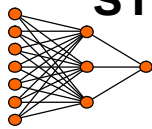


Stéphane Tufféry
Statisticien - Data Miner - Enseignant

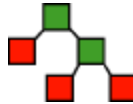


DATA MINING - SCORING

STATISTIQUE DÉCISIONNELLE



APPLICATION AU CRM



19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

1

Plan du cours

- Qu'est-ce que le data mining ?
- A quoi sert le data mining ?
- Les 2 grandes familles de techniques
- Le déroulement d'un projet de data mining
- Coûts et gains du data mining
- Facteurs de succès - Erreurs à éviter
- Informatique décisionnelle et de gestion
- La préparation des données
- Techniques descriptives de data mining
- **Techniques prédictives de data mining**
- Logiciels et consultants
- CNIL et limites légales du data mining
- Le text mining
- Le web mining

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

2



Techniques prédictives

Points forts et points faibles

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



3



Techniques prédictives de data mining :

Classement - prédiction

(généralités)

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



4

Classement & classification

- Le **classement** consiste à placer chaque individu de la population dans une *classe*, parmi plusieurs classes prédéfinies, en fonction des caractéristiques de l'individu indiquées comme variables explicatives
- Le plus souvent, il y a 2 classes prédéfinies (« sains » et « malades », par exemple), et le résultat du classement est un ensemble de règles permettant d'affecter chaque individu à l'une ou l'autre des classes
- La **classification** consiste à regrouper les individus d'une population en un nombre limité de *classes* qui :
 - ne sont pas prédéfinies mais déterminées au cours de l'opération, contrairement aux classes du classement
 - regroupent les individus ayant des caractéristiques similaires et séparent les individus ayant des caractéristiques différentes

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



5

Prédiction

- La **prédiction** consiste à estimer
 - la valeur d'une variable continue (dite « à expliquer », « cible », « réponse », « dépendante » ou « endogène »)
 - en fonction de la valeur d'un certain nombre d'autres variables (dites « explicatives », « de contrôle », « indépendantes » ou « exogènes »)
- Cette variable « cible » est par exemple :
 - le poids (en fonction de la taille)
 - la taille des ailes d'une espèce d'oiseau (en fonction de l'âge)
 - le prix d'un appartement (en fonction de sa superficie, de l'étage et du quartier)
 - la consommation d'électricité (en fonction de la température extérieure et de l'épaisseur de l'isolation)

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



6

Classement et prédiction

- **Classement** : la variable à expliquer est *catégorielle*
 - on parle aussi de **classification** (dans l'école anglo-saxonne) ou de **discrimination**
- **Prédiction** : la variable à expliquer est *continue*
 - on parle aussi de **régression**
 - ou d'**apprentissage supervisé** (réseaux de neurones)
- **Scoring** : classement appliqué à une problématique d'entreprise

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

7



Tableau synoptique

explicatives à expliquer	1 quantitative (covariable)	n quantitatives (covariables)	1 qualitative (facteur)	n qualitatives (facteurs)	mélange
1 quantitative	rég. linéaire simple	rég. linéaire multiple, rég. robuste, PLS, arbres, réseaux de neurones	ANOVA	ANOVA, arbres de décision, réseaux de neurones	ANCOVA, arbres de décision, réseaux de neurones
n quantitatives (représentent des quantités ¹)	régression PLS2	régression PLS2, réseaux de neurones	MANOVA	MANOVA, réseaux de neurones	MANCOVA, réseaux de neurones
1 qualitative nominale	ADL, régression logistique, reg. logistique PLS	ADL, régress. logistique, reg. logistique PLS, arbres, réseaux de neurones, SVM	régression logistique	régression logistique, DISQUAL, arbres, réseaux de neurones	régression logistique, arbres, réseaux de neurones
1 qualitative ordinaire (au moins 3 gp)	régression logistique ordinaire	régression logistique ordinaire	régression logistique ordinaire	régression logistique ordinaire	régression logistique ordinaire
n quantitatives (représentent des mesures répétées d'une même quantité)				modèle à mesures répétées	modèle à mesures répétées

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

8



Techniques inductives et transductives

- Dans les **techniques inductives** :
 - une phase d'apprentissage (phase inductive) pour élaborer un modèle, qui résume les relations entre les variables
 - et qui peut ensuite être appliqué à de nouvelles données pour en déduire un classement ou une prédiction (phase déductive)
- Les **techniques transductives**
 - ne comprennent qu'une seule étape (éventuellement répétée), au cours de laquelle chaque individu est directement classé (ou objet d'une prédiction) par référence aux autres individus déjà classés
 - il n'y a pas élaboration d'un modèle

19/08/2005

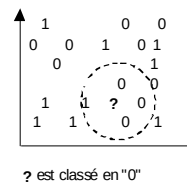
© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



9

k-plus proches voisins

- La plus connue des *techniques transductives*
- Le classement (prédiction) de chaque individu s'opère en regardant, parmi les individus déjà classés, la classe des k individus qui sont les plus proches voisins (ou en calculant la moyenne dans le voisinage de la variable à prédire)
- La valeur de k sera choisie en sorte d'obtenir le meilleur classement (prédiction) possible :
 - ce choix est la principale difficulté de cet algorithme !
- Ainsi, dans l'exemple ci-contre, l'individu « ? » est classé en « 0 », car entouré en majorité de « 0 »



19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



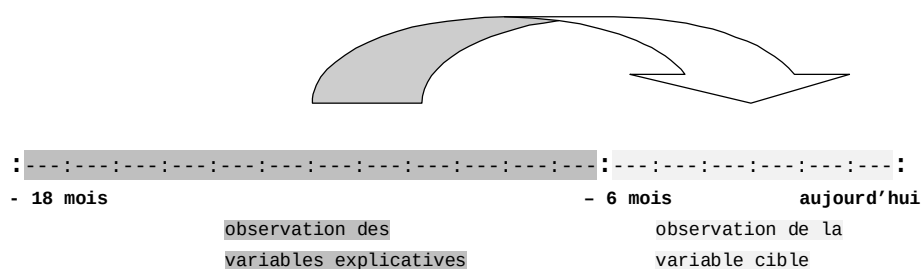
10

Limites des méthodes transductives

- Une technique inductive *résume dans un modèle l'information* contenue dans les données
 - ce qui permet d '*appliquer rapidement* ce modèle à de nouvelles données
- Une technique transductive *manipule l'ensemble des individus* déjà classés, pour tout nouveau classement
 - ce qui nécessite donc une *grande puissance de stockage et de calcul*
- **On utilise surtout les techniques inductives.**
- Une méthode transductive, comme les k-NN, peut être utilisée dans une étape préalable de détection et de mise à l'écart des individus hors norme, des « outliers ».



Méthodes inductives : schéma



Le **modèle** sera par exemple une fonction f telle que :

$$\text{Probabilité}(\text{variable cible} = x) = f(\text{variables explicatives})$$



Méthodes inductives : 4 étapes

- Apprentissage : **construction du modèle** sur un 1^{er} échantillon pour lequel on connaît la valeur de la variable cible
- Test : **vérification du modèle** sur un 2^d échantillon pour lequel on connaît la valeur de la variable cible, que l'on compare à la valeur prédite par le modèle
 - si le résultat du test est insuffisant (d'après la *matrice de confusion*), on recommence l'apprentissage
- **Validation du modèle** sur un 3^e échantillon, pour avoir une idée du taux d'erreur non biaisé du modèle
- **Application du modèle** à l'ensemble de la population à scorer, pour déterminer la valeur de la variable cible de chaque individu.

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



13

Matrice de confusion

valeur prédite $\hat{e}$	A	B	TOTAL
valeur réelle e			
A	1800	200	
B	300	1700	
TOTAL			4000

- Taux d'erreur = $(200 + 300) / 4000 = 12,5 \%$

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



14

Q de Press

- Pour vérifier que le % d'individus correctement classés est significativement meilleur que par un classement aléatoire : le « Q de Press »

$$Q_{\text{Press}} = \frac{(n - (c \times k))^2}{n \times (k - 1)}$$

- n = taille échantillon
- k = nb de groupes
- c = nb d'individus bien classés
- Suit un chi-2 à 1 degré de liberté
 - valeur critique : 10,8 à 0,1 % - 6,63 à 1 % - 3,84 à 5 %
- Ici on a : Q de Press = $(4000 - 7000)^2 / 4000 = 2250$

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

15

Qualités attendues d'une technique 1/2

- La précision
 - le *taux d'erreur* doit être le plus bas possible, et l'indice de Gini le plus proche possible de 1
- La robustesse
 - être le moins sensible possible aux fluctuations aléatoires de certaines variables et aux valeurs manquantes
 - ne pas dépendre de l'échantillon d'apprentissage utilisé
 - être aussi performant quand on l'applique à de nouveaux individus
- La concision
 - les règles du modèle doivent être les plus simples et les moins nombreuses possible

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

16

Qualités attendues d'une technique

2/2

- Des résultats explicites
 - les règles du modèle doivent être accessibles et compréhensibles
- La diversité des types de données manipulées
 - tous les algorithmes ne sont pas aptes à manipuler les données catégorielles, discrètes et continues
- Les possibilités de paramétrage
 - dans un classement, il est parfois intéressant de pouvoir pondérer les erreurs de classement, pour signifier, par exemple, qu'il est plus grave de classer un patient malade en « non-malade » que l'inverse
- La rapidité de calcul du modèle
 - c'est l'*apprentissage* (non son *application*) du modèle qui peut être trop longue

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

17

Représentation graphique des domaines de classement

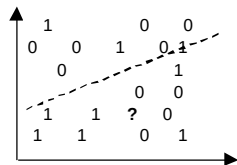
- Toutes les techniques inductives de classement découpent l'espace des variables en régions, dont chacune est associée à une des classes parce qu'elle contient en majorité des individus de cette classe
- L'ensemble des règles de découpage et d'association constitue le *modèle*
- Tout nouvel individu qui tombe dans une de ces régions est classé en conséquence
- La forme de ces régions dépend de la technique employée

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

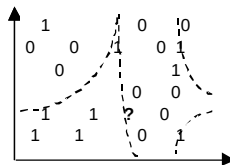
18

Représentation graphique des domaines de classement



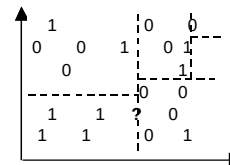
? est classé en "1"

Analyse discriminante



? est classé en "0"

Réseau de neurones



? est classé en "0"

Arbre de décision

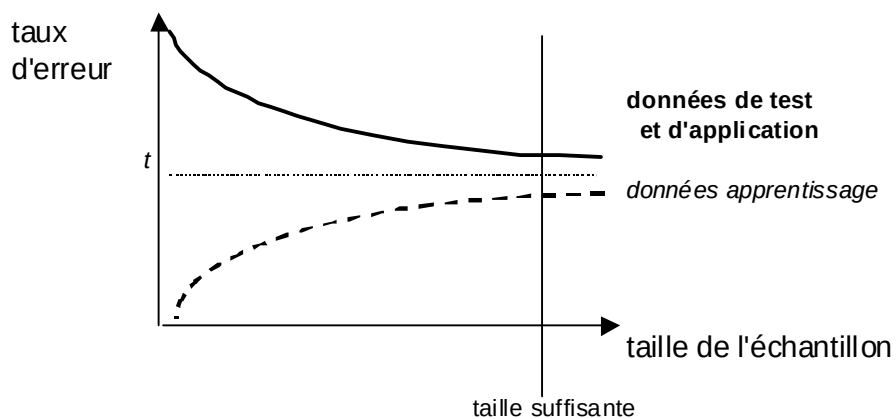
19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



19

Courbes du taux d'erreur en apprentissage et en test



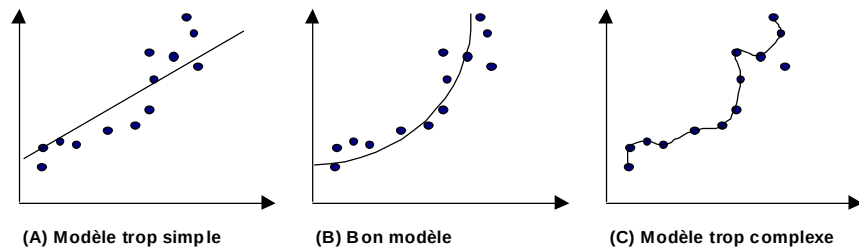
19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



20

Sur-apprentissage en régression



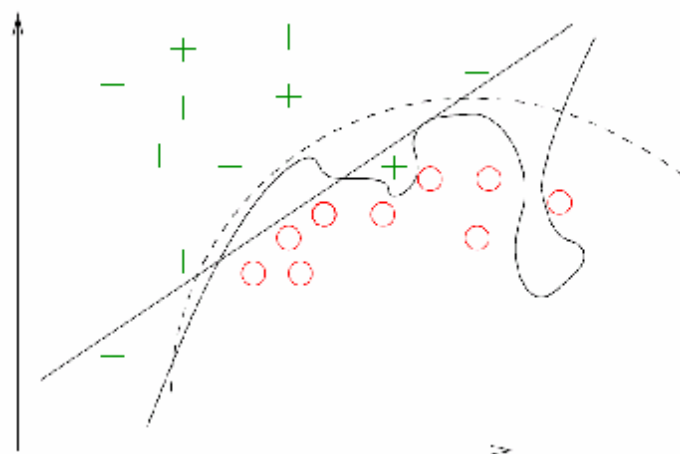
- Modèle trop poussé dans la phase d'apprentissage :
 - il épouse toutes les fluctuations de l'échantillon d'apprentissage,
 - détecte ainsi de fausses corrélations,
 - et les extrapole à tort dans les phases de test et d'application.
- Le sur-apprentissage peut aussi venir d'une confusion des 2 périodes d'observation.

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

21

Sur-apprentissage en classement



Source : Olivier Bousquet

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

22

Méthodes de sélection

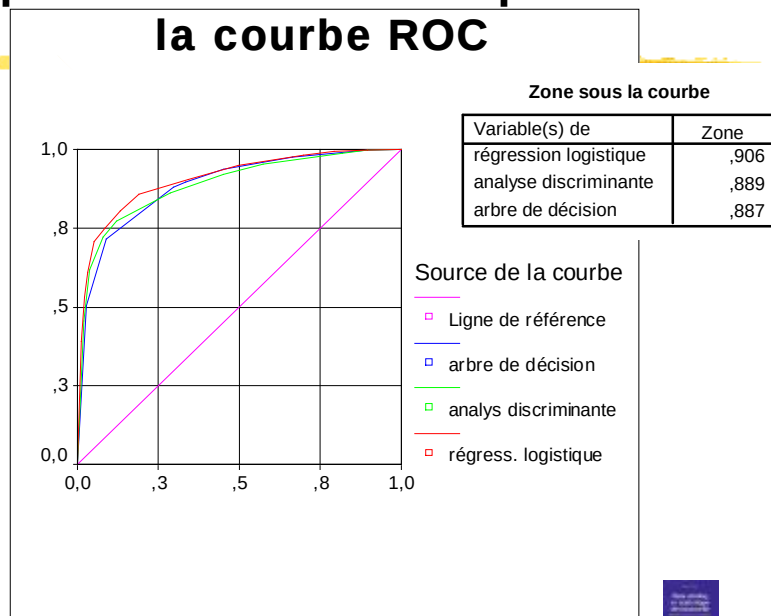
- Pas à pas : Ascendante (« forward »)
 - aucune variable au départ : on ajoute 1 à 1 celles qui augmentent le plus le R^2
- Pas à pas : Descendante (« backward »)
 - toutes les variables au départ : on rejette 1 à 1 celles qui sont insuffisamment corrélées à la cible
- Pas à pas : Mixte (« stepwise »)
 - aucune variable au départ : on ajoute 1 à 1 celles qui sont suffisamment corrélées à la cible, mais on peut retrancher une variable à chaque étape, si son pouvoir discriminant est contenu dans une combinaison des nouvelles variables
- Globale : Algorithme de Furnival et Wilson (si 2 groupes)
 - cherche à ajuster le R^2 en comparant une partie de tous les modèles possibles (élimine les moins intéressants *a priori*)

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

23

Comparaison de modèles prédictifs : la courbe ROC



19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

24

Sensibilité et spécificité

- Pour un score devant discriminer un groupe A (les positifs ; ex : les risqués) par rapport à un autre groupe B (les négatifs ; ex : les non risqués), on définit 2 fonctions du seuil de séparation s du score :
 - sensibilité = $\alpha(s) = \text{Proba}(\text{score} > s / A) = \text{probabilité de bien détecter un positif (ex : proportion de « vrais risqués »)}$
 - spécificité = $\beta(s) = \text{Proba}(\text{score} < s / B) = \text{probabilité de bien détecter un négatif (ex : proportion de « vrais non-risqués »)}$
- Pour un modèle, on cherche s qui maximise $\alpha(s)$ tout en minimisant les faux positifs $1 - \beta(s) = \text{Proba}(\text{score} > s / B)$
 - f.p. : négatifs considérés comme positifs à cause du score
- Le meilleur modèle : permet de capturer le plus possible de vrais positifs avec le moins possible de faux positifs

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



25

Courbe ROC

- La courbe ROC (« receiver operating curve » terme issu du traitement du signal)
 - sur l'axe Y : on a la sensibilité = $\alpha(s)$
 - sur l'axe X : on a $1 - \text{spécificité} = 1 - \beta(s)$
 - proportion y de vrais positifs en fonction de la proportion x de faux positifs, lorsque l'on fait varier le seuil s du score
- Exemple : la courbe ROC de la régression logistique précédente passe par le point (0,3;0,9). Ce point correspond à un seuil s qui est tel que : si on considère « risqués » tous les individus dont le score $> s$, on a détecté :
 - 30% de faux risqués (30% des non-risqués ont un score $> s$)
 - 90 % de vrais risqués (90 % des risqués ont un score $> s$)
 - NB : 0,3 ne correspond pas à 30 % de la population totale !

