Université de Technologie Compiègne

Génie Informatique et Mécanique

TP 3 : Théorie de la décision

Objectif du TP :

L'objectif de ce TP est d'étudier la théorie de la décision. Dans un premier temps, nous allons étudier les performances du classifieur euclidien sur des échantillons issus de deux classes ω_1 et ω_2 de $\mathbb{R}^2$ et dont les distributions sont normales. Dans un second temps nous allons travailler sur un problème de détection de cible, à l'aide de la règle de Bayes.

UV : **SY09**

Semestre : **P14**

1. Classifieur Euclidien

L'objectif de cet exercice est l'évaluation du classifieur Euclidien, pour pouvoir tester ce classifieur nous nous sommes tout d'abord intéressé à la simulation d'un échantillon de taille 600 représentant deux classes : ω_1 et ω_2 dans $\mathbb{R}^2$.

1.1 Simulation d'un échantillon

Pour cela, nous avons créé une fonction `simul(n, pi, mu1, mu2, Sigma1, Sigma2)`. Cette fonction permet, à l'aide de la fonction R `mvnorm` de la bibliothèque MASS de générer aléatoirement deux échantillons. Le premier tiré selon une proportion de π et le second selon $1-\pi$. Ces deux échantillons suivent chacun une loi Normale de paramètres $\mathcal{N}(\mu_1, \Sigma_1)$ et $\mathcal{N}(\mu_2, \Sigma_2)$. Enfin notre fonction mélange les deux échantillons de façon aléatoire à l'aide de la fonction `sample`. Durant ce TP nous avons utilisé pour paramètres fixes $\mu_1 = (0,0)^T$ et $\mu_2 = (10,0)^T$. Ensuite nous avons fait varier les variances Σ_1 et Σ_2. Voici les résultats obtenus :

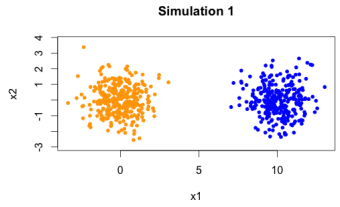

FIGURE 1.1 – Simulation 1

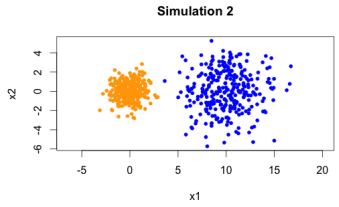

FIGURE 1.2 – Simulation 2

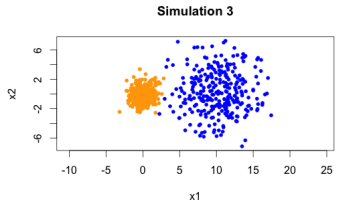

FIGURE 1.3 – Simulation 3

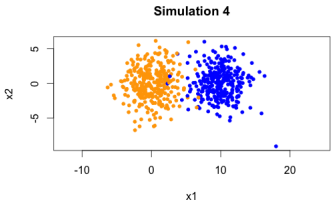

FIGURE 1.4 – Simulation 4

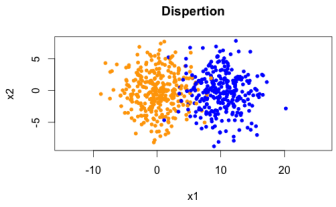

FIGURE 1.5 – Simulation 5

Nous remarquons grâce aux cinq graphiques ci-dessus que plus la variance est élevée, plus la dispersion autour de la moyenne est grande. Ainsi lorsque les 2 variances sont

élevées (simulation 5) nous notons un chevauchement entre les échantillons. Il devient donc plus dur de distinguer les classes ω_1 et ω_2.

1.2 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'echantillon d'apprentissage permettant d'apprendre les moyennes μ_1 et μ_2. La seconde partie est l'ensemble de test permettant d'estimer le taux d'erreur de classifieur Euclidien sur ces données. Si nous n'avions pas utilisé d'ensemble de test, nous n'aurions pu estimer la performance du classifieur.

Nous avons tout d'abord écris une fonction `regleEuclidienne` qui correspond au classifieur Euclidien. Cette fonction compare la distance entre une observation x et l'estimateur μ_1 puis entre x et l'estimateur μ_2. Elle renvoie alors la classe dont x est la 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. Voici les réultats concernant nos 5 simulations différentes :

Variances	Taux d'erreur en %
$\Sigma_1 = 1$ et $\Sigma_2 = 1$	0
$\Sigma_1 = 1$ et $\Sigma_2 = 5$	0,33
$\Sigma_1 = 1$ et $\Sigma_2 = 9$	2,67
$\Sigma_1 = 5$ et $\Sigma_2 = 5$	1,33
$\Sigma_1 = 9$ et $\Sigma_2 = 9$	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 retrouvent mal classifiés : le taux d'erreur est alors plus élevé. En ce qui concerne la simulation 4, où $\Sigma_1 = 5$ et $\Sigma_2 = 5$ nous notons bien que le taux d'erreur est moins élevé que la simulation 3, , où $\Sigma_1 = 1$ et $\Sigma_2 = 9$. Ceci s'explique car la variance de la distribution de classe ω_2 est réduite et compense donc la variance de la distribution de classe ω_1. Ainsi le nombre de points mal classifié a diminué.

