1 Objectif

Présenter les techniques PLS-DA (Partial Least Squares Discriminant Analysis) implémentées dans TANAGRA 1.4.23.

La régression PLS est une technique de régression qui vise à prédire les valeurs prises par un groupe de variables Y (variables à prédire, variables cibles, variables expliquées) à partir d'une série de variables X (variables prédictives, les descripteurs, variables explicatives) (Tenenhaus[1], 1998 ; Garson, http://www2.chass.ncsu.edu/garson/PA765/pls.htm).

Dans le cas où l'on n'a qu'une seule variable cible, elle se positionne comme une alternative à la régression linéaire multiple (méthode des moindres carrés ordinaires - MCO). Elle se démarque par sa capacité à traiter les problèmes où nous avons un grand nombre de descripteurs, non pertinents et/ou colinéaires, voire lorsque le nombre de variables prédictives est supérieur au nombre d'observations. Ce sont des situations que l'on rencontre fréquemment dans l'analyse de données non structurées (images, texte, etc.) où les descripteurs sont générés automatiquement en grand nombre. La méthode des MCO traditionnelle est totalement inefficace dans ce contexte.

Schématiquement, la régression PLS produit itérativement une série de facteurs deux à deux orthogonaux qui sont par la suite présentés à une procédure MCO. Ces axes factoriels sont calculés de manière à maximiser leur covariance avec les variables cibles.

Le choix du nombre de facteurs est très important dans ce processus. Lorsqu'il est élevé, nous privilégions la qualité de la prédiction sur l'échantillon d'apprentissage, au risque de trop coller aux données et de tomber dans le piège du sur-apprentissage[2] : nous modélisons des relations qui n'appartiennent qu'à l'échantillon que nous manipulons. Lorsqu'il est faible, nous privilégions la stabilité de la solution, au risque de ne pas retranscrire la relation qui existe entre les variables cibles et les descripteurs dans la population. Le nombre de facteurs détermine ainsi les performances et la robustesse du modèle que l'on veut produire. Fort heureusement, la plage des valeurs adéquates est suffisamment large pour que nous puissions en proposer a priori sans trop se tromper, quitte à moduler par la suite pour optimiser les performances en prédiction.

La régression PLS a été définie à l'origine pour les problèmes de prédictions sur des variables cibles quantitatives. Il aurait été dommage de ne pas exploiter ses qualités en apprentissage supervisé où, rappelons le, la variable cible est catégorielle.

Dans ce document, nous présentons plusieurs variantes de la régression PLS dédiées à la prédiction d'une variable catégorielle. Elles sont regroupées sous l'appellation générique de « PLS Discriminant Analysis (PLS-DA) ». Elles reposent sur le même principe : dans un premier temps, nous codons la variable à prédire catégorielle à l'aide d'une série d'indicatrices correspondant à ses modalités (codage disjonctif complet) ; dans un second temps, nous présentons le tableau de données, Y composé des indicatrices, X des descripteurs, à l'algorithme PLS. Les variantes diffèrent (1) par le type de codage et la valeur des codes utilisés lors de la constitution du tableau Y ; (2) par l'exploitation des résultats de la régression PLS lors de la phase de classement[3].

[1] Tout au long de didacticiel, nous ferons référence à l'ouvrage de M. Tenehaus qui est la référence francophone en matière de Régression PLS : M. Tenenhaus, « La régression PLS – Théorie et Pratique », Technip, 1998.

[2] Lorsque le nombre de facteur est égal au nombre de variables, nous retombons sur les résultats de la MCO.

[3] Voir S. Chevallier, D. Bertrand, A. Kohler, P. Courcoux, « Application of PLS-DA in multivariate image analysis », in J. Chemometrics, 20 : 221-229, 2006.

Note : Ce didacticiel vise avant tout à présenter les techniques et à donner les repères de lecture des résultats. Nous utiliserons donc un jeu de données très simple pour faciliter la lecture. Dans un prochain document, nous utiliserons des données autrement plus difficiles à appréhender, avec une dimensionnalité élevée au regard du nombre d'observations. Nous constaterons alors l'excellent comportement de la régression PLS qui soutient la comparaison avec des méthodes fortement régularisées telles que les SVM (Support Vector Machine).

