Analyse discriminante linéaire

Une approche pour rendre calculable P(Y/X)

Ricco RAKOTOMALALA

Théorème de Bayes

Probabilité conditionnelle

Estimer la probabilité conditionnelle

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

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

Déterminer la conclusion = déterminer le max.

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

$$\Leftrightarrow$$

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

Probabilité a priori
Estimé facilement par n_k/n

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

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

[1] Hypothèse de normalité

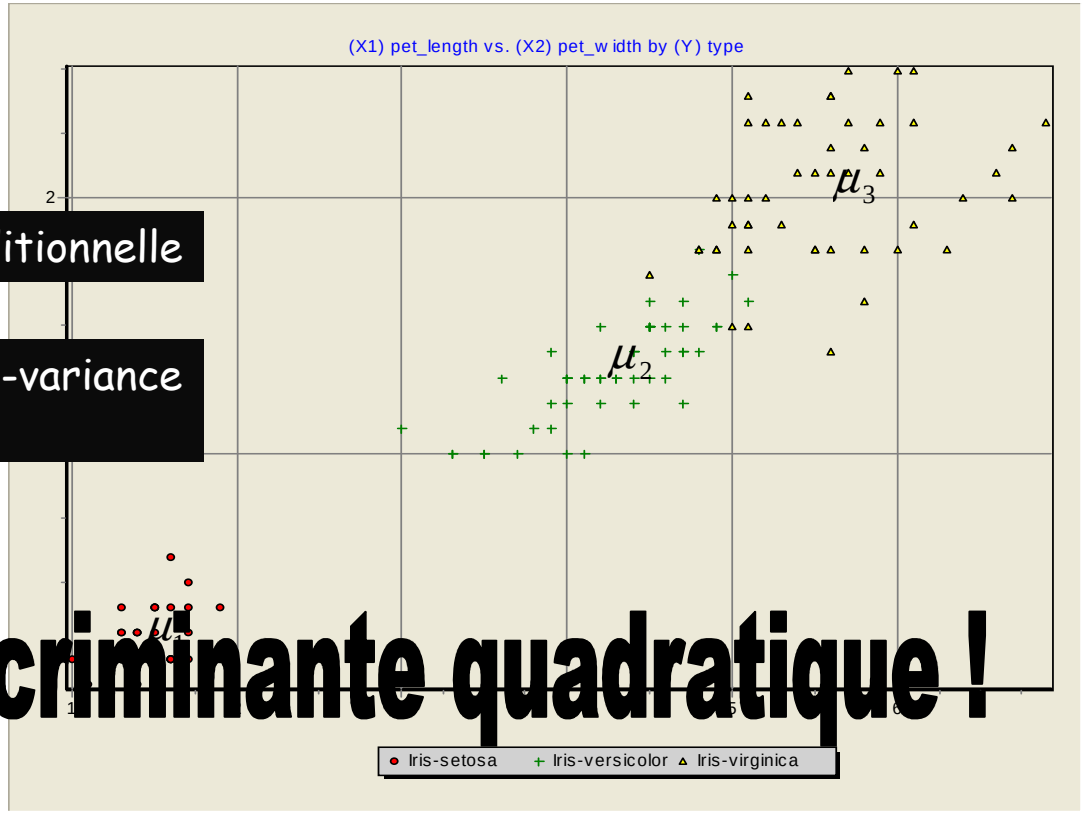
La normalité de la probabilité conditionnelle P(X/Y)

Loi normale multidimensionnelle

$$P\left(X_1 = v_1, \ldots, X_J = v_J / y_k\right) = \frac{1}{\sqrt{2\pi \det(\Sigma_k)}} e^{-\frac{1}{2}(X-\mu_k)\Sigma_k^{-1}(X-\mu_k)'}$$

