

Analyse discriminante, classification supervisée, scoring...

Gilbert Saporta
Conservatoire National des Arts et Métiers

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

Version du 18/10/2013

Bibliographie

- Bardos: « Analyse discriminante », Dunod, 2001
- Hastie, Tibshirani, Friedman : « The Elements of Statistical Learning », 2nd edition, Springer-Verlag, 2009
http://www.stanford.edu/~hastie/local.ftp/Springer/ESLII_print10.pdf
- Nakache, Confais: « Statistique explicative appliquée », Technip, 2003
- Thiria et al. : « Statistique et méthodes neuronales » Dunod, 1997
- Thomas, Edelman, Crook: « Credit scoring and its applications », SIAM, 2002
- Tufféry: « Data Mining et statistique décisionnelle », 4ème édition, Technip, 2012
- Tufféry: « Étude de cas en statistique décisionnelle », Technip, 2009
- Vapnik : « Statistical Learning Theory », Wiley 1998


Plan

- I L'analyse factorielle discriminante
- II Discrimination sur variables qualitatives :
le scoring.
- III Analyse discriminante probabiliste
- IV Régression logistique
- V SVM
- VI Validation
- VII Choix de modèles et théorie de l'apprentissage
statistique
- VIII Arbres de décision

Objet d'étude

- Observations multidimensionnelles réparties en k groupes définis a priori.
- Autre terminologie: classification supervisée
- *Exemples d'application :*
 - Pronostic des infarctus (J.P. Nakache)
 - *2 groupes : décès, survie (variables médicales)*
 - Iris de Fisher :
 - *3 espèces : 4 variables (longueur et largeur des pétales et sépales)*
 - Risque des demandeurs de crédit
 - *2 groupes : bons, mauvais (variables qualitatives)*
 - Autres :
 - *Météo, publipostage, reclassement dans une typologie.*

Quelques dates :



■ P.C. Mahalanobis	1927
■ H. Hotelling	1931
■ R. A. Fisher	1936
■ J. Berkson	1944
■ C.R. Rao	1950
■ T.W. Anderson	1951
■ D. Mc Fadden	1973
■ V. Vapnik	1998

Objectifs



Y variable à expliquer qualitative à k catégories

X_1, X_2, \dots, X_p variables explicatives

■ Objectif 1 : Décrire

- Étude de la distribution des X_i / Y
- Géométrie : Analyse factorielle discriminante AFD
- Tests : Analyse de variance multidimensionnelle MANOVA

■ Objectif 2 : Classer

- Étude de $P(Y / X_1, X_2, \dots, X_p)$
- Modélisation fonctionnelle : Approche bayésienne
- Modélisation logique : Arbre de décision
- Méthodes géométriques.

1^{ère} partie : L'analyse factorielle discriminante



1. Réduction de dimension, axes et variables discriminantes.
2. Cas de 2 groupes.
3. Méthodes géométriques de classement.

Représentation des données

■ 2 cas :

- prédicteurs numériques
- prédicteurs qualitatifs

	1	2	...	k	1	2	j	p
1	$\overline{0}$	1	...	$\overline{0}$	$\overline{X_1^1}$	X_1^2	X_1^j	$\overline{X_1^p}$
2	1	0	...	0				
			...					
i	0	0	...	1	X_i^1	X_i^2	X_i^j	X_i^p
			...					
n	$\underline{1}$	0	...	$\underline{0}$	$\underline{X_n^1}$	X_n^2	X_n^j	$\underline{X_n^p}$

indicatrices des groupes

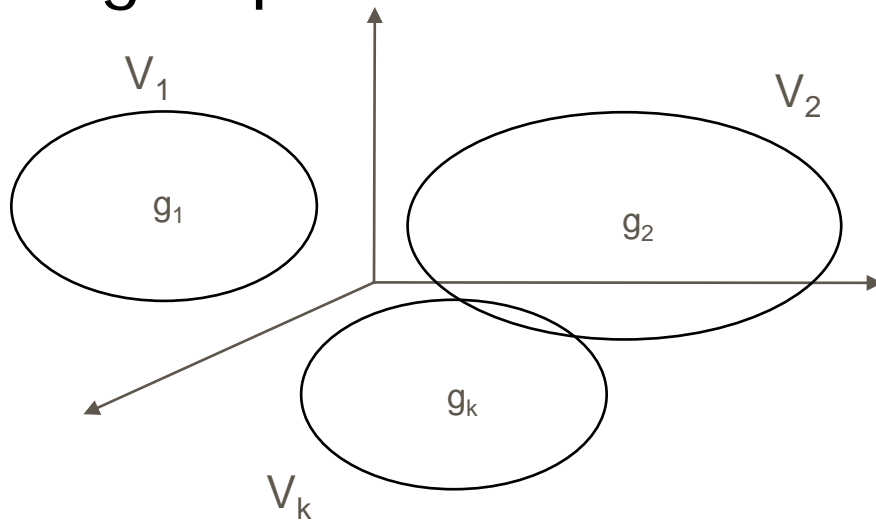
variables explicatives

- n points dans \mathbb{R}^p appartenant à k groupes.

I.1 Réduction de dimension.

Recherche d'axes et de variables discriminantes.

- Dispersion intergroupe et dispersion intra groupe.



W = matrice variance intra

- $W = 1/n \sum n_i V_i$

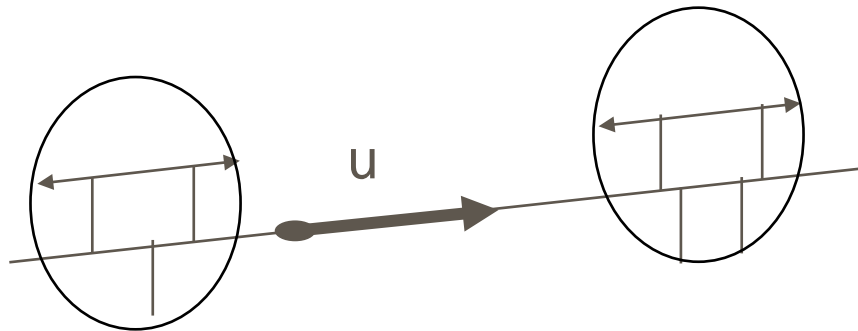
B = matrice variance inter

- $B = 1/n \sum n_i (g_i - g) (g_i - g)'$

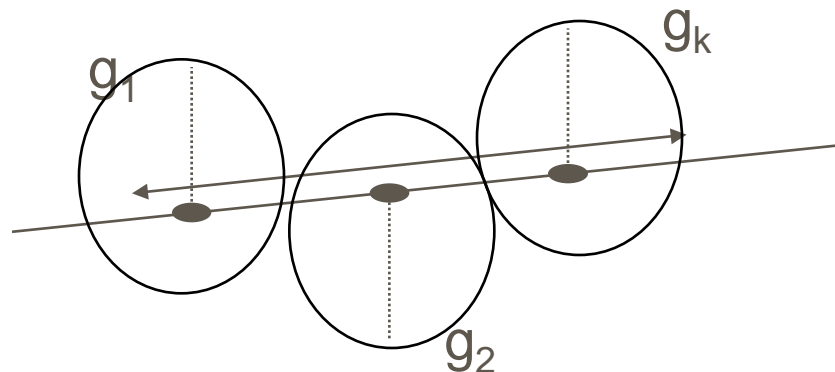
$V = W + B$ variance totale

Axes discriminants : deux objectifs

- Dispersion intraclasse minimale : $\min u'Wu$



- Dispersion interclasse maximale : $\max u'Bu$



Axes discriminants : deux objectifs

■ Simultanéité impossible

$$\min u' W u \Rightarrow W u = \alpha u \quad \alpha \min i$$

$$\max u' B u \Rightarrow B u = \beta u \quad \beta \max i$$

$$V = W + B$$

■ Compromis : $u' V u = u' W u + u' B u$

$$\begin{array}{ccc} \min & & \max \\ \max & \left(\frac{u' B u}{u' V u} \right) & \text{ou} & \left(\frac{u' B u}{u' W u} \right) \\ & V^{-1} B u = \lambda u & & W^{-1} B u = \mu u \end{array}$$

Axes discriminants : deux objectifs

$$\text{a) } V^{-1} B u = \lambda u$$

$$B u = \lambda V u$$

$$B u = \lambda (W + B) u$$

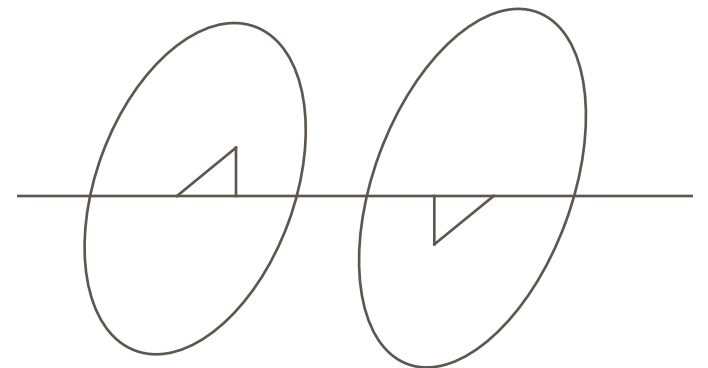
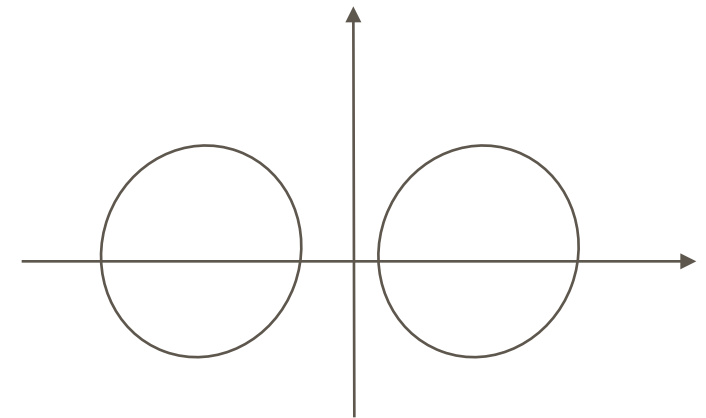
$$(1 - \lambda) B u = \lambda W u$$

$$\text{b) } W^{-1} B u = \frac{\lambda}{1 - \lambda} u = \mu u$$

■ ACP du nuage des g_i avec :

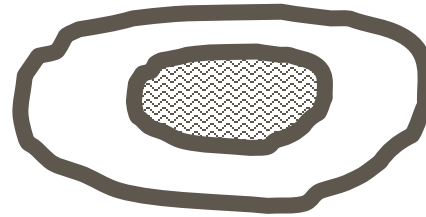
■ Métrique V^{-1}

■ Métrique W^{-1} Mahalanobis

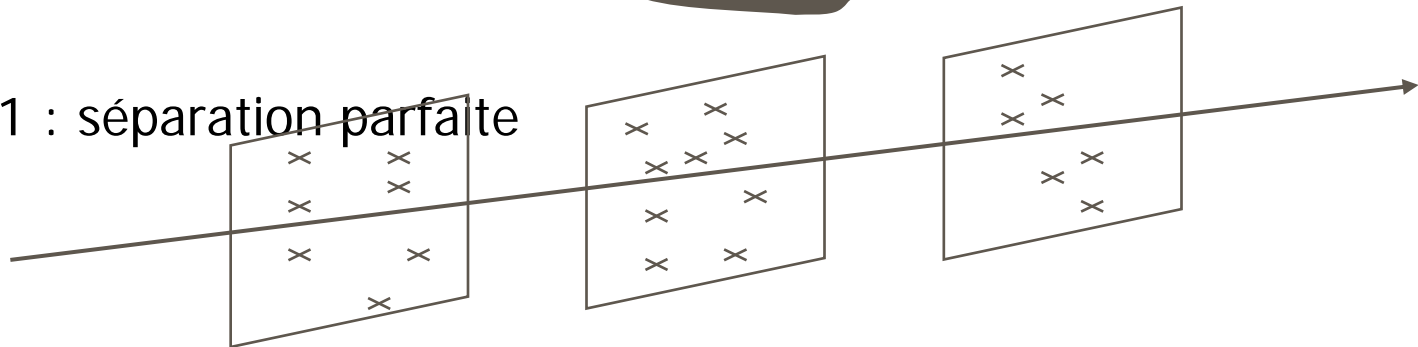


Les différents cas selon λ_1

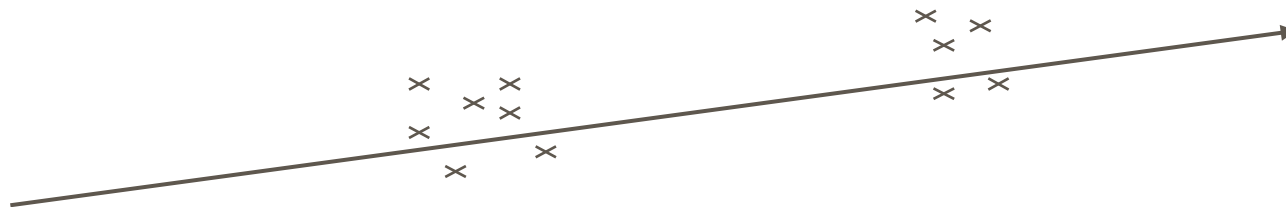
1. $\lambda_1 = 0$: aucune séparation linéaire n'est possible, groupes concentriques



2. $\lambda_1 = 1$: séparation parfaite



3. Mais $0 < \lambda_1 < 1$: séparation possible avec groupes non recouvrants



Nombre d'axes discriminants

- ACP des groupes : dimension de l'espace contenant les centres des groupes g_i
- Si $n > p > k$ (cas fréquent), $k-1$ axes discriminants

Exemple célèbre : Iris de Fisher

- $K = 3$ *Setosa, Versicolor, Virginica*
- $P=4$ *longueur pétale, longueur sépale, largeur pétale, largeur sépale*
- $n_1 = n_2 = n_3 = 50$

Donc deux axes

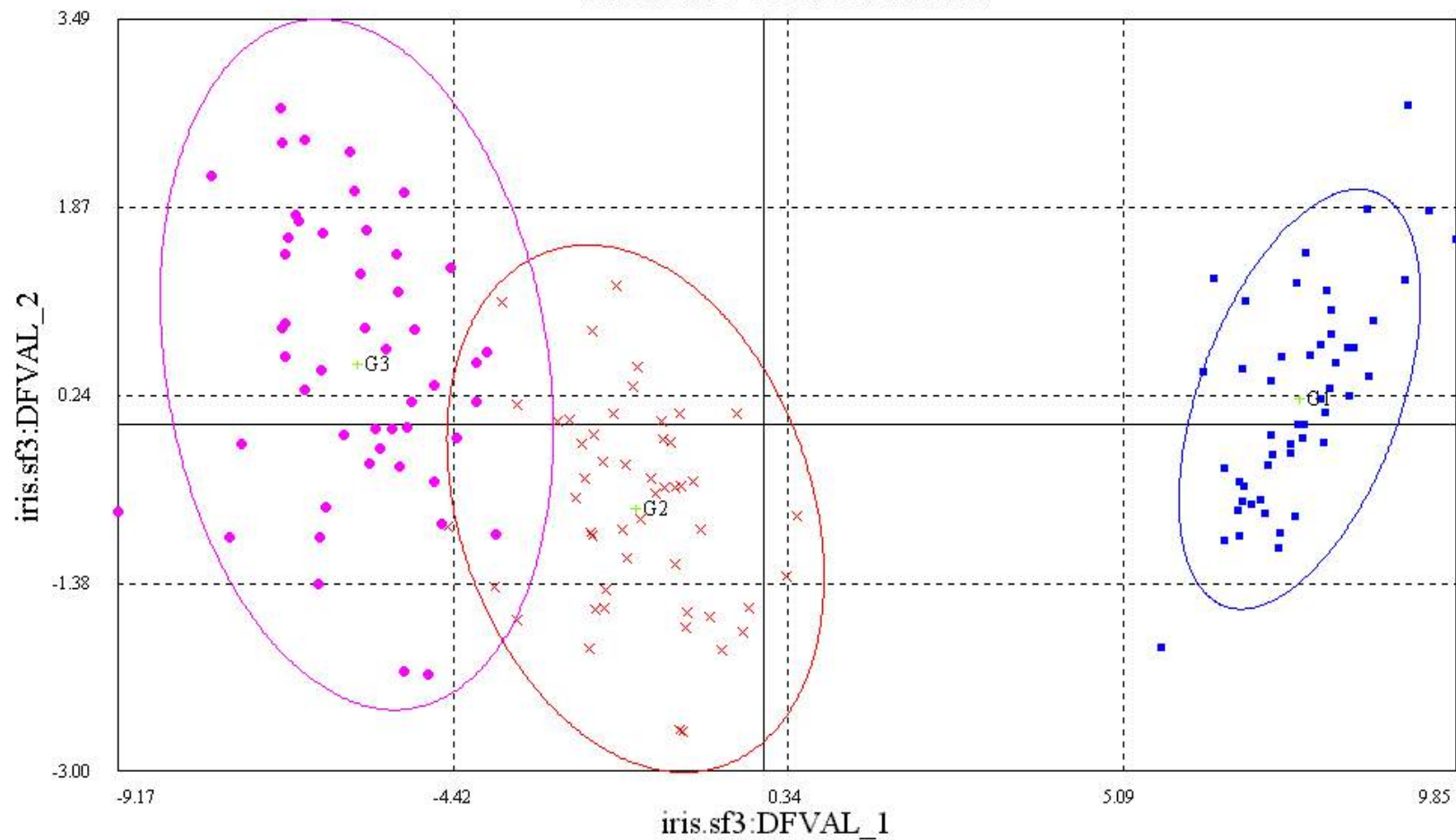


Iris setosa

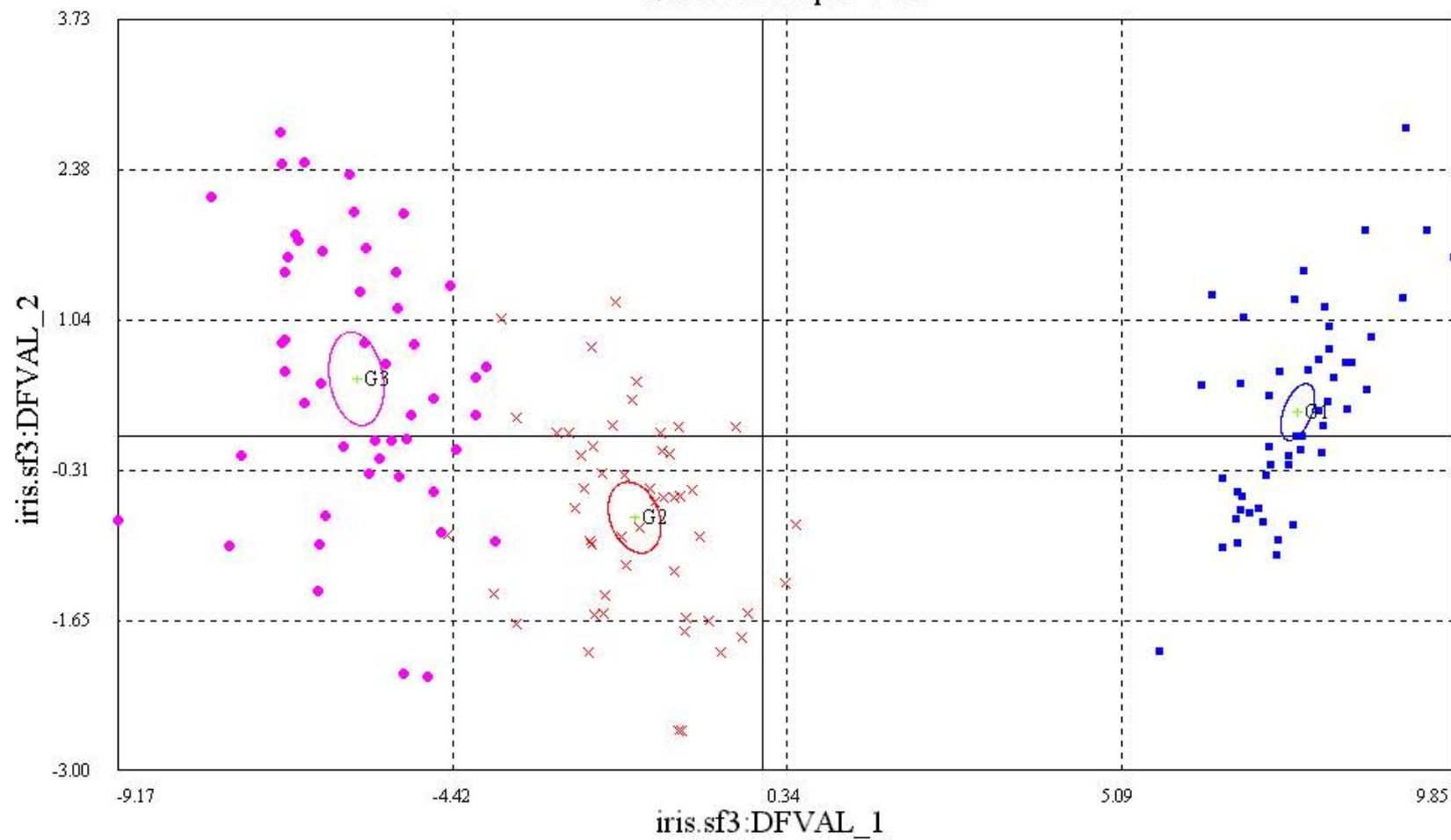
Iris versicolor

Iris virginica

Ellipses de tolérance
contenant 95 % des observations



Ellipses de confiance
Niveau de risque: 5 %.



Distance de MAHALANOBIS



Distance au sens de la métrique W^{-1} .

$$D_p^2 = (g_1 - g_2)' W^{-1} (g_1 - g_2)$$

1. pour $p=1$:
$$\frac{n_1 n_2}{n_1 + n_2} \left(\frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}} \right)^2 = \frac{n_1 n_2}{n_1 + n_2} D_1^2 \sim F(1; n_1 + n_2 - 2)$$

2. p quelconque :

$$D_p^2 = (g_1 - g_2)' W^{-1} (g_1 - g_2)$$
$$D_p^2 = (g_1 - g_2)' W^{-1/2} \underbrace{W^{-1/2} (g_1 - g_2)}_{W^{-1/2} X}$$

- *Standardisation de chaque composante x_j*
- *Décorrélation...*

Interprétation probabiliste

Le Δ^2 théorique : $\Delta_p^2 = (\underline{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} (\underline{\mu}_1 - \underline{\mu}_2)$

2 populations $N_p (\underline{\mu}_1, \Sigma)$ et $N_p (\underline{\mu}_2, \Sigma)$

D_p^2 estimation (biaisée) de Δ_p^2

$$W = \frac{n_1 V_1 + n_2 V_2}{n - 2} = \hat{\Sigma}$$

$$D_p^2 = (\underline{g}_1 - \underline{g}_2)' W^{-1} (\underline{g}_1 - \underline{g}_2)$$

Interprétation probabiliste

$$E(D_p^2) = \frac{n-2}{n-p-1} \left(\Delta_p^2 + \frac{pn}{n_1 n_2} \right)$$

$$\text{Si } \Delta^2 = 0 \quad \underline{\mu}_1 = \underline{\mu}_2$$

$$\frac{n_1 n_2}{n} \frac{n-p-1}{p(n-2)} D_p^2 \sim F(p; n-p-1)$$

Distances de Mahalanobis entre 2 groupes parmi k

■ Théoriques : $\Delta_p^2 = (\mu_i - \mu_j)' \Sigma^{-1} (\mu_i - \mu_j)$

■ Estimées : $D_p^2 = (\underline{g}_i - \underline{g}_j)' \left(\frac{n}{n-k} W \right)^{-1} (\underline{g}_i - \underline{g}_j)$

Si $\Delta^2 = 0$

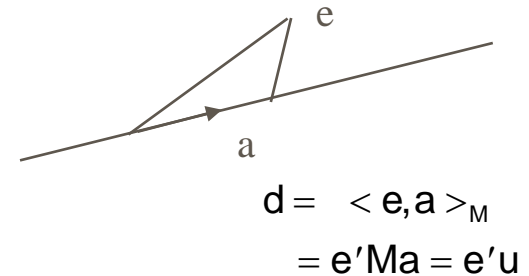
$$\frac{n_i n_j}{n_i + n_j} \cdot \frac{n-k-p+1}{(n-k)p} \cdot D_p^2 = F(p; n-k-p+1)$$

I.2 Cas de deux groupes

- g_1 et g_2 sont sur une droite : 1 seul axe discriminant :

$$a = \alpha (g_1 - g_2)$$

- RAPPEL : en ACP axe a , facteur $u = M a$



- Combinaison discriminante proportionnelle à

$$M (g_2 - g_1) = W^{-1} (g_2 - g_1) \text{ ou } V^{-1} (g_2 - g_1)$$

- FONCTION DE FISHER :

$$W^{-1} (g_2 - g_1) = W^{-1} \begin{pmatrix} \bar{X}_2^1 - \bar{X}_1^1 \\ \bar{X}_2^p - \bar{X}_1^p \end{pmatrix}$$

Historique

Historiquement : $d = \sum_{j=1}^p u_j x^j = X u$

Test (de Student) de comparaison de 2 moyennes : $T = \frac{\bar{d}_1 - \bar{d}_2}{S_d}$

Fisher (1936)

Trouver u_1, u_2, \dots, u_p tel que T maximal.

Solution : u proportionnel à $W^{-1}(g_1 - g_2)$

Nota : $W^{-1}(g_1 - g_2) = \alpha V^{-1}(g_1 - g_2)$ avec : $\alpha = 1 + \frac{n_1 n_2}{n(n-2)} D_p^2$

■ Une régression « incorrecte »

■ \mathbf{y} à 2 valeurs (-1; +1) ou (0;1) ou (a;b)

■ $a=n/n_1$ $b=-n/n_2$

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

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

■ D_p distance de Mahalanobis entre groupes

■ Incompréhensions et controverses!

Modèle linéaire usuel non valide : $y / \mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}; \sigma^2 \mathbf{I})$
en discriminante c'est l'inverse que l'on
suppose :

$$\mathbf{X} / y = j \sim N_p(\boldsymbol{\mu}_j; \boldsymbol{\Sigma})$$

Conséquences

- *Pas de test,*
- *pas d'erreurs standard sur les coefficients*
- MAIS possibilité d'utiliser les méthodes de pas à pas en régression.

FONCTION LINEAIRE DISCRIMINANTE

VARIABLES NUM LIBELLES	CORRELATIONS VARIABLES AVEC F.L.D. (SEUIL= 0.20)	COEFFICIENTS FONCTION REGRESSION DISC.		ECARTS TYPES (RES. TYPE REG.)	T STUDENT REG.)	PROBA
3 FRCAR	0.232	0.0588	0.0133	0.0092	1.44	0.154
4 INCAR	-0.697	-6.1539	-1.3887	0.4966	2.80	0.006
5 INSYS	-0.673	0.1668	0.0376	0.0374	1.01	0.317
6 PRDIA	0.474	-0.0203	-0.0046	0.0351	0.13	0.897
7 PAPUL	0.431	0.1650	0.0372	0.0271	1.37	0.173
8 PVENT	0.269	0.0469	0.0106	0.0176	0.60	0.549
9 REPUL	0.650	-0.0002	0.0000	0.0002	0.19	0.849
CONSTANTE		-1.604374	-0.367565	0.9373	0.3922	0.6958
.....						
R2 =	0.55759	F =	16.74489	PROBA =	0.000	
D2 =	4.94213	T2 =	124.77643	PROBA =	0.000	
.....						

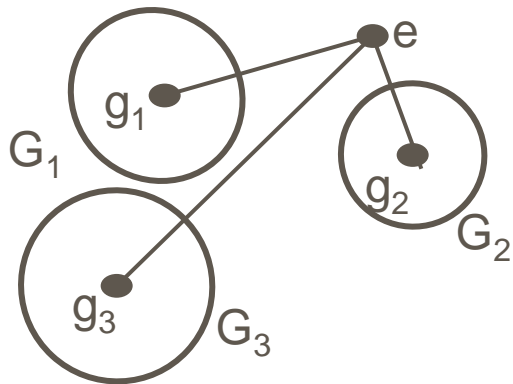
I-3 Méthodes géométriques de classement

y	x ₁
1	
1	
2	
.	
.	
.	
1	

■ Échantillon d'apprentissage

e	
?	

■ e observation de groupe inconnu



■ e classé dans le groupe i tel que:
 $d(e ; g_i)$ minimal

Utilisation des fonctions discriminantes

$$d^2(e; g_i) = (e - g_i)'W^{-1}(e - g_i) = e'W^{-1}e - 2g_i'W^{-1}e + g_i'W^{-1}g_i$$

$$\min d^2(e; g_i) = \max \left(2g_i'W^{-1}e - \underbrace{g_i'W^{-1}g_i}_{\alpha_i} \right)$$

k groupes \Rightarrow k fonctions discriminantes

	1	2	k
	α_1	α_2		α_k
X^1	β_{11}	β_{21}		β_{k1}
X^2				
X^p	β_{1p}	β_{2p}		β_{kp}

- On classe dans le groupe pour lequel la fonction est maximale.



Linear Discriminant Function for Species

		Setosa	Versicolor	Virginica
Constant		-85.20986	-71.75400	-103.26971
SepalLength	Sepal Length in mm.	2.35442	1.56982	1.24458
SepalWidth	Sepal Width in mm.	2.35879	0.70725	0.36853
PetalLength	Petal Length in mm.	-1.64306	0.52115	1.27665
PetalWidth	Petal Width in mm.	-1.73984	0.64342	2.10791

Number of Observations Classified into Species

From Species	Setosa	Versicolor	Virginica	Total
Setosa	50	0	0	50
Versicolor	0	48	2	50
Virginica	0	1	49	50
Total	50	49	51	150
Priors	0.33333	0.33333	0.33333	

pour deux groupes

- On classe dans G_1 si:

$$2g_1'W^{-1}e - g_1'W^{-1}g_1 > 2g_2'W^{-1}e - g_2'W^{-1}g_2$$

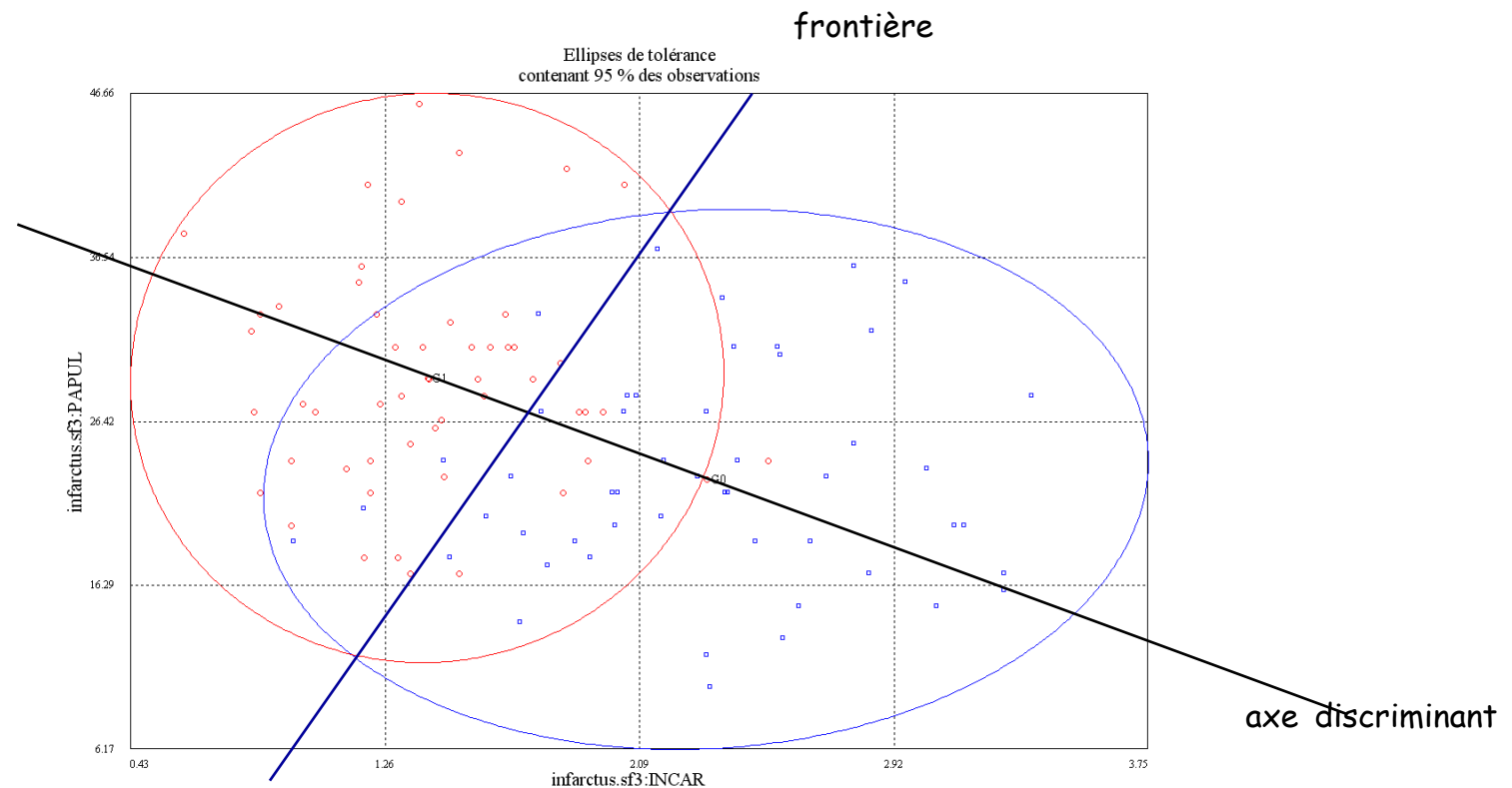
$$(g_1 - g_2)'W^{-1}e > \frac{1}{2}(g_1'W^{-1}g_1 - g_2'W^{-1}g_2)$$

- Fonction de Fisher $> c$

- Score de Fisher: $(g_1 - g_2)'W^{-1}e - \frac{1}{2}(g_1'W^{-1}g_1 - g_2'W^{-1}g_2)$

Interprétation géométrique

- Projection sur la droite des centres avec la métrique W^{-1}
- Dualité axe-frontière plane