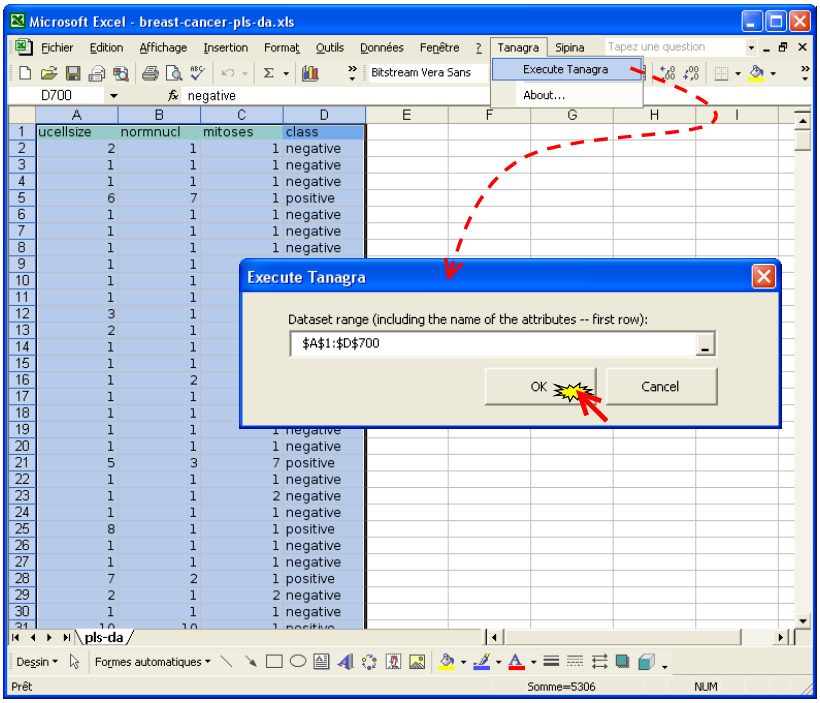
2 Données

Nous utilisons le fichier BREAST-CANCER-PLS-DA.XLS (http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/breast-cancer-pls-da.xls). Il est formé d'une fraction des données « breast-cancer-wisconsin » accessible sur le site UCI[4]. Nous voulons prédire le caractère cancéreux de cellules (CLASS) à partir 3 descripteurs UCELLSIZE, NORMNUCL et MITOSES. Le fichier comporte 699 observations. La variable cible CLASS comporte 2 modalités NEGATIVE (les cellules sont saines) et POSITIVE (elles sont cancéreuses).

3 Préparer les traitements

3.1 Chargement des données et lancement de Tanagra

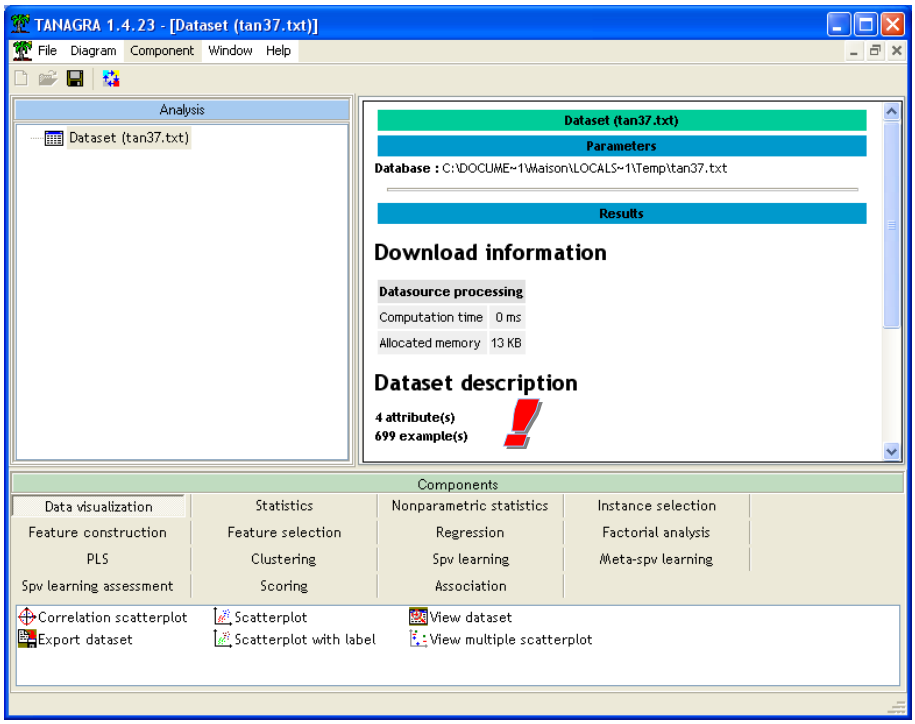
Le plus simple pour lancer Tanagra et charger les données est d'ouvrir le fichier XLS dans le tableur EXCEL. Nous sélectionnons la plage de données. La première ligne doit correspondre au nom des variables. Puis nous activons le menu TANAGRA / EXECUTE TANAGRA qui a été installé avec la macro complémentaire TANAGRA.XLA[5]. Une boîte de dialogue apparaît. Nous vérifions la sélection. Si tout est en règle, nous validons en cliquant sur le bouton OK.



[4] http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)

