SY09

TP03

THEORIE DE LA DECISION

P13

Cynthia NZENGANG

Elsa JULIARD

Objectif

Dans cette partie, nous allons étudier des techniques de classification supervisée. A cet effet, nous disposerons d'un ensemble d'apprentissage qui nous permettra de déterminer une fonction de décision. L'objectif du TD sera en outre, de mettre en pratique les différentes méthodes de classifications supervisées à travers un problème de discrimination en deux classes. Dans un premier temps, on étudiera le classifieur euclidien et ensuite la règle de Bayes.

1. Classifieur euclidien

Dans cet exercice, nous allons simuler un échantillon de données issu de deux classes {1,2} suivant des distributions normales bidimensionnelles et nous évaluerons les performances du classifieur mis en place.

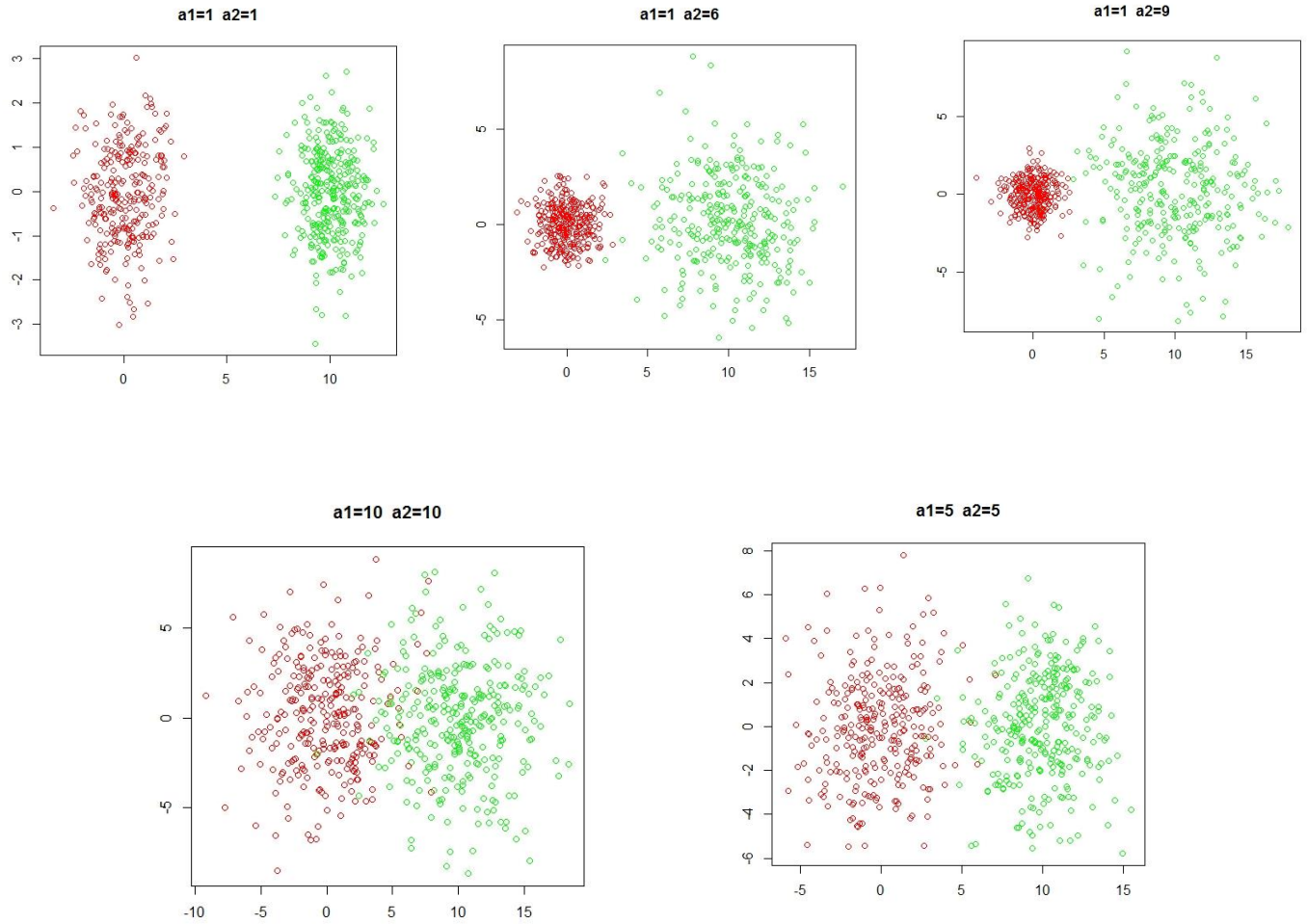
a) Simulation d'un échantillon

Pour cette partie, nous allons générer un jeu de données de taille n suivant les paramètres suivants :

