1. Supports

[REF 1] Rakotomalala R., « Analyse discriminante linéaire » ;
http://eric.univ-lyon2.fr/~ricco/cours/slides/analyse_discriminante.pdf

[REF 2] Wikipédia, « Analyse discriminante linéaire » ;
https://fr.wikipedia.org/wiki/Analyse_discriminante_linéaire

[REF 3] Chavent M., « Analyse discriminante linéaire et quadratique », 2015 (en particulier les sections 3 et 5) ;
http://www.math.u-bordeaux.fr/~machaven/wordpress/wp-content/uploads/2013/10/Analyse_discrim.pdf

2. Outils – Excel + Tanagra

1. Chargez et installer le logiciel Tanagra sur votre ordinateur

http://chirouble.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html

2. Intégrez Tanagra dans Excel en tant que macro complémentaire

http://tutoriels-data-mining.blogspot.fr/2010/08/ladd-in-tanagra-pour-excel-2007-et-2010.html

3. Données

Nous utilisons les données « **breast_train_test.xlsx** » en provenance du serveur UCI (Breast Cancer Wisconsin). Il s'agit de prédire le caractère malin ou non d'une cellule (variable CLASSE) à partir de ses propriétés (CLUMP … MITOSES).

a. Le classeur Excel dispose de 2 feuilles : « Apprentissage » contient l'échantillon d'apprentissage ; « Test » correspond à l'échantillon test.
b. On considère que « begnin » de CLASSE est la modalité positive. Ce commentaire est important lorsqu'il s'agira de calculer les ratios d'évaluation (sensibilité, précision) à partir de la matrice de confusion.

4. Modélisation, prédiction, évaluation

Appuyez-vous sur le tutoriel suivant pour l'élaboration du modèle prédictif sous Tanagra : http://tutoriels-data-mining.blogspot.fr/2008/04/analyse-discriminante-linaire.html. Nous ne procédons pas à la sélection de variables dans cette première étape.

A rendre : un fichier Excel avec des commentaires et le détail des calculs sur la feuille TEST du classeur.

1. Ouvrez le fichier « **breast_train_test.xlsx** » dans Excel. Enumérez les variables disponibles et leurs types (quantitative ou qualitative). De combien d'observations disposons-nous dans l'échantillon d'apprentissage ? Dans l'échantillon test ?

2. En suivant le tutoriel : importez les données d'apprentissage dans Tanagra, définissez les variables INPUT et TARGET de l'étude, puis lancer la modélisation avec le composant LINEAR DISCRIMINANT ANALYSIS. Vous devriez obtenir le résultat suivant :

Classifier characteristics

Data description

Target attribute	classe (2 values)
# descriptors	9

MANOVA

Stat	Value	p-value
Wilks' Lambda	0.1677	-
Bartlett -- C(9)	702.6140	0.0000
Rao -- F(9, 390)	215.0572	0.0000

LDA Summary

	Classification functions		Statistical Evaluation			
Attribute	begnin	malignant	Wilks L.	Partial L.	F(1,390)	p-value
clump	0.754582	1.713715	0.191282	0.876741	54.82902	0.000000
ucellsize	-0.199456	0.467517	0.171002	0.980717	7.66830	0.005888
ucellshape	-0.031990	0.459733	0.169424	0.989852	3.99836	0.046238
mgadhesion	-0.031578	-0.091141	0.167778	0.999562	0.17104	0.679413
sepics	0.812397	1.383152	0.171620	0.977187	9.10491	0.002716
bnuclei	0.255668	1.430277	0.215831	0.777020	111.91780	0.000000
bchromatin	0.694874	1.223167	0.171575	0.977445	8.99955	0.002874
normnucl	-0.084429	0.232631	0.170015	0.986410	5.37306	0.020967
mitoses	0.230869	0.214804	0.167708	0.999981	0.00746	0.931211
constant	-3.130793	-23.324683	-			

3. A l'aide du menu COMPONENT / COPY RESULTS de Tanagra, copiez les sorties de l'analyse discriminante **dans la feuille TEST** du classeur Excel. Conservez uniquement le tableau contenant les coefficients des fonctions de classement « Classifications functions ».
4. Calculez, pour chaque individu de l'échantillon test, les scores « begnin » et « malignant », puis en déduire la prédiction du modèle (qui correspond au max. des deux scores) [REF 1, page 6]. Les premières lignes de votre feuille de calcul TEST devrait se présenter comme suit :

