SY09 Printemps 2014
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 $(\boldsymbol{\mu}_1, \Sigma_1)$ et $(\boldsymbol{\mu}_2, \Sigma_2)$.

Simulation d'un échantillon

En utilisant la fonction mvrnorm de la bibliothèque MASS, écrire la fonction simul de paramètres n, pi, mu1, mu2, Sigma1 et Sigma2 qui retourne un échantillon de taille n tiré suivant une proportion π d'exemples issus de la classe ω_1. Ainsi, pour chaque exemple, il s'agira dans un premier temps de tirer au hasard la classe dont il est issu, avant de le générer en utilisant les paramètres adéquats. La valeur retournée par cette fonction sera une matrice de dimension $(n \times 3)$ contenant les deux variables correspondant à l'espace de simulation et la variable de classe.
On utilisera cette fonction pour les cinq situations suivantes : $n = 600$, $\pi = 1/2$, $\boldsymbol{\mu}_1 = (0,0)^\top$, $\boldsymbol{\mu}_2 = (10,0)^\top$, $\Sigma_1 = a_1\,\mathrm{Id}$ et $\Sigma_2 = a_2\,\mathrm{Id}$ avec $(a_1 = 1, a_2 = 1)$, $(a_1 = 1, a_2 = 5)$, $(a_1 = 1, a_2 = 9)$, $(a_1 = 5, a_2 = 5)$ et $(a_1 = 9, a_2 = 9)$. À chaque fois, on visualisera les données simulées.

Estimation de la probabilité d'erreur

Pour chacune des cinq situations, on cherche à estimer la probabilité d'erreur associée 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(ech,regle,mu1,mu2) qui retourne la probabilité d'erreur estimée sur l'échantillon ech. Utiliser dans cette fonction la commande apply de la manière suivante : `classement=apply(ech,1,regle,mu1=mu1,mu2=mu2)`

Probabilité d'erreur moyenne

Répéter 10 fois les questions 1 et 2, et estimer la moyenne et la variance des résultats ainsi obtenus. Proposer un intervalle de confiance de niveau 5% sur l'espérance de la probabilité d'erreur.

Exercice 2. Règle de 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 :

$$f_{11}(x_1) \sim \mathcal{N}(-1,1), \quad f_{21}(x_1) \sim \mathcal{N}(1,1),$$
$$f_{12}(x_2) \sim \mathcal{N}(-1,1), \quad f_{22}(x_2) \sim \mathcal{N}(1,1).$$

On suppose l'indépendance conditionnelle de X_1 et X_2. Les densités conditionnelles du vecteur $\mathbf{X} = (X_1, X_2)^\top$ sont donc $f_1(\mathbf{x}) = f_{11}(x_1)f_{12}(x_2)$ dans la classe ω_1 et $f_2(\boldsymbol{x}) = f_{21}(x_1)f_{22}(x_2)$ dans la classe ω_2. Dans tout cet exercice, on suppose que les distributions sont connues : elles ne doivent donc pas être estimées à partir d'un échantillon.

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.
2. En utilisant la fonction `simul`, générer un échantillon de n réalisations issues des deux classes ω_1 et ω_2 en proportions égales ($\pi_1 = \pi_2 = 0.5$) ; pour chacun des échantillons, déterminer 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. Soient π_1 et π_2 les probabilités a priori des deux classes, et $c_{\ell k}$ 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, donner une estimation des risques $\alpha = \mathbb{P}(\delta^*(\mathbf{X}) = a_2|\omega_1)$ et $\beta = \mathbb{P}(\delta^*(\mathbf{X}) = a_1|\omega_2)$ en utilisant la fonction `simul`. Commenter ces résultats.