Linear Discriminant Analysis

(Predictive Discriminant Analysis)

Ricco RAKOTOMALALA

Maximum A Posteriori Rule

Bayes Theorem

Calculating the posterior probability

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

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

MAP - Maximum A Posteriori rule

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

$$\Leftrightarrow$$

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

Prior probability of class k: $P(Y=y_k)$
Estimated by empirical frequency n_k/n

How to estimate $P(X/Y=y_k)$

Assumptions are introduced in order to obtain a convenient calculation of this distribution.

Assumption 1: (X_1, …, X_J / y_k) is assumed multivariate normal

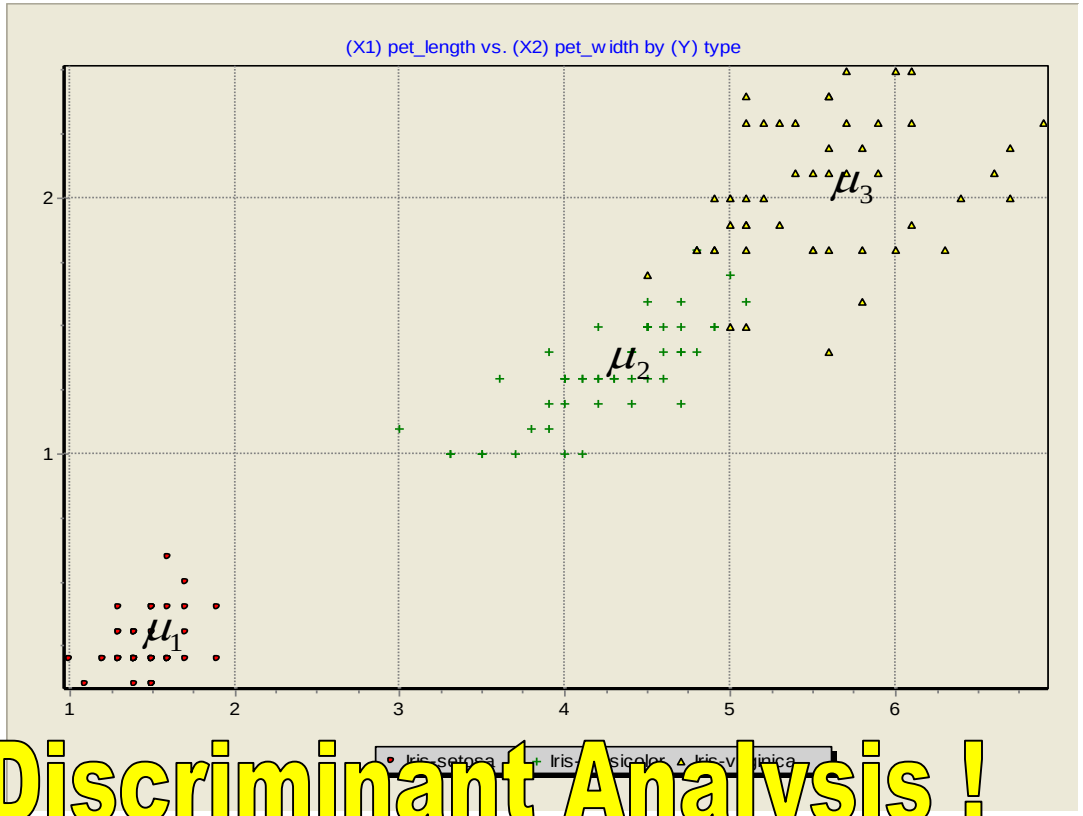
(Multivariate Gaussian Distribution – Parametric method)

