Rapport SY09 TP03 - Théorie de la décision

Jean-Baptiste Audibert & Yueqing Qin

27 mai 2011

1 Classifieur euclidien

1.1 Création de l'échantillon

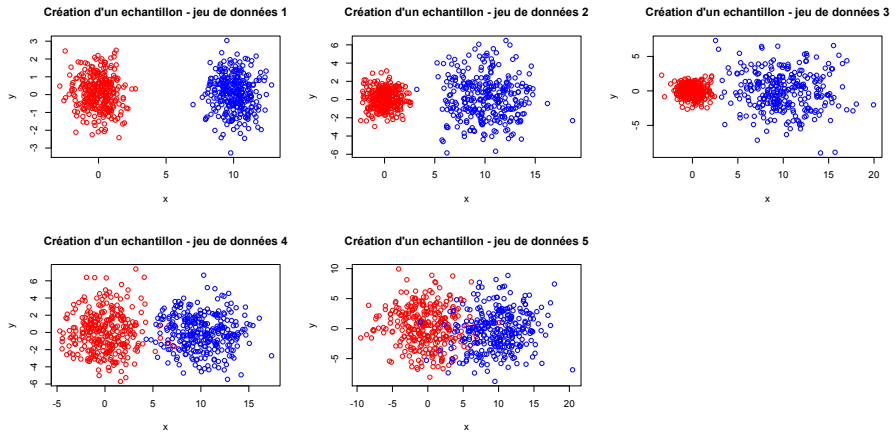
On souhaite disposer d'un échantillon de deux classes ω_1 et ω_2 de taille $n = 600$, et dont la proportion des classes est de 0.5. On créé donc une fonction **simul** pour ce faire.
On a alors l'échantillon suivant :

- Classe $\omega_1 \sim \mathcal{N}(\mu_1, \sum_1),\ \mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sum_1 = a_1 Id$
- Classe $\omega_2 \sim \mathcal{N}(\mu_2, \sum_2),\ \mu_2 = \begin{pmatrix} 10 \\ 0 \end{pmatrix}, \sum_2 = a_2 Id$

Où a_1 et a_2 varient :

- $a_1 = 1,\ a_2 = 1$ - Jeu de données (1)
- $a_1 = 1,\ a_2 = 6$ - Jeu de données (2)
- $a_1 = 1,\ a_2 = 9$ - Jeu de données (3)
- $a_1 = 5,\ a_2 = 5$ - Jeu de données (4)
- $a_1 = 10,\ a_2 = 10$ - Jeu de données (5)

Qui donne l'affichage suivant :



On observe ici que lorsque la variance augmente, la dispersion des points dans les nuages de points augmente également, ce qui logique. On va s'intéresser ici au fait que certains points de la classe ω_1 (resp. ω_2), vont, lorsque la variance est suffisamment élevée, être si proche de la classe ω_2 (resp. ω_1), que l'on peut se tromper sur la classe à laquelle ils appartiennent.

1.2 Estimation de la probabilité d'erreur

A l'aide des fonctions **regleEuclidienne** et **erreurEstime**, nous allons la probabilité d'erreur sur le choix de la classe pour différents points x_i appartenant à la classe ω_1 ou ω_2. On calcule donc cette probabilité pour les cinq jeux de données présentés précédemment, on obtient les résultats suivants :

Jeu (1)	Jeu (2)	Jeu (3)	Jeu (4)	Jeu (5)
0 %	1.67 %	2.33 %	1.03 %	5.35 %

1.3 Estimation de l'erreur moyenne

On souhaite désormais étudier cette probabilité sur plusieurs essais. On répète donc *10 fois* la génération et l'estimation de la probabilité d'erreur. On obtient les résultats suivants :

Jeu de donné	Erreur Moyenne	Variance	Intervalle de confiance
Jeu 1	0	0	/
Jeu 2	$0,87\%$	$3,03.10^{-5}$	$[0,47; 1,26]$
Jeu 3	$2,53\%$	$23,73.10^{-5}$	$[1,43; 3,64]$
Jeu 4	$1,71\%$	$8,79.10^{-5}$	$[1,03; 2,37]$
Jeu 5	$6,34\%$	$68,64.10^{-5}$	$[4,48; 8,21]$

On voit ici qu'il est très intéressant de réaliser plusieurs essais afin de caractériser au mieux la probabilité d'erreur. En effet, dans le premier essai, la moyenne, pour un jeu donné, n'était pas ou à peine dans l'intervalle de confiance obtenu avec 10 essais.