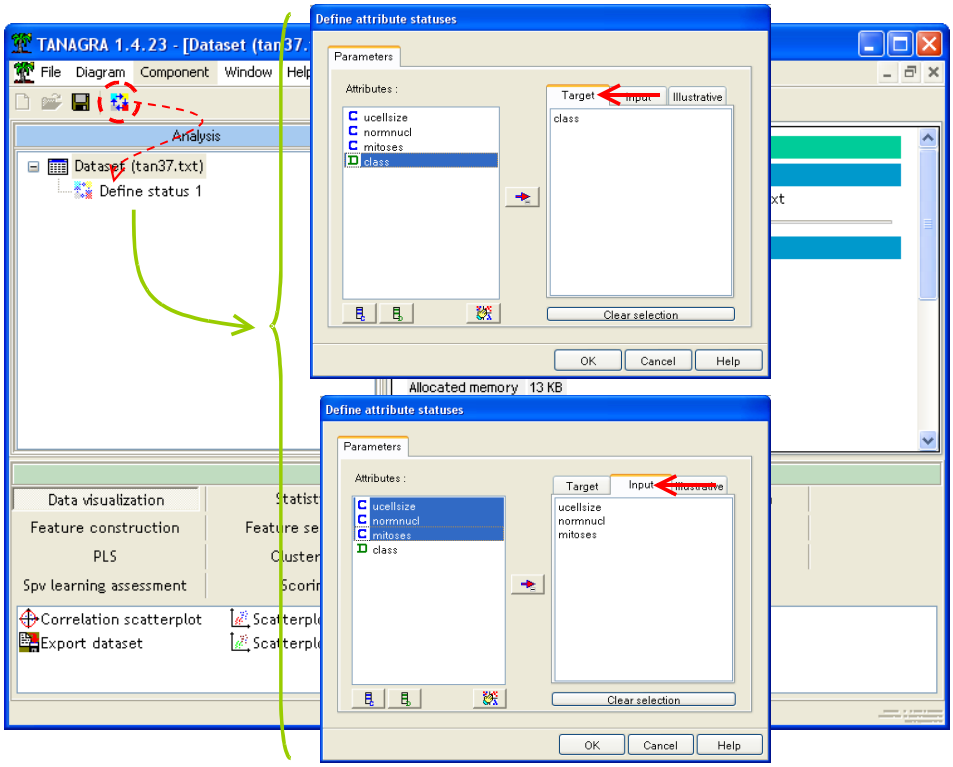
[5] Voir http://tutoriels-data-mining.blogspot.com/2008/03/importation-fichier-xls-excel-macro.html concernant l'installation et l'utilisation de la macro complémentaire TANAGRA.XLA.

Tanagra est automatiquement démarré. Un nouveau diagramme est créé, les données sont chargées : nous disposons de 699 observations décrites à l'aide de 4 variables.



3.2 Définir le problème à traiter

Nous voulons prédire les valeurs de la variable CLASS à partir des autres variables. Nous utilisons le composant DEFINE STATUS pour le spécifier. Le plus simple est de passer par le raccourci dans la barre d'outils. Nous plaçons CLASS en TARGET, les autres variables en INPUT.



4 Partial Least Squares Discriminant Analysis

4.1 Le composant C-PLS

4.1.1 Description de la méthode

Le composant C-PLS est dédié au traitement de problèmes binaire c.-à-d. **la variable TARGET doit être discrète à 2 modalités**.

Il applique la régression PLS après avoir recodé convenablement la variable à prédire. Nous utilisons les codes suggérés dans l'ouvrage de Tomassone et al. (1988)[6]. Ces auteurs montrent l'analogie entre l'analyse discriminante prédictive et la régression lorsque l'on traite une variable cible binaire. Avec un codage approprié, nous retrouvons une règle d'affectation assez commode. Résumons l'idée.

La variable à prédire Y a deux modalités $Y = \{+,-\}$. Si n_+ (resp. n_-) est l'effectif des POSITIVE (resp. NEGATIVE), avec $n = n_+ + n_-$; la variable recodée Z est définie de la manière suivante :

$$Z = \begin{cases} \dfrac{n_-}{n}, \text{si Y} = + \\ -\dfrac{n+}{n}, \text{si Y} = - \end{cases}$$

Ainsi, la régression PLS fournit une fonction linéaire discriminante D(X). Elle permet de prédire la valeur de la variable TARGET à partir des descripteurs (*p* est le nombre de descripteurs) en utilisant la règle usuelle :

$$D(X) = a_0 + a_1 X_1 + \ldots + a_p X_p \begin{cases} \geq 0 \Rightarrow Y = + \\ < 0 \Rightarrow Y = - \end{cases}$$

Première option concernant le choix du nombre de facteurs, l'utilisateur sait précisément ce qu'il souhaite obtenir, il a la possibilité de fixer lui même le nombre d'axes à utiliser.

Deuxième option, c'est la méthode C-PLS qui va détecter automatiquement le nombre d'axes PLS à retenir. La stratégie est relativement simple, elle teste pour chaque nouveau facteur calculé si la variabilité expliquée additionnelle des Y, que l'on appelle *redondance*, est supérieure à un seuil (0.025 par défaut[7]). C'est une heuristique un peu simple, une approche par validation croisée basé sur le PRESS (Tenenhaus, 1998 ; page 83) par exemple serait plus précise. Néanmoins, dans une grande majorité de cas, cela permet d'initier une première analyse qui tient la route.