μ_k Centre de gravité conditionnelle

Σ_k Matrice de variance co-variance conditionnelle

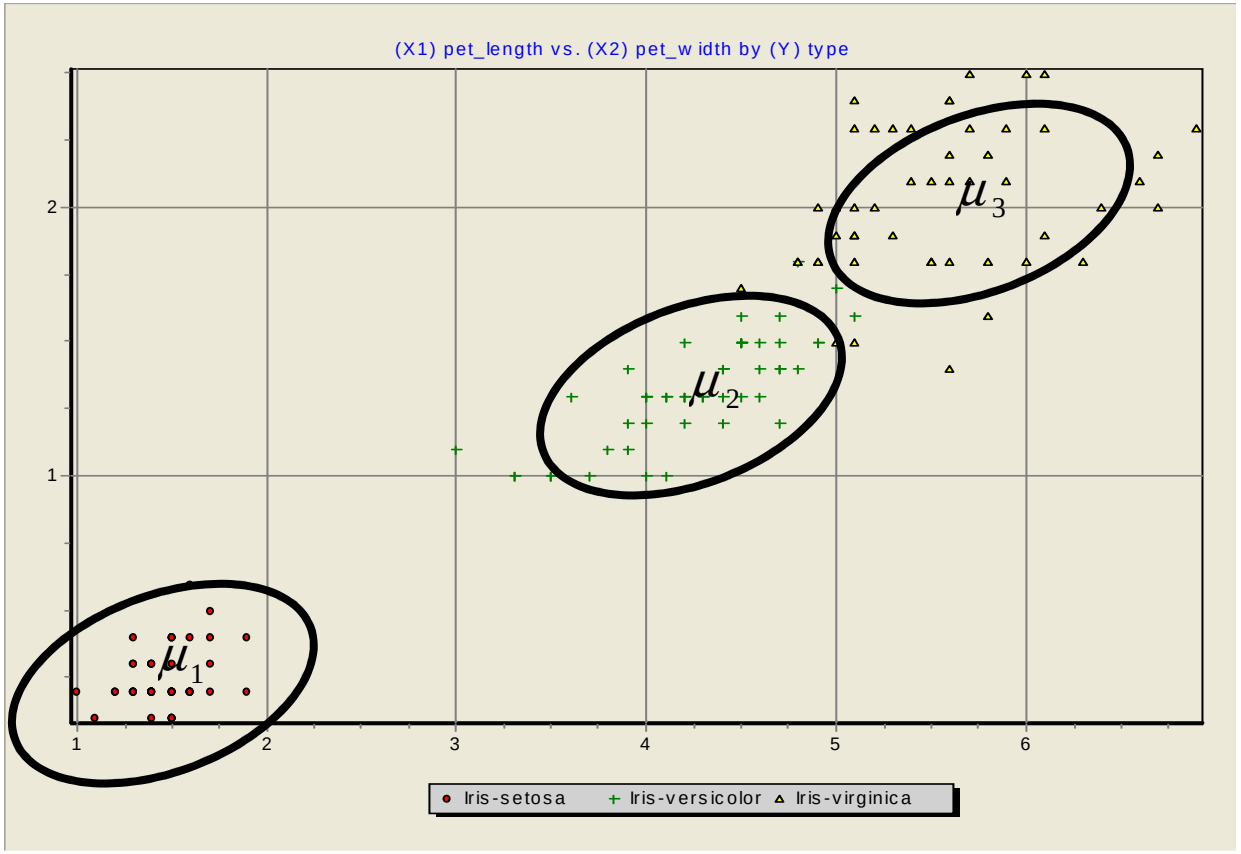


analyse discriminante quadratique !

[2] Homoscédasticité

Égalité des matrices de variance co-variance conditionnelles

$$\Sigma = \Sigma_k, k = 1, \ldots, K$$



analyse discriminante linéaire !

