**SY09 Printemps 2010 - TP 3**

**Théorie de la décision**

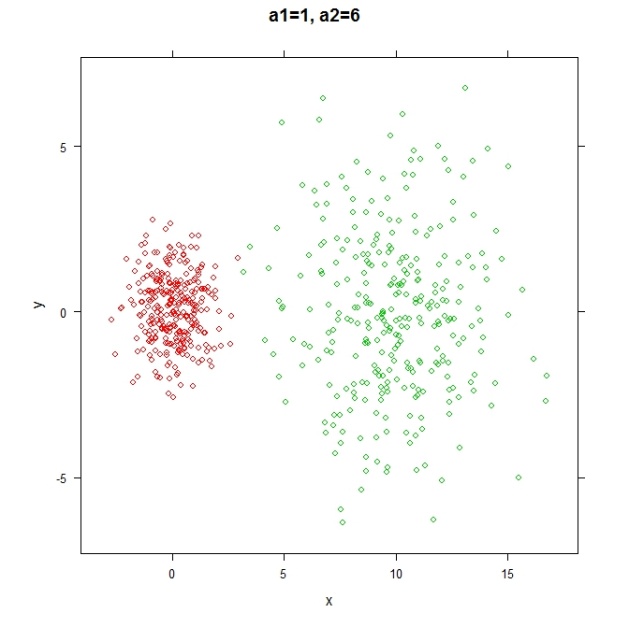
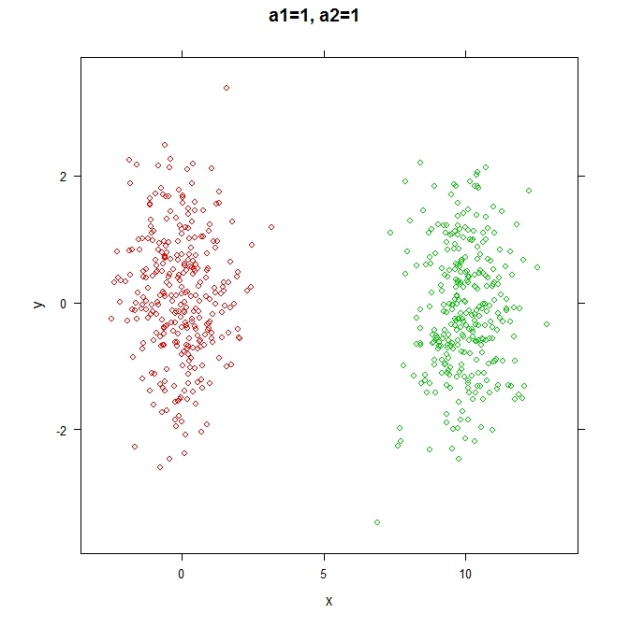
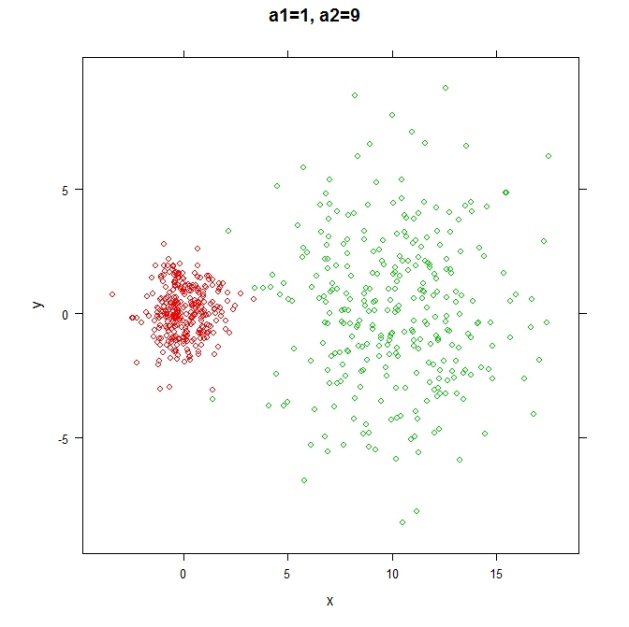
# **Classificateur euclidien**