Règle de classement des plus proches voisins

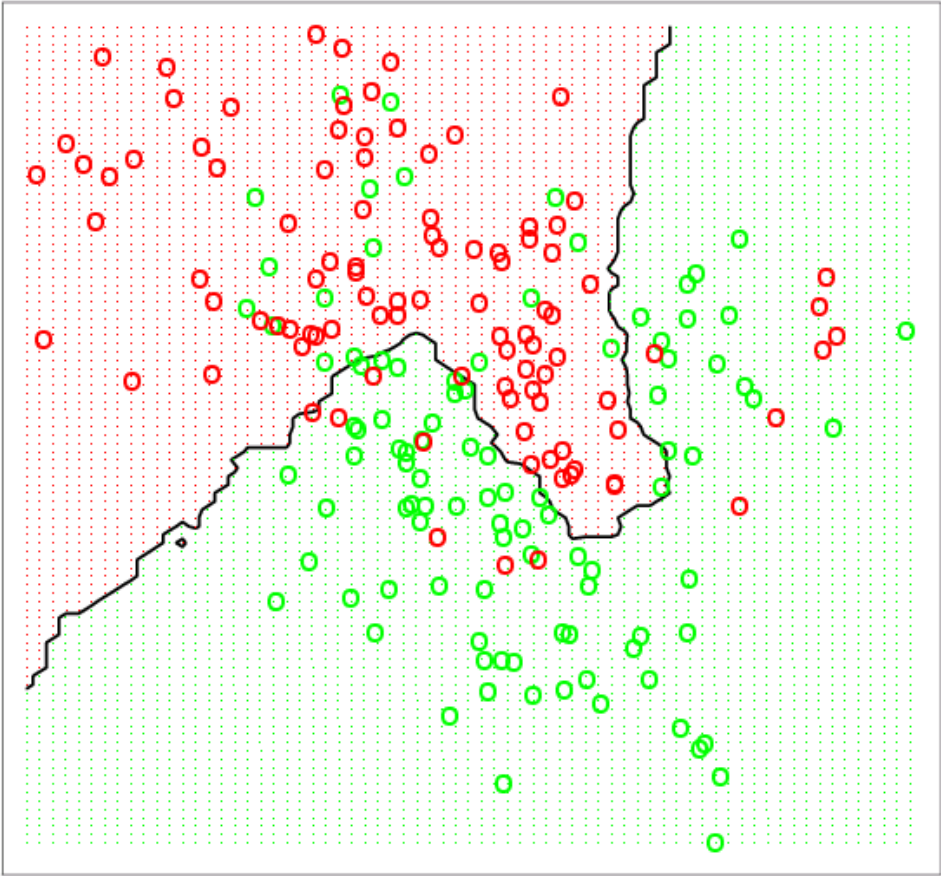
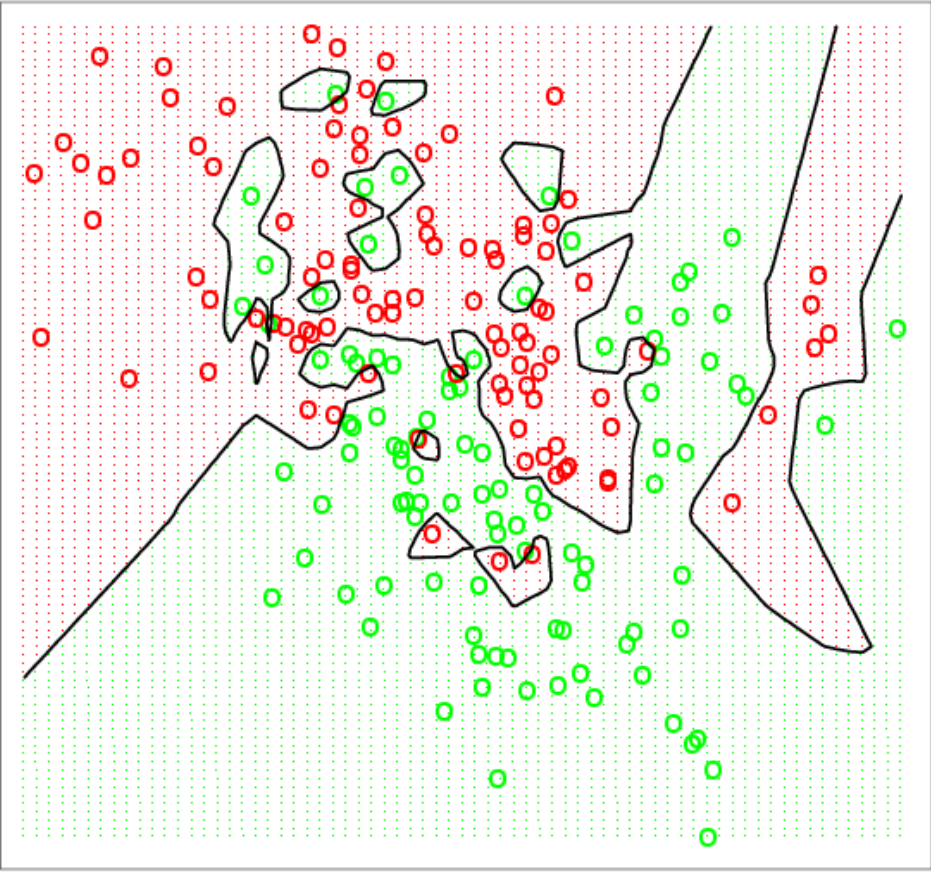
- On compte le nombre d'observations de G_1 , G_2 , ... parmi les k plus proches voisins et on classe dans le groupe le plus fréquent.
- Cas limite $k = 1$

Méthode des plus proches voisins (Hastie and al)

1-Nearest Neighbor Classifier



15-Nearest Neighbor Classifier



BANQUE DE FRANCE

DIRECTION GÉNÉRALE DES ÉTUDES

Direction de la Conjoncture

et de la

Centrale de Bilans

L'ANALYSE DES DÉFAILLANCES D'ENTREPRISES

Rapport présenté à la
IX^e Journée d'Étude des Centrales de Bilans
le 16 Juin 1983

par : Sylvie GHESQUIERE
Bernard MICHA

Informatique : Michel PACHOT

$$100Z = -1,255 R_1 + 2,003 R_2 - 0,824 R_3 + 5,221 R_4 - 0,689 R_5 \\ - 1,164 R_6 + 0,706 R_7 + 1,408 R_8 - 85,544$$

Avec :

R_1 : Part des frais financiers dans les résultats	%	$\frac{\text{Frais financiers}}{\text{Résultat économique brut}}$
R_2 : Couverture des capitaux investis	%	$\frac{\text{Ressources stables}}{\text{Capitaux investis}}$
R_3 : Capacité de "remboursement"	%	$\frac{\text{Capacité d'autofinancement}}{\text{Endettement}}$
R_4 : Taux de marge brute d'exploitation	%	$\frac{\text{Résultat économique brut}}{\text{Chiffre d'affaires HT}}$
R_5 : Délai crédit fournisseurs	jours	$\frac{\text{Dettes commerciales}}{\text{Achats TTC}}$
R_6 : Taux de variation de la valeur ajoutée	%	$\frac{\text{Stocks de travaux en cours} - \text{Avances clients} + \text{Créances d'exploitation}}{\text{Production}}$
R_7 : Délai découvert clients	jours	$\frac{\text{Investissements physiques}}{\text{Valeur ajoutée}}$
R_8 : Taux d'investissements physiques Calculé en moyenne pluriannuelle, 2 ou 3 ans	%	

Lexique comptable de calcul de la fonction Annexe 19

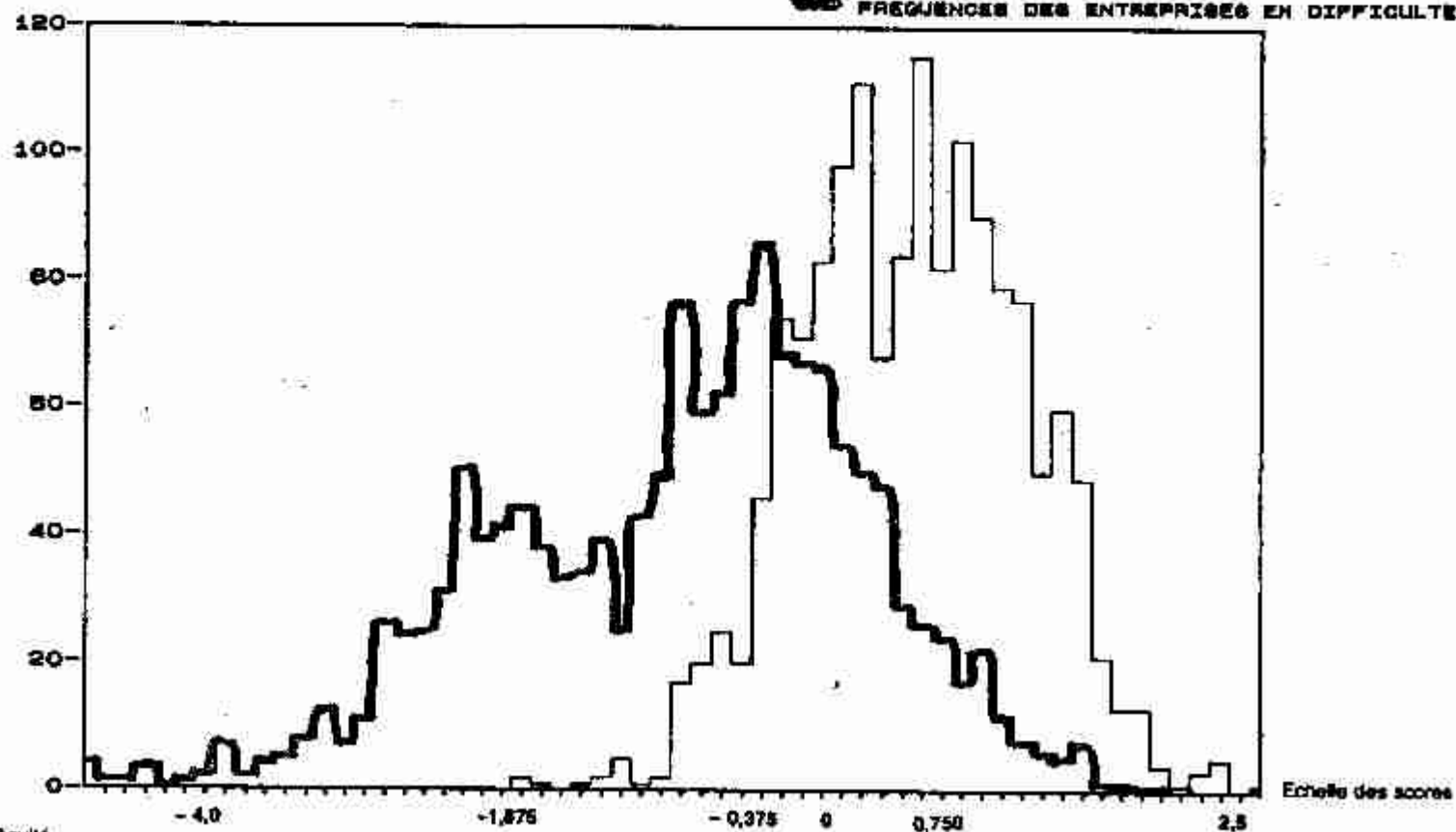
DISTRIBUTION DES ENTREPRISES NORMALES ET EN DIFFICULTE

8

NOMBRE D'ENTREPRISES

— FREQUENCES DES ENTREPRISES NORMALES

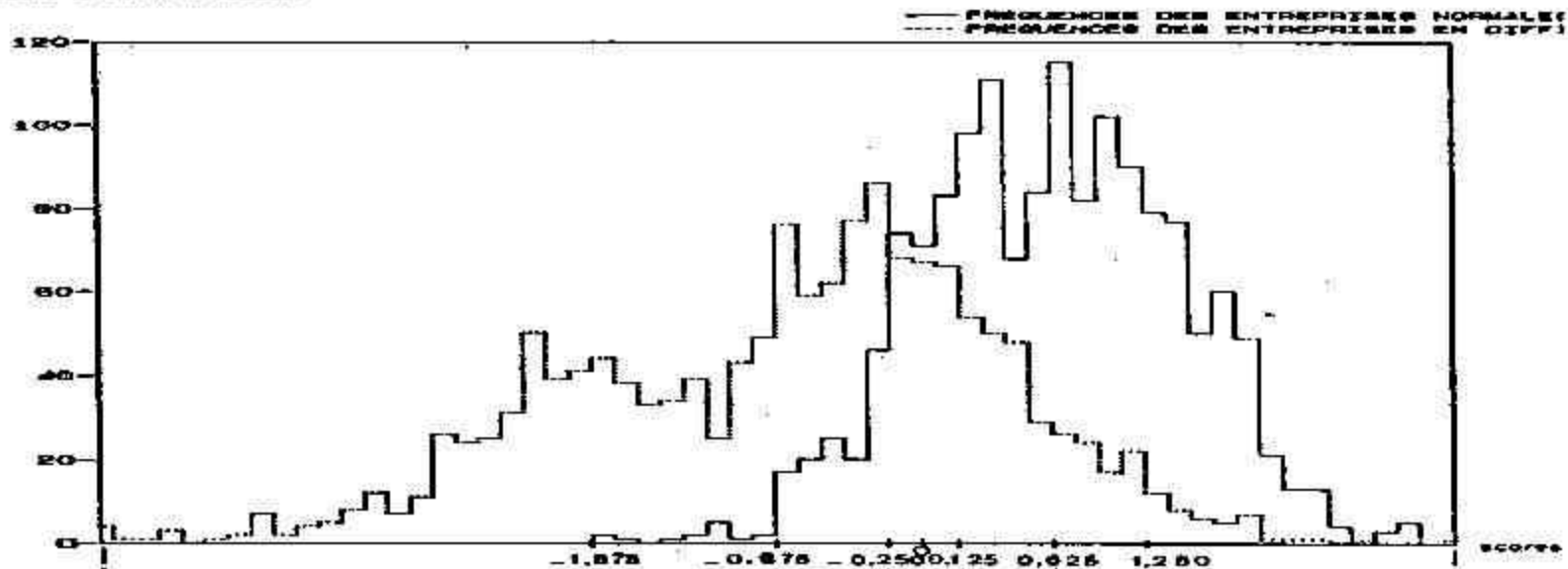
▨ FREQUENCES DES ENTREPRISES EN DIFFICULTE



• 1542 entreprises en difficulté
• 1494 entreprises "normales"

**DISTRIBUTION DES ENTREPRISES
NORMALES ET EN DIFFICULTE**

NUMBRE D'ENTREPRISES



	PROBABILITE D'ETRE UNE ENTREPRISE EN DIFFICULTE (en %)							SYST
P ₁	79,4			23,9				2 z0
P ₂	87,2			46,9	21,-			3 z0
P ₃	100,-	95,8	73,8	46,9	33,4	17,7	9,5	7 z0

zone d'incertitude

	PROBABILITE D'ETRE UNE ENTREPRISE NORMALE (en %)							SYSTE
P ₁	20,6			76,1				2 z0r
P ₂	12,8			53,1	79,-			3 z0r
P ₃	0	4,4	26,2	53,1	66,6	82,3	90,5	7 z0r

POURCENTAGE DE BONS CLASSEMENTS									
AVEC LA FONCTION MOYENNE ETABLIE A	Sur les entreprises en difficulté						Moyennes générales * sur les entreprises		
	Période 1974-1976 à			Période 1977-1979 à			en difficulté	normales	ensemble
	3 ans	2 ans	1 an	3 ans	2 ans	1 an			
	de la défaillance								
PERIODE 1974-1976									
3 ans de la défaillance	77,0	82,0	83,2	77,6	79,6	85,2	82,4	67,3	75,0
2 ans de la défaillance	61,2	75,6	87,6	65,9	70,5	75,8	73,0	81,3	77,3
1 an de la défaillance	40,2	52,4	79,6	46,8	53,4	65,9	56,4	95,9	76,6
PERIODE 1977-1979									
3 ans de la défaillance	65,1	72,4	86,8	67,9	72,4	80,3	74,2	80,1	77,3
2 ans de la défaillance	47,9	62,4	84,0	60,6	68,6	73,9	66,2	88,2	77,2
1 an de la défaillance	37,8	49,6	78,0	51,6	61,4	71,6	58,3	92,3	75,3

- * périodes et nombre d'années avant la défaillance confondus,
- soit 1 542 entreprises en difficulté
- soit 1 494 entreprises normales.

Deuxième partie:

Discrimination sur variables qualitatives et scoring

1. Le problème
2. Disqual
3. Les objectifs du credit scoring

II.1 Discrimination sur variables qualitatives

Y variable de groupe

X_1, X_2, \dots, X_p Variables explicatives à m_1, m_2, \dots, m_p modalités

Exemples

- Solvabilité d'emprunteurs auprès de banques

Y : bon payeur
 mauvais payeur

X_1 : sexe, X_2 : catégorie professionnelle etc.

- Risque en assurance automobile

Y : bon conducteur (pas d'accidents)
 mauvais conducteur

X_1 : sexe, X_2 : tranche d'âge, X_3 : véhicule sportif ou non ...

- Reclassement dans une typologie

Y numéro de groupe

Un peu de (pré)histoire

■ Fisher (1940)

- Un seul prédicteur
- Equations de l'AFC
- Introduction du vocable « Scores »

THE PRECISION OF DISCRIMINANT FUNCTIONS *

* See Author's Note, Paper 155.

I. 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					Total
	Fair	Red	Medium	Dark	Black	
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 introduced 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.

Cas de 2 groupes : le 'scoring'

- Deux idées équivalentes :

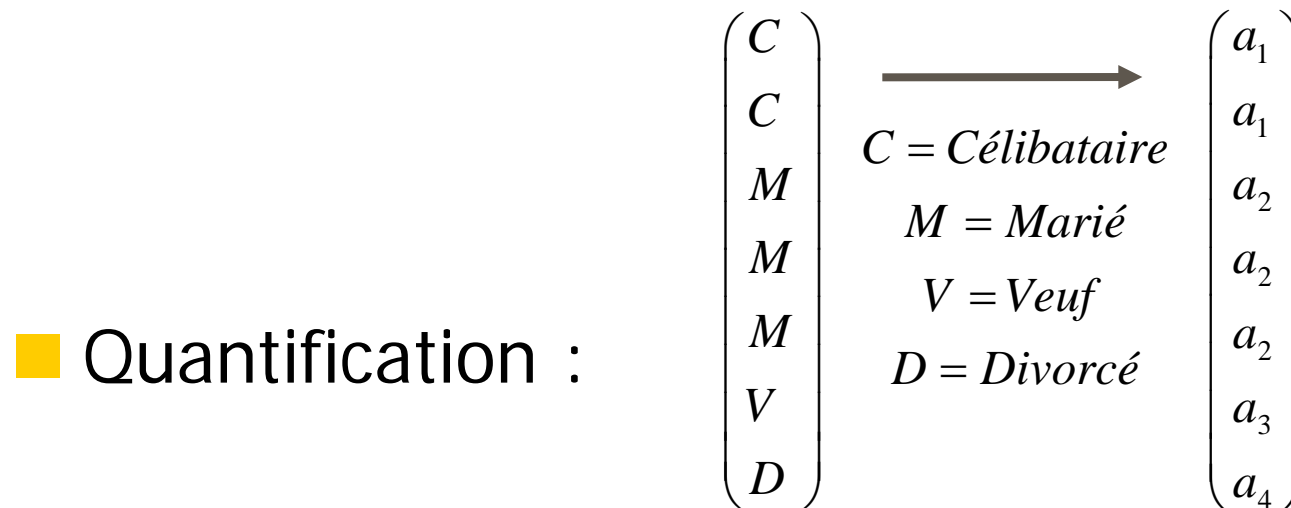
- Transformer les variables qualitatives explicatives en variables quantitatives. Donner des valeurs numériques (notes ou scores) aux modalités de façon optimale: maximiser la distance de Mahalanobis dans \mathbb{R}^p
- Travailler sur le tableau disjonctif des variables explicatives

- *Une réalisation : Passage par l'intermédiaire d'une analyse des correspondances multiples.*

$$\left(\begin{array}{cc|cc|c} X_1 & & X_2 & & \\ \hline 0 & 1 & 1 & 0 & 0 \\ \hline \cdot & & & & \dots \\ \hline \cdot & & & & \\ \hline \cdot & & & & \end{array} \right)$$

VARIABLES explicatives qualitatives

- **Quantification** : Transformer une variable qualitative en une variable numérique et se ramener au cas précédent.
- Exemple : État matrimonial de 7 individus



Quantification

X Tableau disjonctif des variables indicatrices

	C	M	V	D
	1	0	0	0
	1	0	0	0
	0	1	0	0
	0	1	0	0
	0	1	0	0
	0	0	1	0
	0	0	0	1

$$\underline{x} = \begin{pmatrix} a_1 \\ a_1 \\ a_2 \\ a_2 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} = X \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} = X \underline{a}$$

La fonction de Fisher est une combinaison linéaire des variables quantifiées

$$S = \sum_{I=1}^p \alpha_i \tilde{X}_i$$

■ S est une combinaison linéaire des $(m_1 + m_2 + \dots + m_p)$ indicatrices des variables

$$\tilde{X}_i = \sum_{j=1}^{m_i} \beta_j 1_j$$

- X n'est pas de plein rang: $\text{rank}(X) = \sum m_i - p$
- Solution classique: éliminer une indicatrice par prédicteur (GLM , LOGISTIC de SAS)
- **Disqual** (Saporta, 1975):
 - ADL effectuée sur une sélection de facteurs de l'ACM de X. Analogue de la régression sur composantes principales
 - Composantes sélectionnées de manière experte selon inertie et pouvoir discriminant

II.2 DISQUAL

1^{ère} étape

- Analyse des correspondances du tableau des prédicteurs.

$$X = \begin{array}{c} \text{Profession} \\ \text{P}_1 \quad \text{P}_2 \quad \text{P}_3 \quad \text{P}_4 \\ \text{Logement} \\ \text{Prop.} \quad \text{Loc.} \end{array} \begin{array}{c} 1 \\ 2 \\ \vdots \\ \vdots \\ \vdots \\ n \end{array} \left(\begin{array}{cccc|cc} 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ \cdot & & & & \cdot & \\ \cdot & & & & \cdot & \\ \cdot & & & & \cdot & \\ \cdot & & & & \cdot & \end{array} \right) \dots$$

variables indicatrices

$$Z = \begin{array}{c} z^1 \quad \dots \quad z^k \\ 1 \\ 2 \\ \vdots \\ \vdots \\ \vdots \\ n \end{array} \left(\begin{array}{c} | \\ | \\ | \\ | \\ | \\ | \end{array} \right)$$

- *k variables numériques : garder les coordonnées factorielles les plus discriminantes*

2^{ème} étape :

- Analyse discriminante linéaire (Fisher). Score $\mathbf{s} = \sum_{j=1}^k d_j \mathbf{z}^j$
- Score = combinaison linéaire des coordonnées factorielles = combinaison linéaire des indicatrices des catégories
- Coefficients = grille de notation
- $\mathbf{z}^j = \mathbf{X}\mathbf{u}^j$ \mathbf{u}^j : coordonnées des catégories sur l'axe $n^{\circ}j$

$$s = \sum_{j=1}^k d_j X u^j = X \underbrace{\sum_{j=1}^k d_j u^j}_{\text{grille de score}} \quad \begin{pmatrix} \cdot \\ d_j \\ \cdot \end{pmatrix} = \mathbf{V}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) = \begin{pmatrix} \cdot \\ \frac{\bar{z}_1^j - \bar{z}_2^j}{V(\mathbf{z}^j)} \\ \cdot \end{pmatrix}$$

Sélection des axes

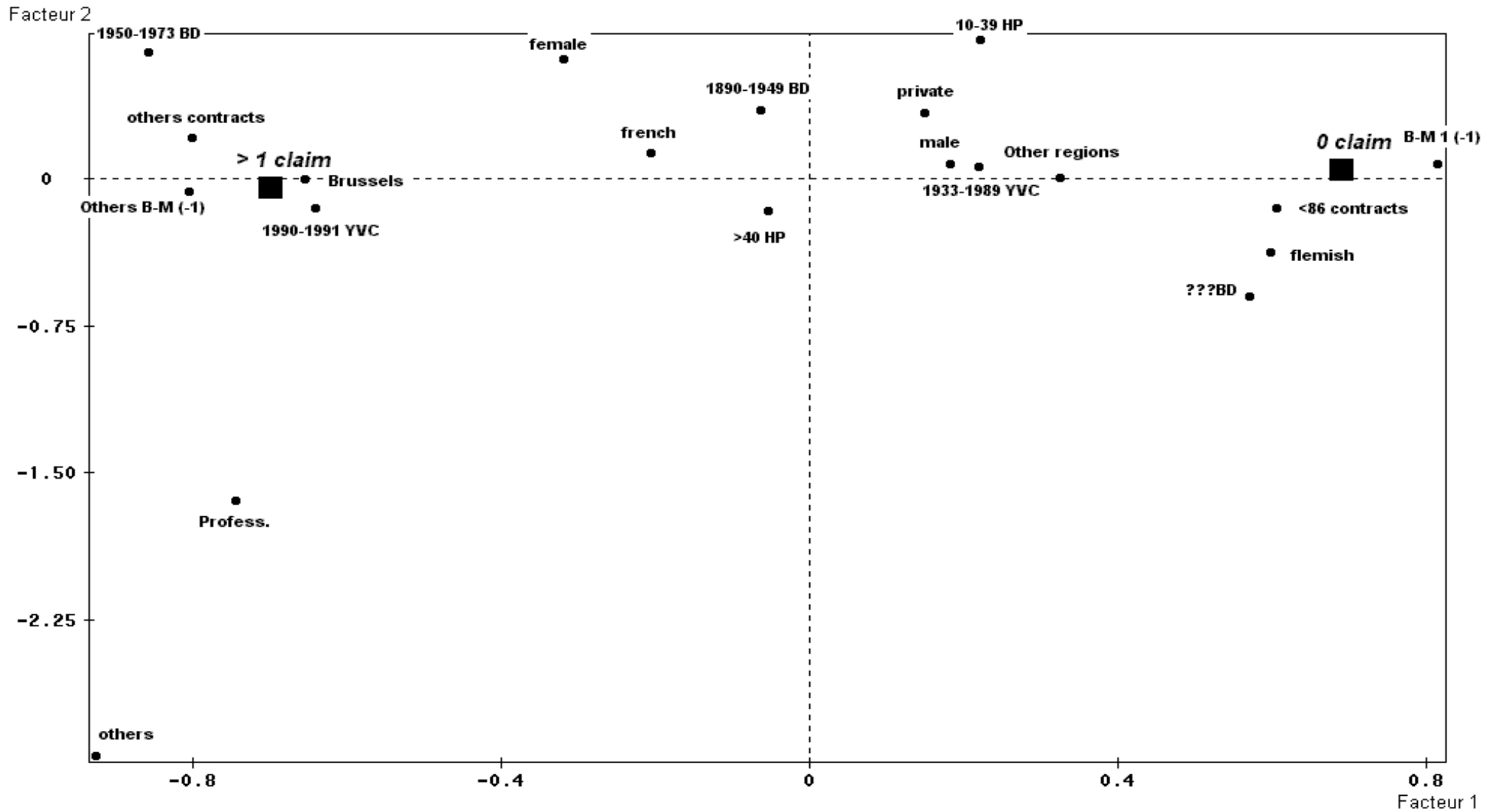


- Selon l'ordre de l'ACM
 - % d'inertie
- Selon le pouvoir discriminant
 - Student sur 2 groupes, F sur k groupes
- Régularisation, contrôle de la VC dimension

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

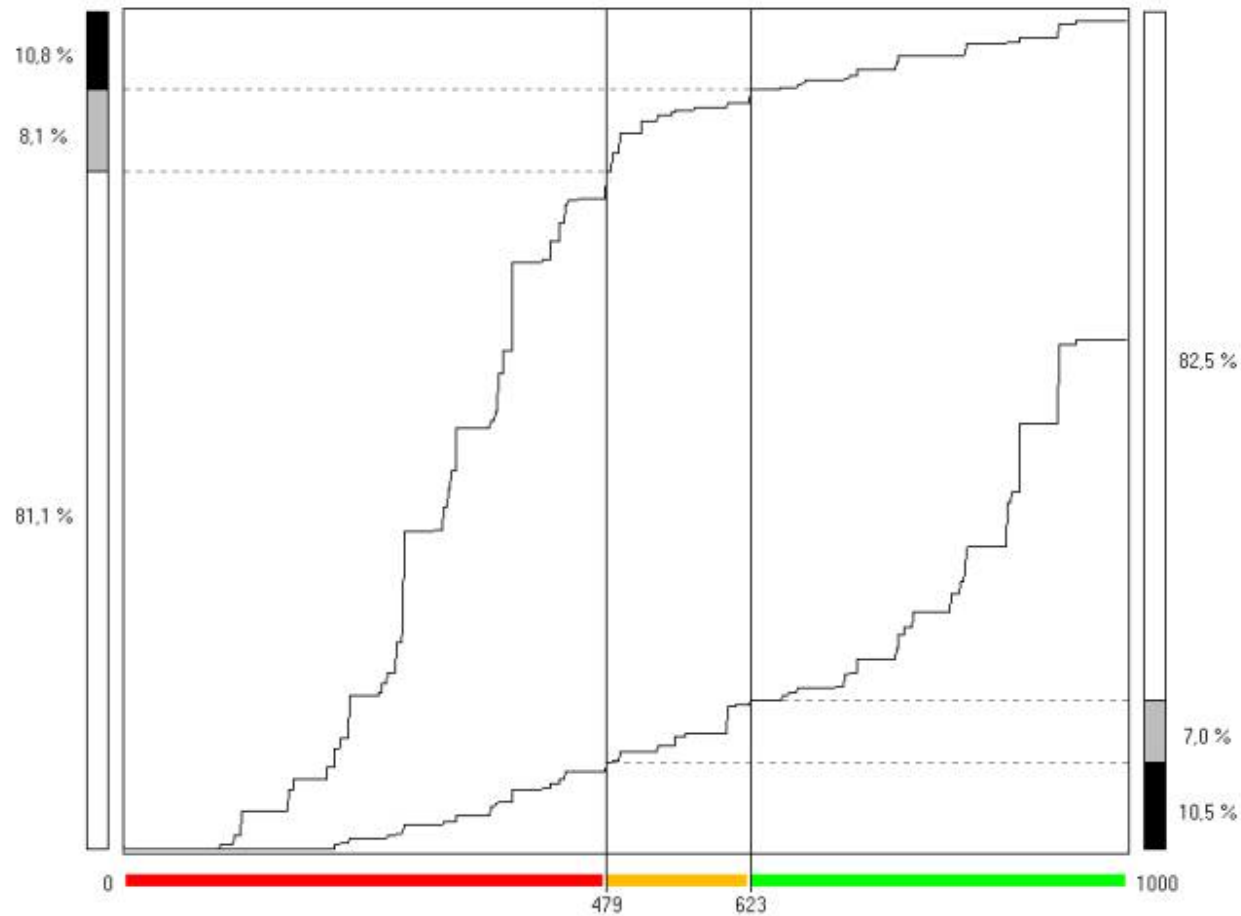
- Echelle de 0 à 1000

- 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

Fonctions de répartition



Cas des prédicteurs numériques



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

Prise en compte des interactions

- Amélioration considérable de l'efficacité du score

Rappel: $Score = f_1(x_1) + f_2(x_2) + \dots$

Modèle additif **sans** interaction

- *Exemple : État matrimonial et nombre d'enfants.*

$$\begin{array}{l}
 \text{2 catégories} \quad \text{3 catégories} \\
 (M_1 \quad M_2) \quad (E_1 \quad E_2 \quad E_3) \\
 \text{variable croisée à 6 catégories} \\
 (M_1E_1 \quad M_1E_2 \quad \quad \quad \quad M_2E_3)
 \end{array}$$

$$\begin{matrix}
 1 \\
 2 \\
 \cdot \\
 \cdot \\
 \cdot \\
 \cdot \\
 n
 \end{matrix}
 \begin{pmatrix}
 1 & 0 & \cdot & \cdot & \cdot & 0 \\
 0 & 1 & \cdot & \cdot & \cdot & 0 \\
 \cdot & \cdot & & & & \\
 \cdot & \cdot & & & & \\
 \cdot & \cdot & & & & \\
 \cdot & \cdot & & & & \\
 \cdot & \cdot & & & &
 \end{pmatrix}$$

Un exemple bancaire

- 15 000 dossiers de demandes de prêt
 - 1000 passés en contentieux
- Variables:
 - Taux d'endettement
 - Revenu disponible par personne du ménage
 - Situation dans le logement
 - Statut matrimonial
 - Nombre d'enfants
 - Profession
 - Ancienneté dans l'emploi

Grille de score

■ Ratio d'endettement :

Inférieur A 10%	Entre 10 et 20%	Entre 20 et 30%	Plus de 30%
+20	+16	+8	0

■ Revenu disponible par personne du ménage :

Inférieur A 1500F	Entre 1500 et 3000F	Plus de 3000F
0	+12	+20

■ Situation dans le logement :

Propriétaire	Locataire
+10	0

Grille de score (suite)

état matrimonial et enfants à charge :

Enfants	Etat matrimonial	
	Marié	Autres
0	+10	+8
1 ou 2	+20	+5
3 et +	+16	0

Grille de score (suite)

profession et stabilité dans l'emploi :

Profession	Travaille dans le même emploi depuis		
	Moins de 4 ans	4 à 10ans	Plus de 10 ans
Fonctionnaires, Retraités	+18	+30	+30
Industriels, Gros commerçant, Profession libérales, cadres supérieurs, employés de bureau	+15	+22	+25
Artisans, Petit commerçant, Exploitants agricoles, Cadres moyens	+5	+12	+17
Employés de commerce, Ouvriers, Autres	0	+5	+10

Exemple :

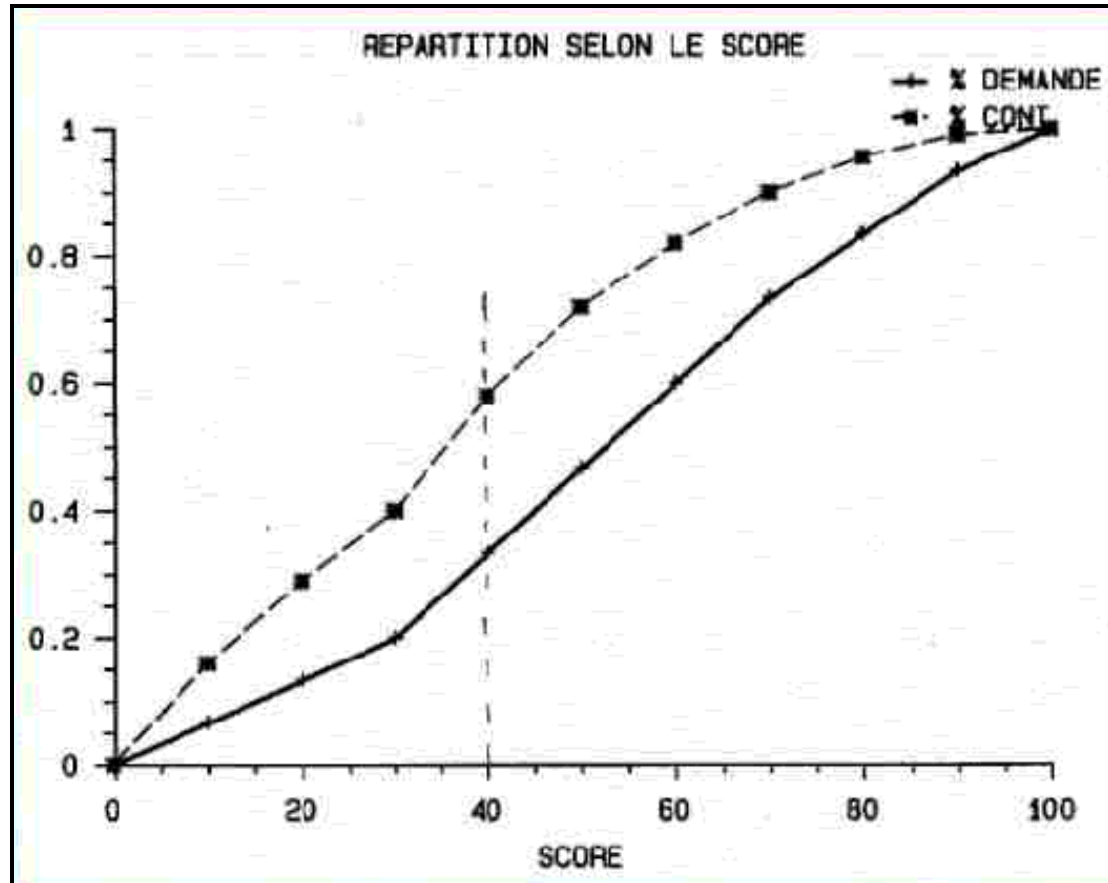
Ratio d'endettement	15%	+16
Revenu disponible par Personne	2300F	+12
Situation dans le logement	Locataire	+0
Etat matrimonial , nombre d'enfants à charge	Marlé, sans enfants	+10
Travaille dans le même emploi depuis	Employé de bureau depuis 6 ans	+22