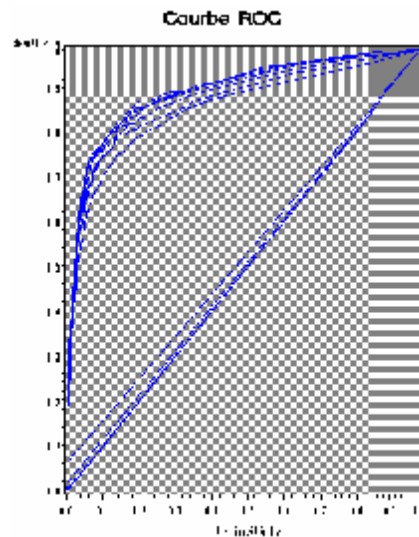
19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



26

Courbes ROC avec entrée progressive des variables du modèle



19/08/2005

©

data.mining.free.fr

27

Aire sous la courbe ROC

- Aire théorique A sous la courbe ROC = probabilité que $\text{score}(x) > \text{score}(y)$, si x est tiré au hasard dans le groupe A (à prédire) et y dans le groupe B
- 1^{ère} méthode d'estimation : par la méthode des trapèzes
- 2^e méthode d'estimation : par les paires concordantes
 - soit n_1 (resp. n_2) le nb d'observations dans A (resp. B)
 - on s'intéresse aux $n_1 n_2$ paires formées d'un x dans A et d'un y dans B
 - parmi ces t paires : on a concordance si $\text{score}(x) > \text{score}(y)$; discordance si $\text{score}(x) < \text{score}(y)$
 - soient nc = nb de paires concordantes ; nd = nb de paires discordantes ; $n_1 n_2 - nc - nd$ = nb d'ex-æquo
 - aire sous la courbe ROC $\approx (nc + 0,5[t - nc - nd]) / n_1 n_2$

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

28

Aire sous la courbe et indice de Gini

- Le modèle est d'autant meilleur que A est plus proche de 1
- Si l'aire $A = 0,5$: modèle pas meilleur qu'une notation aléatoire. Il existe un intervalle de confiance sur A et un test associé :

Aire sous la courbe

Variable(s) de résultats tests	Zone	Erreur Std. ^a	Signif. asymptotique ^b	Intervalle de confiance 95% asymptotique	
				Borne inférieure	Borne supérieure
arbre de décision	,887	,008	,000000	,872	,902
régression logistique	,906	,007	,000000	,892	,921
analyse discriminante	,889	,008	,000000	,873	,904

a. Dans l'hypothèse non-paramétrique

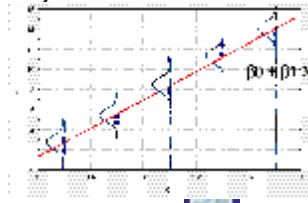
b. Hypothèse nulle /: zone vraie = 0.5

- Indice de Gini associé à la courbe ROC = $2.(A - 0,5)$

Technique de prédiction : La régression linéaire

Hypothèses du modèle linéaire

- Dans le modèle simple : X et Y deux variables continues
- Il n'y a pas d'erreur de mesure sur X
- Il existe une relation linéaire entre X et l'espérance de Y :
 - > on peut écrire $y_i = \alpha + \beta x_i + \varepsilon_i$, pour tout i entre 1 et n
 - n = nb d'observations et ε_i = « résidu » de l'observation i
- La moyenne des résidus $E(\varepsilon_i) = 0$ en tout point X_0
- La variance des résidus est la même pour toutes les valeurs de X (homoscédasticité) : $V(\varepsilon_i) = s^2$
- Les résidus ε_i sont indépendants ($\text{cov}(\varepsilon_i, \varepsilon_j) = 0 \forall i \neq j$) et normalement distribués ($\varepsilon_i \sim N(0, s^2)$)
- **Régression linéaire multiple :**
 - $Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \varepsilon$
 - (avec indépendance des X_i)



19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

31

Hypothèses du modèle linéaire

- On a équivalence de $\varepsilon_i \sim N(0, s^2)$ et $Y/X=x_i \sim N(\alpha + \beta x_i, s^2)$
- Cette hypothèse de normalité classe la régression linéaire dans la famille des modèles linéaires généraux (GLM)
- Dans les modèles linéaires généralisés (voir + loin), la loi de $Y/X=x_i$ n'est plus nécessairement normale
- Les fluctuations de ε_i $\hat{O}$ $Y/X=x_i$ correspondent au fait que :
 - des individus avec même valeur x_i peuvent avoir des réponses Y différentes (variation synchronique)
 - OU un même individu mesuré à plusieurs reprises avec la même valeur x_i peut avoir des réponses Y différentes (variation diachronique)



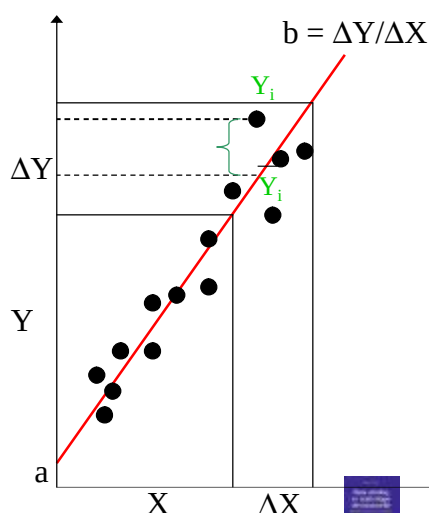
19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

32

Méthode des moindres carrés ordinaires (MCO)

- On recherche des estimateurs a et b de α et β qui minimisent les résidus $\varepsilon_i^2 = (Y_i - \bar{Y}_i)^2$, où $\bar{Y}_i$ est prédit par la droite $Y = a + bX$
- L'estimateur b de la pente vaut :
- $b = \text{cov}(X, Y) / (\sigma_X)^2$
- L'estimateur a de la constante vaut :
- $a = Y_{\text{moy}} - b \cdot X_{\text{moy}}$
- $Y = a + bX$ ajuste le nuage de points



19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

33

Propriétés des estimateurs MCO

- MCO est optimale car les estimateurs des coefficients sont :
 - sans biais : $E(a) = \alpha$ et $E(b) = \beta$
 - de variance minimale
 - constante : $\sigma_a^2 = s^2 [1/n + X_{\text{moy}}^2 / \sum (X_i - X_{\text{moy}})^2]$
 - avec : $s^2 = \text{variance résiduelle définie plus loin}$
 - > IC au niveau $100(1-\alpha)\%$: $a \pm t_{\alpha/2, n-p-1} \cdot \sigma_a$
 - pente : $\sigma_b^2 = s^2 [1 / \sum (X_i - X_{\text{moy}})^2]$
 - > IC au niveau $100(1-\alpha)\%$: $b \pm t_{\alpha/2, n-p-1} \cdot \sigma_b$
 - parmi tous les estimateurs linéaires, et même parmi tous les estimateurs en ajoutant l'hypothèse de normalité $\varepsilon_i \sim N(0, s^2)$
- Pour diminuer les variances :
 - diminuer la variance résiduelle s^2 de l'échantillon
 - augmenter la taille n de l'échantillon
 - augmenter l'étendue des valeurs observées de X

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

34

Que signifie la variance des estimateurs ?

- On n'atteint jamais les véritables coefficients α et β car :
 - le modèle linéaire n'est le plus souvent qu'une approximation de la réalité
 - on ne travaille que sur des échantillons et non la population entière
 - on commet des erreurs de mesure
- Des modèles sur des échantillons différents donneront des estimations a_m et b_m différents
- D'où une variance des estimateurs a et b
- Les estimateurs fournis par la MCO sont « BLUE » : best linear unbiased estimators
 - mais : on accepte parfois (régression ridge) des estimateurs légèrement biaisés si cela permet de diminuer notablement leur variance

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

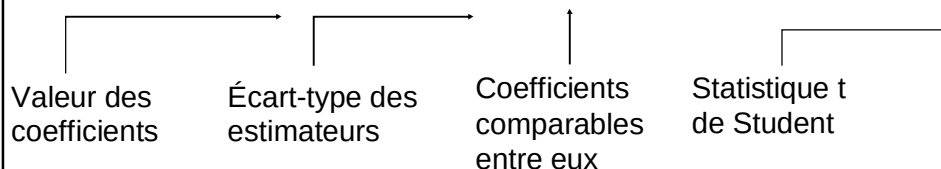
35

Coefficients de régression et tests

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Signification
		B	Erreur standard	Bêta		
1	(constante)	1467,643	62,422		23,512	,000
	TEMPERAT	-37,060	2,295	-,866	-16,147	,000
	ISOLATIO	-29,774	3,492	-,457	-8,526	,000

a. Variable dépendante : CONSOMMA



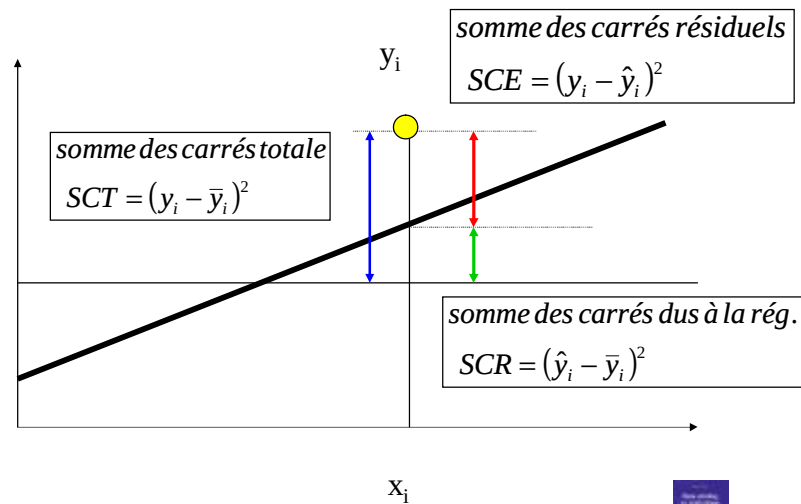
Une valeur $t > 2$ ou $t < -2$ est significative à 95 % d'un coeff $\neq 0$

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

36

Sommes des carrés



19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

37

Test global du modèle

ANOVA^a

Modèle		Somme des carrés	ddl	Carré moyen	F	Signification
1	Régression	3267046,7	2	1633523,333	167,933	,000 ^a
	Résidu	116727,068	12	9727,256		
	Total	3383773,7	14			

a. Valeurs prédites : (constantes), ISOLATIO, TEMPERAT

b. Variable dépendante : CONSOMMA

2 prédicteurs => régression linéaire multiple

$$F = \frac{\frac{SCR}{p}}{\frac{SCE}{n-p-1}}$$

SCR = somme des carrés « Régression »

SCE = somme des carrés « Erreurs »

p = nombre de variables

n = nombre d'observations

suit une loi F de ddl (p,n-p-1)

sous l'hypothèse nulle (H_0) :
($b_1 = b_2 = 0$)

$$R^2 = SCR / SCT = 1 - (SCE / SCT)$$

variance s^2 du terme d'erreur = 98,627²

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

38

Coefficient de détermination

- $R^2 = SCR / SCT$
- R^2 = proportion de variation de la variable cible expliquée par tous les prédicteurs (syn : régresseurs)
- Bon ajustement si R^2 proche de 1
- R^2 est biaisé (optimiste car croissant avec le nb de variables) et on lui substitue le R^2 ajusté :

$$R^2_{ajusté} = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1}$$

- R^2 ajusté est toujours $< R^2$ et peut être < 0

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,983 ^a	,966	,960	98,627

19/08

a.

39

Intervalles de confiance

- $\hat{Y}_0 = a + bX_0$ est une prévision de Y et de la moyenne $E(Y)$ en tout point X_0 de l'intervalle de mesure (car $E(\varepsilon_i) = 0$)
- D'après les formules sur les variances des estimateurs, les IC à $(100-\alpha) \%$ de $E(Y)$ et Y au point X_0 sont :
 - $\hat{Y}_0 \pm t_{\alpha/2, n-p-1} \cdot s \cdot [1/n + (X_0 - X_{moy})^2 / \sum (X_i - X_{moy})^2]^{1/2}$ pour $E(Y)$
 - $\hat{Y}_0 \pm t_{\alpha/2, n-p-1} \cdot s \cdot [1 + 1/n + (X_0 - X_{moy})^2 / \sum (X_i - X_{moy})^2]^{1/2}$ pour Y (on a ajouté la variance du terme d'erreur)
- Autrement dit, la variance de la valeur prédite pour une observation est :
 - $s^2 [1 + 1/n + (X_0 - X_{moy})^2 / \sum (X_i - X_{moy})^2]$
- > Plus difficile d'estimer une valeur possible de Y sachant X_0 que la moyenne des valeurs possibles sachant X_0
- > L'IC augmente quand X_0 s'éloigne de X_{moy}

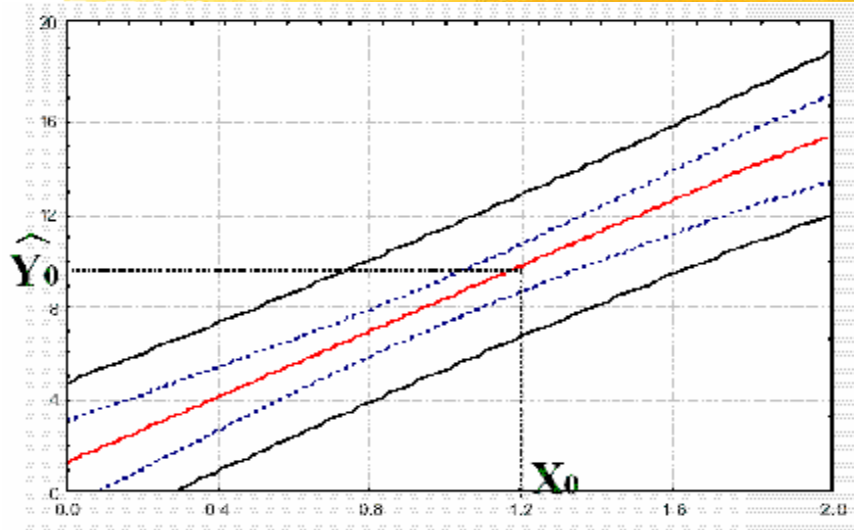
19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



40

IC de la moyenne et des observations



19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

41

Précautions d'utilisation

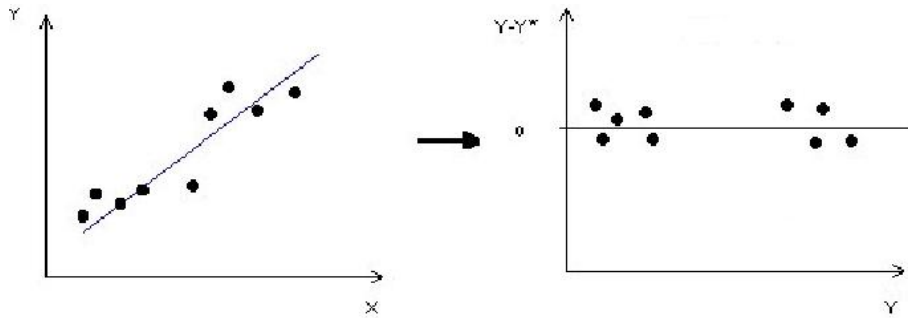
- Le modèle n'est valide que sur l'étendue des observations, et surtout près de la moyenne de X
- Un petit échantillon (<20) ne détecte que les relations fortes ; un grand détecte toutes les relations même faibles (rejet de H_0 malgré petit R^2)
- Minimum de 5 observations (mieux vaut en avoir > 15)
- Attention aux résidus standardisés (résidu / s) > 3
- Pour savoir si les extrêmes ont une influence : les enlever et voir les coeff. restent dans les IC des coeff. initiaux
- Attention aux distances de Cook > 1
 - la distance de Cook d'une observation i mesure l'écart des coefficients avec et sans cette observation
- Régression multiple : vérifier l'absence de multicollinéarité

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

42

Attention aux données éparpillées



Mieux vaut faire 2 modèles de régression

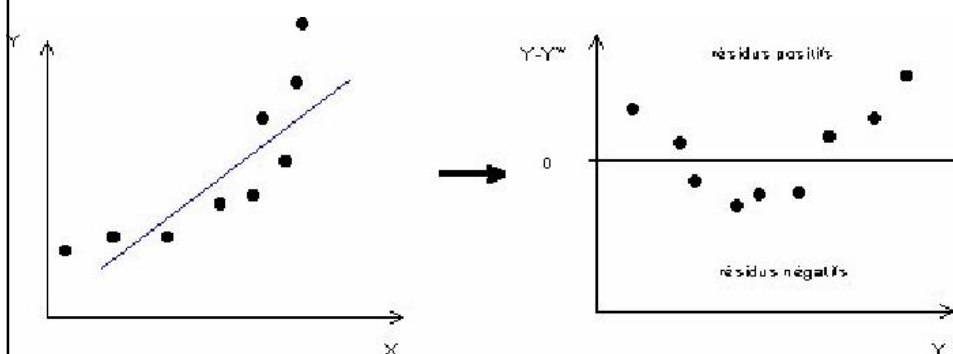
19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



43

Problème 1 : Autocorrélation des résidus



Les valeurs moyennes de Y sont sur-estimées ; les autres sont sous-estimées

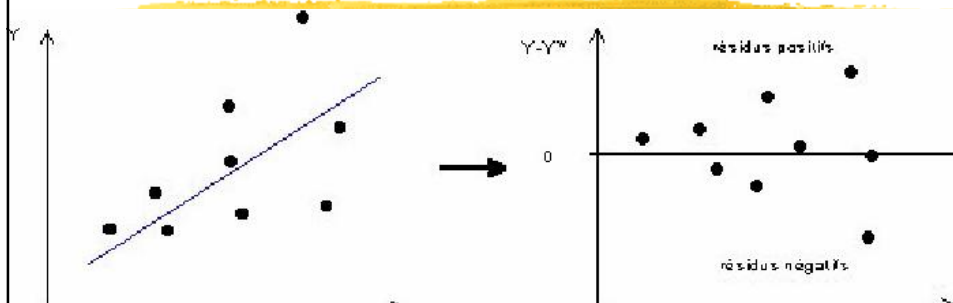
19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



44

Problème 2 : Hétéroscédasticité des résidus



Appliquer le test de Levene en regroupant en classes les valeurs de Y

Estimation précise de Y en fonction de X lorsque Y est petit ; grande incertitude quand Y est grand

=> remplacer Y par son log, son inverse ou sa racine carrée

=> ou utiliser la méthode des moindres carrés pondérés

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

45

Homoscédasticité et autocorrélation des résidus

- Utiliser un diagramme des résidus pour vérifier l'homoscédasticité et l'absence d'autocorrélation
- Statistique de Durbin-Watson pour l'autocorrélation :
 - $= \sum (e_i - e_{i-1})^2 / \sum e_i^2$
 - vaut entre 0 et 4
 - proche de 2 si pas d'autocorrélation (OK entre 1,5 et 2,5)
 - < 2 pour des corrélations positives
 - > 2 pour des corrélations négatives

	R	R-deux	R-deux ajusté	Durbin-Watson
	,983 ^a	,966	,960	1,819

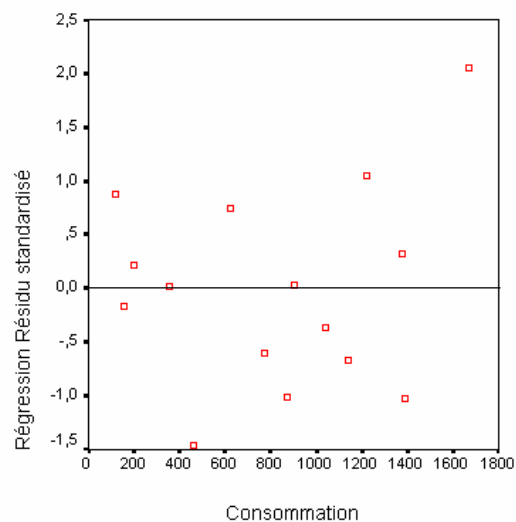
a.