Multivariate Gaussian Density

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

μ_k Conditional centroids

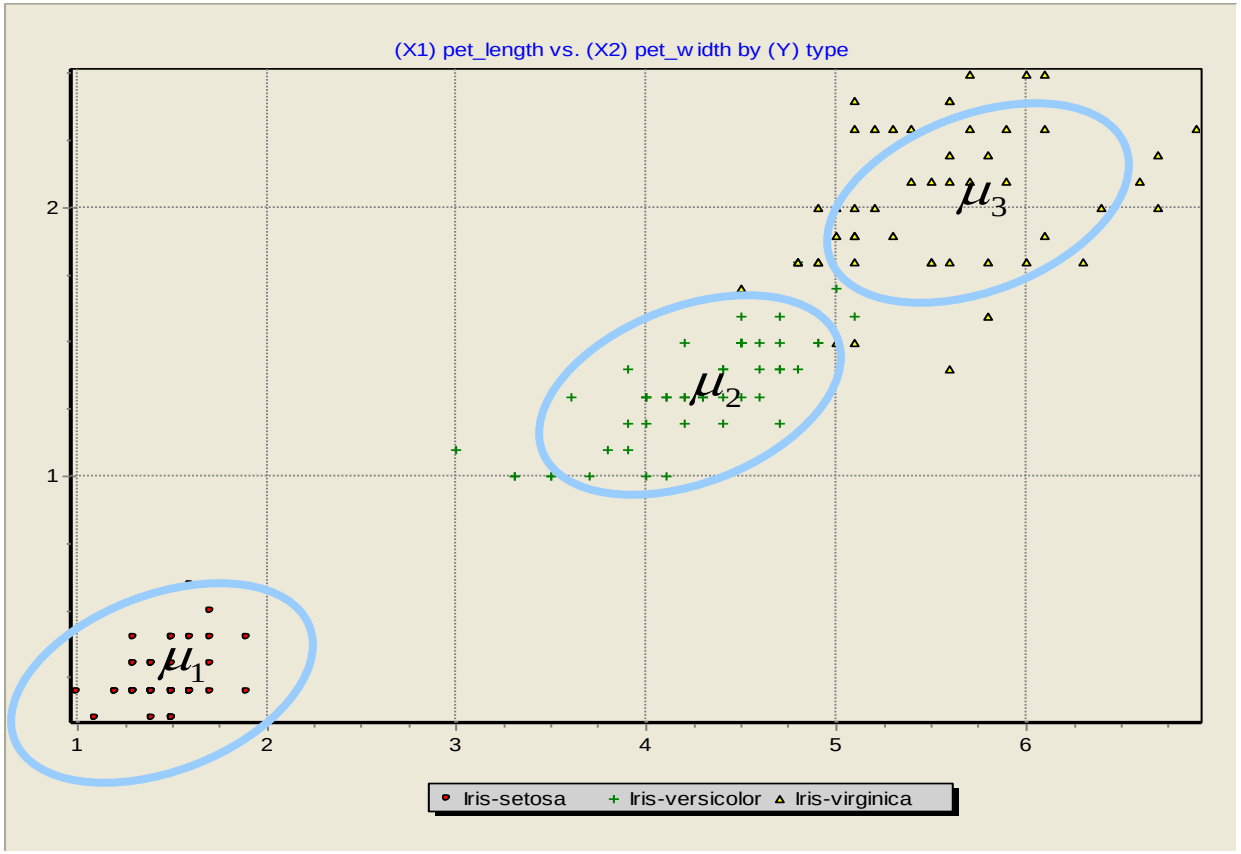
$\sum_k$ Conditional covariance matrices



Quadratic Discriminant Analysis !

Assumption 2: Population covariance matrices are equal

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



Linear Discriminant Analysis (LDA) !

Linear classification functions

(under the assumptions [1] and [2])

The natural logarithm of the conditional probability is proportional to:

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

From a sample with n instances, K classes and J predictive variables

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

Conditional centroids

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

Pooled variance covariance matrix

Linear classification functions

(an explicit classification model that can classify an unseen instance)

The classification function for y_k is proportional to $P(Y=y_k/X)$

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

Takes into account the prior probability of the group

Decision rule

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

Advantages et shortcomings

LDA - in general - is as effective as the other linear methods (e.g. logistic regression)