■ Note de score : + 60

Répartitions par tranches de score

Tranche de score	Nbre de deman.	Nbre de conten.	Taux de conten.	Nbre de deman. cumul.	Nbre de conten. cumul.	Taux de conten. cumul.
90-100	1000	10	1 %	1000	10	1 %
80-90	1500	35	2,3 %	2500	45	1,8 %
70-80	1500	55	3,6 %	4000	100	2,5 %
60-70	2000	80	4 %	6000	180	3 %
50-60	2000	100	5 %	8000	280	3,5 %
40-50	2000	140	7 %	10000	420	4,2 %
30-40	2000	180	9 %	12000	600	5 %
20-30	1000	110	11 %	13000	710	5,4 %
10-20	1000	130	13 %	14000	840	6 %
0-10	1000	160	16 %	15000	1000	6,6 %

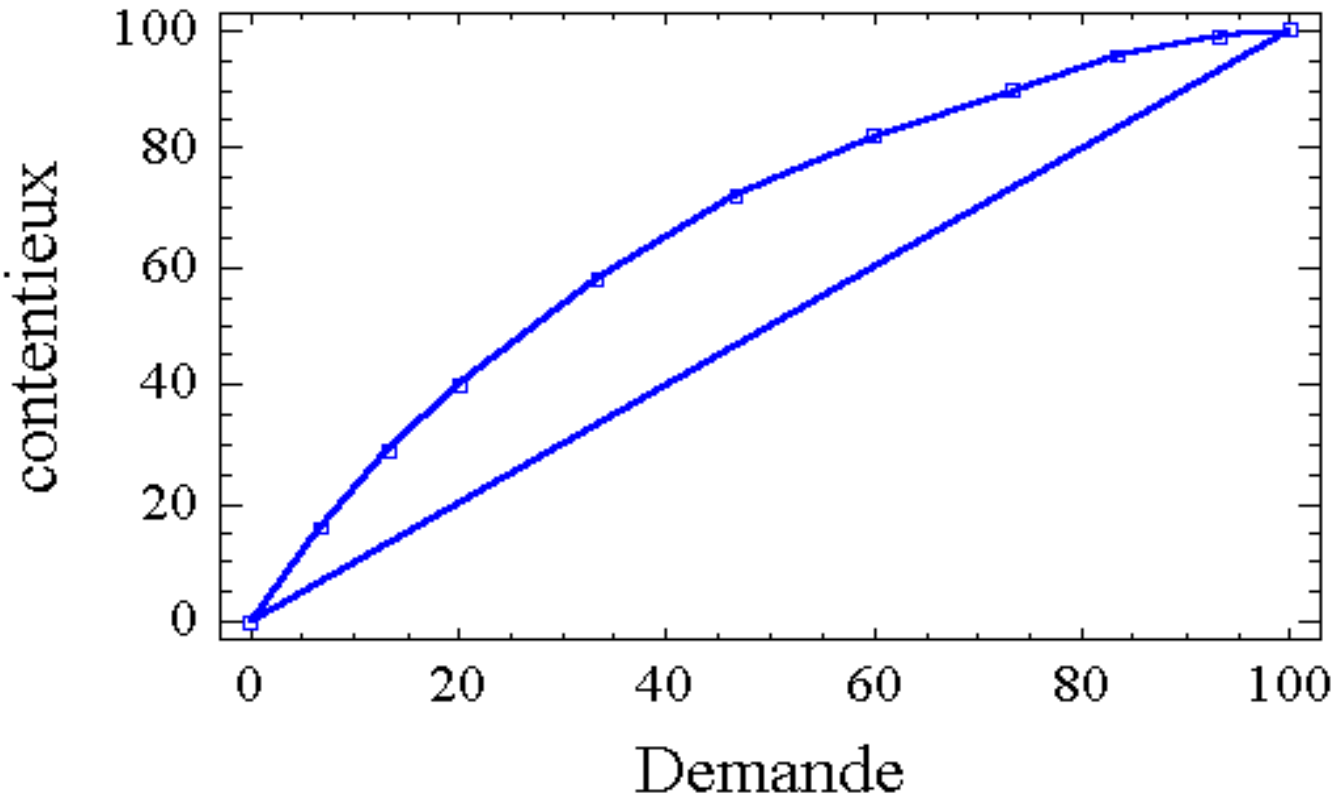
Répartition selon le score



Simulation

	Avant le score	Poli-tique n°1	Poli-tique n°2	Poli-tique n°3
Seuil de sélection	-	Note supér. à 50	Note supér. à 40	Note supér. à 30
Nombre de dossiers produits	10 000	8 000	10 000	12 000
Taux de refus	33 %	46 %	33 %	20 %
Contentieux sur la production	500	280	420	600
Taux de contentieux	5 %	3,5 %	4,2 %	5 %

Courbe de 'lift' (efficacité du ciblage)



II.3 Les objectifs du credit scoring



Sélection des risques
Prévision des impayés
Suivi et contrôle

credit scoring



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 » à partir de 2007

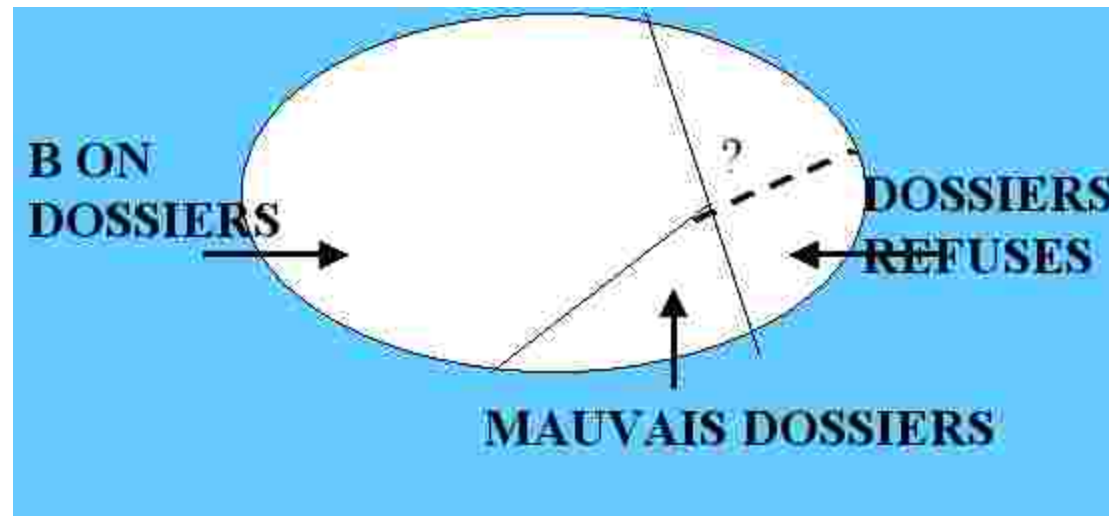
LES DIFFERENTES ETAPES DE REALISATION



- ECHANTILLONNAGE
- COLLECTE DE L'INFORMATION
- REDRESSEMENT
- SELECTION DES CRITERES
- CONSTRUCTION DU MODELE
- SIMULATION
- MISE EN OEUVRE

1. ECHANTILLONNAGE

- OBJECTIF :
- CONSTRUIRE UN ECHANTILLON REPRESENTATIF DE LA DEMANDE ET DU COMPORTEMENT DU PAYEUR.
- 1.1. PRISE EN COMPTE DES DOSSIERS REFUSES



- LES TROIS STRATES DE LA DEMANDE

PROBLEME



- UN SCORE CALCULE UNIQUEMENT SUR LES DOSSIERS ACCEPTES NE S'APPLIQUE PAS A L'ENSEMBLE DE LA DEMANDE.

PRISE EN COMPTE DE LA DIMENSION TEMPORELLE



■ DEUX POSSIBILITES :

■ A) OBSERVER UNE COUPE INSTANTANEE

- INCONVENIENT:

- CERTAINS DOSSIERS SONT CONSIDERES COMME « BONS » ALORS QU'ILS DEVIENDRONT « MAUVAIS » PAR LA SUITE.

■ B) OBSERVER UNE POPULATION DE DOSSIERS TERMINEES

- INCONVENIENT:

- LA STRUCTURE DE LA POPULATION OBSERVEE NE CORRESPOND PAS A LA STRUCTURE ACTUELLE.

2. LA COLLECTE DE L'INFORMATION



■ OBJECTIF:

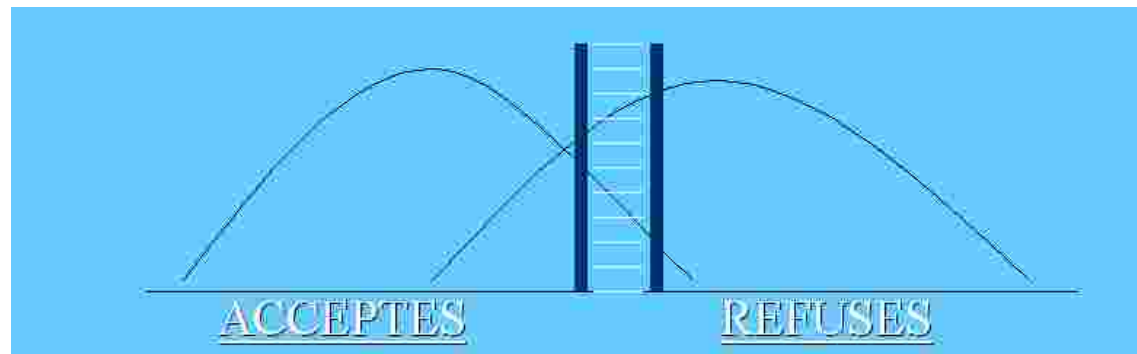
- BATIR UN FICHER CONTENANT TOUTES LES INFORMATIONS CONNUES SUR LES REFUSES AINSI QUE LES BONS ET MAUVAIS PAYEURS.

■ PROBLEMES:

- PAS DE STOCKAGE INFORMATIQUE DES OBSERVATIONS INDIVIDUELLES
- PAS DE CONSERVATION DES DOSSIERS REFUSES
- PAS DE STATISTIQUES PERMETTANT D'ELABORER LE PLAN DE SONDAGE
- HISTORIQUE TROP COURT OU ABSENT

3. REDRESSEMENT

- OBJECTIF: REDONNER A L'ECHANTILLON LA STRUCTURE DE LA DEMANDE ACTUELLE.
- DEUX FAMILLES DE METHODES :
 - A) SCORE ACCEPTE/REFUSE



- HYPOTHESE: LES REFUSES D'UN TRANCHE ONT LE MEME COMPORTEMENT QUE LES ACCEPTES.

3. REDRESSEMENT



■ B) SIMULATION DU COMPORTEMENT

- PRINCIPE : CHAQUE DOSSIER REFUSE SERAIT DEVENU BON (OU MAUVAIS) AVEC UNE PROBABILITE A ESTIMER.

4. SELECTION DES CRITERES



■ OBJECTIF:

- CHOISIR LES VARIABLES ET LES INTERACTIONS A INTRODUIRE DANS LE MODELE.

■ LES PROBLEMES :

- DECOUPAGE/REGROUPEMENT EN CATEGORIES.
- CHOIX DES INTERACTIONS.
- CHOIX DES VARIABLES LES PLUS EXPLICATIVES.
- CHOIX DES VARIABLES LES MOINS CORRELEES ENTRE ELLES.

7. LA MISE EN ŒUVRE




■ OBJECTIF:

- INTRODUIRE LE SCORE COMME OUTIL DE SELECTION, DE PREVISION ET DE SUIVI.

■ LES PROBLEMES :

- FORMATION DES UTILISATEURS.
- MISE EN PLACE DES OUTILS INFORMATIQUES.
- REACTUALISATION.

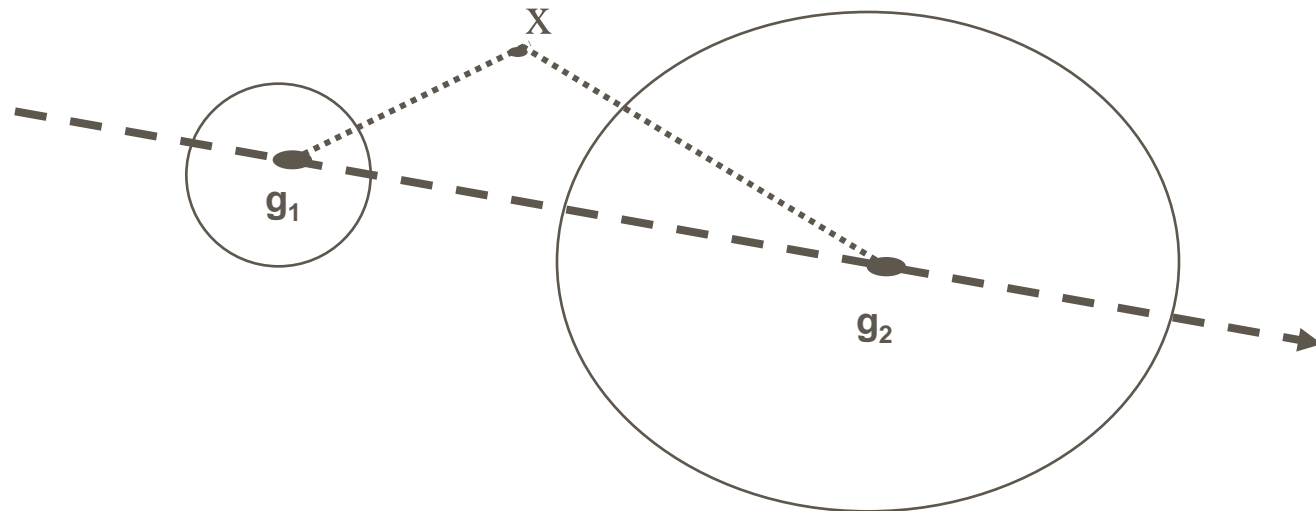
3^{ème} partie : Analyse discriminante probabiliste.



1. Règle bayésienne et loi normale.
2. Méthodes non paramétriques.

Insuffisances des règles géométriques

- Mesures de distances ?
- Risques d'erreurs ?
- Probabilités d'appartenance ?



III.1 Règle bayésienne

p_j probabilité *a priori* d'appartenir au groupe j
 $f_j(\mathbf{x})$ loi des x_i dans le groupe j

$$\text{Formule de Bayes : } P(G_j / \mathbf{x}) = \frac{p_j f_j(\mathbf{x})}{\sum_{j=1}^k p_j f_j(\mathbf{x})}$$

Problème : estimer les $f_j(\mathbf{x})$

■ 3 possibilités :

⇒ *Paramétrique : lois normales avec égalité ou non des Σ_j*

⇒ *Non paramétrique : noyaux ou plus proches voisins*

⇒ *Semi-paramétrique : régression logistique estimation directe de :*

$$P(G_j / x) = \frac{\exp(\theta'x + \theta_0)}{1 + \exp(\theta'x + \theta_0)}$$

La règle bayésienne naïve dans le cadre normal

$f_j(x)$ densité d'une $N(\mu_j; \Sigma_j)$

$$f_j(x) = \frac{1}{(2\pi)^{p/2} |\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_j)' \Sigma_j^{-1}(x - \mu_j)\right)$$

$\max p_j f_j(x) \Rightarrow$ attribuer x au groupe le plus
probable a posteriori

$$\max \left[\ln p_j - \frac{1}{2}(x - \mu_j)' \Sigma_j^{-1}(x - \mu_j) - \frac{1}{2} \ln |\Sigma_j| \right]$$

règle quadratique

La règle bayésienne

Hypothèse simplificatrice : $\Sigma_1 = \Sigma_2 \dots = \Sigma$

On attribue x au groupe j tel que :

$$\max \left[\text{Ln } p_j - \underbrace{\frac{1}{2} x' \Sigma^{-1} x}_{\substack{\text{indépendant} \\ \text{du groupe}}} - \frac{1}{2} \mu_j' \Sigma^{-1} \mu_j + x' \Sigma^{-1} \mu_j \right]$$

$$\text{donc : } \max \left[\underbrace{\text{Ln } p_j - \frac{1}{2} \mu_j' \Sigma^{-1} \mu_j}_{a_j} + x' \Sigma^{-1} \mu_j \right]$$

Règle linéaire équivalente à la règle géométrique si équiprobabilité, après estimation de μ_j par g_j et de Σ par W .

Analyse discriminante probabiliste: cas de deux groupes

Affecter au groupe 1 si $p_1 f_1(\mathbf{x}) > p_2 f_2(\mathbf{x})$

$$f_i(\mathbf{x}) = \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}} (2\pi)^{\frac{p}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)\right)$$

$$\boldsymbol{\mu}_1' \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_1' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 + \ln(p_1) > \boldsymbol{\mu}_2' \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_2' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 + \ln(p_2)$$

$$\underbrace{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \mathbf{x}}_{\text{fonction de Fisher}} > \ln\left(\frac{p_2}{p_1}\right) + \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$$

Fonction de score et probabilité

■ Fonction de score $S(\mathbf{x})$:

$$S(\mathbf{x}) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \mathbf{x} + \ln\left(\frac{p_1}{p_2}\right) - \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$$

Règle : affecter au groupe 1 si $S(\mathbf{x}) > 0$

■ Probabilité d'appartenance au groupe 1 :

$$P(G_1 / \underline{x}) = \frac{p_1 e^{-1/2(\underline{x} - \underline{\mu}_1)' \boldsymbol{\Sigma}^{-1} (\underline{x} - \underline{\mu}_1)}}{p_1 e^{-1/2(\underline{x} - \underline{\mu}_1)' \boldsymbol{\Sigma}^{-1} (\underline{x} - \underline{\mu}_1)} + p_2 e^{-1/2(\underline{x} - \underline{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\underline{x} - \underline{\mu}_2)}}$$

$$1/p = 1 + p_2 / p_1 e^{-1/2(\underline{x} - \underline{\mu}_1)' \boldsymbol{\Sigma}^{-1} (\underline{x} - \underline{\mu}_1) + 1/2(\underline{x} - \underline{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\underline{x} - \underline{\mu}_2)}$$

probabilité

$$\ln(1 / (P(G_1 / \mathbf{x})) - 1) = -S(\mathbf{x}) \quad 1 / P(G_1 / \mathbf{x}) = 1 + e^{-S(\mathbf{x})}$$

$$P(G_1 / \mathbf{x}) = \frac{1}{1 + e^{-S(\mathbf{x})}} = \frac{\exp(S(\mathbf{x}))}{1 + \exp(S(\mathbf{x}))}$$

Fonction logistique du score

- *Expression en fonction des distances de Mahalanobis aux centres :*

$$P = \frac{1}{1 + \frac{p_2}{p_1} \exp\left(-\frac{1}{2} \left((\Delta^2(\mathbf{x}; \boldsymbol{\mu}_2)) - (\Delta^2(\mathbf{x}; \boldsymbol{\mu}_1)) \right)\right)}$$

S(x)

- S(x) suit une loi normale car combinaison linéaire des composantes d'un vecteur normal

$$E(S(\underline{x})) = \frac{\Delta_p^2}{2} - \text{Log} \frac{p_2}{p_1} \text{ pour le groupe 1}$$

$$E(S(\underline{x})) = -\frac{\Delta_p^2}{2} - \text{Log} \frac{p_2}{p_1} \text{ pour le groupe 2}$$

$$\text{et } V(S(\underline{x})) = \Delta_p^2$$

- Probabilité d'erreur de classement de G2 en G1 :
On classe en G1 si $S(x) > 0$

$$P(S(x) > 0) = P\left(U > \frac{\Delta_p}{2} + \frac{1}{\Delta_p} \ln\left(\frac{p_2}{p_1}\right)\right)$$

Règle de Bayes avec coûts d'erreur

- Maximiser la probabilité *a posteriori* peut conduire à des règles absurdes.
 - Coûts d'erreurs :
 - $C(1/2)$ si on classe en G1 un individu de G2
 - $C(1/1) = 0$
 - Coût moyen *a posteriori* d'un classement en G1 : $C(1/2) P(G2/x)$
 - Coût moyen *a posteriori* d'un classement en G2 : $C(2/1) P(G1/x)$
 - On classera x en G1 si $C(1/2) P(G2/x) < C(2/1) P(G1/x)$

$$c(1/2) \frac{p_2 f_2}{p_1 f_1 + p_2 f_2} < c(2/1) \frac{p_1 f_1}{p_1 f_1 + p_2 f_2} \text{ donc si : } \frac{p_1 f_1}{p_2 f_2} > \frac{c(1/2)}{c(2/1)}$$

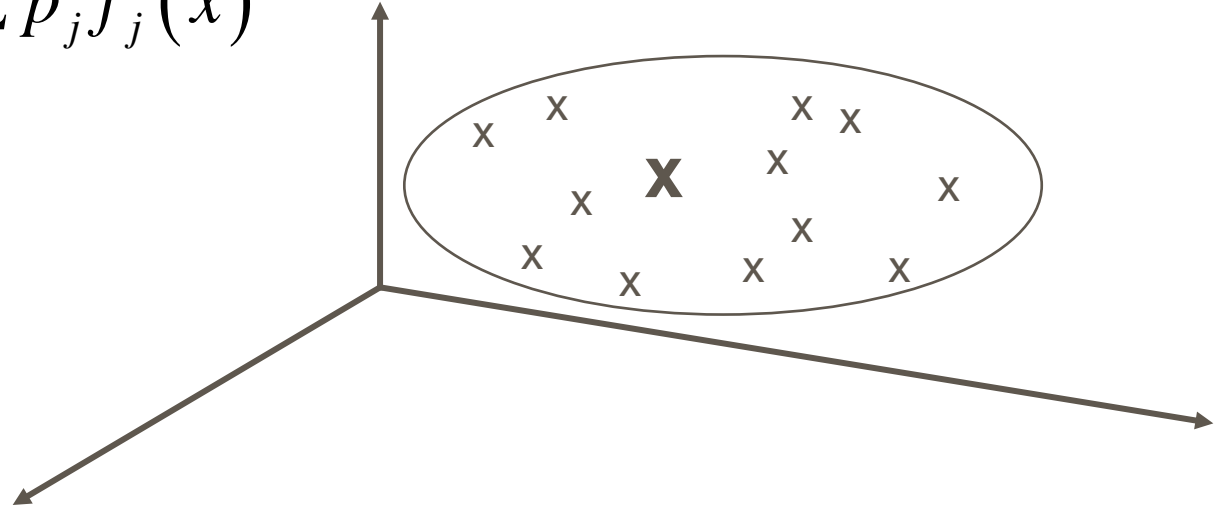
Règle habituelle avec $p'_1 = p_1 c(2/1)$ et $p'_2 = p_2 c(1/2)$

III - 2 : Discriminante non paramétrique

$$\text{Bayes} \quad P(G_j / x) = \frac{p_j f_j(x)}{\sum p_j f_j(x)}$$

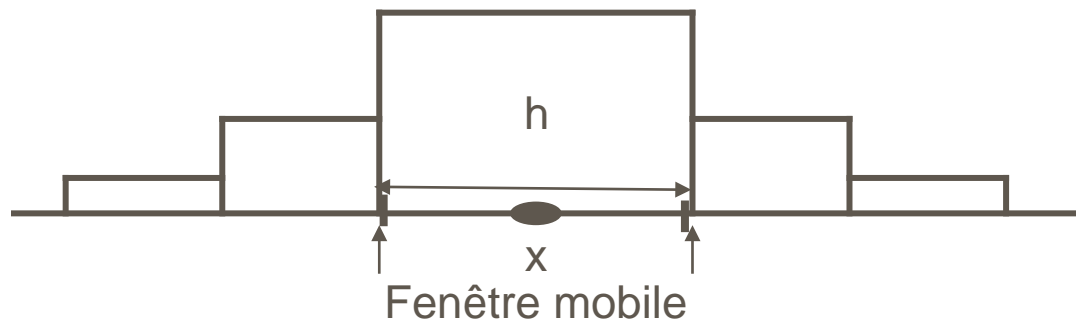
$$\hat{f}_j(x)$$

$$f_j(x) = \frac{\text{Fréquence}}{\text{Volume}}$$



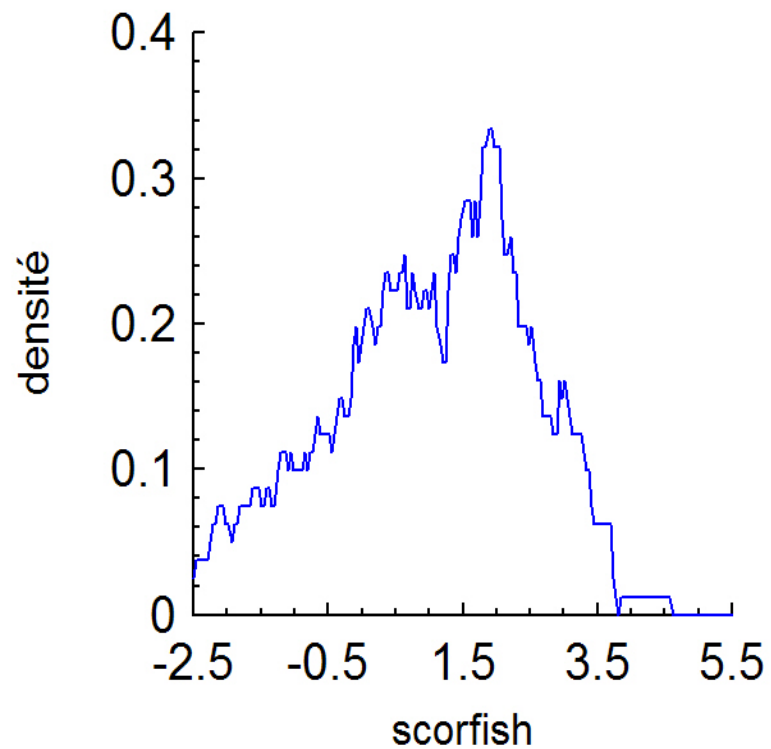
Fenêtre mobile: cas unidimensionnel

- Idée (Parzen-Rosenblatt): un histogramme où chaque classe serait centrée sur l'observation courante



Fenêtre mobile

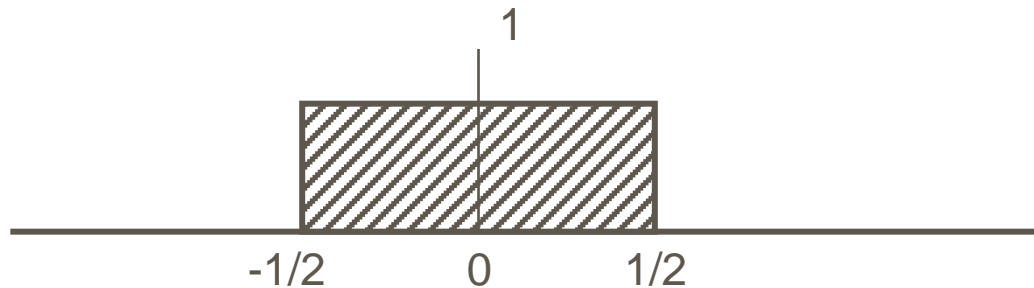
Graphique de densité



$$\hat{f}(x) = n_x / nh$$

Estimateur
discontinu.

densité



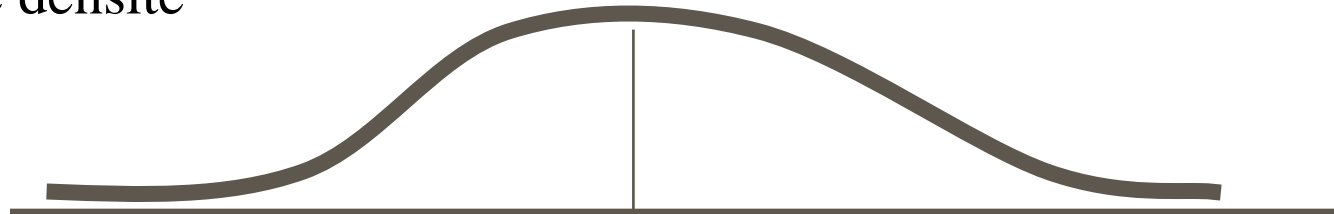
$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n k\left(\frac{x - x_i}{h}\right)$$

$$\begin{cases} K(t) = 1 & \text{si } t \in]-1/2 ; 1/2[\\ K(t) = 0 & \text{sinon} \end{cases}$$

$$K\left(\frac{x - x_i}{h}\right) = 1 \text{ si } x_i \in \left[x - \frac{h}{2} ; x + \frac{h}{2} \right]$$

Méthode du noyau $\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n k\left(\frac{x - x_i}{h}\right)$

k fonction de densité



Choix du noyau

- K continue, paire, unimodale $\int_{-\infty}^{+\infty} K(x)dx = 1$

- Exemples

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) \quad K(u) = \frac{3}{4\sqrt{5}} \left(1 - \frac{u^2}{5}\right) \text{ pour } |u| < \sqrt{5} \text{ Epanechnikov}$$

- K pas forcément positif

$$K(u) = \frac{105}{64} (1 - u^2)^2 (1 - 3u^2) \text{ pour } |u| \leq 1 \text{ Lejeune}$$

Quelques résultats théoriques

- Il n'existe pas d'estimateur sans biais d'une densité qui soit continu, symétrique en x_i

$$E(\hat{f}(x)) = f(x) \quad \forall x \text{ est impossible}$$

- Critère du MISE

$$E\left(\int_{-\infty}^{+\infty} \left(\hat{f}(x) - f(x)\right)^2 dx\right)$$

- Si $\int_{-\infty}^{+\infty} K(x)dx = 1$ $\int_{-\infty}^{+\infty} xK(x)dx = 0$ et $\int_{-\infty}^{+\infty} x^2 K(x)dx = k_2$

$$MISE \approx \frac{h^4}{4} k_2 \int_{-\infty}^{+\infty} (f''(x))^2 dx + \frac{1}{nh} \int_{-\infty}^{+\infty} (K(x))^2 dx$$

$$h_{optimal} = k_2^{-\frac{2}{5}} \left[\int_{-\infty}^{+\infty} (K(x))^2 dx \right]^{\frac{4}{5}} \left[\int_{-\infty}^{+\infty} (f''(x))^2 dx \right]^{-\frac{1}{5}} n^{-\frac{1}{5}}$$

- En substituant h_{opt} qui dépend de f...

$$MISE \approx \frac{5}{4} k_2^{\frac{2}{5}} \left[\int_{-\infty}^{+\infty} (K(x))^2 dx \right]^{\frac{4}{5}} \left[\int_{-\infty}^{+\infty} (f''(x))^2 dx \right]^{\frac{1}{5}} n^{-\frac{4}{5}}$$

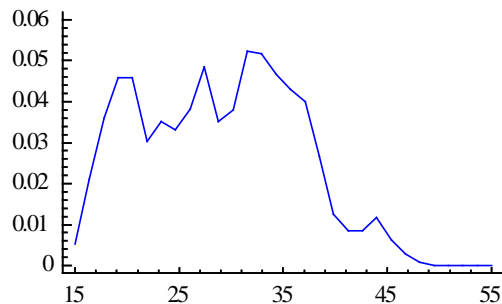
- Calcul des variations:

K optimal = Epanechnikov

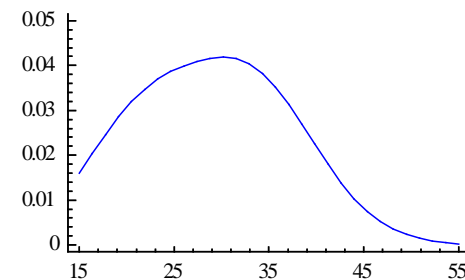
- Noyau moins influent que la constante de lissage

Paramètre de lissage h

- h (ou r) Joue le même rôle que la largeur de classe dans l'histogramme.
- Estimation de h :
 - *Méthodes visuelles (si $p = 1$)*
 - *Maximum de vraisemblance*
- h petit :

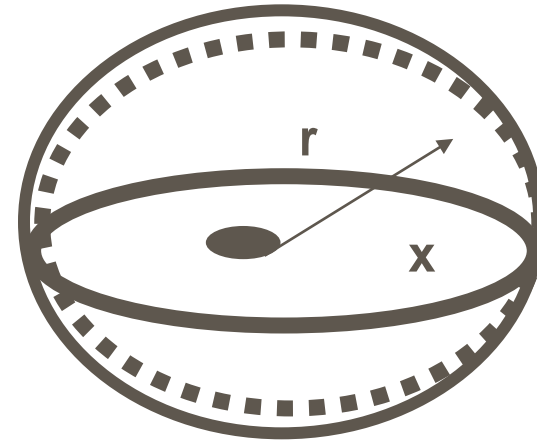


h grand :



Estimation de densité par la méthode du noyau élémentaire

- \Rightarrow Noyau uniforme
 - On compte le nombre d'observations appartenant à la boule de rayon r .
 - Ce nombre est aléatoire.