clump	ucellsiz	ucellsh	mgadh	sepics	bnuclei	bchron	normnu	mitose	classe	d(begnin)	d(malignant)	prediction
1	1	3	1	2	1	1	1	1	begnin	0.018561	-13.98821	begnin
3	1	1	1	2	1	1	1	1	begnin	1.591705	-11.480246	begnin
5	1	4	1	2	1	3	2	1	begnin	4.310218	-3.994652	begnin
6	2	3	1	2	1	1	1	1	begnin	3.592015	-4.952118	begnin
5	1	3	1	2	1	1	1	1	begnin	3.036889	-7.13335	begnin
4	1	1	1	2	1	3	2	1	begnin	3.651606	-7.087566	begnin
8	6	5	4	3	10	6	1	1	malignant	10.73242	21.362903	malignant
1	1	3	1	2	4	2	1	1	begnin	1.480439	-8.474212	begnin
1	1	3	1	1	1	2	1	1	begnin	-0.098962	-14.148195	begnin
5	1	1	1	2	1	3	1	1	begnin	4.490617	-5.606482	begnin
1	1	1	2	1	3	1	1	7	begnin	1.135116	-12.232591	begnin
6	1	1	1	2	1	3	1	1	begnin	5.245199	-3.892767	begnin
7	5	6	10	4	10	5	3	1	malignant	9.904501	19.719805	malignant

LDA Summary

	Classification functions		Statistical Evaluation			
Attribute	begnin	malignant	Wilks L.	Partial L.	F(1,390)	p-value
clump	0.754582	1.713715	0.191282	0.876741	54.82902	0
ucellsize	-0.199456	0.467517	0.171002	0.980717	7.6683	0.005888
ucellshape	-0.03199	0.459733	0.169424	0.989852	3.99836	0.046238
mgadhesion	-0.031578	-0.091141	0.167778	0.999562	0.17104	0.679413
sepics	0.812397	1.383152	0.17162	0.977187	9.10491	0.002716
bnuclei	0.255668	1.430277	0.215831	0.77702	111.9178	0
bchromatin	0.694874	1.223167	0.171575	0.977445	8.99955	0.002874
normnucl	-0.084429	0.232631	0.170015	0.98641	5.37306	0.020967
mitoses	0.230869	0.214804	0.167708	0.999981	0.00746	0.931211
constant	-3.130793	-23.324683	-			

5. A partir des colonnes CLASSE et PREDICTION, construisez la matrice de confusion et calculez les ratios de performance : taux d'erreur, sensibilité (rappel), précision [INTRO APPRENTISSAGE, page 10], en considérant que « classe = begnin » est la modalité positive de l'étude.
6. La qualité du modèle est-elle satisfaisante ?

5. LDA sous R

Nous réitérons l'analyse sous R en utilisant la procédure **lda()** du package MASS. Inspirez-vous de http://tutoriels-data-mining.blogspot.fr/2012/07/analyse-discriminante-lineaire.html, à partir de la **page 19**.

A rendre : un rapport .docx issu d'un projet R Markdown sous RStudio

1. Importez la première feuille « Apprentissage » de « **breast_train_test.xlsx** » dans un premier data frame que vous nommerez DFApp.
2. Construire le modèle prédictif, « classe » est la cible, les autres variables sont les explicatives. Utilisez la procédure lda() du package MASS qu'il faut charger au préalable. MASS est installé par défaut, il n'est pas nécessaire de l'importer à partir du web.
3. Affichez les coefficients du modèle. Ne vous attardez pas dessus, la présentation de lda() est différente des autres logiciels.
4. Importez la seconde feuille « Test » du classeur dans DFTest.
5. Réalisez la prédiction sur l'échantillon test (predict).
6. Pour construire la matrice de confusion, croisez les valeurs observées de la cible (classe) avec celles prédites par le modèle (table).
7. Calculez alors les différents indicateurs de performance des modèles (taux d'erreur, sensibilité, précision). Vous devez obtenir exactement les mêmes résultats que sous Excel + Tanagra.

6. LDA sous Python

Nous réitérons l'analyse sous Python avec la classe LinearDiscriminantAnalysis de « scikit-learn » (http://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html).
Inspirez-vous du tutoriel : http://tutoriels-data-mining.blogspot.com/2015/09/python-machine-learning-avec-scikit.html

A rendre : un fichier PDF imprimé à partir d'un notebook JUPYTER

Remarque : La séquence des traitements est exactement identique à celle sous R, avec des commandes différentes bien sûr. Petite différence par rapport à R quand même, à l'issue de l'apprentissage, les coefficients et la constante fournies par l'outil sont comparables avec ceux fournis par les autres logiciels. Faites le rapprochement avec les sorties de TANAGRA par exemple. Que constatez-vous ?