Traitement de réponses qualitatives

Gilbert Saporta
Conservatoire National des Arts et Métiers

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

Plan

1. Introduction
2. Mesures d'efficacité
3. Les modèles à réponse binaire
   - 3.1 Analyse discriminante
   - 3.2 Régression logistique
   - 3.3 SVM
   - 3.4 Arbres de décision
4. Réponse à plus de deux catégories
5. Plans d'expériences et modèles
6. Combinaison de modèles

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

Bibliographie

- 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
- Tufféry: « Data Mining et statistique décisionnelle »,4ème édition, Technip, 2012

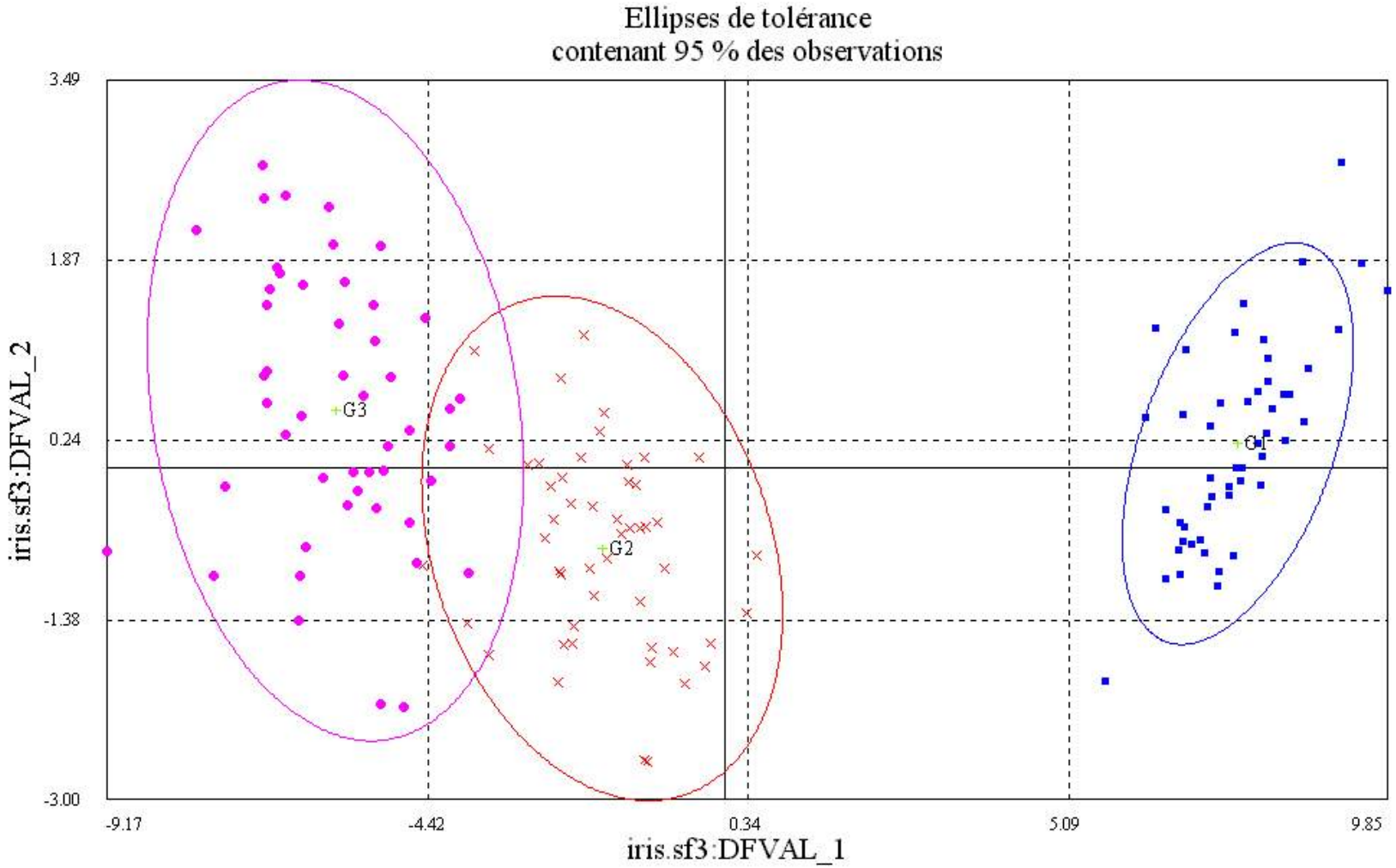
1. Introduction

- Réponse Y qualitative, prédicteurs numériques X_1, .., X_p
- Géomètrie:
  - Observations multidimensionnelles réparties en k groupes définis a priori.

Iris de Fisher (1936):

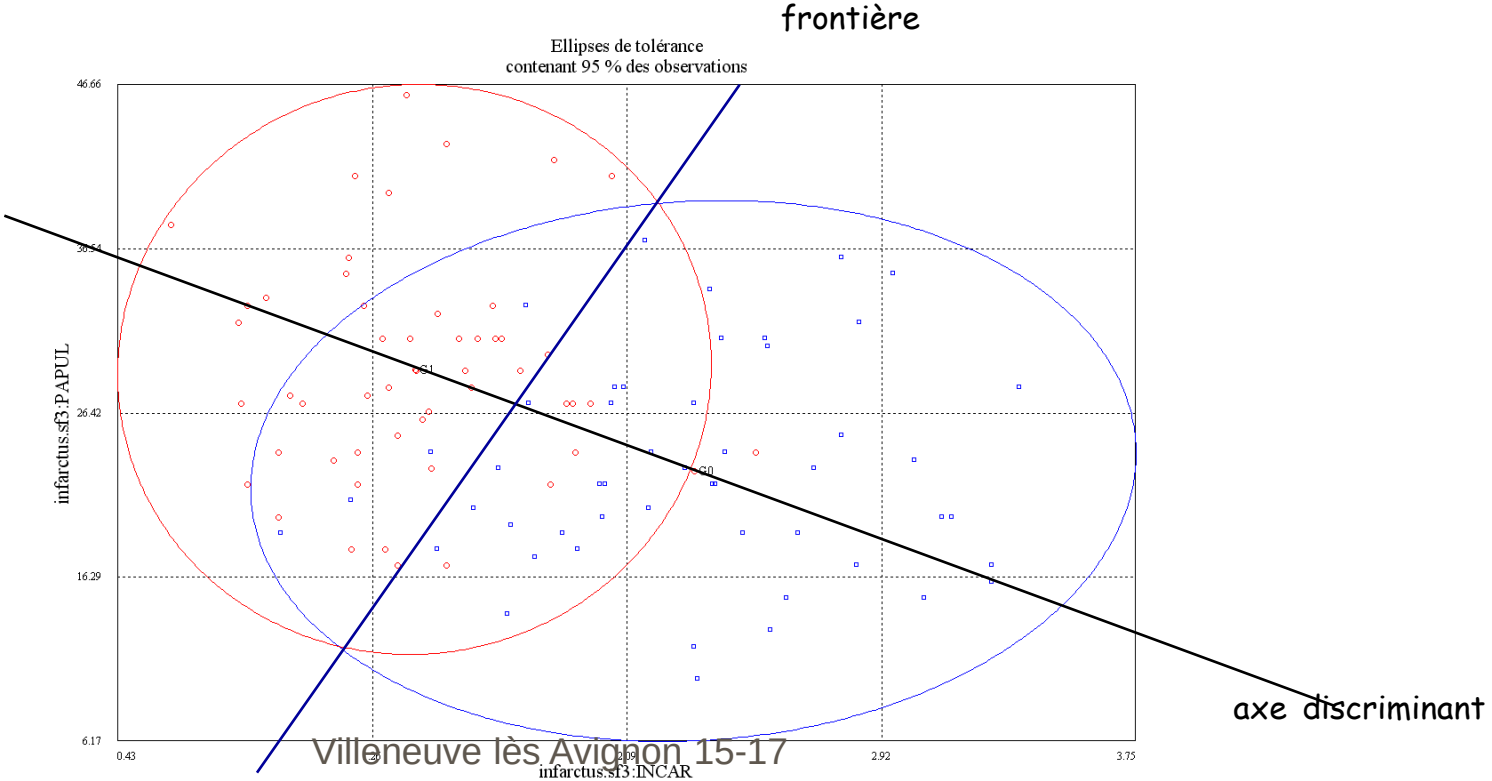
3 espèces : 4 variables (longueur et largeur des pétales et sépales)

		
Iris setosa	Iris versicolor	Iris virginica


Ellipses de tolérance
contenant 95 % des observations
3.49
1.87
0.24
-1.38
-3.00
iris.sf3:DFVAL_2
G3
G2
G1
-9.17
-4.42
0.34
5.09
9.85
iris.sf3:DFVAL_1

Classifieur et frontière

- Classifieur ou fonction discriminante f(x) souvent comparée à zéro
- Classifieur linéaire: dualité axe-frontière plane



Règle des k 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

Classifieur et probabilité a posteriori

- Toute méthode calculant une probabilité est équivalente à un classifieur (ou score) à valeurs comprises entre 0 et 1.
- Seuil à 0.5 ?
  - Très discutable en cas de déséquilibre
  - Rôle des probabilités a priori

2. Mesures d'efficacité

- Taux de bien classés
- Courbe ROC et AUC

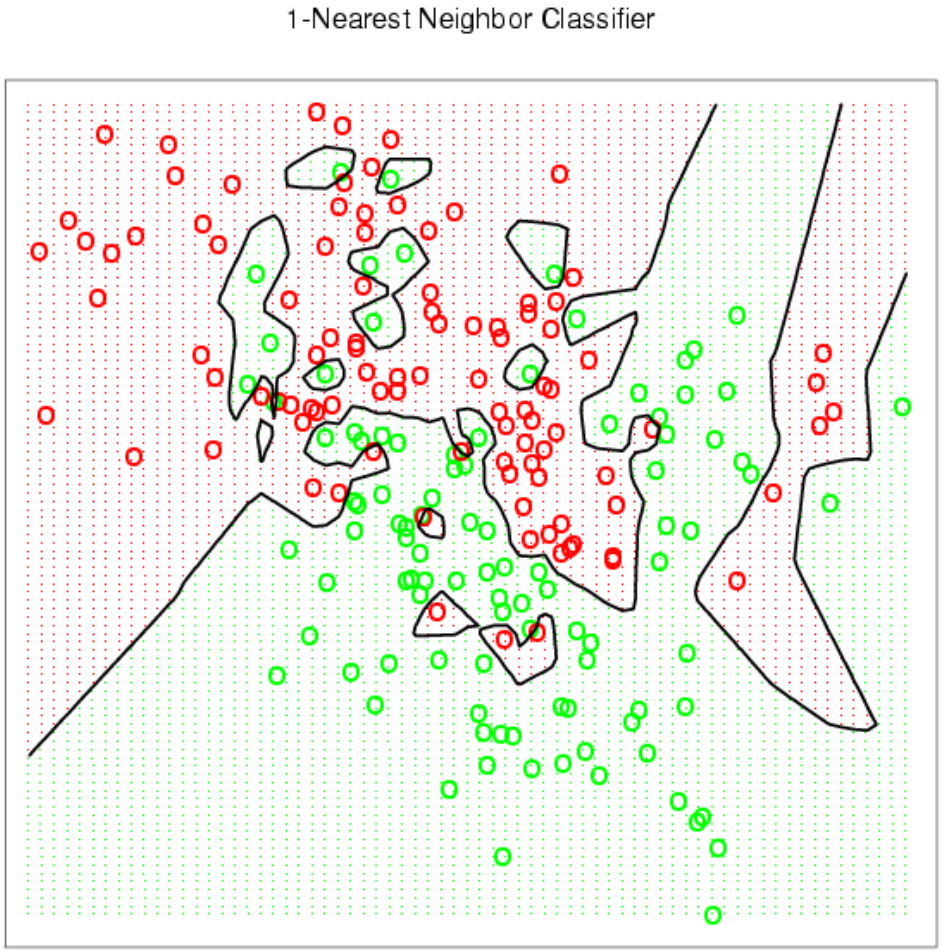
- **Tableau de classement :**
  - On classe des observations dont le groupe est connu :

