**SY09 Printemps 2009 – UTC**

**Théorie de la décision**

Exercice I : Classifieur euclidien

Le but de cet exercice va être d’étudier les performances du classifieur euclidien sur des données de 2 classes, avec des formes géométriques différentes.

1. *Génération des jeux de données*

On cherche à générer 2 jeux de données différents :

* **D1App  ->** de taille *n1App* = 300, correspondant à la classe *w1*, et suivant la loi normale *N* (µ1, ∑1 = I)
* **D2App  ->** de taille *n2App* = 300, correspondant à la classe *w2*, et suivant la loi normale *N* (µ2, ∑2 = *a*² I)

Pour cela, nous utilisons la fonction *mvrnorm*, qui nous permet de générer un échantillon issu d’une loi normale multidimensionnelle. Comme paramètres, il suffit alors de rentrer les valeurs de n, de l’espérance et de la variance.

Pour le jeu de données **D2App,** nous décidons de créer une fonction permettant de changer aisément les paramètres pouvant varier (µ2 et ∑2 = *a*² I):

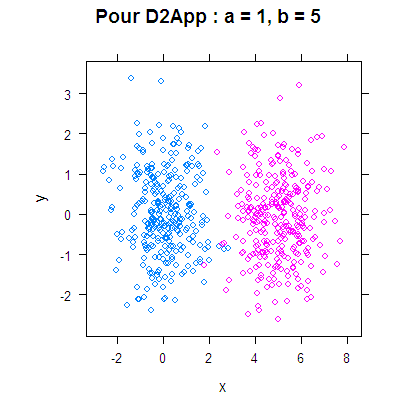
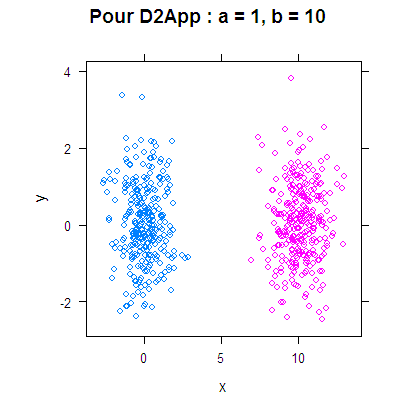
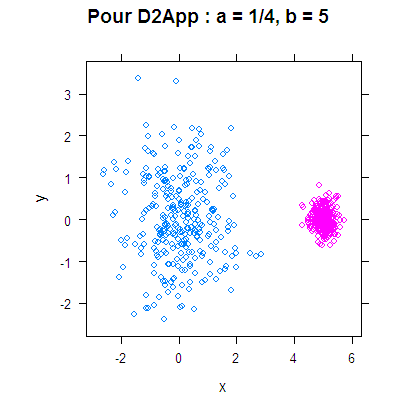
*LoiNormale2 <- function(a,b)*

*{ Sigma2 <- diag(rep(a\*a,2))*

*mu2 <- matrix(c(b,0),nrow =2)*

*res <- mvrnorm(n=300, mu2, Sigma2,tol = 1e-6, empirical = FALSE)*

*res}*

Nous obtenons donc 2 jeux de données (**D1App** et **D2App)** que nous pouvons représenter côté à côté :

Ainsi, en variant les paramètres a et b, et donc µ2 et ∑2 du 2nd jeu de données, nous pouvons obtenir plusieurs représentations des 2 classes dans l’espace. Celles-ci sont alors plus ou moins éloignées : par exemple, plus la variance est grande, plus le nuage de points est étalés (ce qui est en accord avec nos connaissances en statistique, puisque la variance représente le carré de la dispersion des 300 valeurs autour de leur moyenne ; donc plus la variance est grande, plus la dispersion est grande).

Cependant, il faut bien comprendre que pour générer ces jeux de données, nous avons fixé nos propres µ et nos propres ∑. Ce qui n’est pas le cas dans la réalité puisque nous ne connaissons ni la variance, ni l’espérance des échantillons. Le but de la 2nde partie de cet exercice va donc être de générer deux ensembles de test, dont nous ne fixons ni la variance, ni l’espérance, mais juste Theta, une liste composées des µ1 et µ2 obtenus précédemment.