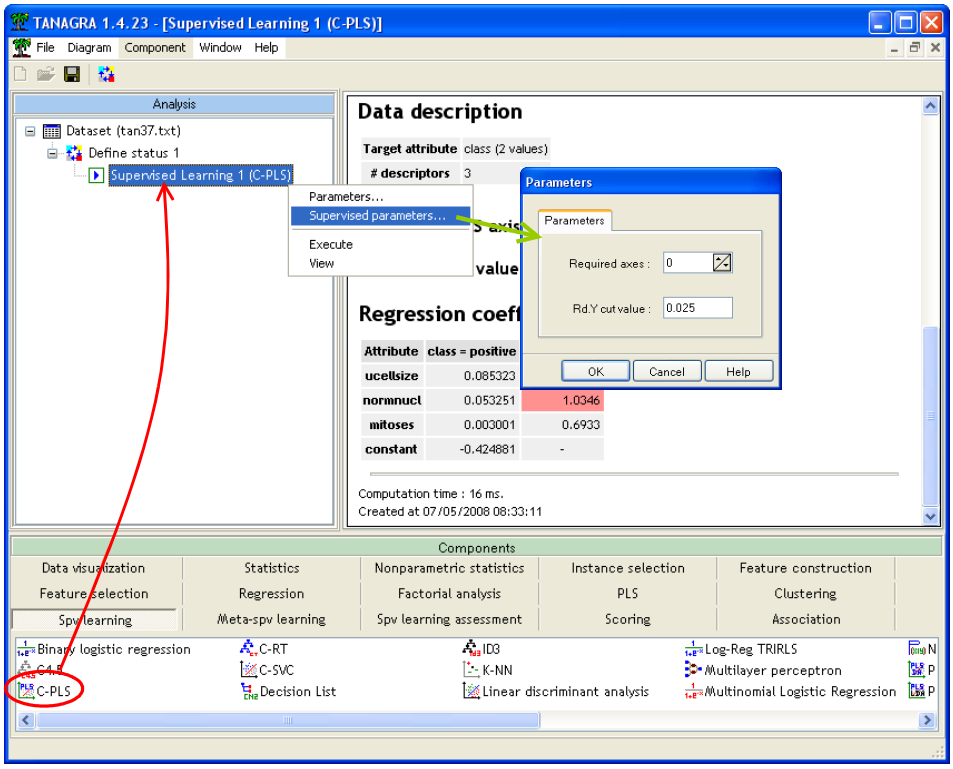
4.1.2 Mise en œuvre et lecture des résultats

Nous insérons le composant C-PLS (onglet SPV LEARNING) dans le diagramme.

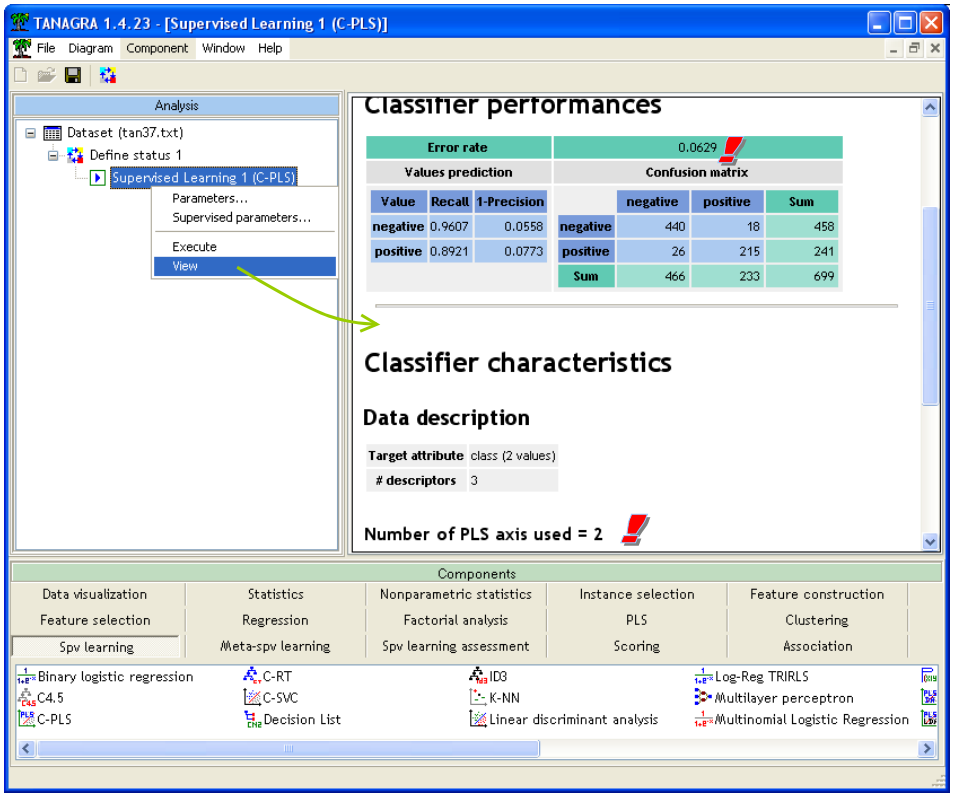
[6] R. Tomassone, M. Danzart, J.J. Daudin, J.P. Masson, « Discrimination et classement », Masson, 1988 ; page 38.

[7] Choix complètement arbitraire, il faut en convenir, mais fonctionne correctement dans la grande majorité des cas. Nous avons bien sûr la possibilité de moduler ce seuil.

Nous activons menu contextuel SUPERVISED PARAMETERS pour accéder à la boîte de dialogue de paramétrage. Le nombre d'axes demandé est égal à zéro, cela indique qu'il sera calculé automatiquement ; la valeur seuil de la redondance sur Y est 0.025, si un facteur supplémentaire introduit une variabilité expliquée additionnelle inférieure à ce seuil, il n'est pas accepté.



Nous validons les paramètres (Bouton OK), puis nous activons le menu contextuel VIEW du composant pour accéder aux résultats.



Nous retrouvons la matrice de confusion, usuelle en apprentissage supervisé. Le taux d'erreur en resubstitution (estimé sur les données d'apprentissage[8]) est proposé. Il est égal à 6.29%. La méthode a retenu les deux premiers axes factoriels pour élaborer le modèle de prédiction.

Coefficients de la fonction discriminante. Plus bas dans la fenêtre, les coefficients de la fonction discriminante sont affichés. Ils nous permettent de réaliser des prédictions sur de nouveaux individus. La modalité positive de la variable à prédire est « CLASS = POSITIVE ».

