SY09 - TP04
Analyses discriminantes quadratique et linéaire

Bertrand Bon - Antoine Hars

June 15, 2013

Introduction

Dans le cadre de ce tp, nous avons étudié les analyses discriminantes quadratique et linéaire.

Exercice 1 : Règle de Bayes.

On suppose que la population est répartie en deux classes, en proportions π_1 et $\pi_2 = 1$ - π_1, issues des distributions gaussiennes bivariées $\mathcal{N}(\mu_1, \Sigma_1)$ et $\mathcal{N}(\mu_1, \Sigma_1)$.

1. Donner une équation de la frontière de décision de la règle de Bayes dans chacun des cas suivants :

(a) $\pi_1 = \mathbf{0.5}$, $\mu_1 = \mathbf{(0,0)}'$, $\mu_2 = \mathbf{(1,1)}'$, $\Sigma_1 = \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$:

(b) $\pi_1 = \mathbf{0.1}$, $\mu_1 = \mathbf{(0,0)}'$, $\mu_2 = \mathbf{(1,1)}'$, $\Sigma_1 = \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$:

(c) $\pi_1 = \mathbf{0.5}$, $\mu_1 = \mathbf{(0,0)}'$, $\mu_2 = \mathbf{(1,1)}'$, $\Sigma_1 = \Sigma_2 = \begin{pmatrix} 1 & -0.3 \\ -0.3 & 1 \end{pmatrix}$:

(d) $\pi_1 = \mathbf{0.6}$, $\mu_1 = \mu_2 = \mathbf{(1,1)}'$, $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}$:

(e) $\pi_1 = \mathbf{0.6}$, $\mu_1 = \mathbf{(0,0)}'$, $\mu_2 = \mathbf{(1,1)}'$, $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$:

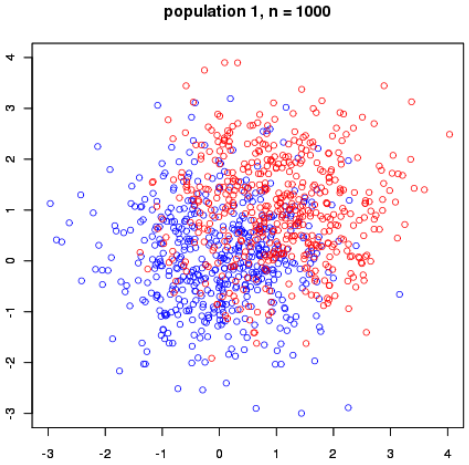
2. Simulation de la règle de Bayes dans R :

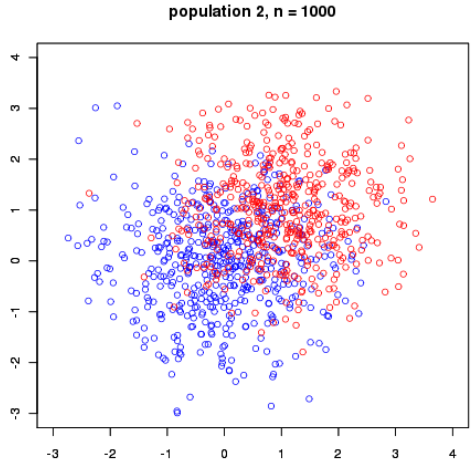
Pour chacune des cinq populations précédentes, en utilisant la fonction *simul* réalisée au TD3 (Théorie de la décision), nous avons généré un échantillon de taille $n = 1000$.
Pour rappel, le code de la fonction *simul* est comme suit :