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

46

Diagramme des résidus pour vérification de l'homoscédasticité et de l'absence d'autocorrélation

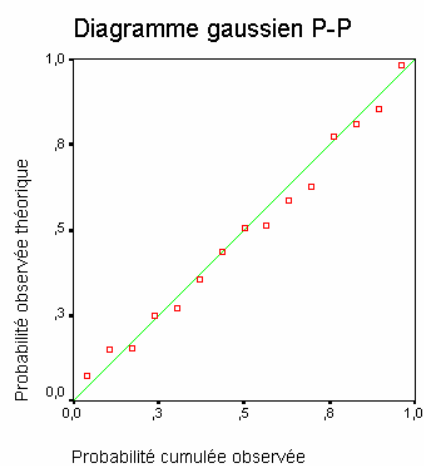
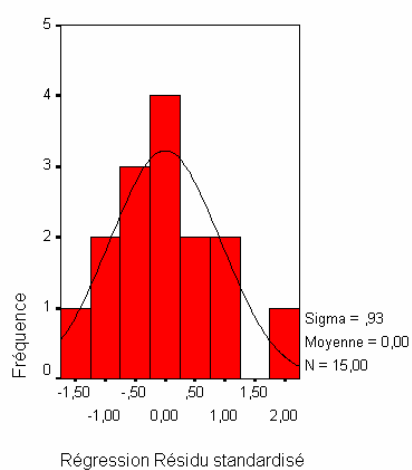


19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

47

Normalité des résidus



19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

48

Attention à la multicollinéarité

- Multicollinéarité = plusieurs variables explicatives (fortement) corrélées entre elles.
- Cela entraîne :
 - des coefficients de régression très sensibles aux fluctuations même faibles des données
 - des écarts-types élevés pour les coefficients de régression
 - une dégradation de la précision des prévisions
- Mesurée par :
 - tolérance $X_i = 1 -$ (coefficient de détermination de la régression de X_i sur les autres variables)
 - doit être $> 0,2$
 - VIF = $1 / \text{tolérance}$
 - doit être < 5

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

49

Attention à la multicollinéarité

- Autre mesure possible : les indices de conditionnement de la matrice des corrélations
 - on a multicollinéarité modérée (resp. forte) si présence d'indices $\eta_k > 10$ (resp. 30)
 - on regarde si on peut relier la valeur propre correspondante à une forte contribution ($> 50\%$) de la composante à la variance de 2 ou plusieurs variables

Modèle	Dimension	Valeur propre	Indice de conditionnement	Proportions de la variance		
				(constante)	TEMPERAT	ISOLATIO
1	1	2,145	1,000	,03	,07	,03
	2	,766	1,673	,02	,92	,02
	3	,089	4,915	,95	,01	,95

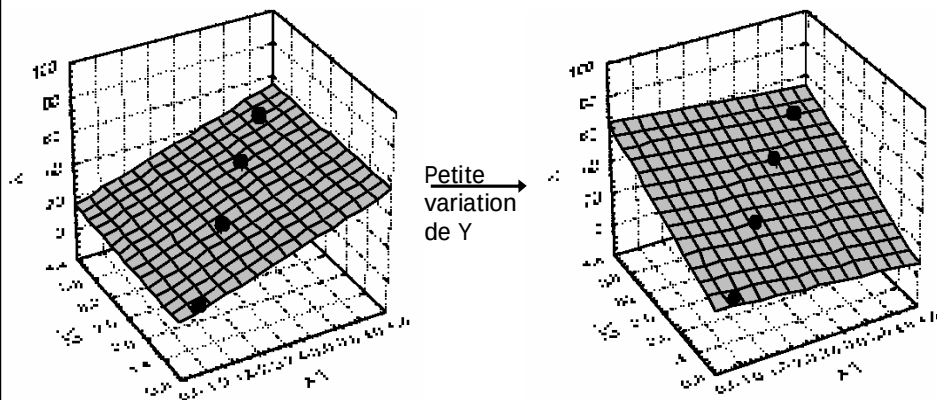
19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

50

Effets de la multicollinéarité

- X_1 et X_2 presque colinéaires $\Rightarrow$ coefficients de la régression très sensibles à de petites variations de Y



19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

51

Solutions à la multicollinéarité

- Suppression des variables concernées
 - accepter de baisser un peu R^2 pour baisser la multicollinéarité
- Transformation (logarithme...) des variables concernées
- Régression biaisée (ridge)
 - l'erreur quadratique de l'estimation de la pente β de la régression = $\text{variance_estimateur} + (\text{biais_estimateur})^2$, d'où une « erreur quadratique avec biais » < « erreur sans biais » si le biais est compensé par une faible variance
- Régression sur composantes principales
 - passer ensuite des coefficients de régression des composantes principales à ceux des variables initiales
- Régression PLS (Partial Least Squares)
 - utilisable même si : nb observations $\ll$ nb variables
 - on démontre (De Jong, 1993) que la régression PLS sur k composantes est toujours plus prédictive que la régression sur les k premières composantes principales

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

52

Principe de la régression PLS

- Le choix des variables transformées résulte d'un compromis entre :
 - maximisation de la variance des X_i (ACP)
 - maximisation de la corrélation entre X_i et Y (régression)
 - donc : on cherche les combinaisons linéaires T_j des X_i maximisant $\text{cov}^2(T_j, Y) = r^2(T_j, Y) \cdot \text{var}(T_j) \cdot \text{var}(Y)$
- 1ère étape : on cherche une combinaison $T_1 = \sum_i \lambda_{1i} X_i$ des X_i qui maximise la variance de T_1 et la corrélation entre T_1 et Y
 - **6** maximiser $\text{cov}^2(T_1, Y) = r^2(T_1, Y) \cdot \text{var}(T_1) \cdot \text{var}(Y)$
 - la solution : $\lambda_{1i} = \text{cov}(Y, X_i)$ en normant $\|(\lambda_{11}, \dots, \lambda_{1p})\| = 1$
- Puis on régresse Y sur t_1 : $Y = c_1 T_1 + Y_1$
- Et on réitère en remplaçant Y par Y_1 et les X_i par les résidus X_{i1} de leur régression sur T_1 : $X_i = c_{1i} T_1 + X_{i1}$
- On réitère jusqu'à ce que le nb de composantes T_k donne un résultat satisfaisant (vérifié par validation croisée)

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



53

Généralisation de la régression PLS

- Régression PLS inventée par Herman et Svante Wold (1983)
- Régression PLS2 développée pour prédire plusieurs Y_j simultanément
 - on peut avoir $\text{nb}(Y_j) \gg \text{nb observations}$
- Régression logistique PLS développée par Michel Tenenhaus (2000)
 - on commence par une régression PLS de l'indicatrice de Y sur les X_i (ou des indicatrices de Y , si Y a plus de 2 modalités)
 - on obtient k composantes PLS (éventuellement : $k = 1$)
 - puis on effectue une régression logistique de Y sur les composantes PLS

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



54

Utilisation de la régression PLS

- Utile en présence d'un grand nombre de variables présentant de la colinéarité ou des valeurs manquantes
- Algorithme simple (suite de régressions simples, sans inversion ni diagonalisation de matrices) => efficace sur de grands volumes de données
- Utilisation en chimie, pétrole, cosmétique, biologie, médecine, agroalimentaire
 - en cosmétique : conserver tous les ingrédients d'un produit => très nombreuses variables explicatives
 - en agroalimentaire (analyse sensorielle) : expliquer le classement d'un produit par plusieurs dégustateurs (var. Y), en fonction de ses propriétés (jusqu'à plusieurs centaines) physico-chimiques et de saveur



Technique de prédiction : La régression robuste



Régression robuste

- Valides quand les résidus des observations ne suivent pas une loi normale
- Peu sensibles aux « outliers »
- De plus en plus répandues dans les logiciels statistiques
 - SAS, R, S-PLUS, STATA...



Algorithmes de régression robuste

- Moindres médianes de carrés
- Moindres carrés winsorisés (least winsored squares)
 - remplacement des x centiles extrêmes par Q_x
- Moindres carrés écrêtés (least trimmed squares)
 - suppression des x centiles extrêmes
- Moindres carrés pondérés
 - par l'inverse de la variance de la variable à expliquer, pour compenser l'hétéroscédasticité, en posant par ex. $p_i = s^2/s_i^2$ au voisinage d'un point x_i
- Moindres carrés localement pondérés sur les voisins (LOESS)
- Doubles moindres carrés
- Régression spline
- Méthode du noyau



Autres algorithmes de régression

- Moindres valeurs absolues

$$\sum_i |x_i - \bar{x}|$$

- Régression polynomiale
- Régression sur variables catégorielles par codage optimal (moindres carrés alternés)

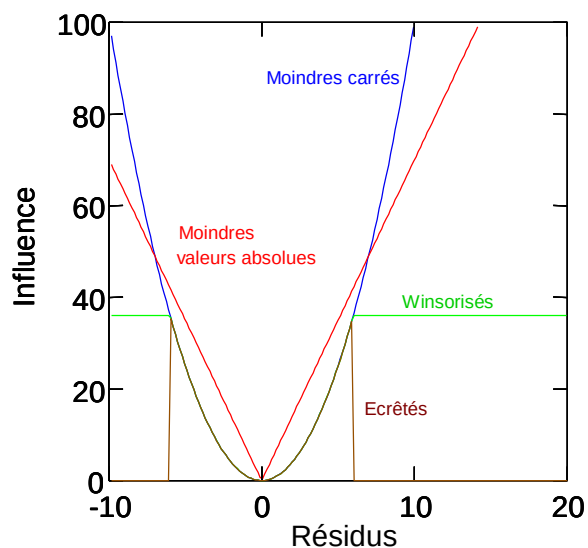
19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



59

Influence des résidus



19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



60

Principe de la régression LOESS

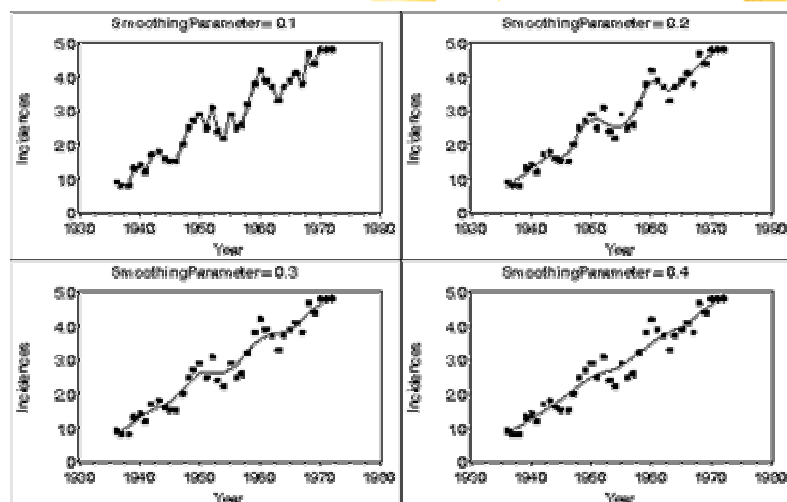
- Pour chaque point x : on prend les n voisins
 - le nombre n est choisi pour représenter un certain % de l'ensemble des points
 - ce % est appelé « paramètre de lissage » (« smoothing parameter »)
 - il existe des critères pour le choix de ce paramètre
- On pondère chacun de ces n points selon une fonction décroissante de leur distance à x
- On calcule la régression pondérée sur les n voisins pour prédire x
- LOESS utilisable avec plusieurs régresseurs
- Initiateur : Cleveland (1979)

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

61

Exemples de régressions LOESS



19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

62

Technique de classement : Analyse discriminante



Deux problématiques

- Situation : on a un ensemble d'individus appartenant chacun à un groupe, le nb de groupes étant fini et > 1
- Analyse discriminante descriptive : trouver une représentation des individus qui sépare le mieux les groupes
- Analyse discriminante prédictive : trouver des règles d'affectation des individus à leur groupe
- L'analyse discriminante offre des solutions à ces deux problématiques



Autre formulation

- Situation : on a un ensemble d'individus caractérisés par une variable à expliquer Y qualitative et des variables explicatives X_i quantitatives
- Analyse discriminante descriptive : trouver une représentation des liaisons entre Y et les X_i
- Analyse discriminante prédictive : trouver des règles de prédiction des modalités de Y à partir des valeurs des X_i
- Cette formulation est équivalente à la précédente



Les différentes formes d'analyse discriminante

	Méthode descriptive (représenter les groupes)	Méthode prédictive (prédire l'appartenance à un groupe)
Approche géométrique	Oui analyse factorielle discriminante	Oui analyse discriminante linéaire ↑ { multinormalité homoscédasticité
Approche probabiliste (bayésienne)	Non	Oui ↑ { équiprobabilité analyse discriminante linéaire a. d. quadratique a. d. non paramétrique régression logistique



Technique de classement : Analyse discriminante géométrique

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



67

L'analyse discriminante géométrique

- Y variable cible catégorielle à k modalités
 - correspondant à k groupes G_i
- X_j p variables explicatives continues
- Principe de l'analyse factorielle discriminante : remplacer les X_j par des axes discriminants : combinaisons linéaires des X_j prenant les valeurs les + différentes possibles pour des individus différant sur la variable cible
- Remarquer l'analogie avec l'ACP
- On a k-1 axes (si nb individus $n > p > k$)
- Exemple historique : les iris de Fisher (3 espèces – 4 variables, longueur et largeur des pétales et des sépales)

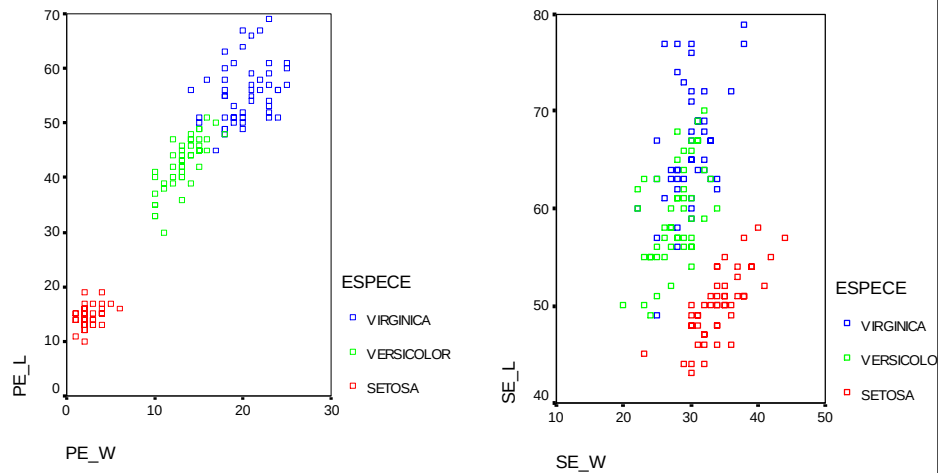
19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



68

Exemple historique : les iris de Fisher



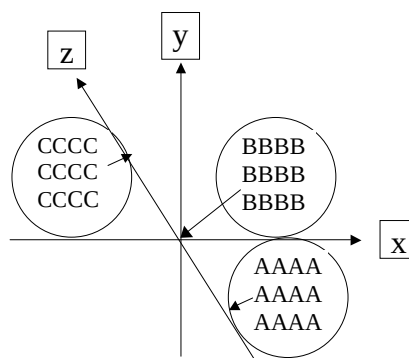
19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

69

Illustration de la problématique descriptive

- Dans l'exemple suivant :
 - l'axe « x » différencie bien les groupes « B » et « C » mais non les groupes « A » et « B »
 - l'axe « y » différencie bien les groupes « A » et « B » mais non les groupes « B » et « C »
 - en revanche l'axe « z » différencie bien les trois groupes.
- La droite :
 - $z = +1$ sépare les « B » et « C »
 - $z = -1$ sépare les « A » et « B »
- z est une fonction de score



