Rapport SY09 TP04 - Théorie de la décision

Jean-Baptiste Audibert & Yueqing Qin

15 juin 2011

1 Etude de la règle de Bayes

Dans cet exercice, nous allons étudier la règle de Bayes pour un échantillon de la population, répartis en deux classe ω_1 et ω_2, de proportions π_1 et $\pi_2 = 1 - \pi_1$. Ces classes sont issus de distributions gaussiennes bivariées :

- Classe $\omega_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$,
- Classe $\omega_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$.

Où π_1, π_2, μ_1, μ_2, Σ_1 et Σ_2 varient. On étudie 5 cas, pour chacun on génère un nuage échantillon de taille totale $n = 1000$, on calcule la frontière entre les deux classes, et on affiche le nuage de points ainsi que la frontière associée.

1.1 Equation des frontières

Pour retrouver l'équation des frontières, on calcule pour chaque cas et pour chaque classe la **fonction discriminante** $g_k(x)$. L'expression de celle ci, pour différentes classes, issus de distributions gaussiennes bivariées, de paramètres μ_k et Σ_k différents est la suivante :

$$g_k(x) = -\tfrac{1}{2}(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k) - \tfrac{1}{2}\ln(\det\Sigma_k) + \ln\pi_k - \tfrac{p}{2}\ln(2\pi). \quad (1)$$

1.1.1 Cas 1

On se place dans le cas où les variances Σ_1 et Σ_2 sont égales - **hypothèse d'homoscédasticité** :

$$g_1(x) = -\tfrac{1}{2}(x_1 \; x_2)\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix} + \ln\pi_1$$

$$g_2(x) = -\tfrac{1}{2}(x_1 - 1 \;\; x_2 - 1)\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}x_1 - 1\\ x_2 - 1\end{pmatrix} + \ln\pi_2$$

Pour obtenir l'équation de la frontière, on pose

$$\begin{aligned}
g_1(x) &= g_2(x)\\
-\tfrac{1}{2}x_1^2 - \tfrac{1}{2}x_2^2 + \ln\pi_1 &= -\tfrac{1}{2}(x_1-1)^2 - \tfrac{1}{2}(x_2-1)^2 + \ln\pi_2\\
x_1^2 + x_2^2 &= (x_1^2 - 2x_1 + 1) + (x_2^2 - 2x_2 + 1) + \ln\left(\tfrac{\pi_2}{\pi_1}\right)\\
x_1 &= 1 - x_2 + \ln\left(\tfrac{\pi_2}{\pi_1}\right) = 1 - x_2
\end{aligned}$$

C'est l'équation d'une **droite**.

1.1.2 Cas 2

On réalise le même développement, mais cette fois ci $\pi_1 \neq \pi_2$:

$$\begin{aligned} x_1 &= 1 - x_2 + \ln\left(\frac{\pi_2}{\pi_1}\right) \\ x_1 &= 1 - x_2 + \ln\left(\frac{0,1}{0,9}\right) \\ x_1 &\approx -x_2 - 1,2 \end{aligned}$$

Qui est également l'équation d'une **droite**.

1.1.3 Cas 3

On obtient la même équation que dans le *Cas 1*, l'équation d'une **droite**. :

$$x_1 = 1 - x_2$$

1.1.4 Cas 4

On repart cette fois ci de l'équation (1) pour trouver les fonctions discriminantes :

$$\begin{aligned} g_1(x) &= -\tfrac{1}{2}(x_1 - 1)^2 - \tfrac{1}{2}(x_2 - 1)^2 + \ln\pi_1 - \ln 2\pi \\ g_2(x) &= -\tfrac{1}{10}(x_1 - 1)^2 - \tfrac{1}{10}(x_2 - 1)^2 - \tfrac{1}{2}\ln(25) + \ln\pi_2 - \ln 2\pi \end{aligned}$$

Soit l'équation de la frontière :

$$(x_1 - 1)^2 + (x_2 - 2)^2 = \tfrac{5}{4}\ln(25) + \tfrac{5}{2}\ln\left(\frac{\pi_2}{\pi_1}\right)$$