- Distribution de la classe w_1 : Normale de paramètres $(\mu_1, \sum_1)$
- Distribution de la classe w_2 : Normale de paramètres $(\mu_2, \sum_2)$

Afin de pouvoir construire notre classifieur euclidien, nous allons dans un premier temps utiliser la fonction **rbinom** qui va nous permettre d'effectuer des tirages aléatoires afin d'obtenir la taille des échantillons pour chaque classe. Une fois la taille de chaque classe obtenue, nous allons utiliser la fonction **mvrnorm** qui va prendre en paramètre la variance et l'espérance de chaque distribution, et va générer des observations pour les classe 1 et 2. Enfin, à chacun de ces échantillons, on associera la classe appropriée {1,2] en faisant attention de les mélanger grâce à la fonction **sample**.

On obtient en sortie une matrice contenant les données générées avec la classe associée à chaque individu et on crée une sortie graphique adaptée.



On remarque sur ces graphes que les nuages de points représentant les classes forment des sortes de cercle centrés en μ et dont le rayon augmente en fonction de a. On peut remarquer que plus a, qui représente le coefficient de la variance augmente, plus le nuage de points s'épaissit et plus les individus des deux classes se confondent ce qui augmente la probabilité d'erreur moyenne. Cette probabilité d'erreur est également plus importante pour les individus situées entre μ_1 etμ_2.

b) Estimation de la probabilité d'erreur

Dans cette partie, nous allons pouvoir vérifier les hypothèses faites après observation des sorties graphiques en calculant la probabilité d'erreur. Il sera question d'appliquer aux données générées le classifieur euclidien qui va consister à associer à chaque individu la classe dont le centre est le plus proche (au sens de la distance euclidienne). Dans un premier temps, nous allons séparer nos données en deux ensembles égaux : l'ensemble d'apprentissage et l'ensemble de test. Nous allons calculer à partir de notre ensemble d'apprentissage les estimateurs (moyenne empirique) qui représenteront les centres des classes. Pour s'assurer de la pertinence de notre classifieur, nous allons réaliser les tests sur l'ensemble de test afin que les données ne soient pas biaisées. La probabilité d'erreur se calcule en faisant le rapport des résultats erronés par le nombre total d'observations.

On obtient les résultats suivants pour le premier passage :

	μ_1	μ_2	Probabilité d'erreur
a1= 1 & a2 =1	(-0.072,-0.0383)	(10.152, -0.0892)	0
a1= 1 & a2 =6	(0.0378, -0.0140)	(10.084, -0.328)	0.01
a1= 1 & a2 =9	(0.041, -0.1084)	(10.094, 0.384)	0.03
a1= 5 & a2 =5	(-0.055,-0.4162)	(10.251, 0.034)	0.01
a1= 10 & a2 =10	(0.722, 0.0903)	(10.427, -0.201)	0.06

c) Probabilité d'erreur moyenne

En répétant une dizaine de fois l'expérience, on peut calculer la moyenne empirique des résultats obtenus et sa variance. La probabilité d'erreur moyenne dans notre cas va nous permettre d'étudier la performance du classifieur.

Après avoir répété 10 fois, nous obtenons le tableau ci-dessous :

	a1= 1 & a2 =1	**a1= 1 & a2 =6**	**a1= 1 & a2 =9**	**a1= 5 & a2 =5**	**a1= 10 & a2 =10**
Moyenne	**0**	**0.0117**	**0.0257**	**0.0087**	**0.0427**
Variance	**0**	**2.2831e-05**	**7.1720e-05**	**2.0247e-05**	**1.7975e-04**

Nous pouvons remarquer d'après ce tableau que dans le cas où les matrices de variance sont égales, cas où a1=1 & a2=1, a1=5 & a2=5, on a les probabilités d'erreur qui sont les plus basses. Cela s'explique par le fait que les données vont former des disques dont les rayons augmentent en fonction de a1 & a2. On constate d'après ce tableau que plus les centres sont proches et les dispersions importantes, plus le classifieur euclidien a des chances de se tromper.

