Carolina BRUNO
Clara PAGANI

TP03 – SY09

Théorie de la décision

Printemps 2013

1 Classificateur euclidien

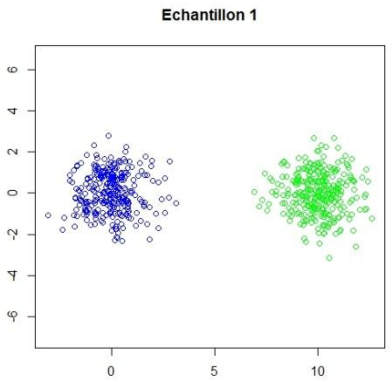
Le but de cette partie est d'étudier le classificateur euclidien sur des échantillons de données issues de deux classes distinctes dans R2 et dont les distributions sont normales.

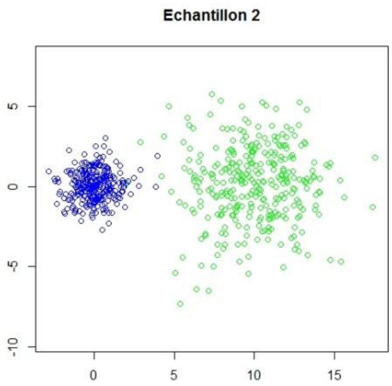
Premièrement nous créons une fonction *simul()* qui permet de simuler un ensemble de données aléatoire. Cet ensemble est constitué d'un échantillon de chaque classe de taille aléatoire (le nombre d'observations totales étant 600. Ces deux échantillons sont mélangés aléatoirement pour former l'échantillon final de taille 600. Chaque échantillon suit une loi normale respectivement de paramètres (u1,E1) et (u2,E2).

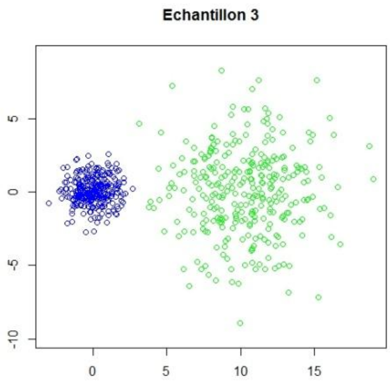
Nous avons : u1 = (0,0) et u2 = (10,0). E1 = a1.I E2 = a2.I

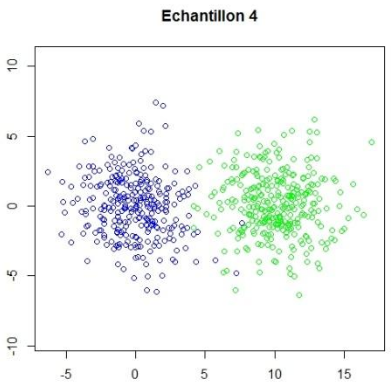
Nous allons simuler 5 types d'échantillons différents en faisant varier les matrices de covariances, i.e. en faisant varier a1 et a2.

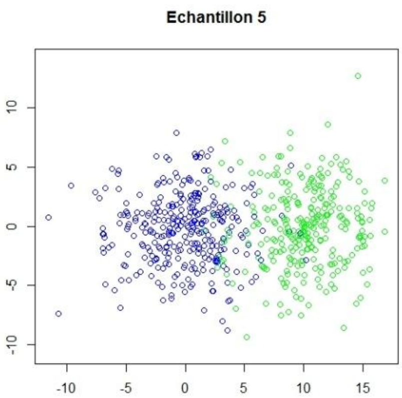
Les 5 graphes ci-dessous représentent les 5 échantillons simulés selon la valeur du couple (a1,a2) = {(1,1),(1,6),(1,9),(5,5),(10,10)}.











Nous remarquons que la variation des paramètres a1 et a2 change la dispersion des échantillons. Ainsi plus la variance est grande, plus les données de chaque échantillon sont dispersées. De plus, nous remarquons qu'il est aisé de distinguer les deux classes dans les quatre premiers échantillons. Cela est plus difficile dans l'échantillon 5 (10,10).

Nous allons maintenant scinder (en deux parties égales) chaque échantillon afin d'obtenir un échantillon d'apprentissage et un échantillon de test. L'échantillon d'apprentissage nous permet de déterminer les estimateurs de u1 et u2. L'échantillon de test nous permet d'estimer le taux d'erreur du classificateur euclidien.