1. *Estimation de la probabilité d’erreur du classifieur euclidien*

Pour la réalisation de cette estimation, nous programmons 2 fonctions :

* La fonction *regleEuclidienne* **->** elle va retourner les valeurs « 1 » ou « 2 » suivant l’action à prendre pour l’observation x à classer. C'est-à-dire qu’elle va regarder la distance euclidienne entre x et chaque centre (µ1 et µ2) des 2 classes. Autrement dit, si un x de la classe *w1* est plus éloigné de µ1 que de µ2, alors la fonction lui assimilera la valeur 2 ; ou si un x de la classe *w1* est plus proche de µ1 que de µ2, alors la fonction lui assimilera la valeur 1. Cela permet de voir si une observation x appartenant à une classe est « correctement classée ».
* La fonction *erreurEstimee* **->** elle va alors permettre d’estimer la probabilité d’erreur du classifieur euclidien. Dans la fonction précédente, on considère que si un échantillon x appartenant à la classe *w1* ressort avec une valeur de 2, alors il y a eu une erreur de classification (de même pour un individu x appartenant à la classe *w2* et ressortant avec une valeur de 1). Cette seconde fonction va donc permettre de sommer les erreurs et d’en tirer ainsi une probabilité d’erreur estimée.

Nous observons alors que cette probabilité correspond à 26% d’erreurs (avec a = 1, b = 10).

Rappelons que pour obtenir ces résultats, nous avions fixé un Theta correspondant à une liste composé des µ1 et µ2 obtenus lors de la question précédente. En réalité, pour obtenir le Theta, nous avions opté pour la fonction suivante :

*theta <- function(D1App,D2App)*

*{ rbind(apply(D1App,2,mean),apply(D2App,2,mean))}*

*theta <- theta(D1App,D2App)*

Ce qui revient à obtenir une moyenne des observations par classe, et ce qui revient donc à obtenir 2 nouvelles espérances. C’est à partir des ces nouvelles valeurs qui nous avons réalisé la 2nde question.

Il paraît donc normal que la probabilité d’erreur soit à ce point élevée : en effet, si nous avions gardé les µ1 et µ2 de la 1ière question, nous aurions obtenu des probabilités d’erreur quasi nulles puisque la classification de l’échantillon a été réalisée **en fonction de** l’espérance de la classe, les observations gravitaient donc autour de ce point. Or, dans ce cas, nous avons **redéfini** une espérance. La probabilité d’erreur que les observations d’une classe soient plus éloignées de cette nouvelle espérance et donc augmentée (rappelons également que nous avons gardé la même classification que dans la 1ière question).

1. *Estimation de la probabilité d’erreur du classifieur euclidien*

(Avec a = 1 et b = 10)

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Essais | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Moyenne | Var |
| % d’erreur | 26,5 | 24,0 | 26,3 | 25,7 | 26,2 | 23,3 | 27,2 | 25,9 | 25,2 | 29,5 | 25,9 | 0,03 |

La probabilité moyenne d’erreur du classifieur euclidien est donc de 25,9%.

Exercice II : **Règles de Neyman-Pearson et Bayes**

Le but de cet exercice va être de pouvoir différencier les avions des missiles.

1. *Génération des vecteurs pour les lois f1 et f2.*

*f1(x1) ~ N(−1, 1), f2(x1) ~ N(1, 1),*

*f1(x2) = f2(x2) ~ N(0, 1).*

*f1(x) = f1(x1)\* f2(x1) dans la classe ω1 et f2(x) = f2(x1)\*f2(x2) dans la classe ω2*

Nous utilisons donc la fonction *rnorm*. Nous obtenons donc 4 vecteurs : vect1X1 et vect1X2 pour la loi *f,1* et vect2X1 et vect2X2 pour la loi *f2*

