Le classifieur bayesien naïf (Modèle d'Indépendance Conditionnelle)

Une approche pour rendre calculable P(Y/X)

Ricco RAKOTOMALALA

Théorème de Bayes
Probabilité conditionnelle

Estimer la probabilité conditionnelle

$$P(Y = y_k / \aleph) = \frac{P(Y = y_k) \times P(\aleph / Y = y_k)}{P(\aleph)}$$

$$= \frac{P(Y = y_k) \times P(\aleph / Y = y_k)}{\sum_{k=1}^{K} P(Y = y_k) \times P(\aleph / Y = y_k)}$$

Déterminer la conclusion = déterminer le max.

$$y_{k*} = \arg \max_k P(Y = y_k / \aleph)$$

$$\Leftrightarrow$$

$$y_{k*} = \arg \max_k P(Y = y_k) \times P(\aleph / Y = y_k)$$

Probabilité a priori
Estimée facilement par n_k/n

Comment estimer $P(X/Y=y_k)$?

Impossibilité à estimer avec des fréquences
Le tableau croisé serait trop grand et rempli de zéros

Cas des variables prédictives qualitatives

Modèle d'indépendance conditionnelle

Hypothèse d'indépendance conditionnelle

$$P(\mathbf{X} / Y = y_k) = \prod_{j=1}^{J} P(X_j / Y = y_k)$$

Les descripteurs sont deux à deux indépendants conditionnellement aux valeurs prises par Y

Pour un descripteur X discret quelconque, la probabilité conditionnelle pour qu'elle prenne la valeur x_l s'écrit

$$P(X = x_l / Y = y_k) = \frac{P(X = x_l \wedge Y = y_k)}{P(Y = y_k)}$$

Et son estimation par les fréquences (profil ligne)

$$\hat{P}(X = x_l / Y = y_k) = \frac{\#\{\omega \in \Omega, X(\omega) = x_l \wedge Y(\omega) = y_k\}}{\#\{\omega \in \Omega, Y(\omega) = y_k\}} = \frac{n_{kl}}{n_k}$$

$Y \setminus X$		x_l		Σ
y_k		n_{kl}		n_k
Σ				n

On lui préfère souvent l'estimateur « laplacien » des probabilités !!!

$$\hat{P}(X = x_l / Y = y_k) = p_{l/k} = \frac{n_{kl} + 1}{n_k + L}$$

(1) « Lissage » des estimations sur les petits effectifs.
(2) Eviter le problème du n_{kl}=0

Un exemple

Maladie	Marié	Etud.Sup
Présent	Non	Oui
Présent	Non	Oui
Absent	Non	Non
Absent	Oui	Oui
Présent	Non	Oui
Absent	Non	Non
Absent	Oui	Non
Présent	Non	Oui
Absent	Oui	Non
Présent	Oui	Non

Estimation directe

$$\hat{P}(Maladie = Absent / Marié = oui, Etu = oui) = \frac{1}{1} = 1$$

$$\hat{P}(Maladie = \text{Présent} / Marié = oui, Etu = oui) = \frac{0}{1} = 0$$

→ Si Etu = oui et Marié = oui Alors Maladie = Absent

(+) Calcul sans hypothèses restrictives, (-) effectifs indigents

NB Maladie	
Maladie	Total
Absent	**5**
Présent	5
Total général	10

NB Maladie	Marié		
Maladie	Non	**Oui**	Total général
Absent	2	**3**	5
Présent	4	1	5
Total général	6	4	10

NB Maladie	Etud.Sup		
Maladie	Non	**Oui**	Total général
Absent	4	**1**	5
Présent	1	4	5
Total général	5	5	10

Indépendance conditionnelle

$$\hat{P}(Maladie = Absent / Marié = oui, Etu = oui)$$

$$= \hat{P}(Maladie = Absent) \times \hat{P}(Marié = oui / M = Abs.) \times \hat{P}(Etu = oui / M = Abs.)$$

$$= \frac{5+1}{10+2} \times \frac{3+1}{5+2} \times \frac{1+1}{5+2} = 0.082$$

$$\hat{P}(Maladie = présent / Marié = oui, Etu = oui)$$

$$= \hat{P}(Maladie = présent) \times \hat{P}(Marié = oui / M = Abs.) \times \hat{P}(Etu = oui / M = Abs.)$$

$$= \frac{5+1}{10+2} \times \frac{1+1}{5+2} \times \frac{4+1}{5+2} = 0.102$$

→ Si Etu = oui et Marié = oui Alors Maladie = Présent

(-) Hypothèse discutable, (+) estimations des probas (effectifs) plus fiables

Avantages et inconvénients (fin du cours ?)

+

>> Simplicité, rapidité de calcul, capacité à traiter de très très très grandes bases (lignes , colonnes) (aucun risque de « plantage », cf. la régression logistique ou l'ADL)

>> Incrémentalité (table des probas conditionnelles à maintenir)

>> Robustesse (performant même si hypothèse non-respectée)

>> C'est un modèle linéaire → même niveau de performances
(cf. les nombreuses expérimentations dans les publications scientifiques)

-

>> Pas de sélection (mise en évidence) des variables pertinentes (sûr, sûr ?)

>> Nombre de règles égal au nombre de combinaisons de descripteurs
(dans la pratique, les règles ne sont pas formées, nous conservons les probas conditionnelles que nous appliquons pour chaque individu à classer ; cf. quelques logiciels + format PMML)

>> Pas de modèle explicite (sûr, sûr ?) → très utilisé en recherche, peu en marketing

On s'en tient souvent à ces conclusions dans les ouvrages…
On ne peut pas aller plus loin ?

Dériver un modèle explicite à partir du bayesien naïf

Passage au logarithme

$$y_{k^*} = \arg\max_k P(Y = y_k) \times \prod_{j=1}^{J} P(X_j / Y = y_k)$$

$$\Leftrightarrow y_{k^*} = \arg\max_k \left[\ln P(Y = y_k) + \sum_{j=1}^{J} \ln P(X_j / Y = y_k) \right]$$

Cas d'une seule prédictive X à L modalités

$$d(y_k, X) = \ln P(Y = y_k) + \ln P(X / Y = y_k)$$

A partir de X on peut dériver L indicatrices

$$\begin{aligned} d(y_k, X) &= \ln P(Y = y_k) + \sum_{l=1}^{L} \ln P(X = x_l / Y = y_k) \times I_l \\ &= \ln P(Y = y_k) + \sum_{l=1}^{L} \ln P(X = x_l / Y = y_k) \times I_l \\ &= a_{0,k} + \sum_{l=1}^{L} a_{l,k} \times I_l \end{aligned}$$

On obtient une combinaison linéaire d'indicatrices : un modèle explicite facile à déployer
Exactement comme avec la régression logistique ou l'ADL !!!
→ Fonction de classement