19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

70

Double objectif de l'analyse factorielle discriminante

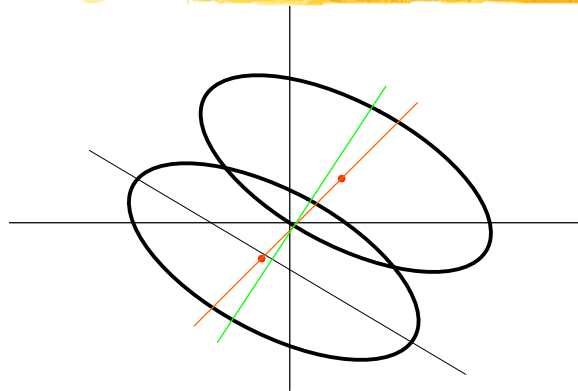
- Les n individus forment un nuage de n points dans R^p , formé des k sous-nuages G_i à différencier
- Variance interclasse (« between ») = variance des barycentres g_i (« centroïdes ») des classes G_i
 - $B = 1/n \sum n_i(g_i - g)(g_i - g)' =$ matrice de covariance « between »
- Variance intraclasse (« within ») = moyenne des variances des classes G_i
 - $W = 1/n \sum n_i V_i =$ matrice de covariance « within »
- Théorème de Huygens : $B + W =$ variance totale V
- Impossible de trouver un axe u qui simultanément :
 - maximise la variance interclasse sur u : $\max u'Bu$
 - minimise la variance intraclasse sur u : $\min u'Wu$

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

71

Visualisation du double objectif



Maximum de dispersion interclasse : u parallèle au segment joignant les centroïdes

Minimum de dispersion intraclasse : u perpendiculaire à l'axe principal des ellipses (on suppose l'homoscédasticité)

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

72

Compromis entre les 2 objectifs

- On reformule l'objectif : au lieu de maximiser $u'Bu$ ou minimiser $u'Wu$, on maximise $u'Bu/u'Wu$
- δ maximiser $u'Bu/u'Vu$ (Huygens)
- On montre que :
 - la solution u est le vecteur propre de $V^{-1}B$ associé à λ la plus grande valeur propre de $V^{-1}B$
 - u vecteur propre de $V^{-1}B$ δ u vecteur propre de $W^{-1}B$, de valeur propre $\lambda/1-\lambda$
- On dit que les métriques V^{-1} et W^{-1} sont équivalentes
 - la métrique W^{-1} (de Mahalanobis) est plus utilisée par les Anglo-saxons et les éditeurs de logiciels
- Distance d de 2 points x et y : $d^2(x,y) = (x-y)' W^{-1}(x-y)$

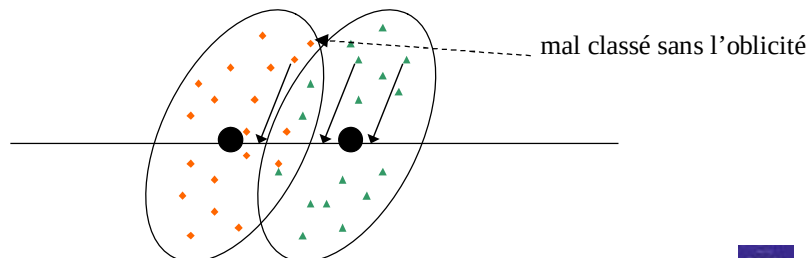
19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

73

Autre formulation de la solution

- ACP du nuage des centroïdes g_i avec :
 - métrique V^{-1}
 - ou métrique W^{-1} équivalente
- Ces métriques correspondent à une projection oblique
- Sans cette obliquité, il s'agirait d'une simple ACP mais les groupes seraient mal séparés

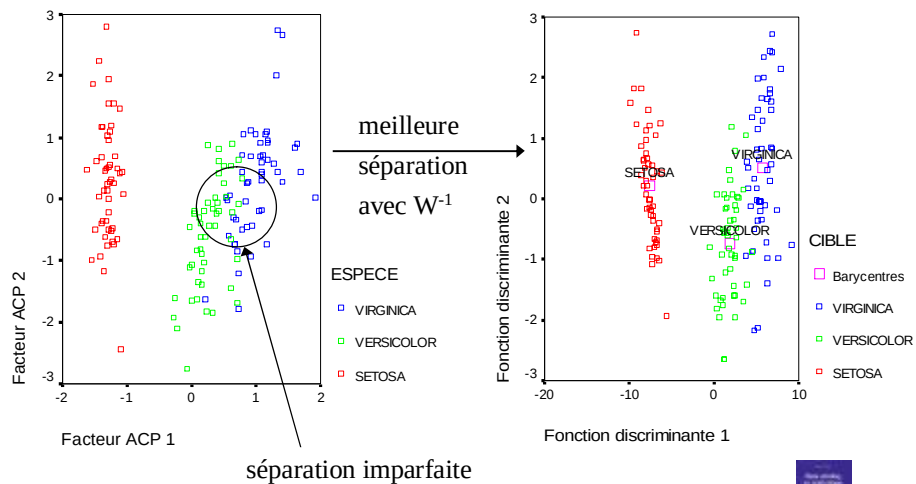


19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

74

ACP avec métrique usuelle et avec W^{-1}

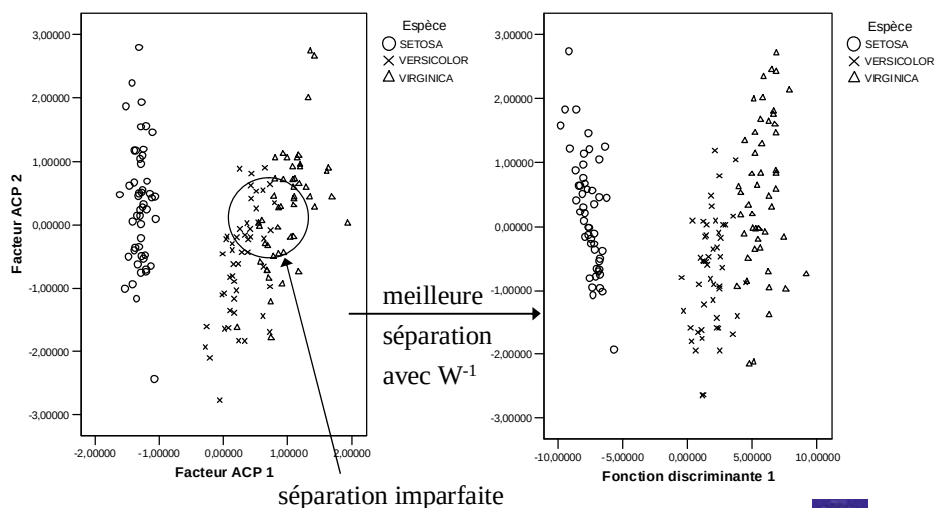


19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

75

ACP avec métrique usuelle et avec W^{-1}



19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

76

Analyse discriminante prédictive et fonctions de Fisher

- On classe x dans le groupe G_i pour lequel la distance au centre g_i est minimale :
- $d^2(x, g_i) = (x - g_i)' W^{-1} (x - g_i) = x' W^{-1} x - 2g_i' W^{-1} x + g_i' W^{-1} g_i$
- Minimiser $d^2(x, g_i)$ équivaut à maximiser $(2g_i' W^{-1} x - g_i' W^{-1} g_i)$
- $g_i' W^{-1} g_i = \alpha_i$ est une constante ne dépendant pas de x
- Pour chacun des k groupes G_i , on a une fonction discriminante de Fisher :
 - $\alpha_i + \beta_{i,1}X_1 + \beta_{i,2}X_2 + \dots + \beta_{i,p}X_p$
- et on classe x dans le groupe pour lequel la fonction est maximale

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

77



Exemple des iris de Fisher

Coefficients des fonctions de classement

	CIBLE		
	SETOSA	VERSICOLOR	VIRGINICA
SE_L	2,35442	1,56982	1,24458
SE_W	2,35879	,70725	,36853
PE_L	-1,64306	,52115	1,27665
PE_W	-1,73984	,64342	2,10791
(Constante)	-86,30847	-72,85261	-104,36832

Fonctions discriminantes linéaires de Fisher

			Classe d'affectation pour analyse 1			Total
			SETOSA	VERSICOLOR	VIRGINICA	
CIBLE	SETOSA	Effectif	50	0	0	50
	VERSICOLOR	Effectif	0	48	2	50
	VIRGINICA	Effectif	0	1	49	50
Total		Effectif	50	49	51	150

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

78



Règle d'affectation dans le cas de 2 groupes

- L'aspect descriptif est simple (l'axe discriminant joint les 2 centroïdes) et on s'intéresse à l'aspect prédictif
- On classe x dans le groupe G_1 si :
 - $2g_1' W^{-1}x - g_1' W^{-1}g_1 > 2g_2' W^{-1}x - g_2' W^{-1}g_2$
 - $\Leftrightarrow (g_1 - g_2)' W^{-1}x - \frac{1}{2} (g_1' W^{-1}g_1 - g_2' W^{-1}g_2) > 0$
- $$f(x)$$
- $f(x)$: fonction de score de Fisher
- D^2 de Mahalanobis : $d^2(g_1, g_2) = (g_1 - g_2)' W^{-1}(g_1 - g_2)$
- $W^{-1}(g_1 - g_2)$ = axe discriminant proportionnel à $V^{-1}(g_1 - g_2)$

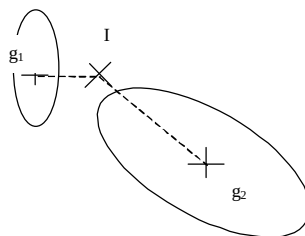
19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

79

Limite de la règle géométrique d'affectation

- Règle géométrique : affecter chaque individu au groupe dont il est le + proche (distance de l'individu au centroïde du groupe)
 - ce n'est pas trivial car il faut prendre la métrique W^{-1} (faire une projection oblique de x sur l'axe discriminant)
- A éviter si les 2 groupes ont des probabilités *a priori* ou des variances différentes



- Dans ce cas : analyse discriminante quadratique (voir plus loin)

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

80

Technique de classement : Analyse discriminante probabiliste

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



81

L'approche probabiliste (bayésienne)

- Pour tout $i \leq k$, soient :
 - $P(G_i/x)$ = proba *a posteriori* d'appartenance à G_i sachant x (connaissant les caractéristiques de x , son « dossier »)
 - $p_i = P(G_i)$ = proba *a priori* d'appartenance à G_i (proportion de G_i dans la population)
 - $f_i(x) = P(x/G_i)$ = densité conditionnelle de la loi de x connaissant son groupe G_i
- D'après le théorème de Bayes :
$$P(G_i / x) = \frac{P(G_i)P(x / G_i)}{\sum_j P(G_j)P(x / G_j)}$$
- Règle de classement bayésienne :
 - on classe x dans le groupe G_i où $P(G_i/x)$ est maximum

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



82

3 possibilités pour estimer $P(G_i/x)$

- En commençant par calculer $P(x/G_i)$ selon une méthode paramétrique (on suppose la multinormalité de $P(x/G_i)$ avec éventuellement égalité des Σ_i , donc le nb de paramètres du problème est fini : ADL ou ADQ)
- En commençant par estimer $P(x/G_i)$ selon une méthode non paramétrique (pas d'hypothèse sur la densité $P(x/G_i)$: méthode du noyau ou des plus proches voisins)
- Directement par une approche semi-paramétrique (régression logistique) où on écrit $P(G_i/x)$ sous la forme :

$$P(G_i / x) = \frac{e^{a'x+b}}{1 + e^{a'x+b}}$$



1^e possibilité : Hypothèse de multinormalité

- La densité d'une loi multinormale $N(\mu_i, \Sigma_i)$ est :

$$f_i(x) = \frac{1}{(2\pi)^{p/2} \sqrt{\det(\Sigma_i)}} \exp \left[-\frac{1}{2} (x - m_i)' \Sigma_i^{-1} (x - m_i) \right]$$

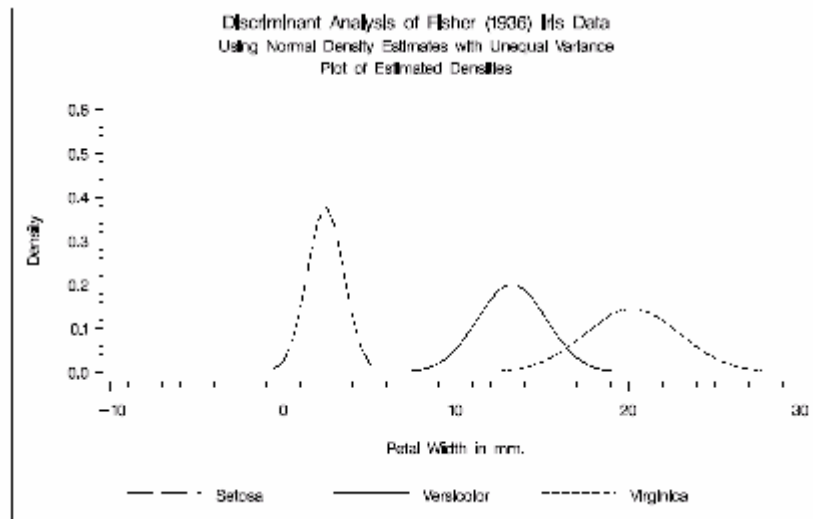
- D'après Bayes, maximiser $P(G_i/x)$ équivaut à maximiser $p_i f_i(x)$:

$$\max_i \left[\log(p_i) - \frac{1}{2} (x - m_i)' \Sigma_i^{-1} (x - m_i) - \frac{1}{2} \log(\det(\Sigma_i)) \right]$$

> On obtient une règle quadratique en x



Multinormalité



19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



85

Hypothèse d'homoscédasticité

- Sous cette hypothèse, on a : $\Sigma_1 = \Sigma_2 = \dots = \Sigma_k = \Sigma$
- On classe x dans le groupe G_i pour avoir :

$$\underset{i}{\text{maximum}} \left[\text{Log}(p_i) - \frac{1}{2} x' \Sigma^{-1} x - \frac{1}{2} m_i' \Sigma^{-1} m_i + x' \Sigma^{-1} m_i \right]$$

- Soit, puisque $x' \Sigma^{-1} x$ est indépendant de i :

$$\underset{i}{\text{maximum}} \left[\underbrace{\text{Log}(p_i) - \frac{1}{2} m_i' \Sigma^{-1} m_i}_{\text{Les probabilités a priori ne changent qu'une constante additive}} + x' \Sigma^{-1} m_i \right]$$

- Homoscédasticité (+ multinormalité) $\Rightarrow$ on passe d'une fonction quadratique à une fonction linéaire
- Avec en + l'équiprobabilité $\Rightarrow$ on a équivalence des règles géométrique (maximiser la fct de Fisher) et bayésienne

19/08/2005

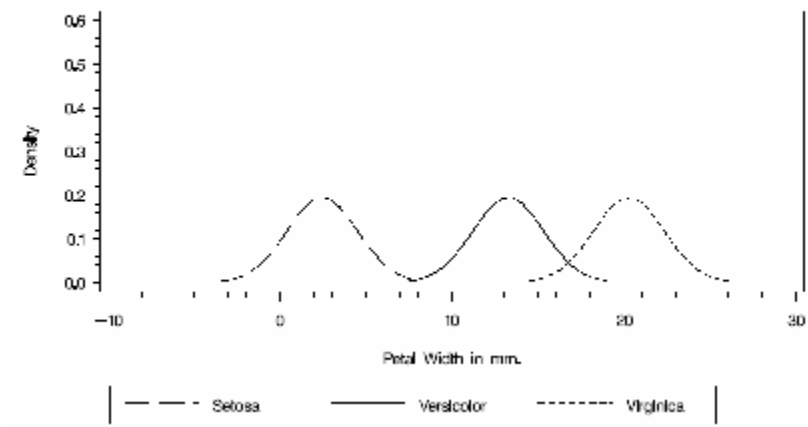
© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



86

Homoscédasticité

Discriminant Analysis of Fisher (1936) Iris Data
Using Normal Density Estimates with Equal Variance
Plot of Estimated Densities



19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



87

Cas de 2 groupes (hypothèses de multinormalité et homoscédasticité)

- Probabilité d'appartenance au groupe 1 :

$$P(G_1 / x) = \frac{p_1 \exp\left[-\frac{1}{2}(x - m_1)' \Sigma^{-1}(x - m_1)\right]}{p_1 \exp\left[-\frac{1}{2}(x - m_1)' \Sigma^{-1}(x - m_1)\right] + p_2 \exp\left[-\frac{1}{2}(x - m_2)' \Sigma^{-1}(x - m_2)\right]}$$

$$\frac{1}{P(G_1 / x)} = 1 + \frac{p_2}{p_1} \exp\left[-\frac{1}{2}(x - m_1)' \Sigma^{-1}(x - m_1) + \frac{1}{2}(x - m_2)' \Sigma^{-1}(x - m_2)\right]$$

- On peut écrire $1/P(G_1/x) = 1 + (p_2/p_1)e^{-f(x)}$
 - avec $f(x) = \frac{1}{2}(x - \mu_1)' \Sigma^{-1}(x - \mu_1) - \frac{1}{2}(x - \mu_2)' \Sigma^{-1}(x - \mu_2)$
- On classe x dans G_1 si $P(G_1/x) > 0,5$
- $\Leftrightarrow (p_2/p_1)e^{-f(x)} < 1 \Leftrightarrow f(x) > \log(p_2/p_1)$

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



88

Cas de 2 groupes (suite)

- Développons la fonction $f(x)$:
 - $f(x) = (\mu_1 - \mu_2)' \Sigma^{-1}x - \frac{1}{2}(\mu_1' \Sigma^{-1} \mu_1 - \mu_2' \Sigma^{-1} \mu_2)$
- On reconnaît la fonction de score de Fisher
- > La règle bayésienne précédente équivaut à la règle :
 - fonction de Fisher $> \log(p_2/p_1)$
- qui généralise la règle géométrique $f(x) > 0$ lorsque les probabilités *a priori* p_1 et p_2 sont différentes
- De plus, la probabilité *a posteriori* $P(G_1/x)$ s'écrit :

$$P(G_1/x) = \frac{1}{1 + \left(\frac{p_2}{p_1}\right) e^{-f(x)}} = \frac{e^{f(x)}}{\left(\frac{p_2}{p_1}\right) + e^{f(x)}}$$

- Généralisation de la fonction logistique !