- \Rightarrow Plus proches voisins.
 - k nombre de voisins est fixé.
 - Volume de la boule : aléatoire.

$$f(x) = \frac{k}{nV} \quad k \text{ paramètre à fixer}$$

noyaux

- noyaux
$$f_t(x) = \frac{1}{n_t} \sum_y k_t(x - y)$$
- uniforme
$$k_t(z) = \begin{cases} \frac{1}{V_r(t)} & \text{si } z' V_t^{-1} z \leq r^2 \\ 0 & \text{sinon} \end{cases}$$
- normal
$$k_t(z) = \frac{1}{C_0(t)} \exp\left(-\frac{1}{2} \frac{z' V_t^{-1} z}{r^2}\right)$$
- Epanechnikov
$$k_t(z) = C_1(t) \left(1 - \frac{z' V_t^{-1} z}{r^2}\right) \text{ si } z' V_t^{-1} z \leq r^2$$
- Biweight
$$k_t(z) = C_2(t) \left(1 - \frac{z' V_t^{-1} z}{r^2}\right)^2$$
- Triweight
$$k_t(z) = C_3(t) \left(1 - \frac{z' V_t^{-1} z}{r^2}\right)^3$$

Estimation de densité versus discrimination linéaire



- Discrimination linéaire :
 - simplicité, robustesse, interprétation
 - inefficace si non linéarités fortes
- Estimation de densité :
 - précision, adaptation aux données
 - calculs complexes, absence d'interprétation

4 ème partie: La régression logistique

IV.1 Le modèle logistique simple

IV.2 Odds ratios


IV.3 Interprétation économétrique

IV.4 Estimation

IV.5 Tests

IV.6 Régression logistique multiple

IV.7 Comparaison avec l'analyse discriminante

- 
- Berkson (biostatistique) 1944
 - Cox 1958
 - Mc Fadden (économétrie) 1973

IV.1 Le modèle logistique simple

- Réponse dichotomique : $Y = 0 / 1$
- Variable explicative : X
- Objectif : Modéliser

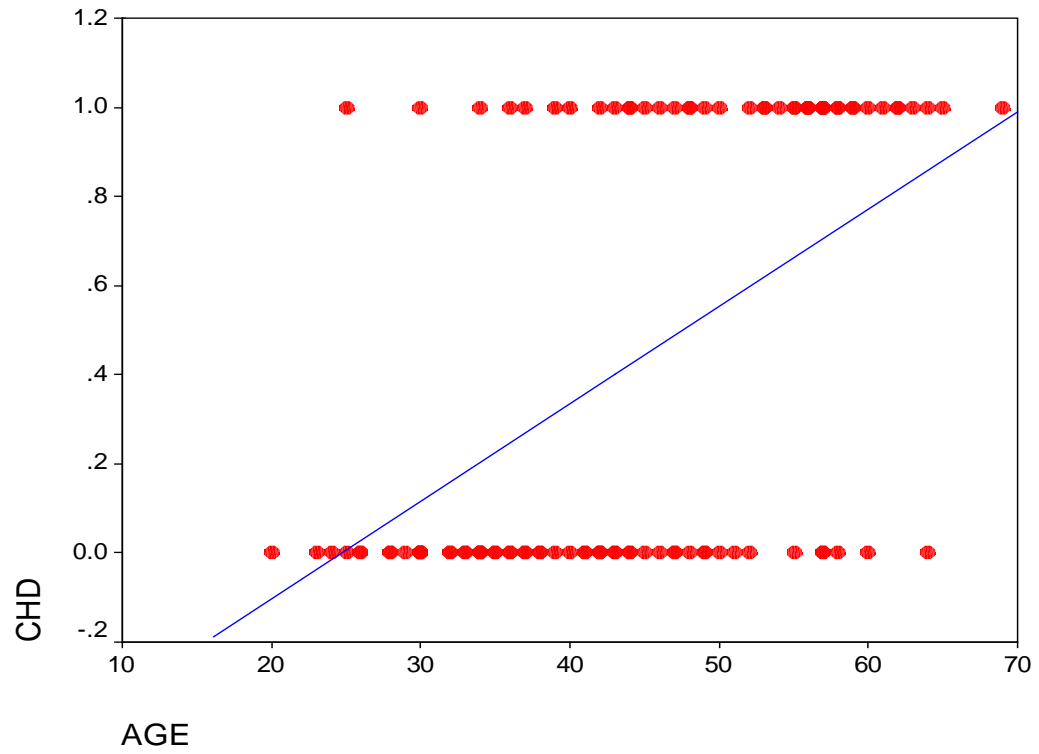
$$\pi(x) = \text{Prob}(Y = 1/X = x)$$

- Le modèle linéaire $\pi(x) = \beta_0 + \beta_1 x$ convient mal lorsque X est continue.
- Le modèle logistique est plus naturel

Exemple : Age and Coronary Heart Disease Status (CHD) (Hosmer & Lemeshow; M. Tenenhaus)

Les données

ID	AGRP	AGE	CHD
1	1	20	0
2	1	23	0
3	1	24	0
4	1	25	0
5	1	25	1
⋮	⋮	⋮	⋮
97	8	64	0
98	8	64	1
99	8	65	1
100	8	69	1

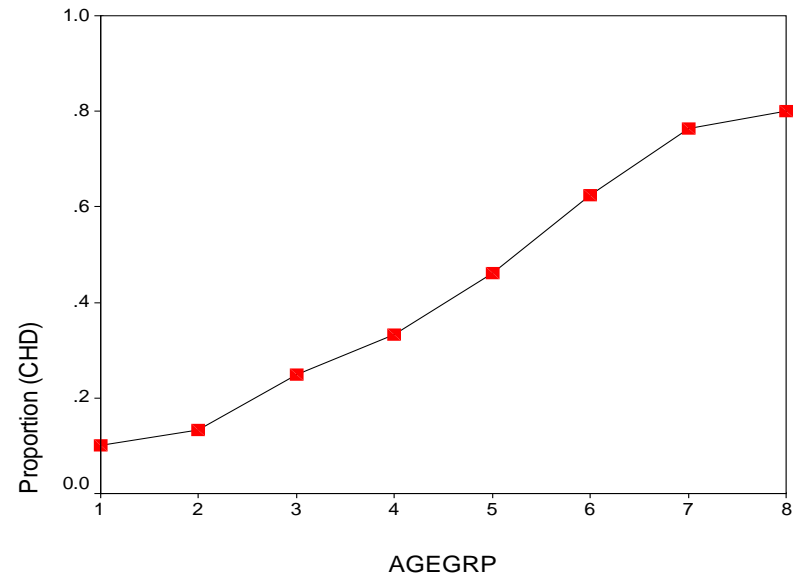


Description des données regroupées par classe d'âge

Tableau des effectifs
de CHD par classe d'âge

Age Group	n	CHD absent	CHD present	Mean (Proportion)
20 – 29	10	9	1	0.10
30 – 34	15	13	2	0.13
35 – 39	12	9	3	0.25
40 – 44	15	10	5	0.33
45 – 49	13	7	6	0.46
50 – 54	8	3	5	0.63
55 – 59	17	4	13	0.76
60 – 69	10	2	8	0.80
Total	100	57	43	0.43

Graphique des proportions
de CHD par classe d'âge



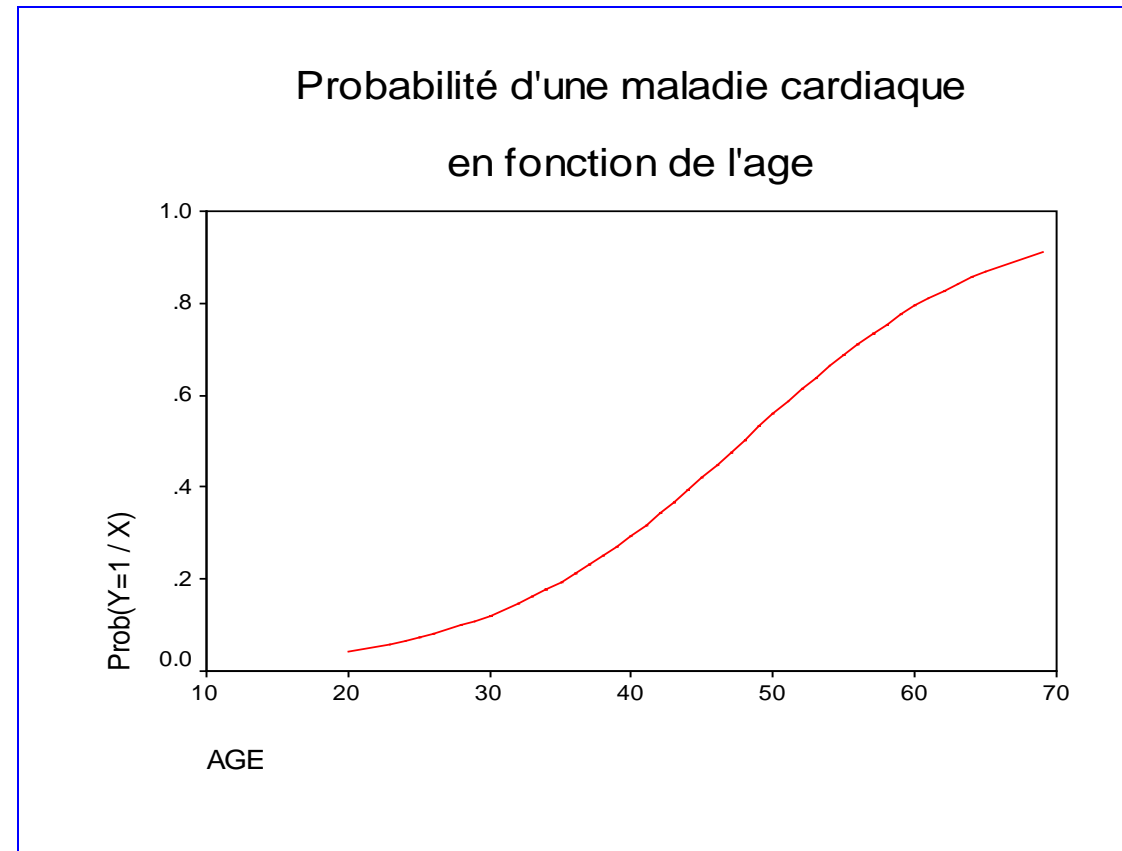
Le modèle logistique simple

$$\pi(\mathbf{x}) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

ou

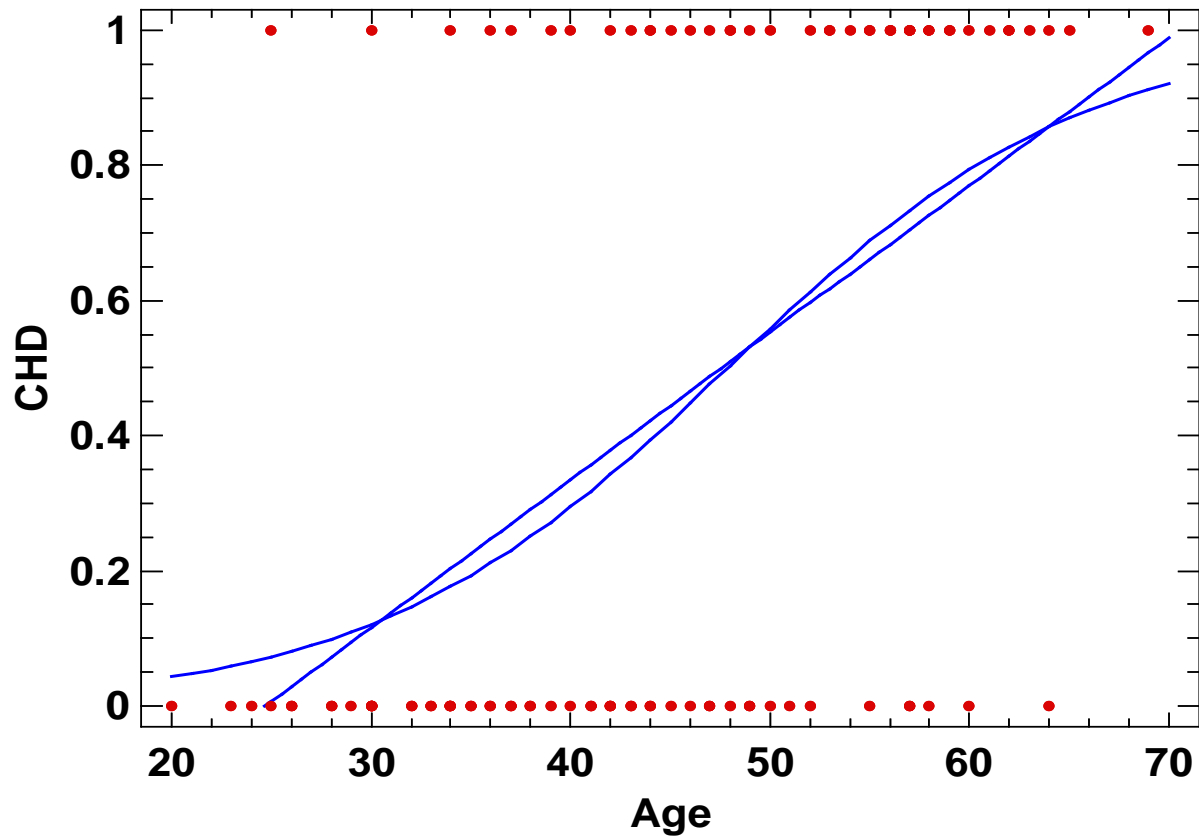
$$\text{Log}\left(\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}\right) = \beta_0 + \beta_1 x$$

Fonction de lien : Logit



- Il s'agit bien d'un problème de régression:
- Modélisation de l'espérance conditionnelle
- $E(Y/X=x) = f(x)$
- Choix de la forme logistique en épidémiologie:
 - S'ajuste bien
 - Interprétation de β_1 en termes d'odds-ratio

■ comparaison régressions linéaire et logistique



IV.2 Odds-Ratio

Si X binaire (sujet exposé $X=1$, non exposé $X=0$)

$$P(Y = 1 / X = 1) = \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}} \quad P(Y = 1 / X = 0) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

$$OR = \frac{P(Y = 1 / X = 1) / P(Y = 0 / X = 1)}{P(Y = 1 / X = 0) / P(Y = 0 / X = 0)} = e^{\beta_1}$$

Odds-Ratio

- Mesure l'évolution du rapport des chances d'apparition de l'événement $Y=1$ contre $Y=0$ (la cote des parieurs) lorsque X passe de x à $x+1$.
- Formule générale:

$$OR = \frac{\pi(x+1)/(1-\pi(x+1))}{\pi(x)/(1-\pi(x))} = e^{\beta_1}$$

IV.3 Interprétation économétrique

- Y possession d'un bien durable par un ménage: manifestation visible d'une variable latente Z inobservable continue.
- Z est l'« intensité du désir » de posséder le bien
- Si $Z < \text{seuil}$ $Y = 0$, sinon $Y = 1$
- Le seuil peut être choisi égal à 0

Modèle d'utilité

- pour le ménage i de caractéristiques x_i (âge, sexe, revenu, CSP...), la possession du bien procure un niveau d'utilité $U(1, x_i)$, la non possession $U(0, x_i)$.

$$Y_i = 1 \Leftrightarrow U(1, x_i) > U(0, x_i)$$

$$Y_i = 0 \Leftrightarrow U(0, x_i) > U(1, x_i)$$

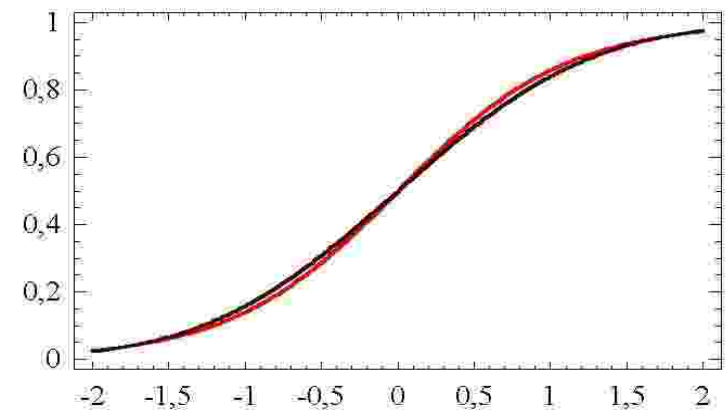
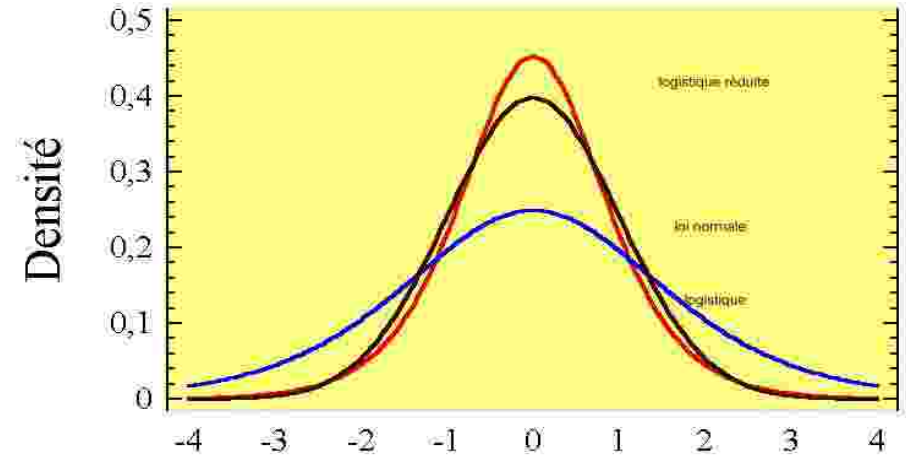
- Variable latente $Z_i = U(1, x_i) - U(0, x_i)$.

Modèle d'utilité (suite)

- $Z_i = x_i \beta + \varepsilon_i$
- $\pi_i = P(Y_i=1 | x_i) = P(Z_i > 0) = P(x_i \beta > -\varepsilon_i) = F(x_i \beta)$
- F fonction de répartition de $-\varepsilon_i$
- Choix de F:
 - Logistique : modèle logit, régression logistique
 - Normal: modèle probit

Comparaison logit-probit

- Logit: $F(x) = 1/(1+e^{-x})$
 $E(X)=0$ $V(X)=\pi^2/3$
- Peu différent en pratique
- Logit plus simple numériquement



IV.4 Estimation des paramètres

Les données

X	Y
x₁	y₁
⋮	⋮
x_i	y_i
⋮	⋮
x_n	y_n

$y_i = 1$ si caractère présent,
0 sinon

Le modèle

$$\begin{aligned}\pi(x_i) &= P(Y = 1 / X = x_i) \\ &= \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}\end{aligned}$$

Vraisemblance (conditionnelle!)

Probabilité d'observer les données

$$[(x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n)]$$

$$= \prod_{i=1}^n \text{Prob}(Y = y_i / X = x_i) = \prod_{i=1}^n \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i}$$

$$= \prod_{i=1}^n \left(\frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right)^{y_i} \left(1 - \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right)^{1-y_i} = L(\beta_0, \beta_1)$$

maximum de vraisemblance

■ $\hat{\beta}_0$ et $\hat{\beta}_1$ maximisent $L(\beta_0, \beta_1) = L(\boldsymbol{\beta})$

■ Maximisation de la log-vraisemblance

$$\ell(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}) = \sum_{i=1}^n [y_i \log \pi_i(x) + (1 - y_i) \log(1 - \pi_i(x))]$$

$$\begin{cases} \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_0} = \sum_{i=1}^n (y_i - \pi_i(x)) = 0 \\ \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_1} = \sum_{i=1}^n x_i (y_i - \pi_i(x)) = 0 \end{cases}$$

■ Estimateurs obtenus par des procédures numériques: pas d'expression analytique

Précision (asymptotique) des estimateurs

■ La matrice

$$V(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} V(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_0, \hat{\beta}_1) & V(\hat{\beta}_1) \end{bmatrix}$$

est estimée par la matrice

$$\left[-\frac{\partial^2 \text{Log } L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} \right]_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}^{-1}$$

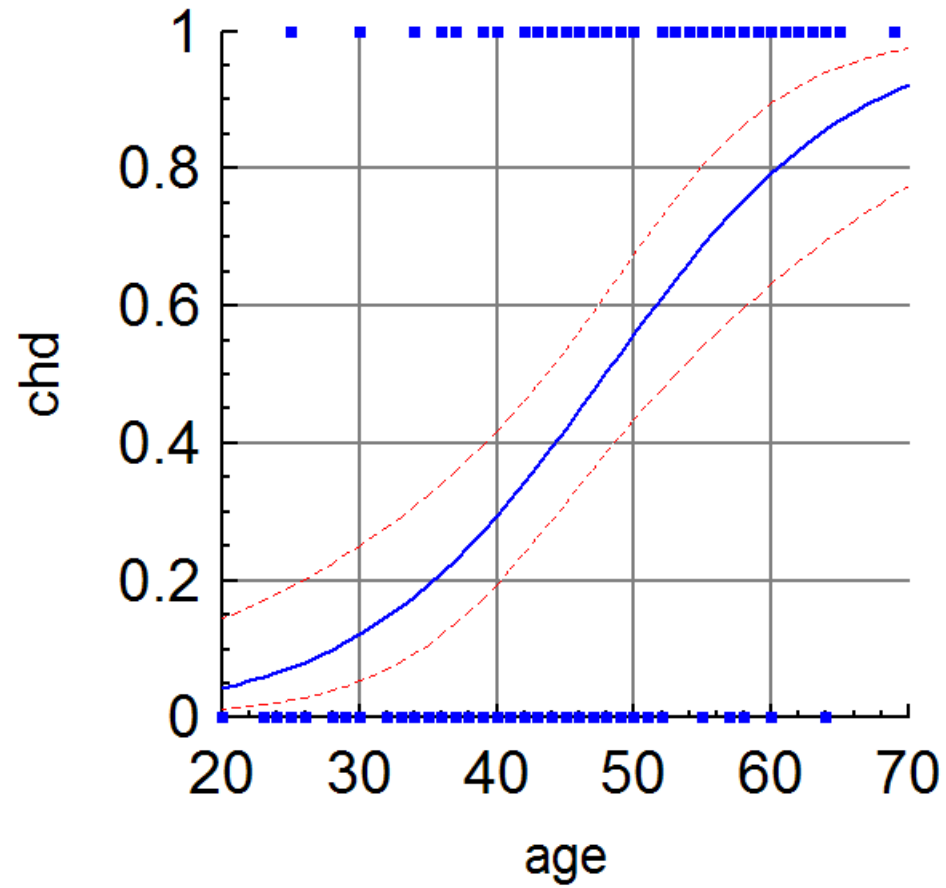
$$\begin{aligned}
V(\hat{\boldsymbol{\beta}}) &= \left[\frac{-\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} \right]_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}^{-1} \\
&= \left[\begin{array}{cc} \sum_{i=1}^n \hat{\pi}_i(1-\hat{\pi}_i) & \sum_{i=1}^n x_i \hat{\pi}_i(1-\hat{\pi}_i) \\ \sum_{i=1}^n x_i \hat{\pi}_i(1-\hat{\pi}_i) & \sum_{i=1}^n x_i^2 \hat{\pi}_i(1-\hat{\pi}_i) \end{array} \right]^{-1} \\
&= \left(\begin{array}{c} \left[\begin{array}{cc} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{array} \right]' \left[\begin{array}{ccc} \hat{\pi}_1(1-\hat{\pi}_1) & & 0 \\ & \ddots & \\ 0 & & \hat{\pi}_n(1-\hat{\pi}_n) \end{array} \right] \left[\begin{array}{c} 1 \\ \vdots \\ 1 \end{array} \right] \\ \left[\begin{array}{c} 1 \\ \vdots \\ 1 \end{array} \right] \left[\begin{array}{c} x_1 \\ \vdots \\ x_n \end{array} \right] \end{array} \right)^{-1} \\
&= (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}.
\end{aligned}$$

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-5.3095	1.1337	21.9350	0.0001	.	.
AGE	1	0.1109	0.0241	21.2541	0.0001	0.716806	1.117

$$\pi(x) = \frac{e^{-5,3095+0,1109x}}{1 + e^{-5,3095+0,1109x}}$$

Graphique du modèle ajusté
avec intervalles de confiance à 95.0%



IV.5 Tests sur les paramètres

- Trois méthodes sont disponibles pour tester l'apport de la variable X au modèle :

1. Le test de Wald

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

2. La méthode du rapport de vraisemblance

3. Le test du score

Test de Wald

- analogue à un test de Student en régression usuelle, si l'on considère la statistique w définie par :

$$w = \frac{\hat{\beta}_1}{\hat{s}(\hat{\beta}_1)}$$

- $\hat{s}(\hat{\beta}_1)$ représente l'estimation de l'écart-type de l'estimateur de β_1 .
- Sous l'hypothèse H_0 , w^2 suit approximativement une loi du khi-deux à un degré de liberté .
- Rejet de H_0 si $w^2 \geq \chi_{1-\alpha}^2(1)$

Test du rapport des vraisemblances

- L'apport de la variable X est mesuré à l'aide de la statistique :

$$G = -2 \log \left[\frac{\text{Vraisemblance sans la variable}}{\text{Vraisemblance avec la variable}} \right]$$

sous l'hypothèse H_0 G suit asymptotiquement une loi du khi-deux à un degré de liberté.

- Vraisemblance sans la variable:

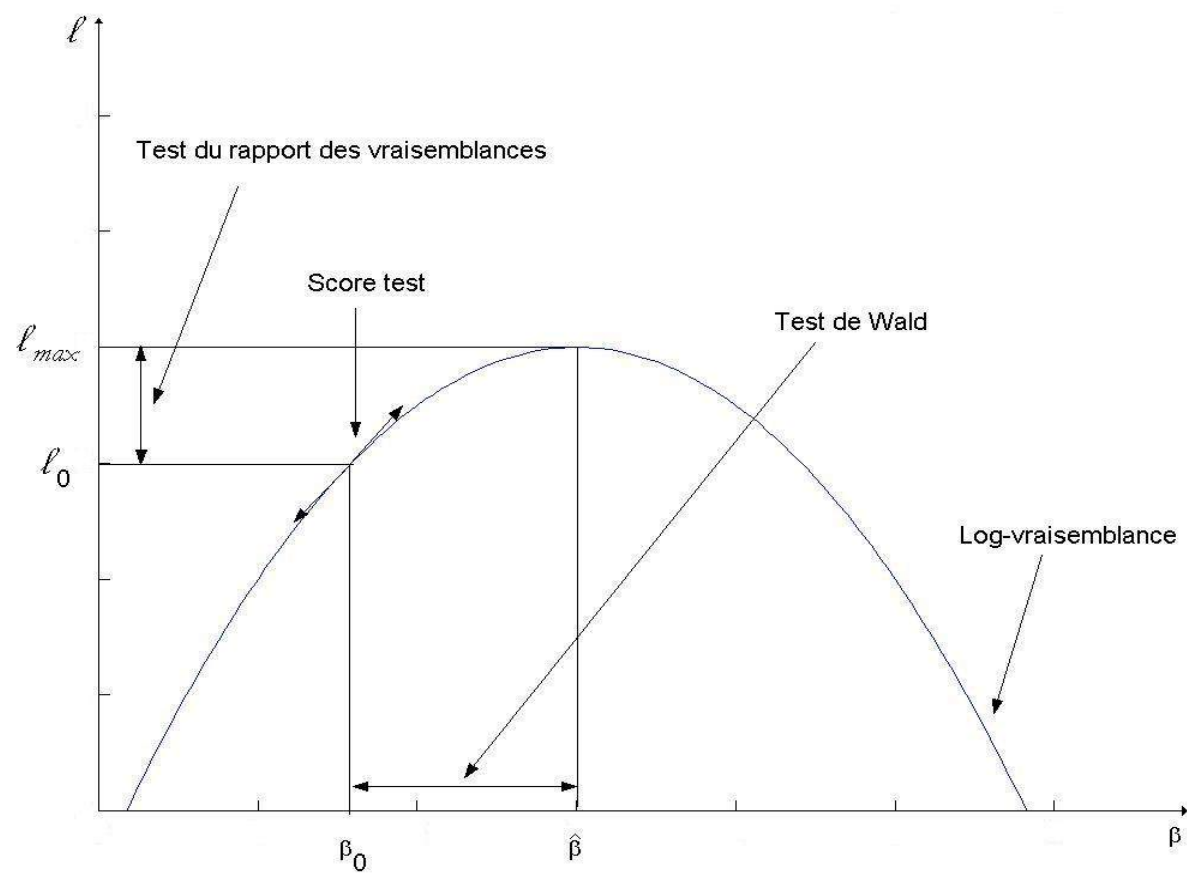
$$\left(\frac{n_1}{n} \right)^{n_1} \left(\frac{n_0}{n} \right)^{n_0}$$

Test du score

$$score = U(\beta)'_{\hat{\beta}_{H_0}} \left[J(\hat{\beta}_{H_0}) \right]^{-1} U(\beta)_{\hat{\beta}_{H_0}}$$

- U vecteur des dérivées partielles de la log-vraisemblance estimées
- Le score suit également asymptotiquement sous H_0 une loi du khi-deux à un degré de liberté
- En régression logistique simple, le score est égal à nr^2 , où r est le coefficient de corrélation linéaire (abusif!) entre Y et X

Comparaison des 3 tests



Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	138.663	111.353	.
SC	141.268	116.563	.
-2 LOG L Score	136.663	107.353	29.310 with 1 DF (p=0.0001) 26.399 with 1 DF (p=0.0001)

Intervalle de confiance de l'odds-Ratio

$$\text{Var}(\hat{\beta}_1) = s_1^2$$

D'où l'intervalle de confiance de OR(1) au niveau 0.95:

$$[e^{\hat{\beta}_1 - 1.96s_1}, e^{\hat{\beta}_1 + 1.96s_1}]$$

Intervalle de confiance de $\pi(x)$ au niveau 95%

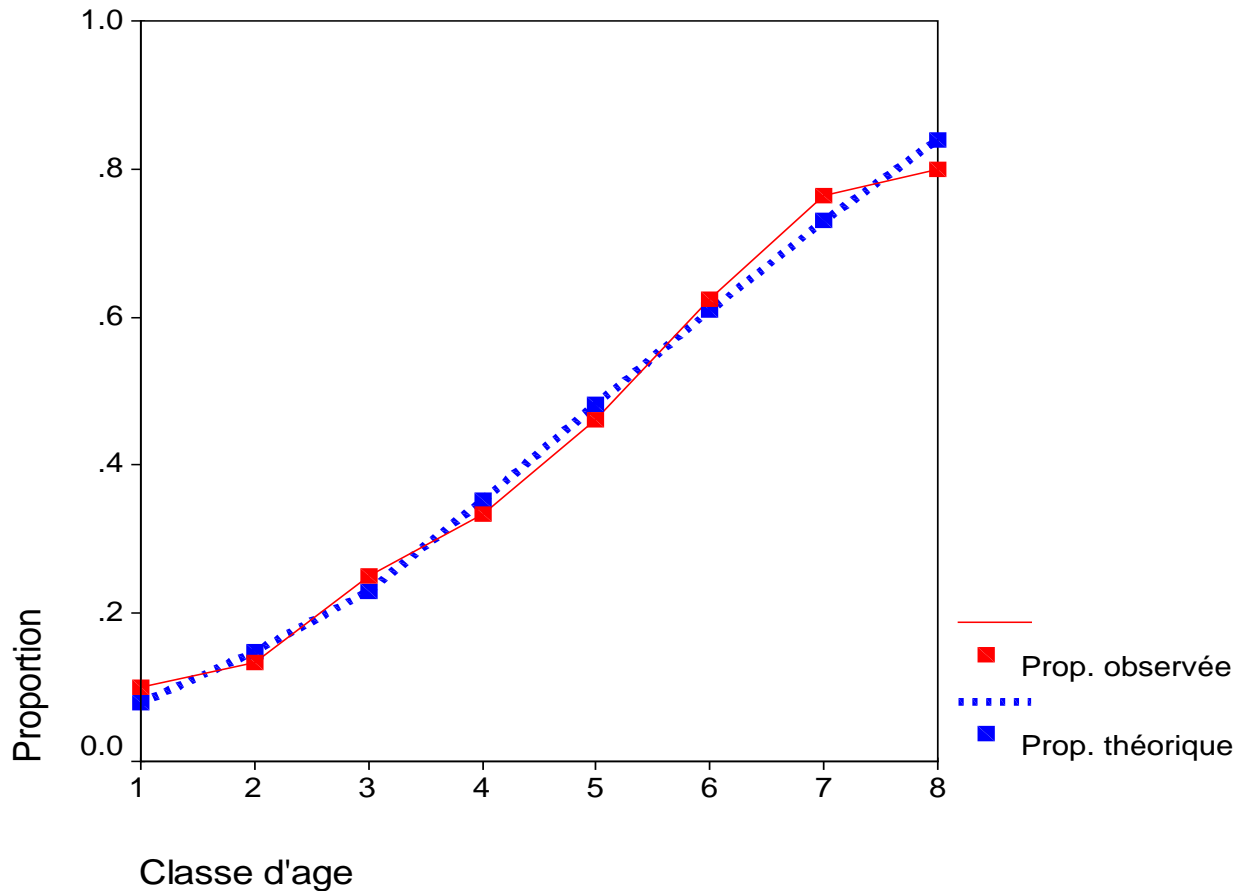
De $\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x) = s_0^2 + s_1^2 x^2 + 2s_{01} x$

on déduit l'intervalle de confiance de

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

$$\left[\frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x - 1.96 \sqrt{\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x)}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x - 1.96 \sqrt{\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x)}}}; \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x + 1.96 \sqrt{\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x)}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x + 1.96 \sqrt{\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x)}}} \right]$$

Comparaison entre les proportions observées et théoriques



Proportion observée :

$$\sum_{i \in \text{Classe}} y_i / n_{\text{Classe}}$$

Proportion théorique :

$$\sum_{i \in \text{Classe}} \hat{\pi}_i / n_{\text{Classe}}$$

puisque $E(y_i) = \pi_i$
estimé par $\hat{\pi}_i$

IV.6 Régression logistique multiple

- Généralisation à p variables explicatives X_1, \dots, X_p .

$$\pi(x) = P(Y = 1 / X = x) = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}$$

- Estimation par le maximum de vraisemblance
 - Ne converge pas toujours: cas de la séparation complète

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: on ajoute $\ln\left(\frac{p_1}{p_2}\right)$
 - Proc DISCRIM
 - PRIORS statement
 - Proc LOGISTIC
 - PEVENT option MODEL statement (SAS 8)
 - PRIOR (ou PRIOREVENT) option SCORE statement (SAS 9)
- Important pour les probabilités , pas pour un score

Tests

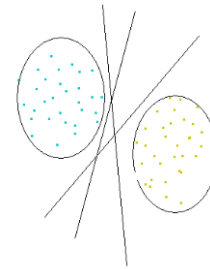
- Tests d'absence d'effet de toutes les variables: $H_0 : \beta_1 = \dots = \beta_p = 0$
 - Rapport de vraisemblance G
 - Score test U
 - Test de Wald
 - Sous H_0 , suivent tous trois asymptotiquement une loi du χ^2 à p ddl

IV.7 Comparaison avec l'analyse discriminante

- Avantages proclamés:
 - Unicité et interprétabilité des coefficients (odds-ratios)
 - Erreurs standard calculables
 - Modélisation des probabilités
 - Hypothèses plus générales qu'en AD gaussienne
 - Maximum de vraisemblance au lieu de moindres carrés (régression linéaire de Y sur les X_j)
 - Prise en charge facile des X qualitatifs (logiciels)

■ Mais:

