentrique

- Première composante PLS : p régressions simples séparées sur chaque prédicteur (variables indicatrices)
- Chaque régression PLS de Y sur les indicatrices de X_j est identique à la régression usuelle (Y standardisé, pas X, pas de terme constant)

- Une composante PLS: équivalent à l'AFC du tableau de contingence concaténé

		good	bad
1	cusag1	29	96
2	cusag2	344	272
3	sexe1	288	253
4	sexe2	76	78
5	sexe3	9	37
6	clang1	250	295
7	clang2	123	73
8	age3m1	118	99
9	age3m2	40	163
10	age3m3	215	106
11	cpost2m1	75	172
12	cpost2m2	298	196
13	bm2m_11	298	59
14	bm2m_12	75	309
15	puis2m1	91	47
16	puis2m2	282	321
17	dpoli2m1	277	137
18	dpoli2m2	96	231

Technique dite “discrimination barycentrique”:

• Score d’un individu: somme des *p* probabilités conditionnelles d’appartenance au groupe 2.

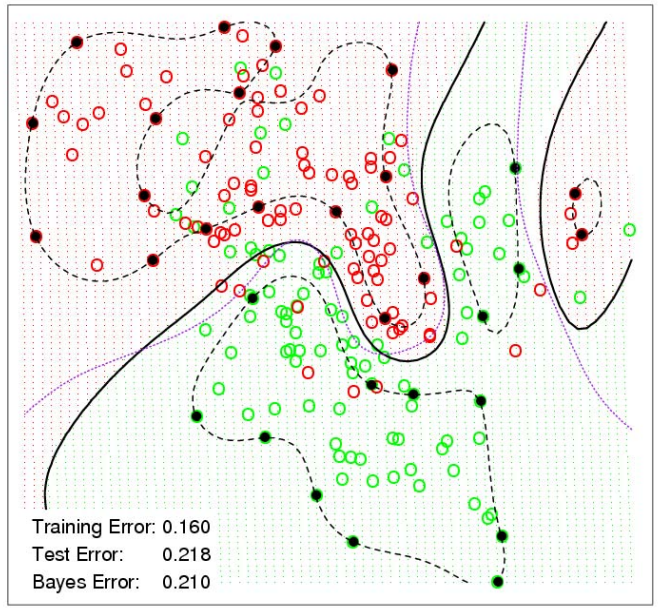
• Semblable au classifieur “naïf Bayes” : score multiplicatif.

• Equivalent à *Disqual* si les prédicteurs sont indépendants deux à deux

4. Scoring direct

- Toute méthode donnant un score ou une probabilité d'appartenance: Estimation de densité, Réseaux de neurones,SVM non linéaires etc.
  - Une probabilité est un score compris entre 0 et 1...

$$f(\mathbf{x}) = \sum_{i \in \text{supports}} \alpha_i y_i K(\mathbf{x}_i; \mathbf{x}) + b = 0$$



Boîtes-noires:

- Manque d'interprétabilité
  - Ne peut être utilisé en crédit à la consommation: obligations légales d'expliquer le refus
  - Doit être adapté aux prédicteurs qualitatifs c
  - Composantes d'ACM, ou pré-scores

- Et les arbres?

