SY09 Printemps 2019
TP 8
Éléments de théorie bayésienne de la décision

Dans ce TP, on souhaite étudier les stratégies de Neyman-Pearson et de Bayes pour résoudre un problème de décision, dans le cas où les distributions conditionnelles voire les probabilités a priori sont connues.

Problème et modélisation

On considérera un problème (simplifié) d'identification de produits chimiques à partir de leur temps de dégradation. Plus particulièrement, on supposera être en présence de $g = 2$ produits, présents en proportions initiales π_1 (classe ω_1), et π_2 (classe ω_2).

On suppose pouvoir tester le temps de dégradation des produits selon deux protocoles ; on notera X^1 et X^2 les temps de dégradation selon le premier et le second protocole, respectivement. Pour un même produit, on supposera indépendants ces temps X^1 et X^2 mesurés par chacun des deux protocoles. On suppose en outre que les distributions de temps de dégradation sont modélisés par des lois exponentielles

$$\begin{aligned} X^1 &\underset{\omega_1}{\sim} \mathcal{E}(\lambda_1), \qquad & X^2 &\underset{\omega_1}{\sim} \mathcal{E}(\lambda_2); \\ X^1 &\underset{\omega_2}{\sim} \mathcal{E}(\theta_1), \qquad & X^2 &\underset{\omega_2}{\sim} \mathcal{E}(\theta_2). \end{aligned}$$

Questions préliminaires

1. Quelle est la densité jointe du vecteur aléatoire $\boldsymbol{X} = (X^1, X^2)^T$ dans chacune des classes ?
2. Montrer que la frontière de décision obtenue en appliquant la stratégie de Neyman-Pearson est une droite dont on précisera les paramètres.
3. Quelle frontière de décision obtient-on si l'on applique la stratégie de Bayes ?

Simulation

On cherche à générer un échantillon de taille n suivant le modèle génératif décrit ci-dessus. Intuitivement, pour chaque nouvel individu, il faudrait (1) déterminer sa classe puis (2) générer un vecteur aléatoire (composante par composante, les variables X^1 et X^2 étant indépendantes conditionnellement à la classe) suivant les paramètres correspondants.

On propose donc le protocole de simulation suivant :

1. calculer le nombre de points n_1 présents dans la classe ω_1 : on verra n_1 comme la réalisation d'une v.a. $N_1 \sim \mathcal{B}(n, \pi_1)$;
2. générer n_1 vecteurs $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_1}$ en concaténant n_1 réalisations de variables aléatoires $X_1^1 \sim \mathcal{E}(\lambda_1)$ et $X_1^2 \sim \mathcal{E}(\lambda_2)$;
3. générer $n_2 = n - n_1$ vecteurs $\boldsymbol{x}_{n_1+1}, \ldots, \boldsymbol{x}_n$ en concaténant n_2 réalisations de variables aléatoires $X_2^1 \sim \mathcal{E}(\theta_1)$ et $X_2^2 \sim \mathcal{E}(\theta_2)$.

Questions.

1. Implémenter ce protocole de simulation, et générer un échantillon étiqueté de taille $n = 10000$, en utilisant $\pi_1 = 0.6$, $\lambda_1 = 1$, $\lambda_2 = 2$, $\theta_1 = 2$, $\theta_2 = 4$.
2. Estimer le taux d'erreur de Bayes au moyen de cet échantillon.
3. En utilisant le protocole d'évaluation des performances vu précédemment, appliquer la stratégie des K plus proches voisins à ce problème de décision. Comparer les performances obtenues au taux d'erreur de Bayes.
4. (Subsidiaire) Justifier rigoureusement l'estimation de la probabilité d'erreur de Bayes $\epsilon^\star$ par la moyenne empirique des erreurs commises par la règle de Bayes $\delta^\star$ sur un échantillon donné.