d) Intervalle de confiance

Si on considère n le nombre d'erreur, π la probabilité d'erreur et N le nombre d'expériences réalisées alors

$n \sim B(N,\pi)$ donc $E[n] = N\pi$ or $m = \frac{n}{N}$ et d'où $E[m] = \pi$ $Var[m] = \frac{\pi(1-\pi)}{N}$ on a donc d'après le théorème de la limite centrale, l'expression suivante $m \sim B(\pi, \frac{\pi(1-\pi)}{N})$ => $\frac{(m-\pi)\sqrt{N}}{\sqrt{\pi(1-\pi)}} \sim N(0,1)$

On a donc $P(\pi - u_{1-\frac{\alpha}{2}}\sqrt{\frac{\pi(1-\pi)}{N}} \leq m \leq \pi + u_{1-\frac{\alpha}{2}}\sqrt{\frac{\pi(1-\pi)}{N}}) = 1-\alpha$, on a π= 0.0427, $u_{1-\frac{\alpha}{2}}$ = 1.96 et N =10, on trouve ainsi le résultat suivant $IC_{0.05}$ = [-0.0826, 0.168] pour le cas **a1=10 & a2= 10.**

2. Règle de Bayes

Dans cette seconde partie, nous nous intéressons à un problème de classification de cibles en deux classes ω_1 et ω_2 (missiles et avions) à partir de leur description par deux variables issues de deux capteurs différents. Nous sommes donc confrontés à un problème de discrimination de deux classes.

1. Montrer que les distributions f_1 et f_2 sont des distributions normales dont on précisera les espérances et les matrices de variance-covariance.

Pour pouvoir résoudre cette partie, nous allons utiliser ce théorème : Soient U_1, U_2, deux variables aléatoires, normales, centrées-réduites et indépendantes, et U= (U_1, U_2)'.
X = μ + B**U** suit une loi normale à 2 dimensions avec $\mu \in R^2$ et B $\in M_{2,2}$.

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

Par conséquent, les densités conditionnelles du vecteur X= $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ sont:
$f\omega_1(x) = f_{11}(x_1) \cdot f_{12}(x_2)$
$f\omega_2(x) = f_{21}(x_1) \cdot f_{22}(x_2)$

Ainsi, $X_1+1 \sim N(0, 1)$ et $X_2 \sim N(0, 1)$; donc X= $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \mu + \mathbf{B}\begin{pmatrix} X_1 + 1 \\ X_2 \end{pmatrix}$ avec $\mu = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$ et **B**= $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

$f\omega_1$ étant une distribution normale, on a: $f\omega_1(x) = \frac{1}{2\pi} \exp\{\frac{-1}{2}[(x_1 + 1)^2 + x_2^2]\}$

On identifie alors:

$\boldsymbol{E}[\boldsymbol{X} | \boldsymbol{\omega} = \boldsymbol{\omega_1}] = \begin{pmatrix} \mathbf{-1} \\ \mathbf{0} \end{pmatrix}$ et $\boldsymbol{Var}[\boldsymbol{X} | \boldsymbol{\omega} = \boldsymbol{\omega_1}] = \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{pmatrix}$

De même, $f\omega_2$ étant une distribution normale, on a: $f\omega_1(x) = \frac{1}{2\pi} \exp\{\frac{-1}{2}[(x_1 - 1)^2 + x_2^2]\}$

Et $\boldsymbol{E}[\boldsymbol{X} | \boldsymbol{\omega} = \boldsymbol{\omega_2}] = \begin{pmatrix} \mathbf{1} \\ \mathbf{0} \end{pmatrix}$ et $\boldsymbol{Var}[\boldsymbol{X} | \boldsymbol{\omega} = \boldsymbol{\omega_2}] = \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{pmatrix}$

2. En utilisant la fonction simul, générer un échantillon de n réalisations issues de deux classes ω_1 et ω_2 en proportions égales; pour chacun des échantillons, déterminer les estimations des différents paramètres de f_1 et f_2. On effectuera ce travail pour n= 10, 100, 1 000, 10 000, 100 000.

Ci-dessous le récapitulatif des différents paramètres pour chaque échantillon.

