SY 09 - P14
TP 3: Théorie de la décision

Ricard Tatiana, Mehr Jean-Christophe

OBJECTIF DU TP

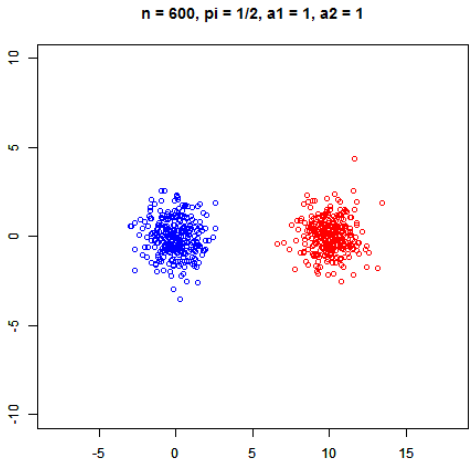
L'objectif de ce TP est de mettre en pratique deux théories de la décision qui sont le classifieur euclidien d'une part duquel nous allons tenter d'estimer son efficacité, et d'autre part la règle de Bayes en étudiant un problème de reconnaissance de cible.

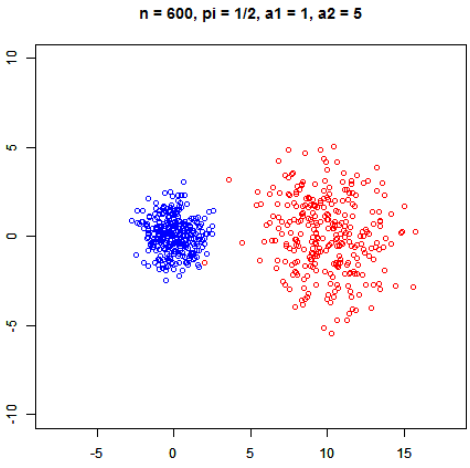
Exercice 1 : Classifieur euclidien

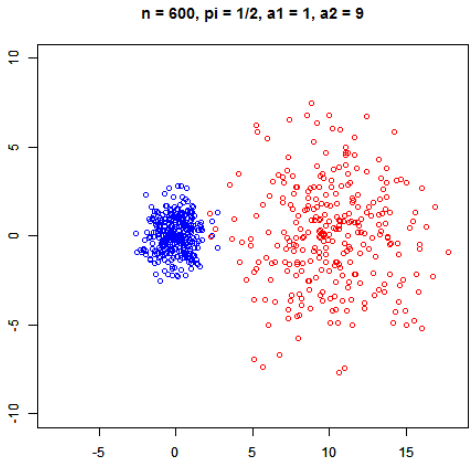
Le but de cet exercice est d'étudier les performances du classificateur euclidien sur des échantillons issus des deux classes ω_1 et ω_2 de $\mathbb{R}^2$ dont les distributions sont normales et de paramètres (μ_1, Σ_1) et (μ_2, Σ_2).

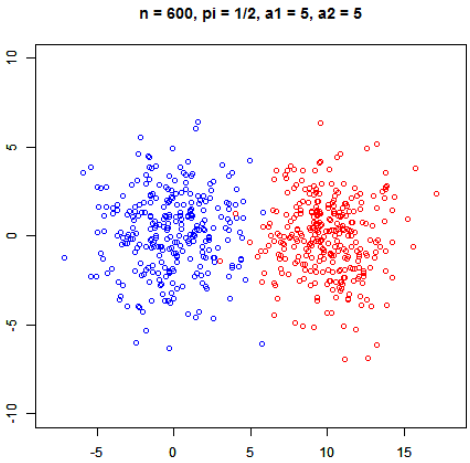
1) Simulation d'un échantillon

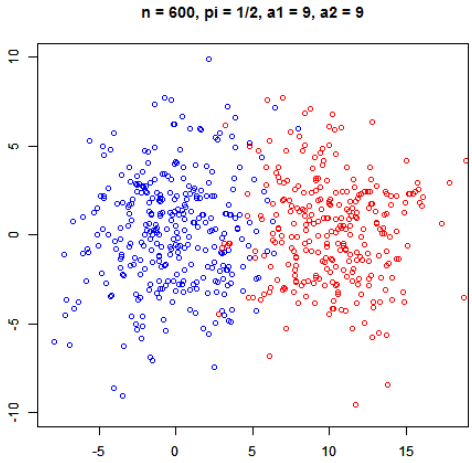
En utilisant la fonction **mvrnorm** de la bibliothèque **MASS**, nous avons écrit la fonction **simul** qui nous retourne un échantillon de taille n tiré suivant une proportion π d'exemples issus de la classe ω_1. Pour chaque exemple, nous avons, dans un premier temps, tiré au hasard la classe dont il est issu, avant de le générer en utilisant les paramètres adéquats.
Voilà les résultats obtenus graphiquement : Nous avons utilisé pour cette fonction les cinq situations suivantes :