5 Validation et choix de modèle

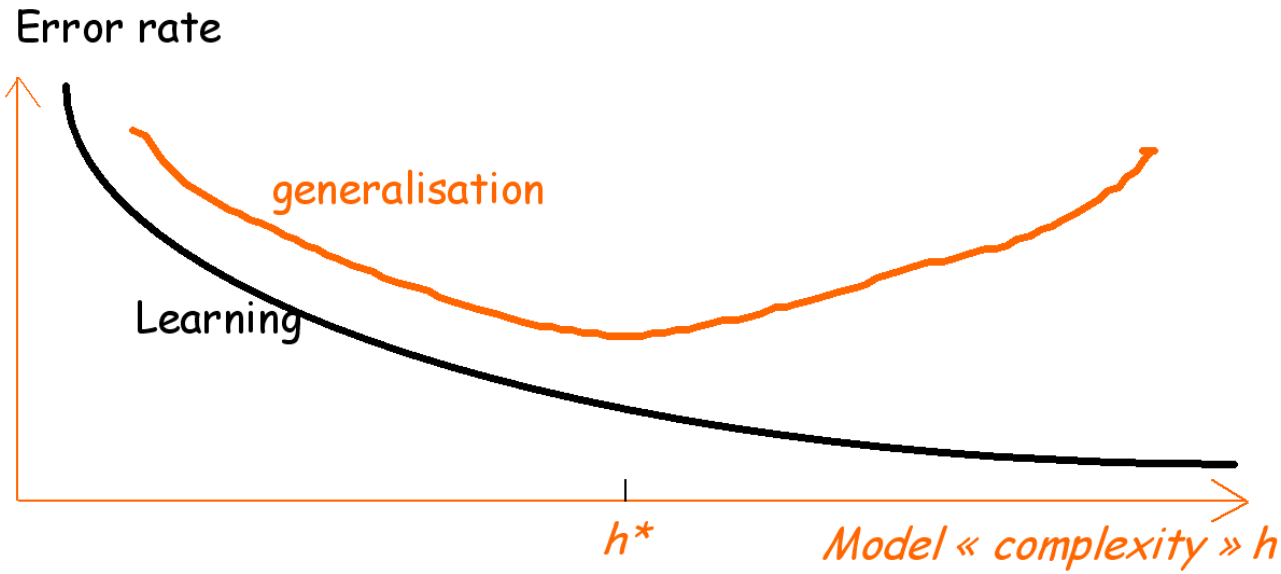
- **5.1 Peut-on et doit-on utiliser des critères statistiques?**
  - D^2, vraisemblance:
    - adéquation du modèle aux données d'apprentissage
    - Non reliés à la capacité prédictive mais facile à optimiser..
  - Vraisemblance pénalisée (AIC,BIC): trop restrictive
    - Réseaux de neurones, ridge régression, PLS?

- Le credit scoring: pas de la science, juste du business...

- On ne cherche pas le « vrai » modèle mais une procédure efficace

5.2 Taux d'erreur et théorie statistique de l'apprentissage

- Taux d'erreur et complexité



Risque empirique et VC dimension

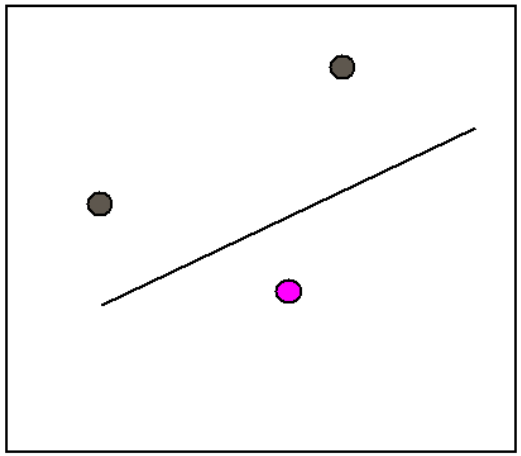
- Risque empirique= erreur en apprentissage R_{emp}
- Erreur en généralisation = R
  - Les deux sont des espérances
- Inégalité de Vapnik
  - Avec une probabilité 1-q

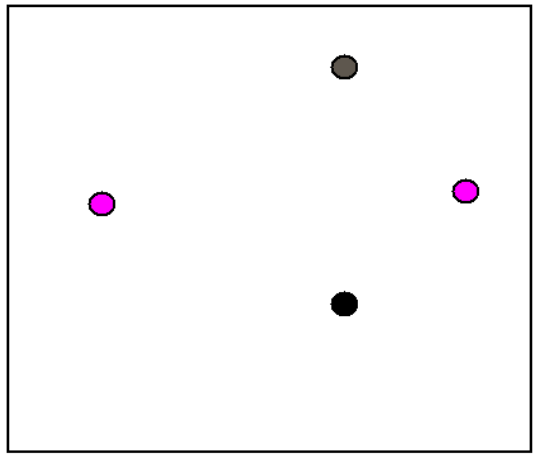
$$R < R_{\text{emp}} + \sqrt{\frac{h\left(\ln\left(2n/h\right)+1\right)-\ln q/4}{n}}$$

« Intervalle de confiance »