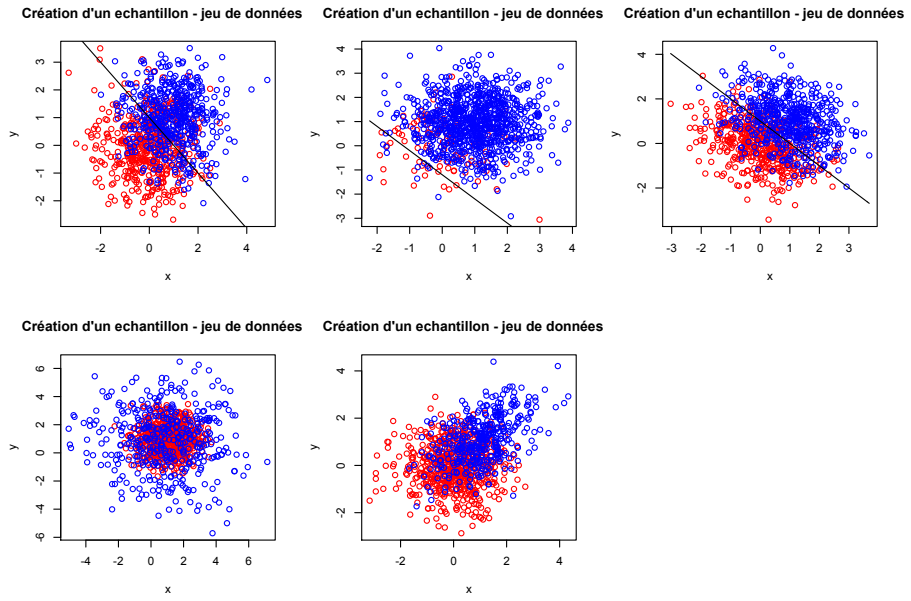
Qui est donc un **cercle**, de centre $\mu = (1, 1)^T$ et de rayon $r = \frac{5}{4}\ln(25) + \frac{5}{2}\ln\left(\frac{\pi_2}{\pi_1}\right)$

1.1.5 Cas 5

On utilise la même démarche que dans le Cas 4 et on obtient

$$-\tfrac{1}{2} + \ln(\pi_1) = -\tfrac{4}{6}(x_1^2 - x_1 x_2^2) - \tfrac{2}{3}(x_1 + x_2) - \tfrac{2}{3} - \tfrac{1}{2}\ln\left(\tfrac{3}{4}\right) + \ln(\pi_2)$$

Ceci est l'équation d'une **ellipse**.
Une fois les frontières calculées, on affiche les échantillons et les frontières :



1.2 Etude de la probabilité d'erreur

Les probabilités d'erreur théoriques nous sont données par la probabilité d'erreur de Bayes et la borne de Bhattacharyya (l'un ou l'autre selon les hypothèses), avec les paramètres c, le nombre de classes, Σ_k, les matrice de variance, ϕ la fonction de répartition de la loi normale, $\Delta = (\mu_2 - \mu_1)'\Sigma^{-1}(\mu_2 - \mu_1)$ le carré de la distance de Mahalanobis, et $\Delta_B^2 = \frac{1}{8}(\mu_2-\mu_1)'\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{-1}(\mu_2-\mu_1)+\frac{1}{2}\ln\frac{det\frac{\Sigma_1+\Sigma_2}{2}}{\sqrt{det\Sigma_1 det\Sigma_2}}$, le carré de la distance de Bhattacharyya entre les deux classes

- Probabilité d'erreur de Bayes : $c = 2$, $\Sigma_k = \Sigma$:
  $\epsilon^* = \phi\left(\frac{ln(\pi_1/\pi_2) - \Delta^2/2}{\Delta}\right)\pi_2 + \left[1 - \phi\left(\frac{ln(\pi_1/\pi_2) + \Delta^2/2}{\Delta}\right)\right]\pi_1$

- Borne de Bhattacharyya (borne supérieur de l'erreur de Bayes) $c = 2$ et Σ_k quelconques :
  $\epsilon^* = \sqrt{\pi_1\pi_2}e^{-\Delta_B^2}$

On obtient alors les résultats suivants :

Cas	Réalisation	Erreur théorique
Cas 1	$0,222$	$0,228$
Cas 2	$0,312$	$0,223$
Cas 3	$0,204$	$0,183$
Cas 4	$0,390$	$0,393$
Cas 5	$0,335$	$0,357$

2 Analyse discriminante sur les données Crabes

2.1 Présentation des fonctions

La fonction **lda** permet de réaliser *directement* **l'analyse discriminante linéaire**.

La fonction **qda** permet de réaliser *directement* textbfl'analyse discriminante quadratique.

La fonction **contour** permet de tracer, sur un graphique, les lignes de niveaux associées à une fonctions dépendant de deux variables.

Enfin, la fonction **sample** permet de tirer, de manière aléatoire un sous échantillon issu d'un échantillon existant.

La différence entre les fonctions **predict** et **predict.lda** réside dans le fait que la première est utilisée selon le type de données étudié, alors que **predict.lda** réalise le même travaille mais uniquement pour des données issues de la fonction **lda**.