Grâce à ces simulations, nous observons que plus la variance est élevée plus la dispersion des point est grande (autour de la moyenne), ils sont moins concentrés. Il apparaît qu'à partir d'une certaine valeur de variance (exemple de la cinquième figure) il existe un chevauchement entre les classes ω_1 et ω_2.
Il devient alors plus difficile de les distinguer et on peut d'ores et déjà supposer que la probabilité d'erreur sera plus importante.

2) Estimation de la probabilité d'erreur

Désormais nous allons chercher à estimer la probabilité d'erreur associée au classifieur Euclidien. Pour cela nous avons séparé en deux parties de même cardinal nos échantillons initiaux contenant les classes ω_1 et ω_2.
La première partie est l'échantillon d'apprentissage permettant d'apprendre les moyennes μ_1 et μ_2. La seconde partie est l'ensemble de tests permettant d'estimer le taux d'erreur de classifieur Euclidien sur ces données.

Utiliser un ensemble de test nous permet d'estimer la performance du classifieur. Nous avons pour celà tout d'abord écris une fonction règle Euclidienne qui correspond au classifieur Euclidien. Cette fonction compare les distances entre d'une observation x et l'estimateurμ_1 puis μ_2 et renvoie alors la classe dont x est le plus proche. Ensuite nous avons réalisé une fonction erreurEstimee qui calcul le nombre d'erreur de classification que notre classifieur a commis sur l'ensemble de test divisé par le cardinal de cet ensemble. Cette fonction renvoie donc le pourcentage d'erreur estimé.

Le calcul de la probabilité d'erreur en fonction des estimations de moyennes μ_1 et μ_2 nous donne les valeurs suivantes :

Variance	Probabilité d'erreur (%)
($\Sigma_1 = 1$, $\Sigma_2 = 1$)	0
($\Sigma_1 = 1$, $\Sigma_2 = 5$)	0.33
($\Sigma_1 = 1$,$\Sigma_2 = 9$)	2.67
($\Sigma_1 = 5$, $\Sigma_2 = 5$)	1.33
($\Sigma_1 = 9$, $\Sigma_2 = 9$)	5.67

Les résultats ci-dessus nous montrent bien que plus la variance est élevée, plus le taux d'erreur augmente et donc moins le classifieur Euclidien est performant. Ceci s'explique par le fait que la règle Euclidienne évalue la distance euclidenne entre chacun des points et les centres des deux classes. Il attribue ensuite à ce point la classe dont le centre est le plus proche.

De ce fait, lorsque la dispersion d'une classe est grande, les échantillons ont tendance à se chevaucher. Ainsi certains points se retrouve mal classifiés car il peut se retrouver plus proche du centre de l'autre classe si c'est un point éloigné du centre de sa propre classe (grande dispersion), le taux d'erreur est alors plus élevé.
En ce qui concerne la simulation 4, où $\Sigma_1 = 5$ et $\Sigma_2 = 5$, nous remarquons que le taux d'erreur est moins élevé que la simulation 3 où $\Sigma_1 = 1$ et $\Sigma_2 = 9$.
Cela laisse à penser qu'en terme d'efficacité il vaut mieux disposer de deux classes de variances modérées que d'une classe de variance faible et une autre très élevée. En effet bien que les deux sommes des variances de chaque simulation soient égales (10) une variance très élevée rendra un taux d'erreur plus important car se rapprochera plus facilement du centre de la classe à variance réduite (il n'y aura qu'un seul type d'erreur).

3) Probabilité d'erreur moyenne

Nous cherchons maintenant à observer la probabilité d'erreur moyenne, nous avons créé une fonction permettant de répéter les opérations précédentes 10 fois. À partir de ces dix simulations pour chacun des cas nous avons calculé la moyenne, la variance et un intervalle de confiance de niveau 5

Cette fonction nous permet de récupérer les données suivantes :

Variance	Moyenne du taux d'erreur	Variance du taux d'erreur (e^{-05})	Interface de confiance de niveau 5%
($\Sigma_1 = 1$, $\Sigma_2 = 1$)	0%	0	[0;0]
($\Sigma_1 = 1$, $\Sigma_2 = 5$)	0.57%	2.481481	[0.21%;0.923%]
($\Sigma_1 = 1$,$\Sigma_2 = 9$)	2.03%	1.481481	[1.276%;2.791%]
($\Sigma_1 = 5$, $\Sigma_2 = 5$)	1.3%	3.567901	[0.873%;1.727%]
($\Sigma_1 = 9$, $\Sigma_2 = 9$)	4.6%	2.04444	[3.577%;5.623%]

Nous remarquons que répéter l'opération permet d'abaisser le pourcentage d'erreur et de rendre la classifieur euclidien plus efficace.