1. *Estimation des différents paramètres des lois*

Nous allons comparer la moyenne des résultats obtenus avec les paramètres des lois.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Vect1X1 | | Vect1X2 | |
| Taille du jeu | Moyenne | Variance | Moyenne | Variance |
| 300 | -0.9392797 | 0.849767 | -0.01189009 | 0.9643012 |
| 1000 | -0.9614003 | 1.038214 | 0.03490163 | 0.9682516 |
| 2000 | -0.9992178 | 1.041234 | 0.003959581 | 1.032255 |

On remarque que plus il y a de valeurs, plus les valeurs empiriques se rapprochent des valeurs théoriques. On peut en conclure que plus la taille du jeu est important, plus on se rapproche des paramètres de la loi générée. Ceci est explicable grâce à la loi des grands nombres.

1. *Détermination des expressions f1(x) et f2(x)*

Chaque variable suit une loi Normale *N* (µ,σ) d’équation :

=> Pour f1(x) : => Pour f2(x)

Pour montrer que ces courbes sont des cercles d’iso-densités, il faut résoudre les 2 équations suivantes :

* f1(x) = M
* f2(x) = N où M et N des constantes

pour f1(x)

pour f2(x)

Soit :

Or l’équation d’un cercle est de la forme : où r est le rayon du cercle, a et b les coordonnées respectives du centre de ce cercle.

Nous observons que nous obtenons des équations de même type que celle d’un cercle. La courbe de f1(x) est donc un cercle centré en (1,0), de rayon et celle de f2(x) est un cercle centré en (-1,0) et de rayon .

1. *Etude du problème avec la règle de Neyman-Pearson*

**

* 1. **La règle pour ce problème s’exprime en fonction d’une seule variable**

On a, grâce à la règle de Neyman-Pearson :

Donc, à partir de , on trouve :

Après simplification de l’équation, on obtient :

On peut donc en conclure que ce problème s‘exprime en fonction d’une seule variable, qui est x1.

* 1. **Expression de la règle en fonction de α\***

L’équation de départ pour avoir l’expression est :

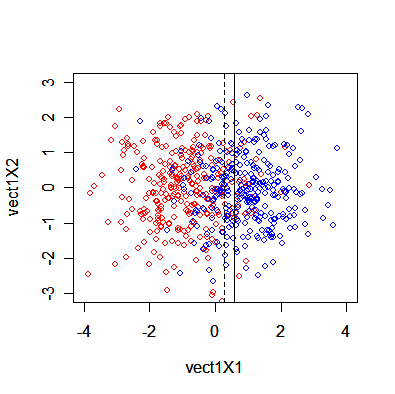
Or,

On obtient donc :

* 1. **Fonction qui permet de construire la frontière**

La fonction utilise l’équation obtenue précédemment :

frontiere = function(alpha)exp(-2\*qnorm(1-alpha)+2)



Nous obtenons donc les 2 frontières à l’aide de la fonction *abline*. La ligne en pointillé correspond à celle du plus petit α\*. On peut donc remarquer que plus α\* est petit, plus la frontière tend vers la gauche, et moins on a de point bleus qui se trouvent à gauche de la frontière.

α correspond à la probabilité d’erreur. Or α ≤ α\*, donc plus α\* est petit, plus la probabilité d’erreur sera faible (où la probabilité d’erreur est la probabilité de considérer un élément comme appartenant à la classe 1 alors qu’il appartient à la classe 2).

* 1. **Estimation de α et β**

Pour estimer α, on calcule la part des vecteurs vect1X1 qui sont supérieur à la valeur de la frontière et, pour β, on calcule la part des vecteurs vect2X1 qui sont inférieur à la valeur de la frontière.

|  |  |  |
| --- | --- | --- |
| α\* | 0,05 | 0,1 |
| α | 0.1166667 | 0.06666667 |
| β | 0.8833333 | 0.9333333 |