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

89

En résumé :

- Avec l'hypothèse de multinormalité :
 - La règle bayésienne est quadratique
- Avec les hypothèses de multinormalité et d'homoscédasticité :
 - La règle bayésienne est linéaire
 - Dans le cas de 2 groupes, elle s'écrit $f(x) > \log(p_2/p_1)$, où $f(x)$ est la fonction de Fisher obtenue par un raisonnement géométrique
- Avec les hypothèses de multinormalité, d'homoscédasticité et d'équiprobabilité :
 - La règle bayésienne est linéaire et équivalente à la règle géométrique
 - Dans le cas de 2 groupes, elle s'écrit $f(x) > 0$ et la probabilité *a posteriori* $P(G_1/x)$ s'écrit sous la forme logistique $P(G_1/x) = 1 / (1 + e^{-f(x)})$

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

90

Coûts de mauvais classement

- On peut introduire des coûts d'erreurs
 - $C(i/j)$ = coût de classement dans G_i au lieu de G_j
 - $C(i/i) = 0$
- Coût moyen de classement en $G_i = \sum_j C(i/j) P(G_j/x)$
- On classe x dans le G_i qui minimise le coût
- Cas de 2 groupes :
 - Coût moyen d'un classement en G_1 : $C(1/2) P(G_2/x)$
 - Coût moyen d'un classement en G_2 : $C(2/1) P(G_1/x)$
 - On classe x en G_1 si $C(1/2) P(G_2/x) < C(2/1) P(G_1/x)$

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



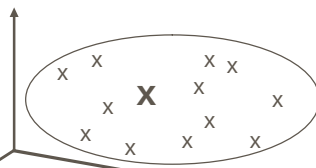
91

An. Discriminante non paramétrique

- D'après Bayes :

$$P(G_i / x) = \frac{p_i f_i(x)}{\sum_j p_j f_j(x)}$$
- Problème d'estimation de la densité :

$$f_i(x) = \frac{\text{fréquence}}{\text{volume}}$$
- Méthodes :
 - noyau (on fixe le diamètre)
 - k-plus proches voisins (on fixe le nb de voisins)
- Condition : avoir un échantillon de grande taille



19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



92

D² de Mahalanobis

- Définition : $D^2 = d^2(g_1, g_2) = (g_1 - g_2)' W^{-1}(g_1 - g_2)$
- Le carré D^2 de la distance de Mahalanobis fournit une mesure de la distance entre les deux groupes à discriminer, et donc de la qualité de la discrimination
- Analogue au R^2 d'une régression
- Plus D^2 est grand, mieux c'est
- On peut faire un test de Fisher sur l'hypothèse nulle que tous les centroïdes sont égaux
- Il peut servir de critère dans une régression pas à pas



R²

- Corrélation canonique = coefficient de corrélation entre la fonction de score et la moyenne par classe (pour chaque individu : on prend la moyenne de la fonction discriminante dans sa classe)
- Carré de la corrélation canonique $R =$ coefficient de détermination $R^2 =$ proportion de la variance de la fonction discriminante expliquée par l'appartenance à l'une ou l'autre classe à discriminer
- Autrement dit $R^2 =$ variance interclasse / variance totale
 - Le but de l'analyse discriminante est de maximiser ce rapport



Lambda de Wilks

- Lambda de Wilks = variance intraclasse / variance totale
 - varie entre 0 et 1 (var. totale = var. intra + var. inter)
 - $\lambda = 1 \Rightarrow$ tous les centroïdes sont égaux
- Plus λ est bas, mieux c'est
- Test de Fisher sur le lambda de Wilks $\Leftrightarrow$ Test de l'hypothèse nulle que tous les centroïdes sont égaux
- Il peut servir de critère dans une régression pas à pas

	Lambda de Wilks	F	ddl1	ddl2	Signification
SE_L	,381	119,265	2	147	,000
SE_W	,599	49,160	2	147	,000
PE_L	,059	1180,161	2	147	,000
PE_W	,071	960,007	2	147	,000

Les groupes diffèrent beaucoup sur la longueur des pétales

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

95

Matrice de confusion Validation croisée

Matrice de confusion^{a,c}

CIBLE			Classe(s) d'affectation prévue(s)			Total
			SETOSA	VERSICOLOR	VIRGINICA	
Original	Effectif	SETOSA	50	0	0	50
		VERSICOLOR	0	48	2	50
		VIRGINICA	0	1	49	50
	%	SETOSA	100,0	,0	,0	100,0
		VERSICOLOR	,0	96,0	4,0	100,0
		VIRGINICA	,0	2,0	98,0	100,0
Validé-croisé ^b	Effectif	SETOSA	50	0	0	50
		VERSICOLOR	0	48	2	50
		VIRGINICA	0	1	49	50
	%	SETOSA	100,0	,0	,0	100,0
		VERSICOLOR	,0	96,0	4,0	100,0
		VIRGINICA	,0	2,0	98,0	100,0

a. Dans la validation croisée, chaque observation est classée par les fonctions dérivées de toutes les autres observations.

b. 98,0% des observations originales classées correctement.

c. 98,0% des observations validées-croisées classées correctement.

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

96

Résumé des critères statistiques

- D^2 de Mahalanobis : test de Fisher
- Lambda de Wilks = $1 - R^2$: test de Fisher

Pas	Nombre de variables	Lambda	F exact			
			Statistique	ddl1	ddl2	Signification
1	1	,059	1180,161	2	147,000	,000
2	2	,037	307,105	4	292,000	,000
3	3	,025	257,503	6	290,000	,000
4	4	,023	199,145	8	288,000	,000

- Matrice de confusion : test Q de Press
- Coefficients discriminants standardisés (sur var. centrées réduites)
 - pour comparer l'importance des variables explicatives

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

97



Syntaxe SAS de l'analyse discriminante

```
ods rtf file="c:\fisher_sas.doc";
proc stepdisc data=matable.ascorer;
class cible;
var var1 var2 ... vari; run;
proc discrim data=matable.ascorer method=normal pool=yes
  crossvalidate all canonical out=matable.scoree
  outstat=matable.destat;
class cible;
priors proportional;
var var1 var2 ... vari; run;
proc discrim data=matable.destat testdata=matable.test
  testout=tout;
class cible;
var var1 var2 ... vari; run;
ods rtf close ;
```

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

98



Fichier en sortie OUTSTAT

Obs	cible	_TYPE_	_NAME_	nbproduits	abonnement1	nbenfants	abonnement2	evolconso	nsorties
1	.	N		6385.00	6385.00	6385.00	6385.00	6385.00	6385.00
2	0	N		5306.00	5306.00	5306.00	5306.00	5306.00	5306.00
3	1	N		1079.00	1079.00	1079.00	1079.00	1079.00	1079.00
4	.	MEAN		8.94	371.28	1.34	23.11	1.16	6.48
5	0	MEAN		8.47	281.68	1.38	19.62	1.14	5.96
6	1	MEAN		11.23	811.86	1.15	40.28	1.25	9.05
...	...	...	...	...	...	...	...	...	...
119	0	LINEAR	_LINEAR_	0.38	-0.00	1.12	-0.00	8.42	0.05
120	0	LINEAR	_CONST_	-7.50	-7.50	-7.50	-7.50	-7.50	-7.50
121	1	LINEAR	_LINEAR_	0.48	0.00	0.83	0.01	9.14	0.09
122	1	LINEAR	_CONST_	-11.27	-11.27	-11.27	-11.27	-11.27	-11.27

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

99

Avantages de l'analyse discriminante

- Problème à solution analytique directe (inverser W)
- Optimale quand les hypothèses de non colinéarité, homoscedasticité et multinormalité sont vérifiées
- Les coefficients des combinaisons linéaires constituent un résultat relativement explicite
- Modélise très bien les phénomènes linéaires
- Aptitude à détecter les phénomènes globaux
- Ne nécessite pas un gros ensemble d'apprentissage
- Rapidité de calcul du modèle
- Possibilité de sélection pas à pas
- Facilité d'intégrer des coûts d'erreur de classement
- Technique implémentée dans de nombreux logiciels

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

100

Inconvénients de l'analyse discriminante

- Ne détecte que les phénomènes linéaires
- Ne s'applique pas à tout type de données (données numériques sans valeurs manquantes)
 - mais possibilité d'utiliser une ACM (méthode DISQUAL)
- Hypothèses contraignantes, et pour s'en rapprocher :
 - normaliser les variables
 - sélectionner soigneusement les variables les + discriminantes
 - éliminer les variables colinéaires
 - éliminer les individus hors norme
 - s'il reste de l'hétéroscédasticité, mieux vaut avoir des classes de tailles comparables
 - travailler sur des populations homogènes
 - il vaut donc mieux préalablement segmenter

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



101

Technique de classement : La régression logistique

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



102

La régression logistique binaire

- Y variable cible binaire $Y = 0 / 1$
- X_j p variables explicatives continues ou binaires
 - $p = 1$ régression logistique simple
 - $p > 1$ régression logistique multiple
- Généralisation : régression logistique polytomique
 - la variable cible Y est catégorielle à k modalités
 - cas particulier : Y ordinaire (régression logistique ordinaire)
- Pb de régression : modéliser l'espérance conditionnelle

$$E(Y/X=x) = \text{Prob}(Y=1/X=x)$$
 sous la forme $E(Y/X=x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$
- Difficulté ! X_i continues $\Rightarrow$ terme de droite non borné alors que $\text{Prob}(Y=1/X=x) \in [0,1] \Rightarrow$ il faut le transformer !
 - en régression linéaire : $E(Y/X=x)$ n'est pas bornée

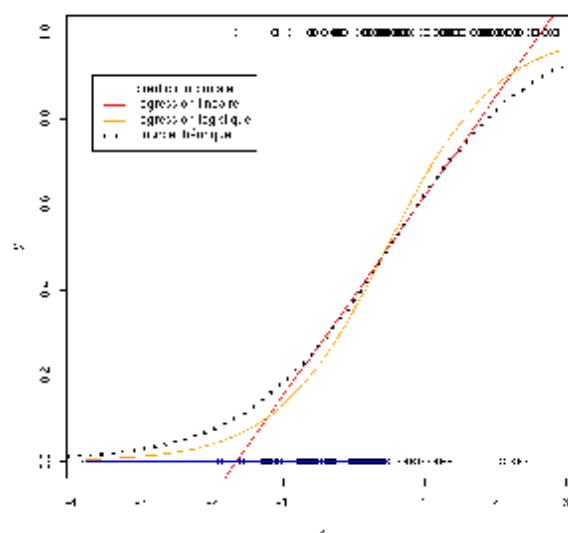
19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

103

Prédiction d'une variable binaire

Comparaison des régressions linéaire et logistique



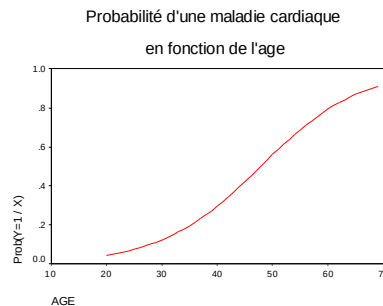
19/08/2005

l.fr

104

La régression logistique binaire

- Visiblement la régression linéaire ne convient pas (distribution des résidus !)
- La figure fait pressentir que ce n'est pas une fonction linéaire de $\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$ qu'il faut appliquer, mais une courbe en S
- Les courbes en S sont courantes en biologie et en épidémiologie



19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

105

Age and Coronary Heart Disease (CHD)

(source : Hosmer & Lemeshow - Michel Tenenhaus)

CHD = maladie coronarienne (rétrécissement des artères du muscle cardiaque)

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

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

106

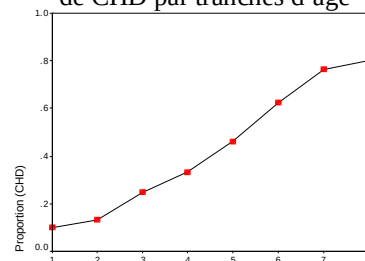
La régression logistique binaire

- Ici, difficile de calculer $\pi(x) := \text{Prob}(Y=1/X=x)$ car trop peu de valeurs de Y pour une valeur x donnée
- On regroupe les valeurs de X par tranches :
 - proportion des Y = 1 sachant x : meilleur estimateur de la probabilité que Y = 1 sachant x
 - procédure de regroupement en classes : classique en scoring !

Tableau des effectifs
de CHD par tranches 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 tranches d'âge



19/08/2005

© Stéphane Tufféry - Data Mining & Scoring

Fonction de lien

- On écrit donc $\pi(x) = \text{Prob}(Y=1/X=x)$ sous la forme :

$$p(x) = \frac{e^{b_0 + \sum_j b_j x_j}}{1 + e^{b_0 + \sum_j b_j x_j}}$$

- 6 $\text{Log}\left(\frac{p(x)}{1-p(x)}\right) = b_0 + b_1 x_1 + \dots + b_p x_p$

Fonction de lien : $\text{Logit}(p(x))$

- Cohérent avec la règle bayésienne de l'analyse discriminante et le calcul de la probabilité *a posteriori* dans le cas gaussien homoscédastique

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



108

Autres régressions : probit et log-log

Modèle	Fonction de lien	Fonction de transfert
Logit	$\text{Log}(\mu / [1 - \mu])$	$\exp(t) / [1 + \exp(t)]$
Probit (normit)	fonction inverse de la fonction de répartition d'une loi normale centrée réduite	$s(t) = \int_{-\infty}^t \frac{e^{-z^2/2}}{\sqrt{2\pi}} dz$
Log-log	$\text{Log}[-\text{Log}(1-\mu)]$	$1 - \exp[-\exp(t)]$

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

109

Logit : odds-ratio d'un régresseur X_i

- Mesure l'évolution du rapport des probas d'apparition de l'événement $Y=1$ contre $Y=0$ (odds = « cote » des parieurs) lorsque X_i passe de x à $x+1$. Dans ce cas, $\text{logit}(\pi(x))$ augmente du coefficient β_i de $X_i \Rightarrow$ la cote $\pi(x)/[1 - \pi(x)]$ est multipliée par $\exp(\beta_i)$
- Formule générale :

$$OR = \frac{p(x+1)/[1-p(x+1)]}{p(x)/[1-p(x)]} = e^{\beta_i}$$

- Si X_i est binaire 0/1, la formule devient :

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

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

110

Interprétation du odds-ratio OR

- Attention : odds-ratio $\neq$ du risque relatif $\pi(x+1)/\pi(x)$
 - sauf si $\pi(x)$ est petit (détection de phénomène rare)
- Un seul OR pour X binaire
 - ex : comparer les hommes ($x=1$) et les femmes ($x=0$)
- Un seul OR est plus douteux pour X continue
 - ex : comparer l'âge 61 et 60, 60 et 59... avec le même OR ?
Risque de manque de robustesse par manque de données (voir CHD ci-dessus). Non détection de la non-linéarité.
- OR à n'utiliser sur des variables qualitatives qu'après dichotomisation (nb indicatrices = nb modalités - 1, en prenant une modalité comme référence)
 - ex : comparer « petites villes » et « campagne » avec un OR1 et comparer « grandes villes » et « campagne » avec un OR2, car aucune raison d'avoir $OR1 = OR2$
 - indicatrices créées automatiquement par certains logiciels

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

111

Odds-ratio d'une variable qualitative

- Exemple : comparaison de la probabilité $\pi(x)$ d'apparition d'un événement dans les grandes villes, les petites villes et à la campagne
 - quand on passe de la modalité de référence (« campagne ») à la modalité « petite ville », la cote $\pi(x)/[1 - \pi(x)]$ est multipliée par l'exponentielle 0,573 de la différence des coefficients B associés à la modalité « petite ville » ($B = -0,558$) et à la modalité de référence ($B = 0$)
 - autrement dit, la cote $\pi(x)/[1 - \pi(x)]$ de l'événement (différent de sa probabilité $\pi(x)$!) est presque 2 fois plus faible dans une petite ville qu'à la campagne

	B	E.S.	Wald	ddl	Signif.	Exp(B)	IC pour Exp(B) 95,0%	
							Inférieur	Supérieur
campagne			36,671	2	0,000			
petite ville	-0,55767728	0,136	16,784	1	0,000	0,573	0,438	0,748
grande ville	0,28802599	0,143	4,057	1	0,044	1,334	1,008	1,765
Constante	-1,25610388	0,236	28,363	1	0,000	0,285		

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

112

Illustration du découpage en classes

- Un même modèle de score avec 4 variables explicatives :
 - continues
 - découpées en classes considérées comme var. ordinales
 - découpées en classes considérées comme var. nominales
- Comparaison des performances

Aire sous la courbe ROC

Variable(s) de résultats tests	Zone	Erreur Std. ^a	Signif. asymptotique ^b	Intervalle de confiance 95% asymptotique	
				Borne inférieure	Borne supérieure
Var explicatives en classes ordinales	,834	,008	,000	,818	,850
Var explicatives en classes nominales	,836	,008	,000	,820	,852
Var explicatives continues	,820	,010	,000	,801	,839

a. Dans l'hypothèse non-paramétrique

b. Hypothèse nulle : zone vraie = 0.5

- Le découpage en classes nominales l'emporte

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



113

Estimation des paramètres

Les données

vecteur X	Y
x^1	y^1
$\vdots$	$\vdots$
x^i	y^i
$\vdots$	$\vdots$
x^n	y^n

$y^i = 0 \text{ ou } 1$

Le modèle

$$p(x^i) = P(Y = 1 / X = x^i)$$

$$= \frac{e^{b_0 + \sum_j b_j x_j^i}}{1 + e^{b_0 + \sum_j b_j x_j^i}}$$

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



114

Recherche du maximum de vraisemblance

- Proba d'observer les données $[(x^1, y^1), (x^2, y^2), \dots, (x^n, y^n)]$

$$= \prod_{i=1}^n \text{Prob}(Y = y^i / X = x^i) = \prod_{i=1}^n p(x^i)^{y^i} (1 - p(x^i))^{1-y^i}$$

$$= \prod_{i=1}^n \left(\frac{e^{b_0 + \sum_j b_j x_j^i}}{1 + e^{b_0 + \sum_j b_j x_j^i}} \right)^{y^i} \left(1 - \frac{e^{b_0 + \sum_j b_j x_j^i}}{1 + e^{b_0 + \sum_j b_j x_j^i}} \right)^{1-y^i} = L(b_0, b_1, \dots, b_p)$$

- On cherche les coefficients β_i maximisant la vraisemblance
- Pas de solution analytique => utiliser une méthode numérique itérative (ex : Newton-Raphson)

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

115

Cas de la régression logistique simple

- On cherche 2 coefficients β_0 et β_1 maximisant la vraisemblance

$$L(\beta_0, \beta_1) = \prod_{i=1}^n \left(\frac{e^{b_0 + b_1 x^i}}{1 + e^{b_0 + b_1 x^i}} \right)^{y^i} \left(1 - \frac{e^{b_0 + b_1 x^i}}{1 + e^{b_0 + b_1 x^i}} \right)^{1-y^i}$$

- Pour ces coefficients, la matrice des covariances

$$V(b) = \begin{bmatrix} V(b_0) & \text{Cov}(b_0, b_1) \\ \text{Cov}(b_0, b_1) & V(b_1) \end{bmatrix}$$

- est estimée par la matrice

$$\left[-\frac{\partial^2 \text{Log } L(b)}{\partial b^2} \right]_{b=(b_0, b_1)}^{-1}$$

intervient dans la statistique de Wald (voir + loin)

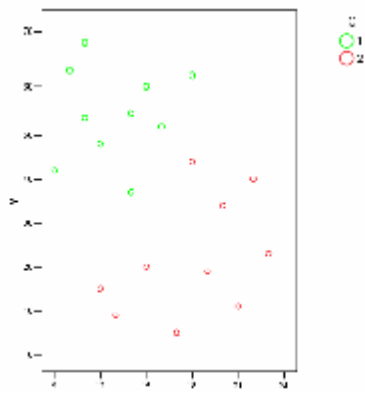
- Il faut inverser la matrice hessienne $H = \partial^2 \text{Log } L(\beta) / \partial \beta^2$
 - impossible en cas de séparation complète des groupes

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

116

Séparation complète des groupes



Variables dans l'équation

		B	E.S.	Wald	ddl	Signif.	E
Etape 1	X	13,184	2237,865	,000	1	,995	53
	Y	-2,726	441,662	,000	1	,995	
	Constante	-100,184	21856,781	,000	1	,996	

a. Variable(s) entrées à l'étape 1 : X, Y.