1.3 Probabilité d'erreur moyenne

Dans cette troisième parties de l'exercice, nous avons créé un 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% sur l'espérance de la probabilité d'erreur. L'intervalle de confiance a été calculé grâce au test de Student.

Variances	Moyenne du taux d'erreur	Variance du taux d'erreur	Intervalle de confiance de niveau 5%
$\Sigma_1 = 1$ et $\Sigma_2 = 1$	0%	0	[0; 0]
$\Sigma_1 = 1$ et $\Sigma_2 = 5$	0,57%	$2.481481e^{-05}$	[0.21%; 0.923%]
$\Sigma_1 = 1$ et $\Sigma_2 = 9$	2,03%	$1.1222e^{-04}$	[1.276%; 2.791%]
$\Sigma_1 = 5$ et $\Sigma_2 = 5$	1,3%	$3.567901e^{-05}$	[0.873%; 1.727%]
$\Sigma_1 = 9$ et $\Sigma_2 = 9$	4,6%	$2.04444e^{-04}$	[3.577%; 5.623%]

Nous remarquons que les intervalles de confiance obtenus sont restreints. Par curiosité, nous avons modifié notre fonction pour répéter les questions 1 et 2, 100 fois. Nous avons alors remarqué que les intervalles de confiance deviennent de plus en plus restreints. La précision de l'estimation est donc plus importante selon le nombre de simulation généré.

2. Règle de Bayes

Dans cet exercice nous considérons un problème de détection de cible dans lequel la classe ω_1 correspond aux missiles et la classe ω_2 correspond aux avions. Chacune des cibles est décrite par deux variables X_1 et X_2. Ces variables sont considérées indépendantes conditionnellement et suivent les lois normales suivantes :

$$f_{11} \sim \mathcal{N}(-1, 1),\ f_{21}\ \mathcal{N}(1, 1)$$
$$f_{12} \sim \mathcal{N}(-1, 1),\ f_{22}\ \mathcal{N}(1, 1)$$

Les densités conditionnelles du vecteur $X = (X_1; X_2)^T$ sont donc $f_1(x) = f_{11}(x1)f_{12}(x2)$ dans la classe ω_1 et $f_2(x) = f_{21}(x1)f_{22}(x2)$ dans la clasee ω_2.

2.1 Normalités des distibutions f_1 et f_2

Une loi normale $X \sim \mathcal{N}(\mu, \sigma)$ a pour densité : $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2})$
Ainsi nous avons 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 $f(x)$ bidimensionnelle a pour forme :
$f(x) = \frac{1}{2\pi \times det(\Sigma)^2} \exp(-\frac{1}{2}(x-\mu)^T \times \Sigma^{-1} \times (x-\mu))$
Par identification nous trouvons tout d'abord $\Sigma_1 = Id$, puis $\mu_1 = (-1,-1)^T$
De la même façon pour $f_2(x)$: $f_2(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(x_1-1)^2) \times \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(x_2-1)^2)$
$f_2(x) = \frac{1}{2\pi} \exp(-\frac{1}{2}(x_1-1)^2 - \frac{1}{2}(x_2-1)^2)$
$f_2(x) = \frac{1}{2\pi} \exp(-\frac{1}{2}[(x_1-1)^2 + (x_2-1)^2])$
Par identification avec la formule vue précedemment nous trouvons donc $\Sigma_2 = Id$ et $\mu_2 = (1,1)^T$
Les lois f_1 et f_2 suivent donc des lois normales bidimensionnelles respectivement de paramètres $\mu_1 = (-1,-1)^T$, $\Sigma_1 = Id$ et $\mu_2 = (1,1)^T$, $\Sigma_2 = Id$

2.2 Simulation des échantillons et estimation des paramètres de f_1 et f_2

À l'aide de la fonction `simul` utilisé lors de l'exercice précédent nous avons pu générer des échantillons de n réalisations issues des deux classes ω_1 et ω_2 en proportions égales, $n \in \{10, 100, 1000, 10000, 100000\}$
Voici les résultats obtenus :

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

En regardant attentivement ces résultats, nous remarquons que, de façon tout à fait logique, plus la taille de l'échantillon généré lors de la simulation est élevé plus les valeurs

empiriques sont proches des valeurs théoriques trouvées plus haut. Ainsi si nous pouvions générés un échantillon de taille infini nous retrouverions les résultats théoriques. Nous pouvons donc conclure que la taille d'un échantillon joue un rôle important dans la fiabilité d'une étude.