VC dimension h

- Une mesure de complexité liée au pouvoir séparateur d'une famille de classifieurs différente du nombre de paramétres





h= nombre maximal de points parfaitement séparable

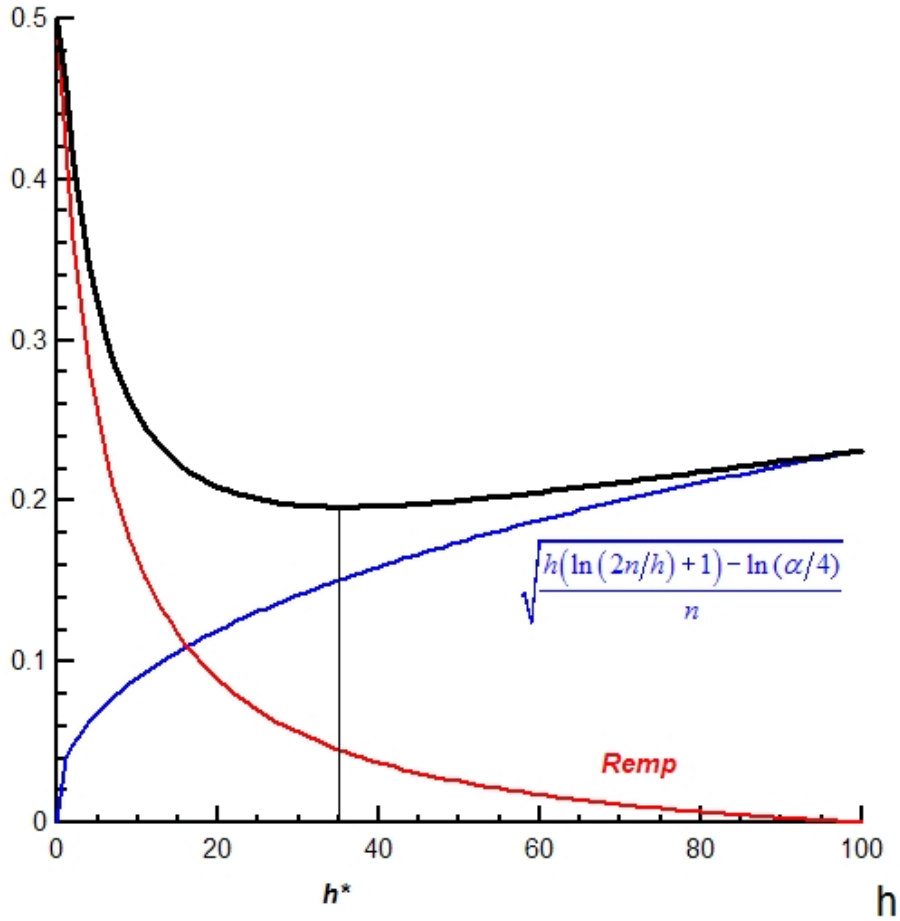
h=3 pour des frontières linéaires en 2D

Conséquences de l'inégalité de Vapnik

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

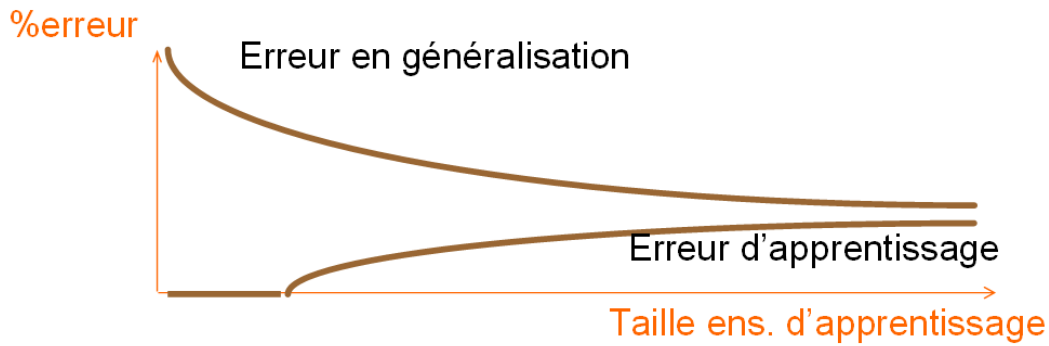
- Dépend de n/h: on peut augmenter h avec n (mais moins vite...)
- Minimiser la borne: compromis entre ajustement et robustesse
- Choix entre modèles de complexité croissante: SRM (Structured Risk Minimization)

dimension optimale *h*à n fixé*



Contrôle de h

- *h* doit être fini pour assurer la convergence



- *h* décroit avec:
  - Réduction de dimension (cf. Disqual)
  - La marge (SVM)
  - *k* en régression ridge
- Mais *h* difficile à obtenir