Historique des itérations^{a,b,c,d}

		-2log-vrais emblance	Coefficients		
Itération			Constante	X	Y
Etape 1	1	9,271	-,132	,182	-,071
	2	5,000	-,750	,344	-,119
	3	2,974	-2,082	,563	-,172
	4	1,747	-4,940	,908	-,237
	5	,816	-10,239	1,505	-,339
	6	,319	-16,448	2,252	-,478
	7	,121	-22,508	3,017	-,629
	8	,045	-28,505	3,789	-,785
	9	,017	-34,483	4,567	-,944
	10	,006	-40,456	5,349	-1,105
	11	,002	-46,429	6,131	-1,267
	12	,001	-52,401	6,914	-1,429
	13	,000	-58,374	7,698	-1,591
	14	,000	-64,346	8,481	-1,753
	15	,000	-70,319	9,265	-1,915
	16	,000	-76,292	10,049	-2,077
	17	,000	-82,265	10,833	-2,239
	18	,000	-88,238	11,617	-2,401
	19	,000	-94,211	12,400	-2,564
	20	,000	-100,184	13,184	-2,726

a. Méthode : Entrée

b. La constante est incluse dans le modèle.

c. -2log-vraisemblance initiale : 27,726

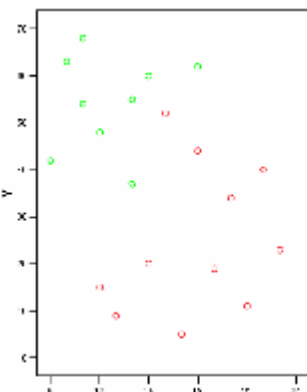
d. L'estimation a été interrompue au numéro d'itération 20 parce que le nombre maximal d'itérations a été atteint. Solution finale introuvable.

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

117

Séparation incomplète des groupes



Variables dans l'équation

		B	E.S.	Wald	ddl	Signif.	Exp(B)	IC pour Exp(B) 95,0%	
Etape 1	X	1,417	1,379	1,056	1	,304	4,124	,276	61,535
	Y	-,247	,189	1,696	1	,193	,781	,539	1,133
	Constante	-10,841	13,949	,604	1	,437	,000		

a. Variable(s) entrées à l'étape 1 : X, Y.

Historique des itérations^{a,b,c,d}

			Coefficients		
Itération		-2log-vraisemblance	Constante	X	Y
Etape 1	1	11,036	-,620	,204	-,062
	2	7,473	-1,523	,373	-,100
	3	5,973	-3,054	,583	-,136
	4	5,323	-5,345	,840	-,172
	5	5,079	-7,956	1,113	-,207
	6	5,020	-9,952	1,321	-,234
	7	5,014	-10,746	1,406	-,245
	8	5,014	-10,840	1,417	-,247
	9	5,014	-10,841	1,417	-,247
	10	5,014	-10,841	1,417	-,247

a. Méthode : Entrée

b. La constante est incluse dans le modèle.

c. -2log-vraisemblance initiale : 27,526

d. L'estimation a été interrompue au numéro d'itération 10 parce que les estimations de paramètres ont changé de moins de ,001.

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

118

Vraisemblance et déviance d'un modèle

- Soit $L(\beta_0)$ = vraisemblance du modèle réduit à la constante
- Soit $L(\beta_n)$ = vraisemblance du modèle saturé (avec toutes les variables explicatives et toutes les interactions pour en avoir autant que d'observations) = vraisemblance maximale
- Soit $L(\beta_k)$ = vraisemblance du modèle avec k variables
- On définit la déviance :
 - > $D(\beta_k) = -2 \log [L(\beta_k) - L(\beta_n)] = \log [L(\beta_n) / L(\beta_k)]^2$
- But de la régression logistique : maximiser la vraisemblance $L(\beta_k) \Leftrightarrow$ minimiser la déviance $D(\beta_k)$
 - $L(\beta_k)$ petit $\in [0,1] \Rightarrow -2 \log L(\beta_k) \in [0, +\infty[$ avec un terme « 2 » pour avoir l'analogie entre déviance et $\Sigma(\text{erreurs})^2$

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



119

Comparaison de modèles

- Pour savoir s'il convient d'ajouter l variables explicatives à un modèle qui en contient déjà k
- On calcule la différence des déviiances
 - > $D(\beta_k) - D(\beta_{k+l}) = -2 (\log L(\beta_k) - \log L(\beta_{k+l}))$
- Sous l'hypothèse H_0 de la nullité des l derniers coefficients, $D(\beta_k) - D(\beta_{k+l})$ suit un chi-2 à l d° de liberté
 - > Sous le seuil critique du chi-2 : on rejette les l nouvelles variables
- Méthode la plus utilisée en régression pas à pas

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



120

Autres indicateurs

- Cas particulier
 - > $D(\beta_0) - D(\beta_k) = -2 \log [L(\beta_0) - L(\beta_k)]$
- suit une loi du χ^2 à k degrés de liberté sous l'hypothèse H_0 de la nullité de tous les coefficients $\beta_1, \dots, \beta_k$. Rejet de H_0 si cette différence dépasse le seuil critique du χ^2 .
- Critère d'Akaike AIC = $-2 \log L(\beta_k) + 2(k+1)$
 - k = nb de ddl = nb de paramètres à expliquer
- Critère de Schwartz BIC = $-2 \log L(\beta_k) + (k+1) \cdot \log n$
 - n = nb total d'individus
 - pénalise les modèles complexes
- Ces 2 critères permettent de comparer 2 modèles
 - ils doivent être le plus bas possible



Le chi-2 de Wald

- Statistique de Wald = $(\beta_i / \text{écart-type}(\beta_i))^2$
- suit un chi-2 à 1 degré de liberté sous l'hypothèse nulle H_0 : le coefficient $\beta_i = 0$
- > teste la significativité de chaque coefficient β_i
 - en comparant le sous-modèle excluant X_i avec le modèle incluant toutes les variables
 - on doit avoir Wald > 4
- Méthode utilisée en régression pas à pas
- NB : Éviter le chi-2 de Wald si peu d'observations ou si les coefficients β_i sont grands



Test de Hosmer et Lemeshow

Tableau de contingence pour le test de Hosmer-Lemeshow

		CHD = 0		CHD = 1		Total
		Observé	Théorique	Observé	Théorique	
Etape 1	1	9	9,213	1	,787	10
	2	9	8,657	1	1,343	10
	3	8	8,095	2	1,905	10
	4	8	8,037	3	2,963	11
	5	7	6,947	4	4,053	11
	6	5	5,322	5	4,678	10
	7	5	4,200	5	5,800	10
	8	3	3,736	10	9,264	13
	9	2	2,134	8	7,866	10
	10	1	,661	4	4,339	5

On découpe les observations en $g = 10$ groupes, ordonnés par probabilité croissante (selon le modèle)

On calcule le χ^2 du tableau $g \times 2$ des fréquences observées et attendues pour l'événement modélisé (ici CHD = 1), que l'on compare à la loi du χ^2 à $(g - n)$ d° de libertés, avec en général $n = 2$.

Si le χ^2 est grand, le modèle ne s'ajuste pas bien aux données

Test de Hosmer-Lemeshow

Etape	Khi-deux	ddl	Signif.
1	,890	8	,999

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

123

Matrice de confusion

Tableau de classification

Observé			Prévu		
			CHD		Pourcentage correct
			0	1	
Etape 1	CHD	0	45	12	78,9
		1	14	29	67,4
Pourcentage global					74,0

a. La valeur de césure est ,500

Variables dans l'équation

		B	E.S.	Wald	ddl	Signif.	Exp(B)	IC pour Exp(B) 95,0%	
								Inférieur	Supérieur
Etape 1	AGE	,111	,024	21,254	1	,000	1,117	1,066	1,171
	Constante	-5,309	1,134	21,935	1	,000	,005		

a. Variable(s) entrées à l'étape 1: AGE.

> 4

1 ∉ IC

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

124

Autres tests

Récapitulatif du modèle

Etape	-2log-vrais emblance	R-deux de Cox & Snell	R-deux de Nagelkerke
1	107,353 ^a	,254	,341

a.

$-2 \text{ Log } L(\beta_k)$

« Équivalent »
du R² de la
régression
linéaire - ne
peut atteindre 1

R² ajusté pour
varier entre 0
et 1

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

125

Tests sous SAS (proc logistic)

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	138.663	111.353
SC	141.268	116.563
-2 Log L	136.663	107.353

R-Square	0.2541	Max-rescaled R-Square	0.3410
----------	--------	-----------------------	--------

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	29.3099	1	<.0001
Score	26.3989	1	<.0001
Wald	21.2541	1	<.0001

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

126

Syntaxe SAS de la régression logistique

```
ods rtf file= « c:\logistic_sas.doc »;
proc logistic data=matable.ascorer outmodel=mon.modele;
  class
    var_quali_1 (ref='A1') ... var_quali_i (ref='Ai') / param=ref;
  model cible (ref='0')=
    var_quali_1 ... var_quali_i var_quanti_1 var_quanti_j
    / selection=forward sle=.05 maxiter=25 outroc=roc rsquare lackfit;
  output out=matable.scoree predprobs=(individual);
run;
symbol1 i=join v=none c=blue;
proc gplot data=roc;
  where _step_ in (1 7);
  title 'Courbe ROC';
  plot _sensit*_1mspec=1 / vaxis=0 to 1 by .1 cframe=ligr;
run;
ods rtf close ;
proc logistic inmodel=mon.modele; score data= autretable.ascorer;run;
```

Hosmer-Lemeshow

R²

niv. de signif. en entrée

enregistre les probas prédites
et les var. _FROM_ et _INTO_

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

127

Tests de concordance

- Soit n_1 (resp. n_2) le nb d'observations où $Y=0$ (resp. $Y=1$)
- Soit $n = n_1 + n_2$ le nb total d'observations
- On s'intéresse aux $t = n_1 n_2$ paires formées d'une observation où $Y = 1$ et d'une observation où $Y = 0$
- Parmi ces t paires : on a concordance si la proba estimée que $Y = 1$ est + grande quand $Y = 1$ que quand $Y = 0$
- Soient nc = nb de paires concordantes ; nd = nb de paires discordantes ; $t - nc - nd$ = nb d'ex-æquo (« tied »)
- D de Somer = $(nc - nd) / t$
- Gamma = $(nc - nd) / (nc + nd)$
- Tau-a = $2 (nc - nd) / n(n-1)$
- $c = (nc + 0,5[t - nc - nd]) / t$ = aire sous la courbe ROC
- Plus ces indices sont proches de 1, meilleur est le modèle

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	79.0	Somers' D	0.600
Percent Discordant	19.0	Gamma	0.612
Percent Tied	2.0	Tau-a	0.297
Pairs	2451	c	0.800

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

128

Effet de la multicollinéarité

- Régression logistique avec 2 variables VAR1 et VAR2 fortement corrélées :

		VAR1	VAR2
VAR1	Corrélation de Pearson	1	,975**
	N	36841	36300
VAR2	Corrélation de Pearson	,975**	1
	N	36300	36300

** La corrélation est significative au niveau 0.01

- On constate une dégradation de la prédictivité de VAR1 avec l'introduction de VAR2 :

		B	E.S.	Wald	ddl	Signif.	Exp(B)	IC pour Exp(B) 95,0%	
Etape 1	VAR1	,098	,004	759,291	1	,000	1,103	Inférieur	Supérieur
	Constante	-4,898	,062	6290,898	1	,000	,007		

a. Variable(s) entrées à l'étape 1: VAR1.

		B	E.S.	Wald	ddl	Signif.	Exp(B)	IC pour Exp(B) 95,0%	
Etape 2	VAR1	,020	,014	2,125	1	,145	1,020	Inférieur	Supérieur
	VAR2	,092	,015	39,280	1	,000	1,096	1,065	1,129
	Constante	-4,993	,065	5867,055	1	,000	,007		

a. Variable(s) entrées à l'étape 2: VAR2.

Résumé des tests

- Test du chi-2 sur $-2 \log [L(\beta_0) - L(\beta_k)]$
- Test du chi-2 sur indicateur de Wald (> 4)
- Test du chi-2 sur le score
- $1 \notin$ IC à 95 % de l'odds-ratio = $\exp(a_i \pm 1,96\sigma(a_i))$
- Test de Hosmer et Lemeshow sur comparaison des proportions observées et théoriques
- R^2 de Cox-Snell et R^2 ajusté de Nagelkerke
- AIC et BIC
- Multicollinéarité (tolérance, VIF, indices de conditionnement)
- Matrice de confusion, tests de concordance, courbe ROC, indice de Gini
- Moins de 10 variables explicatives sont souvent retenues

Influence de l'échantillonnage 1/2

- La régression logistique consiste à écrire $\pi(x) := P(Y=1/X=x)$ sous la forme

$$\text{Log}\left(\frac{p(x)}{1-p(x)}\right) = b_0 + b_1x_1 + \dots + b_px_p$$

- avec des coefficients maximisant la vraisemblance
- Si l'on effectue un échantillonnage E indépendant de X, alors la probabilité $\pi_E(x) := P(Y=1/X=x, X \in E)$ vérifie

$$\text{Log}\left(\frac{p_E(x)}{1-p_E(x)}\right) = b'_0 + b_1x_1 + \dots + b_px_p$$

- avec $\beta'_0 = \beta_0 + \text{constante} (= \log(p_{1,E}/p_{0,E}) + \log(p_0/p_1))$
- p_i = proportion de cas $Y=i$ dans la population totale
- $P_{i,E}$ = proportion de cas $Y=i$ dans l'échantillon E
- Ceci est vrai de logit mais non de probit !