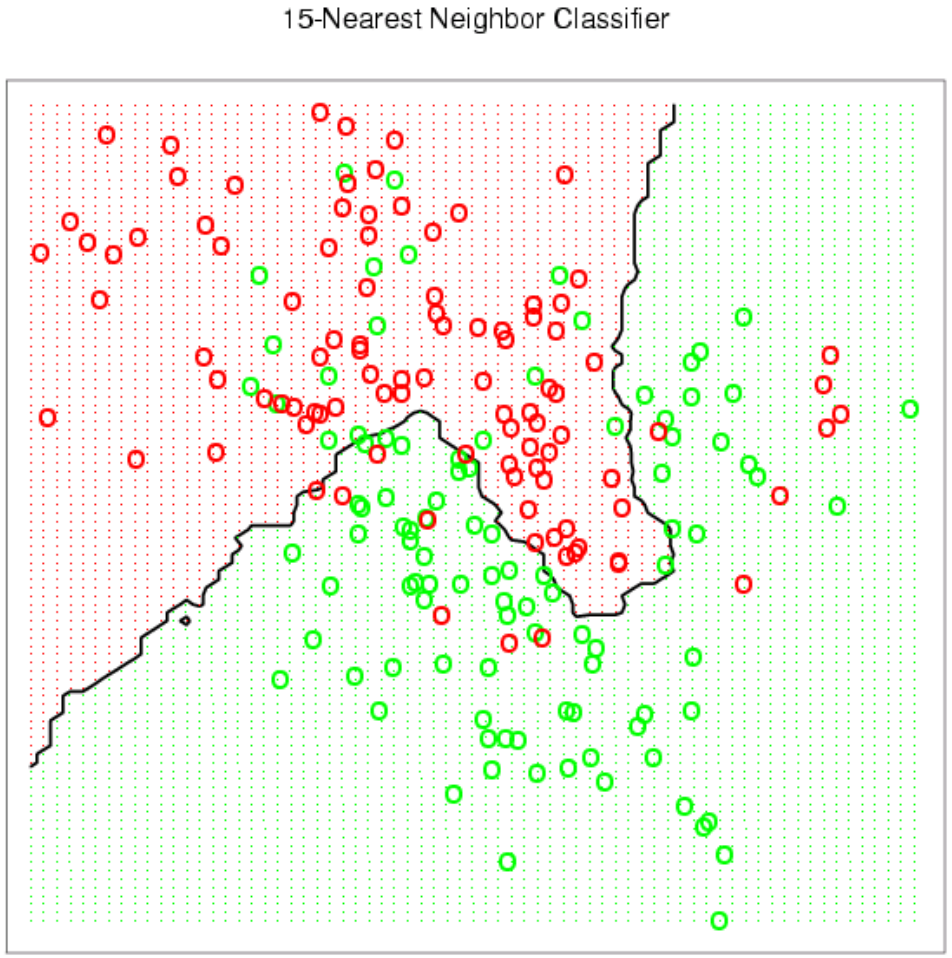
		groupe 1	prédit 2
groupe réel	1	n_{11}	n_{12}
	2	n_{21}	n_{22}

  - Pourcentage de bien classés : $\frac{n_{11}+n_{22}}{n}$
  - Taux d'erreur de classement : $\frac{n_{12}+n_{21}}{n}$

Exemple

Méthode des plus proches voisins (Hastie and al)





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» $\Rightarrow$ 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)*

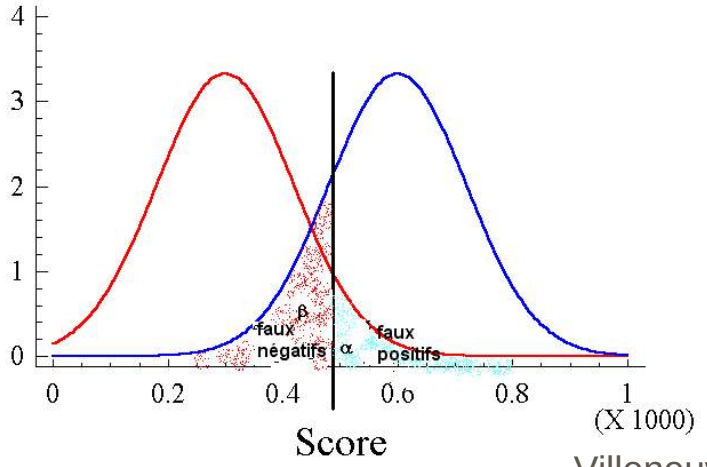
- Nécessité de fixer un seuil pour le classifieur ou pour la probabilité a posteriori
- Retour sur les risques d'erreur:

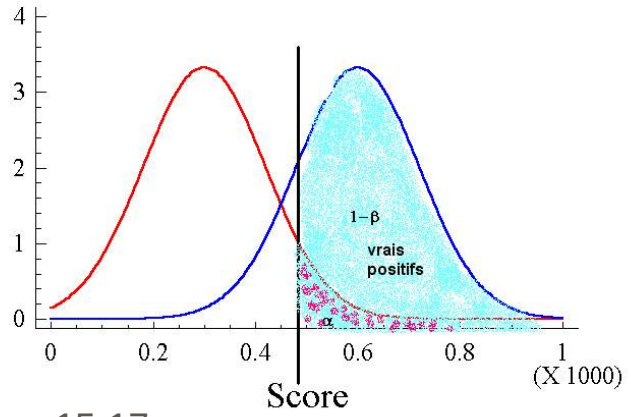
  Groupe d'intérêt G1

  - Faux positifs: risque α
  - Faux négatifs: risque β
  - Sensibilité : 1-β
  - Spécificité: 1-α
- Coûts d'erreurs différents: une autre histoire...

Variation selon un seuil s

- 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



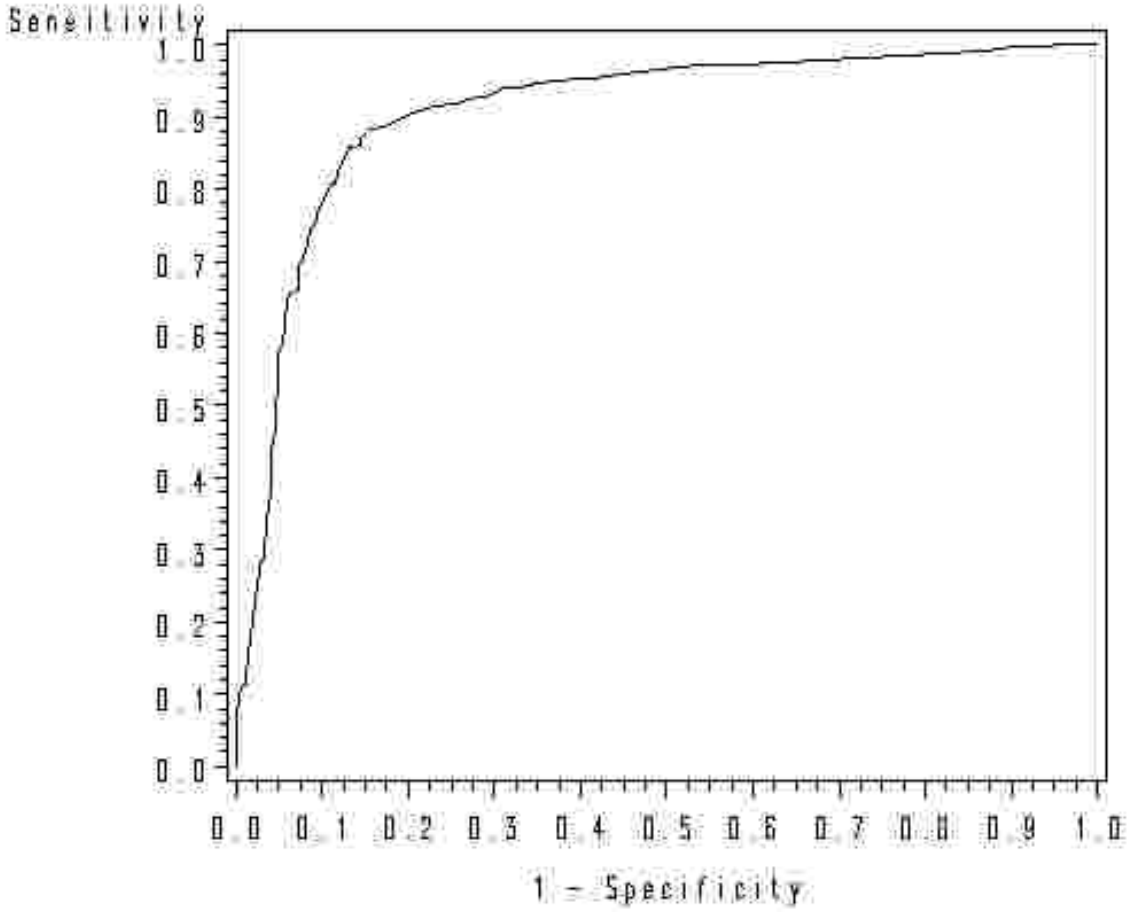


- Facile d’avoir une bonne sensibilité:
  - On classe tout le monde dans le groupe d’intérêt...
  - Dans les cas déséquilibrés on a même un bon taux global de bien classés;
- Une bonne méthode doit avoir une sensibilité et une spécificité élevées