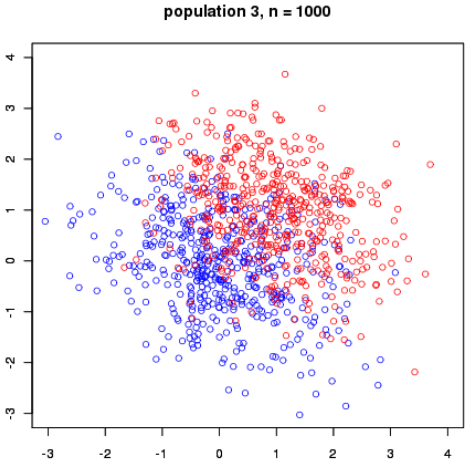
```
simul <- function (n, pi, mu1, mu2, sigma1, sigma2) {

  # On crée la matrice contenant l'échantillon résultat de la fonction.
  result = matrix(nrow = n, ncol = 3)

  # Affectation de chaque élément de l'échantillon final à la classe 0 ou 1.
  for (k in 1:n) {

    rand = sample(0:1, 1)
    result[k, 3] = rand

    if (result[k, 3] == 0) {
```

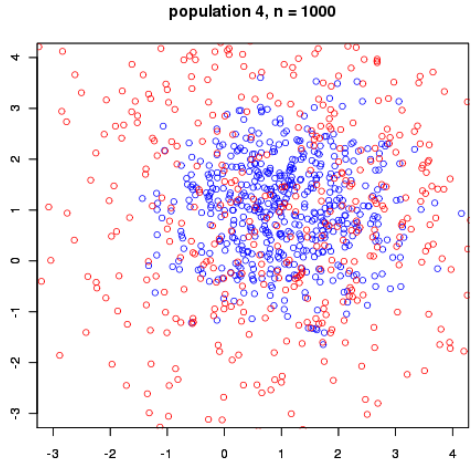
```
      result[k, c(1, 2)] = mvrnorm(1, mu1, sigma1)
    } else {
      result[k, c(1, 2)] = mvrnorm(1, mu2, sigma2)
    }
  }

  return (result)
}
```

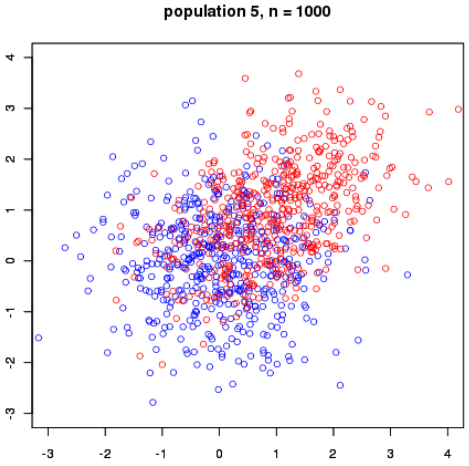
Pour chacun des échantillons de population, nous avons tracé les nuages suivants associés, avec le tracé de la frontière de décision pour les trois premiers cas :











Pour chaque cas de figure, nous avons déterminé l'expression d'un estimateur de la probabilité d'erreur, ainsi que sa réalisation sur l'échantillon correspondant :
Sa réalisation sur les échantillons nous donne les valeur suivantes :

population	μ_1	μ_2	Erreur estimée (%)	Erreur théorique (%)
1	$\begin{pmatrix} -0.028 \\ -0.093 \end{pmatrix}$	$\begin{pmatrix} 1.062 \\ 0.981 \end{pmatrix}$	27.6	NA
2	$\begin{pmatrix} 0.346 \\ 0.006 \end{pmatrix}$	$\begin{pmatrix} 1.015 \\ 0.982 \end{pmatrix}$	31.1	NA
3	$\begin{pmatrix} -0.063 \\ -0.045 \end{pmatrix}$	$\begin{pmatrix} 1.036 \\ 0.915 \end{pmatrix}$	30.6	NA
4	$\begin{pmatrix} 1.018 \\ 0.919 \end{pmatrix}$	$\begin{pmatrix} 1.221 \\ 0.965 \end{pmatrix}$	49	NA
5	$\begin{pmatrix} -0.127 \\ -0.018 \end{pmatrix}$	$\begin{pmatrix} 1.001 \\ 1.002 \end{pmatrix}$	28.2	NA

Cela nous a permis de le comparer avec la probabilité d'erreur théorique.