Echantillon	$\widehat{\mu_1}$	$\widehat{\mu_2}$	$\widehat{\sum}_1$	$\widehat{\sum}_2$
10	$\begin{pmatrix} -0.87 \\ 0.67 \end{pmatrix}$	$\begin{pmatrix} 0.93 \\ 0.88 \end{pmatrix}$	$\begin{pmatrix} 1.39 & 0.42 \\ 0.42 & 1.43 \end{pmatrix}$	$\begin{pmatrix} 0.08 & -0.09 \\ -0.09 & 0.89 \end{pmatrix}$
100	$\begin{pmatrix} -0.94 \\ 0.05 \end{pmatrix}$	$\begin{pmatrix} 0.95 \\ -0.08 \end{pmatrix}$	$\begin{pmatrix} 1.08 & 0.15 \\ 0.15 & 0.94 \end{pmatrix}$	$\begin{pmatrix} 0.89 & 0.2 \\ 0.2 & 1.48 \end{pmatrix}$
1 000	$\begin{pmatrix} -0.98 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0.97 \\ 0.04 \end{pmatrix}$	$\begin{pmatrix} 1.03 & -0.05 \\ -0.05 & 0.98 \end{pmatrix}$	$\begin{pmatrix} 1.06 & 0.11 \\ 0.11 & 0.97 \end{pmatrix}$
10 000	$\begin{pmatrix} -0.99 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0.99 \\ -0.02 \end{pmatrix}$	$\begin{pmatrix} 1.01 & 0.02 \\ 0.02 & 0.98 \end{pmatrix}$	$\begin{pmatrix} 1.01 & 0.01 \\ 0.01 & 0.99 \end{pmatrix}$
100 000	$\begin{pmatrix} -1 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 1 \\ -0.01 \end{pmatrix}$	$\begin{pmatrix} 1 & 0.01 \\ 0.01 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 0.99 \end{pmatrix}$

On remarque facilement, avec l'estimation de ces paramètres, que plus la taille de l'échantillon est important, plus les estimateurs se rapprochent de la valeur théorique calculée à la question précédente.
Avec un échantillon de taille 100 000 ou plus, nous pouvons garantir l'exactitude du résultat proposé par le classifieur.

3. Montrer que les courbes d'iso-densité sont des cercles dont on précisera les rayons.

Calculer les courbes d'iso-densité revient à résoudre l'équation:

$f\omega_1(x) =$ C avec C une constante soit: $\frac{1}{2\pi}\exp\{\frac{-1}{2}[(x_1+1)^2 + x_2^2]\} = C_1$

$$\Leftrightarrow \quad \frac{-1}{2}[(x_1+1)^2 + x_2^2]\} = C_1$$

$$\Leftrightarrow \quad -\frac{1}{2}(x_1+1)^2 + x_2^2 = \ln(2\pi\, C_1)$$

$$\Leftrightarrow \quad (x_1+1)^2 + x_2^2 = -2\cdot\ln(2\pi C_1)$$

Même raisonnement avec $f\omega_2(x) =$ C $\Leftrightarrow (x_1-1)^2 + x_2^2 = -2\cdot\ln(2\pi\, C_2)$

On retrouve ici les équations de cercles centrés en $\mu_1 = \begin{pmatrix}-1\\0\end{pmatrix}$ et $\mu_2 = \begin{pmatrix}1\\0\end{pmatrix}$ et de rayon: $\sqrt{-2\ln(2\pi C)}$.

La fonction racine carré étant positive, on a, par définition:

$-2\ln(2\pi C) \geq 0$, ce qui donne donc: $0 < C \leq \frac{1}{2\pi}$.

En effet, lorsque le maximum est atteint, la courbe d'iso-densité n'est plus qu'un point. Ceci s'explique également par le fait que les fonctions de densité admettent un maximum qui vaut $\frac{1}{2\pi}$.

4. Soient π_1 et π_1 les probabilités a priori des deux classes et c_{lk} le coût associé à l'action a_l lorsque la vraie classe est ω_k. On suppose $c_{11} = c_{22} = 0$.

La règle de Bayes s'écrit:

$$\delta^*(x) = \begin{cases} a_1 \; si \;\; \frac{f\omega_1(x)}{f\omega_2(x)} > \frac{c_{12}\pi_2}{c_{21}\pi_1} \\ a_2 \; sinon \end{cases}$$

Nous utilisons la fonction **simul** pour générer des échantillons que l'on représentera dans le plan (X_1, X_2).