Regression coefficients

Attribute	class = positive	VIP
ucellsize	0.085323	1.2038
normnucl	0.053251	1.0346
mitoses	0.003001	0.6933
constant	-0.424881	-

Figure 1 - Fonction de classement C-PLS

Prenons un exemple simple pour préciser cette idée. Un individu présente les valeurs suivantes (UCELLSIZE = 2 ; NORMNUCL = 1 ; MITOSES = 1). Nous appliquons la fonction D(X) :

$$D(X) = -0.425 + 0.085 \times 2 + 0.053 \times 1 + 0.003 \times 1 = -0.198 < 0 \Rightarrow Y = -$$

Nous concluons que les cellules ne sont pas cancéreuses.

Importance des variables. Enfin, dernière information très importante dans ce même tableau, l'importance des descripteurs dans la projection (VIP : Variable Importance in Projection). Elles nous indiquent la pertinence de chaque descripteur pour la prédiction des valeurs de Y à travers les h premiers facteurs retenus. (Tenenhaus, 1998 ; page 139).

La VIP nous permet de hiérarchiser les variables selon leur pouvoir explicatif sur les variables cibles. Plus la valeur est grande, plus la variable est intéressante. De manière très simplifiée, nous dirons qu'une variable est à considérer avec attention dès lors que (VIP $\geq$ 1).

Une autre règle de sélection (http://www2.chass.ncsu.edu/garson/PA765/pls.htm) serait de supprimer une variable si (a) son VIP < 0.8 ; (b) le coefficient de régression associé est très petit en valeur absolue.

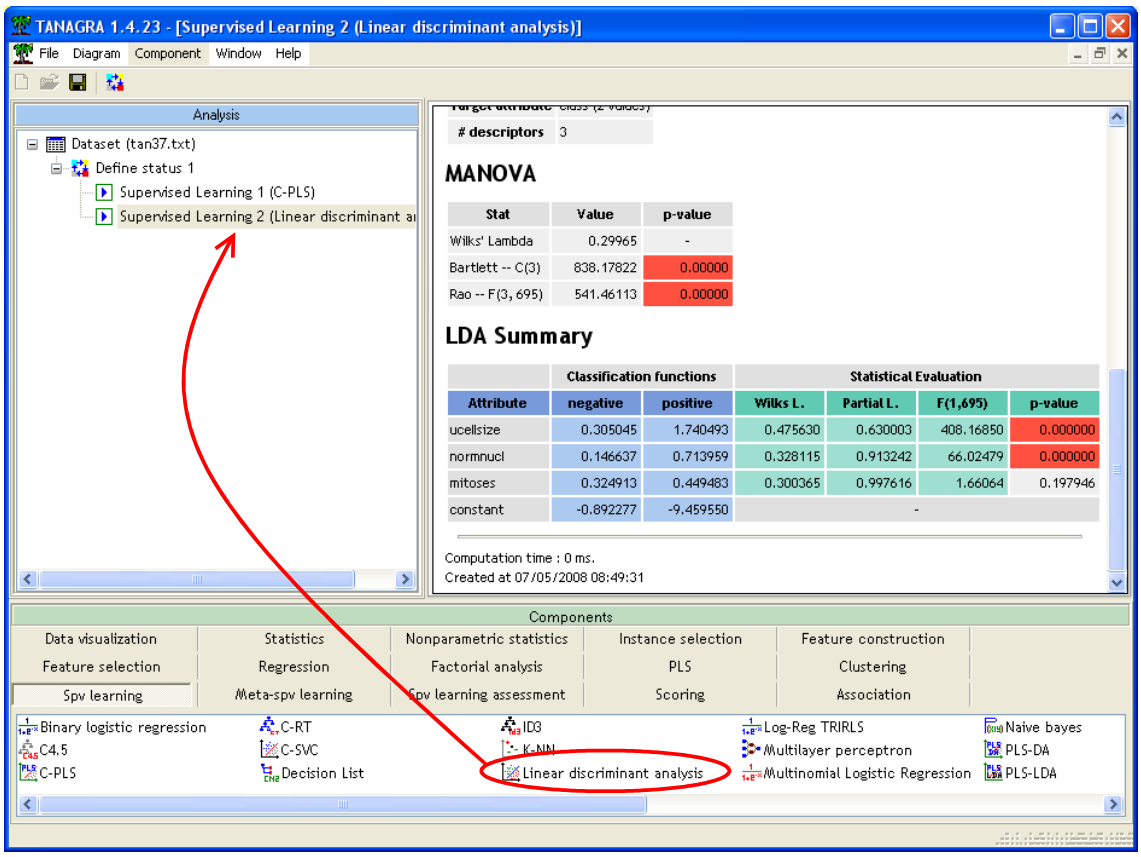
Note : Ces règles de sélection sont à manier avec beaucoup de précautions. Le VIP sert avant tout à positionner les variables les unes par rapport aux autres. Une variable en dernière position ne veut pas nécessairement dire qu'elle est inopérante dans la discrimination.

Dans notre exemple (Figure 1), UCELLSIZE et NORMNUCL semblent déterminantes. On peut se poser la question de conserver ou non la variable MITOSES (avec les réserves émises précédemment).