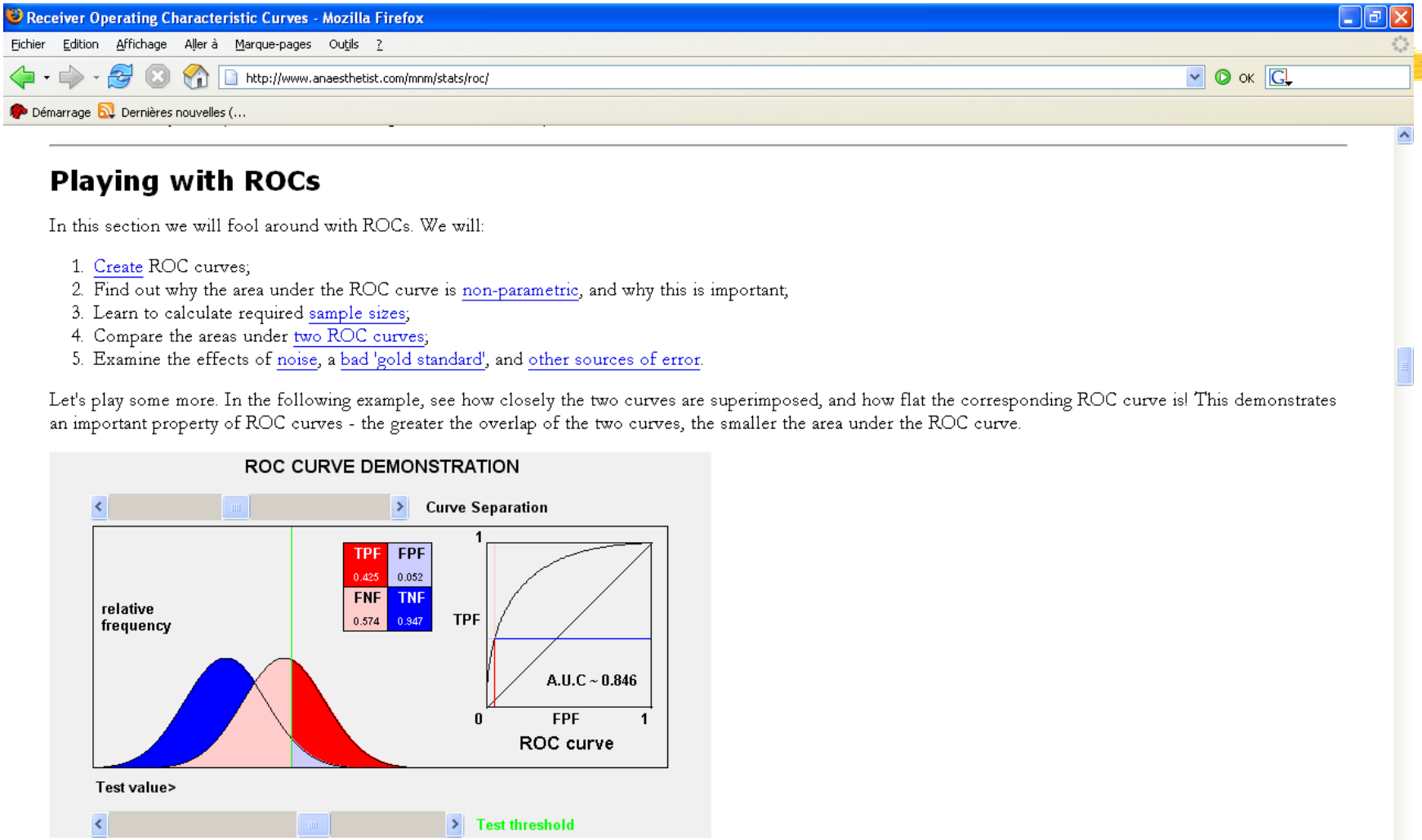
Courbe ROC:

- Evolution de 1-β 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

Courbe ROC


Sensitivity
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
1 - Specificity

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



Surface sous la courbe ROC

- Surface théorique sous la courbe ROC: $P(X_1>X_2)$ si on tire au hasard et indépendemment 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 $C = \frac{n_c}{n_1 n_2}$

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 $n_c = U$

  $U+W = n_1 n_2 + 0.5 n_1 (n_1 + 1)$
- $AUC = U/n_1 n_2$

Comparer des courbes ROC correspondant à plusieurs méthodes

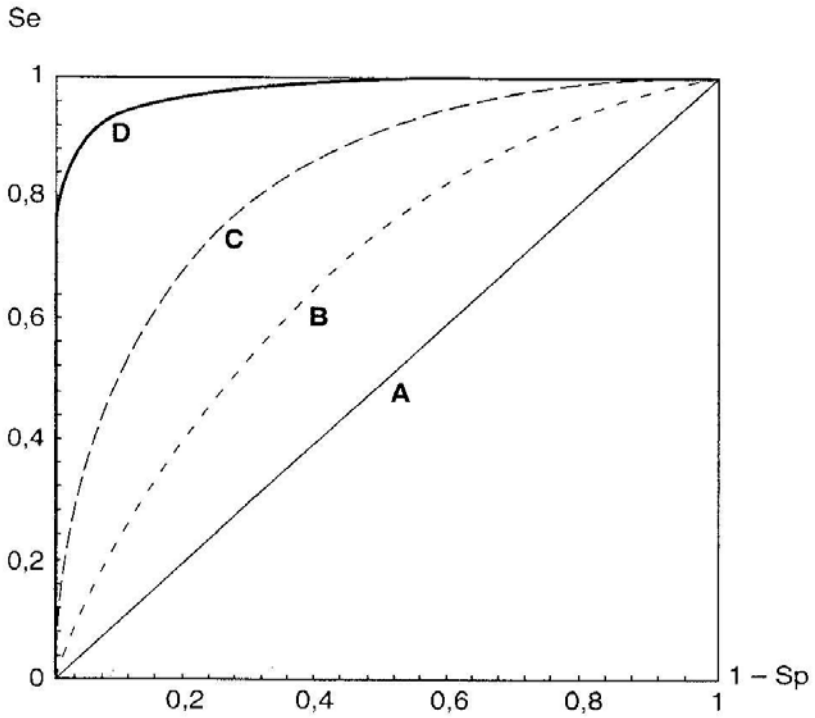
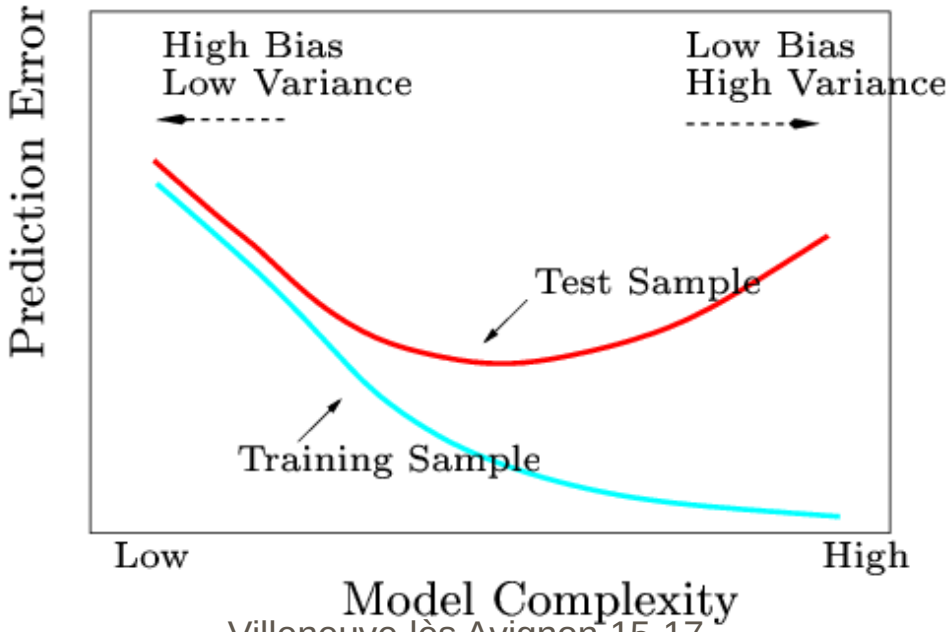


Figure 3-7 **Différentes courbes ROC.** A : test sans intérêt diagnostique. B : mauvais test diagnostique. C : test meilleur que B. D : bon test diagnostique.

- Comparer les courbes ou comparer les AUC?
  - petit problème si les courbes se croisent…
  - Sur échantillon test pour éviter de favoriser la méthode qui a la plus grande complexité



Les 3 échantillons:

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

- Rééchantillonner: validation croisée, bootstrap

Modèle final: avec toutes les données disponibles

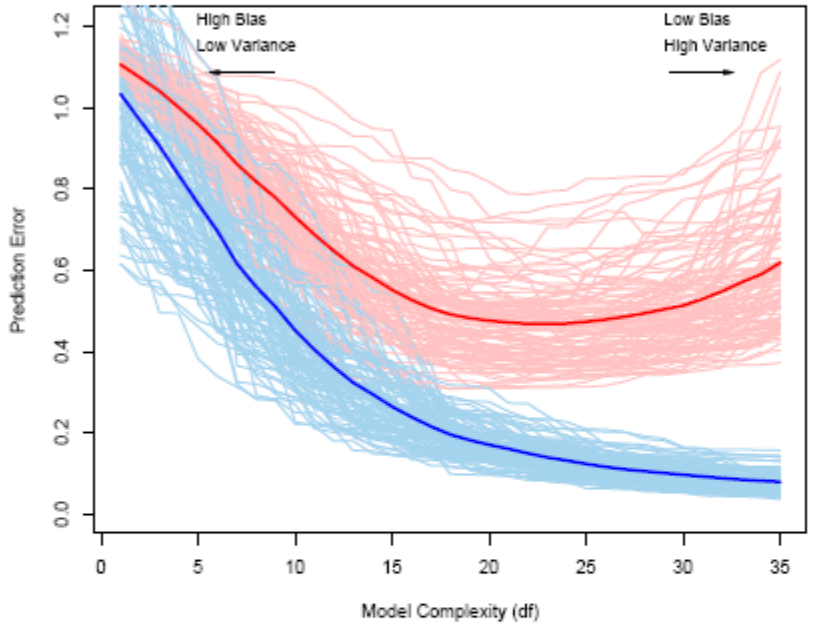


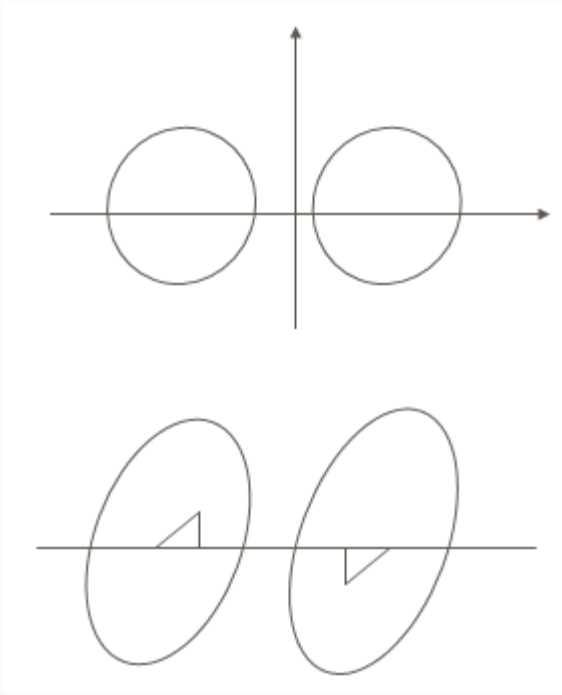
FIGURE 7.1. *Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error* $\overline{\text{err}}$*, while the light red curves show the conditional test error* $\text{Err}_{\mathcal{T}}$ *for* 100 *training sets of size* 50 *each,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