Fonction linéaire discriminante

Simplification des formules sous [1] et [2]

La probabilité conditionnelle est donc proportionnelle à

$$\ln P(X/y_k) \propto -\tfrac{1}{2}(X-\mu_k)\Sigma^{-1}(X-\mu_k)'$$

Avec les estimations sur l'échantillon de taille n, K classes et J descripteurs

$$\hat{\mu}_k = \begin{pmatrix} \overline{x}_{k,1} \\ \vdots \\ \overline{x}_{k,J} \end{pmatrix}$$

Moyennes conditionnelles

$$\hat{\Sigma} = \frac{1}{n-K}\sum_{k=1}^{K} n_k \times \hat{\Sigma}_k$$

Matrice de variance co-variance intra-classes

Fonction linéaire discriminante

Linéarité du modèle d'affectation

Fonction discriminante linéaire proportionnelle à $P(Y=y_k/X)$

$$d(Y_k, X) = \ln\left[P\left(Y = y_k\right)\right] + \mu_k \Sigma^{-1} X' - \frac{1}{2} \mu_k \Sigma^{-1} \mu_k'$$

Règle d'affectation

$$\left.\begin{array}{l} d(Y_1, X) = a_{1,0} + a_{1,1} X_1 + a_{1,2} X_2 + \cdots + a_{1,J} X_J \\ d(Y_2, X) = a_{2,0} + a_{2,1} X_1 + a_{2,2} X_2 + \cdots + a_{2,J} X_J \\ \ldots \end{array}\right\} \quad y_{k*} = \arg \max_k d(Y_k, X)$$

Avantages et inconvénients

ADL propose les mêmes performances que les autres méthodes linéaires
>> Elle est assez robuste par rapport à l'hypothèse de normalité
>> Elle est gênée par la forte violation de l'homoscédasticité (formes de nuages très différentes)
>> Elle est sensible à une forte dimensionnalité et/ ou la corrélation des descripteurs (inversion de matrice)
>> Elle n'est pas opérationnelle si la distribution est multimodale (ex. 2 ou + « blocs » de nuages pour Y=Yk)

Montrer exemple avec le fichier BINARY WAVES – Lecture des résultats et déploiement

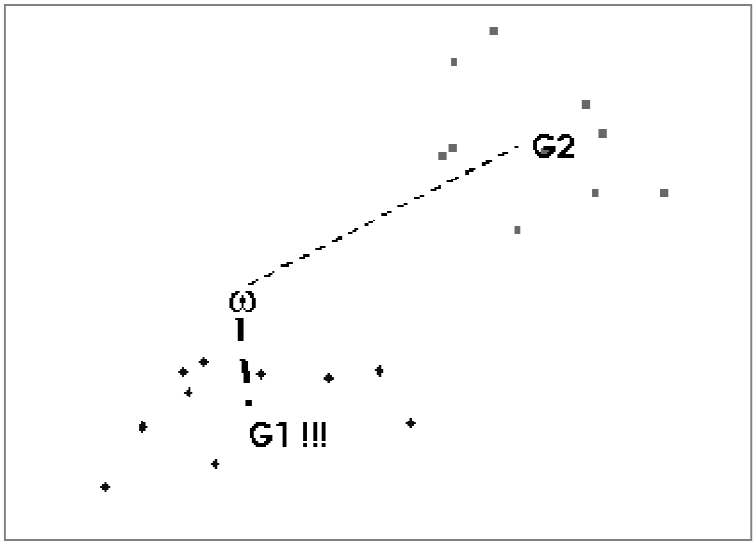
ADL

Caractéristiques géométriques de la règle d'affectation

d(Y_k,X) pour un individu à classer ω dépend de

$$(X(\omega)-\mu_k)\Sigma^{-1}(X(\omega)-\mu_k)'$$

Règle géométrique d'affectation : *Distance par rapport aux centres de gravité avec la métrique de Mahalanobis*