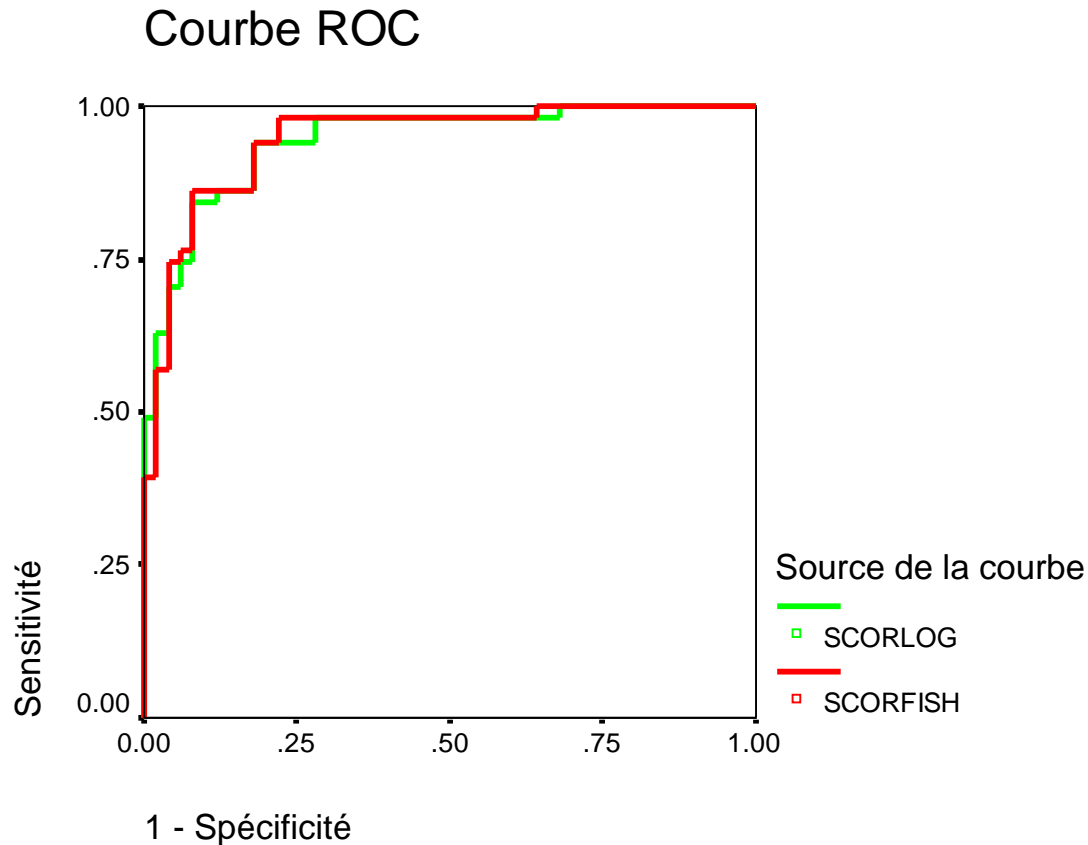
- Erreurs standard asymptotiques , bootstrap en AD
- Non convergence en cas de séparation parfaite.
Fisher existe toujours



- Maximum de vraisemblance conditionnel: non optimal dans le cas gaussien standard
- L'AD peut aussi traiter les variables qualitatives, et de manière plus robuste grâce aux contraintes de sous-espace (Disqual)

- Querelle largement idéologique (modélisation versus analyse des données)
 - L'AD est aussi un modèle, mais sur les lois des X/Y , la logistique sur les lois de Y/X
- En pratique différences peu nettes: fonctions de score souvent très proches
 - « 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)

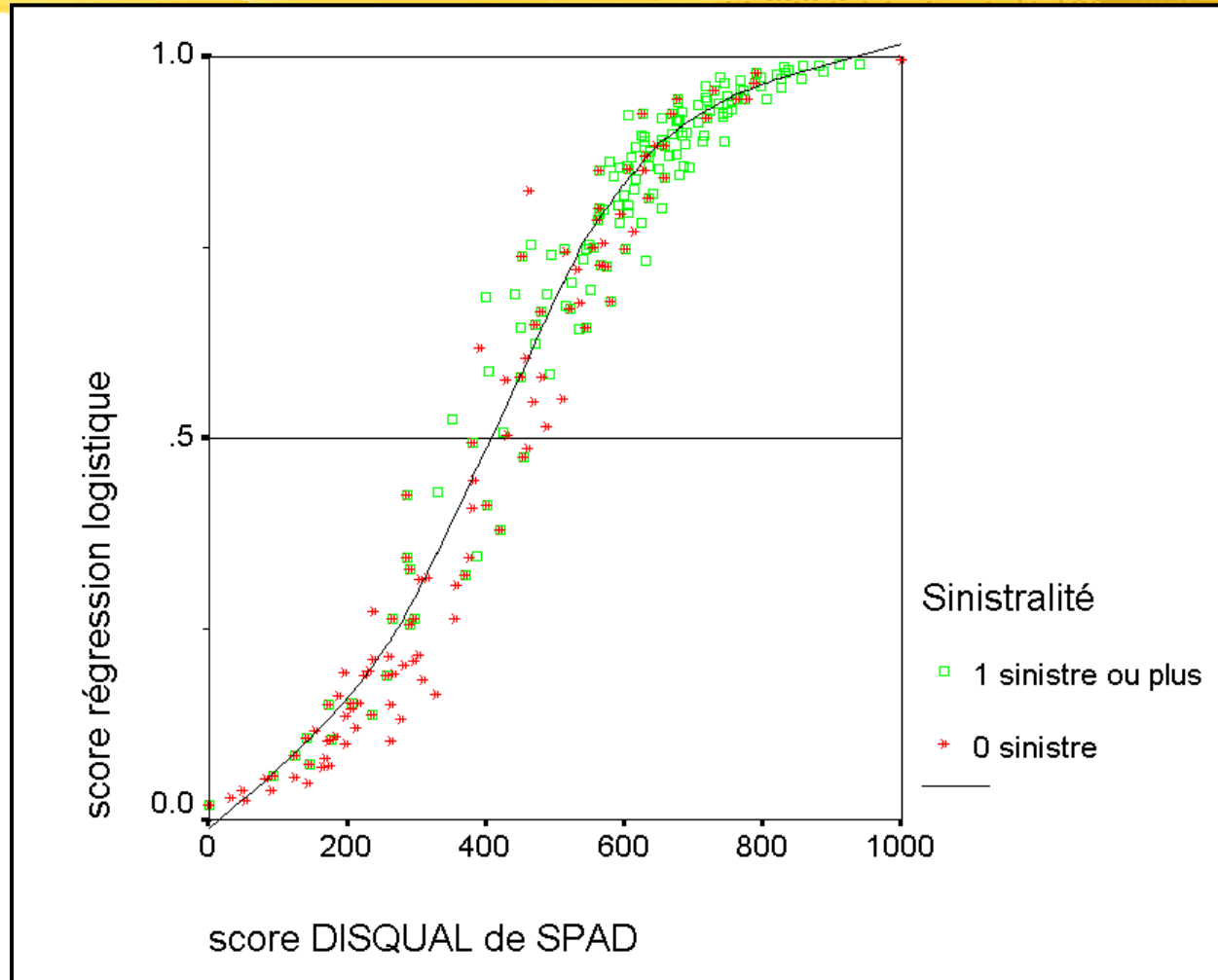
Infarctus: comparaison Fisher et logistique




Zone sous la courbe

Variable(s) de	Zone
SCORFISH	.945
SCORLOG	.943

Assurance



- 
- Usages souvent différents: AD pour classer, logistique pour modéliser (facteurs de risque)
 - Logistique aussi utilisée en scoring
 - Si l'objectif est de classer:
 - On ne fait plus de la science mais de l'aide à la décision
 - Mieux vaut essayer les deux méthodes.
 - Mais comment les comparer?
 - Le vrai critère de choix est la performance en généralisation

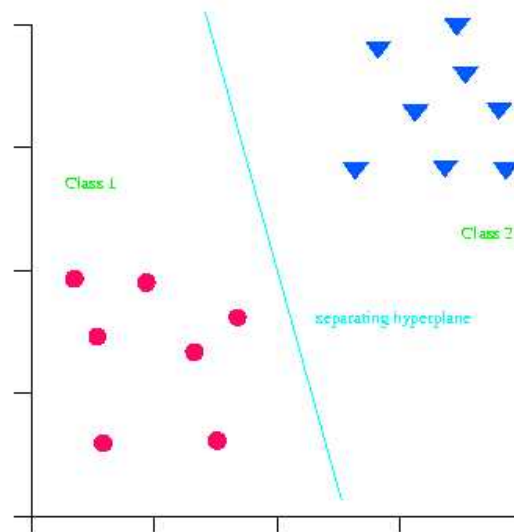
5ème partie:

les SVM (séparateurs à vaste
marge ou support vector
machines)

V.1 Du perceptron aux SVM

- Algorithme de Rosenblatt (1958), la première « machine à apprendre »

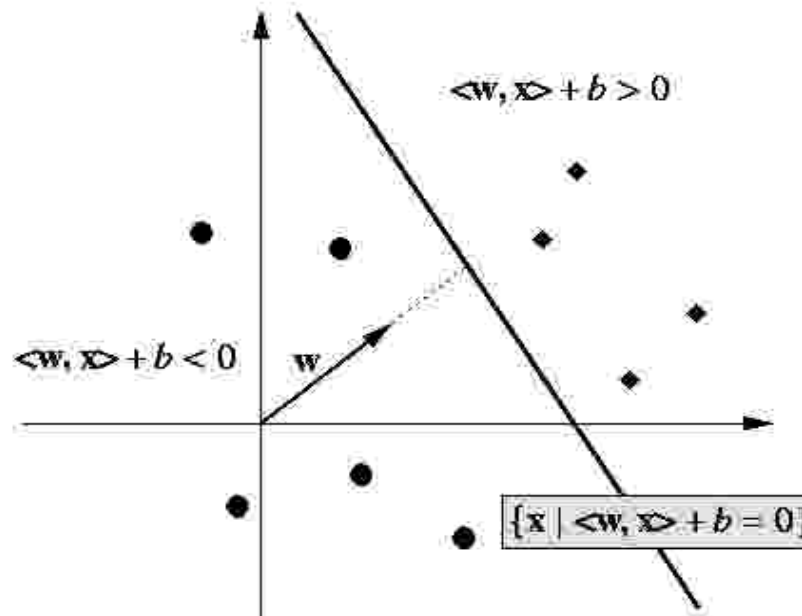
A perceptron algorithm separates linearly separable training data with no error.



Du perceptron aux SVM

■ Equation de l'hyperplan séparateur

$$f(\mathbf{x}) = \mathbf{w}'\mathbf{x} + b = 0$$



Б. Шабалин, МГУ, 8 Decembar 2011

Un peu de géométrie

- Equation d'un hyperplan:

$$f(\mathbf{x}) = \mathbf{w}'\mathbf{x} + b = \mathbf{x}'\mathbf{w} + b = 0$$

- Coefficients définis à un facteur près:

- $b=1$ ou $\|\mathbf{w}\|=1$

- Distance à l'hyperplan:

$$d = \frac{|\mathbf{w}'\mathbf{x} + b|}{\|\mathbf{w}\|}$$

- Minimiser la somme des distances au plan des observations mal classées

$Y_i = 1$ mal classé si $w'x_i + b < 0$

$Y_i = -1$ mal classé si $w'x_i + b > 0$

$$\min \left(- \sum_{\text{mal classés}} y_i (\mathbf{w}'\mathbf{x}_i + b) \right)$$

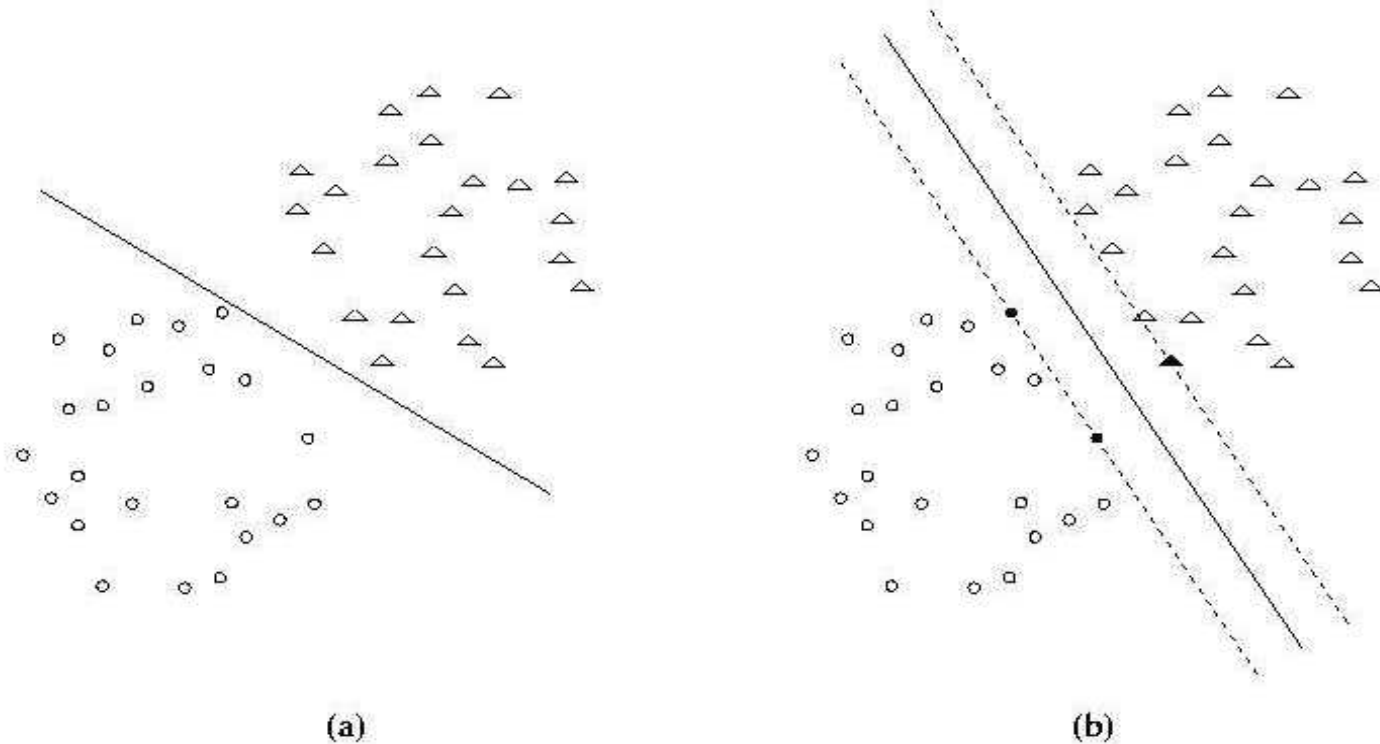
$$\text{gradient } \frac{\partial}{\partial \mathbf{w}} = - \sum_{\text{mal classés}} y_i \mathbf{x}_i \quad \frac{\partial}{\partial b} = - \sum_{\text{mal classés}} y_i$$

■ Gradient stochastique (obs. par obs.)

$$\begin{pmatrix} \mathbf{w} \\ b \end{pmatrix}_{n-1} + \rho \begin{pmatrix} y_i \mathbf{x}_i \\ y_i \end{pmatrix} \rightarrow \begin{pmatrix} \mathbf{w} \\ b \end{pmatrix}_n$$

- ρ coefficient d'apprentissage
- Solutions multiples dans le cas séparable selon l'initialisation
- Non convergence dans le cas non séparable

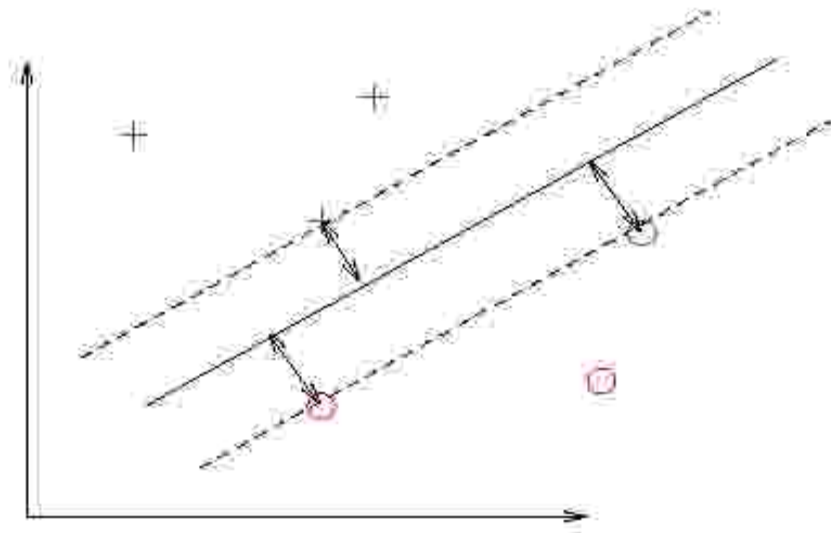
V.2 L'hyperplan optimal (Vapnik)



**Frontière avec « no man's land » maximal,
Hyperplan « épais »**

Hyperplan optimal

- Maximise la « marge » ou rayon du corridor: distance du point le plus proche à l'hyperplan



■ Cas séparable

- Marge C : tous les points sont à une distance $> C$

$$\max C \quad \text{sous } y_i(\mathbf{x}_i' \mathbf{w} + b) \geq C \text{ et } \|\mathbf{w}\| = 1$$

$$\text{contrainte équivalente: } y_i(\mathbf{x}_i' \mathbf{w} + b) \geq C \|\mathbf{w}\|$$

$$\text{ou } \|\mathbf{w}\| = \frac{1}{C} \quad \text{car } \mathbf{w} \text{ et } b \text{ définis à l'échelle près}$$

$$\min \|\mathbf{w}\| \quad \text{sous } y_i(\mathbf{x}_i' \mathbf{w} + b) \geq 1$$

Programme quadratique

■ Lagrangien:

$$\|\mathbf{w}\|^2 - 2 \sum \alpha_i [y_i (\mathbf{x}_i' \mathbf{w} + b) - 1]$$

■ D'où:

$$\mathbf{w} = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \quad \text{et} \quad \sum_{i=1}^n \alpha_i y_i = 0$$

■ Dual de Wolfe

$$\max \left[\sum \alpha_i - \frac{1}{2} \sum \sum \alpha_i \alpha_k y_i y_k \mathbf{x}_i' \mathbf{x}_k \right]$$

$$\text{avec } \alpha_i \geq 0 \text{ et } \sum_{i=1}^n \alpha_i y_i = 0$$

■ Conditions de Kuhn et Tucker:

$$\alpha_i \left[y_i (\mathbf{x}_i' \mathbf{w} + b) - 1 \right] = 0$$

$$\text{Si } \alpha_i > 0 \text{ alors } y_i (\mathbf{x}_i' \mathbf{w} + b) = 1$$

$$\text{Si } y_i (\mathbf{x}_i' \mathbf{w} + b) > 1 \text{ alors } \alpha_i = 0$$

- W , donc l'hyperplan, ne dépend que des points **supports** où les α_i sont non nuls.

■ Solution

$$\mathbf{w} = \sum_{\alpha_i > 0}^n \alpha_i y_i \mathbf{x}_i$$

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

- $f(\mathbf{x})$ ne dépend que des points supports
- est une combinaison linéaire des variables (score)
- règle de décision selon le signe de $f(\mathbf{x})$

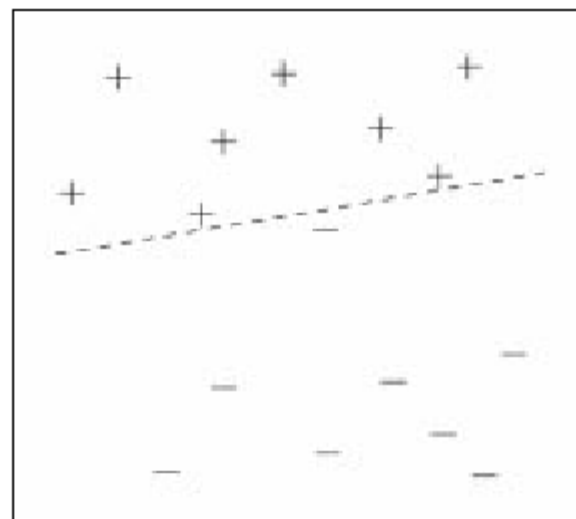
■ L'hyperplan optimal ne dépend que des points proches (différent de Fisher)

■ VC dimension:

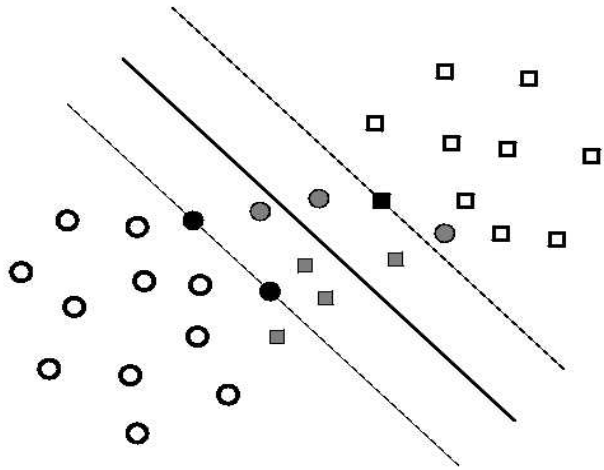
$$h \leq \frac{R^2}{C^2} \quad \text{où } \|x\| \leq R$$

■ Plus la marge est grande, meilleure est la robustesse en principe.

■ Mais pas toujours :



V.3 Le cas non séparable

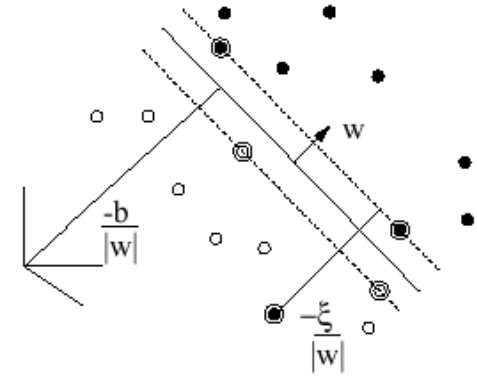


Deux solutions:

- modifier le critère
- changer d'espace pour rendre le problème linéairement séparable

■ Variables d'écart

$$\min \|\mathbf{w}\| \quad \text{sous } y_i (\mathbf{x}_i' \mathbf{w} + b) \geq 1 - \xi_i$$
$$\text{et } \sum \xi_i < \gamma$$



- On borne la proportion de points tombant du mauvais côté.
- La solution ne dépend que des points supports où : $y_i (\mathbf{x}_i' \mathbf{w} + b) > 1 - \xi_i$

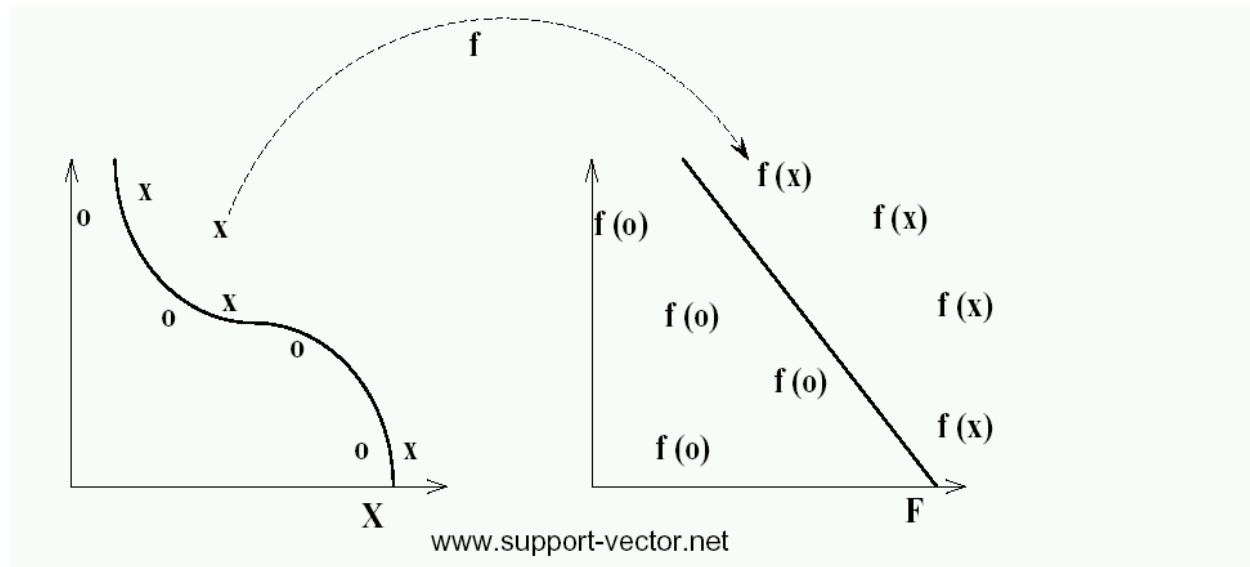
- 
- Formulation équivalente:

$$\min \left[\|\mathbf{w}\|^2 + C \sum \xi_i \right] \quad \text{avec } y_i (\mathbf{x}_i' \mathbf{w} + b) \geq 1 - \xi_i$$

- C contrôle le trade-off entre la marge et l'erreur.
- $0 < \alpha_i < \gamma$

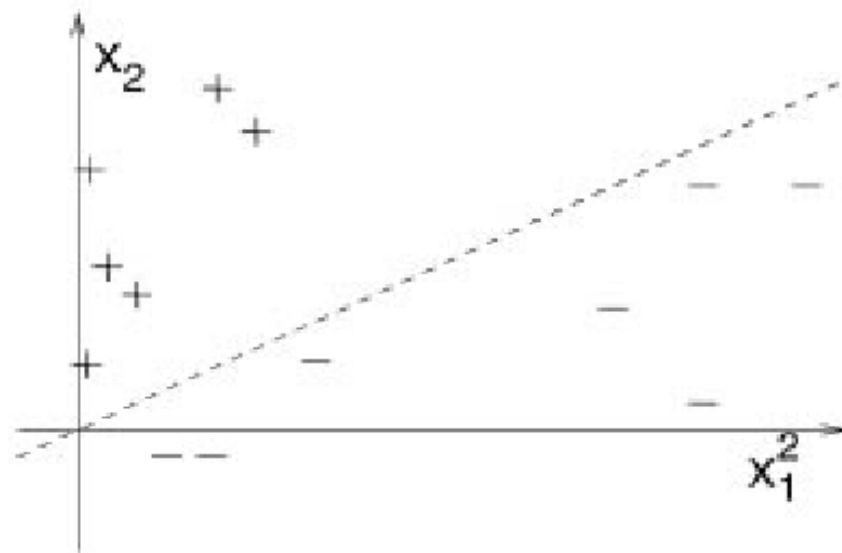
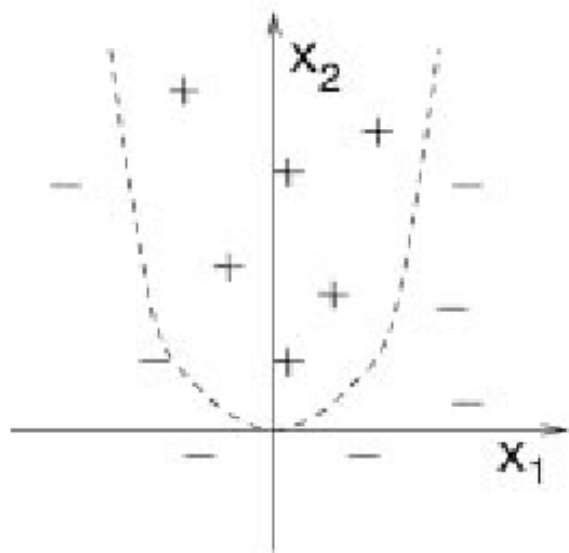
SVM non-linéaires

- Passage dans un espace de données transformées (« feature space ») de grande dimension
- Un séparateur linéaire dans $\Phi(E)$ donne un séparateur non-linéaire dans E .

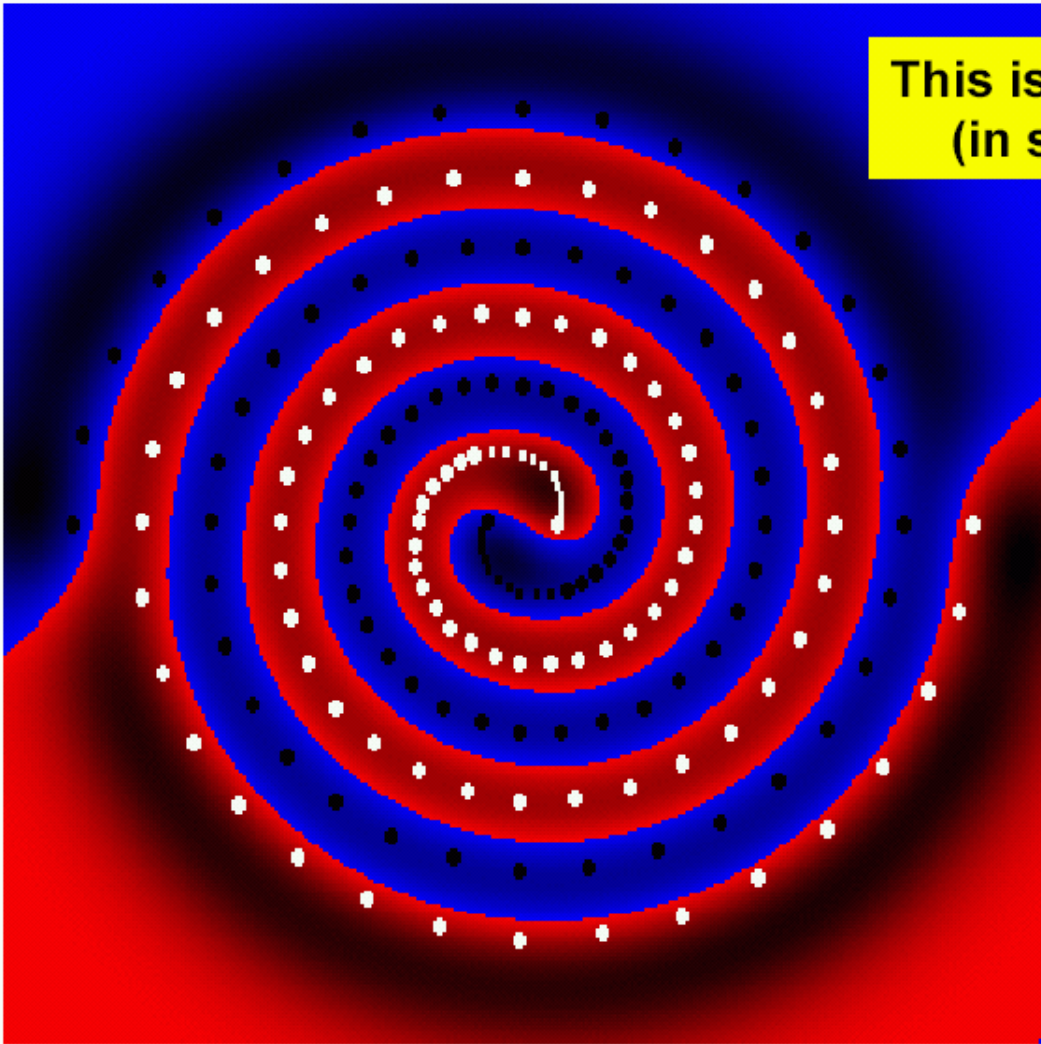


Input Space: $\vec{x} = (x_1, x_2)$ (2 Attributes)

Feature Space: $\Phi(\vec{x}) = (x_1^2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, 1)$ (6 Attributes)



**This is a hyperplane!
(in some space)**



www.support-vector.net/nello.html

Solution

$$\left\{ \begin{array}{l} \max \left[\sum \alpha_i - \frac{1}{2} \sum \sum \alpha_i \alpha_k y_i y_k \langle \Phi(\mathbf{x}_i) | \Phi(\mathbf{x}_k) \rangle \right] \\ 0 < \alpha_i < C \quad \text{et} \quad \sum \alpha_i y_i = 0 \end{array} \right.$$

$$\text{Solution } f(\mathbf{x}) = \sum_{i=1}^n \alpha_i y_i \langle \Phi(\mathbf{x}_i) | \Phi(\mathbf{x}) \rangle + b$$

- Ne dépend que des produits scalaires

Espaces de Hilbert à noyaux reproduisants

- Noyaux $K(x, x') = \Phi(x) \Phi(x')$
- Le « kernel trick »: choisir astucieusement K pour faire les calculs uniquement dans l'espace de départ.
- Exemple: $x = (x_1; x_2)$ $\Phi(x) = (x_1^2; \sqrt{2}x_1x_2; x_2^2)$
- Dans l'espace d'arrivée:

$$\begin{aligned}\Phi(x)\Phi(x') &= x_1^2 x_1'^2 + 2x_1x_2x_1'x_2' + x_2^2 x_2'^2 \\ &= (x_1x_1' + x_2x_2')^2 = (xx')^2\end{aligned}$$

- On peut donc calculer le produit scalaire dans $\Phi(E)$ sans utiliser Φ

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

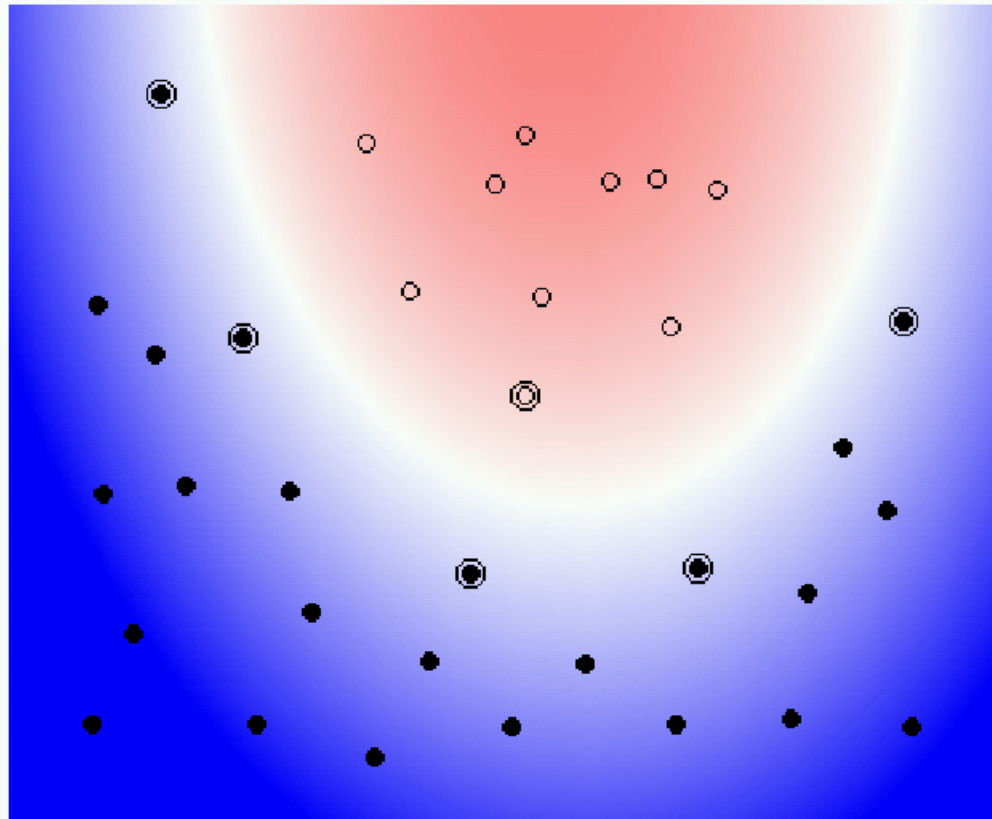
- Conditions de Mercer pour avoir un noyau:
 - $k(x_i; x_j)$ terme général d'une matrice sdp