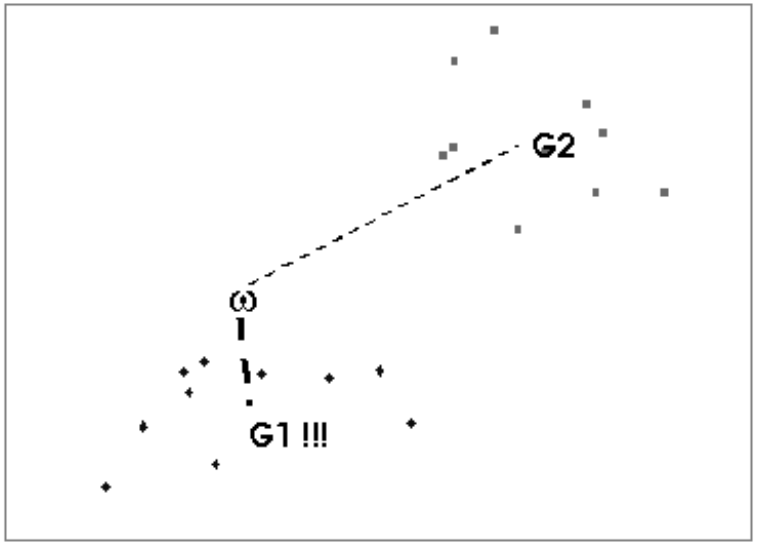
- >> It is robust to the deviation from the Gaussian assumption
- >> It may be disturbed by a strong deviation from the homoscedasticity assumption
- >> It is sensitive to the dimensionality and/or the presence of redundant variables
- >> The multimodal conditional distributions constitute a problem (e.g. 2 or more « clusters » for $Y=Y_k$)
- >> Sensitivity to outliers

Classification rule – Distance to the centroids

The classification function $d(Y_k,X)$ computed for the individual ω is based on

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

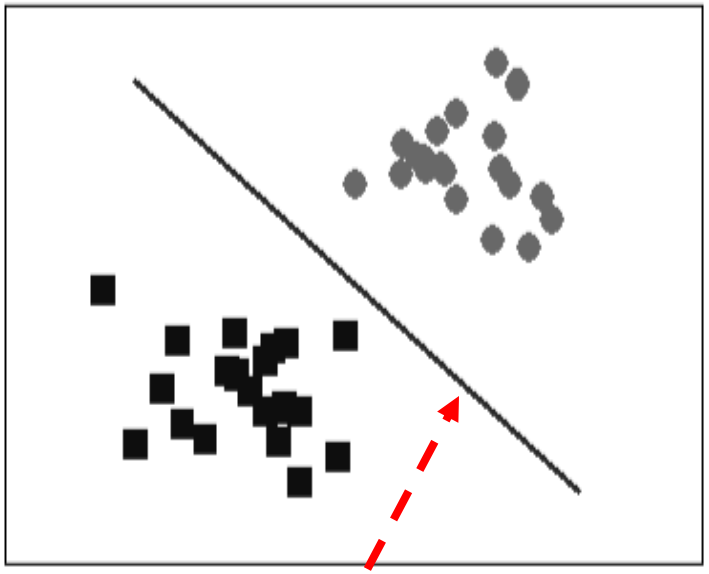
Distance-based classification : Assign ω to that the population to which it is closest (1) in the sense of the distance to the centroids, (2) using the Mahalanobis distance



We understand that LDA fails in some situations: (a) when we have multimodal conditional distributions, the group centroids are not reliable; (b) when the conditional covariance matrices are very different, the pooled covariance matrix is not appropriate for the calculation of distances.