Exercice 2 : Analyse discriminante sur les données *Crabes.*

Dans cet exercice, nous désirons utiliser l'analyse discriminante linéaire et l'analyse discriminante quadratique sur les données *crabs* afin de déterminer une fonction permettant de distinguer le sexe à partir des mesures *FL* et *RW*.

1. Expliquer ce que font les fonctions suivantes :

lda : La fonction lda sert à effectuer l'analyse discriminante linéaire de données (elle prend en paramètre une formule, un data frame ou une matrice). Elle cherche à détecter si la matrice de covariance d'une classe est singulière.

qda : Cette fonction est utilisée pour exécuter une analyse discriminante quadratique sur des données en utilisant une décomposition QR qui retournera un message d'erreur si la variance du groupe est singulière pour chaque groupe.

contour : Il s'agit d'une fonction générique utile pour créer un graphe de contour ou pour ajouter des lignes de contour à un graphe existant. Dans notre cas, elle est utile pour tracer les frontières de décision sur nos graphiques

sample : Cette fonction nous permet de récupérer un échantillon de taille spécifiée d'éléments de l'ensemble X en remettant en place ou non les éléments.

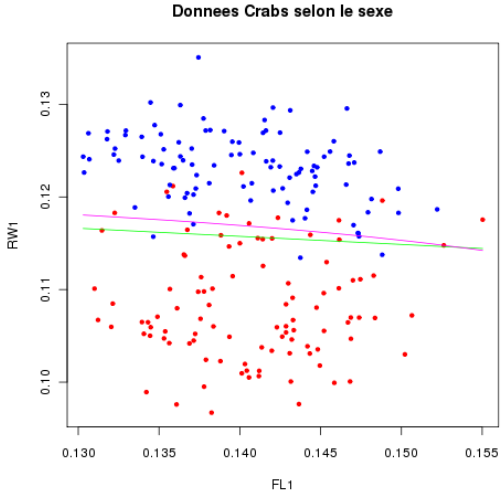
predict : Predict() est une fonction générique de prédictions à partir des résultats des diverses fonctions de création de modèles. La forme retournée dépend de la classe entrée en paramètre.

predict.lda : Cette fonction classifie des observations multi-variables en utilisant l'analyse discrimante linéraire et projette les données sur les discriminantes linéaires. Cette fonction centre les discriminants linéaires de sorte que le nombre moyen pondéré des centres de gravité du groupe soit à l'origine.

Comparaison entre predict et predict.lda : predict.lda() est une méthode de la fonction générique predict() pour la classe *lda*. On peut soit appeler predict() sur une classe lda d'un objet spécifié ou appeler la fonction predict.lda() sans se soucier de la classe de l'objet.

2. L'analyse discriminante quadratique et l'analyse discriminante linéaire des données *crabs*

Nous avons d'abord effectué ces deux analyses sur les données *crabs* en prenant comme échantillon d'apprentissage l'ensemble des données.
nous avons ensuite tracé les frontières de décision que nous avons obtenu (verte pour la lda et magenta pour la qda) :



Nous pouvons remarquer sur le graphique que les frontières de décision de chacune des deux analyses discriminantes sont sensiblement différentes, et que certaines valeurs des données ne sont donc pas du

même côté par rapport aux deux frontières.

Nous avons donc calculé les estimations d'erreur sur cet échantillon d'apprentissage pour les deux analyses :

Type d'analyse	Estimation d'erreur (%)
lda	9.5
qda	8.5

Nous pouvons observer d'après ces deux valeurs que l'analyse discrimante linéaire semble moins précise que l'analyse discrimante quadratique vu que la probabilité d'erreur pour la première est plus grande que la seconde.
Pour notre jeu de données, nous pouvons dire que les crabes ont sensiblement plus de chances d'être mieux classés avec l'analyse discrimante quadratique qu'avec l'analyse discrimante linéaire.

3.

4.

Exercice 3 :

Conclusion