2 Etude de la règle de la règle de Bayes

2.1 Construction de f_1 et f_2

On a, avec $X = (X_1, X_2)^T$:

- $f_{11}(x_1) \sim \mathcal{N}(-1, 1)$,
- $f_{21}(x_1) \sim \mathcal{N}(1, 1)$,
- $f_{12}(x_2) = f_{22} \sim \mathcal{N}(0, 1)$,
- $f_1(x) = f_{11}(x_1)f_{12}(x_2)$, la densité conditionnelles dans la classe ω_1,
- $f_2(x) = f_{21}(x_1)f_{22}(x_2)$, la densité conditionnelles dans la classe ω_2.

Soit :

$$\begin{aligned} f_1(x) &= f_{11}(x_1)f_{12}(x_2) \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_1+1)^2} . \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_2)^2} \\ &= \frac{1}{2\pi} e^{-\frac{1}{2}(x_1+1)^2 - \frac{1}{2}(x_2)^2} \end{aligned}$$

D'où $f_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ avec $\mu_1 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$, et $\sigma_1 = I$

et

$$\begin{aligned} f_2(x) &= f_{21}(x_1) f_{22}(x_2) \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_1-1)^2} . \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_2)^2} \\ &= \frac{1}{2\pi} e^{-\frac{1}{2}(x_1-1)^2 - \frac{1}{2}(x_2)^2} \end{aligned}$$

D'où $f_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ avec $\mu_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, et $\sigma_2 = I$

2.2 Estimation des paramètres de f_1 et f_2

Maintenant que l'on connait les densités conditionnelles dans chacune des classes, on peut estimer les paramètres intrinsèques à ces densités.
Tout d'abord pour la classe ω_1 :

Population	Espérance	Variance
10	$\begin{pmatrix} -1.50 \\ 0.38 \end{pmatrix}$	$\begin{pmatrix} 1.17 & -0.11 \\ -0.11 & 0.72 \end{pmatrix}$
100	$\begin{pmatrix} -1.11 \\ 0.17 \end{pmatrix}$	$\begin{pmatrix} 1.16 & -0.06 \\ -0.06 & 1.04 \end{pmatrix}$
1000	$\begin{pmatrix} -0.96 \\ 0.06 \end{pmatrix}$	$\begin{pmatrix} 1.02 & -0.01 \\ -0.01 & 0.93 \end{pmatrix}$
10000	$\begin{pmatrix} -1.00 \\ 0.01 \end{pmatrix}$	$\begin{pmatrix} 0.99 & 0 \\ 0 & 1.01 \end{pmatrix}$
100000	$\begin{pmatrix} -1 \\ 0.00 \end{pmatrix}$	$\begin{pmatrix} 1.00 & 0 \\ 0 & 1.01 \end{pmatrix}$

Puis pour la classe ω_2 :

Population	Espérance	Variance
10	$\begin{pmatrix} -0.99 \\ 0.37 \end{pmatrix}$	$\begin{pmatrix} 0.18 & -0.41 \\ -0.41 & 1.50 \end{pmatrix}$
100	$\begin{pmatrix} 0.90 \\ 0.28 \end{pmatrix}$	$\begin{pmatrix} 0.88 & -0.32 \\ -0.32 & 1.00 \end{pmatrix}$
1000	$\begin{pmatrix} 0.98 \\ -0.06 \end{pmatrix}$	$\begin{pmatrix} 1.05 & -0.01 \\ -0.01 & 0.99 \end{pmatrix}$
10000	$\begin{pmatrix} 1.02 \\ -0.01 \end{pmatrix}$	$\begin{pmatrix} 0.97 & 0.01 \\ 0.01 & 0.98 \end{pmatrix}$
100000	$\begin{pmatrix} 1.00 \\ 0.00 \end{pmatrix}$	$\begin{pmatrix} 0.99 & 0.00 \\ 0.00 & 0.99 \end{pmatrix}$

On observe que plus la population est élevée, plus on se rapproche des résultats théoriques. Ceci est dû à **la loi des grands nombres**, qui nous dit que plus la taille d'un échantillon aléatoire augmente, plus ses caractéristiques statistiques sont proches de la population réelle.

2.3 Etude des cercles d'isodensité

On souhaite montrer que les courbes d'isodensité de f_1 et f_2 sont des cercles, on pose donc
- $f_1(x) = K_1$ où K_1 est une constante,
- $f_2(x) = K_2$ où K_2 est également une constante.

