**SY09 Printemps 2009 – UTC**

**Analyses discriminantes quadratique et linéaire**

Exercice I : Règle de Bayes : formes linéaires et quadratique

* *Cas (a), (b), (c), (d), (e) : équation des frontières de décision de la règle de Bayes*

Dans le cas général de la règle de Bayes, les fonctions discriminantes s’écrivent sous la forme :

gk(x) = – (x-μk)’∑-1(x-μk) – ln(det∑k) + ln лk – ln (2л) , ou encore :

gk(x) = – (x)’∑-1(x) + (μk)’ ∑-1(x) – (μk)’ ∑-1 (μk) – ln(det∑k) + ln лk – ln (2л)

Pour obtenir la frontière de décision, il suffit alors de résoudre l’équation gk(x) = gl(x)

* Cas (a) : л1= 0.5, μ1= (0,0)’, μ2= (1,1)’, ∑1= ∑2=

x1 = 1 – x2 => correspond à une équation de droite

* Cas (b) : л1= 0.1, μ1= (0,0)’, μ2= (1,1)’, ∑1= ∑2=

x1 = -1,2 – x2 => correspond à une équation de droite

* Cas (c) : л1= 0.5, μ1= (0,0)’, μ2= (1,1)’, ∑1= ∑2=

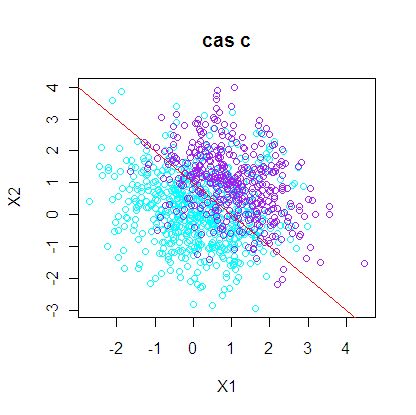
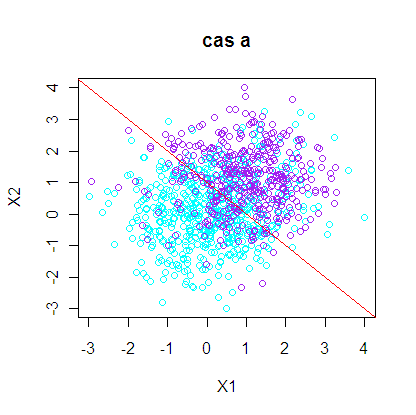
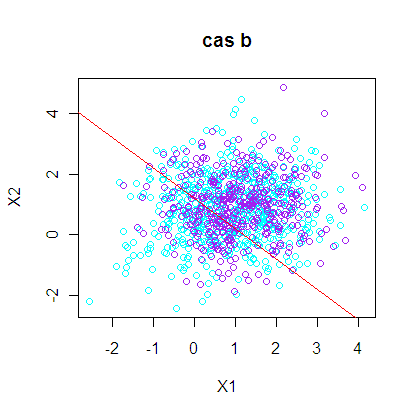
x1 = 1 – x2 => correspond à une équation de droite

* Cas (d) : л1= 0.6, μ1=μ2= (1,1)’, ∑1= ∑2=

(x1 – 1)² - (x2 – 1)² = 5(ln(1,5) + ) => correspond à une équation d’un cercle

* Cas (e) : л1= 0.6, μ1= (0,0)’, μ2= (1,1)’, ∑1= ∑2=

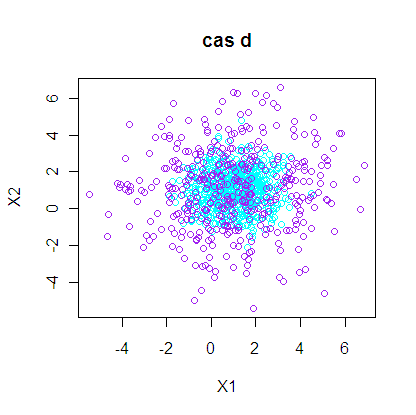
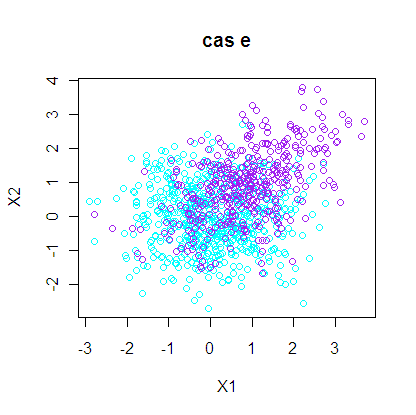
(x1 – 1)² - (x2 – 1)² = 5(ln(1,5) + ) => correspond à une équation d’une ellipse

* *******Cas (a), (b), (c) : représentation des vecteurs, frontières de décisions, estimation de la probabilité d’erreur comparée à la probabilité d’erreur théorique*

(a) erreur estimée : 0,227 (b) erreur estimée : 0,398 (c) erreur estimée : 0,178

erreur théorique : 0,242 erreur théorique : 0,362 erreur théorique : 0,144

* *Cas (d), (e) : représentation des vecteurs, estimation de la probabilité d’erreur comparée à la borne de Bhattacharyya*



(d) erreur estimée : 0,21 (e) erreur estimée : 0,263

borne de Bhattacharyya : 0,365 borne de Bhattacharyya : 0,386

* *Interprétations :*

Pour les 3 premiers cas, les estimations de probabilité d’erreur sont relativement proches des probabilités d’erreur théorique (avec un écart ne dépassant pas les 4%). Les équations des frontières et la répartition dans l’espace des deux classes gaussiennes des chaque cas nous ont permis de conclure quant à une représentation suivant une ADL, séparant ainsi les deux classes d’individus de manière linéaire.