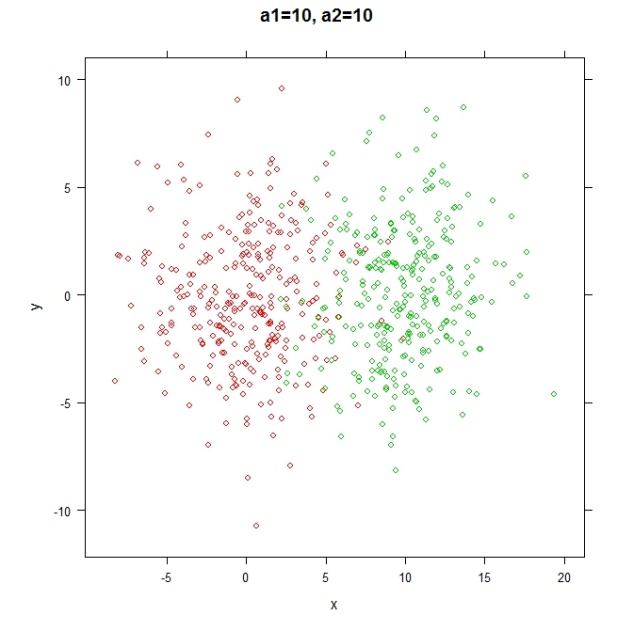
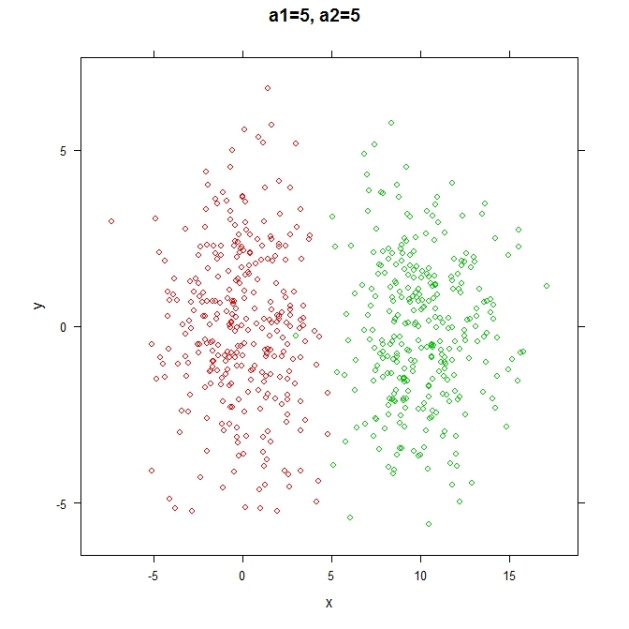
On veut pour ce tp, étudier les performances du classificateur euclidien sur des échantillons issus de deux classes suivant des lois normales différentes

### Simulation d’un échantillon

On génère deux ensembles D1App correspondant à la classe ω1 et D2App correspondant à classe ω2. Ces deux jeux ont la même taille n1App = n2App = 300. Ils suivent deux lois normales respectivement N(μ1,∑1 = I) et N(μ2,∑2 = a2I) dont on va pouvoir changer les paramètres.

On utilise alors la fonction *mvrnorm*, qui permet de générer un échantillon issu d’une loi normale multidimensionnelle. On donne comme paramètre le nombre de valeurs n, l’espérance μ et la variance ∑ = a\*I.

Avec n=600, μ 1=(0,0)’, μ2=(10,0)’, ∑1=a1\*I et ∑2=a2\*I, on obtient les résultats suivants :



On visualise les données dans l’espace qui sont alors plus ou moins dispersées : plus la variance est grande, plus le nuage de points est grand (ce qui est en accord avec nos connaissances en statistique, puisque la variance représente le carré de la dispersion autour de la moyenne).

### Estimation de la probabilité d’erreur

Pour les situations précédentes, on veut estimer l’erreur associée au classificateur euclidien. On va donc chercher les espérances μ ainsi que le taux d’erreur. Pour cela on écrit les fonctions suivantes :

* *regleEuclidienne* **:** qui renvoi les valeurs « 1 » ou « 2 » selon que l’élément appartient à la classe « 1 » ou « 2 ». on va donc calculer la distance euclidienne entre l’élément en question et chaque centre (µ1 et µ2) de classe pour renvoyer la valeur la plut petite. Cela permet de voir si un élément est « correctement classé » car si, appartenant à « 2 », il est plus proche de µ1, il sera considéré comme appartenant à « 1 ».
* *erreurEstimee*: qui calcule la probabilité d’erreur estimée sur un échantillon du classifieur euclidien. Pour cela, on considère que si un élément n’est pas correctement classé par *regleEuclidienne*  alors il y a eu une erreur de classification. On va sommer les erreurs et calculer ainsi une probabilité d’erreur estimée.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Situation** | **a1=1 a2=1** | **a1=1 a2=6** | **a1=1 a2=9** | **a1=5 a2=5** | **a1=10 a2=10** |
| Erreur Estimée | 0 | 0.01166667 | 0.025 | 0.01333333 | 0,05833333 |

On répète cette opération 10 et on calcul l’erreur estimée moyenne ainsi que la variance :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Situation** | **a1=1 a2=1** | **a1=1 a2=6** | **a1=1 a2=9** | **a1=5 a2=5** | **a1=10 a2=10** |
| Erreur Estimée moyenne | 0 | 0,023333375 | 0,02183333 | 0,01266667 | 0,054166667 |
| Variance | 0 | 9,16667E-16 | 1,8796E-05 | 1,3086E-05 | 7,42284E-05 |

On peut noter que malgré la dispersion apparente des points lors de la question 1, la probabilité d’erreur estimée reste assez faible (<5.4% pour a1=10 et a2=10).