De ceci, on obtient :

$$\begin{aligned} f_1(x) &= K_1 \\ \frac{1}{2\pi} e^{-\frac{1}{2}(x_1+1)^2 - \frac{1}{2}(x_2)^2} &= K_1 \\ e^{-\frac{1}{2}(x_1+1)^2 - \frac{1}{2}(x_2)^2} &= 2\pi K_1 \\ (x_1+1)^2 + (x_2)^2 &= -2ln(2\pi K_1) \end{aligned}$$

On a bien ici un cercle de centre $c_1 = (-1, 0)$, et de rayon $r_1 = \sqrt{-2ln(2\pi K_1)}$, avec $K_1 \in]0; 1[$
De même :

$$\begin{aligned} f_2(x) &= K_2 \\ \frac{1}{2\pi} e^{-\frac{1}{2}(x_1-1)^2 - \frac{1}{2}(x_2)^2} &= K_2 \\ e^{-\frac{1}{2}(x_1-1)^2 - \frac{1}{2}(x_2)^2} &= 2\pi K_2 \\ (x_1-1)^2 + (x_2)^2 &= -2ln(2\pi K_2) \end{aligned}$$

On obtient également un cercle de centre $c_2 = (1, 0)$ et de rayon $r_2 = \sqrt{-2ln(2\pi K_2)}$, avec $K_2 \in]0; 1[$

2.4 Utilisation de la règle de Bayes

Tout d'abord, les risques associés à chacune des décisions (choix de la classe ω_1 ou ω_2) sont :

$$\begin{aligned} R(a_1) &= c_{11}\mathbb{P}(\omega_1|x) + c_{12}\mathbb{P}(\omega_2|x) \\ R(a_2) &= c_{21}\mathbb{P}(\omega_1|x) + c_{22}\mathbb{P}(\omega_2|x) \end{aligned}$$

Avec $\mathbb{P}(\omega_2|x) = \frac{\pi_2 f_2(x)}{f(x)}$, $\mathbb{P}(\omega_1|x) = \frac{\pi_1 f_1(x)}{f(x)}$ et $f(x) = \pi_1 f_1(x) + \pi_2 f_2(x)$.

La règle de Bayes nous dit que si l'on décide a_1 :

$$\begin{aligned} c_{11}\mathbb{P}(\omega_1|x) + c_{12}\mathbb{P}(\omega_2|x) &< c_{21}\mathbb{P}(\omega_1|x) + c_{22}\mathbb{P}(\omega_2|x) \\ (c_{11} - c_{21})\,\mathbb{P}(\omega_1|x) &< (c_{12} - c_{22})\,\mathbb{P}(\omega_2|x) \\ (c_{11} - c_{21})\, f_1(x)\pi_1 &< (c_{12} - c_{22})\, f_2(x)\pi_2 \\ (c_{21} - c_{11})\, f_1(x)\pi_1 &> (c_{22} - c_{12})\, f_2(x)\pi_2 \end{aligned}$$

Soit finalement $\delta^*(x)$, la règle de Bayes pour ce cas :

$$\delta^*(x) = \begin{cases} a_1 & \text{si } \frac{f_1(x)}{f_2(x)} > \left(\frac{c_{12}-c_{22}}{c_{21}-c_{11}}\right)\frac{\pi_2}{\pi_1} \\ a_2 & \text{sinon} \end{cases}$$

Or $c_{11} = c_{22} = 0$, donc finalement :

$$\delta^*(x) = \begin{cases} a_1 & \text{si } \frac{f_1(x)}{f_2(x)} > \left(\frac{c_{12}}{c_{21}}\right)\frac{\pi_2}{\pi_1} \quad (1) \\ a_2 & \text{sinon} \end{cases}$$

On peut également calculer :

$$\begin{aligned} \frac{f_1(x)}{f_2(x)} &= \frac{\frac{1}{2\pi}e^{-\frac{1}{2}(x_1+1)^2-\frac{1}{2}(x_2)^2}}{\frac{1}{2\pi}e^{-\frac{1}{2}(x_1-1)^2-\frac{1}{2}(x_2)^2}} \\ &= \frac{e^{-\frac{1}{2}(x_1+1)^2}}{e^{-\frac{1}{2}(x_1-1)^2}} \\ &= \frac{e^{-\frac{1}{2}(x_1^2+2x_1+1)}}{e^{-\frac{1}{2}(x_1^2-2x_1+1)}} \\ &= e^{-2x_1} \end{aligned}$$