Exemples de noyaux

- Linéaire $K(x;x') = \langle x;x' \rangle$
- Polynomial $K(x;x') = (\langle x;x' \rangle)^d$ ou $(\langle x;x' \rangle + 1)^d$
- Gaussien (radial basis)
$$K(x;x') = \exp(-(\|x-x'\|^2)/\sigma^2)$$

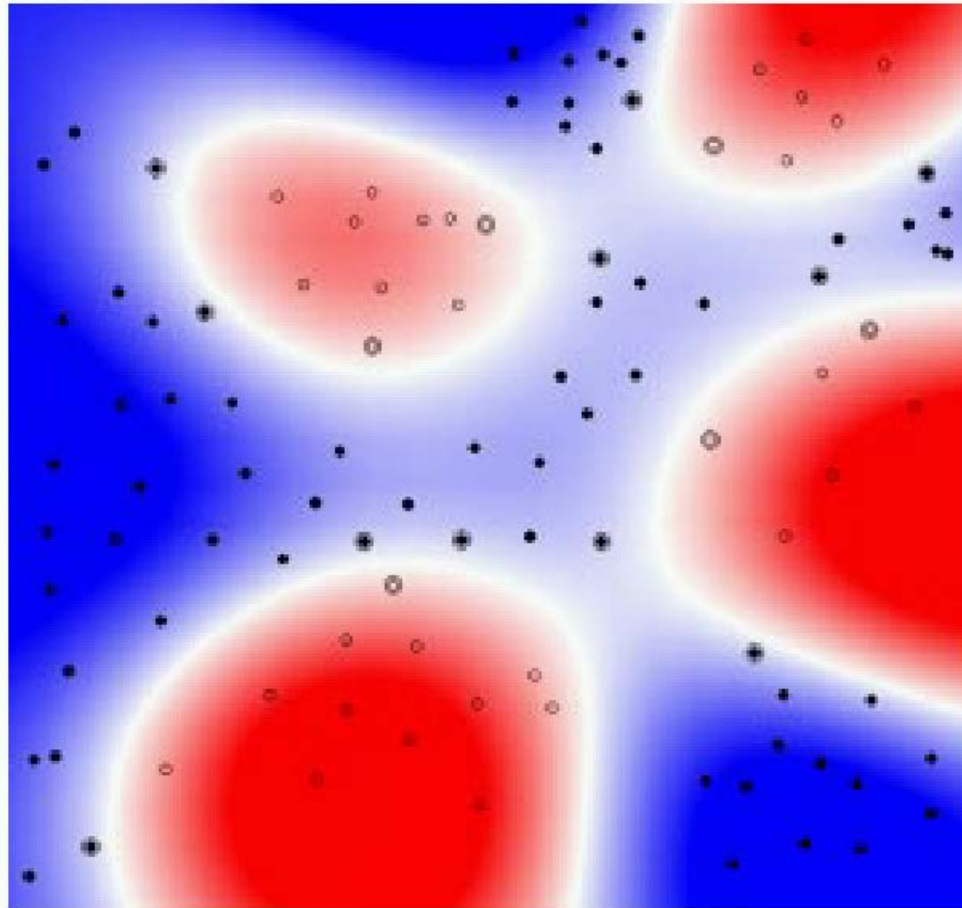
Example: SVM with Polynomial of Degree 2

Kernel: $K(\vec{x}_i, \vec{x}_j) = [\vec{x}_i \cdot \vec{x}_j + 1]^2$

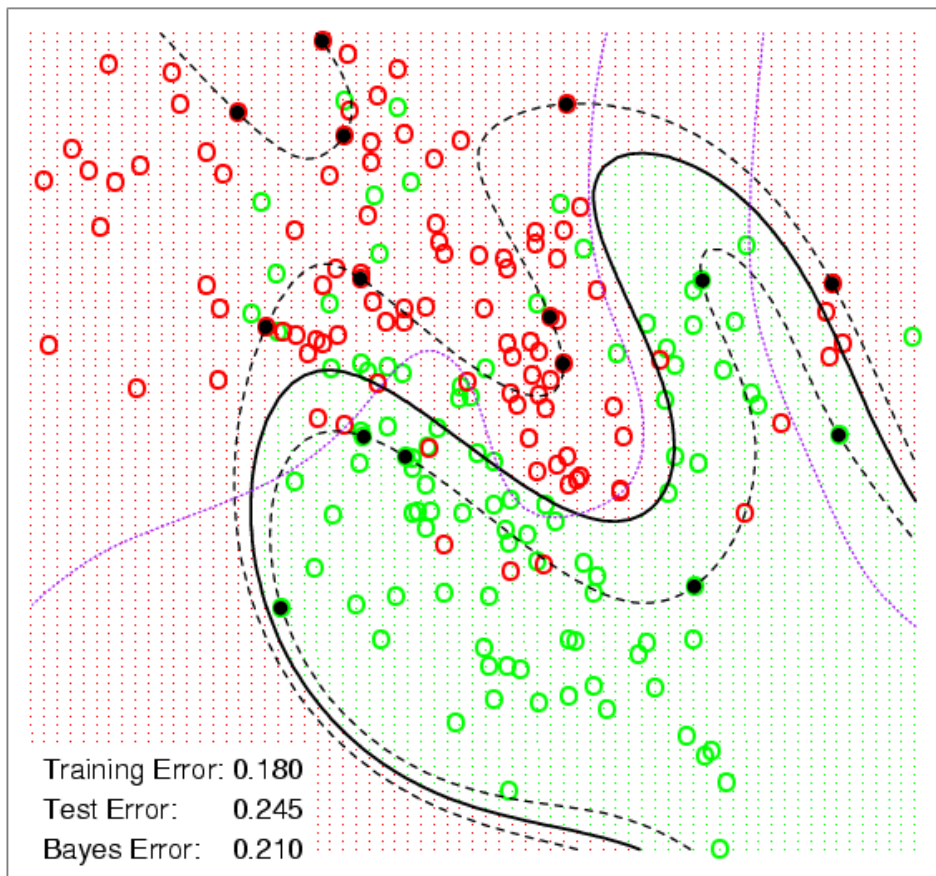


Example: SVM with RBF-Kernel

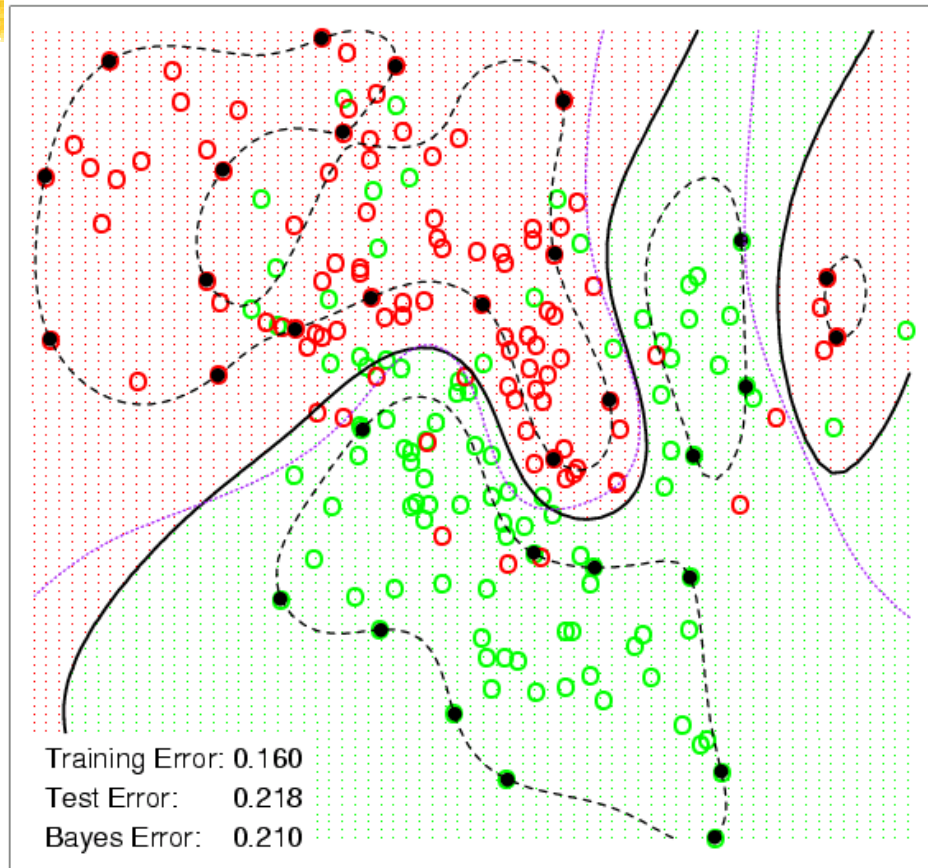
Kernel: $K(\vec{x}_i, \vec{x}_j) = \exp(-\gamma|\vec{x}_i - \vec{x}_j|^2)$



SVM - Degree-4 Polynomial in Feature Space



SVM - Radial Kernel in Feature Space



Le problème de la généralisation.

les SVM évitent :



- Le risque de surapprentissage (« curse of dimensionality »)
- L'infinité de solutions dans le cas séparable (problème mal posé)

Le problème de la généralisation.

les SVM :

- Contrôlent la capacité de généralisation en augmentant la marge car:

$$h \leq \frac{R^2}{C^2} \quad \text{où } \|x\| \leq R$$

- Ne dépend pas de la dimension de l'espace (éventuellement ∞)

Approches voisines



- LS-SVM, GDA (Baudat, Anouar) : fonction de Fisher dans le feature space

Quelques références



- <http://www.kernel-machines.org>
 - Th.Joachims « tutorial SVM »
 - C.Burges « a tutorial on SVM for pattern recognition »
- O.Bousquet « introduction aux SVM », <http://www.math.u-psud.fr/~blanchard/gtsvm/intro.pdf>
- J.Suykens et al. « Least squares support vector machines », World Scientific, 2002
- Logiciels:
 - <http://svm.dcs.rhbnc.ac.uk/pagesnew/GPat.shtml>
 - <http://www.csie.ntu.edu.tw/~cjlin/>

6 éme partie: validation

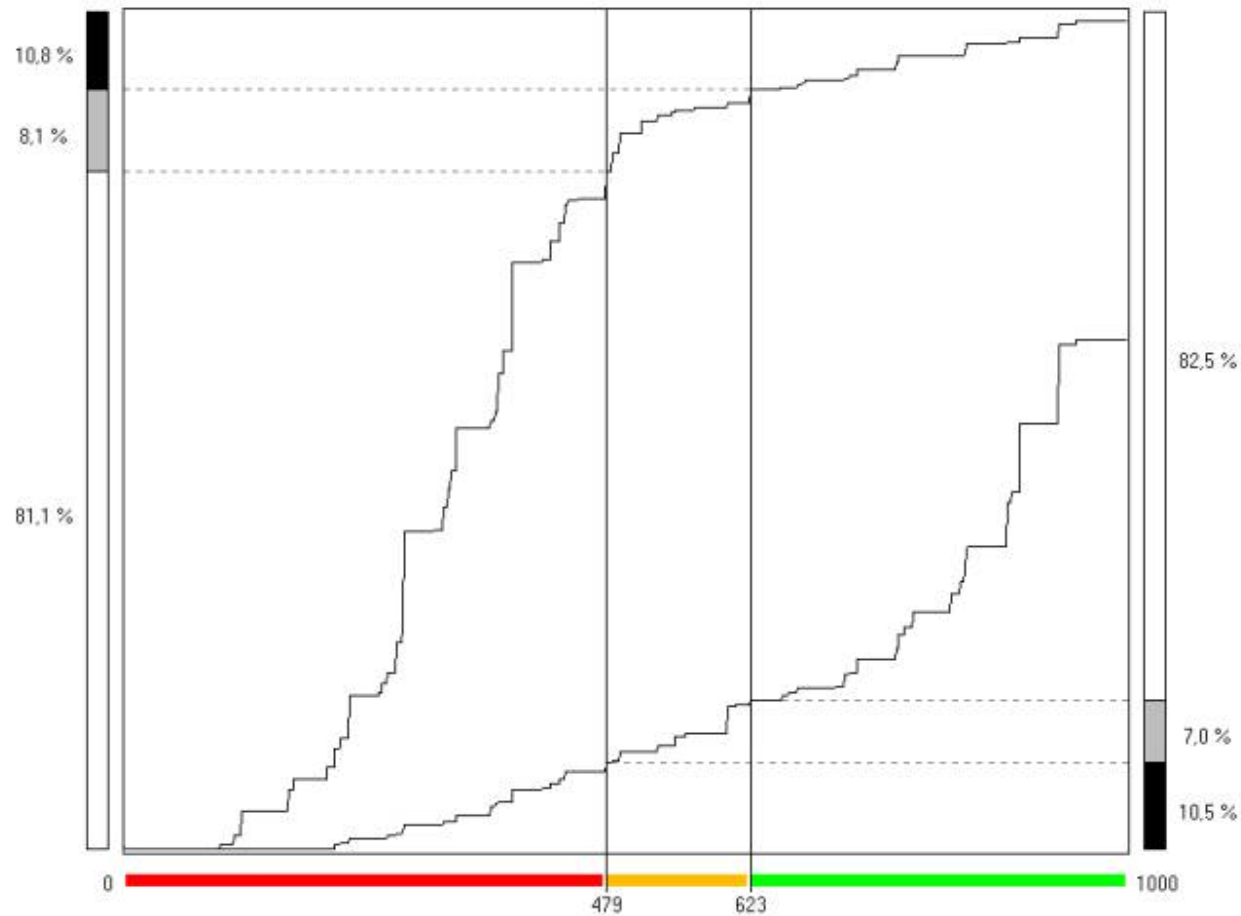
- VI-1 Qualité d'un score
- VI-2 Qualité d'une règle de classement

VI-1 Qualité d'un score

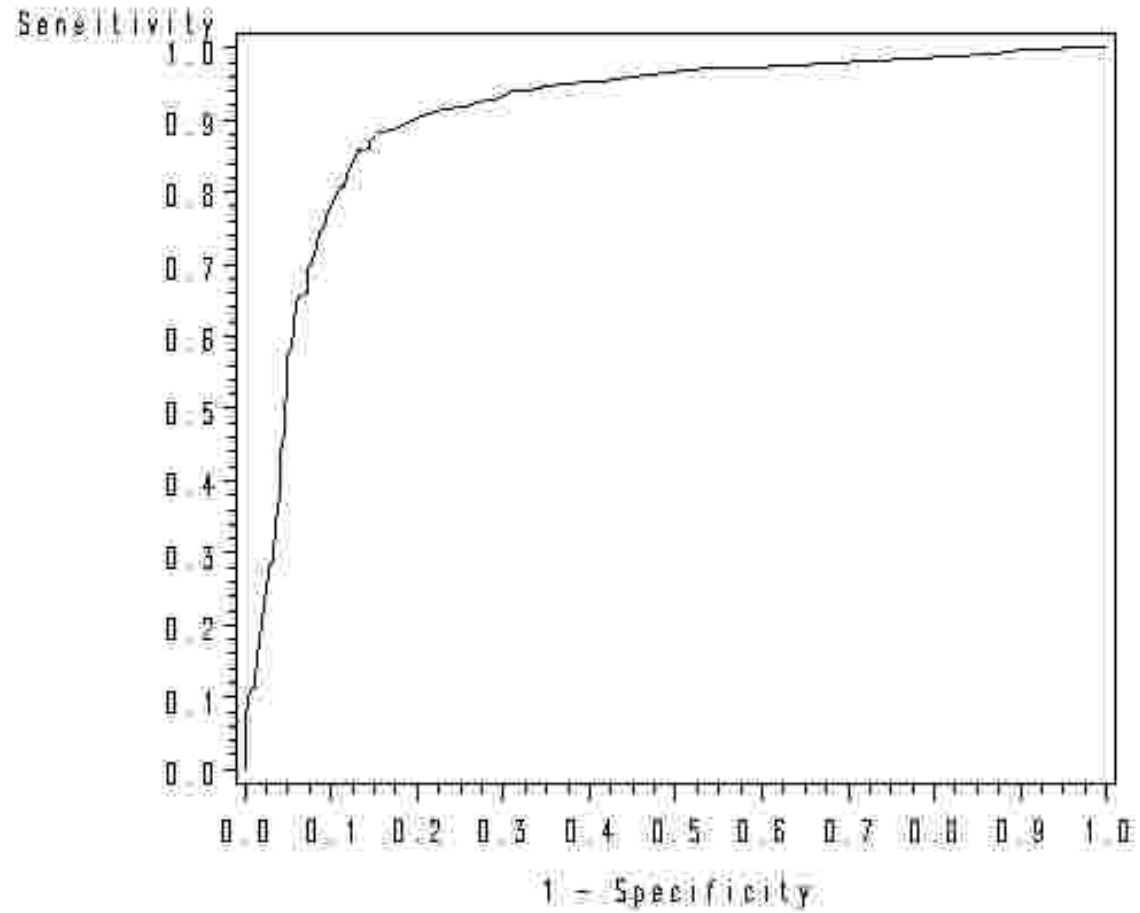


- Qu'il soit obtenu par Fisher, logistique ou autre (une probabilité est un score...)
- Comparaison des distributions du score sur les deux groupes
 - densités
 - fonctions de répartition

Fonctions de répartition

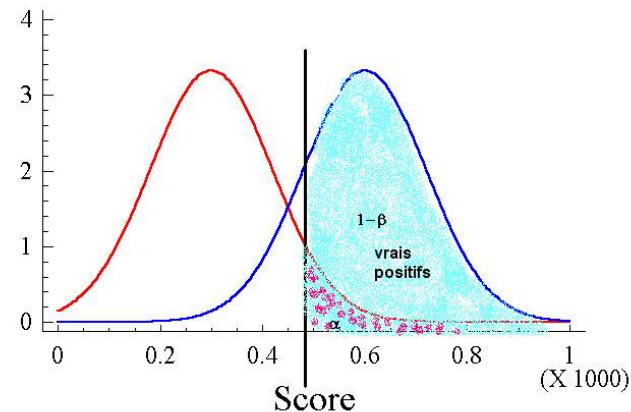
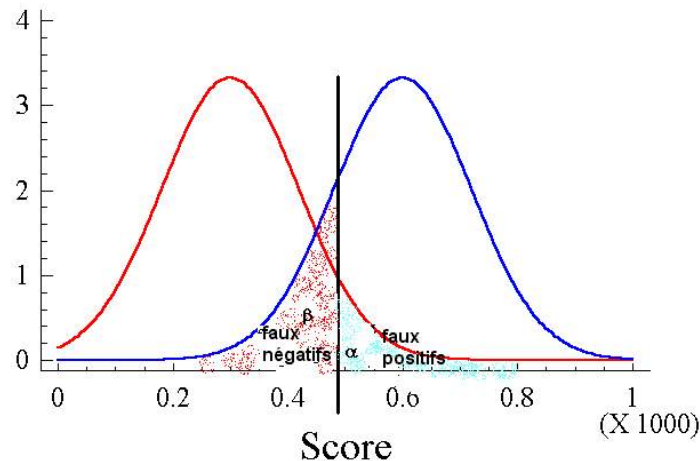


Courbe ROC



Courbe ROC: interprétation

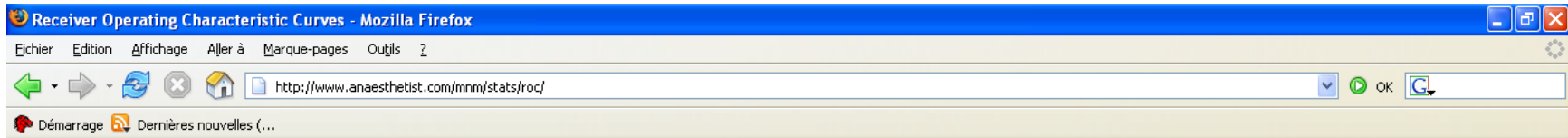
- Groupe à détecter G_1 : scores élevés
- Sensibilité $1-\beta = P(S > s / G_1)$: % de vrais positifs
- Spécificité $1-\alpha = P(S < s / G_2)$: % de vrais négatifs



Courbe ROC: interprétation (2)

- Evolution de $1-\beta$ puissance du test en fonction de α , risque de première espèce lorsque le seuil varie
- Proportion de vrais positifs en fonction de la proportion de faux positifs

Un site: <http://www.anaesthetist.com/mnm/stats/roc/>

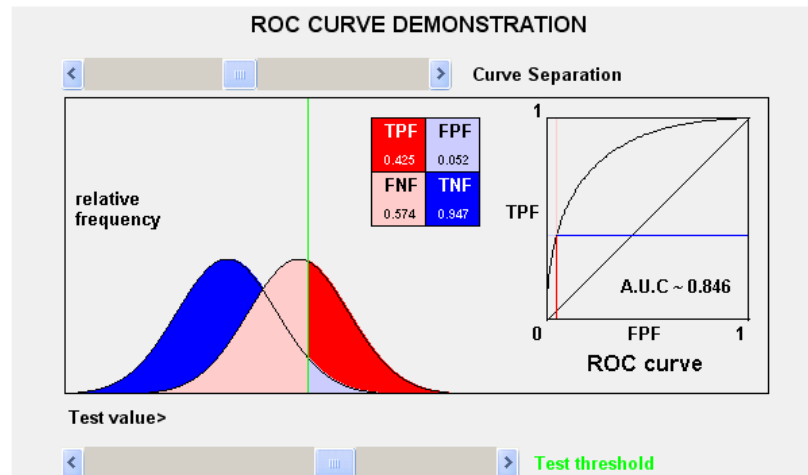


Playing with ROCs

In this section we will fool around with ROCs. We will:

1. [Create](#) ROC curves;
2. Find out why the area under the ROC curve is [non-parametric](#), and why this is important;
3. Learn to calculate required [sample sizes](#);
4. Compare the areas under [two ROC curves](#);
5. Examine the effects of [noise](#), a [bad 'gold standard'](#), and [other sources of error](#).

Let's play some more. In the following example, see how closely the two curves are superimposed, and how flat the corresponding ROC curve is! This demonstrates an important property of ROC curves - the greater the overlap of the two curves, the smaller the area under the ROC curve.



Vary the curve separation using the upper "slider" control, and see how the ROC curve changes. When the curves overlap almost totally the ROC curve turns into a diagonal line from the bottom left corner to the upper right corner. What does this mean?

Surface sous la courbe ROC

- Surface théorique sous la courbe ROC:
 $P(X_1 > X_2)$ si on tire au hasard et indépendamment une observation de G_1 et une observation de G_2

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

- Estimation non-paramétrique de la surface:
 - Proportion de paires concordantes $c = \frac{n_c}{n_1 n_2}$

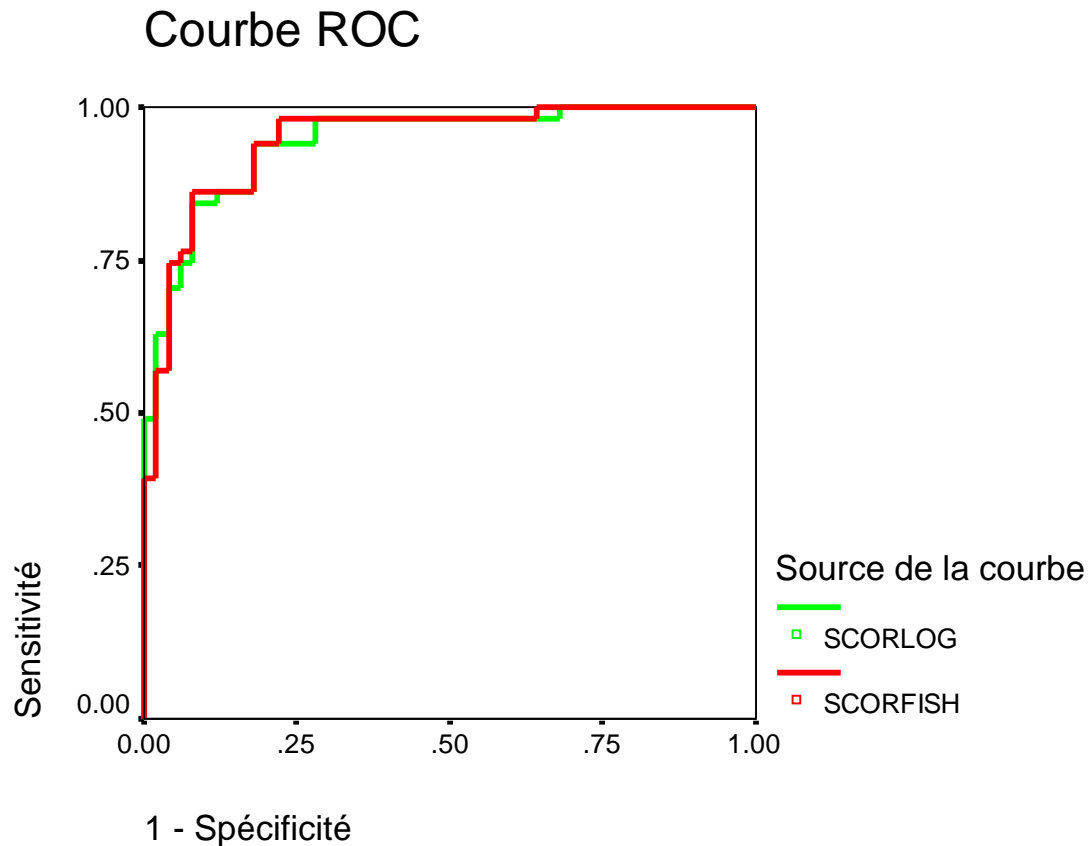
mesures de concordance

- Coefficients d'association entre les probabilités calculées et les réponses observées.
- Paires formées d'une obs où $Y=1$ et d'une où $Y=0$.
 - Nombre de paires $t=n_1n_2$ $n=n_1+n_2$
- Si l'observation telle que « $Y = 1$ » a une probabilité estimée que $Y = 1$, plus grande que celle de l'observation où « $Y = 0$ » la paire est concordante.
- nc = nombre de paires concordantes; nd = nombre de paires discordantes; $t - nc - nd$ = nombre d'ex-aequo

Courbe ROC: propriétés

- Courbe ROC et surface sont des mesures intrinsèques de séparabilité, invariantes pour toute transformation monotone croissante du score
- La surface est liée aux statistiques U de Mann-Whitney et W de Wilcoxon $n_c = U$
$$U + W = n_1 n_2 + 0.5 n_1 (n_1 + 1)$$
- $AUC = U / n_1 n_2$

Infarctus: comparaison Fisher et logistique



Zone sous la courbe

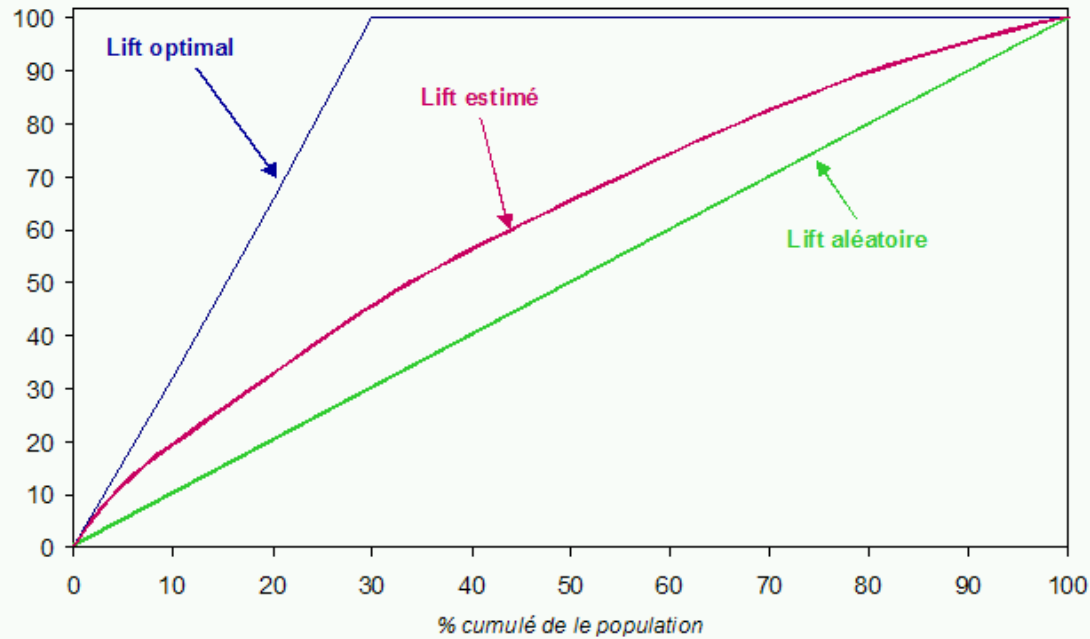
Variable(s) de	Zone
SCORFISH	.945
SCORLOG	.943

Autres mesures

- D de Somers = $(nc - nd) / t$
- Gamma = $(nc - nd) / (nc + nd)$
- Tau-a = $2 (nc - nd) / n(n-1)$
- **Indice de Gini**
 - Double de la surface entre la courbe ROC et la diagonale
$$G = 2AUC - 1$$
 - En l'absence d'ex-aequo: G identique au D de Somers
- La capacité prédictive du modèle est d'autant meilleure que ces indices sont proches de 1.

Courbe de lift

% de la cible



Surface sous la courbe lift

- Pourcentage d'individus ayant un score $> s$

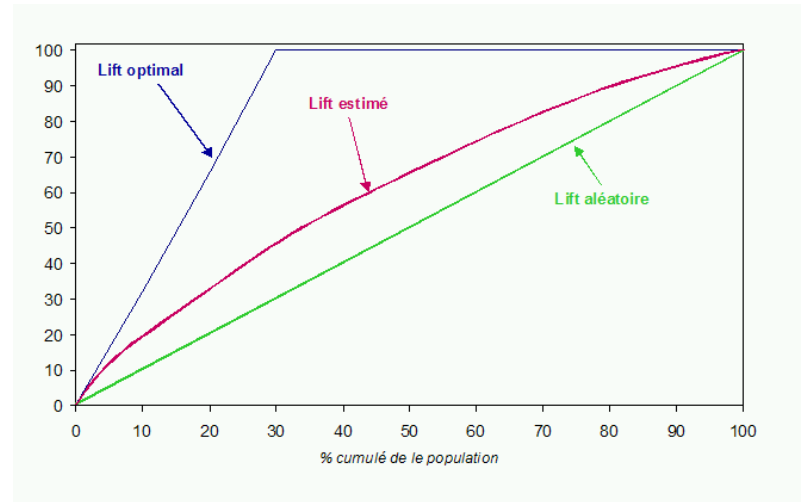
$$p_1(1-\beta) + (1-p_1)\alpha$$

- Surface

$$\begin{aligned} L &= \int (1-\beta) d \{ p_1(1-\beta) + (1-p_1)\alpha \} = \\ &= \left[p_1 \int (1-\beta) d(1-\beta) \right] + \left[(1-p_1) \int (1-\beta) d\alpha \right] \\ &= \frac{p_1}{2} + (1-p_1)AUC \end{aligned}$$

Coefficient Ki (Kxen)

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



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

VI-2 Qualité d'une règle de classement

■ Tableau de classement :

- On classe des observations dont le groupe est connu :

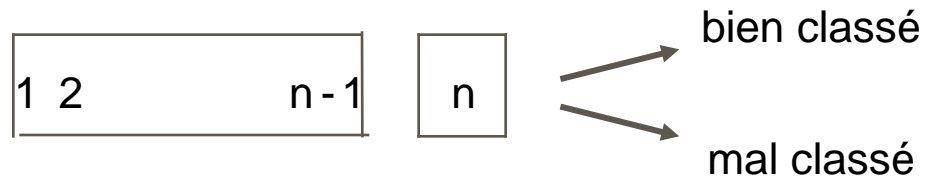
		groupe prédit	
		1	2
groupe réel	1	n_{11}	n_{12}
	2	n_{21}	n_{22}

- Pourcentage de bien classés : $\frac{n_{11} + n_{22}}{n}$

- Taux d'erreur de classement : $\frac{n_{12} + n_{21}}{n}$

Sur quel échantillon faire ce tableau ?

- Échantillon test d'individus supplémentaires.
 - Si on reclasse l'échantillon ayant servi à construire la règle (estimation des coefficients) : «méthode de resubstitution»
⇒ BIAIS
 - *surestimation du pourcentage de bien classés.*
- Solutions pour des échantillons de petite taille :
Validation croisée
 - n discriminations avec un échantillon test d'une unité : *% de bien classés sans biais (mais variance souvent forte)*

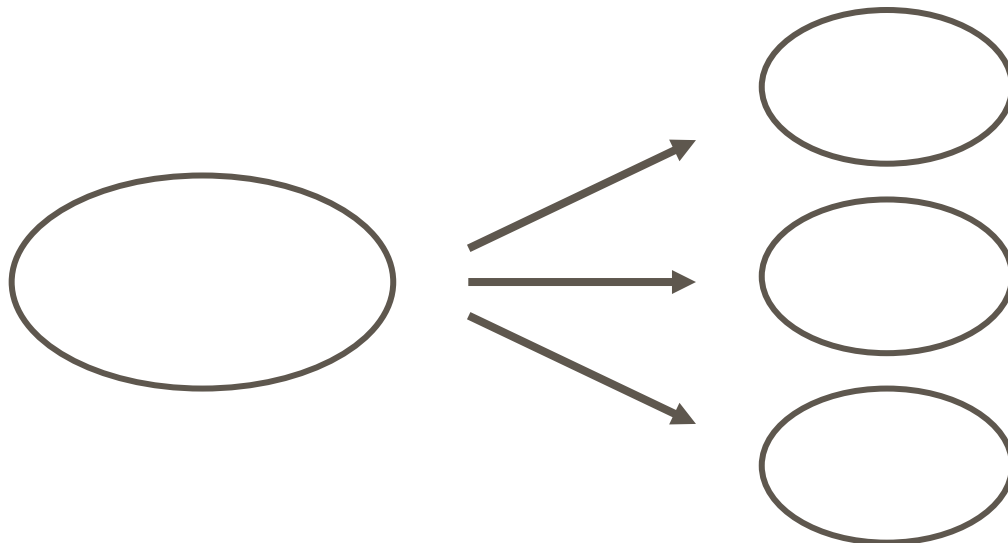


Bootstrap

- B analyses discriminantes d'où distributions empiriques des coefficients et du % de bien classés.