Un exemple

NB Maladie			
Maladie	Total		
Absent	**5**		
Présent	5		
Total général	10		

NB Maladie	Etud.Sup		
Maladie	Non	**Oui**	Total général
Absent	4	**1**	5
Présent	1	4	5
Total général	5	5	10

$$d(absent, X) = \ln\frac{5+1}{10+2} + \ln\frac{4+1}{5+2}\times(X = non) + \ln\frac{1+1}{5+2}\times(X = oui)$$
$$= -0.6931 - 0.3365\times(X = non) - 1.2528\times(X = oui)$$

$$d(present, X) = -0.6931 - 1.2528\times(X = non) - 0.3365\times(X = oui)$$

Pour un individu Etu.Sup = NON

$$d(absent, X) = -0.6931 - 0.3365 = -1.0296$$
$$d(present, X) = -0.9631 - 1.2528 = -1.9495$$

Prédire l'ABSENCE de la maladie

Implémentation dans Tanagra
Utilisation de (L-1) indicatrices pour une variable X à L modalités

Prior distribution of class attribute "Maladie"

Values	Count	Percent	Histogram
Absent	5	50.00 %	
Présent	5	50.00 %	

Model description

	Classification functions	
Descriptors	Absent	Présent
Etud.Sup = Oui	-0.916291	0.916291
constant	-1.029619	-1.945910

puisque $I_1 + I_2 + \cdots + I_L = 1$

$$d(y_k, X) = \ln P(Y = y_k) + \sum_{l=1}^{L} \ln P(X = x_l / Y = y_k) \times I_l$$

$$= \ln P(Y = y_k) + \ln P(X = x_L / Y = y_k) + \sum_{l=1}^{L-1} \ln \frac{P(X = x_l / Y = y_k)}{P(X = x_L / Y = y_k)} \times I_l$$

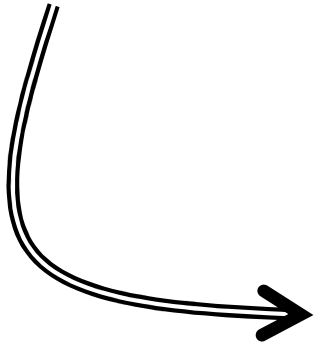
$$= b_{0,k} + \sum_{l=1}^{L-1} b_{l,k} \times I_l$$

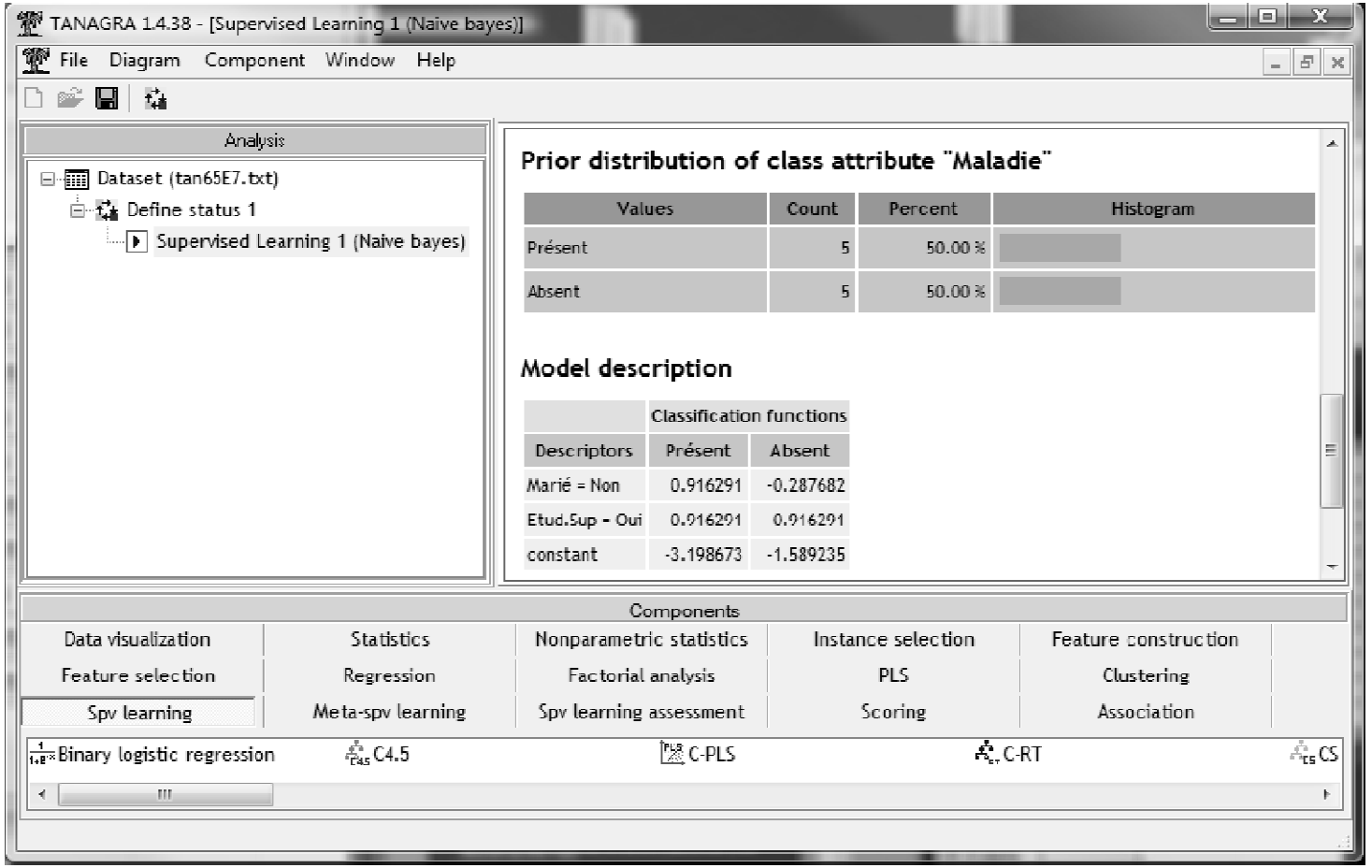
Codage habituellement utilisé pour la régression logistique et l'ADL

Maladie	Marié	Etud.Sup
Présent	Non	Oui
Présent	Non	Oui
Absent	Non	Non
Absent	Oui	Oui
Présent	Non	Oui
Absent	Non	Non
Absent	Oui	Non
Présent	Non	Oui
Absent	Oui	Non
Présent	Oui	Non

Généralisation à J variables prédictives

Chaque explicative est décomposée en un bloc d'indicatrices
X = L modalités → (L-1) indicatrices





La variable à prédire prend deux modalités :: Y={+,-}

$$\frac{\begin{cases} d(+, X) = a_{+,0} + a_{+,1}X_1 + a_{+,2}X_2 + \cdots + a_{+,J}X_J \\ d(-, X) = a_{-,0} + a_{-,1}X_1 + a_{-,2}X_2 + \cdots + a_{-,J}X_J \end{cases}}{d(X) = c + c_1X_1 + c_2X_2 + \ldots + c_JX_J}$$

Règle d'affectation

D(X) > 0 → Y = +

Interprétation

>> D(X) est communément appelé un score, c'est la propension à être positif.