Premièrement, nous créons une fonction *regleEuclidienne()* permettant de déterminer si une observation est plus proche de la classe w1 ou w2 en fonction du calcul de la distance euclidienne entre les coordonnées de cette observation et celles des estimateurs de u1 et u2.

Afin de déterminer le taux d'erreur du classificateur euclidien nous établissons la fonction *erreurEstimee().* Cette fonction permet de comparer, pour chaque observation de l'échantillon test, le classement euclidien des données à la valeur de classe initiale réelle. Le nombre de différence de classement divisé par le nombre d'observations de l'échantillon test nous permet de déterminer le taux d'erreur.

Pour améliorer la fiabilité de nos résultats, nous répétons 10 fois tout ce processus pour chaque valeur de (a1,a2). Nous avons automatisé cette répétition dans un fonction *repeat().* Nous calculons pour chaque cas, la moyenne, la variance ainsi que l'intervalle de confiance (de niveau 5%) du taux d'erreur. Les résultats finaux sont consignés dans le tableau ci-dessous :

(a1,a2)	Moyenne	Variance	Intervalle de confiance
(1,1)	0	0	
(1,6)	0.7 %	0.004 %	[0.24% ; 1.15%]
(1,9)	2.33 %	0.0029%	[1.94% ; 2.72%]
(5,5)	1.5 %	0.0072 %	[0.89% ; 2.10%]
(10,10)	5.4 %	0.010 %	[4.68% ; 6.12 %]

Nous remarquons un taux d'erreur assez élevé (de l'ordre de 5 %) concernant l'échantillon 4, i.e. l'échantillon ayant des matrices de covariance avec les paramètres(10,10).Les autres taux d'erreurs sont relativement faibles. Nous retrouvons bien ce que nous avions observé au point 1, à savoir une distinction des classes plus difficile sur cet échantillon donc un taux d'erreur plus important. L'échantillon 1 à un taux d'erreur de 0% ce qui n'est pas surprenant au vu de la représentation graphique d'un de ces échantillons ci-dessus : les deux classes sont très distinctement séparées.

2 Règle de Neyman-Pearson, règle de Bayes

On considère un problème de détection de cibles dans lequel la classe ω_1 correspond aux missiles et la classe ω_2 correspond aux avions. Chaque variable suit, dans chaque classe, une loi normale avec les paramètres suivants :

$$f_{11}(x_1) \sim N(-1,1), \qquad f_{21}(x_1) \sim N(1,1), \qquad f_{12}(x_2) = f_{22}(x_2) \sim N(0,1)$$

On suppose l'indépendance conditionnelle de x_1 et x_2. Les densités conditionnelles du vecteur X= $(x_1, x_2)^T$ sont donc $f_1(x) = f_{11}(x_1)\, f_{12}(x_2)$ dans la classe ω_1 et $f_2(x) = f_{21}(x_1)\, f_{22}(x_2)$ dans la classe ω_2.

2.1 f_1 ET f_2 SONT DES DISTRIBUTIONS NORMALES

Analyse discriminante avec une règle d'affectation probabiliste :

$$fk(x) = \frac{1}{(2\pi)^{p/2}\sqrt{\text{dét} \sum_k}} \text{ x } \exp\left(-\frac{1}{2}(x-\mu_k)^T \textstyle\sum_k^{-1}(x-\mu_k)\right)$$

Avec $\sum_k$: matrice de covariance de la classe k et μ_k : centre de la gaussienne de la classe k.
Sachant que p = 2, en appliquant cette règle à la distribution $f1$ et $f2$, nous obtenons le résultat suivant:

$f_1(x)$	$f_2(x)$
$f_1(x) = f_{11}(x_1) \times f_{12}(x_2) \sim N(-1,1) \times N(0,1)$ $f_1(x) = \frac{1}{2\pi\sigma_{11}\sigma_{12}} \times e^{-\frac{1}{2\sigma_1}(x_1-\mu_{11})-\frac{1}{2\sigma_2}(x_2-\mu_{12}))}$ Avec, $\mu_{11} = -1$, $\sigma_{11} = 1$, $\mu_{12} = 0$, $\sigma_{12} = 1$ Et $\mu_1 = \binom{\mu_{11}}{\mu_{12}} = \binom{-1}{0}$, $\Sigma_1 = \begin{pmatrix} \sigma_{11}^2 & 0 \\ 0 & \sigma_{12}^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$f_2(x) = f_{21}(x_1) \times f_{22}(x_2) \sim N(1,1) \times N(0,1)$ $f_2(x) = \frac{1}{2\pi\sigma_{21}\sigma_{22}} \times e^{-\frac{1}{2\sigma_1}(x_1-\mu_{21})-\frac{1}{2\sigma_2}(x_2-\mu_{22}))}$ Avec, $\mu_{21} = 1$, $\sigma_{21} = 1$, $\mu_{22} = 0$, $\sigma_{22} = 1$ Et $\mu_1 = \binom{\mu_{21}}{\mu_{22}} = \binom{1}{0}$, $\Sigma_2 = \begin{pmatrix} \sigma_{21}^2 & 0 \\ 0 & \sigma_{22}^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
$f_1(x) = \frac{1}{2\pi} \times e^{(-\frac{1}{2}(x_1+1)+(x_2))}$	$f_2(x) = \frac{1}{2\pi} \times e^{(-\frac{1}{2}(x_1-1)+(x_2))}$

2.2 SIMULATION D'ÉCHANTILLON

Ci-après, les résultats de la simulation des échantillons permettant de déterminer les estimateurs de f1 et f2. Nous avons effectué cette simulation pour les valeurs de n suivantes : 10, 100, 1000, 10000, 100000.

Taille	10	100	1000	10000	100000
μ_1	$\binom{-1.09}{-0.04}$	$\binom{-0.82}{0.18}$	$\binom{-1.02}{0.01}$	$\binom{-1}{0.01}$	$\binom{-0.99}{-0.01}$
μ_2	$\binom{1.15}{-0.63}$	$\binom{1.01}{-0.05}$	$\binom{1.02}{0.03}$	$\binom{0.95}{-0.02}$	$\binom{1}{-0.01}$
$\sum_1$	$\begin{pmatrix} 2.86 & 0.03 \\ 0.03 & 0.35 \end{pmatrix}$	$\begin{pmatrix} 0.83 & -0.01 \\ -0.01 & 0.66 \end{pmatrix}$	$\begin{pmatrix} 0.08 & 0.01 \\ 0.01 & 0.09 \end{pmatrix}$	$\begin{pmatrix} 0.99 & 0.03 \\ 0.03 & 1.03 \end{pmatrix}$	$\begin{pmatrix} 0.99 & 0 \\ 0 & 1 \end{pmatrix}$
$\sum_2$	$\begin{pmatrix} 0.83 & 0.47 \\ 0.47 & 0.98 \end{pmatrix}$	$\begin{pmatrix} 0.99 & 0.90 \\ -0.04 & 0.84 \end{pmatrix}$	$\begin{pmatrix} 0.97 & -0.07 \\ -0.07 & 0.87 \end{pmatrix}$	$\begin{pmatrix} 1.03 & 0 \\ 0 & 0.97 \end{pmatrix}$	$\begin{pmatrix} 0.99 & 0 \\ 0 & 0.99 \end{pmatrix}$

Nous constatons très clairement avec ces résultats que plus la taille de l'échantillon est importante, plus les estimateurs se rapprochent de la valeur théorique réelle.

2.3 LES COURBES D'ISO-DENSITÉ

Pour déterminer les courbes d'iso-densité, nous résolvons l'équation $\boldsymbol{f_i(x)} = \mathbf{K_i}$ avec Ki une constante.
Nous avons pour f1(x) :

$$f1\,(x) = \frac{1}{2\pi} x\, e^{\left(-\frac{1}{2}(x-\mu_1)' + (x-\mu_1)\right)} = K1$$

$$\leftrightarrow\ -ln(2\pi) - \frac{((x-u1)' + (x-u1))}{2} = ln(K1)$$ avec $(x-\mu_1) = \binom{x1-1}{x2-0}$

$$\leftrightarrow (x1+1)^2 + x2^2 = -2\,ln(2\pi K1)$$

Avec $-2\ln 2\pi\, k_1 > 0 \leftrightarrow 2\pi\, k_1 > 1 \leftrightarrow k_1 = k_2 < \frac{1}{2\pi}$

Nous faisons la même démarche pour f2(x) et nous obtenons les résultats ci-dessous :