3. Panorama de la classification binaire

3.1 Analyse linéaire discriminante de Fisher

- Géomètrie
  - On classe selon la distance au centre de gravité des groupes
  - Distance de Mahalanobis

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



Isotropie:mètrique I

Nécessité de tenir compte des corrélations: mètrique W^{-1} , W est la moyenne des matrices de covariances

- On classe dans G_1 si:

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

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

- Fonction de Fisher >c
- Score de Fisher: $(g_1 - g_2)'W^{-1}e - \tfrac{1}{2}(g_1'W^{-1}g_1 - g_2'W^{-1}g_2)$

Historique

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

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

Fisher (1936)

Trouver $u_1, u_2, ..., u_p$ 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 : $\alpha=1+\dfrac{n_1 n_2}{n(n-2)}D_p^2$

Une régression « incorrecte »

- **y** à 2 valeurs (-1;+1) ou (0;1) ou (a;b)
- a=n/n_1 b=-n/n_2

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

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

- D_p distance de Mahalanobis entre groupes
- Incompréhensions et controverses!

Modèle linéaire usuel non valide : $y / \mathbf{X} \quad N(\mathbf{X}\boldsymbol{\beta}; \sigma^2 \mathbf{I})$

en discriminante c'est l'inverse que l'on suppose :

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

Conséquences

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

Analyse discriminante probabiliste

p_j probabilité *a priori* d'appartenir au groupe j

$f_j(\mathbf{x})$ loi des x_i dans le groupe j

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

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

La règle bayésienne dans le cadre normal

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

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

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

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

règle quadratique

Hypothèse simplificatrice : $\sum_1 = \sum_2 \quad \ldots \quad = \sum$

On attribue x au groupe j tel que :

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

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

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

Analyse discriminante probabiliste: cas de deux groupes

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

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

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

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

Fonction de score et probabilité

- Fonction de score S(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(**x**)>0

- Probabilité d'appartenance au groupe 1 :

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

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

Probabilité a posteriori

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

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

Fonction logistique du score

3.2 La régression logistique

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

Le modèle logistique simple

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

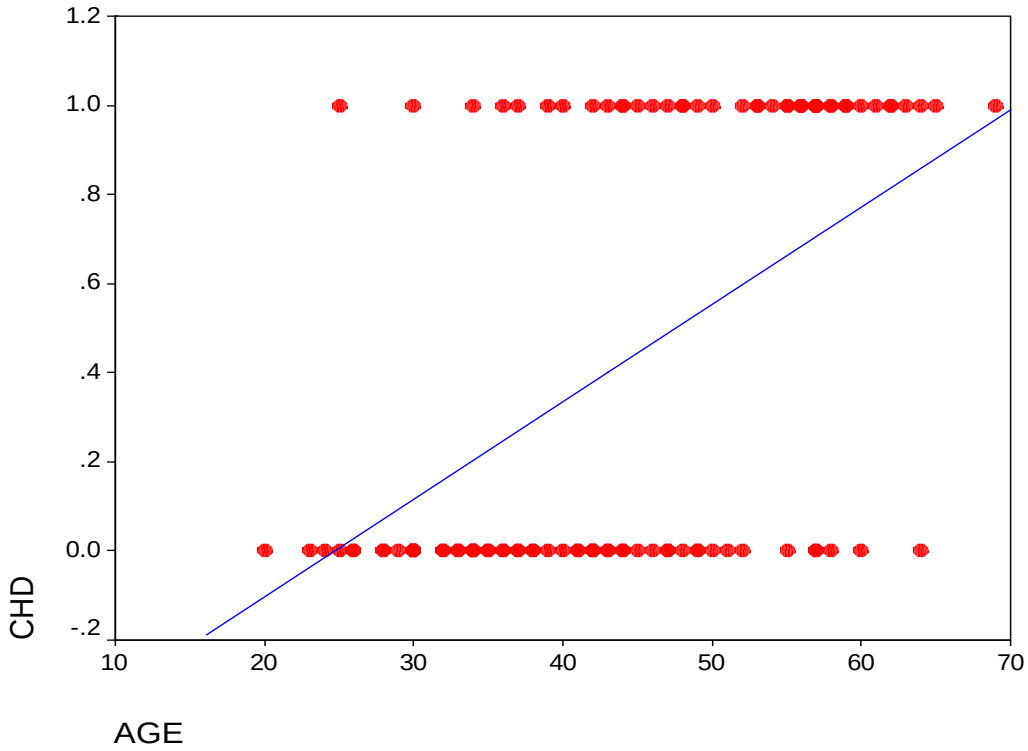
$$\pi(x) = \text{Prob}(Y = 1/X = x)$$

- 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

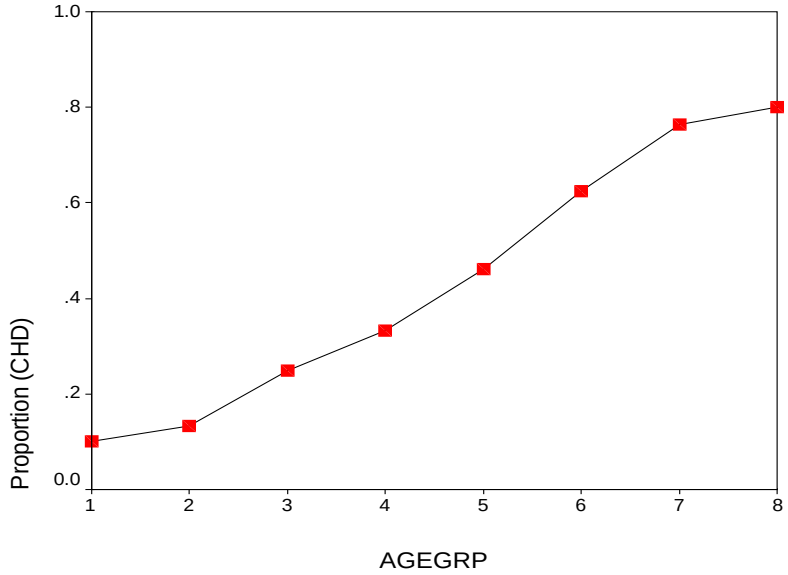

1.2
1.0
.8
.6
.4
.2
0.0
-.2
CHD
10
20
30
40
50
60
70
AGE

Description des données regroupées par classe d'age

Tableau des effectifs de CHD par classe d'age

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'age



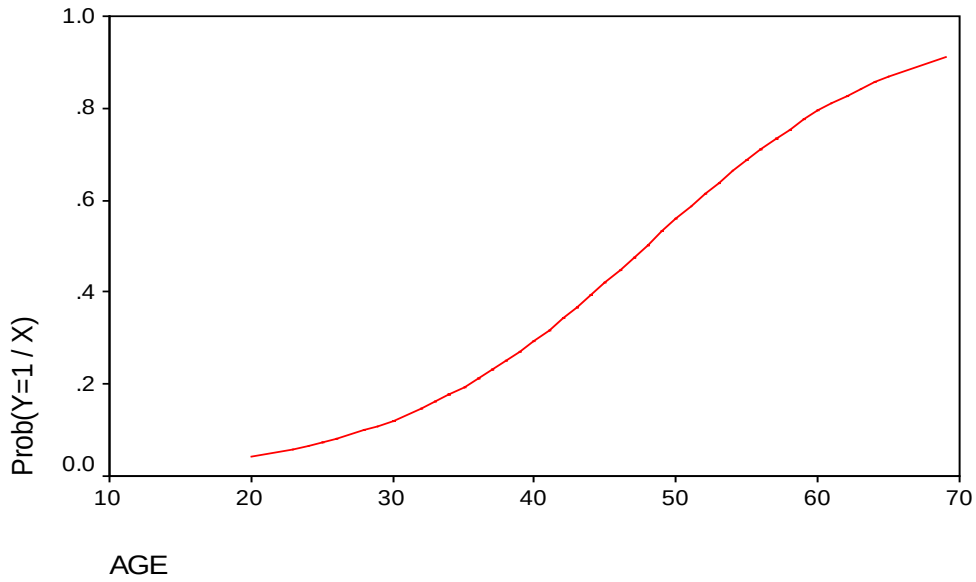
Le modèle logistique simple

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

ou

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

Fonction de lien : Logit

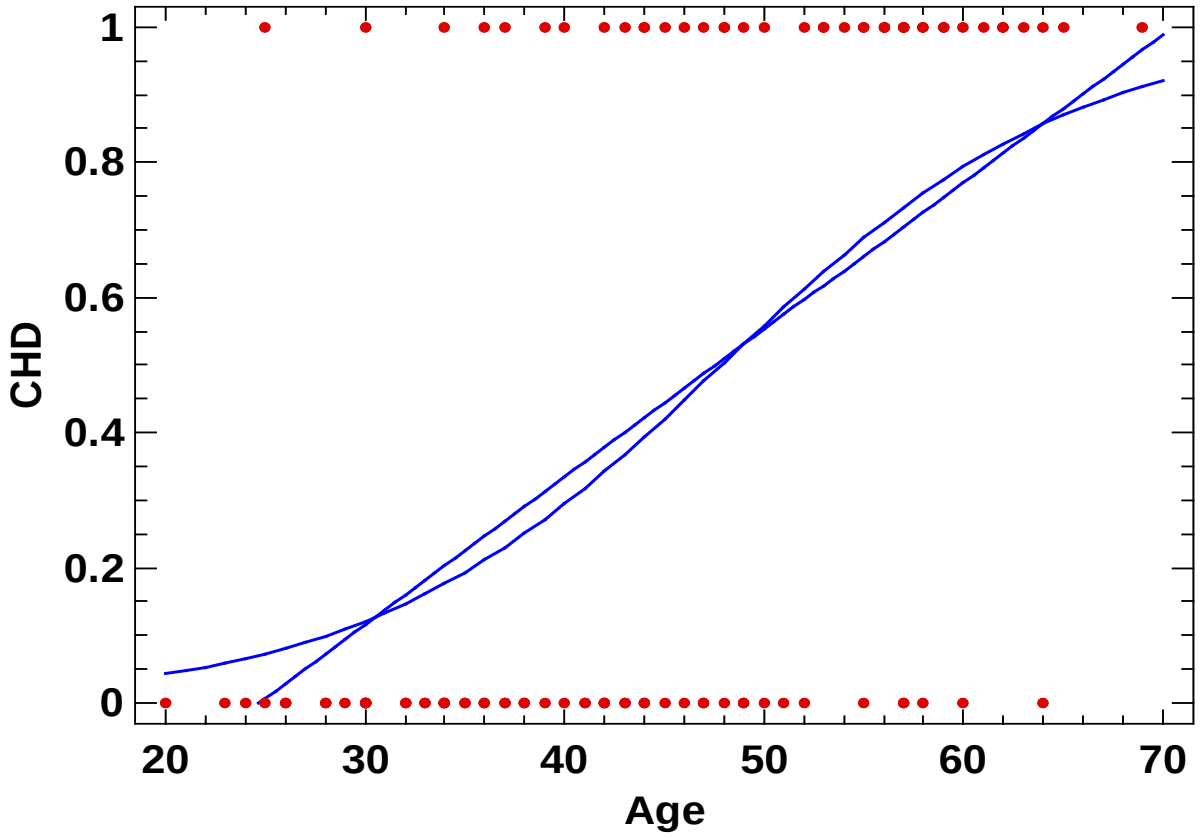


- Il s'agit bien d'un probléme de régression:
- Modélisation de l'espérance conditionnelle