>> Le signe des coefficients « c » donne une idée sur le sens de la causalité.

>> Les Xj étant des indicatrices (0/1), les coefficients « c » peuvent être lus comme des « points » attribués aux individus portant le caractère Xj

Notre exemple :	Classification functions		SCORE
Descriptors	**Présent**	**Absent**	**D(X)**
Marié = Non	0.916291	-0.287682	1.203973
Etud.Sup = Oui	0.916291	-0.916291	1.832582
constant	-3.198673	-1.589235	-1.609438

Ne pas être marié rend malade.
Faire des études rend malade.

→Hum, ça donne à réfléchir…

Lecture des coefficients
Fonction de classement

Revenons à l'estimation fréquentielle des probabilités

Nombre de Maladie	Marié		
Maladie	Non	Oui	Total général
Présent	0.8	0.2	1.0
Absent	0.4	0.6	1.0
Total général	0.6	0.4	1.0

Résultats Naive Bayes

	Classification functions	
Descriptors	**Présent**	**Absent**
Marié = Non	1.38629	-0.4055
constant	-2.3026	-1.204

$$odds(M = N / M = O; Y = present) = \frac{0.8}{0.2} = 4 \Rightarrow \ln(odds) = 1.386294$$

Les personnes malades (maladie = présent) ont 4 fois plus de chances de ne pas être mariées (que d'être mariées)

Le coefficient de la fonction de classement correspond au logarithme de l'odds

$$odds(M = N / M = O; M = absent) = \frac{0.4}{0.6} = 0.667 \Rightarrow \ln(odds) = -0.4055$$

Chez les personnes non malades, on a (1/0.667) = 1.5 fois plus de chances d'être marié (que de ne pas l'être)

Lecture des coefficients
Fonction Score

Nombre de Maladie	Marié		
Maladie	Non	Oui	Total général
Présent	0.8	0.2	1.0
Absent	0.4	0.6	1.0
Total général	0.6	0.4	1.0

	Classification functions		
Descriptors	**Présent**	**Absent**	**SCORE**
Marié = Non	1.38629	-0.40547	1.79176
constant	-2.30259	-1.20397	-1.09861

$$odds-ratio(M=N/M=O;Y=P/Y=A) = \frac{odds(M=N/M=O;Y=P)}{odds(M=N/M=O;Y=A)} = \frac{4}{0.66} = 6$$

Les personnes malades ont 6 fois plus de chances d'être mariées que les personnes non malades.

$$\ln(6) = 1.79176$$

Le coefficient de la fonction score correspond au logarithme de l'odds-ratio

Commentaires :

- La lecture de l'odds-ratio est inversée par rapport à la régression logistique
- Toutes ces analyses ne sont pertinentes que si la liaison entre X et Y est significative !!!

Sélection des variables

Evaluer la pertinence d'une variable
Eliminer les variables non-pertinentes
Supprimer les redondances

Conséquence étonnante de l'hypothèse d'indépendance conditionnelle

Par construction (indépendance conditionnelle, toutes les estimations sont faites de manière individuelle)
➔ le retrait ou l'ajout de variables explicatives ne perturbe pas les autres coefficients (des autres variables).

	Classification functions	
Descriptors	**Présent**	**Absent**
Marié = Non	0.916291	-0.287682
constant	-1.94591	-1.252763

Modèle à 1 variable

	Classification functions	
Descriptors	**Présent**	**Absent**
Marié = Non	0.916291	-0.287682
Etud.Sup = Oui	0.916291	-0.916291
constant	-3.198673	-1.589235

Modèle à 2 variables

Il n'est pas nécessaire de reconstruire le modèle à chaque ajout ou retrait de variables.

Pertinence d'une variable

Une variable a un fort impact si elle permet d'exacerber les différences entre les fonctions de classement $d(y_k,X)$ (selon y_k)
$\Leftrightarrow$ si distributions conditionnelles $P(X/y_k)$ sont très différentes selon les valeurs de y_k
$\Leftrightarrow$ si dist. conditionnelles $P(X/y_k)$ sont très différentes de la dist. marignale $P(X)$

Nombre de Marié	Etud.Sup		
Maladie	Non	Oui	Total général
Absent	0.8	0.2	1.0
Présent	0.2	0.8	1.0
Total général	0.5	0.5	1.0

Nombre de Marié	Marié		
Maladie	Non	Oui	Total général
Absent	0.4	0.6	1.0
Présent	0.8	0.2	1.0
Total général	0.6	0.4	1.0

$$H(X) = \sum_{l=1}^{L} p_{.l} \log_2 p_{.l}$$

$$H(X/Y) = \sum_{k=1}^{K} p_{k.} \sum_{l=1}^{L} p_{l/k} \log_2 p_{l/k}$$

~ variance totale

~ variance intra-classes

→ ~Variance inter-classes
c.-à-d. variance expliquée

$$H(X) - H(X/Y) = I(Y,X) = \sum_{l=1}^{L} \sum_{k=1}^{K} p_{kl} \log_2 \frac{p_{kl}}{p_{l.} \times p_{k.}}$$

Information mutuelle

On peut établir une hiérarchie entre les variables

$$I(Y, ES) = 0.2781$$

Nombre de Marié	Etud.Sup		
Maladie	Non	Oui	Total général
Absent	0.4	0.1	0.5
Présent	0.1	0.4	0.5
Total général	0.5	0.5	1.0

$$I(Y, M) = 0.1245$$

Nombre de Marié	Marié		
Maladie	Non	Oui	Total général
Absent	0.2	0.3	0.5
Présent	0.4	0.1	0.5
Total général	0.6	0.4	1.0

On peut même tester la significativité du lien

Test (H0 : les deux variables sont indépendantes)

$$G = 2 \times n \times \ln 2 \times I(Y, X)$$
$$\sim \chi^2[(K-1)\times(L-1)]$$

$$G(ES) = 3.85$$
$$\Rightarrow p.value = 0.0496$$

Le lien entre Y et ES est significatif

$$G(M) = 1.73$$
$$\Rightarrow p.value = 0.1889$$

Le lien entre Y et Maladie, NON

Utiliser l'indicateur s (symetrical uncertainty)

Parce que varie entre [0 ; 1]

$$s_{Y,X} = 2 \times \frac{I(Y,X)}{H(Y) + H(X)}$$

Exemple « kr-vs-kp » (19 sél. pour α = 0.001)

Calculations details

N°	Attribute	Values	Statistic	Statistic (Histogram)	p-value
1	rimmx	2	0.452284		0.000000
2	bxqsq	2	0.380101		0.000000
3	wknck	2	0.365265		0.000000
4	bkxwp	2	0.232908		0.000000
5	wkna8	2	0.196557		0.000000
6	r2ar8	2	0.164526		0.000000
7	bkxcr	2	0.163828		0.000000
8	mulch	2	0.158337		0.000000
9	wkpos	2	0.146543		0.000000
10	bkxbq	2	0.139802		0.000000
11	skrxp	2	0.130476		0.000000
12	stlmt	2	0.127724		0.000000
13	wkcti	2	0.126350		0.000000
14	rkxwp	2	0.101402		0.000000
15	bkon8	2	0.091163		0.000000
16	rxmsq	2	0.087799		0.000001
17	blxwp	2	0.085171		0.000001

RANKING :

1. Calculer s pour toutes les variables
2. Les trier par ordre décroissant
3. Ne conserver que les variables significativement liées avec Y

Problèmes ennuyeux :

- Fixer le seuil « alpha » très difficile
- Tout devient significatif dès que « n » est important

→*Solution possible : « loi du coude »*

Problème rédhibitoire :

- Ne tient pas compte de la redondance entre les prédicteurs !!!