2.3 Courbes d'iso-densité des distibutions f_1 et f_2

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.

2.4 Règle de Bayes et frontière de décisions

2.4.1 Expression de la règle de Bayes δ^*

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.

2.4.2 Frontières de décision dans le plan (X_1, X_2) et estimation des risques de première et seconde espèce

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 $c_{21} = 1$	$\pi_1 = \pi_2$	$x_1 + x_2 = -\frac{1}{2}ln(10)$	$\alpha = 0.26$	$\beta = 0.015$
$c_{12} = c_{21} = 1$	$\pi_2 = 10\pi_1$	$x_1 + x_2 = -\frac{1}{2}ln(10)$	$\alpha = 0.27$	$\beta = 0.01$

Nous remarquons que pour le cas 1 la frontière se situe au milieu. Ceci s'explique par le fait que le coût d'une mauvaise affectation à une classe et les probabilités à priori sont identiques. Nous avons donc le même risque de classer les points dans w1 que dans w2. Le règle de Bayes minimisant les risques de chaque espèce, le seuil sera alors placé entre les

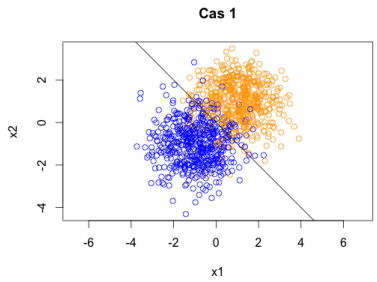

FIGURE 2.1 – Cas 1

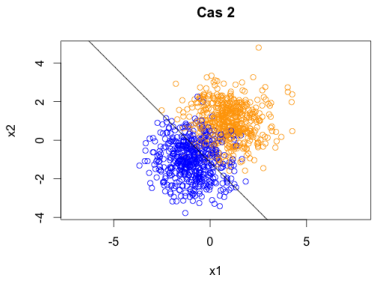

FIGURE 2.2 – Cas 2

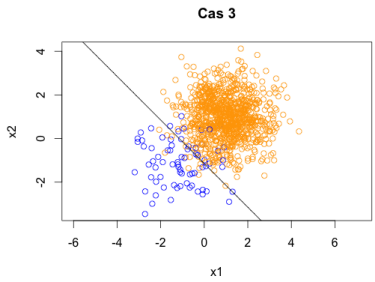

FIGURE 2.3 – Cas 3

deux classes. Nous remarquons de plus que les erreurs de première et de seconde espèce sont bien du même ordre de grandeur.

Pour le cas 2, les coûts diffèrent, c_{12} est 10 fois plus important que c_{21} et les probabilités a priori π_1 et π_2 sont toujours égales. Ainsi il est 10 fois plus coûteux de mal classer un individu appartenant normalement à la classe ω_2. Sur le graphique, la règle de Bayes place la frontière de décision plus proche du centre de la classe ω_1. De cette façon la règle aura peu de chance de mal classifier un élément appartenant à ω_2. En regardant en plus les erreurs α et β, nous voyons qu'α est très élevé. Ceci se comprend par le fait que être sûr de ne pas mal classifier les éléments couteux, la règle de Bayes classifie mal une grande quantité d'élément de la classe ω_1.

Enfin pour le cas 3, les coûts sont à nouveaux les mêmes, néanmoins les proportions diffèrent : $\pi_2 = 10\pi_1 = \frac{10}{11}$. La proportion des individus de ω_2 est 10 fois supérieure à celle des individus de la classe ω_1. Si la frontière était placée au milieu, il y aurait alors beaucoup plus de chances de classer un point de ω_2 dans la classe ω_1. Pour équilibrer les risques la règle de Bayes place donc la frontière de décision plus proche du centre de la classe ω_1. Ceci explique que l'erreur de seconde espèce β soit plus faible et que le risque α soit beaucoup plus important.

Conclusion

Lors de ce TP nous avons pu mettre en application la théorie de la décision à travers différents exercices de classification. Dans un premier temps, nous avons utilisé et mesuré l'efficacité du classifieur euclidien. Nous avons pour ce faire appris à simuler des échantillons. Dans un second temps, nous avons utilisé la règle de Bayes pour trouver une règle de décision qui classifie au mieux nos individus c'est-à-dire minimisant la probabilité d'er-

reur et le coût total. Nous avons donc constaté que les frontières de décision, dictées par la règle de Bayes, varient en fonction des coûts et de la proportion des classes. Durant cet exercice, nous avons aussi pu montrer que la taille de l'échantillon influence directement sur la fiabilité d'une estimation.