5.3 Les 3 échantillons

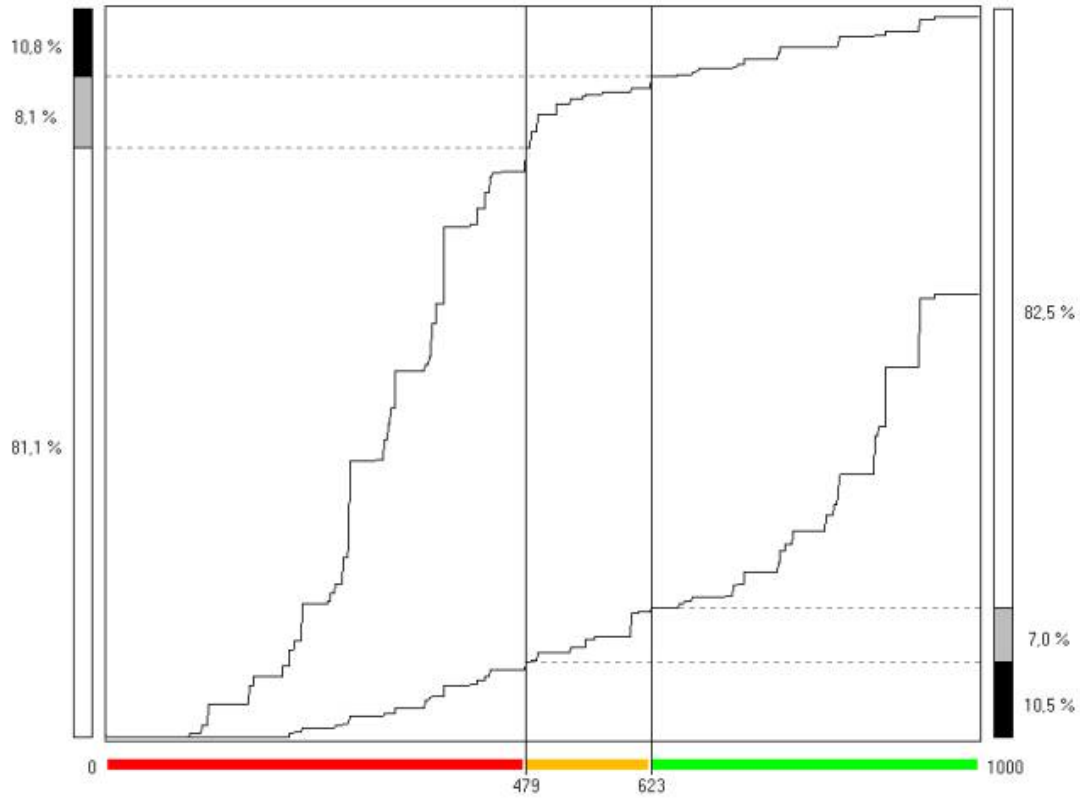
- **Apprentissage**: pour estimer les paramètres des modèles
- **Test** : pour choisir le meilleur modèle
- **Validation** : pour estimer la performance sur des données futures
- Nécessité de faire plusieurs tirages