Sélection tenant compte de la redondance – Méthode CFS

Le « mérite » d'un sous-ensemble de « p » variables prédictives

$$merit = \frac{p \times \overline{s}_{Y,X}}{\sqrt{p + p \times (p+1) \times \overline{s}_{X,X}}}$$

Numérateur : lien des prédicteurs avec la variable cible (pertinence)

Dénominateur : lien entre les prédicteurs (redondance)

→ On veut que les prédicteurs soient fortement liés avec la cible, tout en étant le moins liés entre eux (les plus orthogonaux possible)

Results

INPUT attribute selection

INPUT selection	
Before filtering	34
After filtering	3

Exemple « kr-vs-kp »
3 var. sélectionnées

Keeped into INPUT selection

	Attributes
1	bxqsq
2	rimmx
3	wknck

Calculations details

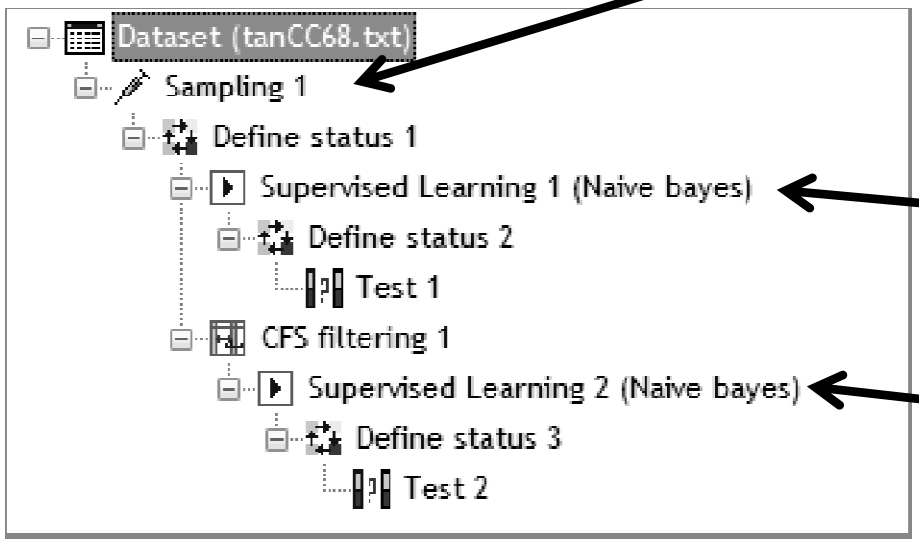
Selected attribute	MERIT(S)
rimmx	0.235390
bxqsq	0.246590
wknck	0.257278

Optimisation « FORWARD »

- Commencer avec 0 variables
- Ajout itératif de variables
c.-à-d. chercher la variable qui augmente le plus le « mérite »
- Etc.

→Arrêt lorsque le critère « mérite » commence à décroître

Pourquoi la sélection de variables ?



1500 apprentissage
1696 test

34 prédicteurs
Taux d'erreur en test = 14.80%

3 prédicteurs
Taux d'erreur en test = 9.67%

La sélection de variables permet de réduire le nombre de variables tout en maintenant le niveau de performances

Parfois, elle peut être bénéfique (comme ici, mais c'est plutôt rare)

Parfois, elle peut être nocive (quand trop restrictive)

Cas des variables prédictives quantitatives (1)

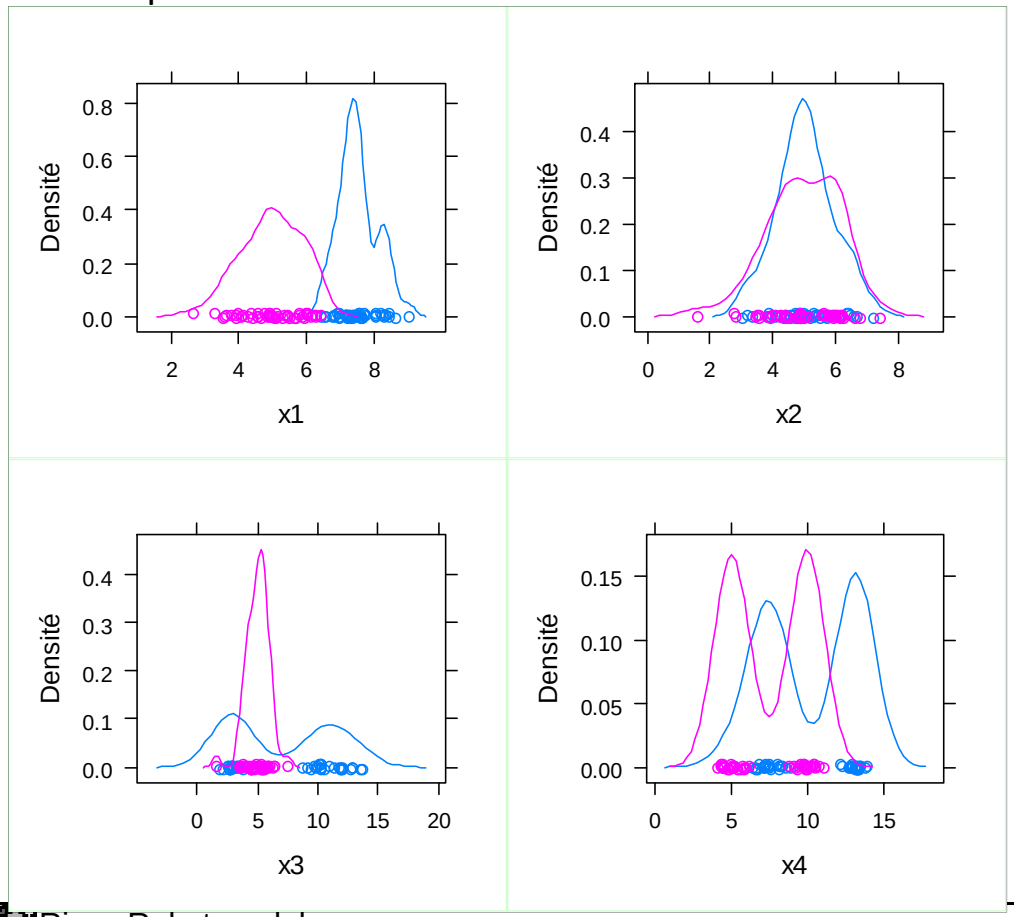
Se ramener au cas des variables qualitatives (précédent)
En discrétisant les prédicteurs quantitatifs

Plusieurs études montrent que cette approche est la pus intéressante
C'est aussi l'approche à privilégier si mélange de prédicteurs quanti et quali

Découpage en intervalles des variables quantitatives
Utilisation d'algorithmes supervisés

Les algorithmes usuels (ex. fréquences égales, largeurs égales) ne tiennent pas compte de la représentation des modalités de Y dans les intervalles → inadaptés pour l'apprentissage supervisé.