2.2 Analyse discriminante linéaire et quadratique

On réalise tout d'abord l'analyse discriminante linéaire sur l'ensemble des crabes, et on affiche les points à partir des variables réduites *FL1* et *RW1*, qui permettent de déterminer le sexe. On a les résultats suivants :

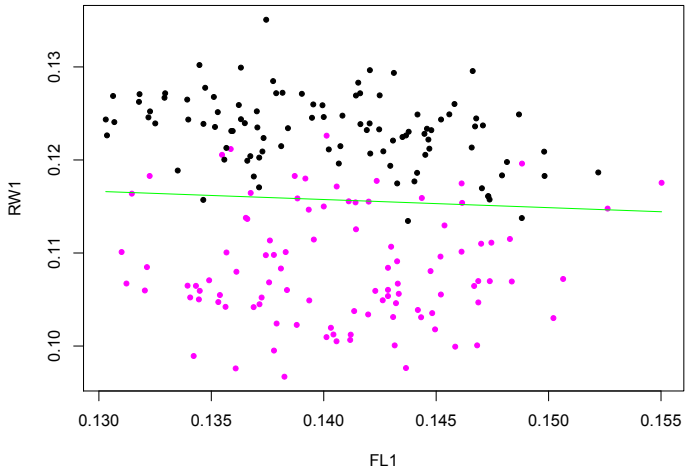


FIGURE 1 – Réalisation de l'analyse discriminante linéaire sur les données crabes

Ainsi qu'une erreur de 9.5%.
Pour l'analyse discriminante quadratique, on a les résultats suivants :

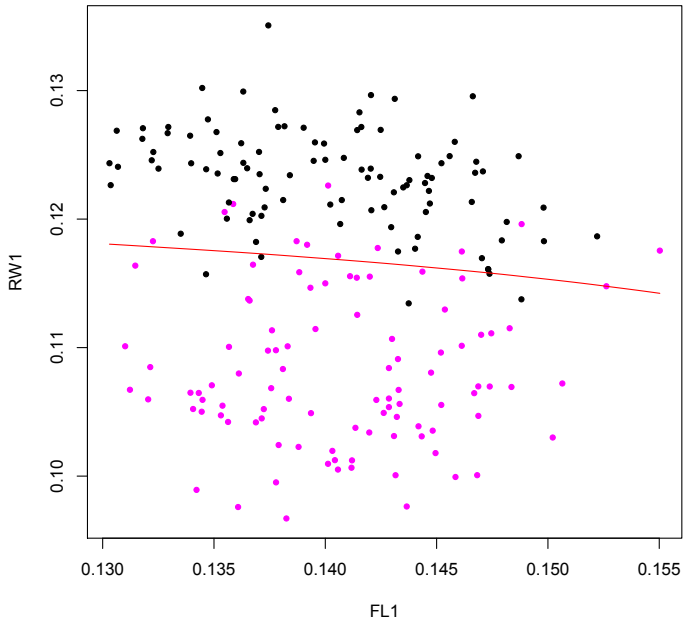


FIGURE 2 – Réalisation de l'analyse discriminante quadratique sur les données crabes

Et une erreur de 8.5%. On remarque que ces erreurs sont très proches, car les frontières de décisions sont également proches. Ceci s'explique par le fait que les deux classes sont **linéairement séparables**, impliquant que l'analyse quadratique, analyse sans hypothèse qui pourrait fausser l'analyse discriminante, n'apporte pas des résultat plus pertinents.

2.3 Estimation de l'erreur d'apprentissage avec un ensemble de test

On souhaite désormais réalise l'analyse discriminante avec un **échantillon d'apprentissage** et un **échantillon de test** distincts. On sélectionne tout d'abord $\frac{2}{3}$ des données *crabs* pour l'échantillon d'apprentissage, et les $\frac{1}{3}$ restants pour l'échantillon de test. On répète ce processus pour différentes proportions de découpages et on obtient les résultats :

Proportion	Nombre d'estimation	LDA	QDA
$\frac{2}{3}$-$\frac{1}{3}$	1	$3,91\%$	4%
$\frac{2}{3}$-$\frac{1}{3}$	200	$2,98\%$	2.35%
$\frac{1}{2}$-$\frac{1}{2}$	1	$5,5\%$	$3,25\%$
$\frac{1}{2}$-$\frac{1}{2}$	200	$4,58\%$	$3,92\%$

On voit ici tout d'abord que la répétition du processus d'analyse permet d'approcher au mieux l'erreur d'apprentissage. De plus, on observe toujours que les résultats des analyses discriminante linéaire et quadratique sont proches, et que cette dernière n'apporte pas de résultats plus précis. Enfin, on observe que plus l'échantillon d'apprentissage est grand, moins l'erreur est importante.