Exercice 2 : Règle de Bayes

1. Montrer que les distributions f_1 et f_2 sont des distributions normales.

La fonction de densité d'une variable aléatoire suivant une loi normale est de la forme $f(x) = \frac{1}{\sigma(2\pi)^{\frac{1}{2}}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$

Nous avons donc pour f_1(x) :

$f_1(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(x_1+1)^2) \times \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(x_2+1)^2)$

$f_1(x) = \frac{1}{2\pi} \exp(-\frac{1}{2}(x_1+1)^2 - \frac{1}{2}(x_2+1)^2)$

$f_1(x) = \frac{1}{2\pi} \exp(-\frac{1}{2}[(x_1+1)^2 + (x_2+1)^2])$

Nous savons qu'une loi normale $f(x)$ bidimensionelle est de la forme :
$f(x) = \frac{1}{2\pi det(\Sigma)^2} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$

Par identification avec la formule vue précedemment nous trouvons donc $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ et $\mu_1 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$

De la meme facon, on trouve pour f_2(x) : $\Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ et $\mu_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

Les lois f_1 et f_2 suivent donc des lois normales bidimensionelles avec les paramètres vu ci dessus.

2. Générer un échantillon de n réalisations issues des deux classes ω_1 et ω_2.

Pour générer un échantillon de nn réalisations issues des deux classes ω_1 et ω_2 en proportions égales ($\pi_1 = \pi_2 = 0.5$), nous avons utilisé la fonction **simul** écrite précédemment.
Pour chacun des échantillons, nous avons ensuite déterminé les estimations des différents paramètres de f_1 et f_2.
Pour les valeurs de n suivantes : 10, 100, 1000, 10 000, 100 000, nous obtenons donc les résultats suivants :

Taille de l'échantillon	$\hat{\mu_1}$	$\hat{\Sigma_1}$	$\hat{\mu_2}$	$\hat{\Sigma_2}$
$n = 10$	$(0.22, -1.4)^T$	$\begin{pmatrix} 0.40 & 0.25 \\ 0.25 & 0.71 \end{pmatrix}$	$(0.8, 2.52)^T$	$\begin{pmatrix} 1.06 & 2.03 \\ 2.03 & 3.88 \end{pmatrix}$
$n = 100$	$(-0.73, -0.75)^T$	$\begin{pmatrix} 0.98 & 0.096 \\ 0.096 & 0.93 \end{pmatrix}$	$(1.1, 1.03)^T$	$\begin{pmatrix} 0.96 & -0.05 \\ -0.05 & 0.87 \end{pmatrix}$
$n = 1000$	$(-0.94, -1.04)^T$	$\begin{pmatrix} 1.02 & 0.03 \\ 0.03 & 0.89 \end{pmatrix}$	$(0.97, 1.07)^T$	$\begin{pmatrix} 0.94 & -0.01 \\ -0.01 & 0.90 \end{pmatrix}$
$n = 10000$	$(-0.99, -1.00)^T$	$\begin{pmatrix} 1.00 & -0.01 \\ -0.01 & 1.00 \end{pmatrix}$	$(1.00, 0.98)^T$	$\begin{pmatrix} 0.96 & 0.02 \\ 0.02 & 0.98 \end{pmatrix}$
$n = 100000$	$(-1.00, -1.00)^T$	$\begin{pmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{pmatrix}$	$(0.99, 1.00)^T$	$\begin{pmatrix} 1.01 & 0.01 \\ 0.01 & 1.01 \end{pmatrix}$

Nous avons pu observer que plus n était grand, plus on se rapprochait de la valeur théorique.