# **Règles de Neyman-Person et Bayes**

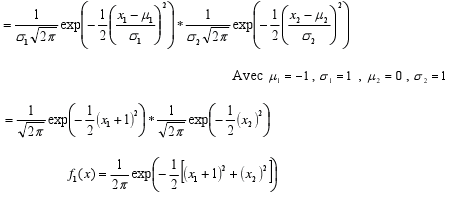
On considère un problème de détection de cibles dans lequel la classe ω1 correspond aux missiles, et la classe ω2 correspond aux avions. Chaque cible est décrite par deux variables X1 et X2 issues de deux capteurs différents. Chaque variable suit, dans chaque classe, une loi normale avec les paramètres suivants :

f11(x1) ∼ N(−1, 1) f21(x1) ∼ N(1, 1) f12(x2) = f22(x2) ∼ N(0, 1)

On suppose l'indépendance conditionnelle de X1 et X2. Les densités conditionnelles du vecteur X= (X1,X2)’ sont donc : f1(x) = f11(x1)f12(x2) dans la classe ω1 f2(x ) = f21(x1)f22(x2) dans la classe ω2.

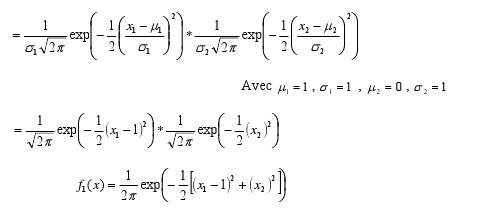
### Calcul de f1 et f2 :

f1 ( x ) = f11(x1) ∗ f12( x 2)



D’où f1(x) ∼ N, loi multi-normale avec µ=(-1,0)’ et ∑=I

f2 ( x ) = f21(x1) ∗ f22( x 2)



D’où f2(x) ∼ N, loi multi-normale avec µ=(1,0)’ et ∑=I

### Simulations

On utilise la fonction mvrnorm pour simuler des échantillons de différentes tailles, on calcule ensuite la moyenne et la variance de chaque vecteur pour constater ou non la tendance vers les valeurs théorique.

On a donc les résultats suivant pour la classe ω1 :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Taille de l’échantillon | 300 | 10 | 100 | 1000 | 10 000 | 100 000 |
| Mean X1 | -0.9859839 | -0.5946 | -1.00335 | -1.03772 | -0.998243 | -0.9988623 |
| Mean X2 | -0.00922008 | 0.26514 | 0.014959 | 0.0063340 | 0.0152704 | 0.00113909 |
| Var X1 | 1.036427 | 1.3699 | 1.02254 | 1.03777 | 0.984495 | 1.004036 |
| Var X2 | 0.8274417 | 0.2812 | 0.80886 | 1.00639 | 0.979752 | 1.001740 |

Et pour la classe ω2 :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Taille de l’échantillon | 300 | 10 | 100 | 1000 | 10 000 | 100 000 |
| Mean X1 | 1.060898 | 1.14333 | 0.95758 | 0.972161 | 0.9972412 | 1.000985 |
| Mean X2 | -0.020016 | -0.03626 | -0.0097831 | 0.0481819 | 0.01261037 | -0.000247 |
| Var X1 | 0.931692 | 0.93537 | 0.92436 | 0.993347 | 0.986927 | 1.002532 |
| Var X2 | 0.8626005 | 1.1693 | 1.116906 | 0.99587 | 0.9943532 | 1.001366 |

On peut voir que la précision empirique est d’autant plus forte que l’échantillon est important. On assimile ceci intuitivement à la loi des Grands Nombres.

