Subject

In this tutorial, we show how to use the CANONICAL DISCRIMINANT ANALYSIS component.

One of the goals of this method is to produce new variables ("latent" variables) from a set of examples classified into groups. These new variables optimize the separation between groups.

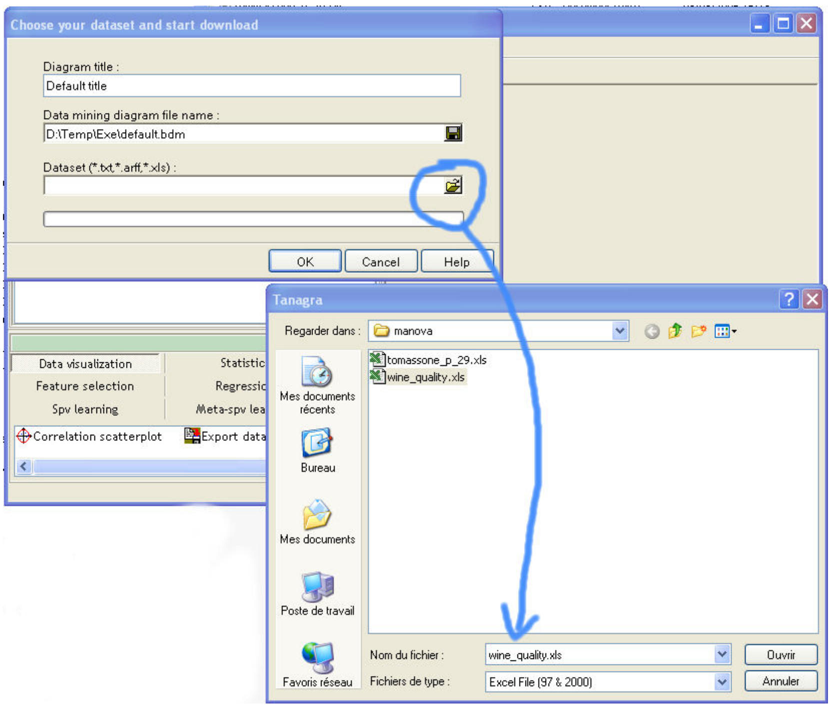
Dataset

We use the WINE_QUALITY.XLS[1] dataset. It contains 34 wines classified into 3 groups "good", "medium" and "bad". We have weather descriptors (sum of daily temperature, sun, heat and rain).

Canonical Discriminant Analysis

Download the dataset

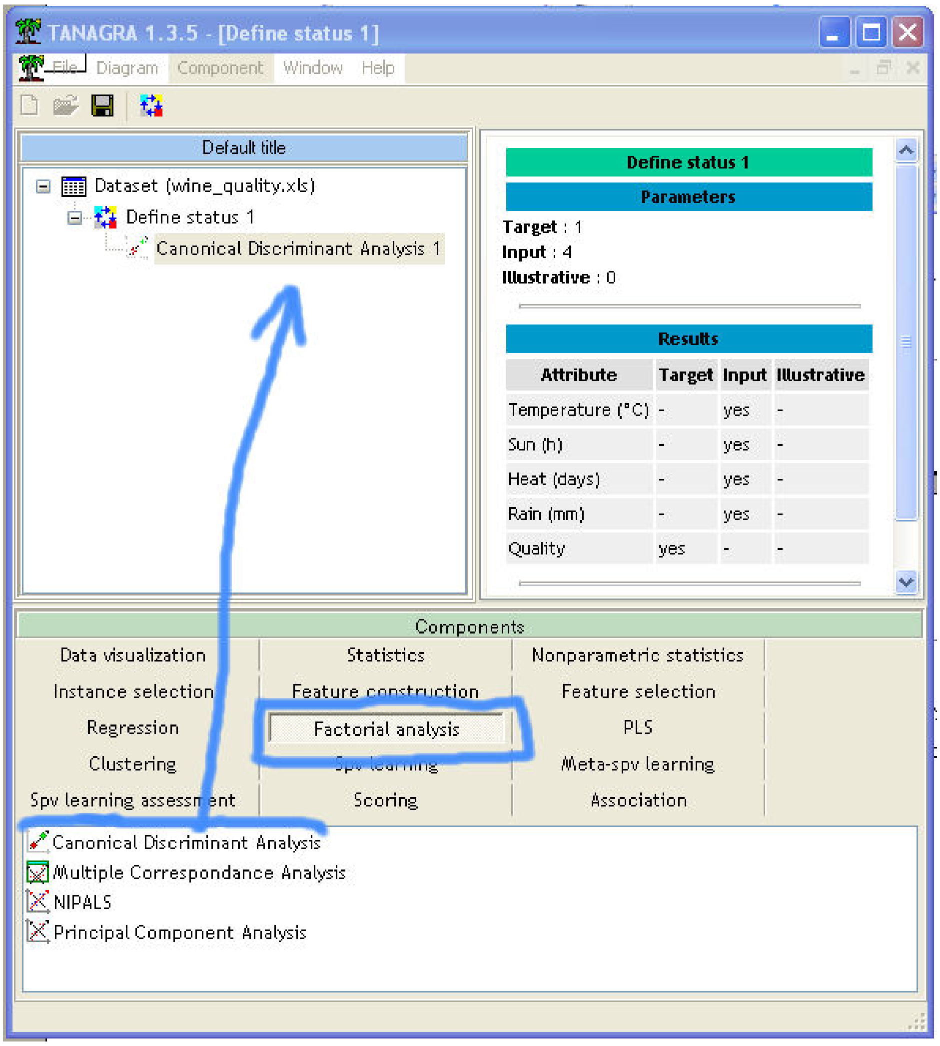
The first step is to create a new diagram and import the dataset (FILE / NEW).