3. Montrer que les courbes d'iso-densité sont des cercles.

Pour trouver les courbes d'iso-densité d'une fonction, il nous faut résoudre une équation de type $f(x) = c$, avec $c \in \mathbb{R}$ constante. Ainsi nous posons :

$f_1(x) = c_1 \Leftrightarrow \frac{1}{2\pi} \exp(-\frac{1}{2}[(x_1+1)^2 + (x_2+1)^2]) = c_1$

$\Leftrightarrow \exp(-\frac{1}{2}[(x_1+1)^2 + (x_2+1)^2]) = 2\pi \times c_1$

$\Leftrightarrow -\frac{1}{2}[(x_1+1)^2 + (x_2+1)^2] = ln(2\pi \times c_1)$

$\Leftrightarrow (x_1+1)^2 + (x_2+1)^2 = -2ln(2\pi \times c_1)$

Les courbes d'iso-densité de f_1 sont alors des cercles de centre $O_1(-1,-1)$ et de rayon $r_1 = \sqrt{-2ln(2\pi \times c_1)}$. Ceci est valide seulement lorsque $-2ln(2\pi \times c_1) \geq 0$

C'est-à-dire pour $c_1 \leq \frac{1}{2\pi}$

De façon très similaire les courbes d'iso-densité sont pour f_2 :

$f_2(x) = c_2 \Leftrightarrow \frac{1}{2\pi} \exp(-\frac{1}{2}[(x_1-1)^2 + (x_2-1)^2]) = c_2$

$\Leftrightarrow \exp(-\frac{1}{2}[(x_1-1)^2 + (x_2-1)^2]) = 2\pi \times c_2$

$\Leftrightarrow -\frac{1}{2}[(x_1-1)^2 + (x_2-1)^2] = ln(2\pi \times c_2)$

$\Leftrightarrow (x_1-1)^2 + (x_2-1)^2 = -2ln(2\pi \times c_2)$

Les courbes d'iso-densité de f_2 sont alors des cercles de centre $O_2(1,1)$ et de rayon $r_2 = \sqrt{-2ln(2\pi \times c_2)}$. Ceci est valide seulement lorsque : $-2ln(2\pi \times c_2) \geq 0$

C'est-à-dire pour $c_2 \leq \frac{1}{2\pi}$

Ainsi les courbes d'iso-densité des distributions f_1 et f_2 sont des cercles.

4. Règle de Bayes.

(a) Donner l'expression de la règle de Bayes δ^* pour ce problème.

Ici, l'ensemble $\mathcal{A}$ des actions est le même que dans la question précédente, c'est-à-dire que les actions peuvent prendre 2 valeurs : a_1 et a_2. Ainsi nous avons pour x fixé :

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

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

La règle de Bayes δ^* minimisant $r(\delta|x)$, x fixé est alors

$$\delta^*(x) = \begin{cases} a_1 & \text{si } r_1(x) < r_2(x) \\ a_2 & \text{sinon} \end{cases}$$

Comme nous n'avons que deux classes, cette règle peut s'exprimer en fonction du rapport de vraisemblance $\frac{f_1(x)}{f_2(x)}$

$\delta^*(x) = a_1 \Leftrightarrow r_1(x) < r_2(x)$

$\Leftrightarrow \frac{f_1(x)}{f_2(x)} > \frac{c_{12}-c_{22}}{c_{21}-c11}\frac{\pi_2}{\pi_1}$

D'après l'énoncé du problème, les coûts c_{11} et c_{22} sont égaux à 0 : $\frac{f_1(x)}{f_2(x)} > \frac{c_{12}}{c_{21}}\frac{\pi_2}{\pi_1}$

Calculons le rapport de vraisemblance $\frac{f_1(x)}{f_2(x)}$:

$\frac{f_1(x)}{f_2(x)} = \frac{\frac{1}{2\pi}\exp(-\frac{1}{2}[(x_1+1)^2+(x_2+1)^2])}{\frac{1}{2\pi}\exp(-\frac{1}{2}[(x_1-1)^2+(x_2-1)^2])}$

$= \exp(-\frac{1}{2}[(x_1+1)^2 + (x_2+1)^2] + \frac{1}{2}[(x_1-1)^2 + (x_2-1)^2]))$

$= \exp(-\frac{1}{2}[(x_1+1)^2 + (x_2+1)^2 - (x_1-1)^2 - (x_2-1)^2])$

À l'aide d'identités remarquables nous trouvons :

$\frac{f_1(x)}{f_2(x)} = \exp(-\frac{1}{2} \times (4(x_1+x_2)) = \exp(-2(x_1+x_2))$

D'où, $\exp(-2(x_1+x_2)) > \frac{c_{12}-c_{22}}{c_{21}-c11}\frac{\pi_2}{\pi_1} \Leftrightarrow -2(x_1+x_2) > ln(\frac{c_{12}-c_{22}}{c_{21}-c11}\frac{\pi_2}{\pi_1})$

$\Leftrightarrow x_1 + x_2 < -\frac{1}{2} \times ln(\frac{c_{12}-c_{22}}{c_{21}-c11}\frac{\pi_2}{\pi_1})$

et donc la règle δ^* de Bayes devient

$$\delta^*(x) = \begin{cases} a_1 & \text{si } x_1 + x_2 < -\frac{1}{2} \times ln(\frac{c_{12}-c_{22}}{c_{21}-c11}\frac{\pi_2}{\pi_1}) \\ a_2 & \text{sinon} \end{cases}$$

Cette règle de décision dépent de x_1 et x_2.