$$E(Y/X=x)=f(x)$$

- Choix de la forme logistique en épidémiologie:
  - S'ajuste bien
  - Interprétation de β_1 en termes d'odds-ratio

comparaison régressions linéaire et logistique



Odds-Ratio

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

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

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

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.4 Estimation des paramètres

Les données

X	Y
x_1	y_1
⋮	⋮
x_i	y_i
⋮	⋮
x_n	y_n

y_i = 1 si caractère présent,
0 sinon

Le modèle

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

Vraisemblance (conditionnelle!)

Probabilité d'observer les données

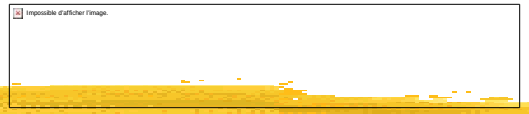
$$[(x_1,y_1), \ldots, (x_i,y_i), \ldots, (x_n,y_n)]$$

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

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

maximum de vraisemblance

- $\hat{\beta}_0$ et $\hat{\beta}_1$ maximisent



- Maximisation de la log-vraisemblance

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

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

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

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

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

Régression logistique multiple

- Généralisation à p variables explicatives $X_1,\ldots, X_p$.

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

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

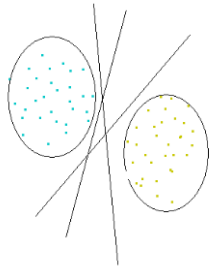
Probabilités a posteriori et stratification

- Estimer P demande de connaître les vraies probabilités a priori
- Les modifier change seulement β_0 en ADL et en logistique:on ajoute $\ln\left(\frac{p_1}{p_2}\right)$
- Important pour les probabilités , pas pour un score

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)

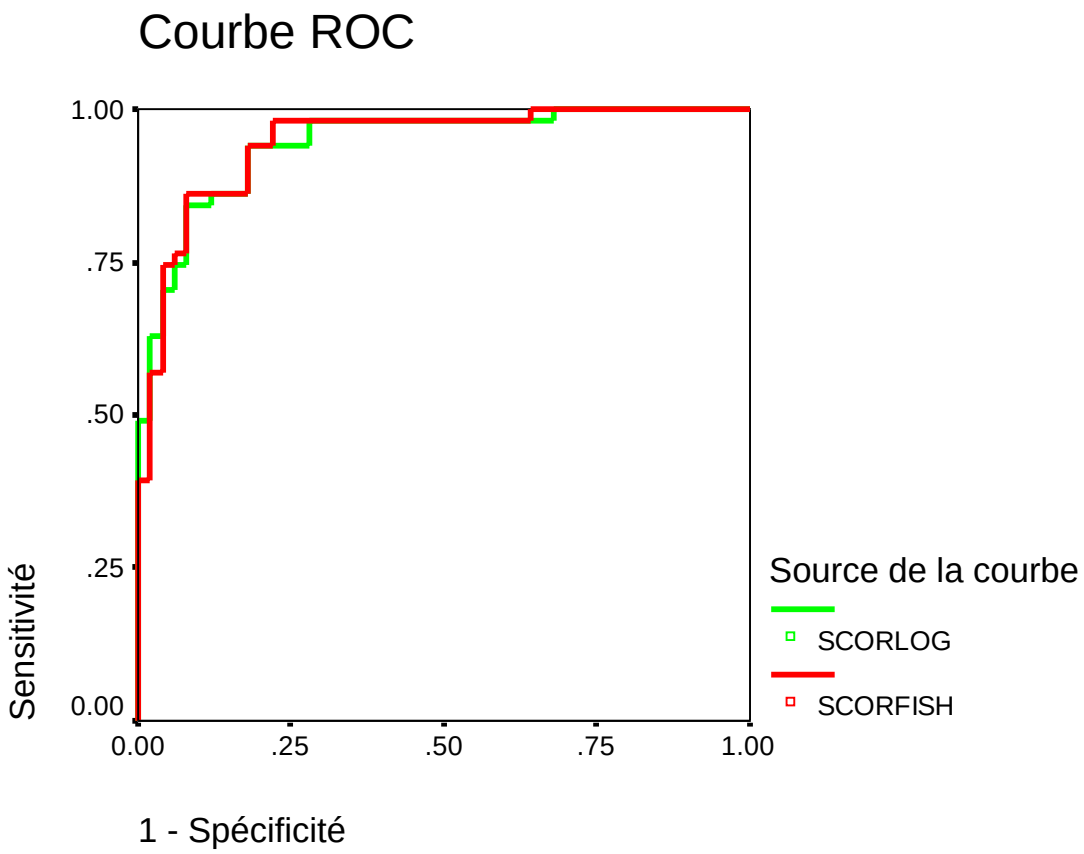
- Mais:
  - Erreurs standard asymptotiques , bootstrap en AD
  - Non convergence en cas de séparation parfaite. Fisher existe toujours



  - 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

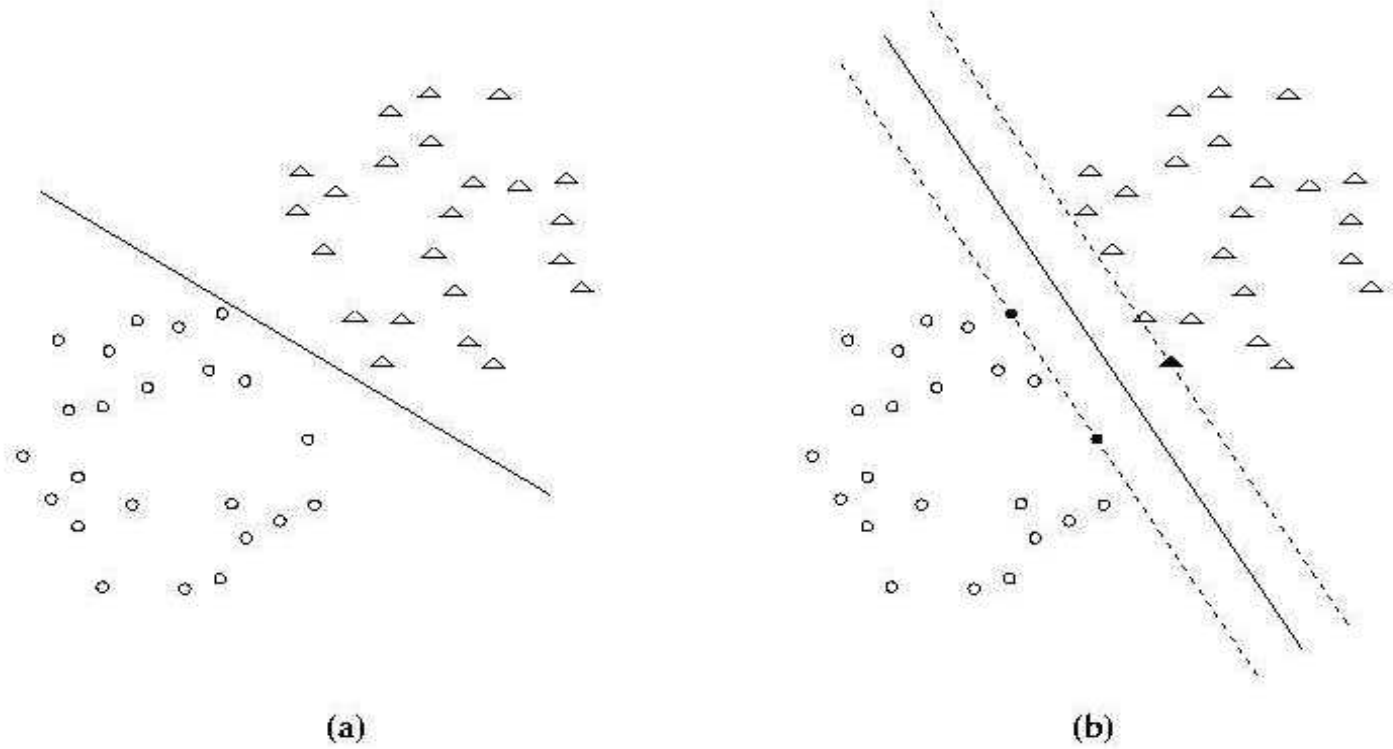
Variable(s) de	Zone
SCORFISH	.945
SCORLOG	.943

- Usages souvent différents: AD pour classer, logistique pour modéliser (facteurs de risque)

- 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 sur des données test

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

L'hyperplan optimal (Vapnik)



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

Un peu de géometrie

- 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:

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

Cas séparable

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

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

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

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

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

Programme quadratique

- Lagrangien: $\|\mathbf{w}\|^2 - 2\sum \alpha_i \left[y_i (\mathbf{x}_i^{'} \mathbf{w} + b) - 1 \right]$
- D'où: $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \quad \text{et} \sum_{i=1}^{n} \alpha_i y_i = 0$

- Conditions de Kühn et Tucker:

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

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

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

- *w*, donc l'hyperplan, ne dépend que des points supports, proches de la frontière, où les α_i sont non nuls.