Pour le cas (d), les données semble éparpillées mais centrées autour d’un même point. Graphiquement, il semble donc impossible de séparées les deux classes de manière linéaire. De plus, l’équation de la frontière nous indique une séparation de la forme circulaire. Tout ceci nous permet de conclure quant à une représentation suivant une ADQ, permettant ainsi d’obtenir une frontière de décision quadratique. La probabilité d’erreur relevée reste assez faible (21%), mais plus éloignée et inférieure à la borne de Bhattacharyya (avec un écart de presque 16%).

Pour le cas (e) tout comme pour le cas (d), une représentation quadratique semble la plus appropriée, même si dans ce dernier cas, les données paraissent moins éparpillées que dans le cas (d). L’équation de la frontière de décision va en accord avec cette observation, puisqu’elle nous donne une séparation de forme elliptique entre les deux classes. La probabilité d’erreur relevée reste également assez faible (26%), mais éloignée et inférieure à la borne de Bhattacharyya (avec un écart de presque 15%).

Exercice II : **LDA, QDA sur les données Crabes, et évaluation des performances**

On désire utiliser les méthodes d’analyse discriminante linéaire et l’analyse discriminante quadratique pour connaître l’espèce des crabes à partir du jeu de données utilisé dans les TPs précédents.

* *Explications d’instructions en langage R :*

**Contour** : trace des lignes de niveaux afin de représenter le troisième paramètre sur un plan déterminé par le premier et le deuxième paramètre.

**Sample :** représente un tirage aléatoire à partir des différents paramètres (taille du tirage, la probabilité du tirage des nombres, les nombres à tirer).

**Lda** : analyse discriminante linéaire. La fonction retourne la probabilité d’obtenir une des variables, le coefficient de linéarité et la moyenne des variables par groupe.

**Qda** : analyse discriminante quadratique. La fonction retourne la probabilité d’obtenir une des variables dans l’un des groupes et la moyenne des variables par groupe.

* *Différence entre predict et predict.lda*

La fonction predict.lda hérite de la fonction predict.

* *Calcul de l’estimateur de la probabilité d’erreur :*

La valeur de l’estimateur de la probabilité d’erreur dans la zone d’apprentissage est de 28,5% pour l’analyse discriminante linéaire (ADL) et de 28% pour l’analyse discriminante quadratique (ADQ).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | 1 | 2 | 3 | 4 | 5 | Moyenne |
| Test | ADL | 26,53% | 31,34% | 31,43% | 26,47% | 29,41% | 29,04% |
| ADQ | 25,17% | 29,85% | 27,86% | 30,15% | 27,94% | 28,19% |

Pour le test, on obtient donc 29% pour l’ADL et 28,2% pour l’ADQ. Nous pouvons dire que les résultats sont plutôt optimistes car ils sont très proches de celle obtenue à partir de l’apprentissage.

On constate aussi que c’est le modèle ADQ qui est le meilleur. On peut observer que l’analyse quadratique est plus fine que l’analyse lineaire

* *Calcul de l’estimateur de la probabilité d’erreur sur les 2/3 de l’échantillon:*

Pour cette expérience, nous prenons seulement 2/3 de l’échantillon à l’aide de la fonction *sample.*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | 1 | 2 | 3 | 4 | 5 | Moyenne |
| Apprentissage | lda | 33,83% | 31,58% | 29,32% | 27,07% | 26,32% | 29,62% |
| qda | 29,32% | 33,08% | 29,32% | 30,08% | 24,81% | 29,32% |
| Test | lda | 47,37% | 35,63% | 29,79% | 23,40% | 24,74% | 32,19% |
| qda | 34,74% | 35,63% | 30,85% | 23,40% | 27,84% | 30,49% |

Nous obtenons donc, pour l’apprentissage, une erreur sur l’ADL de 29,62% et sur l’ADQ de 29 ,32%.

Pour le test, l’erreur estimée de l’ADL est de 32,19% et celui de l’ADQ est de 30,49%.

On constate toujours que le meilleur modèle est celui de l’ADQ.

* *Etude sur l’estimateur de la probabilité d’erreur sur différents échantillon:*

Pour le premier essai, nous allons changer la taille de l’échantillon apprentissage et donc aussi de l’échantillon test. Nous allons donc changer la probabilité à 50% pour l’apprentissage au lieu de 30%.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | 1 | 2 | 3 | 4 | 5 | Moyenne |
| Test | lda | 32,99% | 29,36% | 27,68% | 21,51% | 35,48% | 29,40% |
| qda | 28,87% | 28,44% | 30,36% | 21,51% | 31,18% | 28,07% |

On remarque que les valeurs sont plus précises lors de l’analyse discriminante quadratique. Les valeurs sont à peu près pareilles que lors de l’apprentissage (28,5% pour l’ADL et de 28% pour l’ADQ).

Toutes les erreurs diminuent en fonction de la taille de l’échantillon : Plus le nombre d’observation pour l’apprentissage sera élevé, plus on pourra analyser finement l’ensemble de test, et donc moins l’erreur sera grande.

On aurait pu aussi remarquer une grande différence entre l’ADQ lors de l’apprentissage et lors du test. Lors de l’apprentissage, l’ADQ peut donner un faible pourcentage d’erreur puisque cette analyse adapte trop ses données c’est-à-dire que le résultat de l’analyse peut être une illusion pour l’échantillon d’apprentissage et pour l’échantillon test, le résultat donnera une forte erreur.

L’analyse linéaire donnera toujours, pour les différents échantillons, une valeur conforme à la valeur trouvé à partir de l’échantillon d’apprentissage.