4 exemples de distributions conditionnelles



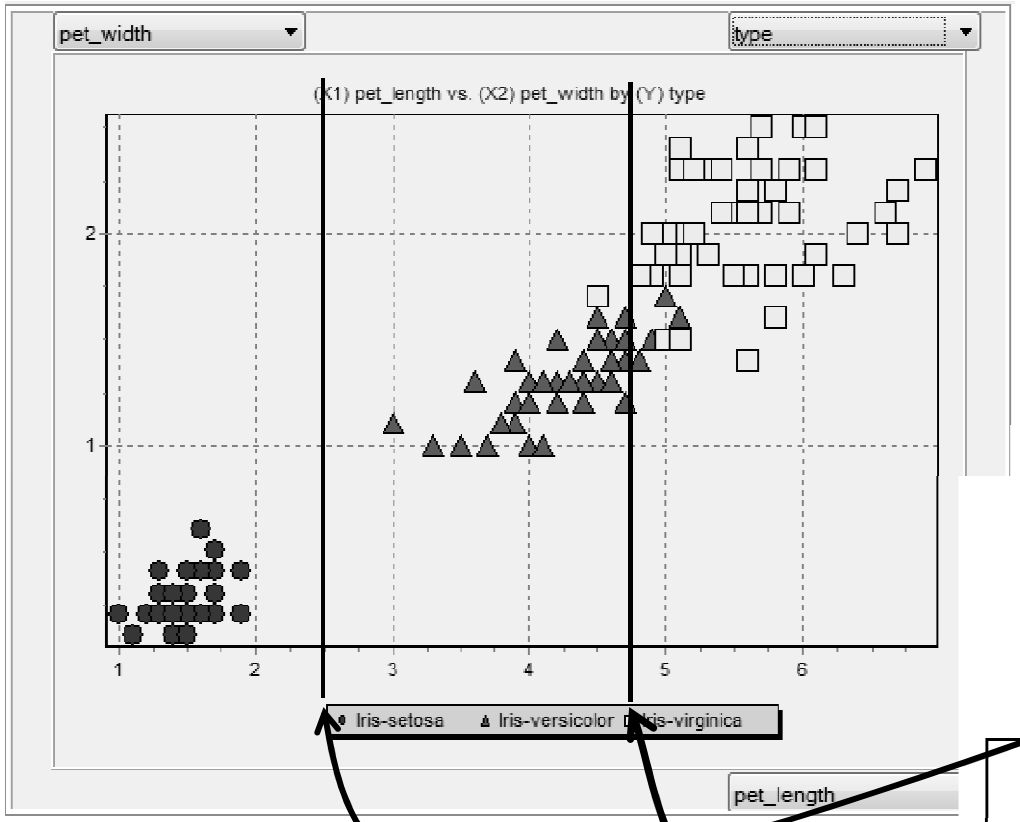
Utilisation des algorithmes supervisés (MDLPC, Fayyad et Irani, 1993 ; Kerber, 1992)

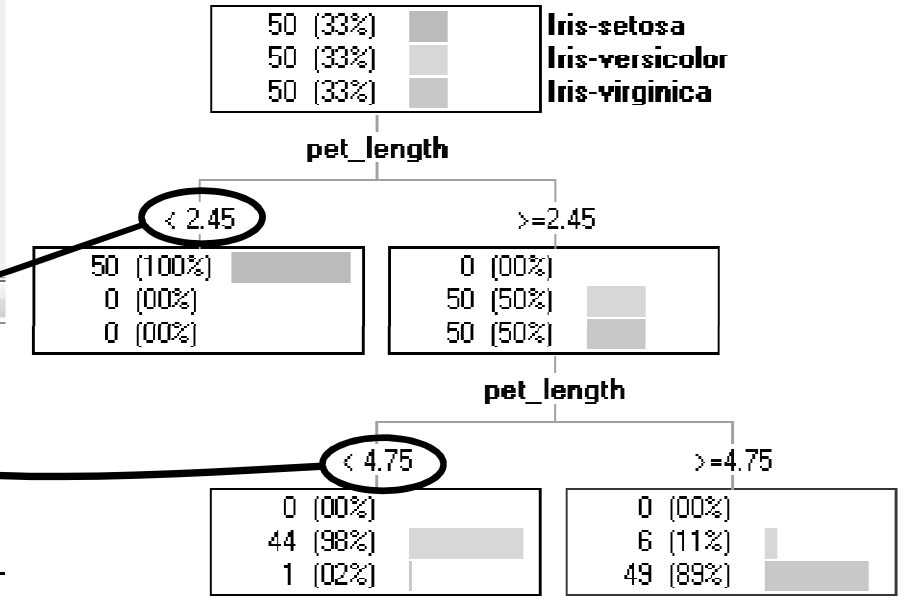
Pourquoi ?

- Création d'intervalles où une des modalités de Y est surreprésentée
- Détection automatique du bon nombre d'intervalles

Découpage en intervalles des variables quantitatives
Une solution simple : les arbres de décision

Utiliser la variable à découper comme seule variable prédictive dans l'arbre





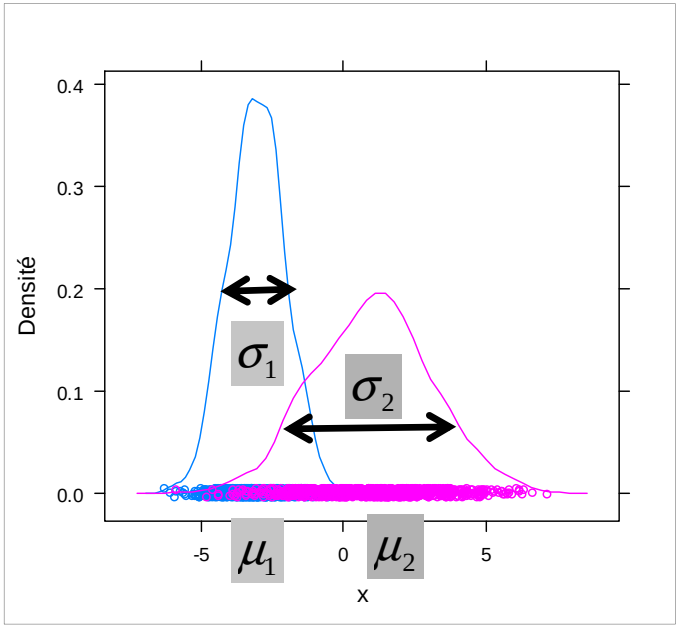
Cas des variables prédictives quantitatives (2)