- Échantillon

B Réplifications par tirage avec remise de n parmi n



Septième partie: du choix de modèles à la théorie de l'apprentissage statistique

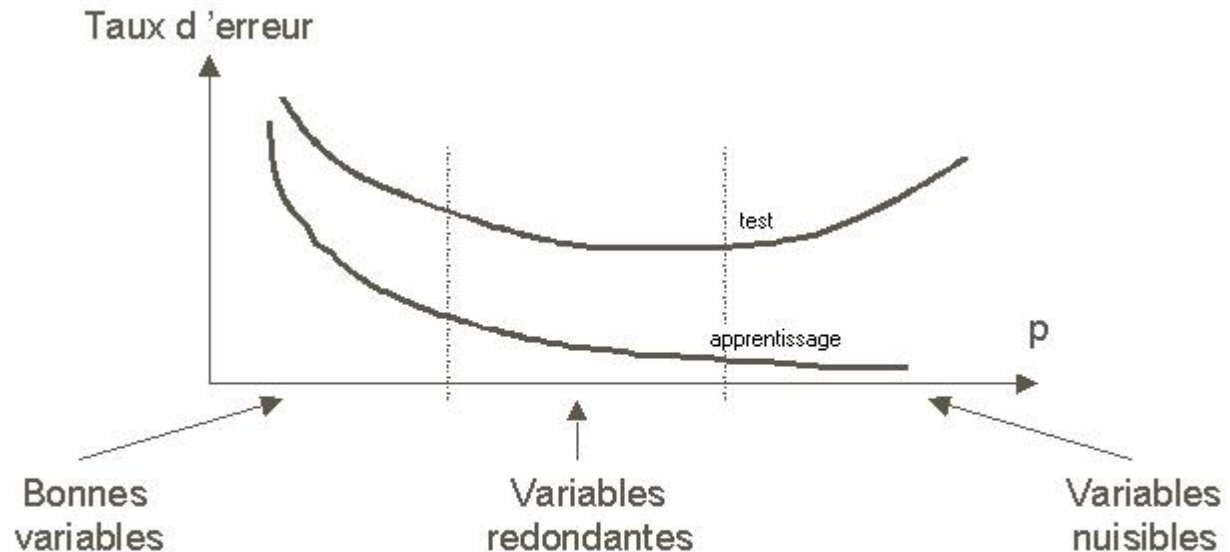
- VII.1 Sélection de variables
- VII.2 Choix de modèles par vraisemblance pénalisée
- VII.3 L'apprentissage selon Vapnik

VII.1 Sélection de variables

Réduire le nombre de prédicteurs

■ Pourquoi ?

- Économie
- Pertinence
- Stabilité



■ Comment ?

- Recherche exhaustive $2^p - 1$ sous-ensembles
- Méthodes pas à pas ascendantes, descendantes

Critères

- • Le % de bien classés n'est pas utilisé dans les logiciels classiques (SAS, SPSS...): trop de calculs.
- • Algorithmes usuels en analyse discriminante:
 - Critère Λ de Wilks : $\Lambda = |\mathbf{W}| / |\mathbf{V}|$
 - *On recherche à minimiser Λ : équivaut à maximiser D^2 pour $k=2$*
 - *Suppose implicitement la normalité*
 - Méthodes pas à pas : non optimales.
 - Pour $k=2$ recherche exhaustive par l'algorithme de Furnival et Wilson.

Tests de variables en AD

- Test d'apport d'une variable : Sous l'hypothèse de non apport :

$$\frac{n - k - p}{k - 1} \left(\frac{\Lambda_p}{\Lambda_{p+1}} - 1 \right) \sim F_{k-1 ; n-k-p}$$

- Test de non discrimination : (analyse de variance multidimensionnelle)

$$k = 3 \quad \frac{1 - \sqrt{\Lambda}}{\sqrt{\Lambda}} = \frac{p}{n - p - 2} F(2p ; n - p - 2)$$

pour $k > 3$ approximations

Sélection de variables en régression logistique

- Méthode ascendante :
 - Selon le score dans la proc logistic de SAS
- Méthode descendante:
 - Selon la statistique de Wald dans la proc logistic de SAS

VII.2 Choix de modèles par vraisemblance pénalisée

- Comparer des modèles ayant des nombres de paramètres différents: K nombre de paramètres à estimer.

Critère d'Akaike :

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

Critère de Schwartz :

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

- On préférera le modèle pour lequel ces critères ont la valeur la plus faible.

■ AIC et BIC ne sont semblables qu'en apparence

■ Théories différentes

■ AIC : approximation de la divergence de Kullback-Leibler entre la vraie distribution f et le meilleur choix dans une famille paramétrée

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

Asymptotiquement:

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

■ BIC : choix bayésien de modèles

m modèles M_i paramétrés par θ_i de probabilités *a priori* $P(M_i)$ égales.

Distribution *a priori* de θ_i pour chaque modèle $P(\theta_i / M_i)$.

Distribution *a posteriori* du modèle sachant les données ou vraisemblance intégrée $P(\mathbf{x}/M_i)$

Choix du modèle le plus probable *a posteriori* revient à maximiser

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

$$P(M_i / \mathbf{x}) = \frac{e^{-0.5BIC_i}}{\sum_{j=1}^m e^{-0.5BIC_j}}$$

Comparaison AIC BIC

- Si n tend vers l'infini la probabilité que le *BIC* choisisse le vrai modèle tend vers 1, ce qui est faux pour l'*AIC*.
- *AIC* va choisir le modèle qui maximisera la vraisemblance de futures données et réalisera le meilleur compromis biais-variance
- L'*AIC* est un critère prédictif tandis que le *BIC* est un critère explicatif.
- Pour n fini: résultats contradictoires. *BIC* ne choisit pas toujours le vrai modèle: il a tendance à choisir des modèles trop simples en raison de sa plus forte pénalisation

AIC BIC réalistes?



- Vraisemblance pas toujours calculable.
- Nombre de paramètres non plus: ridge, PLS etc.
- « Vrai » modèle?
 - « tous les modèles sont faux ; certains sont utiles » G.Box

Vapnik: choisir selon la VC
dimension

VII.3 : La théorie de l'apprentissage statistique

- Une introduction aux théories de V.Vapnik (rédigée en collaboration avec Michel Bera, Kxen)

Un mathématicien russe arrivé aux USA en 92, qui travaille depuis chez NEC après les Bell (aujourd'hui AT&T) Labs.

Premiers papiers en russe dès 1972.

Premier livre chez Springer Verlag en 1982

US Medal en sciences en 1992.

Un troisième livre (800 pages) chez J. Wiley, en 1998

■ Norbert Wiener 1948

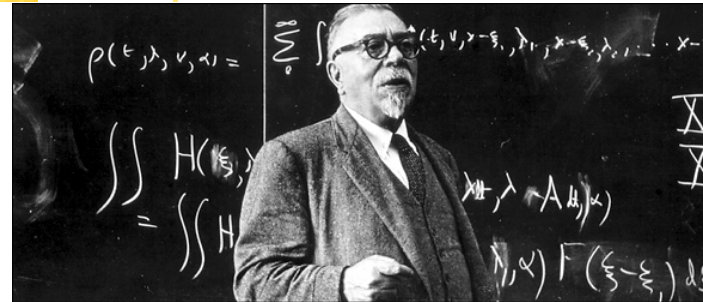
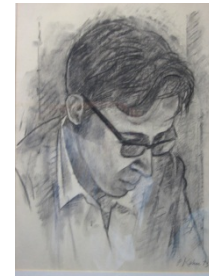


Image courtesy of the Research Laboratory of Electronics at [MIT](#).

■ Frank Rosenblatt 1962



■ Vladimir Vapnik 1982



Le problème de la boîte noire et l'apprentissage supervisé

- Etant donnée une entrée x , un système non déterministe renvoie une variable $y = f(x) + e$. On dispose de n paires (x_i, y_i) et on cherche une fonction qui approxime la fonction inconnue f .
- Deux conceptions:
 - Une bonne approximation est une fonction proche de f
 - Une bonne approximation est une fonction qui donne un taux d'erreur voisin de celui de la boîte noire

Risque d'apprentissage

- Apprentissage "supervisé"
- Y réponse à prédire, X prédicteurs
 - Y numérique **régression** ; binaire (-1;+1) **discrimination**
- Un modèle calcule un prédicteur

■ où: $\hat{y} = f(X, w)$

- f classe de fonction
- w est un paramètre qui définit le modèle, estimé sur l'ensemble d'apprentissage

■ Fonction de perte $L(y;f(\mathbf{x}, \mathbf{w}))$

■ *Régression* $L(y;f(\mathbf{x}, \mathbf{w}))=(y-f(\mathbf{x}))^2$

■ *Discrimination : taux (ou coût) d'erreur de classement*

- y et \hat{y} à valeurs dans $\{-1 ; +1\}$

$$L(y; \hat{y}) = \frac{1}{2}|y - \hat{y}| = \frac{1}{4}(y - \hat{y})^2$$

■ Risque (erreur de généralisation sur de nouvelles données $z = (X, y)$)


$$R = E(L) = \int L(z, w) dP(z)$$

■ Objectif impossible: minimiser sur w le Risque

■ $P(z)$ probabilité **inconnue**

■ On dispose seulement de n cas d'apprentissage (z_1, \dots, z_n) tirés suivant la loi $P(z)$, au lieu de minimiser R , on minimise le Risque Empirique :

$$R_{emp} = \frac{1}{n} \sum_{i=1}^n L(y_i; f(\mathbf{x}_i; \mathbf{w}))$$

- 
- Problème central en théorie de l'apprentissage:

Quelle est la relation entre le Risque R et le risque empirique R_{emp} ?

- Quelle est la capacité de généralisation de ce genre de modèle?

Le dilemme biais-variance

■ Modèle $y=f(x) + \varepsilon$, f estimé sur données d'apprentissage

■ Erreur de prédiction $y_0 - \hat{y}_0 = f(x_0) + \varepsilon - \hat{f}(x_0)$

Doublement aléatoire

■ Erreur quadratique moyenne de prédiction (risque R)

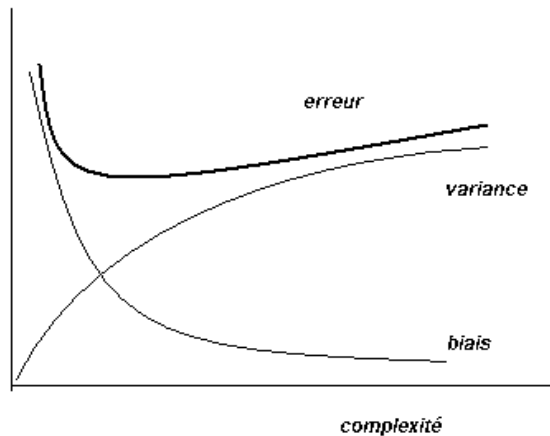
$$E(y_0 - \hat{y}_0)^2 = \sigma^2 + E\left(f(x_0) - \hat{f}(x_0)\right)^2 = \sigma^2 + \underbrace{\left(E\left(\hat{f}(x_0)\right) - f(x_0)\right)^2}_{\text{biais}} + \underbrace{V\left(\hat{f}(x_0)\right)}_{\text{variance}}$$

biais

variance

$$E(y_0 - \hat{y}_0)^2 = \sigma^2 + E\left(f(x_0) - \hat{f}(x_0)\right)^2 = \sigma^2 + \left(E\left(\hat{f}(x_0)\right) - f(x_0)\right)^2 + V\left(\hat{f}(x_0)\right)$$

- premier terme: aléa irréductible
- deuxième terme: carré du biais du modèle
- troisième terme: variance de la prédiction



Plus un modèle sera complexe plus le biais sera faible, mais au détriment de la variance.

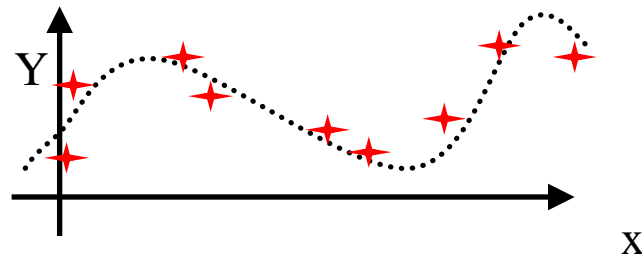
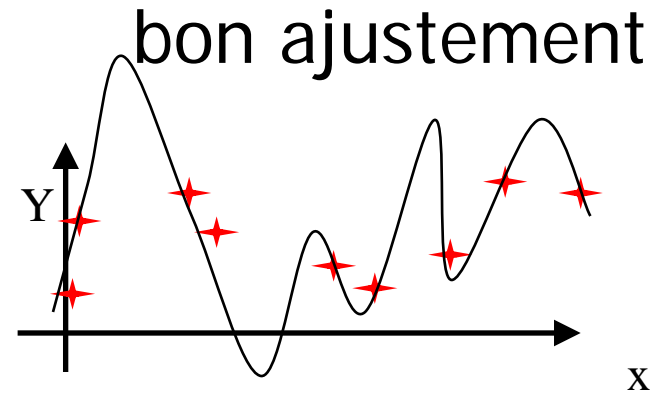
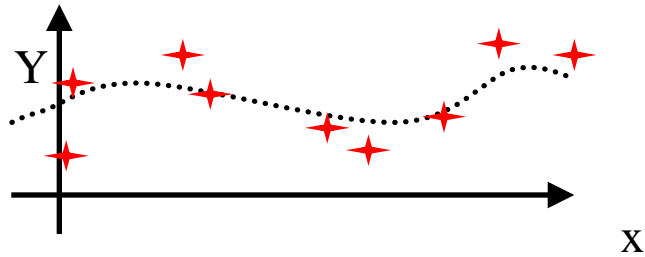
Mais comment mesurer la complexité?

Robustesse



- Modèle robuste: erreurs en apprentissage et en généralisation du même ordre de grandeur

■ Modele robuste



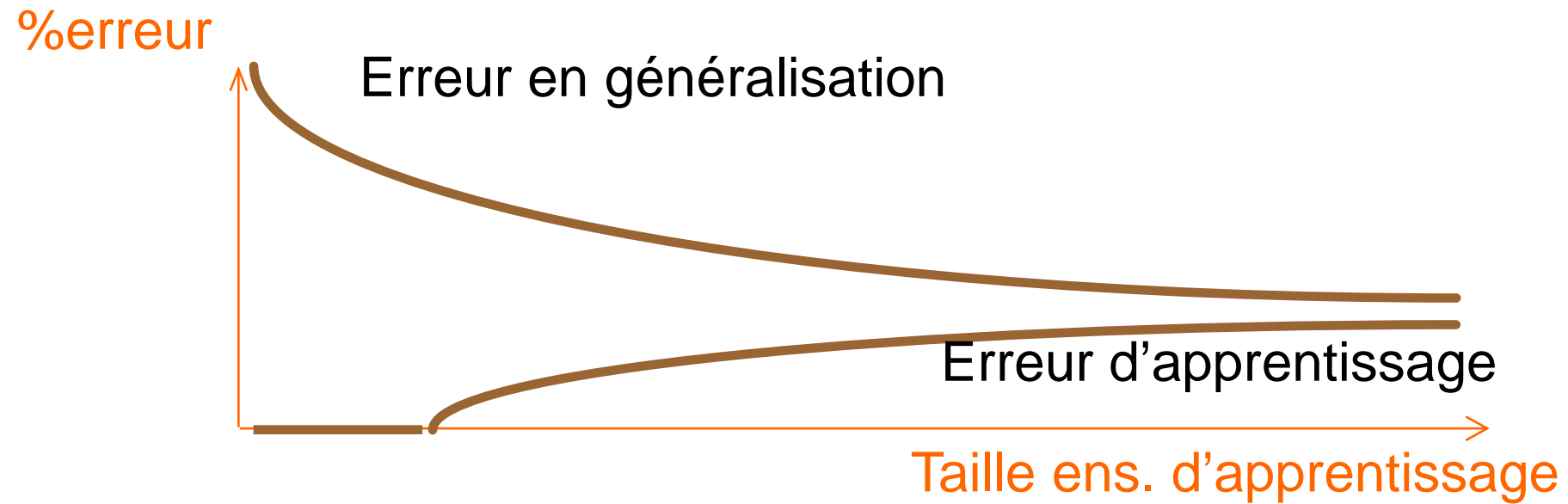
Compromis

Consistence

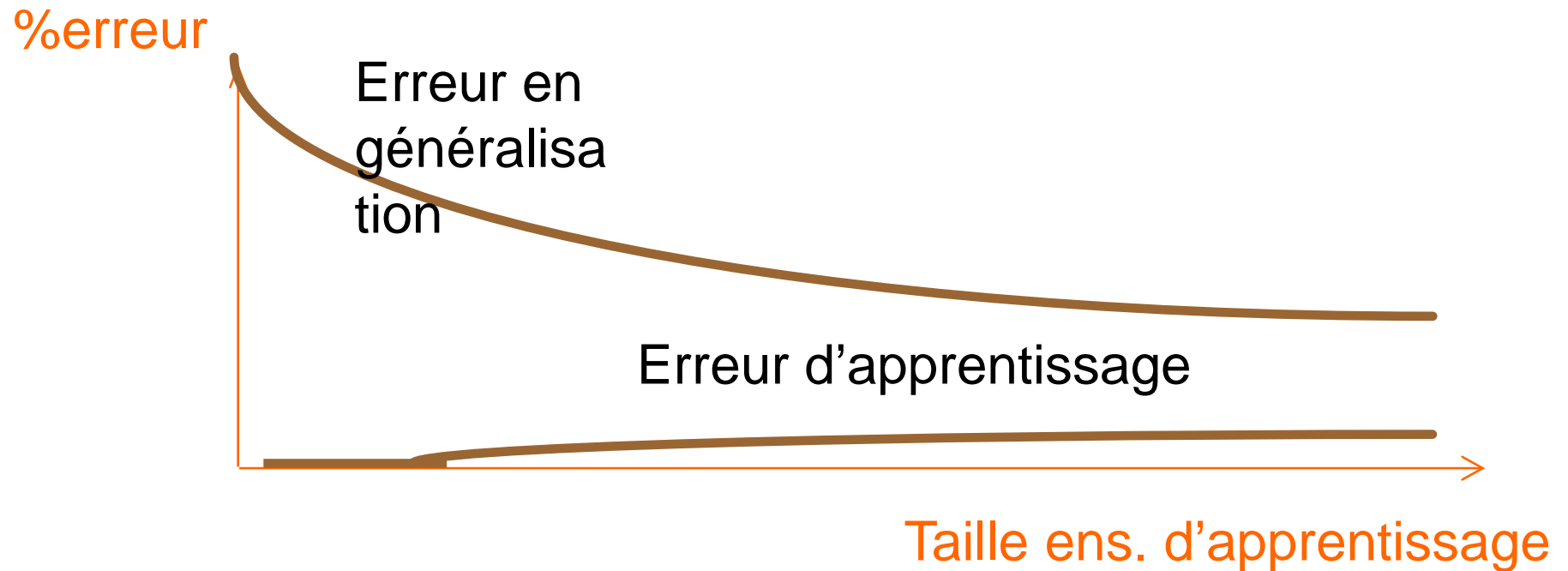


- Un processus d'apprentissage est consistant si l'erreur sur l'ensemble d'apprentissage converge, lorsque la taille de cet ensemble augmente, vers l'erreur en généralisation.

Apprentissage consistant



Apprentissage non consistant



Les quatre piliers de la théorie de l'apprentissage

- **1** Consistence (garantit la généralisation)
 - Sous quelles conditions un modèle peut-il généraliser?
- **2** Vitesse de convergence en fonction du nombre d'exemples (mesure de la généralisation)
 - Comment s'améliore la généralisation lorsque le nombre d'exemples augmente ?

Quatre piliers de la théorie de l'apprentissage



■ 3 Contrôle de la capacité de généralisation

- Comment contrôler efficacement la généralisation à partir de l'information contenue dans un ensemble d'apprentissage de taille finie ?

■ 4 Construire des algorithmes d'apprentissage

- Existe-t-il une stratégie pour construire des algorithmes qui garantissent, mesurent et contrôlent la capacité de généralisation de modèles d'apprentissage ?

La VC dimension

- Dimension de Vapnik-Cervonenkis: une mesure du pouvoir séparateur (complexité) d'une famille de fonctions

$$f(X, w) : \mathbb{R}^p \rightarrow \mathbb{R}$$

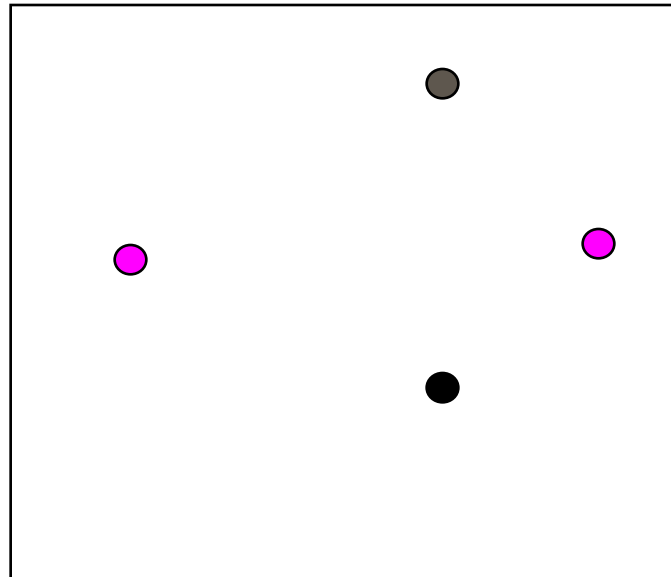
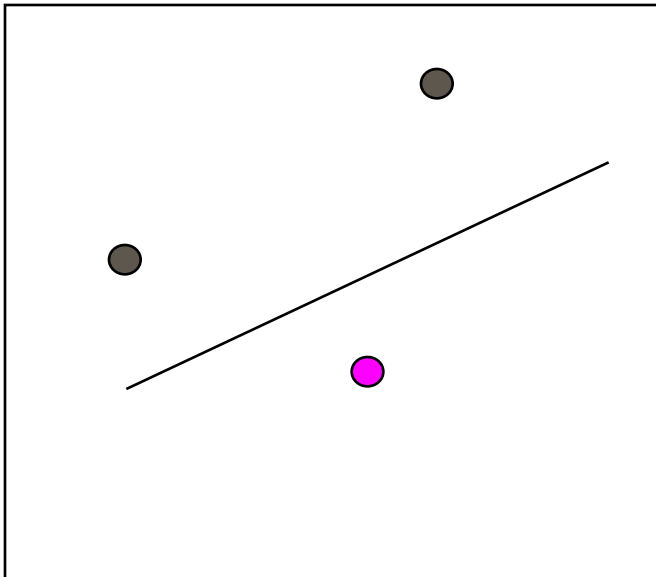
- VC dimension : un nombre entier attaché à une famille F de fonctions
- Chaque f de F – c'est-à-dire, pour un w donné – peut-être utilisé pour de la classification :
 - $f(X, w) \geq 0$: X classé en 1
 - $f(X, w) < 0$: X classé en -1

VC dimension suite

- Pour un échantillon de n points (x_1, \dots, x_n) de \mathbb{R}^p Il existe 2^n manières différentes de séparer cet échantillon en deux sous-échantillons
- Un ensemble F de fonctions $f(X, w)$ "hache" (*shatters*) l'échantillon si les 2^n séparations peuvent être faites par des $f(X, w)$ différentes de la famille F

Exemple

- En 2-D, les fonctions linéaires (droites) peuvent "hacher" 3 points, mais pas 4



Aucune ligne droite ne peut séparer les points noirs des points roses

Un ensemble de fonctions de $\mathbb{R}^p \rightarrow \mathbb{R}$ a la dimension h si :

- Il existe un jeu de h points de \mathbb{R}^p qui peut être "haché", quel que soit l'étiquetage des points
- Aucun ensemble de $h+1$ points ne peut être haché par cet ensemble de fonctions.

Quelques exemples

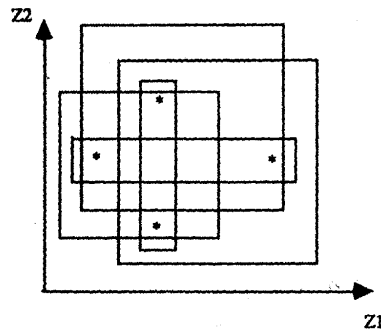
- La VC dimension de l'ensemble des hyperplans de \mathbb{R}^p est $p+1$
- Hyper-rectangles de \mathbb{R}^p parallèles aux axes

$$h=2p$$

(V.Cherkassky, F.Mulier, 1998)

- Sphères de \mathbb{R}^p

$$h=p+1$$



■ Mais les VC dimensions ne sont PAS égales au nombre de paramètres libres

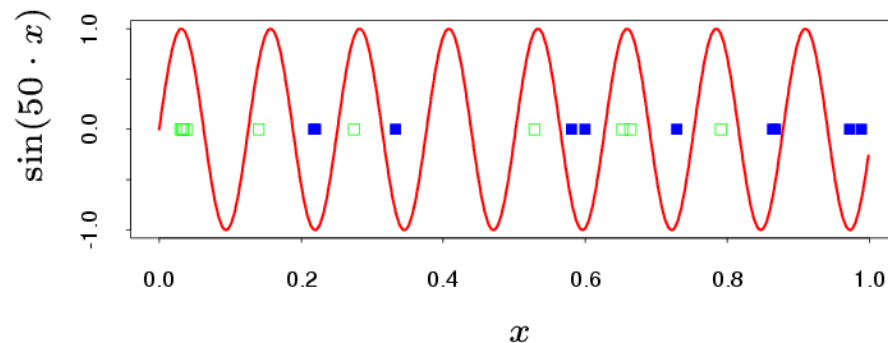
- La VC dimension de l'ensemble de fonctions

$$f(x, w) = \text{sign}(\sin(w \cdot x)),$$

$$c < x < 1, c > 0,$$

avec un paramètre libre w est infinie.

Hastie et al. 2001



Deux cas importants:

a) régression ridge

- La VC dimension de l'ensemble des indicatrices linéaires

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

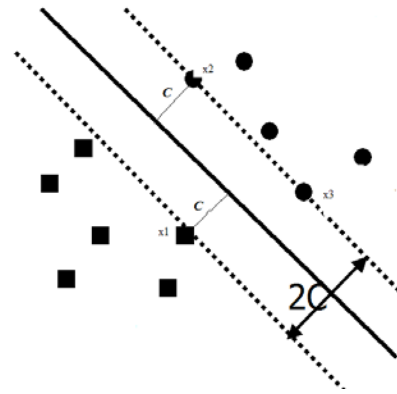
$$\|X\| \leq R$$

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 $p+1$.

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

b) L'hyperplan de marge maximale



■ Même résultat:

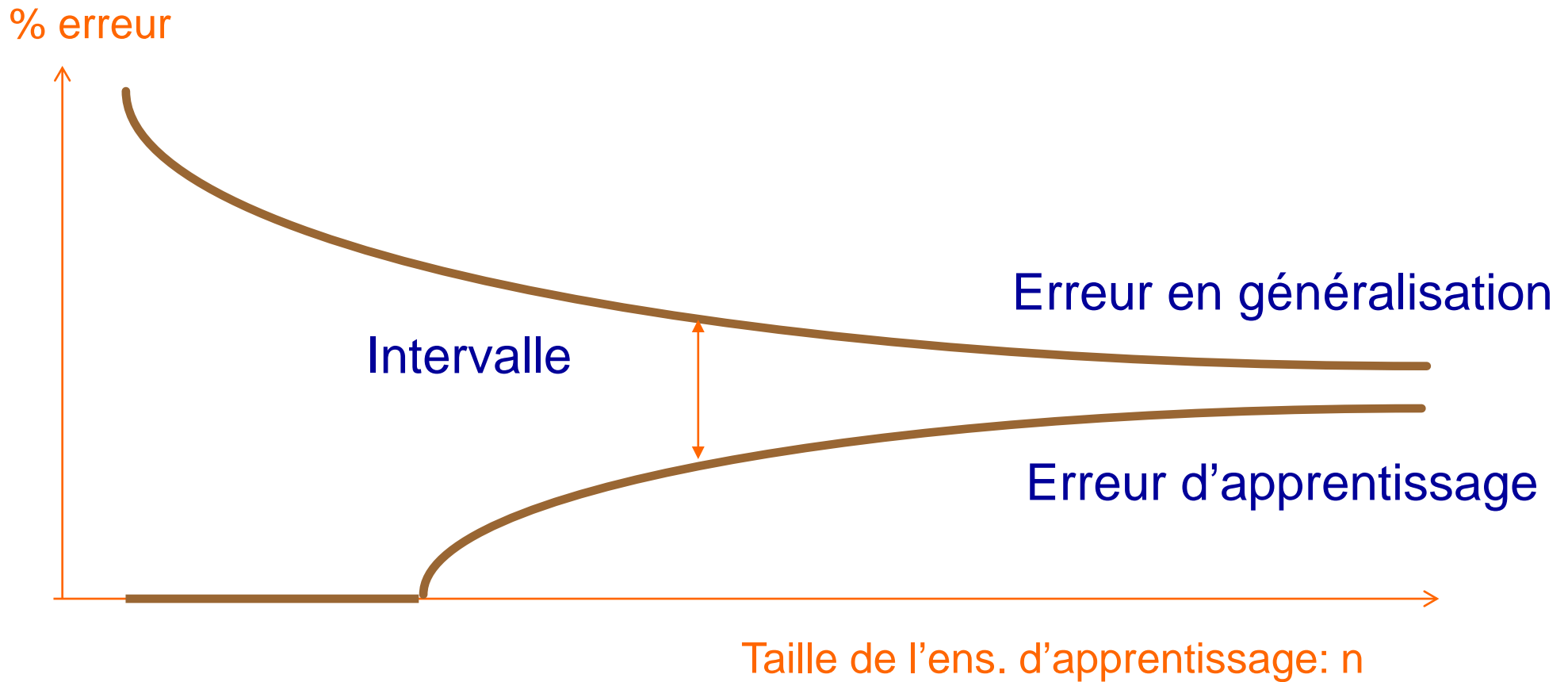
$$h \leq \min \left[\text{ent} \left(\frac{R^2}{C^2} \right); p \right] + 1$$

Théorème de Vapnik :



- Q : Quelles sont les conditions nécessaires et suffisantes pour assurer la consistance ?
- R : Le processus d'apprentissage est consistant si et seulement si la famille de modèles a une VC dimension finie h
- La VC dimension finie ne garantit pas seulement la généralisation, mais c'est LA SEULE MANIERE qui permet à la généralisation de se produire.

Vitesse de convergence



Vitesse de convergence (2)

- Q : Quelle est la différence entre les erreurs d'apprentissage et de test pour une taille donnée de l'ensemble d'apprentissage ?
- R : La différence entre les erreurs d'apprentissage et de test dépend du rapport entre la VC dimension, h , et la taille de l'ensemble d'apprentissage, n .

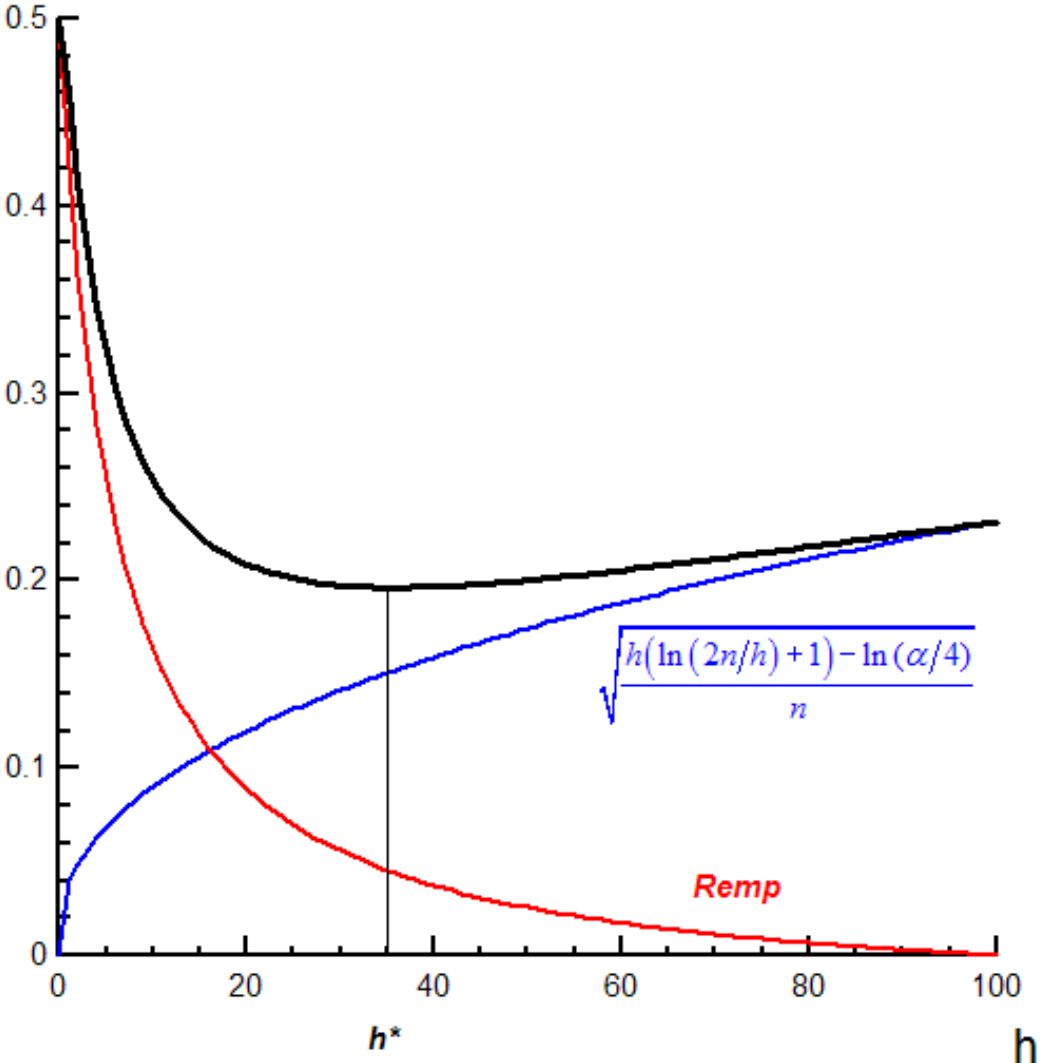
Inégalité de Vapnik

- 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

n fixé



De Guillaume d'Ockham à Vapnik



wikipedia

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, excommunié, se réfugie à Munich, à la cour de Louis de Bavière, lui-même excommunié. Meurt de l'épidémie de peste noire.

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. »

Le rasoir d'Ockham ou principe de parcimonie

Principe de raisonnement attribué à Ockham : « Les multiples ne doivent pas être utilisés sans nécessité » (*pluralitas non est ponenda sine necessitate*).

Rasoir d'Ockham et science moderne

Le rasoir d'Ockham n'est malheureusement pas un outil très incisif, car il ne donne pas de principe opératoire clair pour distinguer entre les hypothèses en fonction de leur complexité : ce n'est que dans le cas où deux hypothèses ont la même vraisemblance qu'on favorisera l'hypothèse la plus simple (ou parcimonieuse). Il s'agit en fait d'une application directe du théorème de Bayes où l'hypothèse la plus simple a reçu la probabilité a priori la plus forte. Des avatars modernes du rasoir sont les mesures d'information du type AIC, BIC où des mesures de pénalité de la complexité sont introduites dans la log-vraisemblance.