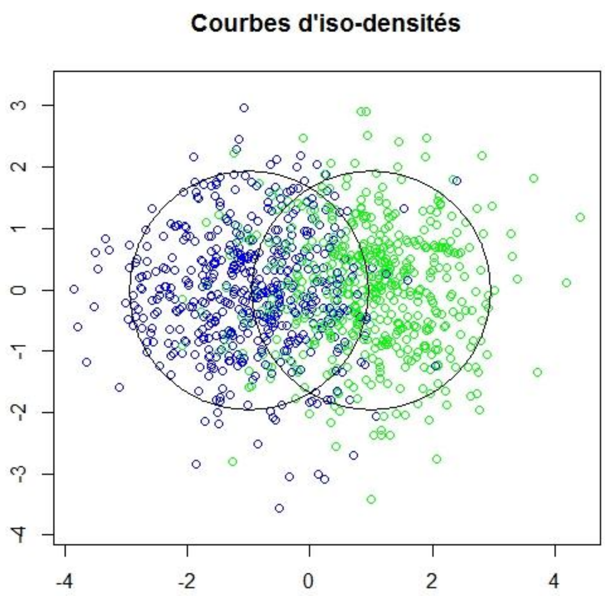
$f_1(x)$ = K_1 $\leftrightarrow (x_1 - (-1))^2 + (x_2 - 0)^2 = R^2 = 2\ln 2\pi$ K_1, où le **rayon R=** $\sqrt{-\mathbf{2ln2\pi}\, \boldsymbol{k_1}}$ et le

centre du cercle $\boldsymbol{c_1} = \binom{-1}{0}$
$f_2(x) = K_2 \leftrightarrow (x_1 - 1)^2 + (x_2 - 0)^2 = R^2 = 2\ln 2\pi\ K_2$, où le **rayon R=**$\sqrt{-2\ln 2\pi\ k_2}$ et le centre du cercle $\boldsymbol{c_2} = \binom{1}{0}$

Ainsi nos deux courbes d'iso-densité sont des cercles de centre respectifs $\boldsymbol{\mu_1} = \binom{-1}{0}$ et $\boldsymbol{\mu_2} = \binom{1}{0}$ et de rayons communs R.

Nous savons que x1 suit une loi normale centrée réduite N(-1,1) et que x2 suit de même un loi normale centrée réduite N(0,1). Ainsi la somme de ces deux distributions élevées au carré va suivre une loi du Chi2 de degré 2.
Nous définissons la distance au centre $D^2 = (x1 - 1)^2 + x2$
Nous pouvons en déduire que $P(D^2 \leq R^2)$ suit une loi du Chi2 de degré 2.
Nous avons donc $R^2 = \frac{(u_\alpha + \sqrt{3})^2}{2}$.
En choisissantu8k, α = 0.950, nous obtenons R = 1.96.

Nous traçons ensuite nos courbes d'iso densité sur le graph obtenu précédemment avec n = 1000.



Nous observons que les courbes d'iso densités de chaque classe semblent bien représenter les zones respectives aux données de chaque classe.

2.4 RÈGLE DE BAYES

Soient π_1 et π_2 les probabilités a priori des deux classes, et c_{lk} le coût associé au choix de l'action a_l lorsque la vraie classe est ω_k. On suppose $c_{11} = c_{22} = 0$. L'ensemble A des actions est le même que dans la question précédente.

La règle de Bayes pour notre problème est la suivante:

$$\delta(x) = a_1 \leftrightarrow r(\delta(x)) = c_{11}\,\mathbb{P}(\omega_1|\,x) + c_{12}\,\mathbb{P}(\omega_2|\,x) = r_1(x)$$

$$\delta(x) = a_2 \leftrightarrow r(\delta(x)) = c_{21}\,\mathbb{P}(\omega_1|\,x) + c_{22}\,\mathbb{P}(\omega_2|\,x) = r_2(x)$$

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

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

Le δ^* minimisant $r(\delta|x)$ pour un x:

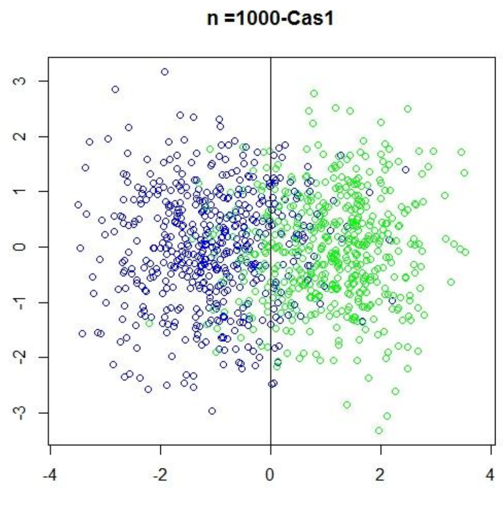
$$\delta^*(x) = \begin{cases} a_1 \text{ si } r_1(x) < r_2(x) \\ a_2 \text{ sinon} \end{cases} \leftrightarrow \delta^*(x) = \begin{cases} a_1 \text{ si } \frac{f_1(x)}{f_2(x)} < \frac{c_{12}c_{22}\pi_2}{c_{21}c_{11}\pi_1} \\ a_2 \text{ sinon} \end{cases}$$