Démarche paramétrique
Hypothèses de distribution pour les probas conditionnelles
(pour simplifier les calculs…)

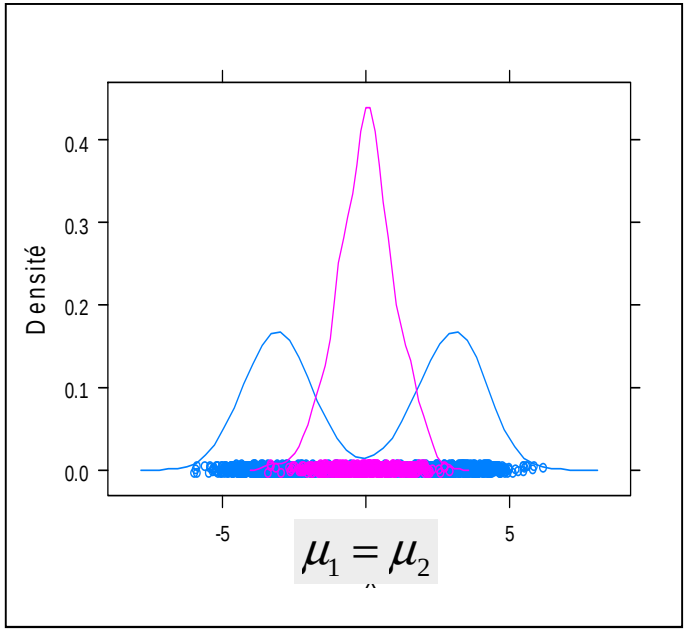
HYP.1 - Distributions conditionnelles gaussiennes

$$P[X_j / Y = y_k] = f_k(X_j) = \frac{1}{\sigma_{k,j}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_j - \mu_{k,j}}{\sigma_{k,j}}\right)^2}$$

Distribution d'un des prédicteurs conditionnellement à yk



Compatible avec l'hypothèse de distribution gaussienne



Incompatible avec l'hypothèse de distribution gaussienne → seule solution possible : discrétiser

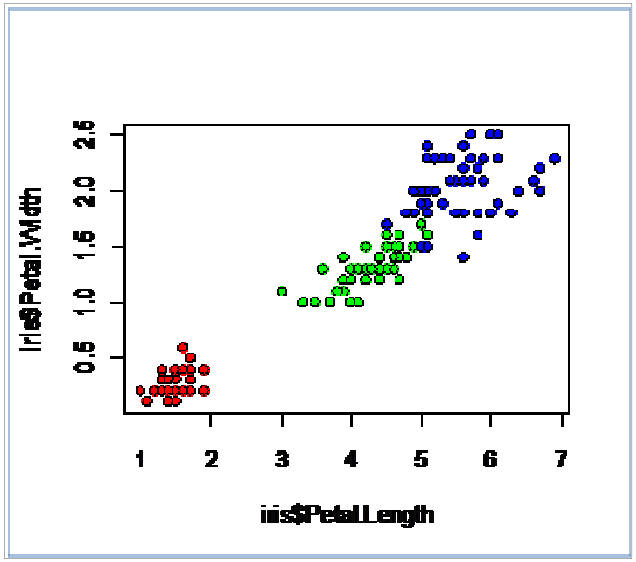
Remarque : C'est **un cas particulier de l'analyse discriminante** où l'on considère que les éléments hors diagonale de la matrice de variance covariance sont tous nuls ! (cf. Cours d'Analyse discriminante prédictive).

Modèle quadratique

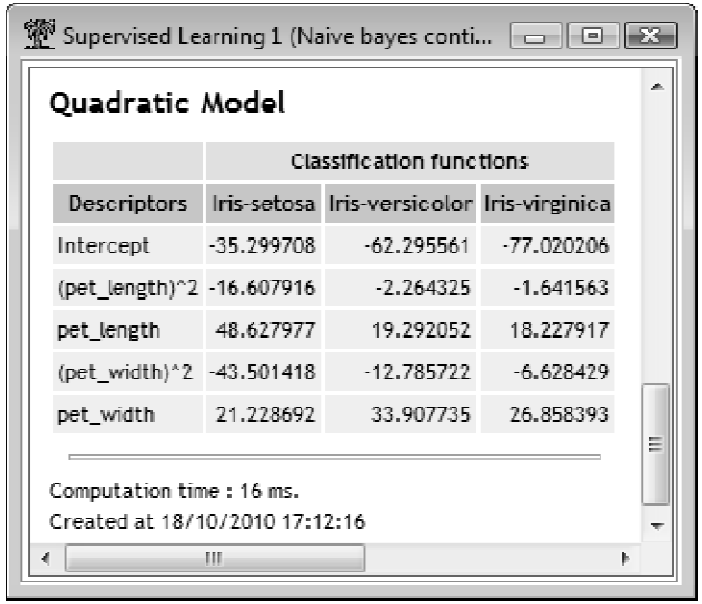
$$d(y_k, \aleph) \propto \ln p_k + \sum_j \left\{ -\frac{1}{2\times\sigma_{k,j}^2} x_j^2 + \frac{\mu_{k,j}}{\sigma_{k,j}^2} x_j - \left(\frac{\mu_{k,j}^2}{2\times\sigma_{k,j}^2} + \ln(\sigma_{k,j}) \right) \right\}$$

$$\propto \ln p_k + \sum_j a_{k,j} x_j^2 + b_{k,j} x_j + c_{k,j}$$

Règle d'affectation inchangée c.-à-d. $\hat{y}(\omega) = y_{k^*} \Leftrightarrow y_{k^*} = \arg\max_k d[y_k, \aleph(\omega)]$



Fichier IRIS (2 variables)



Interprétation très difficile !!!