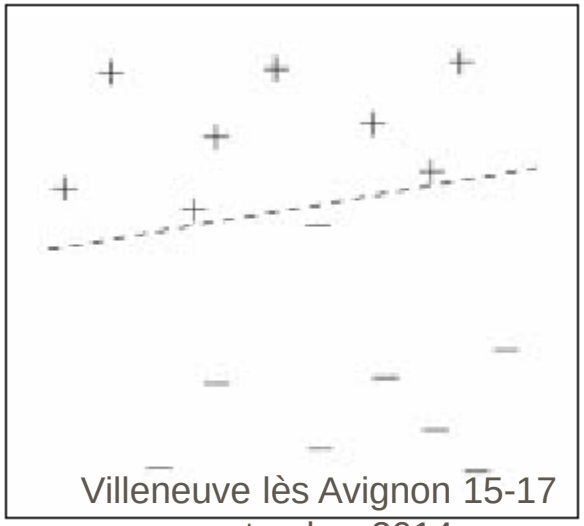
- Solution

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

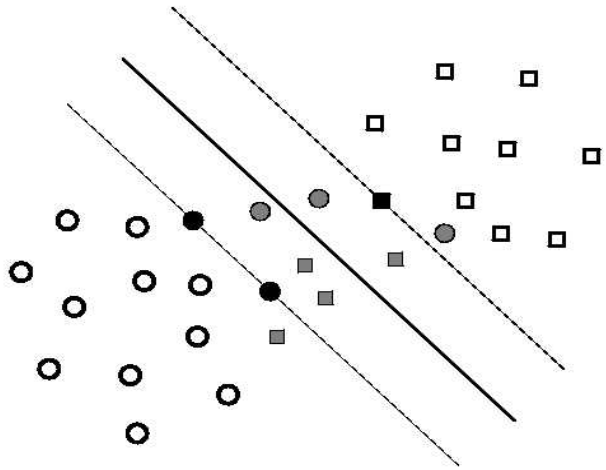
$$f(\mathbf{x}) = \langle \mathbf{w} | \mathbf{x} \rangle + b = \sum_{\alpha_i > 0}^{n} \alpha_i y_i \langle \mathbf{x_i} | \mathbf{x} \rangle + b = \sum_{\alpha_i > 0}^{n} \alpha_i y_i \mathbf{x}_i^{'} \mathbf{x} + b$$

- $f(\mathbf{x})$ 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 (différe de Fisher)
- Plus la marge est grande, meilleure est la robustesse en principe.
- Mais pas toujours :



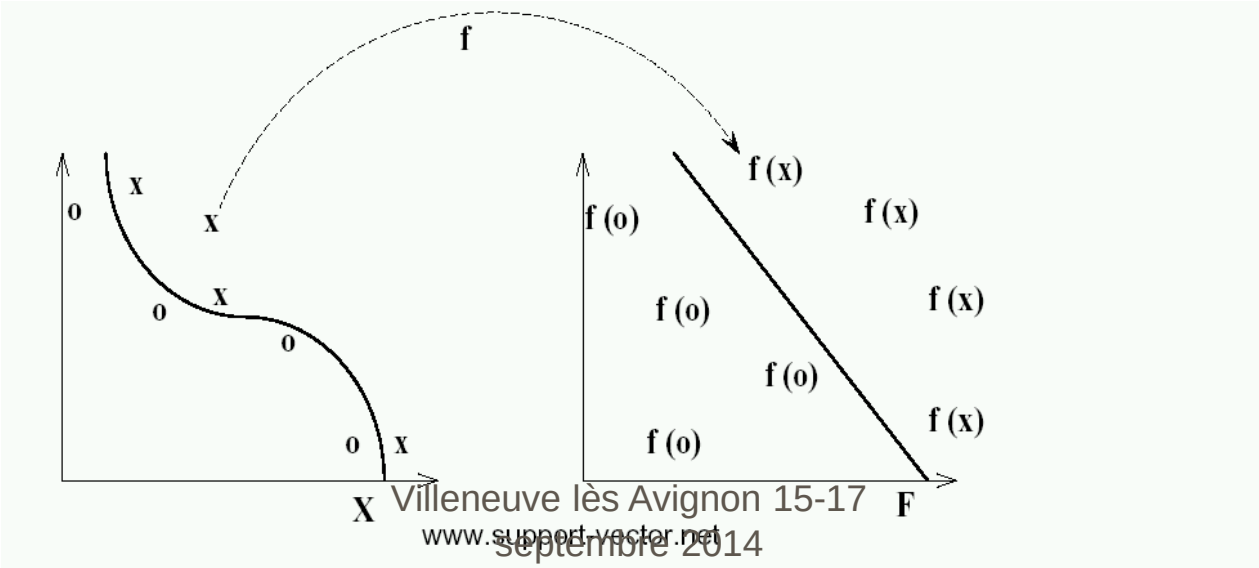
Le cas non séparable



changer d’espace pour rendre le problème linéairement séparable

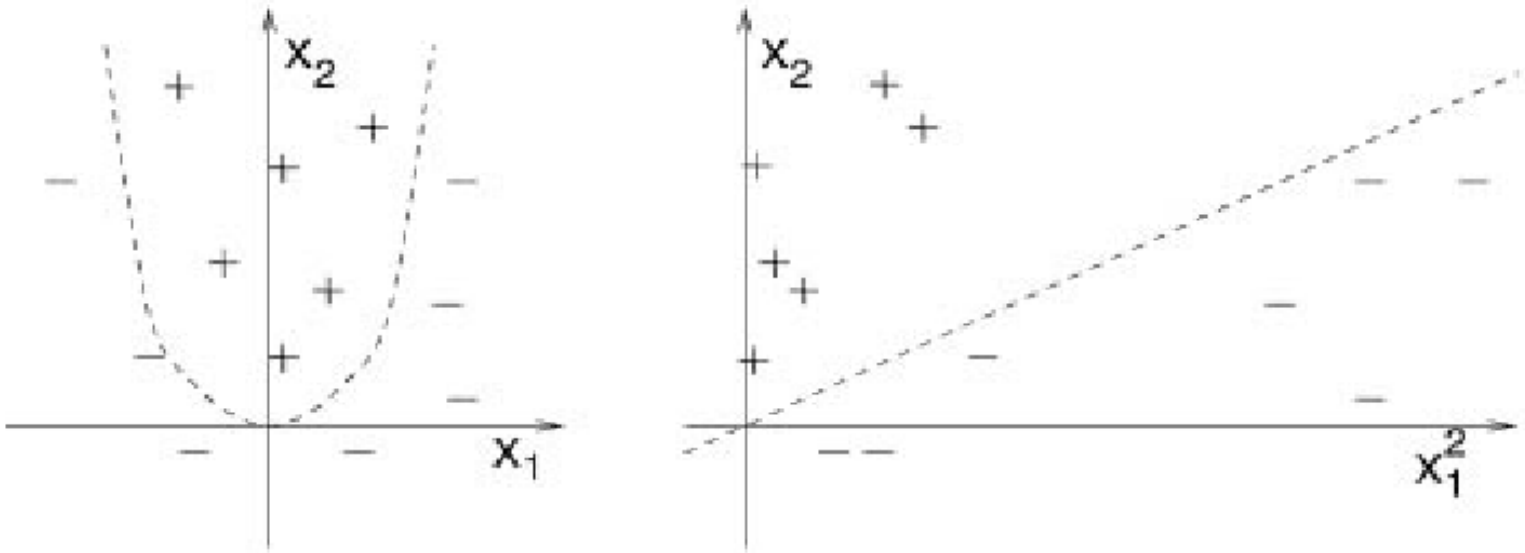
SVM non-linéaires

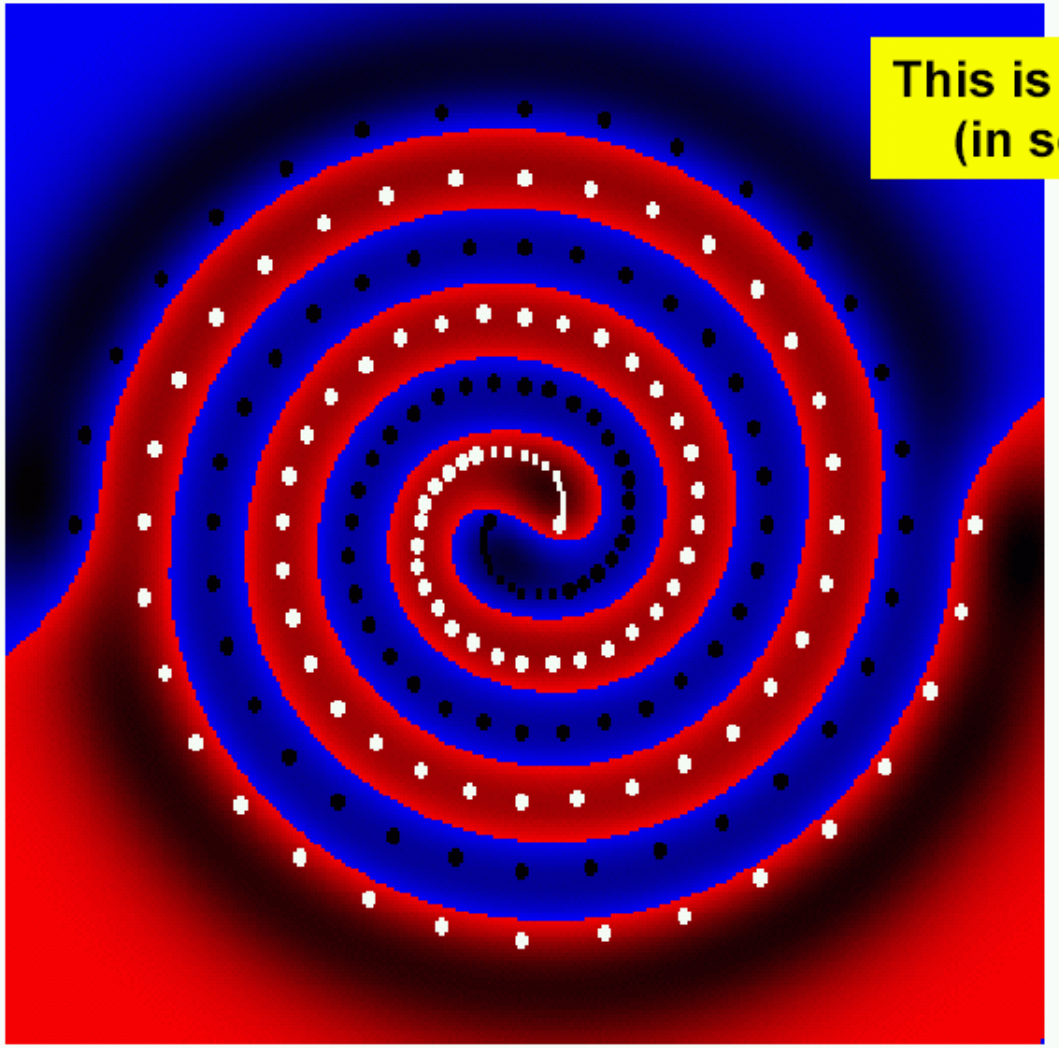
- Passage dans un espace de données transformées (« feature space ») de grande dimension
- Un séparateur linéaire dans $\Phi(E)$ donne un séparateur non-linéaire dans E.