3 Etude des différents classifieurs

3.1 Estimation des paramètres du modèle

3.1.1 Cas 1

On suppose $\pi_1 = \pi_2$ et $\Sigma_1 = \Sigma_2 = \sigma^2 I$, et on estime les paramètres :

- $\hat{\pi}_1 = \frac{n_1}{n} = 0.5$, $\hat{\pi}_2 = 0.5$
- $\hat{\mu}_1 = \frac{1}{n_1} \sum_{i=1}^{n} t_{i1} x_i = (0.64, 0.50)^T$, $\hat{\mu}_2 = (2.12, 2)^T$
- $S = \frac{1}{n-c} \sum_{k=1}^{c} (n_k - 1) S_k$
  $S = \begin{pmatrix} 1.38 & -0,66 \\ -0,66 & 0.86 \end{pmatrix}$, d'où l'estimateur $\hat{\sigma}^2 = \frac{1}{2} trace(S) = 1,12$

3.1.2 Cas 2

On suppose $\Sigma_1 = \Sigma_2$, et on estime les paramètres :

- $\hat{\pi}_1 = \hat{\pi}_2 = 0.5$
- $\hat{\mu}_1 = (0.64, 0.50)^T$, $\hat{\mu}_2 = (2.12, 2)^T$
- On utilise l'estimateur sans biais pour estimer la variance :
  $S = \frac{1}{n-c} \sum_{k=1}^{c} (n_k - 1) S_k$ avec $S_k = \frac{n_k}{n_k - 1} \hat{\Sigma}_K$
  $S = \begin{pmatrix} 1.38 & -0.66 \\ -0.66 & 0.86 \end{pmatrix}$

3.1.3 Cas 3

- $\hat{\pi}_1 = \frac{n_1}{n} = 0.5$, $\hat{\pi}_2 = 0.5$
- $\hat{\mu}_1 = (0.64, 0.50)^T$, $\hat{\mu}_2 = (2.12, 2)^T$
- $S_1 = \frac{n_1}{n_1 - 1} diag(\hat{\Sigma}_1) = \begin{pmatrix} 1.45 & 0 \\ 0 & 1.18 \end{pmatrix}$
  $S_2 = \begin{pmatrix} 1.31 & 0 \\ 0 & 0.535 \end{pmatrix}$

3.1.4 Cas 4

- $\hat{\pi}_1 = \frac{n_1}{n} = 0.5$, $\hat{\pi}_2 = 0.5$
- $\hat{\mu}_1 = (0.64, 0.50)^T$, $\hat{\mu}_2 = (2.12, 2)^T$

– On utilise l'estimateur sans biais pour estimer la variance :
$S_1 = \frac{n_1}{n_1-1}\hat{\Sigma}_1 = \begin{pmatrix} 1.45 & -0.91 \\ -0.91 & 1.18 \end{pmatrix}$
$S_2 = \begin{pmatrix} 1.31 & -0.42 \\ -0.42 & 0.535 \end{pmatrix}$

3.2 Lien avec les différents classifeurs

– Le cas 1 correspond au **classifieur euclidien**,

– Le second cas correspond à **l'analyse discriminante linéaire**

– Le troisième cas correspond au **classifieur bayésiens naïfs**

– Le dernier à **l'analyse discriminante quadratique**

3.3 Fonction discriminantes

L'expression de la fonction de décision nous est donné par :

$$g(x) = \frac{1}{2}(x-\mu_1)'\Sigma_1^{-1}(x-\mu_1) - \frac{1}{2}(x-\mu_2)'\Sigma_2^{-1}(x-\mu_2) + \frac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_2|} - \frac{\pi_1}{\pi_2}$$

On l'applique dans les différents cas, et on trouve les résultats suivants :

3.4 Classifieur euclidien

$g(x) = 1,48x_1 + 1,5x_2 - 3,92$

3.5 Analyse disciminante linéaire

$g(x) = 3.87x_1 + 5,23x_2 - 11,87$

3.6 Classifieur bayésiens naïfs

$g(x) = -0,04x_1^2 - 0,52x_2^2 + 1,18x_1 + 3,35 - 4,79$

3.7 Analyse discriminante quadratique

$g(x) = 0,17x_1^2 - 0,41x_2^2 + 0,26x_1x_2 + 2,31 + 5,12x_2 - 9,48$