(b et c) Tracer les frontières de décision correspondantes dans le plan (X_1, X_2) et donner une estimation des risques α et β.

Nous sommes à la frontière de décision lorsque, $x_1 + x_2 = -\frac{1}{2} \times ln(\frac{c_{12}-c_{22}}{c_{21}-c11}\frac{\pi_2}{\pi_1})$
Le tableau ci-dessous présente les différentes frontières de décision selon les coûts associés aux actions et les probablités à priori. Pour estimer les erreurs de secondes et de premières espèces nous avons réutilisé la fonction `simul` de l'exercice 1.

Coûts associés	Probabilités à priori	Frontière de décision	α	β
$c_{12} = c_{21} = 1$	$\pi_1 = \pi_2$	$x_1 + x_2 = 0$	$\alpha = 0.102$	$\beta = 0.065$
$c_{12} = 10$ et $c_{21} = 1$	$\pi_1 = \pi_2$	$x_1 + x_2 = -\frac{1}{2}ln(10)$	$\alpha = 0.26$	$\beta = 0.015$
$c_{12} = c_{21} = 1$	$\pi_2 = 10\pi_1$	$x_1 + x_2 = -\frac{1}{2}ln(10)$	$\alpha = 0.27$	$\beta = 0.01$

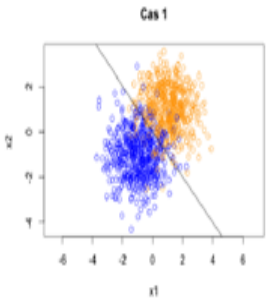


FIGURE 2.1 – Cas 1

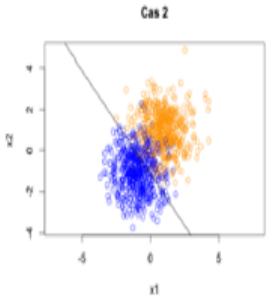


FIGURE 2.2 – Cas 2

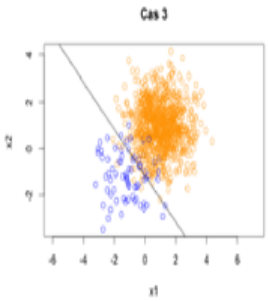


FIGURE 2.3 – Cas 3

Nous remarquons que pour la cas 1, la frontière se situe quasiment au milieu. On peut l'expliquer par le afit que le coût d'une erreur d'affectation à une classe et les probabilités sont identiques.
Le risque de classer les points dans ω_1 et ω_2 est donc identique.
Pour le cas 2, les coûts sont différents : en effet, c12 est 10 fois plus important que c21 et les probabibilés semblent toujours égales. Ainsi, il est dix fois plus coûteux de mal classer un individu appartenant à la classe ω_2. Sur le graphique, la règle de Baynes place la frontière plus près d'ω_1. La règle de Bayes classifie donc mal une grande quantité d'éléments de cette classe afin d'être sûr de ne pas mal classifier d'éléments coûteux (corroboré par les valeurs d'α et β avec α très grand).

Pour finir, dans le cas 3 les coûts sont une nouvelle fois les memes, cependant les proportions sont différentes. La proportion des individus dans ω_2 est quasiment dix fois plus grande que celle d'ω_1. Pour équilibrer les risques, la règle de Bayes place donc la frontière plus proche d'ω_1. Ceci explique donc pourquoi l'erreur de second type β soit faible et que le risque *alpha* soit beaucoup plus grand !

Conclusion

Nous avons finalement pu dans ce TP étudier un classifieur euclidien et les effets des valeurs des variances sur le pourcentage d'erreur. En outre nous avons pu observer que comme dans beaucoup de cas, répéter l'expérience permettait de réduire ce pourcentage d'erreur.
Dans le deuxième exercice nous avons pu étudier la règle de Bayes et observer l'influence de la proportion des classes et des coûts sur les frontières de décisions, la rêgles de Bayes tentant de minimiser le pourcentage d'erreur et le coût global.
Les règles de décisions permettent donc de classifier des éléments en fonction d'observations et de paramètres prédéfini, il est alors possible de jouer sur ces paramètres et sur la taille de l'échantillon afin d'influencer ces décisions.