- **Modèle final:** avec toutes les observations

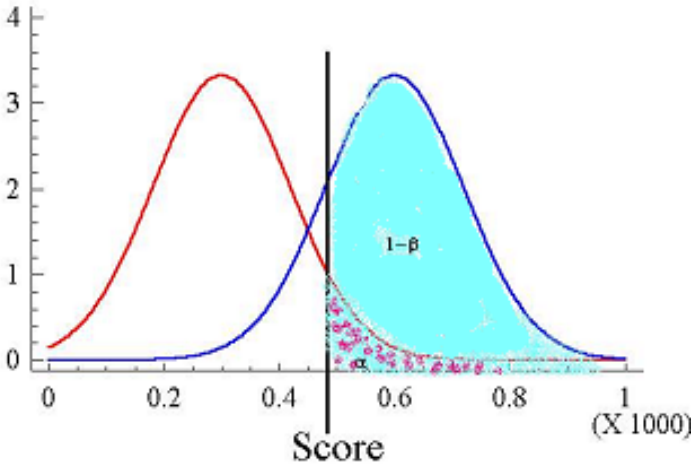
5.4 ROC, lift et autres mesures

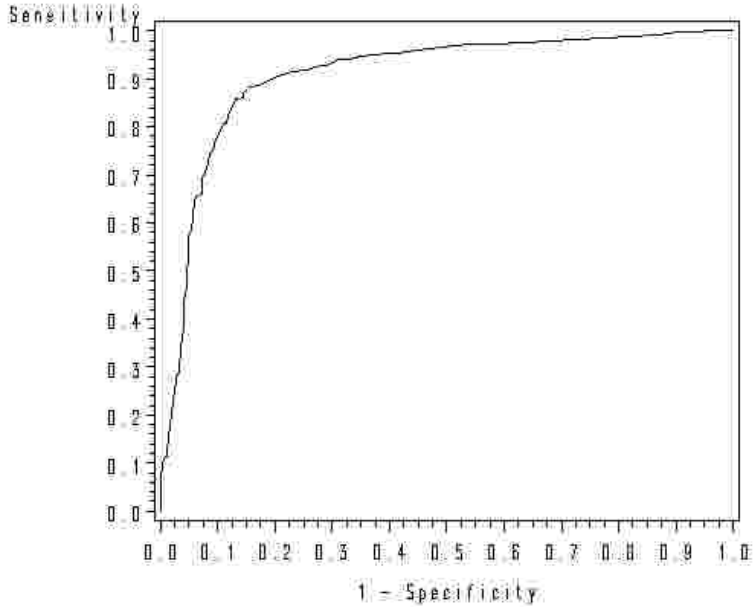
- Taux d'erreur pas toujours pertinent:
  - Nécessite le choix d'un seuil
    - Probabilité a posteriori >0.5 discutable
    - Minimisation d'un coût (mais coûts souvent inconnus)
- Etudier les variations de la performance quand on change le seuil

Traffic light zones



Courbe ROC





% de vrais « bons » (1-β)
contre % de faux « bons » (α)

- Courbe ROC invariante pour toute transformation monotone croissante
- Surface sous la courbe: mesure de performance permettant de comparer (partiellement) des modèles
- $$AUC = \int_{s=+\infty}^{s=-\infty} (1-\beta(s))d\alpha(s) = P(X_1 > X_2)$$

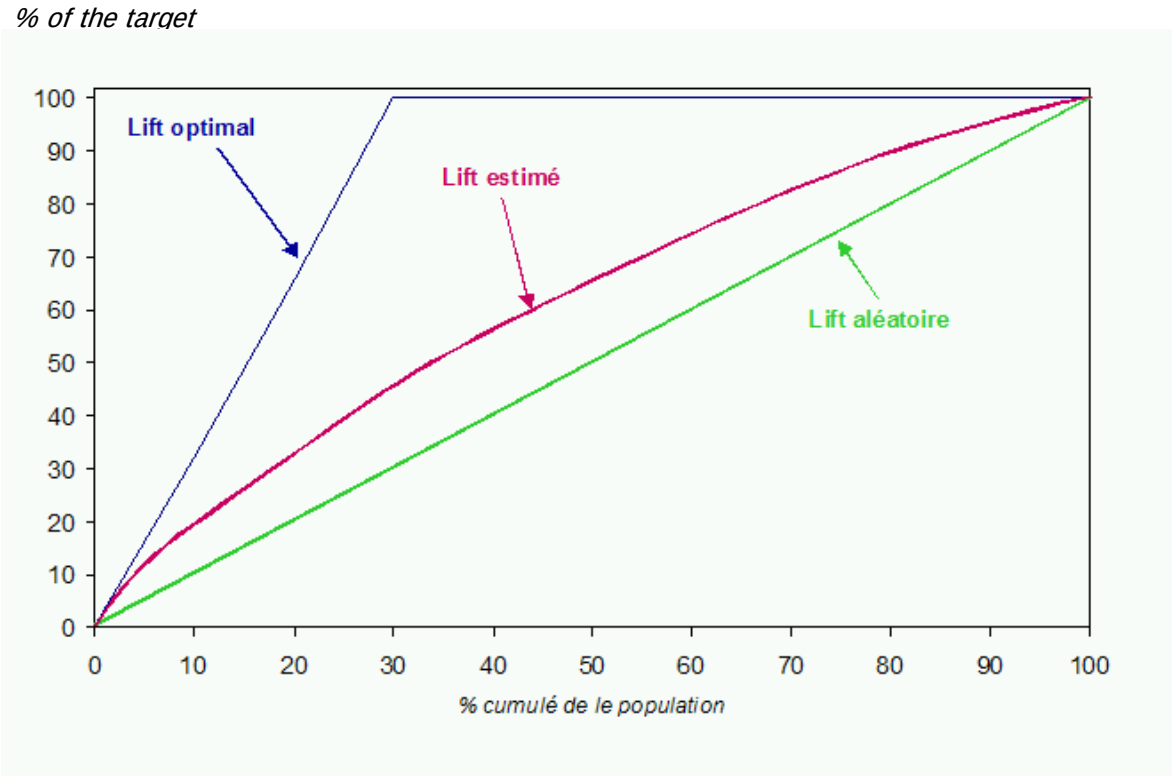
  On tire une obs de G_1 et une de G_2

  - AUC estimée par la proportion de paires concordantes

    $$c = n_c / n_1 n_2$$

- n_c statistique de Wilcoxon-Mann-Whitney

Lift chart


% of the target
100
90
80
70
60
50
40
30
20
10
0
Lift optimal
Lift estimé
Lift aléatoire
0
10
20
30
40
50
60
70
80
90
100
% cumulé de le population

