

## SY09

### TP — Éléments de théorie bayésienne de la décision

numpy=1.26.4; seaborn=0.13.2; matplotlib=3.8.3; pandas=2.2.0

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.

## 1 Partie théorique

### 1.1 Problème et modélisation

On considérera un problème très 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  $\pi_1$  (classe  $\omega_1$ ) et  $\pi_2$  (classe  $\omega_2$ ).

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 :

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

### 1.2 Questions préliminaires

- 1] Quelle est la densité jointe du vecteur aléatoire  $\mathbf{X} = (X^1, X^2)^T$  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 ?
- 4] Indépendamment des hypothèses de distribution formulées dans cet exercice, justifier rigoureusement l'estimation de la probabilité d'erreur de Bayes  $\epsilon^*$  par la moyenne empirique des erreurs commises par la règle de Bayes  $\delta^*$  sur un échantillon donné.
- 5] Toujours indépendamment des hypothèses de distribution formulées dans cet exercice, on veut en déduire l'expression de l'erreur de Bayes pour les problèmes de classification binaires ( $g = 2$ ).
  - 5 a] Montrer que l'erreur théorique de tout classifieur binaire peut s'écrire

$$\varepsilon(\delta) = \int_{\mathcal{R}_2} \pi_1 f_1(\mathbf{x}) d\mathbf{x} + \int_{\mathcal{R}_1} \pi_2 f_2(\mathbf{x}) d\mathbf{x}.$$

5 b) Montrer que dans le cas du classifieur de Bayes, cette erreur s'écrit

$$\varepsilon(\delta^*) = \int_{\mathcal{X}} \min(\pi_1 f_1(\mathbf{x}), \pi_2 f_2(\mathbf{x})) d\mathbf{x},$$

avec  $f_1$  et  $f_2$  les densités conditionnelles et  $\pi_1$  et  $\pi_2$  les proportions des classes.

## 2 Partie pratique

### 2.1 Protocole de 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  $\omega_1$  : on verra  $n_1$  comme la réalisation d'une v.a.  $N_1 \sim \mathcal{B}(n, \pi_1)$  ;
2. générer  $n_1$  vecteurs  $\mathbf{x}_1, \dots, \mathbf{x}_{n_1}$  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  $\mathbf{x}_{n_1+1}, \dots, \mathbf{x}_n$  en concaténant  $n_2$  réalisations de variables aléatoires  $X_2^1 \sim \mathcal{E}(\theta_1)$  et  $X_2^2 \sim \mathcal{E}(\theta_2)$ .

#### Implémentation

6) Implémenter ce protocole de simulation, et générer un échantillon et ses étiquettes de taille  $n = 1000$ , en utilisant  $\pi_1 = 0.6$ ,  $\lambda_1 = 1$ ,  $\lambda_2 = 2$ ,  $\theta_1 = 2$ ,  $\theta_2 = 1$ .

7) Afficher le jeu de données avec l'information de classe. Quel serait un bon séparateur linéaire ?

8) Estimer le taux d'erreur de Bayes au moyen de cet échantillon. On pourra utiliser la fonction `np.apply_along_axis`.

9) Tracer la frontière de Bayes correspondante. On utilisera la fonction `add_decision_boundary` du TP07 en lui donnant directement une fonction de prédiction (et pas un objet `scikit-learn`) et en précisant l'argument `model_classes`.

```
def predict(X):  
    return ...
```

10) 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.

#### Calcul numérique de l'erreur de Bayes.

11) Créer une fonction en Python prenant en argument une abscisse et une ordonnée et renvoyant la fonction sous le signe intégrale.

On pourra utiliser la fonction `expon` définie dans le module `scipy.stats` pour calculer la densité d'une loi exponentielle.

12) En utilisant la fonction `dblquad` d'intégration numérique d'une intégrale double, donner un encadrement de l'erreur de Bayes à  $10^{-3}$  près.

```
from scipy.integrate import dblquad
```



### Calcul numérique de l'erreur d'un classifieur

Dans cette partie, on se propose de calculer numériquement l'erreur théorique d'un classifieur binaire  $\delta$  quelconque. D'après la question 5, on sait que l'erreur de  $\delta$  peut s'écrire

$$\varepsilon(\delta) = \int_{\mathcal{R}_1} \pi_2 f_2(\mathbf{x}) \, d\mathbf{x} + \int_{\mathcal{R}_2} \pi_1 f_1(\mathbf{x}) \, d\mathbf{x}.$$

En d'autres termes, il suffit d'identifier les deux régions de décision complémentaires de  $\delta$ , en utilisant le mécanisme de classement implémenté par le classifieur ; puis d'intégrer la bonne densité conditionnelle sur chacune des régions pour calculer son erreur théorique.

Nous allons utiliser la méthode de Simpson pour calculer numériquement ces intégrales. Pour cela, il faut commencer par discrétiser l'espace sur lequel on va intégrer. Une discrétisation uniforme ne présentant guère de sens pour notre problème, on choisit une discrétisation basée sur des fractiles exponentiels.

**13** À l'aide de `np.linspace`, générer une discrétisation uniforme des ordres de fractiles entre 0 et 1 exclu.

**14** À l'aide de la méthode `ppf` calculant les fractiles, en déduire une discrétisation exponentielle des abscisses et des ordonnées.

Il faut à présent calculer en chacun des points repérés par leur abscisse et leur ordonnée les deux densités possibles  $f_1(\mathbf{x})$  et  $f_2(\mathbf{x})$ .

**15** Créer un tableau bidimensionnel des densités pour  $f_1$ . Faire de même pour  $f_2$ .

Il faut maintenant identifier les deux régions de décision et prédisant la classe pour chaque point de la discrétisation bidimensionnelle.

**16** Créer un tableau bidimensionnel donnant la classe pour chaque point de la discrétisation bidimensionnelle.

**17** Créer un tableau bidimensionnel calculant la valeur de la fonction à intégrer en utilisant la prédiction de la classe.

On pourra utiliser la fonction `np.where`.

**18** Intégrer numériquement la fonction obtenue. On utilisera deux intégrations successives avec la méthode de Simpson.

```
from scipy.integrate import simps
simps(simps(<tableau bidim à intégrer>, <discrétisation des y>),
      ↪ <discrétisation des x>)
```