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)





www.support-vector.net/nello.html

Solution

$$\begin{cases} \max\left[\sum \alpha_i - \frac{1}{2}\sum\sum \alpha_i \alpha_k y_i y_k \left\langle \mathbf{\Phi(x_i)} \middle| \mathbf{\Phi(x_k)} \right\rangle\right] \\ 0 < \alpha_i < C \quad \text{et} \sum \alpha_i y_i = 0 \end{cases}$$

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

- 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:

$$\Phi(x)\Phi(x') = x_1^2 x_1'^2 + 2x_1x_2x_1'x_2' + x_2^2x_2'^2$$
$$= (x_1x_1' + x_2x_2')^2 = (xx')^2$$

- On peut donc calculer le produit scalaire dans Φ(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)

$$K(x;x')=\exp-(||x-x'||^2)/\sigma^2)$$

Logiciel gratuit: libsvm

http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Approches voisines

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

3.4 Arbres de décision

- A l'origine développées autour de 1960 en sciences sociales pour détecter des interactions (AID puis ChAID), très utilisées en marketing.
- Regain d'intérêt avec les travaux de Breiman & al. (1984) devenus un des outils les plus populaires du **data mining** en raison de la lisibilité des résultats.
- On peut les utiliser pour prédire une réponse 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

- On sélectionne tout d'abord la variable explicative qui explique le mieux la réponse . D'où une première division de l'échantillon en deux (ou plusieurs sous-ensembles). On présentera plus tard des critères permettant de diviser un segment.
- Puis on réitère cette procédure à l'intérieur de chaque sous-ensemble 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 complexité peut être exponentielle.

logiciels gratuits:

- SIPINA http://eric.univ-lyon2.fr/~ricco/sipina.html
- TANAGRA http://eric.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html
- Package R: rpart

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

La méthode CART (Breiman, Friedman, Olshen, Stone)

- 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

Discrimination : critère de division

- Impureté d'un nœud :

$$i(t) = \sum_{r}^{k} \sum_{s}^{k} P(r/t) P(s/t)$$

  - Avec r≠ 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) ≠ 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 :

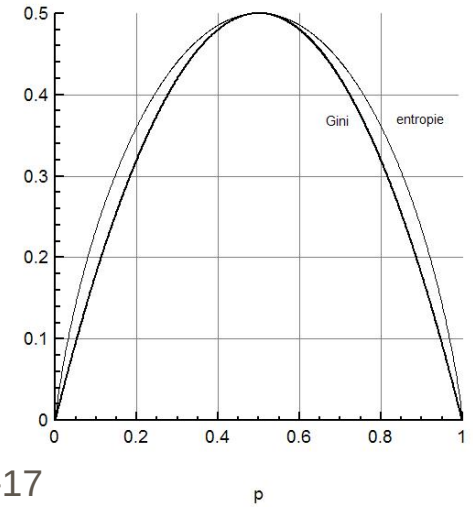
$$\Delta i(s^*,t) = \max\{\Delta i(s,t)\}$$

- Réduction maximale pour l 'ensemble des p variables :

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

Entropie et indice de Gini

- **entropie** $-\sum_{i=1}^{k} p_i \ln(p_i)$
- **indice de diversité de Gini** $\sum_{} p_i p_j$
- **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 noeud terminal est affecté à la classe qui est la mieux représentée (règle majoritaire)

Discrimination : T.E.A.

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

- T.E.A associé à l 'arbre :
  - 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 $\{A_{max}\text{-}1,\ldots A_h, \ldots A_1\}$ avec $1 \leq h < max$*
  - *Le taux d'erreur en apprentissage (TEA) de $_{Ah}$ 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 :

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

Avantages et inconvénients

- Alternative intéressante aux méthodes paramétriques usuelles : pas d'hypothèse sur les données,
- Efficace pour cas non linéaires
- Résultats très lisibles
- MAIS : elles fournissent souvent des arbres instables (les branches coupées ne repoussent jamais...).
- Nécessité de grands échantillons

4. Réponses à plus de 2 catégories

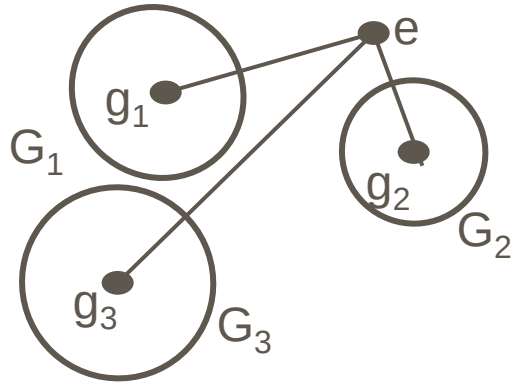
- **4.1 Cas de catégories non ordonnées**
  - Analyse discriminante géomètrique ou probabiliste
  - Régression logistique multinomiale

  - Il n'y a plus un seul classifieur (une seule fonction de score)

Méthodes géométriques de classement

y	x^1 ... x^p
1	
1	
2	
.	
.	
.	
1	

e	
?	



- Échantillon d'apprentissage
- e observation de groupe inconnu
- e classé dans le groupe i tel que: $d(e ; g_i)$ minimal

fonctions discriminantes

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

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

k groupes $\Rightarrow$ k fonctions discriminantes

	1	2	.....	k
1	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	

Discriminante probabiliste

p_j probabilité *a priori* d'appartenir au groupe j

f_j (**x**) loi des x_i dans le groupe j

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

f_j (**x**) gaussiennes

Régression logistique multinomiale

On modélise les probas d'appartenance à k-1 classes.
La dernière s'en déduit aisément.

$$\text{Prob}(Y = i / x) = \frac{e^{\alpha_i + \beta_{i1} x_1 + \cdots + \beta_{ik} x_k}}{1 + \sum_{i=1}^{k-1} e^{\alpha_i + \beta_{i1} x_1 + \cdots + \beta_{ik} x_k}} \quad , \quad i = 1, \ldots, k-1$$

$$\text{Prob}(Y = k / x) = \frac{1}{1 + \sum_{i=1}^{k-1} e^{\alpha_i + \beta_{i1} x_1 + \cdots + \beta_{ik} x_k}}$$

Cas d'un grand nombre de catégories pour Y

- méthodes précédentes déconseillées
- Approche Machine Learning
  - Un contre un: k(k-1)/2 analyses
  - Un contre les autres: k-1 analyses

  - Classement selon règle majoritaire (ou majoritaire pondérée, si coûts d'erreur très différents)

4.2 Réponse à modalités ordonnées

- Régression logistique: modèle à pentes égales

$$\text{Prob}(Y \leq i / x) = \frac{e^{\alpha_i + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\alpha_i + \beta_1 x_1 + \cdots + \beta_k x_k}}$$

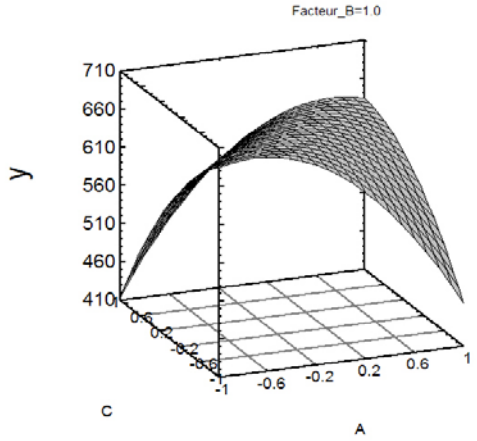
- Odds-ratio proportionnels:

$$\frac{\text{Prob}(Y \leq i/x) / \text{Prob}(Y > i/x)}{\text{Prob}(Y \leq i/x') / \text{Prob}(Y > i/x')} = \frac{e^{\alpha_i + x\beta}}{e^{\alpha_i + x'\beta}} = e^{(x - x')\beta}$$

5. Plans d'expériences et modèles de classification

- Valeurs des x fixées à l'avance, non aléatoires
- Typiquement quelques points expérimentaux avec des répétitions
- Plans classiques optimaux pour estimer des modèles linéaires (y compris avec intéractions et effets quadratiques)
- Points souvent aux extrémités du domaine expérimental

- Surfaces de réponse: inadapté au cas d'une réponse binaire



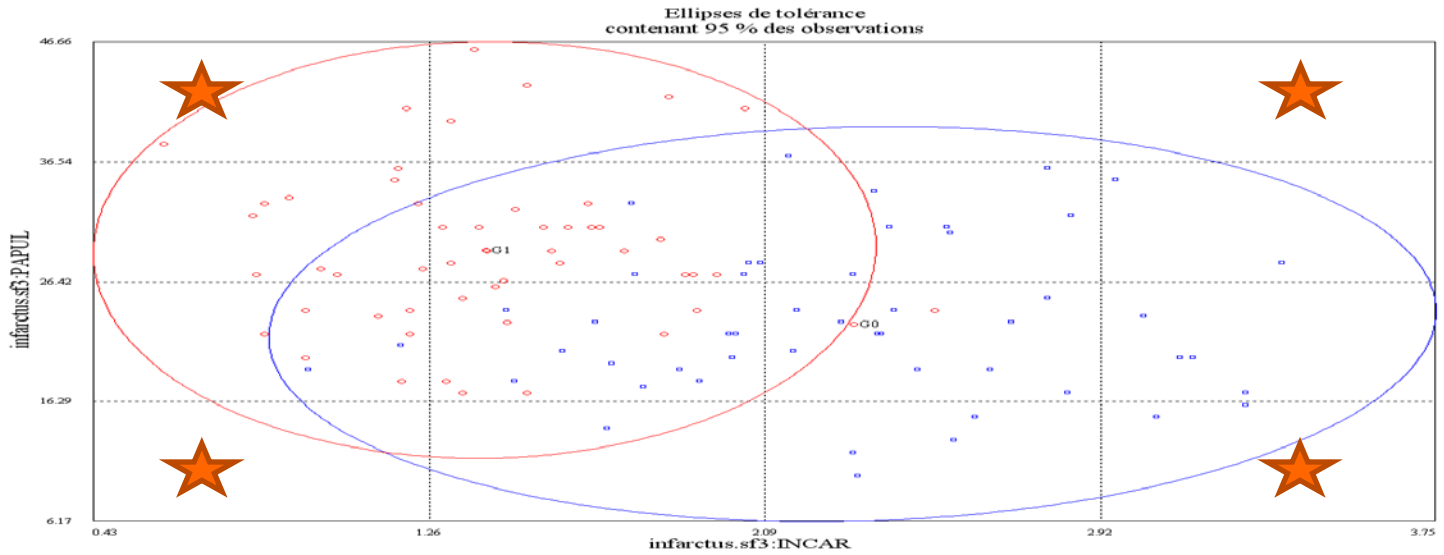
- Permet cependant d'estimer un classifieur avec de meilleurs réultats qu'un échantillonnage