[8] Rappelons à toutes fins utiles que le taux d'erreur en resubstitution est un très mauvais indicateur des véritables performances du modèle de prédiction. On a tout intérêt à : soit réserver une fraction des observations pour la phase de test, soit utiliser les méthodes de ré échantillonnage (bootstrap, validation croisée, etc.). Voir par exemple le didacticiel http://tutoriels-data-mining.blogspot.com/2008/03/validation-croise-bootstrap-leave-one.html à ce sujet.

4.1.3 Comparaison avec l'analyse discriminante

Comparons ces résultats avec ceux de l'Analyse Discriminante Linéaire (voir http://tutoriels-data-mining.blogspot.com/2008/04/analyse-discriminante-linaire.html pour plus de détails sur cette technique). Nous insérons le composant LINEAR DISCRIMANT ANALYSIS (onglet SPV LEARNING) dans le diagramme, à la suite du DEFINE STATUS 1. Nous cliquons sur le menu contextuel VIEW pour obtenir les résultats.



L'analyse discriminante propose une fonction de classement pour chaque modalité de la variable à prédire. Les coefficients ne sont pas comparables avec ceux de C-PLS. En revanche, en ce qui concerne le rôle des descripteurs, les résultats sont cohérents : nous observons également le même ordonnancement des variables, et MITOSES n'est pas significatif dans la discrimination

4.2 Le composant PLS-DA

4.2.1 Description de la méthode

Le composant PLS-DA appréhende les problèmes multi classes c.-à-d. **une variable TARGET comportant K (K ≥ 2) modalités.**

Pour la construction du modèle de prédiction, la variable cible est remplacée par K variables indicatrices définies de la manière suivante :

$$Z_k = \begin{cases} 1, \text{si } Y = y_k \\ 0, \text{sinon} \end{cases}$$

L'algorithme PLS est lancé sur le tableau formé par les Zk et les descripteurs X. Nous obtenons K fonctions de classement qui permettent d'associer, pour un individu à classer, une valeur prédite à chacune des indicatrices

$$\hat{Z}_k = b_{0,k} + b_{1,k}X_1 + \cdots + b_{p,k}X_p$$

Lors de la phase de classement, nous calculons la valeur prédite pour chacune des indicatrices, nous assignons la conclusion correspondant à la valeur la plus élevée c.-à-d.

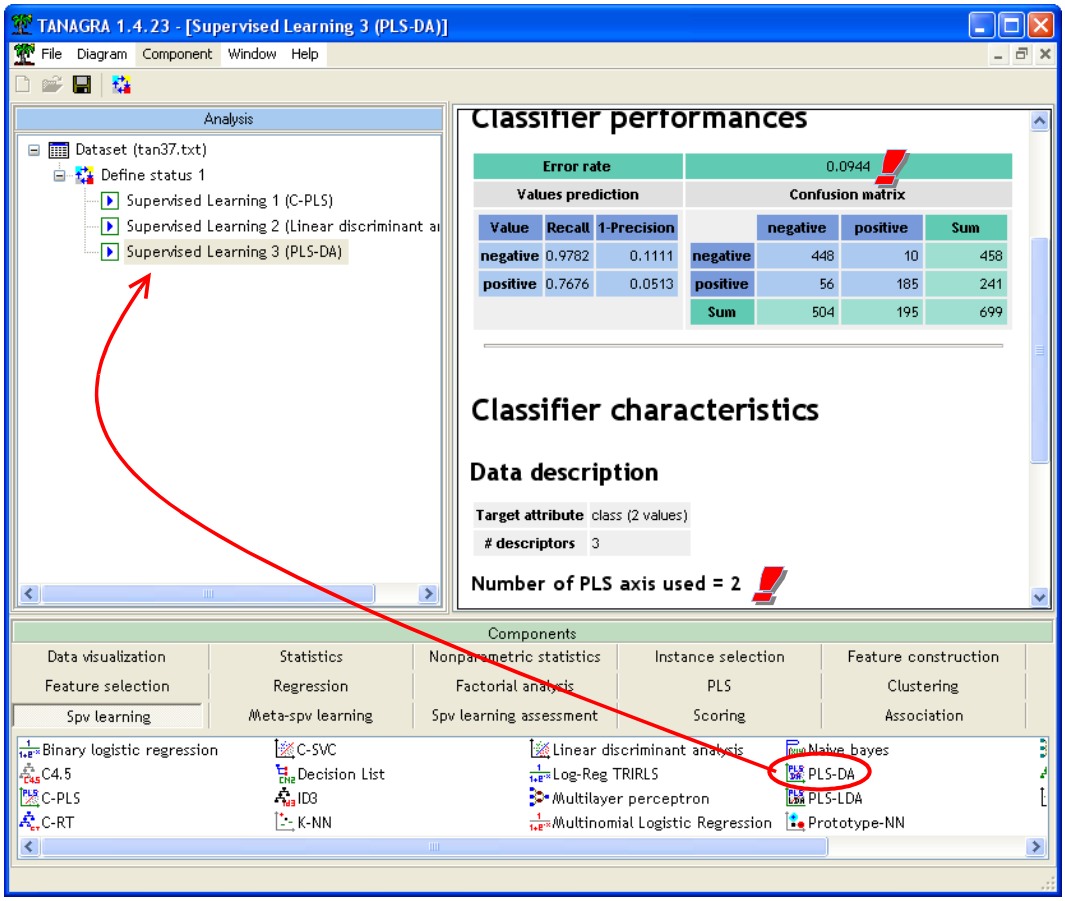
$$Y = y_{k^*} \Leftrightarrow k^* = \arg\max_k \left(\hat{Z}_k\right)$$

Ce composant intègre les mêmes options que le précèdent en ce qui concerne le choix du nombre de facteurs.