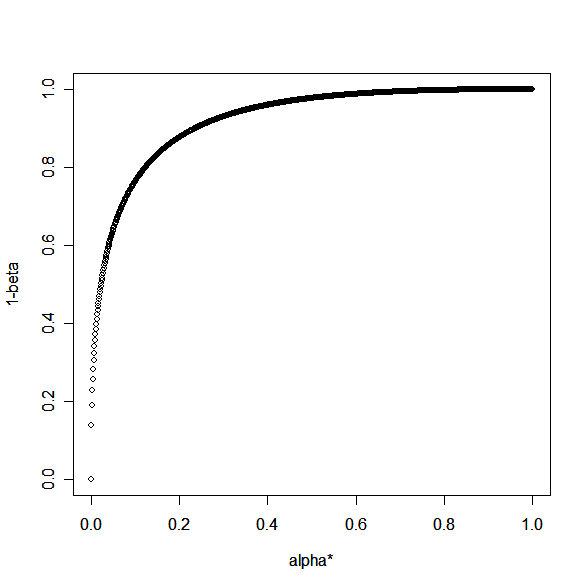
* 1. **Expression de la courbe COR 1-β = g(α\*)**

Pour obtenir l’équation de la courbe, on part de l’expression suivante :

Or, on a : donc

Alors :

On obtient donc :



La représentation de COR est donc :

(Cette courbe permet en général de choisir la valeur de α\*, en fonction de 1-)

1. *Etude du problème avec la règle de Bayles*

**

* 1. **Expression de Bayles**

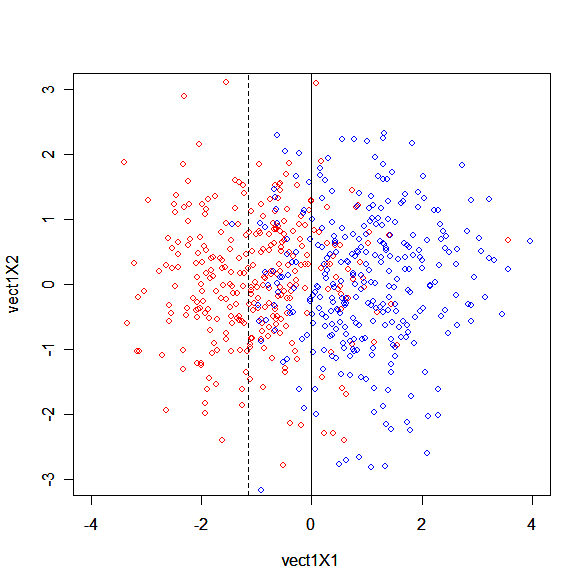
La règle de Bayles permet de minimiser le risque (π) en prenant en compte les coûts (c).

On a, grâce à la règle de Bayles :

* 1. **Graphique avec les frontières de décision correspondantes aux cas suivants :**

L’équation des frontières est égale à : .

Alors le graphe obtenu est le suivant :



On remarque que, pour le premier cas (i) (*frontière en trait continu*), la frontière est au centre des 2 nuages de points, ce qui semble correct puisque π1= π2 et c12=c21, ce qui implique une symétrie entre les 2 nuages.

Pour les 2 autres cas, les frontières (*pointillées*) se chevauchent et elles se situent plus à gauche. Pour un des cas (ii), minimiser le coût revient à minimiser le nombre de points bleus à gauche de la frontière. Pour l’autre cas (iii), minimiser le risque revient à prendre un maximum de points bleus dans la partie droite.

* 1. **Estimation de α et β**

On utilise la même méthode que pour l’estimation précédente.

|  |  |  |
| --- | --- | --- |
|  | Cas i | Cas ii et iii |
| α | 0.1433333 | 0.5633333 |
| β | 0.1566667 | 0.003333333 |

α correspond à la probabilité de considérer un élément comme appartenant à la classe 2 alors qu’il appartient a la classe 1, et inversement pour β.

Pour le premier cas (i), les coûts et les probabilités a priori sont quasi identique donc il est logique que α et β soient quasi identiques.

Pour les cas ii et iii, on remarque que α est nettement supérieur à β car, pour le cas ii, le cout encouru pour une action a1 alors que Z = 2 est très élevé et pour le cas iii, la probabilité a priori de la classe 2 est nettement supérieur à celle de l’autre classe.