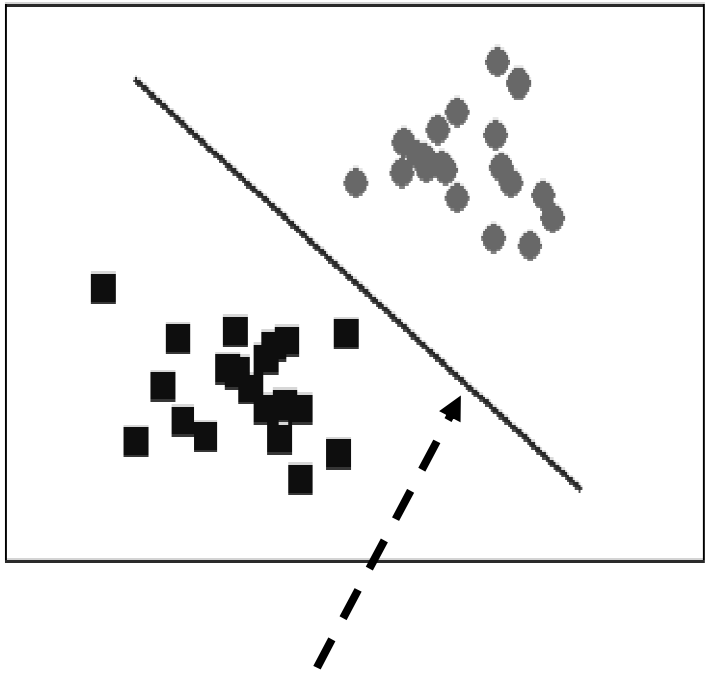


On comprend mieux les problèmes sous hétéroscédasticité (le calcul de la distance est biaisée), et lorsque la distribution est multimodale (le centre de gravité ne représente plus rien)

ADL

Interprétation sous la forme d'un apprentissage par paritionnement

Forme de la frontière entre les groupes
Séparation linéaire entre les groupes



ex. breast cancer = f (cellsize, cellshape)
dans Tanagra

Formé par l'ensemble des points équidistants
par rapport aux deux centres de gravité : c'est une droite !

ADL : Une règle d'affectation probabiliste peut s'interpréter comme une règle géométrique, elle peut également se voir comme une méthode de partitionnement de l'espace de représentation

ADL

Évaluation globale du modèle

(1) Évaluation standard d'un modèle de prédiction
Apprentissage + Test → Matrice de confusion

(2) Évaluation « statistique » utilisant l'hypothèse de normalité

$$H_0 : \mu_1 = \cdots = \mu_K$$
$$H_1 : \text{un au moins s'écarte des autres}$$

Statistique du test : LAMBDA de WILKS

$$\Lambda = \frac{\det(W)}{\det(V)}$$

Matrice de variance co-variance intra

Matrice de variance co-variance totale

! Voir l'analogie avec l'analyse de variance ANOVA (on parle alors de MANOVA)
Sa loi est peu tabulée, on utilise la transformation de RAO qui suit une loi de Fisher

ADL

Évaluation individuelle des coefficients (des variables)

On veut évaluer « la contribution » d'une variable dans le modèle de prédiction.

L'idée est de mesurer la variation du Λ de WILKS en passant du modèle à J variables au modèle à (J-1) variables c.-à-d. sans la variable que l'on cherche à évaluer.

La statistique du test

$$\frac{n-K-J+1}{K-1}\left(\frac{\Lambda_{J-1}}{\Lambda_J}-1\right) \cong F\left(K-1, n-K-J+1\right)$$

! Étant assez lourde à calculer, cette statistique est peu présente dans les logiciels *(on utilise alors un autre artifice)*

ADL

Cas particulier du problème à 2 classes (K = 2)

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

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

Règle d'affectation

D(X) > 0 → Y = +

Interprétation

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

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

>> Lorsque les Xj sont elles-mêmes des variables indicatrices (0/1), les coefficients « c » peuvent être lus comme des « points » attribués aux individus portant le caractère Xj

Évaluation

On peut voir une analogie forte entre la LDA et la régression linéaire multiple dans laquelle la variable endogène Y est codée 1/0 [1 = + ; 0 = -].