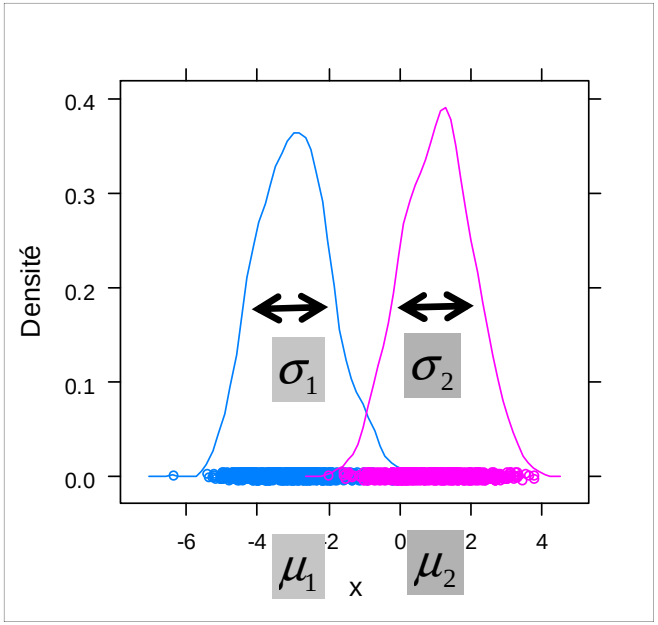
HYP.2 - Homoscédasticité

Les dispersions conditionnelles sont identiques

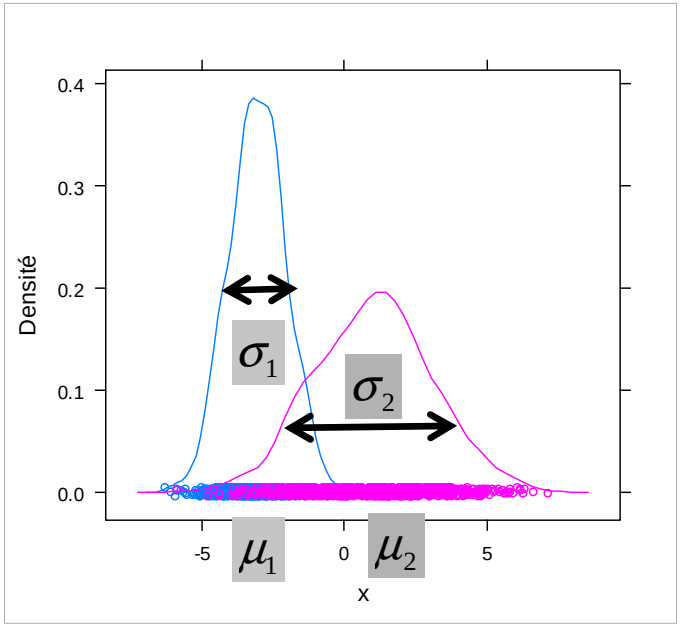
$$\sigma_{k,j} = \sigma_j \ , \forall k$$

Estimation à l'aide de l'écart-type intra-classes !!!

$$P[X_j / Y = y_k] = f_k(X_j) = \frac{1}{\sigma_j \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_j - \mu_{k,j}}{\sigma_j}\right)^2}$$



Compatible avec l'hypothèse d'homoscédasticité



Non compatible avec l'hypothèse d'homoscédasticité

➡ Mais la technique est robuste !!!

Homoscédasticité - Conséquences

Modèle linéaire

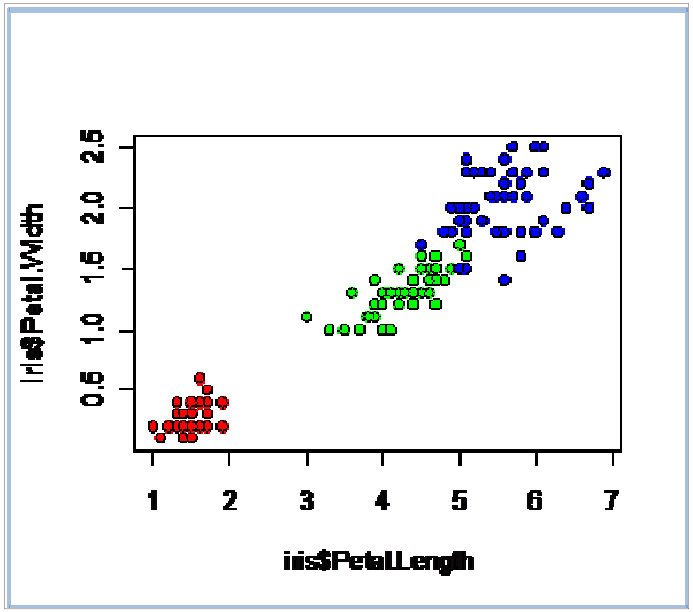
$$d(y_k, \aleph) \propto \ln p_k + \sum_j \left\{ \frac{\mu_{k,j}}{\sigma_j^2} x_j - \frac{\mu_{k,j}^2}{2 \times \sigma_j^2} \right\}$$

$$\propto a_{k,0} + a_{k,1} x_1 + a_{k,2} x_2 + \cdots + a_{k,J} x_J$$

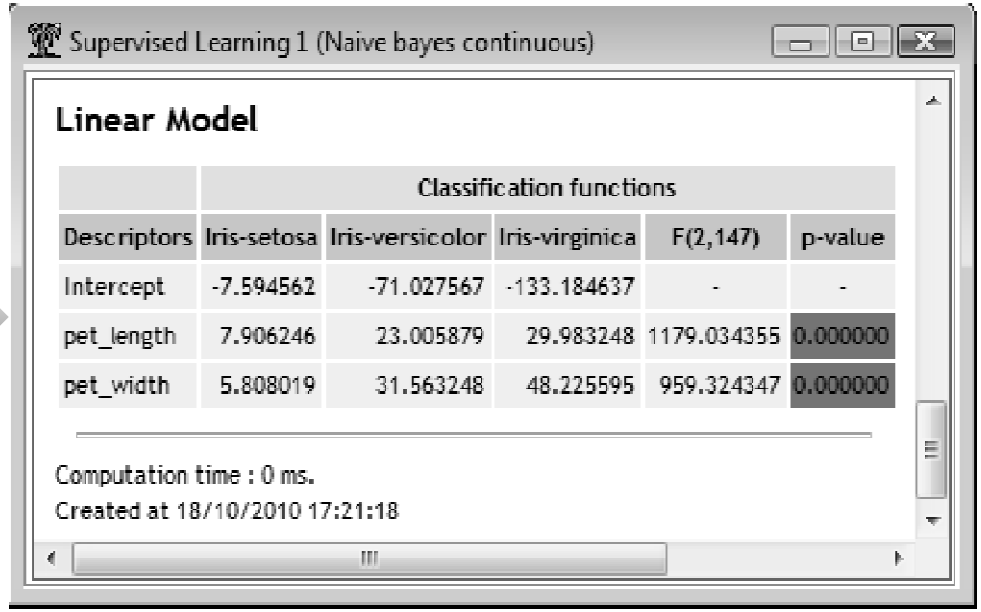
Règle d'affectation inchangée c.-à-d.

$$\hat{y}(\omega) = y_{k^*} \Leftrightarrow y_{k^*} = \arg\max_k d[y_k, \aleph(\omega)]$$

Et si K=2 (Y : + vs. -), on peut construire une fonction SCORE -- D(X)



Fichier IRIS (2 variables)