4.2.2 Mise en œuvre et lecture des résultats

Nous introduisons le composant PLS-DA (onglet SPV LEARNING) dans le diagramme.

Nous lançons les calculs en activant le menu contextuel VIEW du composant. La fenêtre de visualisation comporte plusieurs parties. Tout d'abord la matrice de confusion et le taux d'erreur en resubstitution, il est égal à 9.44% dans notre exemple. PLS-DA nous indique ensuite qu'il a conservé les deux premiers axes factoriels pour la prédiction.



Les deux fonctions de classement (une pour chaque modalité de la variable à prédire, Figure 2) sont disponibles. Le rôle des variables est évalué à l'aide du critère VIP qui se lit de la même manière que pour le composant C-PLS.

Classification functions

Attribute	negative	positive	VIP
ucellsize	-0.085323	0.085323	1.2038
normnucl	-0.053251	0.053251	1.0346
mitoses	-0.003001	0.003001	0.6933
constant	1.080103	-0.080103	-

Figure 2 - Fonctions de classement PLS-DA

Pour un nouvel individu à classer avec la description suivante (UCELLSIZE = 2 ; NORMNUCL = 1 ; MITOSES = 1), nous devons calculer les deux fonctions et prédire la modalité associée à la valeur la plus élevée. Détaillons cela :

$$\begin{cases} D(-,X) = 1.080 - 0.085 \times 2 - 0.053 \times 1 - 0.003 \times 1 = 0.853 \\ D(+,X) = -0.080 + 0.085 \times 2 + 0.053 \times 1 + 0.003 \times 1 = 0.147 \end{cases}$$
$$\text{Ainsi, } D(-,X) > D(+,X) \Rightarrow Y = -$$

Note : Dans un cadre binaire, on peut s'étonner que C-PLS et PLS-DA ne produisent pas exactement les mêmes résultats comme en atteste les deux matrices de confusion (Figure 1 et Figure 2). Il y a des (petites) différences. Pourtant, la variable à prédire étant binaire, si un individu n'est pas positif, il est forcément négatif, le fait d'utiliser deux indicatrices au lieu d'une ne devrait pas modifier le comportement de l'algorithme PLS.

La différence s'explique par la valeur du code adopté lors du recodage de la variable cible : **le seuil d'affectation n'est pas le même selon les méthodes**. Ce que retranscrit d'ailleurs dans les fonctions de classement : les coefficients associés aux variables sont identiques d'une méthode à l'autre (Class = Positive, Figure 1 et Figure 2), en revanche les constantes sont dissemblables.

Si nous avions adopté le codage 0/1 pour la méthode C-PLS, ne tenant pas compte des poids respectifs des positifs et négatifs, les résultats auraient été strictement identiques.

De même, si la proportion des positifs et négatifs est identique (50% positifs et 50% négatifs), la méthode C-PLS, telle qu'elle est implémentée dans TANAGRA, fournira exactement les mêmes coefficients que PLS-DA.

4.3 Le composant PLS-LDA

4.3.1 Description de la méthode

Dernier composant basé que la régression PLS que nous étudions dans ce didacticiel, la méthode PLS-LDA combine la régression PLS et l'analyse discriminante linéaire.

La méthode s'applique à la prédiction d'**une variable TARGET comportant K (K ≥ 2) modalités.** Il intègre le mécanisme de sélection d'axes de C-PLS et PLS-DA.