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

131

Influence de l'échantillonnage 2/2

- Si E est indépendant de X, la même fonction de score permet de décider si $Y=1$ (en changeant seulement le seuil de décision)
 - cas particulier : $p_{1,E}/p_{0,E} = p_1/p_0 \Rightarrow \beta'_0 = \beta_0$
- Un score calculé sur une sous-population E peut s'appliquer à une sous-population E', si la distribution des variables explicatives est la même dans E et E', même si l'événement à prédire est plus rare dans E'
 - en appliquant le calcul de $P(Y=1/X=x, X \in E)$ aux $X \in E'$ et en fixant le même seuil d'acceptation $P(Y=1/X=x, X \in E) > s_0$, on aura le même % d'acceptés dans E' (puisque les var. explicatives ont mêmes distributions dans E et E'), mais la fréquence de l'événement sera plus faible dans les acceptés de E', puisque leur proba $P(Y=1/X=x, X \in E') < P(Y=1/X=x, X \in E)$

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

132

Avantages de la régression logistique

- Odds-ratios facilement interprétables (pour modèle *logit*)
- Permet de traiter les variables explicatives discrètes, qualitatives ou continues
- Permet de traiter une variable cible ordinale ou nominale
- Hypothèses + générales que l'analyse discriminante (pas de multinormalité ni d'homoscédasticité)
- Permet de détecter certains phénomènes non linéaires
- Modélise directement une probabilité
- Fournit des intervalles de confiance
- Nombreux tests statistiques disponibles
- Possibilité de sélection pas à pas des variables
- Technique implémentée dans de nombreux logiciels

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

133

Inconvénients de régression logistique

- Suppose la non-colinéarité des variables explicatives
- Approximation numérique :
 - calcul itératif moins rapide que le calcul direct de l'analyse discriminante
 - moindre précision que l'analyse discriminante quand les hypothèses de cette dernière sont satisfaites
 - ne converge pas toujours vers une solution optimale
 - inopérant dans le cas de la séparation complète des groupes ! puisque la log-vraisemblance s'approche de 0 (iris de Fisher et séparation des Setosa !)
- Ne s'applique qu'aux données sans valeurs manquantes
- Sensible aux individus hors norme (sauf discrétisation)

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

134

La régression logistique ordinaire 1/2

- La variable Y à prédire est ordinale
- Fonctions de lien :
 - logit
 - probit : coefficients moins facilement interprétables que le logit
 - log-log : $\text{Log} [-\text{Log}(1-\mu)]$
 - utilisé quand les valeurs élevées de la cible sont plus probables
 - ex : valeurs 3 à 5 / 5 dans une enquête de satisfaction
 - Cauchit : $\text{tg}[\pi(\mu - 0,5)]$
 - utilisé quand les valeurs extrêmes de la cible sont plus probables
 - ex : valeur 5 / 5 dans une enquête de satisfaction

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



135

La régression logistique ordinaire 2/2

- Y prend m valeurs ordonnées, notées 1, 2, ..., m
- Dans le modèle à pentes égales : on suppose que le logit des probas cumulatives s'écrit sous la forme

$$\text{logit}(\text{Prob}(Y \leq r / X = x)) = a_r + \sum_i b_i x_i, \text{ pour } 1 \leq r < m$$

- On parle de « proportional odds model » car :

$$\frac{\text{Prob}(Y \leq r / X = x) / \text{Prob}(Y > r / X = x)}{\text{Prob}(Y \leq r / X = x') / \text{Prob}(Y > r / X = x')} = \frac{\exp(a_r + \sum_i b_i x_i)}{\exp(a_r + \sum_i b_i x'_i)} = \exp\left(\sum_i b_i (x_i - x'_i)\right)$$

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



136

Techniques de classement :

Le modèle linéaire général

Le modèle linéaire généralisé

Le modèle additif généralisé

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



137

Terminologie

- Covariables = variables explicatives continues (quantitatives)
- Facteurs = variables explicatives catégorielles (qualitatives)
 - niveaux d'un facteur = ses modalités

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



138

Effets fixes et aléatoires 1/2

- Effets fixes des facteurs et covariables
 - contrôlés par l'expérimentateur
 - en prenant toutes les valeurs
 - dont on veut quantifier l'effet sur la variable cible
 - similaire à une analyse de régression – but prédictif
- Effets aléatoires des facteurs et covariables
 - en prenant un échantillon de valeurs
 - on veut quantifier la proportion de la variance de la variable cible qu'ils expliquent
 - similaire à une analyse de corrélation
 - à but descriptif et non prédictif
- Effets mixtes
 - Présence d'effets fixes et aléatoires

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



139

Effets fixes et aléatoires 2/2

- Ex 1 : comparaison de 2 traitements sur plusieurs patients dans plusieurs hôpitaux – mettre la variable « hôpital » en effet aléatoire
 - permet d'éviter le biais dû au lieu où est administré le traitement
 - ne permet pas de prédire le résultat dans un nouvel hôpital
- Ex 2 : comparaison de 2 conditionnements d'un produit sur les achats de plusieurs consommateurs dans plusieurs magasins – mettre la variable « magasin » en effet aléatoire
 - permet d'éviter le biais dû au lieu d'achat
 - ne permet pas de prédire les achats dans un nouveau magasin

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



140

Modèle à mesures répétées 1/2

- Les mesures $y_1, y_2, \dots, y_k$ de Y à prédire sur plusieurs individus sont corrélées (données longitudinales) car
 - il s'agit d'un même individu observé k fois (par ex : avant et après un traitement médical)
 - ou de k individus partageant une caractéristique commune (même famille, même segment)
- On sort des hypothèses de la régression linéaire et de la régression logistique qui supposent l'absence de corrélation des mesures sur plusieurs individus
- Y peut être continue ou discrète
- Un modèle à mesures répétées peut traiter à la fois des effets fixes et aléatoires

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



141

Modèle à mesures répétées 2/2

- Dans un modèle à mesures répétées, on a des effets :
 - intra-individus (« within-subject effects »)
 - influence du temps, du traitement (comparaison du patient avant et après traitement)
 - généralise la comparaison de moyennes sur 2 échantillons appariés
 - inter-individus (« between-subject effects »)
 - influence des caractéristiques du patient, telles que l'âge, le sexe, la formulation sanguine... (comparaison du patient par rapport aux autres)
 - interactions intra-inter (« within-subject-by-between-subject effects »)
 - interactions du traitement et des caractéristiques du patient

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



142

Application aux données de survie 1/2

- Pour chaque individu, les observations sont répétées dans le temps à des instants $t_1, t_2, \dots, t_N$
- On s'intéresse à la survenue d'un événement (par ex : décès, départ) à un instant t_i , modélisée par la var cible :
 - $y_k = 0$ si $k < i$, $y_i = 1$, pas d'observation si $k > i$: on connaît le délai de survenue de l'événement
 - on a $y_k = 0$ pour tout $k \leq N$ si l'événement ne survient pas (et si l'individu est observé jusqu'au bout) : on ne connaît que la limite inférieure du délai de survenue de l'événement (cette donnée est « censurée »)
 - la donnée est aussi censurée si l'individu est perdu de vue avant la fin et avant la survenance de l'événement
- On cherche à expliquer la variable « délai de survie » pour mettre en évidence les facteurs favorables

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

143

Application aux données de survie 2/2

- Un modèle courant (de Cox)

À COMPLETER !

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

144

Modèle de survie de Kaplan-Meier

- Modélise la durée avant l'apparition d'un événement (décès, départ...)
- Certaines données sont censurées (encore vivant), mais on doit en tenir compte (les durées de vie + longues étant + censurées, par définition)
- On cherche des modèles intégrant à la fois les données censurées et non censurées
- Le modèle de Kaplan-Meier permet de calculer une estimation non paramétrique de la fonction de survie : $S(t) = \text{Prob}(\text{durée de vie} > t)$
- Il permet de comparer les fonctions de survie (et les courbes de survie) de plusieurs échantillons (« strates »)
 - correspondant par ex. à plusieurs traitements médicaux différents
 - et d'effectuer des tests

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



145

Modèle de survie de Cox 1/3

- Même champ d'application que le modèle de Kaplan-Meier
- Le modèle de régression de Cox à hasards proportionnels (1972) permet d'ajouter p variables explicatives et d'estimer leurs coefficients dans la fonction de survie, donc leur impact sur la durée de vie
 - ex : sexe / nb cigarettes fumées par jour
- C'est un modèle semi-paramétrique (forme paramétrique (Σ) pour les effets des var. explicatives, et forme non paramétrique de la fonction de survie)
- Pour tout individu i de var. explicatives x_{ij} , la fonction de survie s'exprime sous la forme :

$$S(t, x_i) = S_0(t) \exp\left(\sum_{j=0}^p b_j x_{ij}\right)$$
 - où $x_{i0} = 1 \forall i$ et $S_0(t)$ est la fonction de survie de base (« hasard de base »),
 - et où l'on recherche le vecteur β supposé indépendant de i .

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



146

Modèle de survie de Cox 2/3

- On trouve le vecteur β des coefficients de régression par maximisation d'une fonction de vraisemblance (comme la régression logistique)
 - plusieurs méthodes de sélection des var. explicatives existent (ascendante, descendante, pas à pas)
 - interprétation des « odds-ratios »
- Les données censurées :
 - n'interviennent pas dans le calcul de β
 - interviennent dans le calcul de $S_0(t)$
- Le terme de « hasards proportionnels » vient de ce que le rapport $h_i(t) / h_k(t)$ ne dépend pas de t
 - sauf si les x_{ij} en dépendent

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



147

Modèle de survie de Cox 3/3 (fonctionnalités supplémentaires)

- Les variables explicatives x_{ij} peuvent dépendre ou non de t
 - soit en étant une fonction de t , soit en prenant une valeur différente par valeur de t
- On peut faire des analyses stratifiées (sur des échantillons différents), en supposant que le vecteur β des coefficients de régression est indépendant de l'individu i et de la strate
 - en revanche, le hasard de base $S_0(t)$ dépend de la strate
 - d'où l'utilisation des analyses stratifiées sur une strate X_j quand une variable explicative X_j ne satisfait pas l'hypothèse des hasards proportionnels
 - X_j n'intervient plus dans le terme $\exp()$ mais intervient dans $S_0(t)$

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



148

Modèle linéaire général (GLM)

- Généralise la régression linéaire multiple de plusieurs façons
- Les variables explicatives peuvent non seulement être continues, mais :
 - catégorielles (ANOVA)
 - continues et catégorielles (ANCOVA)
- Il peut y avoir plusieurs variables continues à expliquer
 - MANOVA, MANCOVA
- Prise en compte des modèles à effet fixes, aléatoires ou mixtes
- Prise en compte des modèles à mesures répétées



Modèle linéaire généralisé (GLZ)

- Généralise le modèle linéaire général quand Y à prédire n'est plus forcément continue
- On écrit $g(E(Y/X=x)) = \beta_0 + \sum_i \beta_i x_i$
- g = fonction de lien monotone différentiable (g^{-1} = fonction de transfert)
- La distribution de $Y/X=x$ peut être :
 - normale (continue : régression) $g(\mu) = \mu$
 - gamma (continue positive) $g(\mu) = -1/\mu$
 - Bernoulli (discrète : oui/non) $g(\mu) = \log(\mu/(1-\mu))...$
(logit, probit, log-log)
 - de Poisson (discrète : comptage) $g(\mu) = \log(\mu)$
 - Y = nb de sinistres (assurance) ou effectif d'un tableau de contingence (modèle log-linéaire)
 - multinomiale, etc.



Modèle linéaire généralisé (GLZ)

- Double généralisation du modèle linéaire général : loi de $Y/X=x$ non normale et $g \neq 1$
- Estimation du modèle : par la méthode du maximum de vraisemblance (analogue des moindres carrés)
- Évaluation du modèle : par calcul de la déviance des log-vraisemblances (analogue de la somme des carrés des résidus de la régression) et test du chi-2
- Existence d'une régression logistique à mesures répétées (proc GENMOD de SAS)
- Variable V « offset » : sert à tarer un modèle si la variable cible dépend linéairement de V
 - le nb de sinistres dans une compagnie d'assurance doit être équilibré par la variable offset « nb de contrats »
- Source : Nelder-Wedderburn (1972)

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



151

Modèle additif généralisé (GAM)

- On écrit $g(E(Y/X=x)) = \beta_0 + \sum_i f_i(x_i)$
- g : fonction de lien (g^{-1} : fonction de transfert)
- f_i : fonction quelconque (non-paramétrique : on n'a plus un simple paramètre comme le coefficient β_i) de x_i
 - par ex : f_i = fonction spline
- Mais le modèle reste additif (c'est $\sum_i$ qui combine les f_i)
- La distribution de Y peut être normale, poissonienne ou binomiale
 - ex : modèle logistique additif généralisé si $g(\mu) = \log(\mu/1-\mu)$
- Modélisation puissante mais attention au sur-apprentissage et à l'interprétabilité des résultats
- Source : Hastie - Tibshirani (1990)

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



152

Technique de classement ou prédiction : Arbres de décision

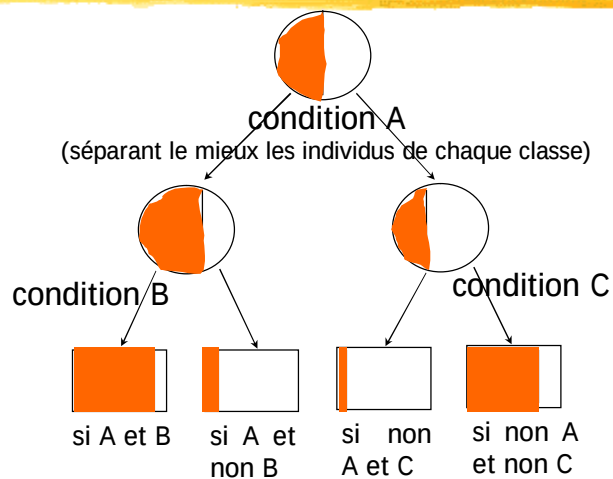
19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



153

Classement par arbre de décision



19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



154

Les arbres de décision : principe

- Pour répartir les individus d'une population en n classes, on commence par choisir la variable séparant le mieux les individus de chaque classe, en sous-populations appelées *nœuds* : le critère précis (C1) de choix de la variable et de sa valeur testée dépend de chaque type d'arbre
- Pour chaque nœud, on répète la même opération, ce qui donne naissance à un ou plusieurs *nœuds fils*. Chaque nœud fils donne à son tour naissance à un ou plusieurs nœuds, et ainsi de suite, jusque ce que :
 - la séparation des individus ne soit plus possible
 - OU un certain critère (C2) d'arrêt d'approfondissement de l'arbre soit satisfait

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



155

Les arbres de décision : Définitions

- Si chaque nœud de l'arbre a au plus deux nœuds fils, on dit que l'arbre est *binaire*
- Le premier nœud de l'arbre est la *racine*
- Les nœuds terminaux sont les *feuilles*
- Le chemin entre la racine et chaque feuille est l'expression d'une règle.
 - par exemple : les clients dont l'âge est $< x$, les revenus $< y$ et le nombre de comptes $> z$ appartiennent dans n % des cas à la classe C
- *Densité* d'un nœud = rapport de son effectif à l'effectif total de la population
 - une valeur courante pour la densité minimale des feuilles d'un arbre est 1 ou 2 %.

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



156

Critère d'arrêt d'un arbre (C2)

- Le critère d'arrêt (C2) dépend du type et du paramétrage de l'arbre. Souvent (C2) combine plusieurs règles :
 - la profondeur de l'arbre a atteint une limite fixée
 - OU le nombre de feuilles (c'est-à-dire de règles) a atteint un maximum fixé
 - OU l'effectif de chaque nœud est inférieur à une valeur fixée (au moins 50 à 100 pour de bons résultats)
 - ou la division de chaque nœud provoque la naissance d'un fils d'effectif inférieur à une valeur fixée
 - OU la qualité de l'arbre est suffisante
 - OU la qualité de l'arbre n'augmente plus de façon sensible.
- C'est bien entendu sur cette dernière règle que les arbres diffèrent le plus
 - exemple : dans les arbres CART la qualité se mesure par la pureté des nœuds

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

157

Principaux critères (C1)

- Le critère du χ^2
 - lorsque les variables explicatives sont catégorielles
 - utilisé dans l'arbre CHAID
- L'indice de Gini, l'indice Twoing et l'entropie
 - pour tous types de variables explicatives
 - l'indice de Gini est utilisé dans l'arbre CART
 - l'indice Twoing est utilisé dans l'arbre CART lorsque la variable cible a ≥ 3 modalités
 - l'entropie est utilisée dans les arbres C4.5 et C5.0
 - plus les classes sont uniformément distribuées dans un nœud, plus l'indice de Gini et l'entropie sont élevés ; plus le nœud est pur, plus ils sont bas

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

158

Les principaux arbres de décision

- CHAID (CHi-Square Automation Interaction Detection)
 - utilise le test du χ^2 pour définir la variable la plus significative et le découpage de ses modalités
 - adapté à l'étude des variables explicatives discrètes
- CART (Classification and Regression Tree)
 - cherche à maximiser la pureté des nœuds
 - adapté à l'étude de tout type de variables explicatives
- C5.0 de J.R. Quinlan
 - cherche à maximiser le gain d'information réalisé en affectant chaque individu à une branche de l'arbre
 - adapté à l'étude de tout type de variables explicatives.

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



159

Arbre CHAID 1/2

- Cet arbre est de conception plus ancienne (principe : 1975, Hartigan ; algorithme : 1980, Kass)
- Il discrétise les variables explicatives continues
- La cible est une variable qualitative à k modalités
- Utilise le χ^2 pour regrouper les modalités de chaque variable dans un nœud, en sélectionnant itérativement la paire de modalités dont le sous-tableau (2 x k) associé a le plus petit χ^2 et en les fusionnant si ce χ^2 n'est pas significatif
- Après avoir regroupé les modalités de façon significative, CHAID réutilise le χ^2 pour retenir la variable la plus significative du nœud (plus faible probabilité associée au χ^2 , si cette proba est < à un seuil fixé)

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



160

Arbre CHAID 2/2

- Il traite l'ensemble des valeurs manquantes comme une seule catégorie
 - pas d'utilisation de variables de substitution
- Il n'est pas forcément binaire (contrairement à CART) et produit des arbres souvent plus larges que profonds
 - utile pour la discrétisation de variables continues
- Il souffre de l'absence de dispositif automatique d'optimisation par élagage : quand l'arbre maximum est élaboré, les critères d'arrêt étant rencontrés, sa construction s'achève



Discrétisation avec CHAID 1/4

- Supposons que nous voulions prédire une variable cible à l'aide de certaines variables, dont l'âge, et que nous voulions découper l'âge en classes pour les raisons déjà indiquées :
 - prise en compte de la non-linéarité de la réponse en fonction de l'âge
 - suppression du problème des extrêmes
 - modèle plus robuste
- Nous allons découper l'âge en 10 tranches (ou plus, si le nb d'individus est grand) et regarder le % d'individus dans la cible pour chaque classe d'âge



Discrétisation avec CHAID 2/4

			cible		Total
			non	oui	
tranche d'âge	18-25 ans	Effectif	127	81	208
		% dans tranche d'âge	61,1%	38,9%	100,0%
	25-29 ans	Effectif	104	126	230
		% dans tranche d'âge	45,2%	54,8%	100,0%
	29-32 ans	Effectif	93	101	194
		% dans tranche d'âge	47,9%	52,1%	100,0%
	32-35 ans	Effectif	113	99	212
		% dans tranche d'âge	53,3%	46,7%	100,0%
	35-38 ans	Effectif	93	94	187
		% dans tranche d'âge	49,7%	50,3%	100,0%
	38-40 ans	Effectif	149	123	272
		% dans tranche d'âge	54,8%	45,2%	100,0%
	40-42 ans	Effectif	108	72	180
		% dans tranche d'âge	60,0%	40,0%	100,0%
	42-45 ans	Effectif	116	97	213
		% dans tranche d'âge	54,5%	45,5%	100,0%
	45-51 ans	Effectif	77	113	190
		% dans tranche d'âge	40,5%	59,5%	100,0%
	> 51 ans	Effectif	71	145	216
		% dans tranche d'âge	32,9%	67,1%	100,0%
Total			1051	1051	2102
			50,0%	50,0%	100,0%