### Courbes d’iso intensité

Pour montrer que les 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 donnée par . D’où le résultat suivant :

**et**

### Neyman-Pearson

### Recherche de la 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 : x1.

### Expression de la règle en fonction α\*

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

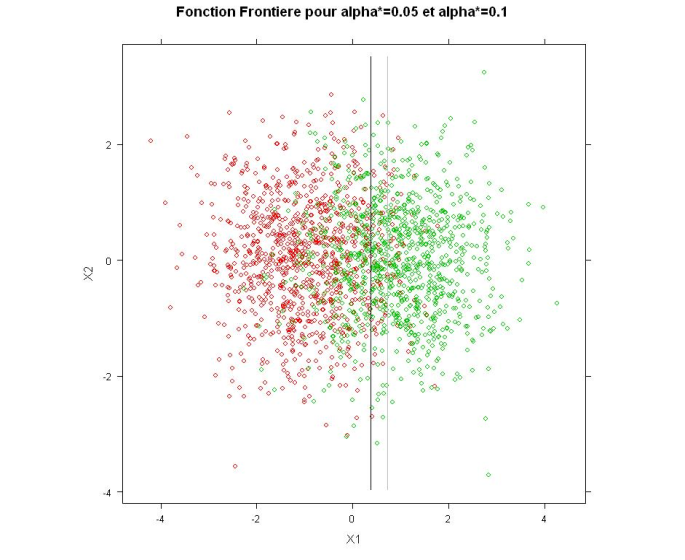
Or,

On obtient donc :

### Fonction frontière

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

La fonction utilise l’équation obtenue précédemment :



On obtient les 2 frontières à l’aide de la fonction *abline*.

La ligne noir correspond à α\*=0.05 et celle en gris à α\*=0.1. On peut donc remarquer que plus α\* est petit, plus la frontière tend vers la gauche, et moins on a de point vert qui se trouvent à gauche de la frontière. α correspondant à la probabilité d’erreur et α ≤ α\*, donc plus α\* est petit, plus la probabilité d’erreur est faible.

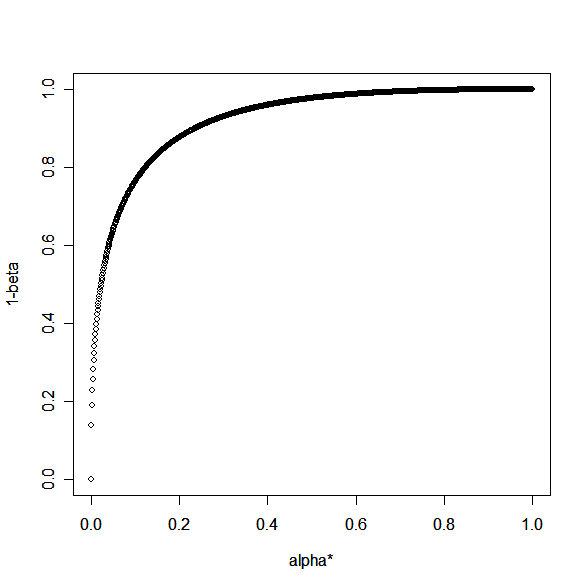
### Estimation de α et β

Pour estimer α, on cherche nombre d’éléments de ω2 qui dépassent la frontière et sont considérés comme appartenant à ω1, ce qui permettra de calculer la probabilité d’erreur.  On effectue le raisonnement inverse pour trouver β.

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

### Courbe COR 1-β =g(α\*)

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



Or : donc

D’où:

Et enfin :

On trouve donc la représentation ci-contre.

### Bayes

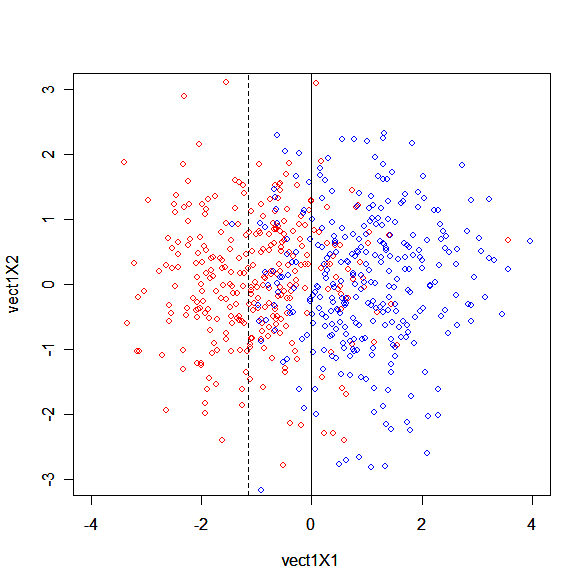
La règle de Bayes permet de minimiser le risque (π) en prenant en compte les coûts (c).

On a, grâce à la règle de Bayes :

### Graphique des frontières de décision :

L’équation des frontières est égale à : .

On obtient le graphe ci-après :



On remarque que, pour le premier cas (i) (*ligne continue*), la frontière est entre les 2 nuages de points, ce qui semble correct puisque π1= π2 et c12=c21, ce qui implique une équidistance.

Pour les 2 autres cas, les frontières (*en pointillées*) se chevauchent et 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.

### Estimation de α et β

On utilise la même méthode que précédemment.

|  |  |  |
| --- | --- | --- |
|  | 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 ω2 alors qu’il appartient a la classe ω1, 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 à1 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.