Classification rule – Linear separator

Linear decision boundaries (hyperplane)
to separate the groups



Defined by the points equally distant to the two conditional centroids

LDA, the decision rule can be interpreted in different ways: (a) MAP decision rule (posterior probability); (b) distance to the centroids; (c) linear separator which defines various regions in the representation space

Evaluation of the classifier

(1) Estimating classification error rate
Holdout scheme: Learning + Test → Confusion matrix

(2) Overall "statistical" evaluation of the classifier

$$H_0 : \mu_1 = \cdots = \mu_K$$

One-way MANOVA statistical test
H_0: the population centroids do not differ

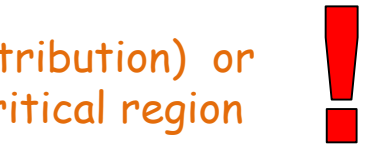
The test statistic: WILKS' LAMBDA

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

Pooled covariance matrix

Global covariance matrix

In practice, we use the Bartlett transformation (χ^2 distribution) or the Rao transformation (F distribution) to define the critical region



Assessing the relevance of the descriptors

Measuring the influence of the variables in the classifier

The idea is to measure the variation of the Wilks' lambda of the model with [J variables] and without [J-1 variables] the variable that we want to evaluate.

The F statistic (loss in separation if the J^{th} variable is deleted)

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

This statistic is often available into the tools from the statistician community (not into the tools from the machine learning community)

The particular case of the binary classification (K = 2)

We have a binary class attribute → Y = {+,-}

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

Decision rule

D(X) > 0 → Y = +

Interpretation

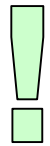
>> d(X) is a SCORE function, it enables to assign a score [proportional to the positive class probability estimate] to each instance

>> The sign of the coefficients allows to understand the sense of the influence of the variable on the class attribute

Evaluation

>> There is an analogy between the logistic regression and the LDA.

>> There is also a strong analogy between the linear regression between the linear regression of an indicator (0/1) response variable and the LDA (we can use some results of the first one for the second one).



LDA with Tanagra software

Statistical overall evaluation

MANOVA

Stat	Value	p-value
Wilks' Lambda	0.1639	-
Bartlett -- C(9)	1252.4759	0
Rao -- F(9, 689)	390.5925	0

LDA Summary

	Classification functions		Statistical Evaluation			
Attribute	**begnin**	**malignant**	**Wilks L.**	**Partial L.**	**F(1,689)**	**p-value**
clump	0.728957	1.615639	0.183803	0.891601	83.76696	0
ucellsize	-0.316259	0.29187	0.166796	0.982512	12.26383	0.000492
ucellshape	0.066021	0.504149	0.165463	0.990423	6.6621	0.010054
mgadhesion	0.057281	0.232155	0.164499	0.99623	2.60769	0.106805
sepics	0.654272	0.869596	0.164423	0.996687	2.29011	0.130659
bnuclei	0.209333	1.427423	0.210303	0.779248	195.18577	0
bchromatin	0.686367	1.245253	0.167816	0.976538	16.55349	0.000053
normnucl	-0.000296	0.461624	0.168846	0.97058	20.88498	0.000006
mitoses	0.200806	0.278126	0.163956	0.99953	0.32432	0.569209
constant	-3.047873	-23.296414				

Classification functions
(Linear Discriminant Functions)

Variable importance

LDA with SPAD software

(1) Only for binary problem
(2) All predictive variables must be continuous
(3) Evaluation of the relevance of the variables by the way of the linear regression

$$D = d(begnin / X) - d(malignant / X)$$

$(9.15...)^2 \approx 83.76696$

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

Overall statistical evaluation of the model
F from the Wilks' lambda, Hotelling's T2