Nous pouvons nous inspirer des résultats de la régression pour évaluer globalement le modèle et surtout pour évaluer individuellement la « significativité » des coefficients « c » → test de Student (*attention, la transposition du test n'est pas totale*)

ADL

Mise en œuvre dans SPAD

(1) Uniquement les problèmes à deux classes
(2) Tous les descripteurs doivent être continus
(3) Évaluation des variables par analogie avec la régression linéaire multiple

$$D = d(Y_1 / X) - d(Y_2 / X)$$

FONCTION LINEAIRE DISCRIMINANTE

NUM	LIBELLES	CORRELATIONS VARIABLES AVEC F.L.D. (SEUIL= 0.08)	COEFFICIENTS FONCTION DISC.	COEFFICIENTS REGRESSION	ECARTS TYPES (RES. TYPE REG.)	T STUDENT	PROBA
1	clump	-0.716	-0.8867	-0.0693	0.0076	9.15	0.000
2	ucellsize	-0.818	-0.6081	-0.0475	0.0136	3.50	0.000
3	ucellshape	-0.819	-0.4381	-0.0342	0.0133	2.58	0.010
4	mgadhesion	-0.697	-0.1749	-0.0137	0.0085	1.61	0.107
5	sepics	-0.683	-0.2153	-0.0168	0.0111	1.51	0.131
6	bnuclei	-0.815	-1.2181	-0.0952	0.0068	13.97	0.000
7	bchromatin	-0.757	-0.5589	-0.0437	0.0107	4.07	0.000
8	normnucl	-0.712	-0.4619	-0.0361	0.0079	4.57	0.000
9	mitoses	-0.423	-0.0773	-0.0060	0.0106	0.57	0.569
CONSTANTE			19.606468	1.258487	0.0347	******	0.0000

R2 = 0.83612 F = 390.59235 PROBA = 0.000
D2 = 22.52031 T2 = 3556.14746 PROBA = 0.000

Indicateurs de « séparabilité » des groupes
F du Lambda de Wilks, T2 de Hotelling

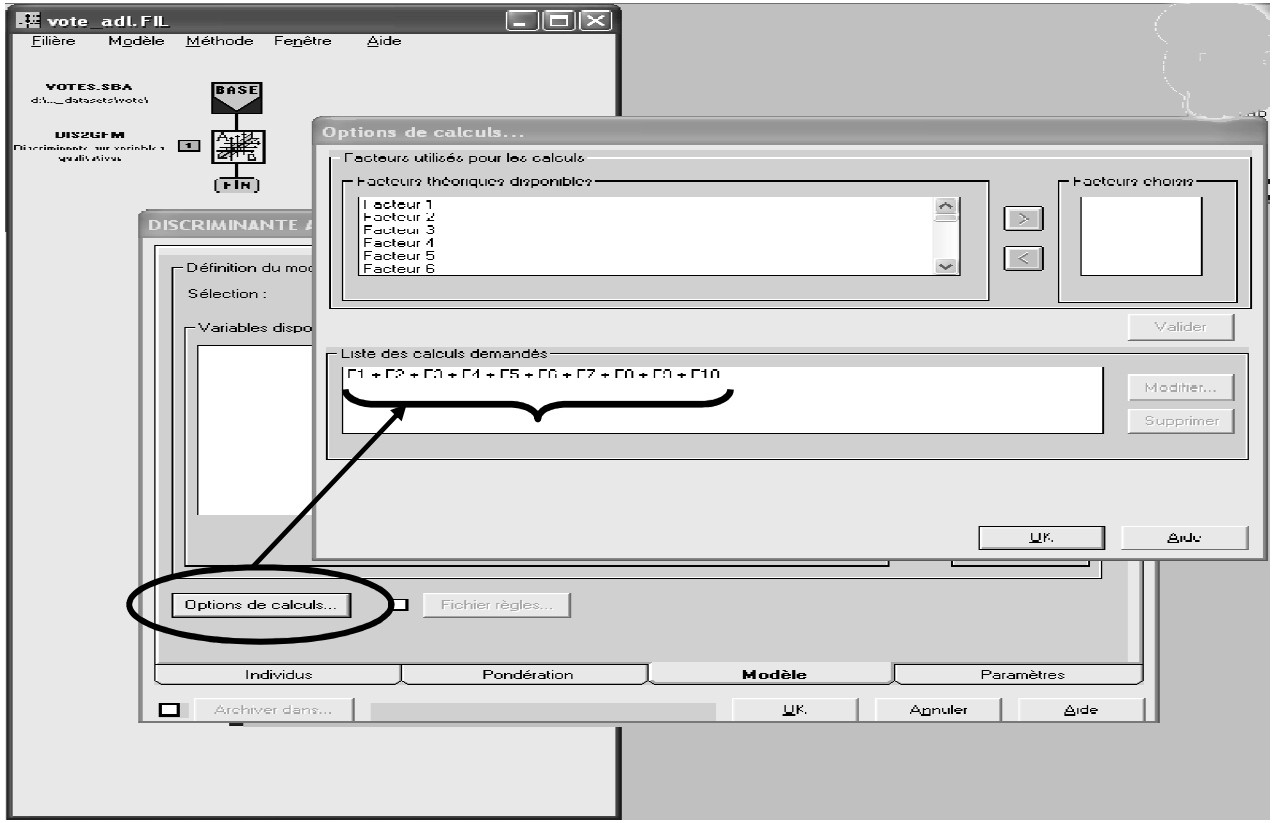
Résultats indicatifs de la régression
(pour l'évaluation individuelle des variables)