- **$c_{12} = c_{21} = 1$ et $\pi_1 = \pi_2$**

La règle de Bayes s'écrit donc $\delta^*(x) = a_1 \Leftrightarrow f\omega_1(x) > f\omega_2(x)$. On trouve la frontière de décision en résolvant l'équation $f\omega_1(x) = f\omega_2(x)$.

On trouve $x_1 = 0$.

- **$c_{12} = 10c_{21}$ et $\pi_1 = \pi_2$**

La règle de Bayes s'écrit alors $\delta^*(x) = a_1 \Leftrightarrow f\omega_1(x) > 10f\omega_2(x)$. On trouve la frontière de décision en résolvant l'équation $f\omega_1(x) = 10f\omega_2(x)$.

On trouve $x_1 = -\frac{\ln 10}{2}$ = -1.15

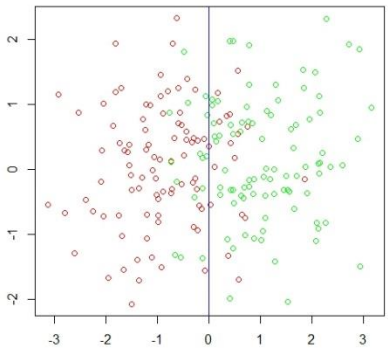
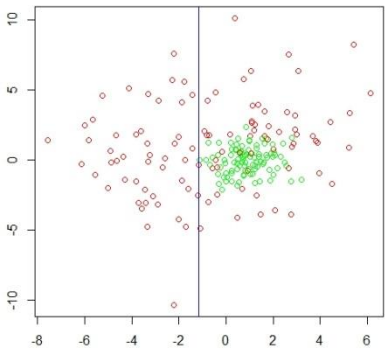
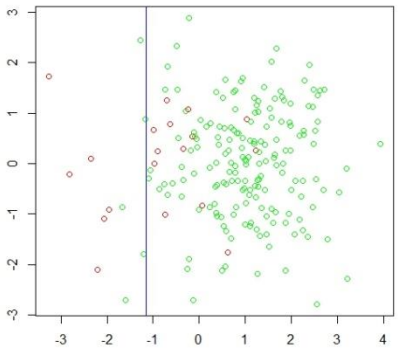
- **$c_{12} = c_{21} = 1$ et $10\pi_1 = \pi_2$**

La règle de Bayes s'écrit alors $\delta^*(x) = a_1 \Leftrightarrow f\omega_1(x) > 10f\omega_2(x)$. On trouve la frontière de décision en résolvant l'équation $f\omega_1(x) = 10f\omega_2(x)$.

De même que précédemment, on trouve $x_1 = -\frac{\ln 10}{2}$ = -1.15 On obtient ainsi le même résultat qu'à la question précédente.

Graphiquement, on obtient les résultats ci-dessous avec les erreurs a =α et b= β mentionnés.

a=0.156 b=0.165 *a=0.37 b=0.005* *a=0.684 b=0.022*

Dans le premier cas, les erreurs de première et de deuxième espèce ont le même coût. Par conséquent la frontière de décision se situe au milieu des deux centres.

Dans le deuxième cas, le coût d'erreur de seconde espèce, c'est-à-dire le coût d'affectation à la classe ω_1 alors que la vraie classe est ω_2, est 10 fois plus grand que le coût d'erreur de première espèce. Par conséquent la règle tend à minimiser l'erreur de seconde espèce. La frontière de décision est alors décalée vers la gauche car affecter une observation à la classe ω_2, même si ce n'est pas la bonne classe, "coûte moins cher".

Dans le troisième cas, la probabilité a priori de la classe ω_2 est dix fois plus grande que celle de la classe ω_1. Ainsi, sans même se préoccuper des coûts ou des observations, on sait que les observations ont "10 fois plus de chances" d'appartenir à la classe ω_2 que la classe ω_1. C'est pourquoi il apparaît cohérent de décaler la frontière vers la gauche une nouvelle fois.

Conclusion

Au cours de ce TP, nous avons découvert deux méthodes de classification supervisée.Nous avons étudié le classifieur euclidien. Il s'agit d'un classifieur très simple mais aussi assez robuste. On note toute de même que si les distributions ne sont pas sphériques, le classifieur montre ses limites.

Avec la règle de Bayes, il s'agit plutôt de classer les individus tout en contrôlant le risque. Ainsi, on admet que des erreurs soit plus ou moins importantes que d'autres. Elle permet non pas de minimiser le nombre d'erreurs mais le risque.