Surface sous la courbe de lift

- Proportion des unités ayant un score>s

$$p_1(1-\beta)+(1-p_1)\alpha$$

- Surface:

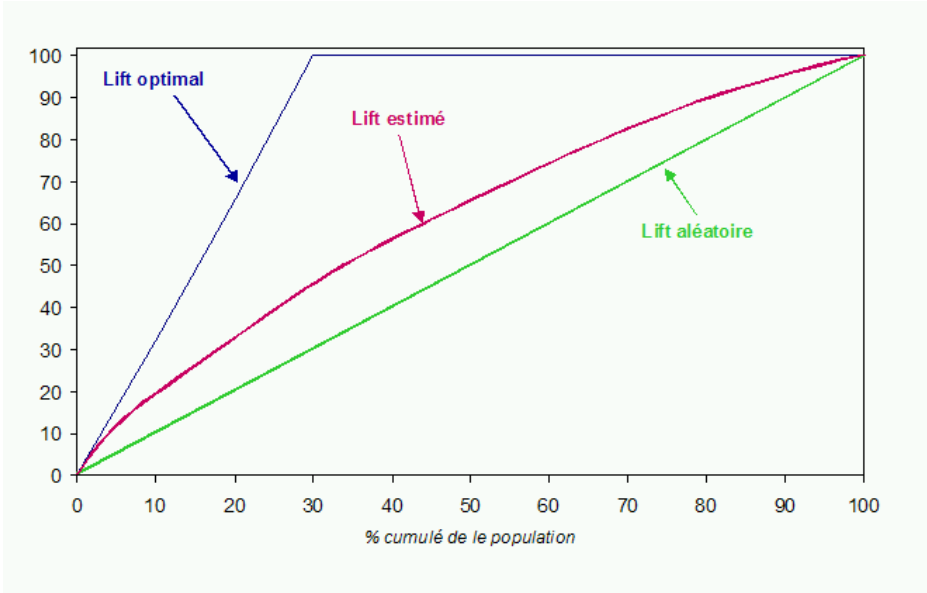
$$L=\int(1-\beta)d\left\{p_1(1-\beta)+(1-p_1)\alpha\right\}=$$

$$\left[p_1\int(1-\beta)d(1-\beta)\right]+\left[(1-p_1)\int(1-\beta)d\alpha\right]$$

$$=\frac{p_1}{2}+(1-p_1)AUC$$

Coefficient Ki (Kxen)

- Ki=(aire entre courbe lift et diagonale) / (aire entre courbe lift et courbe idéale)