Voir le parallèle avec le « vrai » test du LAMBDA de WILKS dans TANAGRA

ADL

Traitement des descripteurs discrets

(1) Codage 0/1 des variables catégorielles (attention à la dernière modalité)
(2) Régularisation : ADL sur les axes factoriels de l'ACM → DISQUAL (G. Saporta)



SPAD fourni directement les coefficients sur les variables originelles codées en disjonctif complet (0/1), nous n'avons pas à réaliser les transformations à partir des facteurs.

ADL

Sélection automatique des variables (1/2) -- Principe

Principe : S'appuyer sur la statistique F

Démarche : Évaluer l'adjonction d'une (J+1)ème variable dans le modèle au sens de F

$$\frac{n-K-J}{K-1}\left(\frac{\Lambda_J}{\Lambda_{J+1}}-1\right) \cong F\left(K-1, n-K-J\right)$$

Sélection FORWARD

J=0
REPETER
    Pour chaque variable candidate, calculer F
    Choisir la variable qui maximise F
    Est-ce que l'adjonction entraîne une amélioration « significative » du modèle ?
    Si OUI, Ajouter la variable dans la sélection
JUSQU'À Plus de variable à ajouter

Remarque :

(1) Il faut s'entendre sur le terme « significatif » → attention à la p-value calculée
(2) Autres stratégies : BACKWARD et STEPWISE
(3) A rapprocher avec l'idée de la corrélation partielle en régression linéaire

