SY09 Printemps 2010
TP 3
Théorie de la décision

Exercice 1. Classifieur euclidien

On veut étudier les performances du classifieur euclidien sur des échantillons issus de deux classes ω_1 et ω_2 de $\mathbb{R}^2$ dont les distributions sont normales et de paramètres $(\boldsymbol{\mu}_1, a_1\mathbf{I})$ et $(\boldsymbol{\mu}_2, a_2\mathbf{I})$.

1. Simulation d'un échantillon : en utilisant la fonction `mvrnorm` de la bibliothèque `MASS`, écrire la fonction `simul` de paramètres n, $mu1$, $mu2$, $a1$ et $a2$ qui retourne un échantillon de taille n dont la moitié est issue de la classe ω_1 et l'autre moitié est issue de la classe ω_2. On supposera que ces deux sous-échantillons sont répartis de manière aléatoire dans l'échantillon total. La valeur retournée par cette fonction sera une matrice de dimension `(n,3)` contenant les deux variables correspondant à l'espace de simulation et la variable de classe.

   On utilisera cette fonction pour les cinq situations suivantes : $n = 600$, $\boldsymbol{\mu}_1 = (0,0)'$, $\boldsymbol{\mu}_2 = (10,0)'$, $\Sigma_1 = a_1\mathbf{I}$ et $\Sigma_2 = a_2\mathbf{I}$ avec $(a_1 = 1, a_2 = 1)$, $(a_1 = 1, a_2 = 6)$, $(a_1 = 1, a_2 = 9)$, $(a_1 = 5, a_2 = 5)$ et $(a_1 = 10, a_2 = 10)$. À chaque fois, on visualisera les données simulées.
2. Estimation de la probabilité d'erreur : pour chacune des cinq sitations, on cherche à estimer la probabilité d'erreur associé au classifieur euclidien. Pour ceci, l'échantillon simulé est coupé en deux : la première moitié forme un échantillon d'apprentissage permettant d'estimer les moyennes μ_1 et μ_2 et la seconde moitié forme un ensemble test permettant d'estimer le taux d'erreur.

   On pourra s'appuyer sur les deux fonctions suivantes :
   - La fonction `regleEuclidienne = function(x,mu1,mu2)` qui retourne la classe retenue par le classifieur euclidien pour l'observation x.
   - La fonction erreurEstimee = function(D,regle,mu1,mu2) qui retourne la probabilité d'erreur estimée sur l'échantillon `D`. Utiliser dans cette fonction la commande apply de la manière suivante : classement=apply(D,1,regle,mu1=mu1,mu2=mu2).
3. Répéter 10 fois la question 2 et calculer la moyenne et la variance des résultats ainsi obtenus.

Exercice 2. Règles de Neyman-Pearson et Bayes

On considère un problème de détection de cibles dans lequel la classe ω_1 correspond aux missiles et la classe ω_2 correspond aux avions. Chaque cible est décrite par deux variables X_1 et X_2 issues de deux capteurs différents. Chaque variable suit, dans chaque classe, une loi normale avec les paramètres suivants :

$$f_{11}(x_1) \sim \mathcal{N}(-1,1), \quad f_{21}(x_1) \sim \mathcal{N}(1,1),$$

$$f_{12}(x_2) = f_{22}(x_2) \sim \mathcal{N}(0,1).$$

On suppose l'indépendance conditionnelle de X_1 et X_2. Les densités conditionnelles du vecteur $\mathbf{X} = (X_1, X_2)'$ sont donc $f_1(\mathbf{x}) = f_{11}(x_1)f_{12}(x_2)$ dans la classe ω_1 et $f_2(\mathbf{x}) = f_{21}(x_1)f_{22}(x_2)$ dans la classe ω_2.

Dans tout cet exercice, on suppose que les distributions sont connues et ne sont donc pas estimées à partir d'un échantillon.

1. Montrer que les distributions f_1 et f_2 sont des distributions normales dont on précisera les vecteurs moyenne et les matrices de variance.
2. En utilisant la fonction `mvrnorm`, simuler deux échantillons de taille $n = 300$ issus des classes ω_1 et ω_2 et déterminer pour chacun des échantillons les estimations des différents paramètres de f_1 et f_2. On effectuera ce travail pour les valeurs de n suivantes : 10,100,1000,10000,100000. Interpréter ces résultats.
3. Montrer que les courbes d'iso-densité sont des cercles dont on précisera les rayons.
4. Soit $\mathcal{A} = \{a_1, a_2\}$ les actions d'affectation aux classes ω_1 et ω_2. Pour une règle de décision δ, on définit les probabilités d'erreur $\alpha = P(\delta(\mathbf{X}) = a_2|\omega_1)$ et $\beta = P(\delta(\mathbf{X}) = a_1|\omega_2)$. On rappelle que la règle de Neyman-Pearson minimise β sous la contrainte $\alpha \leq \alpha^*$ pour une valeur α^* fixée appelée niveau de signification.

(a) Montrer que la règle de Neyman-Pearson pour ce problème s'exprime en fonction d'une seule variable. Interpréter ce résultat.

(b) Donner l'expression de cette règle en fonction de α^*.

(c) Construire une fonction, qui en fonction de α^*, dessine la frontière de décision correspondante dans le plan (X_1, X_2). Application : $\alpha^* = 0.05$ et $\alpha^* = 0.1$.

(d) À partir des données simulées précédemment, donner une estimation de α et β.

(e) Donner l'expression de la courbe COR $1 - \beta = g(\alpha^*)$, et tracer cette courbe avec R.

5. Soient π_1 et π_2 les probabilités a priori des deux classes, et c_{lk} le coût associé au choix de l'action a_ℓ lorsque la vraie classe est ω_k. On suppose $c_{11} = c_{22} = 0$. L'ensemble $\mathcal{A}$ des actions est le même que dans la question précédente.

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

   (b) Tracer avec R les frontières de décision correspondantes dans le plan (X_1, X_2) dans les cas suivants :

      i. $c_{12} = c_{21} = 1$, $\pi_1 = \pi_2$;

      ii. $c_{12} = 10$, $c_{21} = 1$, $\pi_1 = \pi_2$;

      iii. $c_{12} = c_{21} = 1$, $\pi_2 = 10\pi_1$.

   (c) Pour ces différents cas, et à partir des données générées précédemment, donner une estimation de $\alpha = P(\delta^*(\mathbf{X}) = a_2|\omega_1)$ et $\beta = P(\delta^*(\mathbf{X}) = a_1|\omega_2)$. Commenter ces résultats.