Avantages des plans orthogonaux ou quasi orthogonaux

- Les X deviennent « indépendantes »; mais ce ne sont plus des variables aléatoires
- Matrices de covariance matrices (W, V) inversibles
- Sélection de variables facile et non ambigüe

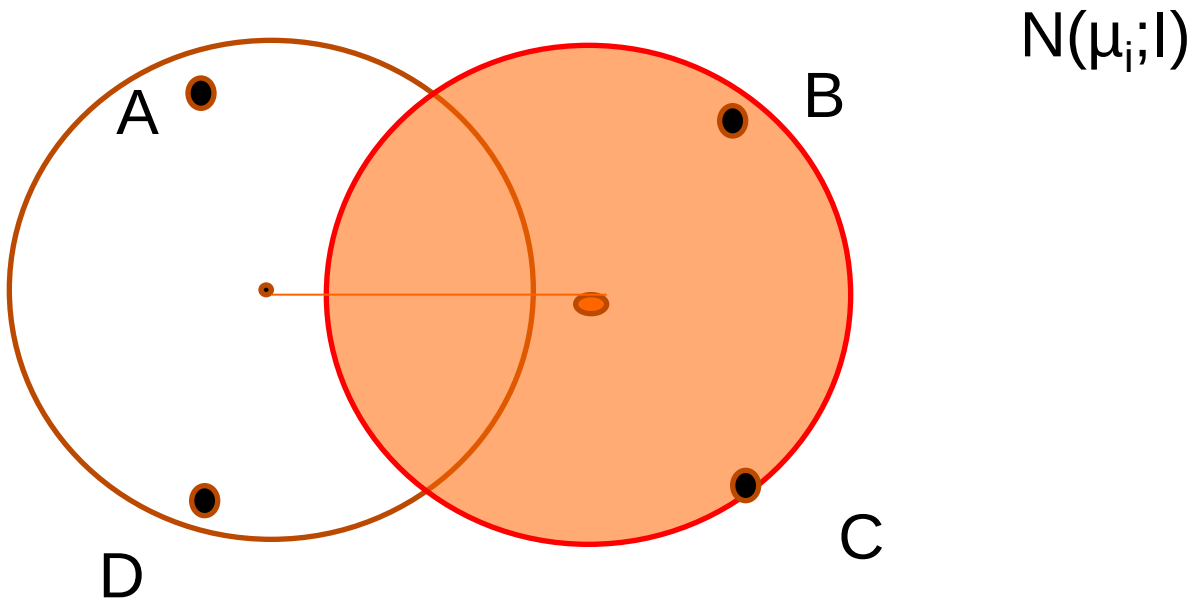
Inconvénients

- Forte probabilité de séparation complète: inadapté à la régression logistique



- Inadapté aux SVM: trop loin des points support

- Probabilité de séparation complète
  - Dépend de la distance de Mahalanobis entre les groupes



- Exemple:
  - 4 points écartés d'un écart-tpe de leurs centroïdes
  - 2 répétitions (n=8)
  - Séparation complète si Y=-1 pour A et D, et Y=1 pour B et C
  - Si Δ= 3, la probabilité de séparation complète vaut 0.99; si Δ=2 prob=0.86, si Δ=1 prob=0.2

- Conclusion:
  - Estimer un classifieur lorsque les données proviennent d'un plan expérimental classique marche bien pour les modèles linéaires (AD de Fisher) mais pas pour les modèles non-linéaires
  - Le plan doit être adapté au modèle
  - Nécessité de critères différents de la D ou A optimalité (X-optimalité ...) . Algorithmes difficiles
  - Peu de travaux:

Min Yang, Bin Zhang and Shuguang Huang (2011)
Optimal designs for generalized linear models with multiple design variables
Statistica Sinica **21**, 1415-1430

J.-P. Vila and J.-P. Gauchi. (2007) Optimal designs based
on exact confidence regions for parameter estimation of a nonlinear regression model.
Journal of Statistical Planning and Inference, **137**(9): 2935-2953

6. Combinaison de modèles

- Bayesian Model Averaging

$$P(y / \mathbf{x}) = \sum_{i=1}^{m} P(y / M_i, \mathbf{x}) P(M_i / \mathbf{x})$$

$$E(y / \mathbf{x}) = \sum_{i=1}^{m} E(y / M_i, \mathbf{x}) P(M_i / \mathbf{x})$$

Moyenne des prévisions de chaque modèle, pondérées par les probabilités a posteriori

- Stacking
  - Combinaison non bayésienne de m prédictions obtenues par des modèles différents
  - Premiere idée : régression linéaire

$$\hat{f}_1(\mathbf{x}),\ \hat{f}_2(\mathbf{x}),\ldots,\ \hat{f}_m(\mathbf{x})$$

$$\min \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m} w_j\, \hat{f}_j(\mathbf{x}) \right)^2$$

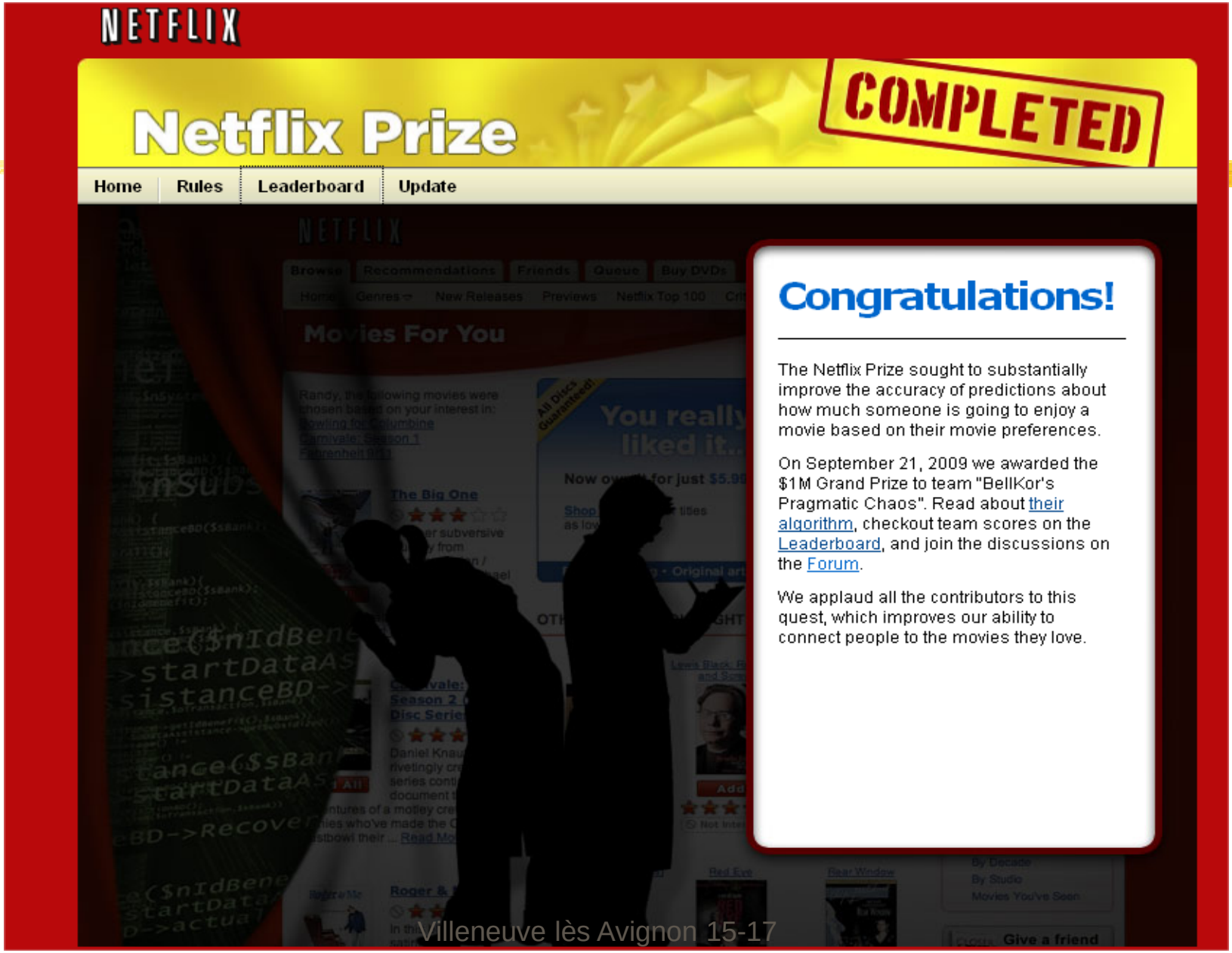
    - Favorise les modèles les plus complexes: surapprentissage

- Solution: utiliser les valeurs prédites en otant à chaque fois l'unité i

$$\min \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} w_j \hat{f}_j^{-i}(\mathbf{x}) \right)^2$$

- Améliorations:
  - Combinaisons linéaires à coefficients positifs (et de somme 1)
  - Régression PLS ou autre méthode régularisée car les m prévisions sont très corrélées

- Avantages
  - Prévision meilleures qu'avec le meilleur modèle
  - Possibilité de mélanger des modéles de toutes natures: arbres , ppv, réseaux de neurones etc. alors que le BMA utilise des modèles paramétrés de la même famille


NETFLIX
Netflix Prize
COMPLETED
Home
Rules
Leaderboard
Update
Congratulations!
The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.
On September 21, 2009 we awarded the $1M Grand Prize to team "BellKor's Pragmatic Chaos". Read about their algorithm, checkout team scores on the Leaderboard, and join the discussions on the Forum.
We applaud all the contributors to this quest, which improves our ability to connect people to the movies they love.

- BellKor's Pragmatic Chaos team bested Netflix's own algorithm for predicting ratings by 10.06%.
- "The Netflix dataset contains more than 100 million datestamped movie ratings performed by anonymous Netflix customers between Dec 31, 1999 and Dec 31, 2005 This dataset gives ratings about m = 480,189 users and n = 17,770 movies The contest was designed in a training-test set format. A Hold-out set of about 4.2 million ratings was created consisting of the last nine movies rated by each user (or fewer if a user had not rated at least 18 movies over the entire period).The remaining data made up the training set."
- Blend of 24 predictions

Application L'Oréal

- Référence: C. Gomes , H. Nocairi , M. Thomas , F. Ibanez , J. Collin , G. Saporta - **Stacking prediction for a binary outcome**, Compstat 2012, August 2012, pp.271-282, Limassol, Chypre,

Téléchargeable à http://cedric.cnam.fr/fichiers/art_2626.pdf

Merci!