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

163

Discrétisation avec CHAID 3/4

- Nous voyons que certaines classes sont proches du point du vue du % dans la cible :

- tranches 2 et 3
- tranches 4 à 8
- tranches 9 et 10

Noeud 0		
Catégorie	%	n
O	50,00	1051
N	50,00	1051
Total	(100,00)	2102

AGE
Prob. ajustée - valeur=0,0000, Khi-deux=50,4032, ddl=3

<=24			(24,32]			(32,45]			>45		
Noeud 6			Noeud 7			Noeud 8			Noeud 9		
Catégorie	%	n	Catégorie	%	n	Catégorie	%	n	Catégorie	%	n
O	38,94	81	O	53,54	227	O	45,58	485	O	63,55	258
N	61,06	127	N	46,46	197	N	54,42	579	N	36,45	148
Total	(9,90)	208	Total	(20,17)	424	Total	(50,62)	1064	Total	(19,31)	406

- Nous voyons que CHAID a fait automatiquement ce que nous avons fait manuellement

19/08/2005

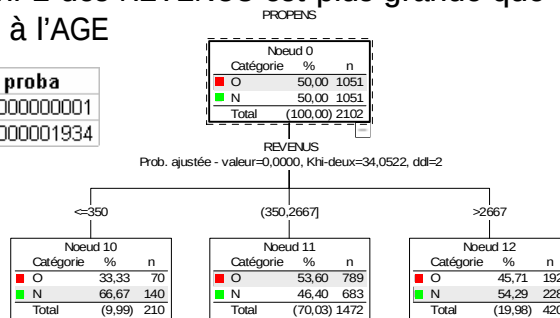
© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

164

Discrétisation avec CHAID 4/4

- Pour la scission de la racine de l'arbre, la variable AGE est retenue devant la variable REVENUS car la proba associée au chi-2 des REVENUS est plus grande que celle associée à l'AGE

variable	chi-2	ddl	proba
âge	50,40	3	0,0000000001
revenus	34,05	2	0,0000001934



- NB : si le nb de ddl n'est pas le même pour 2 variables, il faut comparer les probas et non les chi-2 eux-mêmes

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

165

Indice de Gini

- Indice de Gini d'un nœud = $1 - \sum_i f_i^2$
 - où les f_i , $i = 1$ à p , sont les fréquences relatives dans le nœud des p classes à prédire (variable cible)
 - probabilité que 2 individus, choisis aléatoirement dans un nœud, appartiennent à 2 classes différentes
- Plus les classes sont uniformément distribuées dans un nœud, plus l'indice de Gini est élevé ; plus le nœud est pur, plus l'indice de Gini est bas
- Chaque séparation en k nœuds fils (d'effectifs $n_1, n_2, \dots, n_k$) doit provoquer la plus grande hausse de la pureté, donc la plus grande baisse de l'indice de Gini. Autrement dit, il faut minimiser :

$$\text{Gini (séparation)} = \sum_{i=1}^k \frac{n_k}{n} \text{Gini}(k^{\text{e}} \text{ nœud})$$

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

166

Arbre CART 1/2

- Il est binaire
 - il est moins large que profond, mais parfois trop profond
- Variable cible quantitative ou qualitative $\Rightarrow$ CART sert à la prédiction comme au classement
- Quand dans une feuille, la proportion d'individus dans chaque classe C est notée p_C , la classe C pour laquelle p_C est maximale est considérée comme « la classe » de cette feuille, et les individus de cette feuille sont considérés comme classés dans C
- La proportion $p_C = \text{pureté}$ du nœud, quantité d'autant plus petite dans une partie de la population que toutes les classes y sont uniformément distribuées
- Gère les valeurs manquantes en recourant aux variables équiréductrices
 - différent de CHAID

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



167

Arbre CART 2/2

- Un nœud est considéré comme une feuille lorsque aucune séparation ne conduit à une hausse significative de la pureté
- Une fois l'arbre maximum construit, l'algorithme en déduit plusieurs sous-arbres par élagages successifs, qu'il compare entre-eux, avant de retenir celui pour lequel le taux d'erreur mesuré en test est le plus bas possible
- CART permet la prise en compte de coûts C_{ij} de mauvaise affectation (d'un individu de la classe j dans la classe i) en les intégrant dans le calcul de l'indice de Gini
 - $$\text{Gini}(\text{nœud}) = \sum_{i \neq j} C_{ij} f_i f_j$$

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



168

Mécanisme de scission des nœuds avec Gini (ex : catalogue avec prix article et achat)

Article	Prix	Achat
1	125	N
2	100	N
3	70	N
4	120	N
5	95	O
6	60	N
7	220	N
8	85	O
9	75	N
10	90	O

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



169

Mécanisme de scission des nœuds avec Gini

Achat	N		N		N		O		O		O		N		N		N		N			
Prix	60		70		75		85		90		95		100		120		125		220			
Seuil	55		65		72		80		87		92		97		110		122		172		230	
	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>
O	0	3	0	3	0	3	0	3	1	2	2	1	3	0	3	0	3	0	3	0	3	0
N	0	7	1	6	2	5	3	4	3	4	3	4	3	4	4	3	5	2	6	1	7	0
Gini	0,420		0,400		0,375		0,343		0,417		0,400		0,300		0,343		0,375		0,400		0,420	

19/08/2005

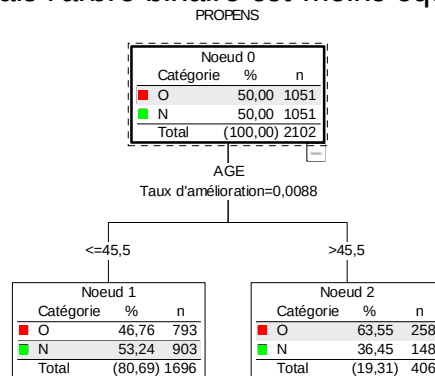
© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



170

Exemple précédent avec CART

- La scission de la racine se fait par l'AGE, comme avec CHAID, mais l'arbre binaire est moins équilibré :



- CART est surtout apte à détecter rapidement des profils très marqués

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



171

Entropie

- Entropie (ou « information ») d'un nœud = $\sum f_i \cdot \log(f_i)$
 - où les f_i , $i = 1$ à p , sont comme ci-dessus les fréquences relatives dans le nœud des p classes à prédire
- Plus les classes sont uniformément distribuées dans un nœud, plus l'entropie est élevée ; plus le nœud est pur, plus l'entropie est basse (elle vaut 0 lorsque le nœud ne contient qu'une seule classe)
- Comme précédemment, il faut minimiser l'entropie dans les nœuds-fils

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



172

Arbre C5.0

- C5.0 (successeur de C4.5) est adapté comme CART à tout type de variables
- Dispositif d'optimisation de l'arbre par construction puis élagage d'un arbre maximum
 - le procédé d'élagage est différent de celui de CART
- C5.0 fonctionne en cherchant à maximiser le gain d'information réalisé en affectant chaque individu à une branche de l'arbre
- C5.0 n'est pas binaire. Les variables catégorielles, au niveau d'un nœud père, donnent naissance à un nœud fils par modalité
 - inconvénient : les nœuds voient plus rapidement leurs effectifs baisser (moindre fiabilité statistique)

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



173

CART et complexité du choix (C1)

- Si une **variable explicative catégorielle** X a un ensemble E de n valeurs possibles $x_1, \dots, x_n$, toute condition de séparation sur cette variable sera de la forme
 - $X \in E'$, où $E' \subset E - \{0\}$
- > $2^{n-1} - 1$ conditions de séparation possibles
- Pour une **variable explicative continue** X , la complexité est liée au tri des valeurs $x_1, \dots, x_n$ de X , puisqu'une fois les variables dans l'ordre $x_1 \leq \dots \leq x_n$,
- il suffit de trouver l'indice k tel que la condition
 - $X \leq \text{moyenne}(x_k, x_{k+1})$
- soit la meilleure (selon le critère choisi, par exemple Gini).

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



174

Traitements des valeurs manquantes

- Variables équidivisantes :
 - celles qui assurent (à peu près) la même pureté des nœuds que la variable optimale
- Variables équiréductrices :
 - celles qui répartissent les individus (à peu près) de la même façon que la variable optimale
- Ces variables servent de variables « de rechange » lorsque la variable optimale a une valeur manquante.
 - Par cohérence, il vaut mieux utiliser les variables équiréductrices

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



175

Pré-élagage et Post-élagage

- Certains arbres (CHAID) effectuent un pré-élagage :
 - si un approfondissement d'une branche dégrade la qualité de l'arbre : on s'arrête là.
- D'autres arbres (CART, C5.0) effectuent un post-élagage :
 - l'approfondissement de l'arbre est mené à son terme
 - AVANT d'élaguer l'arbre
 - > ce qui est + efficace, car parfois le sur-apprentissage a commencé avant d'être détecté par le taux d'erreur en test
 - > l'arbre peut ainsi découvrir des informations importantes plus profondes que ce que montre un élagage prématuré.

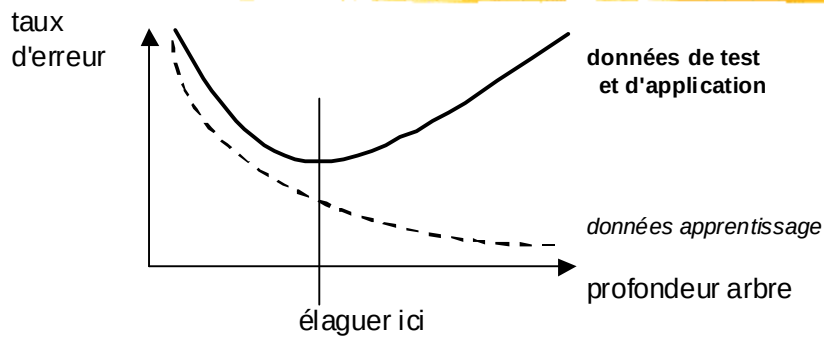
19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



176

Élagage et sur-apprentissage



- Un bon arbre doit être élagué pour éviter la remontée du taux d'erreur due au sur-apprentissage.

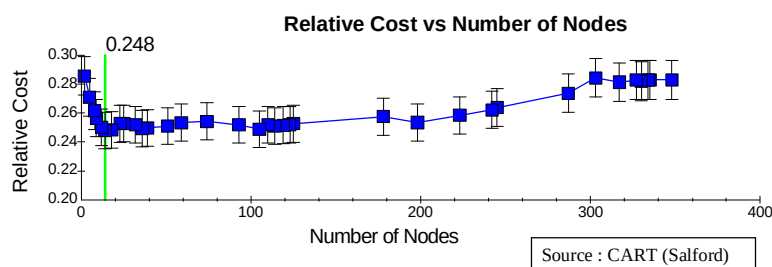
19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



177

Choix du bon niveau d'élagage



19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



178

Validation croisée

- Lorsque la population est trop petite pour en extraire un échantillon d'apprentissage et un de test (courant en pharmacie) :
- On a recours à la **validation croisée** (leave-one-out)
 - La population est scindée en, disons, 10 échantillons de tailles égales, ayant chacun la même distribution pour la classe ou la variable à prédire.
 - On utilise les 9 premiers échantillons comme échantillon d'apprentissage, et le 1/10^e restant comme échantillon de test. On obtient ainsi un taux d'erreur en test.
 - On répète ensuite 9 fois la même opération sur chaque 9/10^e possible, en prenant chaque 1/10^e restant pour échantillon de test.
 - On combine enfin les 10 taux d'erreur obtenus.

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



179

Prédiction par arbre de décision

- Les arbres peuvent s'appliquer à la prédiction : la variable à expliquer X est continue.
- C'est une alternative à la régression linéaire multiple.
- Principe :
 - la variable X doit avoir une variance + faible dans les nœuds fils
 - la variable X doit avoir une moyenne la + distincte possible d'un nœud fils à un autre.

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



180

Avantages des arbres de décision 1

- Ils fournissent des règles :
 - explicites (contrairement aux réseaux de neurones)
 - qui s'écrivent directement avec les variables d'origine (contrairement aux méthodes factorielles)
- Méthode non paramétrique, non perturbée par :
 - la distribution non linéaire ou non monotone des prédicteurs par rapport à la variable cible
 - la colinéarité des prédicteurs
 - les interactions entre les prédicteurs
 - les individus « hors-normes » (isolés dans des règles spécifiques)
 - les fluctuations des prédicteurs non discriminants (l'arbre sélectionne les plus discriminantes)

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



181

Avantages des arbres de décision 2

- Beaucoup traitent (sans recodification) des données hétérogènes (numériques et non numériques, voire manquantes)
 - un arbre CART peut supporter 20 % de valeurs manquantes pour une variable, si la distribution de ces valeurs manquantes est aléatoire.
 - utilisation des variables équadivisantes
- Durée de traitement
 - leur apprentissage peut être un peu long, mais beaucoup moins que pour les réseaux de neurones
 - leur application est très rapide d'exécution

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



182

Inconvénients des arbres de décision

- Les nœuds du niveau $n+1$ dépendent fortement de ceux du niveau n
 - un arbre détecte des optimums locaux et non globaux
 - > la modification d'une seule variable, si elle est placée près du sommet de l'arbre, peut entièrement modifier l'arbre
 - les variables sont testées séquentiellement et non simultanément
 - > manque de robustesse
- L'apprentissage nécessite un nombre suffisant d'individus (pour avoir au moins 30 à 50 individus / nœud)
- Discontinuité de la réponse de la variable cible en fonction des variables explicatives (nb de valeurs du score = nb de feuilles)
- Valeurs du score non uniformément distribuées



Pour améliorer les résultats : Le rééchantillonnage



Rééchantillonnage Bootstrap

- Pour estimer un paramètre statistique dont on ne connaît pas la loi dans un échantillon de n individus
 - ou quand son calcul exige une distribution normale non vérifiée
- On l'approche par une suite de m (souvent $m \gg 30$ et même $m > 100$) tirages aléatoires de n individus avec remise
 - en mesurant le paramètre pour chaque échantillon simulé
 - puis en établissant la distribution des fréquences des valeurs de ce paramètre
 - puis en calculant l'intervalle de confiance du paramètre
 - $(2n-1)!/[n!(n-1)!]$ échantillons bootstrap différents
- Inventé par Bradley Efron (1979)

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



185

Agrégation de modèles : BAGGING

- BAGGING : bootstrap aggregating, Breiman, 1996
- Construction d'une famille de modèles sur n échantillons *bootstrap* (tirages avec remise)
- Ensuite agrégés par un vote ou une moyenne des estimations (ou une moyenne des probabilités en régression logistique)
- FORETS ALEATOIRES, Breiman, 2001
- = Bagging pour les arbres de décision en ajoutant un tirage aléatoire parmi les variables explicatives
- Évite de voir apparaître toujours les mêmes variables

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



186

Agrégation de modèles : BOOSTING

- BOOSTING, Freund et Shapire, 1996
- Version adaptative du Bagging (même principe de départ)
- Mais on travaille sur toute la population
- et à chaque itération, on augmente le poids des individus mal classés dans les itérations précédentes
 - tandis que le poids des bien classés n'augmente pas
 - > convergence plus rapide
- Avec CART, le nb de feuilles est à prendre dans $[4, 8]$ ou $= \sqrt{p}$, où p = nb de variables explicatives

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



187

Agrégation de modèles : ARCING

- ARCING : adaptative resampling and combining, Breiman, 1998
- Variante du boosting
- Mais on ne travaille que sur des échantillons :
 - au lieu de modifier les poids des individus de la population
 - on tire un nouvel échantillon selon des probabilités inversement proportionnelles à la qualité d'ajustement de l'itération précédente
- > Un individu qui a participé à un bon ajustement a moins de chances d'être tiré de nouveau

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



188

Agrégation de modèles : Conclusion

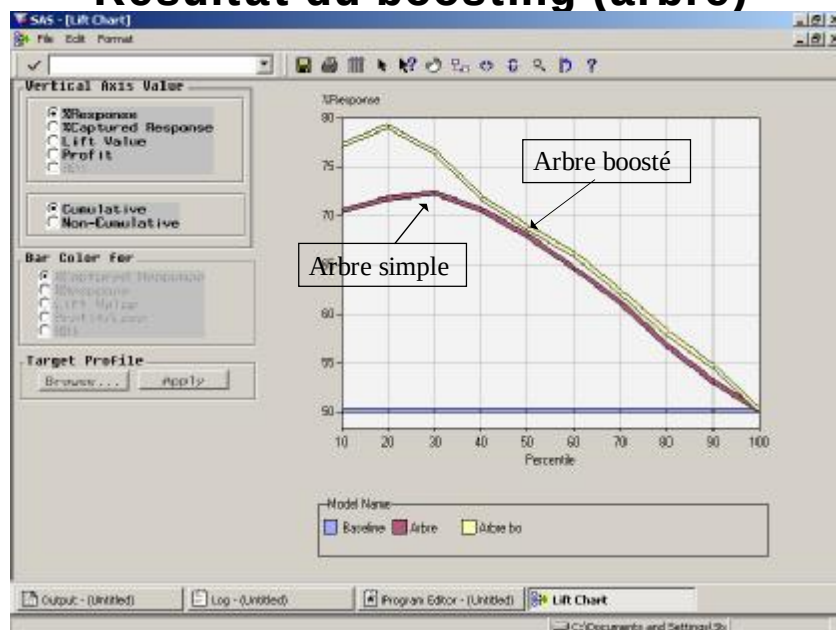
- Ces techniques permettent d'améliorer parfois très nettement la qualité (tx de biens classés) et la robustesse (sur un autre échantillon) des prédictions
 - même avec seulement une centaine d'itérations
 - mais surtout sur les arbres de décision ! et non sur l'analyse discriminante ou la régression logistique pour lesquelles le gain est faible (surtout avec le bagging)
- AVANTAGES
 - bonne résistance au bruit
 - bonne résistance au surapprentissage
- INCONVÉNIENTS
 - perte de lisibilité
 - importance du temps machine de traitement
- Objet de nombreux travaux théoriques en cours

19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

189

Résultat du boosting (arbre)



19/08/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

190