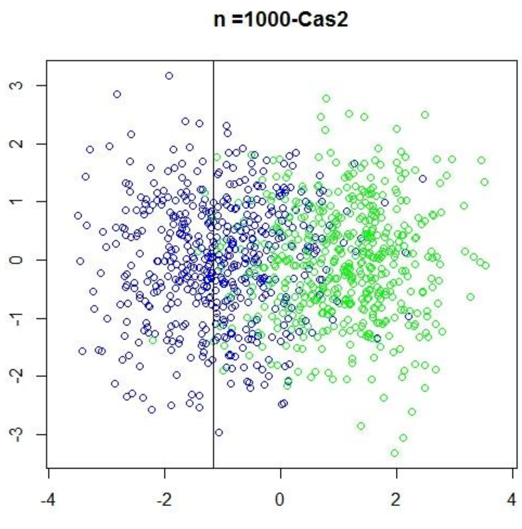
$$\delta^*(x) = \begin{cases} a_1 \text{ si } e^{-2x_1} > \frac{c_{11}\pi_2}{c_{21}\pi_1} \\ a_2 \text{ sinon} \end{cases} \leftrightarrow \delta^*(x) = \begin{cases} a_1 \text{ si } x_1 < \frac{ln(c_{12}) + ln(\pi_2) - ln(c_{21}) - ln\,(\pi_1)}{-2} \\ a_2 \text{ sinon} \end{cases}$$

Nous remarquons alors que la frontière de décision ne dépend pas du paramètre x2. Ainsi ce sera une droite perpendiculaire à l'axe des abscisses.

Afin de déterminer les frontières de décision, nous allons calculer pour chaque cas l'équation de cette droite tel que x1 = $\frac{ln(c_{12}) + ln(\pi_2) - ln(c_{21}) - ln\,(\pi_1)}{-2}$.

Cas 1. c12 = c21 = 1, π1 = π2 x1 = 0
Cas 2. c12 = 10, c21 = 1, π1 = π2 x1 = -1.151
Cas 3. c12 = c21 = 1, π2 = 10π1. X1 = 1.099





Dans le cas 1, nous observons que la frontière sépare plutôt bien les deux classes, ceci paraît logique sachant que les côuts de choix sont les mêmes pour ces deux classes et que la probabilité des deux classes est égales.

Dans le cas 2, nous observons que la frontière est décalée vers la gauche comparé au cas 1. Ceci est dû au fait que le coût de choix de la classe 1 est 10 fois supérieur à celui de la classe 2. Les erreurs d'affectation de choix de classe sont d'autant plus importantes.

Nous évaluons maintenant les estimateurs de risques pour chacun des cas. Pour cela, nous nous appuyons sur les formules suivantes :

$$\hat{\alpha} = \frac{nb\ données\ mal\ classées\ dans\ w1}{nb\ total\ de\ données\ dans\ w1}$$

$$\hat{\beta} = \frac{nb\ données\ mal\ classées\ dans\ w2}{nb\ total\ de\ données\ dans\ w2}$$

Afin d'évaluer ces deux estimateurs pour chacun des cas, nous vérifions dans chacun des échantillons si la classe associé répond bien au critère de choix préalablement déterminé avec la règle de Bayes.

Cas	$\hat{\boldsymbol{\alpha}}$	$\hat{\boldsymbol{\beta}}$
Cas1	0.1653226	0.1673387
Cas2	0.5542406	0.01006036

Nous retrouvons bien des résultats en cohérences avec ceux de la question précédente. Dans le cas 1, l'estimation d'erreur de classement des données est équivalente entre les deux classes. Dans le cas 2, l'estimation d'erreur de classement des données diffèrent grandement selon la classe. Sans surprise, nous constatons que la classe 1, à une estimation d'erreur très largement supérieure.