ECP, SDM 2015-2016
M. Mougeot,C. Malherbe N. Vayatis,

Analyse discriminante
Régression logistique

Rappel: `help(fonc)` pour obtenir de l'aide sur la fonction nommée "fonc".

Analyse Discriminante

Données de l'étude: Les données analysées sont les données iris, disponibles dans le logiciel R sous le nom "iris" dans une structure de type dataframe. Exécuter successivement chacune des instructions dans la console et analyser les résultats affichés. Ce travail doit être réalisé rapidement ($< 15mn$). Il a pour but de vous présenter, par l'exemple, un certain nombre d'instructions utiles dans l'appréhension des données. Aidez-vous de la commande `help()` pour plus de détail si besoin.

1. `iris` -`nrow(iris)` - `col(iris)` - `dim(iris)` - `names(iris)`
2. `plot(iris)` - `plot(iris,col=c("blue","red","green")[iris$Species])`- `mean(iris)` - `sd(iris)`
3. `iris[1,]` - `iris[,5]` - `iris$Species`
4. `levels(iris$Species)` - `nlevels(iris$Species)`
5. `tapply(iris$Sepal.Width,iris$Species,mean)` - `tapply(iris$Species,iris$Species,length)`

Analyse discriminante linéaire prédictive (ADL)(librairie MASS)

1. Charger la librairie MASS dans l'espace de travail à l'aide de l'instruction `library(MASS)`.
2. Etudier l'aide de la fonction: `help(lda)`. Puis effectuer *une analyse discriminante linéaire* sur les données iris à l'aide de l'instruction: `model=lda(Species~.,data=iris)`.
3. Exécuter l'instruction `print(model)` (ou `model`). Analyser et intérpréter les résultats affichés (prior, règles de décision...)
4. Que donne l'instruction `plot(model)`?
5. Partitionner aléatoirement les données iris en deux partitions: une base d'apprentissage contenant 80% des observations et une base de test contenant les 20% restant à l'aide de la fonction `sample()`.
6. En utilisant *uniquement* les données de la base d'apprentissage, effectuer une analyse discriminante linéaire sur ces données. En utilisant la fonction `predict.lda()` (ou également `predict()`), calculer les prédictions pour chacune des observations.
7. Calculer la matrice de confusion sur la base d'apprentissage en vous aidant de la fonction `table()`. Noter la position dans le tableau des valeurs cibles et des prédictions. Calculer l'erreur de prédiction.
8. Calculer les prédictions, la matrice de confusion et l'erreur sur la base de test. Que remarquez-vous?
9. Repartitionner aléatoirement vos deux bases et ré exécuter les questions précédentes. Que remarquez-vous? conclusion.

Analyse discriminante quadratique prédictive (ADQ) (librairie MASS)
Ce travail est à réaliser hors séance de TP.

1. Etudier l'aide de la fonction: `help(lda)`. Puis, effectuer une analyse quadratique discriminante sur les données

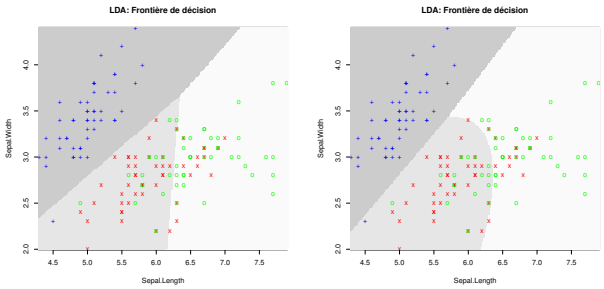

Figure 1: Frontière de discrimination LDA et QDA pour deux variables explicatives Petal.width Petal.Length

2. Comparer les résultats obtenus par ADL et ADQ.

Régression logistique

Applications: le fichier "SAHeart.txt" contient les données historiques d'une étude réaliseé sur les facteurs responsables d'un attaque cardiaque pour $n = 462$ habitants d'Afrique du Sud (`http://www-stat.stanford.edu/~tibs/ElemStatLearn`). Le fichier "SAHeart.info" décrit les variables étudiées et le fichier "SaHeart.txt" contient les données d'étude. Consulter ces deux fichiers. La variable "chd" est la variable de réponse binaire étudiée qui indique si un individu a eu (chd=1) ou pas (chd=0) un incident cardiaque. Les autres variables sont les variables explicatives potentiellement liées à la réponse.

1. Charger les données dans l'environement de travail R. Récupérer dans une structure de type dataframe les variables explicatives suivantes `sbp,tobacco,ldl,famhist,obesity,alcohol,age` et la variable de réponse `chd`. A quoi correspondent ces variables?

2. Visualiser à l'aide d'un scatterplot le jeu de données correspondant aux variables explicatives, en distinguant pour chaque individu le type de réponse (1/0) à l'aide d'un code couleur.

   (`pairs(tab,pch=22,bg=c("red","blue")[unclass(factor(tab[,"chd"]))])`).

3. **Régression logistique:** La fonction `glm()` de R permet d'estimer les paramètres d'un modèle linéaire généralisé. Consulter l'aide cette fonction. Utiliser cette fonction pour estimer les paramètres du modèle. Utiliser la fonction `summary()` pour une description complète de l'objet R.

4. Comparer pour l'ensemble des individus la réponse prédite et la réponse attendue. Calculer la matrice de confusion et le pourcentage de "faux positifs" $P(\hat{Y} = 1/Y = 0)$ et de "faux négatifs" $P(\hat{Y} = 0/Y = 1)$. Conclusion.

5. **Validation croisée:** On souhaite à présent estimer les coefficients du modèle sur 75% des individus (base d'apprentissage), puis évaluer la qualité des résultats obtenus par ce premier modèle sur les 25% des individus restants (base de test). Calculer la matrice de confusion sur les bases de test et d'apprentissage. Répeter cette procédure plusieurs fois et estimer l'erreur min, max, moyenne de classification. Quel est l'intéret d'une telle approche?

6. Effectuer une régression logistique avec sélection de variables en utilisant la fonction `step`. Quels sont les coefficients retenus les plus significatifs, les moins? Que peut-on en conclure?

7. Utiliser la library de R, ROCR, pour construire et visualiser les courbes ROC des 3 modèles suivants: le modèle complet, le modèle sélectionné par la fonction step, le modèle le plus intéressant contenant une seule variable explicative. On visualisera la courbe ROC sur les données d'apprentissage et sur les données de test.