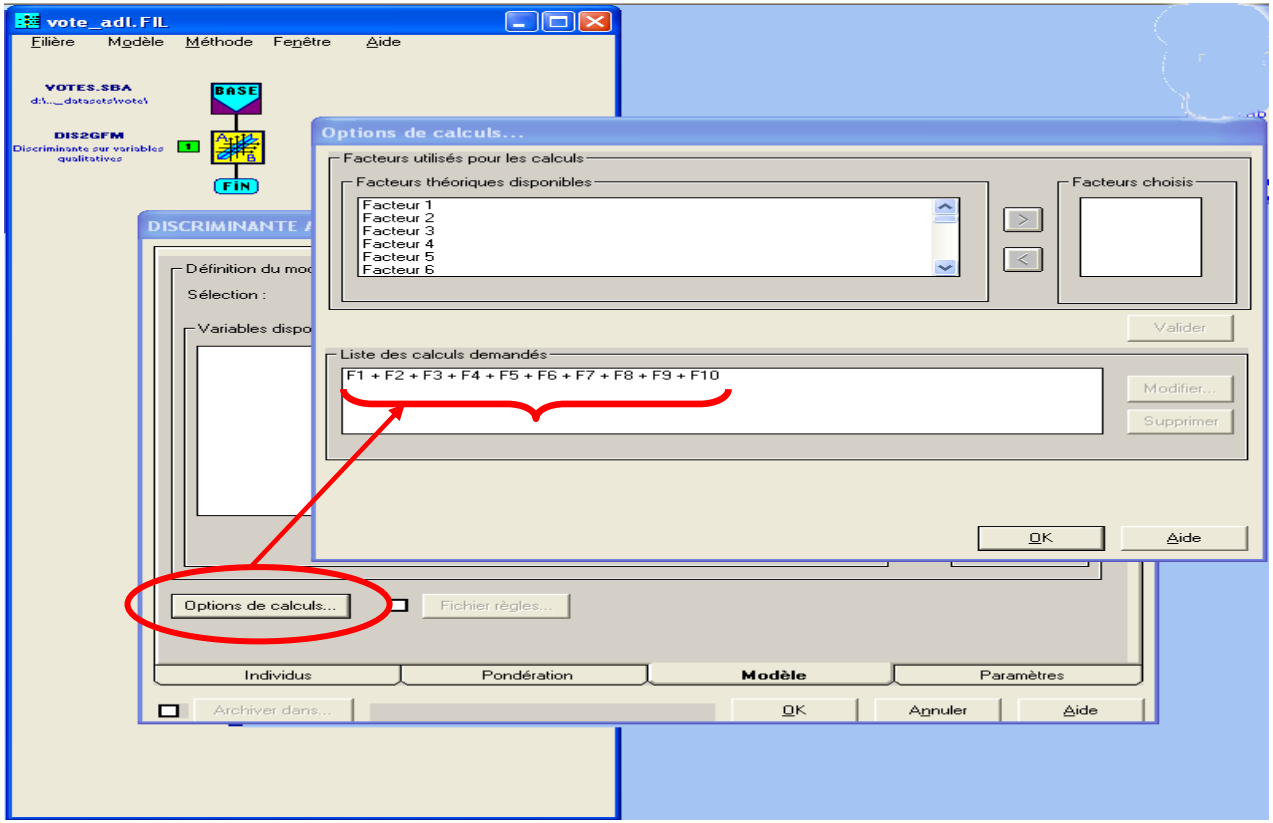
Results of the linear regression on the indicator response variable

Dealing with discrete (categorical) predictive variables

(1) Dummy coding scheme (we must define a fixed reference level)

(2) DISQUAL (Saporta): Multiple Correspondence Analysis + LDA from the factor scores

(This is a kind of regularization which enables to reduce the variance of the classifier when we select a subset of the factors)



Some tools such as SPAD can perform DISQUAL and provide the classification functions on the dummy variables.

Feature selection (1) – The STEPDISC approach

Forward strategy

Principle: Based on the F statistic

Process: Evaluate the addition of the (J+1)th variable into the classifier at each step

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

FORWARD selection

J=0
REPEAT
For each candidate variable, calculate the F statistic
Select the variable which maximizes F
The addition implies a "significant" improvement of the model?
If YES, the variable is incorporated in the model
UNTIL (no variable can be added)

Note:

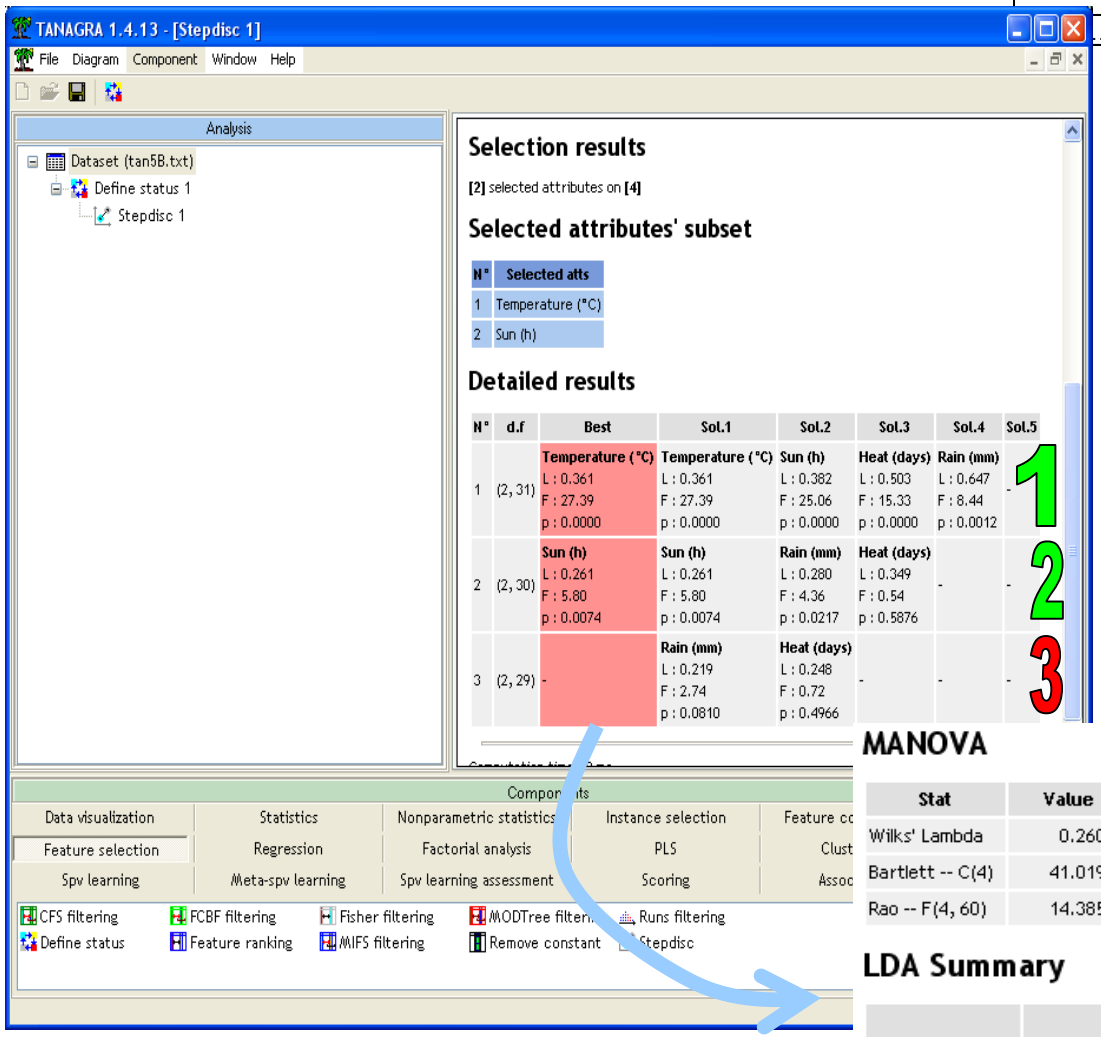
(1) Problems may arise when we define "significant" with the computed p-value (see 'multiple comparison')
(2) Other strategies: BACKWARD and BIDIRECTIONAL
(3) A similar strategy is performed in the linear regression

Feature selection (2)

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
...	...	...	...	...

Wine quality (Tenenhaus, pp. 256-260)

E.g. Stopping rule – Significance level $\alpha = 0.05$



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	-			

Bibliography

STAT 897D – "Applied Data Mining", The Pennsylvania State University, 2014. https://onlinecourses.science.psu.edu/stat857/

G. James, D. Witten, T. Hastie, R. Tibshirani, "An introduction to Statistical Learning", Springer, 2013. http://www-bcf.usc.edu/~gareth/ISL/

SAS/STAT(R) 9.3 User's Guide, "The DISCRIM Procedure". http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#discrim_toc.htm