De Guillaume d'Ockham à Vapnik

Si deux familles de modèles expliquent les données avec une qualité égale, alors la famille **qui a la plus faible VC dimension** doit être préférée.

1re découverte: La VC (Vapnik-Chervonenkis) dimension mesure la complexité d'une famille de modèles.

De Guillaume d'Ockham à Vapnik



Si deux modèles expliquent les données avec une qualité égale, alors celui qui provient d'une famille à plus faible VC dimension a une meilleure performance en généralisation.

2ème découverte: La VC dimension peut être reliée à des résultats de généralisation (résultats sur de nouvelles données).

De Guillaume d'Ockham à Vapnik




Pour construire le meilleur modèle à partir de données, il faut tenter d'optimiser à la fois sa performance sur l'ensemble d'apprentissage, **et** sa performance de généralisation tirée de la VC dimension : pour ce faire, il faut parcourir une suite de familles d'applications pour y construire ce modèle

3ème découverte: Au lieu d'observer des différences entre des modèles, mieux vaut les contrôler..

Contrôle de la Capacité de Généralisation

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

- Risque = Risque d'Apprentissage + Intervalle de Confiance
 - Minimiser la seule erreur d'apprentissage ne donnera pas une espérance d'erreur faible (une bonne généralisation)
-  minimiser la somme de l'erreur d'apprentissage et de l'intervalle de confiance.

Principe de minimisation structurée du risque (SRM) (1)

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

- 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.

Principe de minimisation structurée du risque (SRM)(2)

- Considérons une structure $S_1 \subset S_2 \subset \dots \subset S_L$ sur l'ensemble des fonctions vérifiant la propriété

$$h_1 < h_2 < \dots < h_L$$

- Pour chaque élément S_i de la structure, l'inégalité est valide

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

SRM : Trouver i tel que la somme devienne minimale,

Application du principe SRM

- La structure S_i (familles de modèles) peut être contrôlée par :
 - Architecture de réseaux de neurones
 - Degré d'un polynôme
 - Méthodologie d'apprentissage
 - Contrôle des poids dans un réseau de neurones, ...

Avec/sans l'approche SRM de Vapnik

■ Sans le SRM:

- Hypothèses sur la distribution statistique (inconnue) des données
- Un grand nombre de dimensions signifie un modèle à grand nombre de paramètres, ce qui pose des problèmes de généralisation
- Modéliser revient à chercher le meilleur ajustement

■ Avec le SRM:

- On étudie la famille de modèles, contrôlant sa VC dimension h
- Le nombre de paramètres peut être très grand, car on contrôle par définition la généralisation
- Modéliser c'est rechercher le meilleur compromis entre ajustement et robustesse

Borne supérieure trop grande,

mais:



Théorème (Devroye, Vapnik) :

Pour toute distribution le SRM fournit la meilleure solution possible avec probabilité 1

(universally strongly consistent)

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

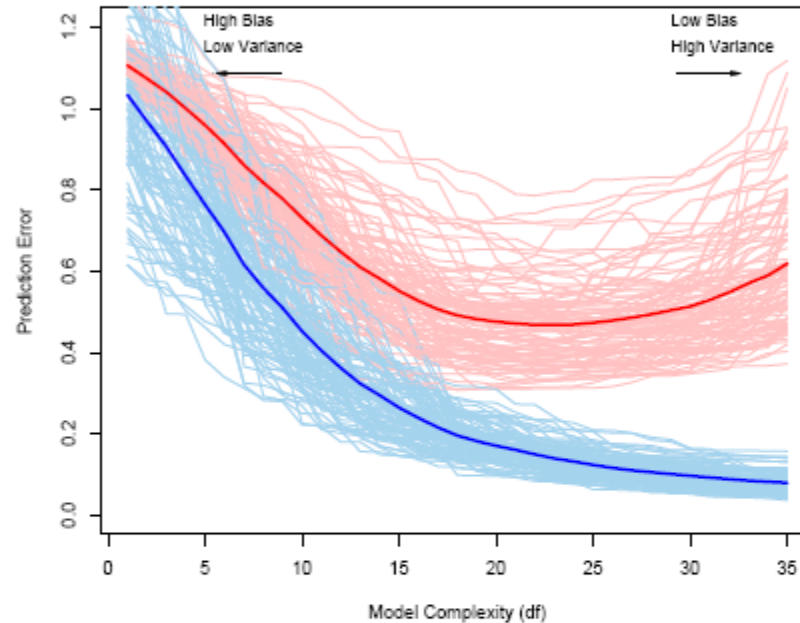


FIGURE 7.1. Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error $\overline{\text{err}}$, while the light red curves show the conditional test error $\text{Err}_{\mathcal{T}}$ for 100 training sets of size 50 each, as the model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

Principes d'induction




- Ne pas chercher à résoudre un problème plus général que nécessaire
 - Ne pas estimer une densité si on veut estimer une fonction
 - Ne pas estimer une fonction si on veut seulement estimer une valeur en un point

8 ème partie : arbres de décision

Les méthodes de segmentation

- Développées autour de 1960 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** ou **fouille de données** 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. Le terme de **partitionnement récursif** est parfois utilisé

- Les méthodes de segmentation sont des méthodes à but explicatif qui résolvent les 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 (on présentera plus tard des critères permettant de diviser un segment).
- 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

- 
- La segmentation est donc en concurrence avec les méthodes explicatives paramétriques (régressions linéaires, logistique, analyse discriminante ...).
 - A la différence de ces méthodes, les variables sont présentées et utilisées séquentiellement et non simultanément.
 - Les méthodes de segmentation sont des techniques non paramétriques, très peu contraintes par la nature des données.
 - Les sorties se présentent sous forme d 'arbres de décision qui fournissent des règles d 'affectation lisibles et facilement interprétables.

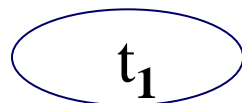
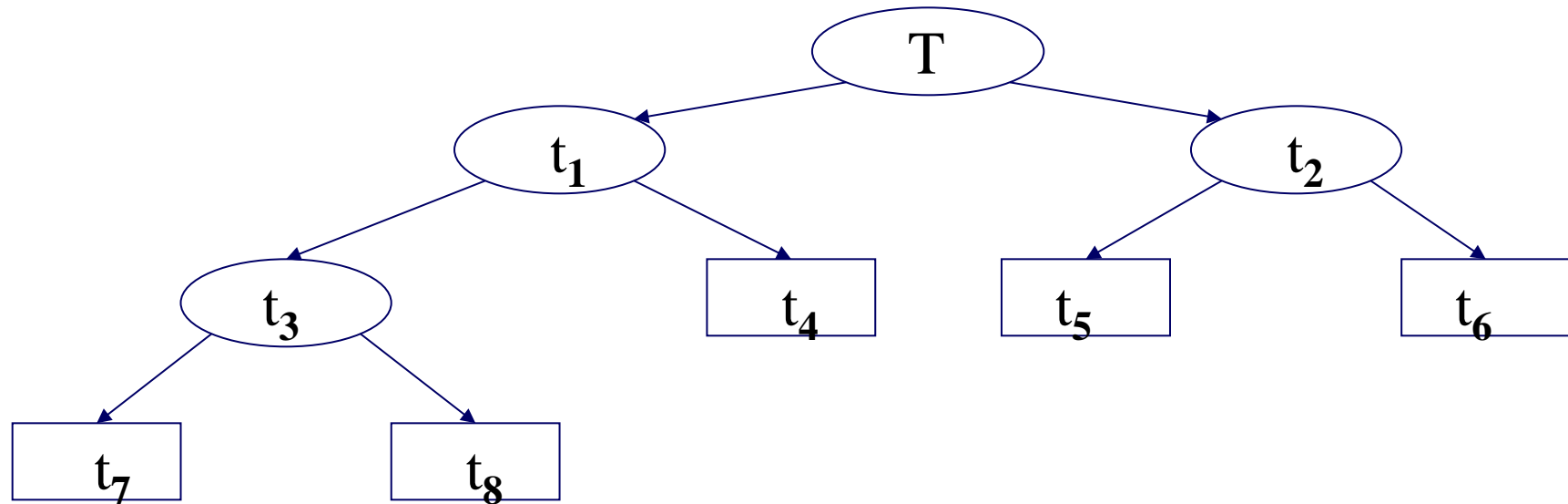
Un logiciel gratuit:



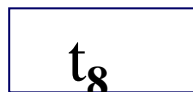
- SIPINA <http://eric.univ-lyon2.fr>

Arbre de décision

- On représente ainsi les divisions successives de l'échantillon (on parcourt l'arbre en le descendant).
- A chaque étape, on divise un segment en plusieurs segments plus purs ou de variances plus faibles (i.e. plus homogènes).



: Segments **intermédiaires**



: Segments **terminaux**

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 .

Divisions d'un nœud (arbres binaires)

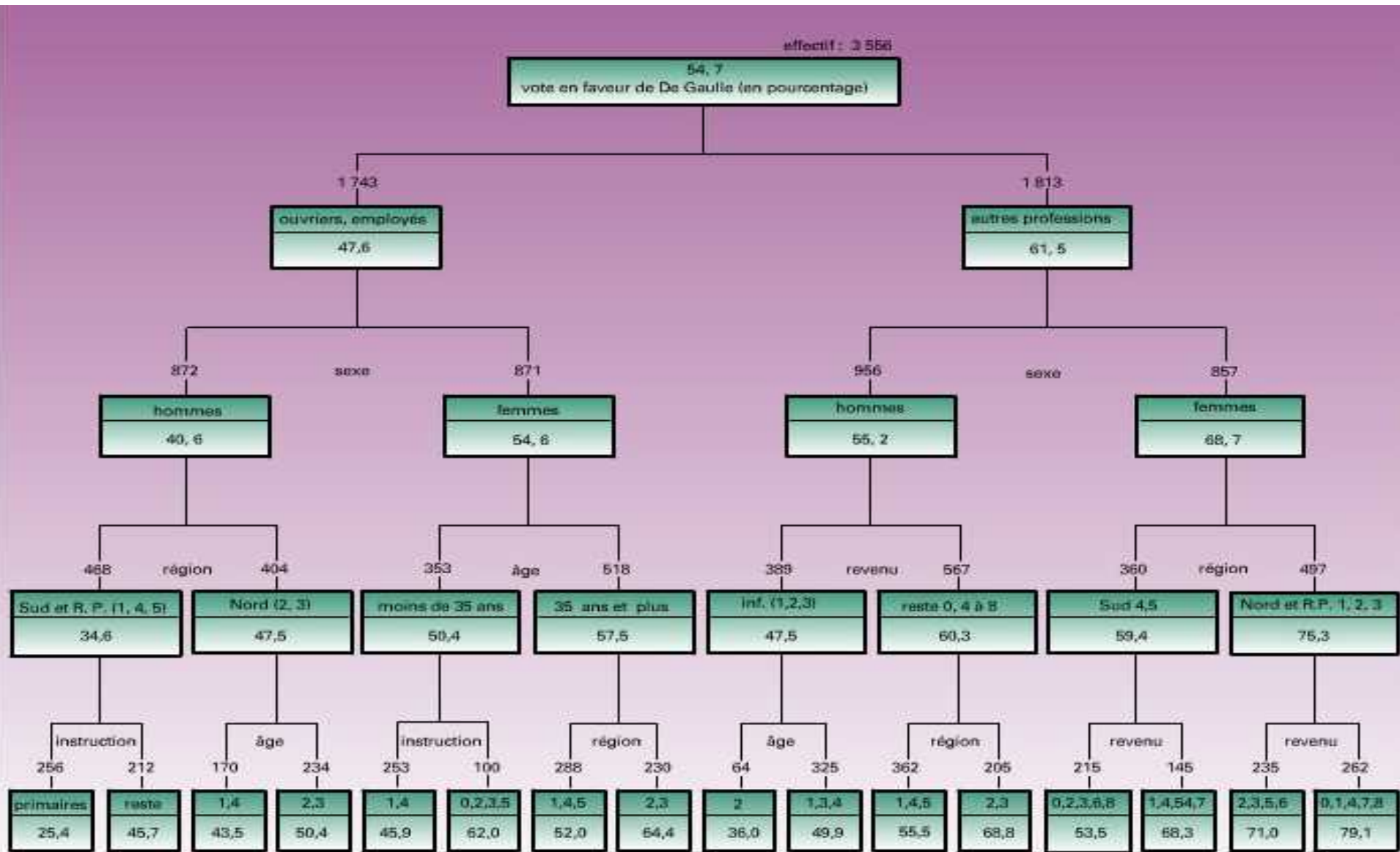
- Les divisions possibles dépendent de la nature statistique de la variable :
 - variable binaire $B(0,1)$: une division possible
 - variable nominale N (k modalités) : $2^{k-1} - 1$ divisions possibles
 - variable ordinale O (k modalités) : $k-1$ divisions possibles
 - variable quantitative Q (q valeurs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

■ Exemple : (3 variables, divisions binaires)

binaire $(b1,b2)$: $(b1)$ $(b2)$

Ordinale $(o1,o2,o3,o4)$:
 $(o1)$ $(o2,o3,o4)$
 $(o1,o2)$ $(o3,o4)$
 $(o1,o2,o3)$ $(o4)$

nominale $(n1)$ $(n2,n3)$
 $(n2)$ $(n1,n3)$
 $(n3)$ $(n1,n2)$



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
- Il est à noter que CART utilise le même principe pour analyser une variable nominale (problème de discrimination) ou une variable continue (régression).

Discrimination : critère de division

- Impureté d'un nœud :

$$i(t) = \sum_r^k \sum_s^k P(r/t)P(s/t)$$

- Avec $r \neq s$ et où $P(r/t)$ et $P(s/t)$ sont les proportions d'individus dans les classes c_r et c_s dans le segment t ($i(t)$ est l'indice de diversité de Gini)
- *Segment pur : ne contient que des individus d'une classe, $i(t) = 0$*
- *Segment mélangé : $i(t) \neq 0$ et $i(t)$ fonction croissante du mélange*

Réduction d'impureté

- Réduction de l'impureté par la division s :

$$\Delta i(s, t) = i(t) - p_g i(t_g) - p_d i(t_d)$$

- Où les p_g sont les proportions d'individus du nœud t respectivement dans les segments descendants t_g et t_d (la fonction $i(t)$ étant concave, l'impureté moyenne ne peut que décroître par division d'un nœud)
- Réduction maximale pour chaque variable :

$$\Delta i(s^*, t) = \max \{ \Delta i(s, t) \}$$

- Réduction maximale pour l'ensemble des p variables :

$$\Delta^* = \max_{j=1 \dots p} \{ \Delta i(s^*, t) \}$$

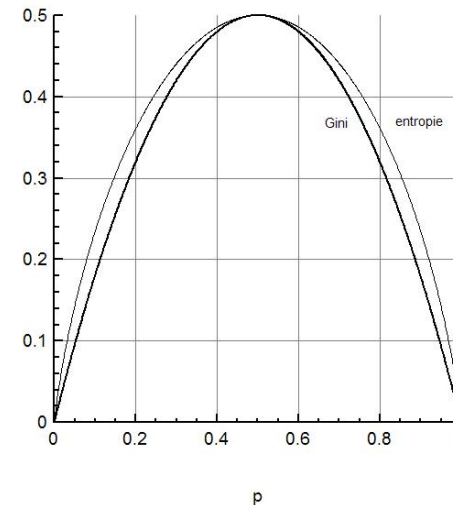
Entropie et indice de Gini

■ entropie $-\sum_{i=1}^k p_i \ln(p_i)$

■ indice de diversité de Gini

$$\sum_{i=1}^k p_i p_j$$

■ Pour deux classes,
indices très proches:



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.

- Taux d'erreur de classement en apprentissage (T.E.A) associé à un segment terminal de l'arbre A :

$$R(s/t) = \sum_{r=1}^k p(r/t)$$

- Avec $r=s$ et où $P(r/t) = n_r(t)/n_t$ est la proportion d'individus du segment t affectés à la classe c_s et qui appartiennent à la classe c_r
- T.E.A associé à l'arbre :

$$TEA(A) = \sum_{t \in A} \frac{n_r(t)}{n} R(s/t) = \sum_{t \in A} \sum_{k=1}^k \frac{n_r(t)}{n}$$

- 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 :

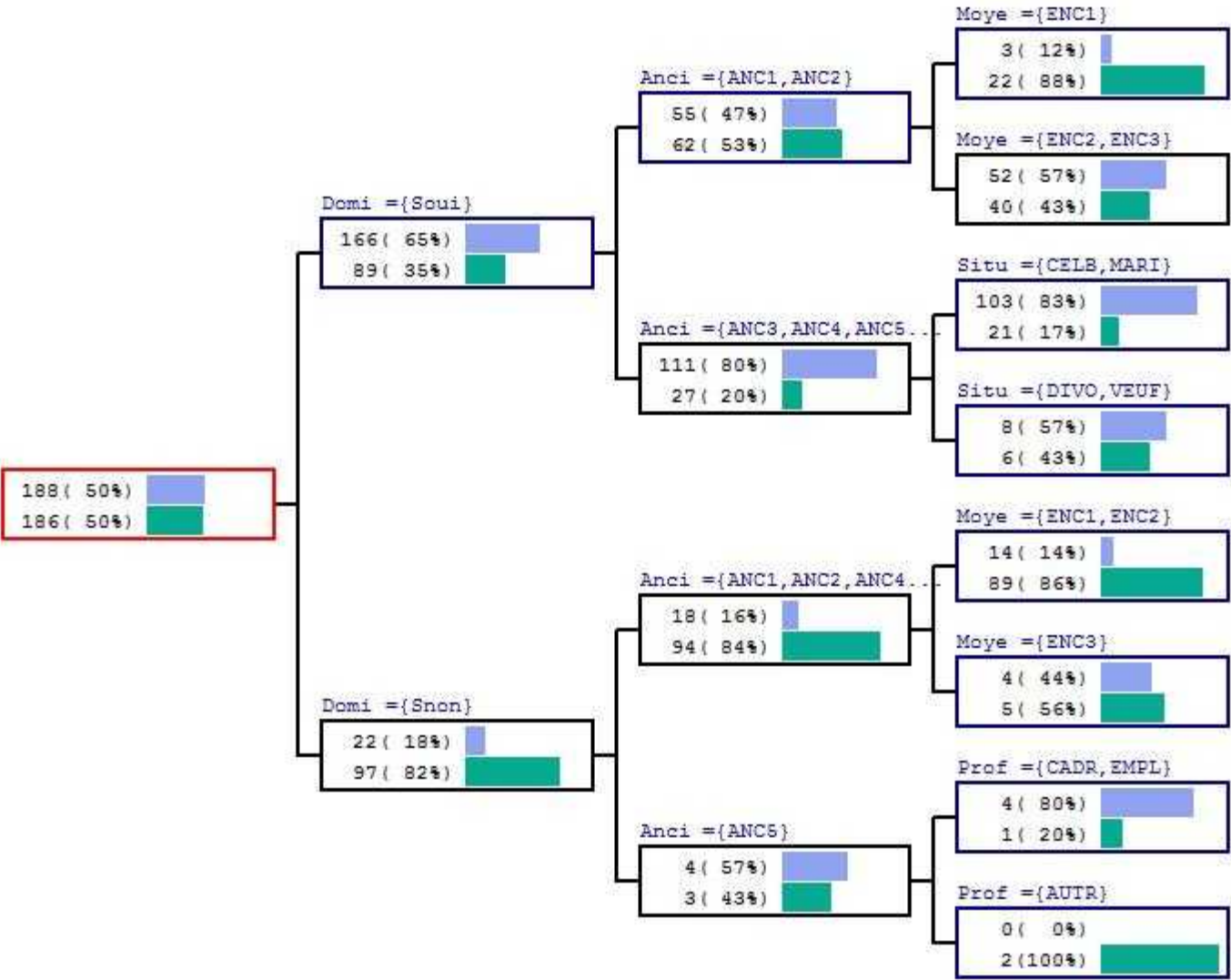
- Choix de A^* tel que l'erreur de classement en test (ETC) vérifie :

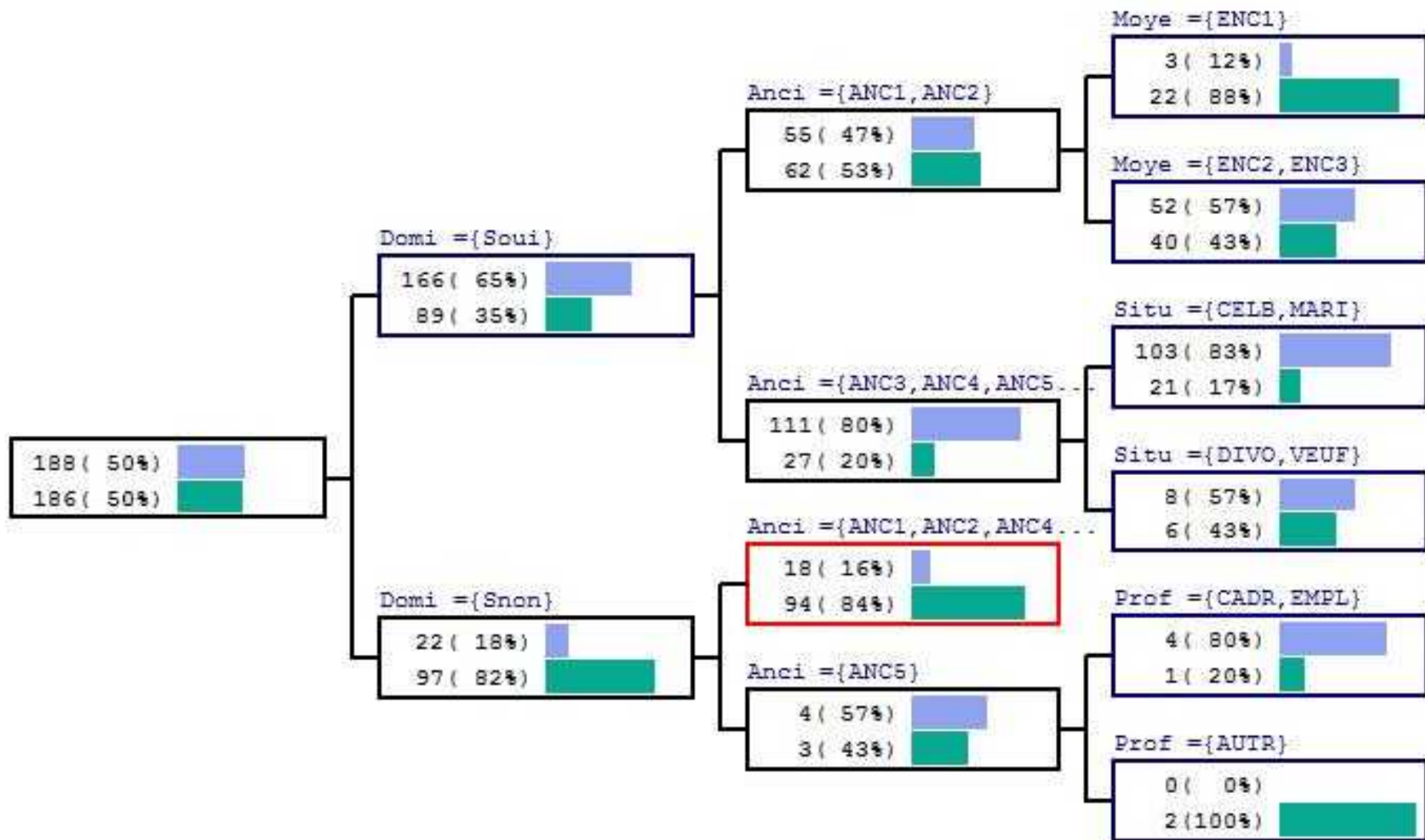
$$ETC(A^*) = \min_{1 \leq h \leq max} \{ETC(A_h)\}$$

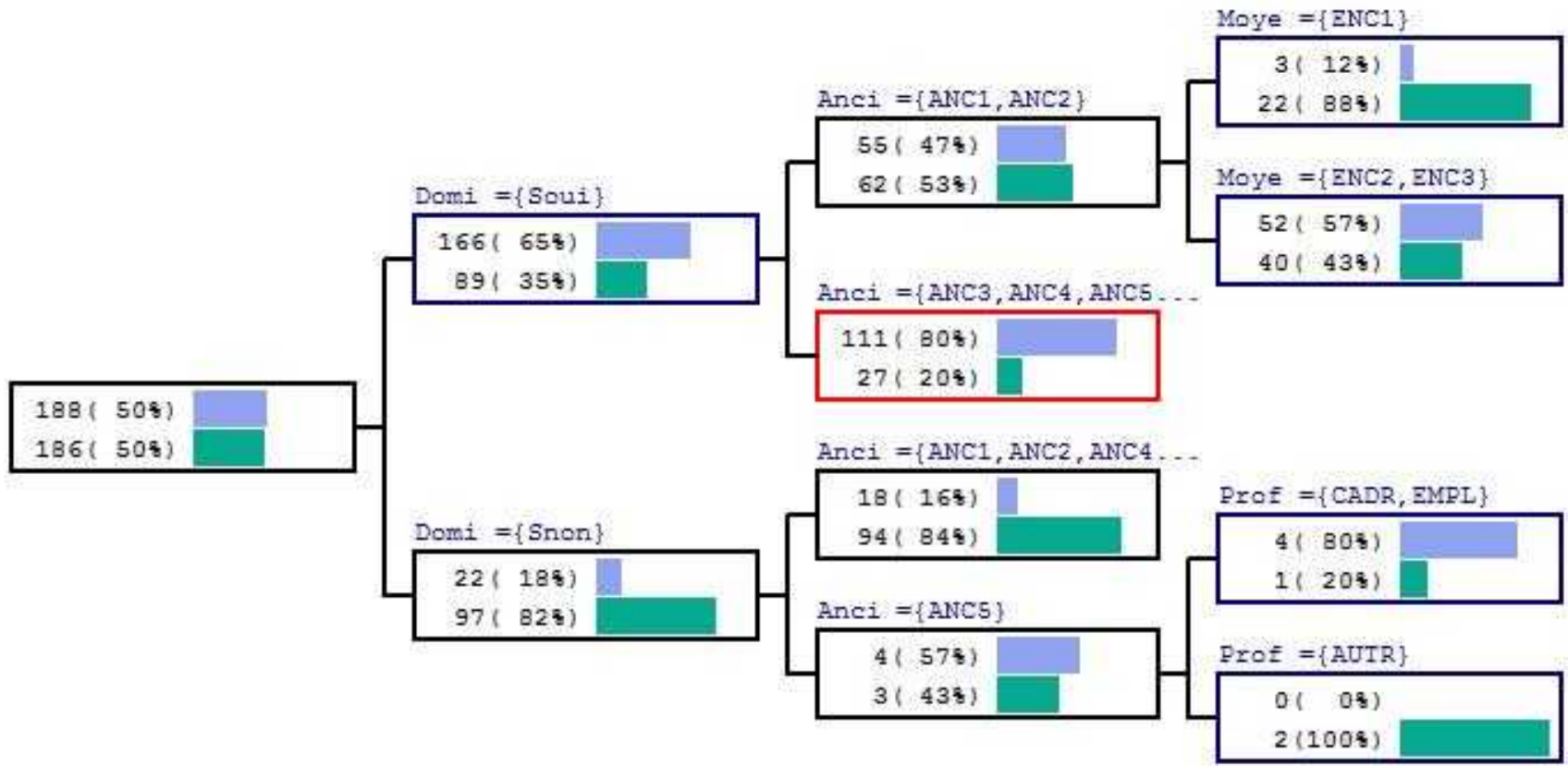
Divisions équiréductrices et équidivisantes

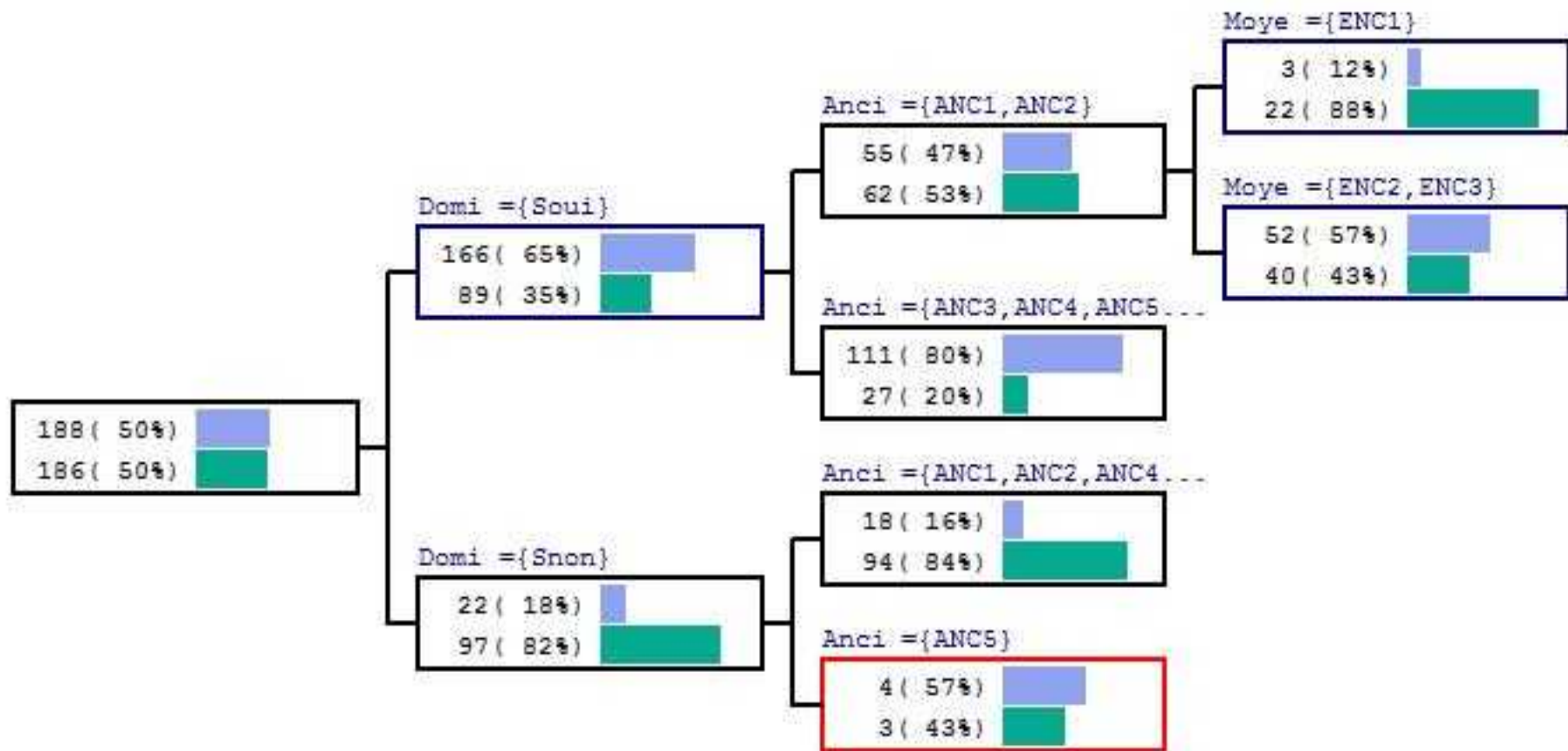
- En plus de la meilleure division d^* (celle assurant la plus grande réduction de l'impureté ou de la variance résiduelle) , on définit :
- Les divisions équiréductrices : celles qui assurent après d^* les plus fortes réduction de l'impureté ou des variances résiduelles ; elles permettent d'autres choix de variables explicatives.
- Les divisions équidivisantes : fournissent les répartitions les plus proches de la meilleure division d^* ; elles permettent de gérer les données manquantes.

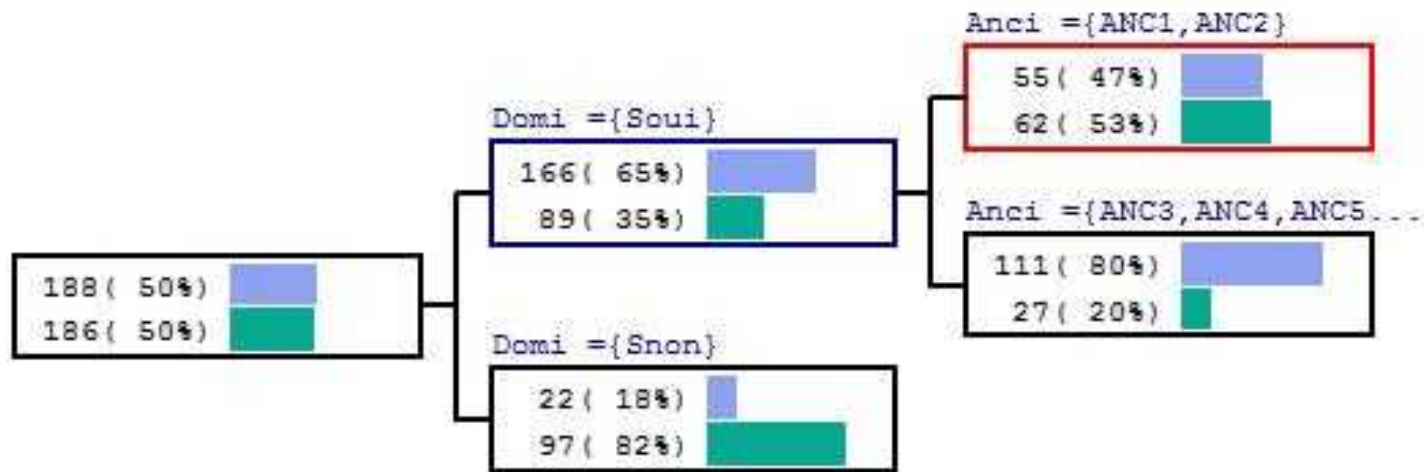
Exemple: bons et mauvais clients d'une banque (SPAD)











Arbres de régression

- Si y est numérique, mesure d'homogénéité = variance de la classe
- Division en deux sous-groupes: minimiser la variance intra-groupe ou maximiser la variance inter-groupe.

$$V_{\text{inter}} = \frac{1}{n} \left(n_1 (\bar{y}_1 - \bar{y})^2 + n_2 (\bar{y}_2 - \bar{y})^2 \right)$$

$$V_{\text{inter}} = \frac{n_1 n_2}{n^2} (\bar{y}_1 - \bar{y}_2)^2$$

- La coupure optimale pour une variable qualitative nominale à m modalités doit respecter l'ordre induit par la moyenne de y .
- On réordonne donc les catégories de x selon \bar{y}_i et il n'y a plus que $m-1$ dichotomies à examiner au lieu de $2^{m-1} - 1$.

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, et ce fait peut être particulièrement gênant si les variables équiréductrices sont 'proches' de la variable qui a servi à faire la division).

Nouvelles tendances :

- « Bagging » ou bootstrap averaging

- B arbres à partir de B répliques: « forêt »

- Procédure de vote

- « Boosting » AdaBoost

- Combinaison de classifieurs faibles

$$\sum_m \alpha_m G_m(x)$$

- Poids croissant avec la précision

- Classifieur G_m : surpondération des observations mal classées de G_{m-1}