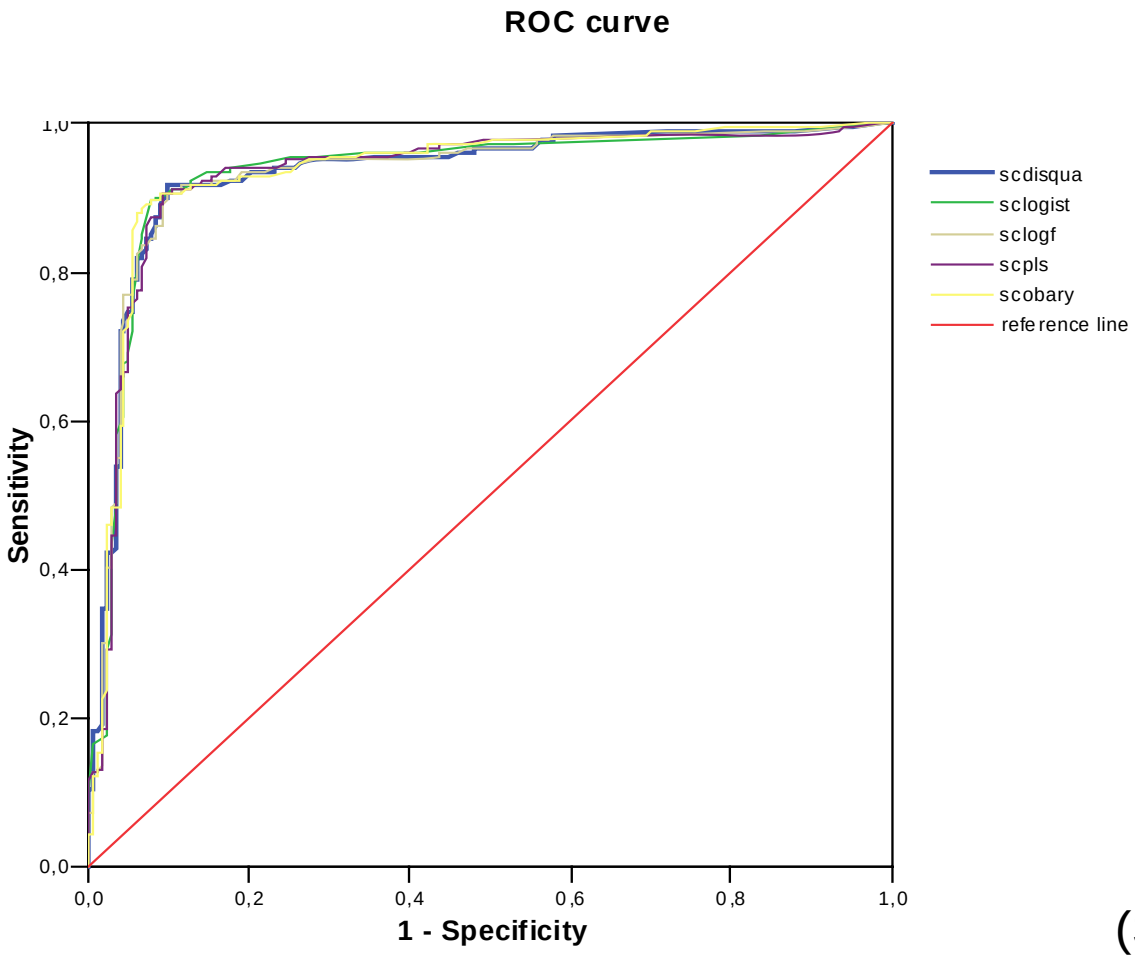


$$Ki = \frac{L - \frac{1}{2}}{\frac{1 - p_1}{2}} = \frac{p_1 + 2(1 - p_1)AUC - 1}{1 - p_1} = 2AUC - 1$$

Ki=Somers' *D* ou Accuracy Ratio *AR*

- Optimiser AUC ou Ki: mêmes résultats.
- Ne prend pas en compte les coûts
- Ne pas comparer sur données d’apprentissage...

5.5 Quelques résultats expérimentaux



AUC

Score	AUC
scdisqua	.934
sclogist	.933
sclogf	.932
scpls	.933
scobary	.935

(Saporta, Niang,2003)

Baesens (2003) 17 techniques sur 8 jeux de données

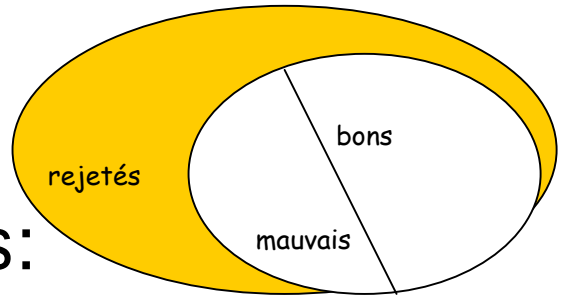
Table 4 Test set AUC on credit scoring data sets