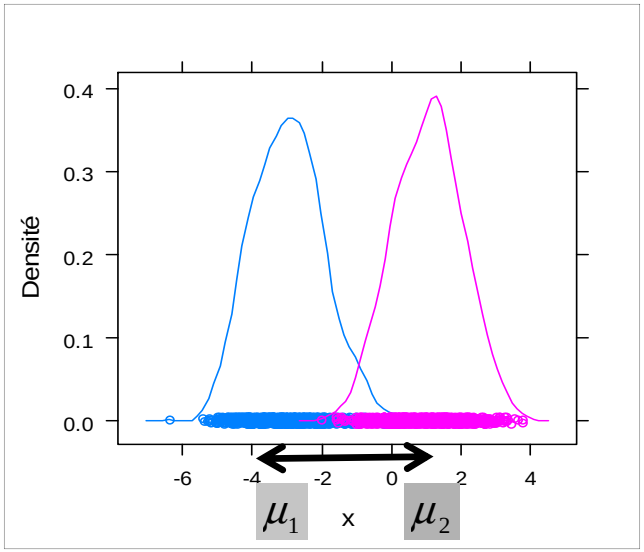
Interprétation facilitée

Ex. PET.LENGTH faible -> Setosa
PET.LENGTH moyen -> Versicolor
PET.LENGTH élevé -> Virginica

Evaluation des variables

Evaluer la pertinence d'une variable
Eliminer les variables non-pertinentes
Supprimer les redondances…

Contribution d'une variable – Test ANOVA



Comparaison des moyennes conditionnelles

$$H_0 : \mu_{k,j} = \mu,\ \forall k$$

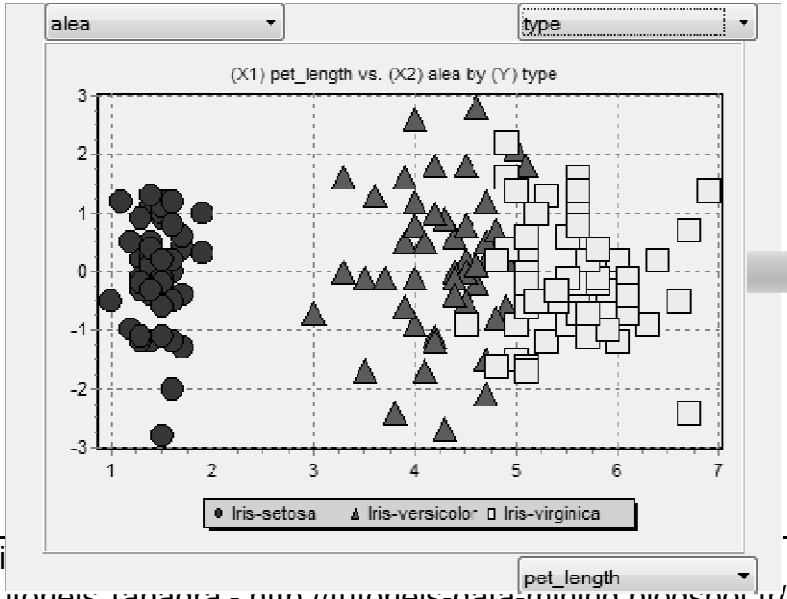
Statistique de test F

$$F = \frac{\frac{\sum_k n_k (\hat{\mu}_k - \hat{\mu})^2}{K-1}}{\frac{\sum_k (n_k - 1)\hat{\sigma}_k^2}{n-K}}$$

Variance Inter

Variance Intra

Sous H0, F ~ Fisher (K-1, n-K) degrés de liberté



Linear Model

	Classification functions				
Descriptors	Iris-setosa	Iris-versicolor	Iris-virginica	F(2,147)	p-value
Intercept	-6.886948	-50.112224	-84.338242	-	-
alea	-0.041928	0.142192	-0.105733	0.909119	0.405132
pet_length	7.906246	23.005879	29.983248	1179.034355	0.000000

Sélection de variables - RANKING

RANKING :
1. Calculer F pour toutes les variables
2. Les trier par ordre décroissant
3. Ne conserver que les variables significativement liées avec Y c.-à-d. test ANOVA = écart significatif entre les moyennes cond.

IRIS + 1 ALEA

Keeped into INPUT selection

	Attributes
1	pet_length
2	pet_width
3	sep_length
4	sep_width

Calculations details

N°	Attribute	F	F (max normalized)	p-value (2,147)
1	pet_length	1179.03		0.000000
2	pet_width	959.32		0.000000
3	sep_length	119.26		0.000000
4	sep_width	47.36		0.000000
5	alea	0.91		0.405132

Mêmes difficultés qu'avec les prédicteurs qualitatifs

→ Fixer le seuil « alpha » ?
→ Gestion de la redondance ?

Pourquoi pas un critère de type MERITE utilisé dans CFS ?

$$merit = \frac{p \times \overline{s}_{Y,X}}{\sqrt{p + p \times (p+1) \times \overline{s}_{X,X}}}$$

- Indice 1 : Mesure le lien entre Y (quali.) et X (quanti.)
- Indice 2 : Mesure le lien entre Xj (quanti) et Xj' (quanti)

→ Indice 1 et Indice 2 ne sont pas comparables !

Quelles pistes ?

➔ Indicateur de type LAMBDA de Wilks utilisé en Analyse discriminante. Généralisation multidimensionnelle de la comparaison de moyenne.
Mais… calculs matriciels lourds, on perd l'avantage de rapidité du Naive Bayes.

➔ Utiliser des techniques de sélection intégrées à d'autres méthodes (ex. sélection des variables par un arbre de décision).
Mais… les variables pertinentes pour un type de classifieur ne le sont pas forcément pour les autres

➔ Passer par la discrétisation et utiliser les techniques pour var. expl. qualitatives

Conclusion

+

>> Très connu et très utilisé des chercheurs (text mining, etc.)

>> Cf. les avantages déjà cités précédemment (surtout Incrémentalité et Traitement des très grandes bases)

>> Possibilité de produire un modèle explicite !!! (Totalement méconnu)

-

>> Méconnu des praticiens (applications marketing)... *parce qu'on ne sait pas qu'on peut en dériver un modèle explicite...*

Bibliographie

Tutoriels Tanagra – « Le bayesien naïf revisité »

http://tutoriels-data-mining.blogspot.com/2010/03/le-classifieur-bayesien-naif-revisite.html

Tutoriels Tanagra – « Bayesien naïf pour prédicteurs continus »

http://tutoriels-data-mining.blogspot.com/2010/10/bayesien-naif-pour-predicteurs-continus.html

Wikipedia, « Naive Bayes Classifier »

http://en.wikipedia.org/wiki/Naive_Bayes_classifier

STATSOFT E-BOOKS, « Naive Bayes Classifier » (cf. autres hypothèses de distributions)

http://www.statsoft.com/textbook/naive-bayes-classifier/