[1] From M. Tenenhaus, « Méthodes Statistiques en Gestion », Edition Dunod, 1996, p. 244 (Tableau 1) – Annual data on 1924 – 1957 period.

Performing the analysis

We insert the DEFINE STATUS component; we set QUALITY as TARGET and the other descriptors as INPUT. Then, we insert the CANONICAL DISCRIMINANT ANALYSIS component.



Reading the results

We obtain 3 tables: the importance of the latent variables in the separation of groups; the coefficients of the discrimination functions; the factor structure matrix i.e. the correlation between the descriptors and the latent variables.

Results

Roots and Wilks' Lambda

Root	Eigenvalue	Proportion	Canonical R	Wilks Lambda	CHI-2	d.f.	p-value
1	3.27886	0.95945	0.875382	0.205263	46.7122	8	0.000000
2	0.13857	1.00000	0.348867	0.878292	3.8284	3	0.280599

Canonical Discriminant Function

Coefficients	Unstandardized		Standardized	
Attribute	Root n°1	Root n°2	Root n°1	Root n°2
Temperature (°C)	-0.0086	0.0000	-0.7509	0.0041
Sun (h)	-0.0068	0.0053	-0.5476	0.4309
Heat (days)	0.0271	-0.1278	0.1984	-0.9362
Rain (mm)	0.0059	-0.0062	0.4456	-0.4690
constant	32.91135	-2.16759	-	

Factor Structure Matrix - Correlations

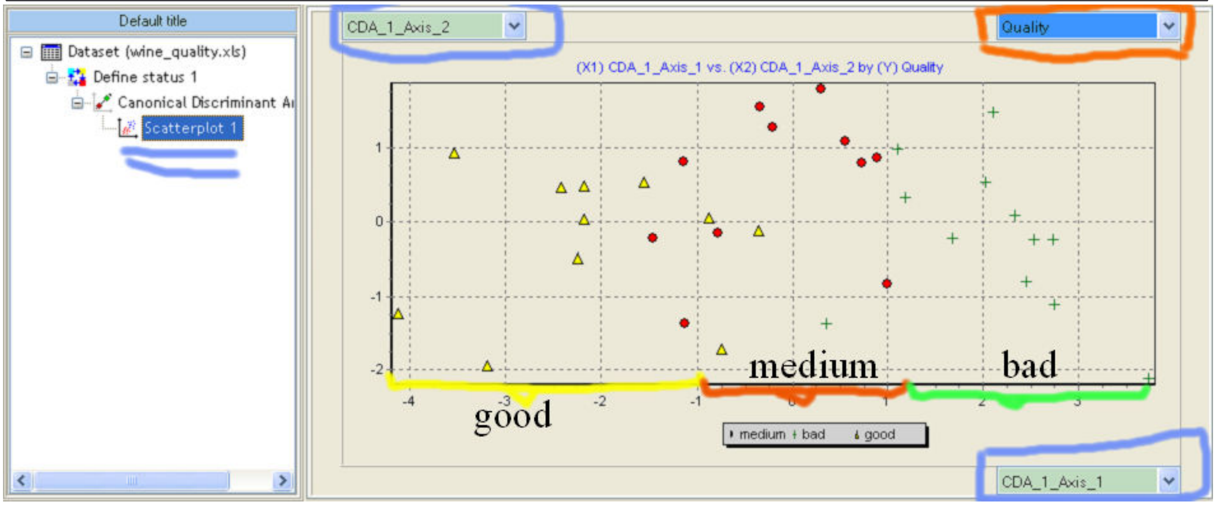
Root	Root n°1			Root n°2		
Descriptors	Total	Within	Between	Total	Within	Between
Temperature (°C)	-0.901	-0.724	-0.987	-0.375	-0.584	-0.164
Sun (h)	-0.897	-0.701	-0.999	0.116	0.176	0.052
Heat (days)	-0.771	-0.525	-0.956	-0.590	-0.780	-0.292
Rain (mm)	0.663	0.398	0.977	-0.361	-0.421	-0.212

The unstandardized discriminant coefficients enable us to perform a projection of a new individual. For instance, on the first factor Z1, with the following values TEMPERATURE = 3000, SUN = 1100, HEAT = 20 and RAIN = 300, we compute the coordinate: -0.0086 x 3000 + -0.0068 x 1100 + 0.0271 x 20 + 0.0059 x 300 + 32.91135 = 1.9435

Graphical representation

We can build a graphical representation of the examples on the 2 axes Z1 and Z2. We add a SCATTERPLOT component in our diagram.

We select the appropriate variables in the list box; we can define the shape of the points according to the group value (QUALITY). We see that the first factor enables to separate the wines according their quality. We see also that the above new observation (Z1 = 1.9435) is probably a “bad” wine.


Default title
Dataset (wine_quality.xls)
Define status 1
Canonical Discriminant A
Scatterplot 1
CDA_1_Axis_2
Quality
(X1) CDA_1_Axis_1 vs. (X2) CDA_1_Axis_2 by (Y) Quality
good
medium
bad
medium bad good
CDA_1_Axis_1