Technique	*Bene1*	*Bene2*	*Germ*	*Austr*	*UK1*	*UK2*	*UK3*	*UK4*	*AR*
LDA	**77.1**	*77.1*	**78.4**	**92.8**	*64.1*	*73.6*	**74.4**	**72.3**	**5.38**
QDA	*73.4*	*72.4*	*71.8*	**91.5**	*63.3*	*72.1*	*68.1*	*68.3*	*10.8*
LOG	**77.0**	*78.0*	**77.7**	**93.2**	*63.9*	*73.0*	**74.6**	**72.7**	**4.38**
LP	*76.1*	*77.5*	**76.3**	**92.6**	*56.4*	*62.3*	*62.0*	*62.2*	*11.9*
RBF LS-SVM	**77.6**	77.8	**77.4**	**93.2**	65.0	74.7	**72.9**	**73.1**	**3.38**
Lin LS-SVM	**76.9**	*77.1*	**78.4**	**92.9**	*64.4*	*73.7*	**73.8**	**72.5**	**5.50**
RBF SVM	*76.7*	*77.1*	**77.2**	**92.6**	*59.3*	*65.4*	*67.3*	*63.4*	9.13
Lin SVM	*75.9*	*77.5*	**76.6**	**93.6**	*56.4*	*63.9*	*62.9*	*62.9*	10.1
NN	**76.9**	**79.1**	**78.7**	**91.7**	**66.4**	**75.8**	**74.6**	**72.9**	**3.25**
NB	**76.5**	*70.6*	**77.2**	**93.1**	**65.8**	*73.7*	*66.9*	67.9	7.88
TAN	75.5	**78.2**	**78.3**	**93.4**	**66.8**	74.5	*64.0*	*66.6*	**5.63**
C4.5	*72.2*	*71.1*	**74.7**	91.6	*56.1*	*65.7*	*50.0*	*49.9*	*14.7*
C4.5rules	*71.6*	*74.2*	*62.0*	*85.3*	*61.7*	*70.4*	*60.3*	68.4	*13.0*
C4.5dis	*73.0*	*73.2*	**74.6**	**93.1**	*50.0*	*50.0*	*50.4*	*49.9*	*13.7*
C4.5rules dis	*73.0*	*71.5*	*64.4*	93.1	65.2	*71.5*	*66.7*	*64.9*	10.8
KNN10	*71.7*	*69.6*	*70.2*	91.4	*58.9*	*65.4*	*63.0*	67.0	*14.1*
KNN100	*74.9*	*71.5*	**76.1**	**93.0**	*62.8*	*69.9*	*70.0*	**70.4**	9.5

« However, it has to be noted that simple, linear classifiers such as LDA and LOG also gave very good performances, which clearly indicate that most credit scoring data sets are only weakly non-linear ».

6. Reject inference

- Analyse sur prêts approuvés :biais de sélection



- Techniques empiriques:
  - Définir les refus comme des mauvais
  - Extrapolation
  - Augmentation ou repondération
- Modèles probabilistes
  - Estimation de données manquantes (EM)
  - Probit bivariés
  - Tobit

- On peut construire un modèle non biaisé dans certains cas si les variables de refus X_1 sont un sous-ensemble des variables de score X :
- Si $X_1 \not\subset X$ c'est impossible

Peu d'évaluations publiées

The scope for improved predictive performance by any form of reject inference is modest . Reject inference in the form of re-weighting applicants within a training sample of accepted cases and adopting a cut-off point based on those accepted cases appears to perform no better than unweighted estimation. In fact where the rejection rate is high, results appear to be quite noticeably worse. Reject inference in the form of extrapolation appears to be both useless and harmless. (Crook, Banasik 2002)

Many methods have been used for tackling this problem. Most of those used in practice are demonstrably ineffective. The best strategies are to build a formal sample selection model to supplement the classification model, and to obtain data about the rejected applicants. This can come from a small sample of people who would normally be rejected (this is done in mail order) or from other sources, such as other supplier (Hand 2005)

7. Vers des scores de durée

- Pas seulement « si » mais « quand » arrive la défaillance
  - Traitement de données censurées. Permet de résoudre le problème des données incomplètes pour les prêts à long terme
  - Utile pour la les calculs de lifetime value et de LGD (Bâle II)
  - Stepanova, Thomas, 2001: modèle de Cox

$$h(t, \mathbf{x}) = h_0(t) \exp(\sum_{j=1}^{p} \beta_j x_j)$$

  - Nécessité d'autres modèles

Conclusions et perspectives

- Credit scoring: toujours un domaine très actif et générateur d'emplois
- Encore de la place pour des recherches
- Discriminante et logistique efficaces mêmes comparées à de nouvelles méthodes
- Mais gains de précision illusoires
  - Si données de mauvaise qualité
  - Si les populations changent

Références

- Baesens: « Developing intelligent systems for credit scoring using machine learning techniques » Ph.D, Leuven, 2003
- Bardos: « Analyse discriminante », Dunod, 2001
- Hastie, Tibshirani, Friedman : « The Elements of Statistical Learning», Springer-Verlag, 2001
- Mays ed. « Handbook of credit scoring » Glenlake, 2001
- Thomas, Edelman,Crook: « Credit scoring and its applications », SIAM, 2002
- Credit Research Center http://www.crc.man.ed.ac.uk
- http://www.defaultrisk.com/
- Basel Committee publications: http://www.bis.org/bcbs/publ.htm