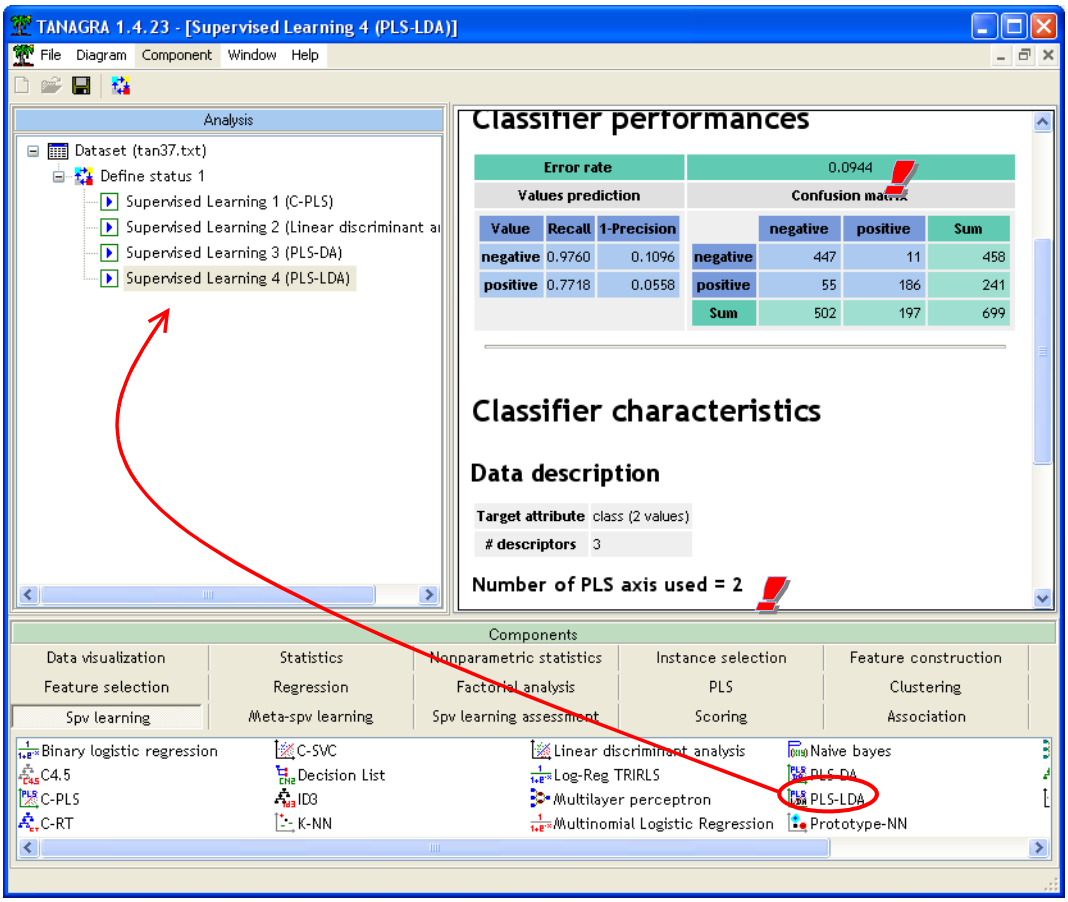
L'apprentissage est réalisé en 2 temps. Dans un premier temps, la variable cible est recodée en K indicatrices 0/1, comme pour PLS-DA. La régression PLS est lancée, nous pouvons spécifier le nombre d'axes, la méthode peut également la déterminer en se basant toujours sur les

redondances de Y. Les axes factoriels sont alors explicitement produits et, dans un deuxième temps, l'analyse discriminante est lancée en prenant en entrée les facteurs de la PLS.

Ces derniers étant deux à deux orthogonaux, l'analyse discriminante est particulièrement stable. Par rapport à l'analyse sur les axes principaux de l'Analyse en Composantes Principales, ce qui constitue une stratégie très populaire de régularisation (voir par exemple http://tutoriels-data-mining.blogspot.com/2008/03/analyse-discriminante-sur-axes.html), les facteurs PLS ont été élaborés en tenant compte de la nature supervisée du problème : la variable cible Y a pesé dans la construction des axes.

4.3.2 Mise en œuvre et lecture des résultats

Nous insérons le composant PLS-LDA (onglet SPV LEARNING dans le diagramme). Nous activons le menu contextuel VIEW pour accéder aux résultats.



La succession PLS + LDA est totalement transparente pour l'utilisateur. La méthode fournit directement les fonctions de classement non standardisés.

La méthode nous indique que 2 axes ont été retenus. La règle de décision est identique à celle de la PLS-DA, à savoir nous appliquons chaque fonction de classement (Figure 3), une par modalité de la variable à prédire, la prédiction correspond à celle qui produit la valeur la plus élevée.

Classification functions

Attribute	negative	positive
ucellsize	-0.426839	0.811171
normnucl	-0.266394	0.506258
mitoses	-0.015013	0.028531
constant	1.102686	-7.271344

Figure 3- Fonctions de classement PLS-LDA

Reprenons notre exemple ci-dessus. Pour un nouvel individu à classer avec la description suivante (UCELLSIZE = 2 ; NORMNUCL = 1 ; MITOSES = 1), nous calculons :

$$\begin{cases} D(-,X) = 1.103 - 0.427 \times 2 - 0.266 \times 1 - 0.015 \times 1 = -0.032 \\ D(+,X) = -7.271 + 0.811 \times 2 + 0.506 \times 1 + 0.029 \times 1 = -5.114 \end{cases}$$

$$\text{Ainsi, } D(-,X) > D(+,X) \Rightarrow Y = -$$

Note : La combinaison des techniques rend difficile l'évaluation du rôle de chaque variable. Pour cette raison, nous n'affichons pas le VIP qui ne traduit pas le rôle de l'analyse discriminante dans la modélisation.

Autre aspect important, le choix du nombre de facteurs incombe uniquement à la Régression PLS dans ce composant. On pourrait imaginer mettre à contribution le processus de sélection de variables de l'analyse discriminante (voir http://tutoriels-data-mining.blogspot.com/2008/03/stepdisc-analyse-discriminante.html) pour déterminer les axes les plus pertinents dans la deuxième phase de calcul. Dans une version prochaine de TANAGRA peut être… Avis aux amateurs.

5 Conclusion

Avec la version 1.4.23 de TANAGRA, nous avons voulu mettre en avant les techniques d'apprentissage supervisé basées sur la Régression PLS, communément appelées « Analyse Discriminante PLS ». Elles sont très populaires dans beaucoup de domaines, elles sont pourtant peu connues de la communauté de l'apprentissage automatique. Pourtant ses caractéristiques sont très intéressantes, voire décisives dans certains contextes, notamment lorsque les descripteurs sont très nombreux, fortement bruités et redondants par exemple. Un canevas qui survient fréquemment dans le DATA MINING.

Dans ce didacticiel, nous montrons comment mettre en œuvre ces méthodes dans TANAGRA, comment lire et exploiter les résultats.