Ce qui nous donne

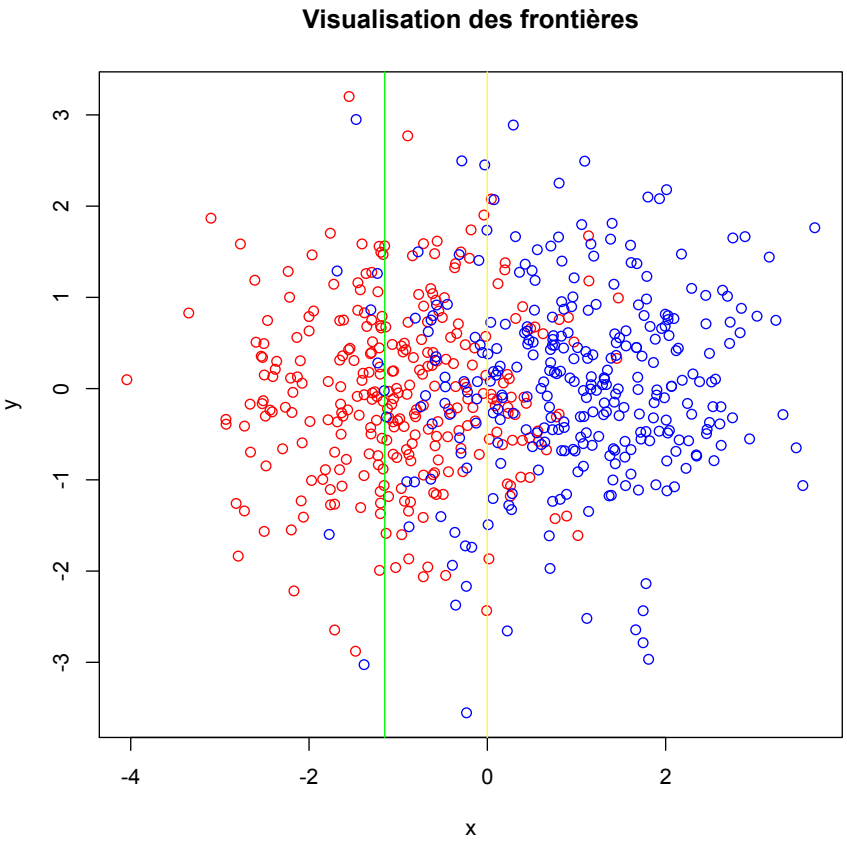
$$\begin{aligned} (1) &\Leftrightarrow -2x_1 > ln\left(\frac{c_{12}}{c_{21}}\frac{\pi_2}{\pi_1}\right) \\ &\Leftrightarrow x_1 < \frac{-1}{2}\ ln\left(\frac{c_{12}}{c_{21}}\frac{\pi_2}{\pi_1}\right) \end{aligned}$$

2.5 Tracée des frontières de décision

L'équation des frontières est donc donnée par $x_1 = \frac{-1}{2}\ ln\left(\frac{c_{12}}{c_{21}}\frac{\pi_2}{\pi_1}\right)$. Traçons ces frontières pour 3 cas :

- $c_{12} = c_{21} = 1,\ \pi_1 = \pi_2$ soit $x_1 = \frac{-1}{2}ln(1) = 0$, (cas 1)
- $c_{12} = 10,\ c_{21} = 1,\ \pi_1 = \pi_2$ soit $x_1 = \frac{-1}{2}ln(10)$, (cas 2)
- $c_{12} = c_{21} = 1,\ \pi_2 = 10\pi_1$ soit $x_1 = \frac{-1}{2}ln(10)$, (cas 3$\Leftrightarrow$ cas 2)

Avec un échantillon composé de 600 individus.



On a donc ici 3 frontières, deux vertes en $x_1 = \frac{-1}{2}ln(10)$ et une en jaune en $x_1 = \frac{-1}{2}ln(1) = 0$

2.6 Estimation des paramètres α et β

On a $\alpha = \mathbb{P}(\delta^*(\mathbf{X}) = a_2|w_1)$ et $\beta = \mathbb{P}(\delta^*(\mathbf{X}) = a_1|w_2)$, les probabilités d'erreur que l'on souhaite estimer dans les 3 cas précédents (en fait 2).

Pour cela, on compte le nombre d'individus de la classe ω_1 considérés, grâce à la frontière, comme étant dans la classe ω_2, et on divise par la population de la classe ω_1 (α) et le nombre d'individus de la classe ω_2 considérés comme étant dans la classe ω_1, divisé par le nombre d'individus de la classe ω_2 (β).
On obtient alors :

- cas 1 : $\hat{\alpha} = 0.172$, $\hat{\beta} = 0.173$
- cas 2 et 3 : $\hat{\alpha} = 0.578$, $\hat{\beta} = 0.014$

Dans le premier cas, $\hat{\alpha} \cong \hat{\beta}$, ce qui est logique puisque les probabilités à priori sont égales, tout comme les couts d'erreur.

Dans le cas 3, $\hat{\alpha} \gg \hat{\beta}$. Ceci est dû est fait que la probabilité à priori de la classe ω_2 est dix fois supérieur à celle de classe ω_1. Les individus de ω_1, identifiés dans ω_2, sont donc très nombreux, et inversement.

Enfin dans le cas 2, on a également un facteur 10 sur le cout d'erreur, d'où l'égalité des résultats.