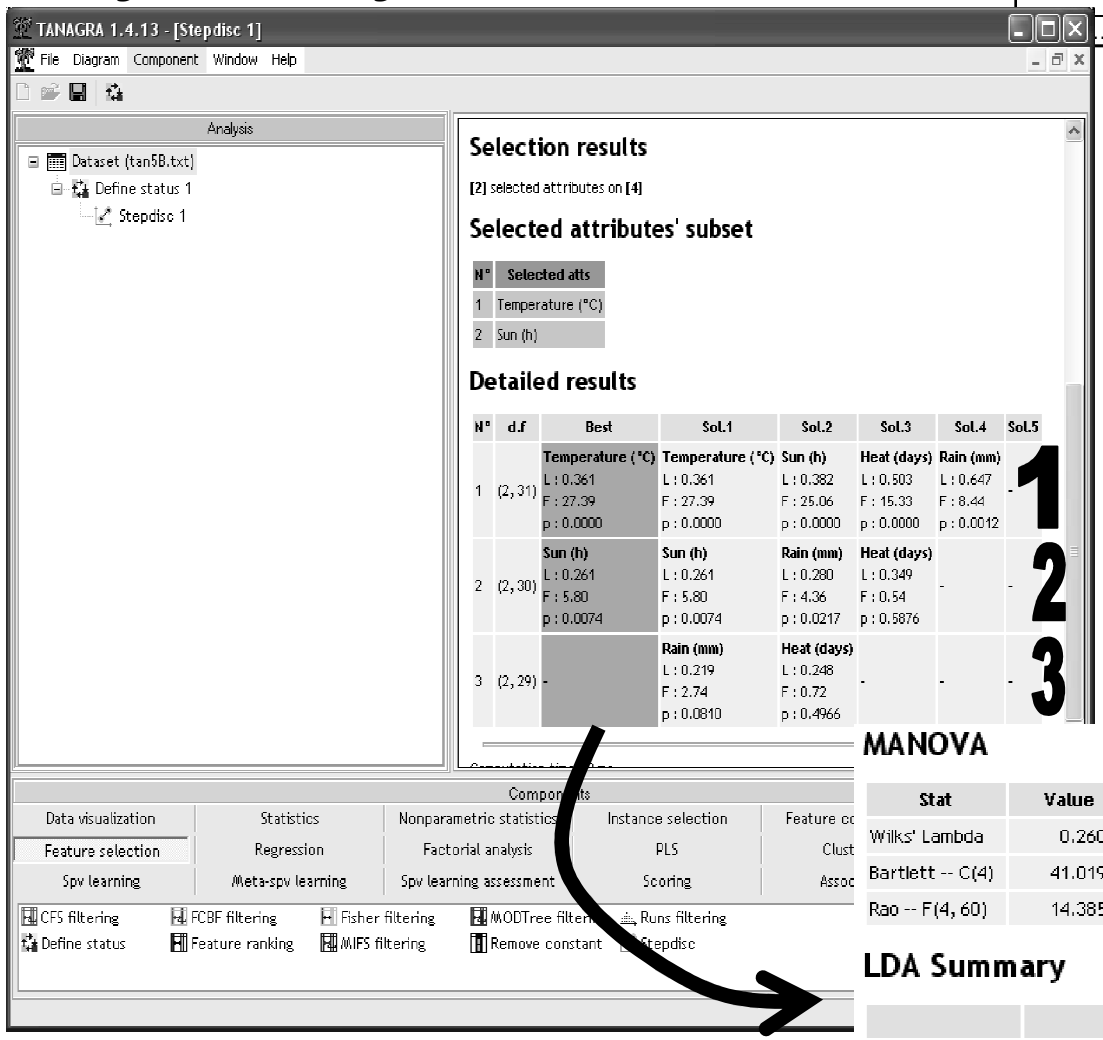
ADL

Sélection des variables (2/2) – Un exemple

Qualité des vins de Bordeaux (Tenenhaus, pp. 256-260)

Ex. Règle d'arrêt – Signif. Value = 0.05

Temperature	Sun (h)	Heat (days)	Rain (mm)	Quality
3064	1201	10	361	medium
3000	1053	11	338	bad
3155	1133	19	393	medium
3085	970	4	467	bad
3245	1258	36	294	good
...	...	...	...	...



MANOVA

Stat	Value	p-value
Wilks' Lambda	0.26057	-
Bartlett -- C(4)	41.01913	0.00000
Rao -- F(4, 60)	14.38531	0.00000

LDA Summary

	Classification functions			Statistical Evaluation			
Attribute	medium	bad	good	Wilks L.	Partial L.	F(2,30)	p-value
Temperature (°C)	0.389654	0.380641	0.408796	0.382143	0.681861	6.99861	0.003202
Sun (h)	0.081305	0.062977	0.091231	0.361395	0.721007	5.80425	0.007398
constant	-664.402665	-614.577030	-739.145806	-			

Bibliographie

M. BARDOS -- « Analyse discriminante – Application au risque et scoring financier », DUNOD, 2001.

L. LEBART, A. MORINEAU, M. PIRON – « Statistique exploratoire multidimensionnelle », DUNOD, 2000 (3ème édition).

M. TENENHAUS – « Méthodes statistiques en gestion », DUNOD, 1996.

R. TOMASSONE, M. DANZART, J.J. DAUDIN, J.P. MASSON –

« Discrimination et classement », Masson, 1988.