SY19 – Examen Final Automne 2006-2007
Partie II : exercices

Documents autorisés. Temps : 1 heure.

Exercice 1

Soit un problème de discrimination à deux classes $\Omega = \{\omega_1, \omega_2\}$ et p variables $\mathbf{X} = (X^1, \ldots, X^p)^t$. Nous supposons que les p variables sont indépendantes conditionnellement à la classe et que la densité conditionnelle de la variable X^j conditionnellement à la classe ω_k est :

$$f_k(x^j) = \begin{cases} \theta_k^j \exp\left(-\theta_k^j x^j\right) & \text{si } x^j \geq 0 \\ 0 & \text{sinon,} \end{cases}$$

où θ_k^j est un paramètre caractérisant la distribution de la variable X^j dans la classe ω_k.

1. Donner l'expression de la densité conditionnelle $f_k(\mathbf{x})$ du vecteur $\mathbf{X}$ conditionnellement à la classe ω_k.
2. Calculer $h(\mathbf{x}) = \log(f_1(\mathbf{x})/f_2(\mathbf{x}))$.
3. Donnez l'expression de la règle de Bayes avec coûts 0-1, en fonction de $h(\mathbf{x})$ et des probabilités a priori des classes π_1 et π_2.

Exercice 2

On considère un problème de discrimination à deux classes $\Omega = \{\omega_1, \omega_2\}$ et une variable $X \in \mathbb{R}$. On suppose que la variable X suit dans une chaque classe une loi normale avec les espérances μ_1 et μ_2, avec $\mu_2 > \mu_1$, et une variance égale à 1. Les probabilités a priori sont notées π_1 et π_2.

1. Montrer que la règle de Bayes avec coûts 0-1 revient à comparer x à un seuil s que l'on précisera.
2. Donner l'expression littérale de la probabilité d'erreur de Bayes ϵ^*.
3. On dispose des données d'apprentissage suivantes :
   - classe ω_1 : 0.1, 0.2, 0.6, 0.7 ;
   - classe ω_2 : 0.5, 0.57, 1.5

   (a) Estimer numériquement les paramètres μ_1, μ_2, π_1 et π_2.

   (b) Calculer le seuil s de la question 1.

   (c) Estimer numériquement la probabilité d'erreur de Bayes ϵ^*, en utilisant la formule de la question 2 (on a $\phi(0.4) \approx 0.7$ et $\phi(0.9) \approx 0.8$).

   (d) Estimer la probabilité d'erreur de la règle du plus proche voisin, par la méthode du "leave-one-out".
4. On applique sur les données précédentes la méthode des arbres de décision, en choisissant l'entropie comme critère d'impureté. Quel est le gain d'information (réduction de l'entropie) associé au test $X < 0.3$?