Classification supervisée

Février 2019, Sans document, 1h30

Préambule : Le sujet est composé de quatre exercices indépendants. La qualité de la rédaction sera prise en compte. Toutes les réponses seront données sur la copie (ne pas rendre le sujet).

Exercice 1

On dispose de n observations $(x_1, y_1), \dots, (x_n, y_n)$ telles que $x_i \in \mathbb{R}$ et $y_i \in \{0, 1\}$ pour $i = 1, \dots, n$. On souhaite expliquer les sorties y_i par les entrées x_i.

1. Rappeler la définition d'une règle de prévision.
2. Rappeler la définition de la règle de Bayes $g^\star$ et de l'erreur de Bayes $L^\star$.
3. En quoi la règle de Bayes est-elle optimale ?
4. On considère (X, Y) un couple aléatoire à valeurs dans $\mathbb{R} \times \{0, 1\}$ tel que

$$X \sim \mathcal{U}(0, 100) \qquad \text{et} \quad (Y|X = x) \sim \begin{cases} \mathcal{B}(1/3) & \text{si } X \leq 20 \\ \mathcal{B}(3/4) & \text{si } X > 20 \end{cases}$$

   où $\mathcal{U}[a, b]$ désigne la loi uniforme sur $[a, b]$ et $\mathcal{B}(p)$ la loi de Bernoulli de paramètre $p \in [0, 1]$. Calculer la règle de Bayes et l'erreur de Bayes.

Exercice 2

On considère un n-échantillon i.i.d. $(x_1, y_1), \dots, (x_n, y_n)$ où $x_i \in \mathbb{R}^2$ et $y_i \in \{0, 1\}$. On désigne par g_0 et g_1 les centres de gravités des 2 groupes :

$$g_0 = \frac{1}{\text{card}\{i : y_i = 0\}} \sum_{i:y_i=0} x_i \quad \text{et} \quad g_1 = \frac{1}{\text{card}\{i : y_i = 1\}} \sum_{i:y_i=1} x_i.$$

1. Rappeler la définition des variances intra (W) et inter (B).
2. Soit u un vecteur de $\mathbb{R}^2$. Exprimer les variances intra $W(u)$ et inter $B(u)$ projetées sur la droite vectorielle engendrée par u en fonction des variances calculées à la question 1. On prendra soin de justifier les résultats.
3. Trouver un vecteur $u^\star$ qui maximise $B(u)$ sous la contrainte $1 - W(u) = 0$. On prendra soin de justifier les résultats.

Exercice 3

On cherche à expliquer une variable aléatoire Y à valeurs dans $\{0, 1\}$ par une variable aléatoire X à valeurs dans $\mathbb{R}$.

1. Ecrire proprement le modèle d'analyse discriminante linéaire dans ce contexte.
2. Quels sont les paramètres à estimer ? On précisera la nature de ces paramètres (réels, entiers, matrices, vecteurs de dimension...).
3. On suppose que $\mathbf{P}(Y = 1) = \mathbf{P}(Y = 0) = 1/2$. On obtient les estimations suivantes des paramètres du modèle LDA :

$$\widehat{\mu}_0 = -3, \quad \widehat{\mu}_1 = 6 \quad \text{et} \quad \widehat{\sigma^2} = 4.$$

   Calculer la règle de prévision issue de ce modèle.
4. Même question lorsque $\mathbf{P}(Y = 1) = 1/3$ et $\mathbf{P}(Y = 0) = 2/3$.

Exercice 4

Soit (X, Y) un vecteur aléatoire à valeurs de $\mathbb{R}^2 \times \{0, 1\}$. Pour $k \in \{0, 1\}$, la loi de $X|Y = k$ est un vecteur Gaussien d'espérance $\mu_k \in \mathbb{R}^2$ et de matrice de variance covariance Σ. On génère 4 échantillons $(x_i, y_i), i = 1, \dots, n$ de taille $n = 100$ selon ce modèle. Pour chaque échantillon on utilise les mêmes valeurs pour les espérances des lois : $\mu_0 = (-3, 0)$ et $\mu_1 = (3, 0)$. Chaque échantillon correspond à une matrice de variance covariance parmi les 4 suivantes :

$$\Sigma_1 = \begin{pmatrix} 3 & -0.95 \\ -0.95 & 1 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 1 & 0.95 \\ 0.95 & 1 \end{pmatrix}, \quad \Sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \Sigma_4 = \begin{pmatrix} 1 & -0.95 \\ -0.95 & 3 \end{pmatrix}.$$

Les 4 échantillons sont représentés sur la figure 1.

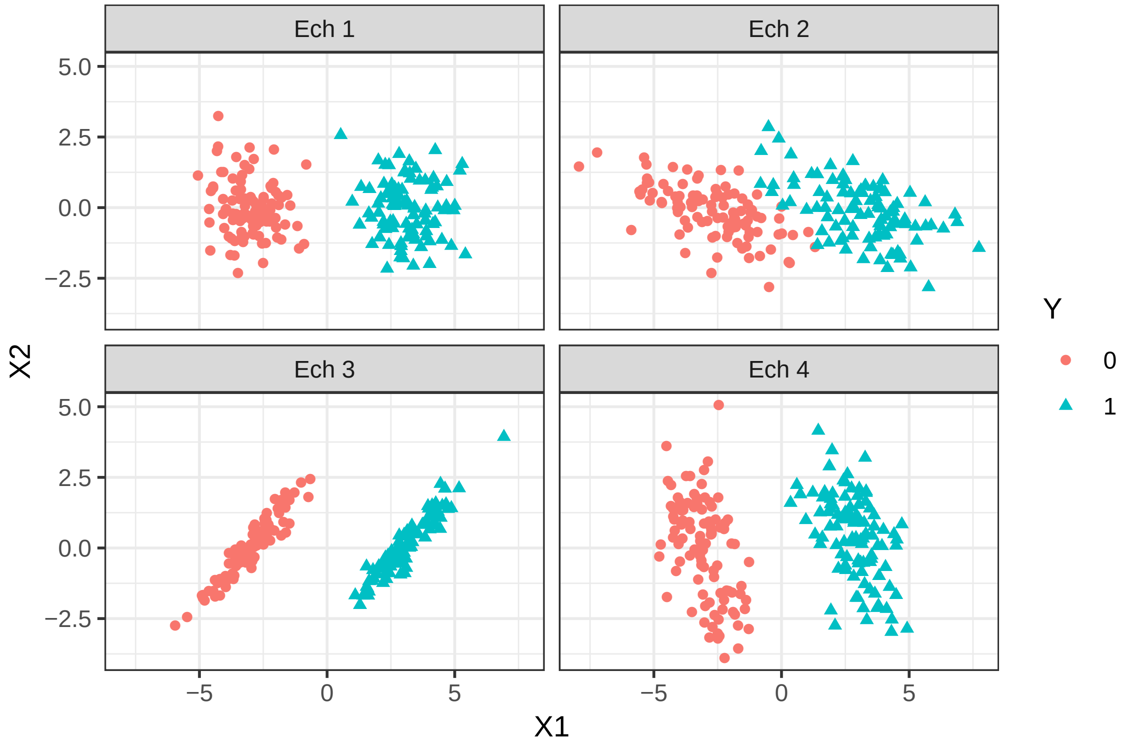

FIGURE 1 – Les 4 échantillons générés.

1. Associer chaque échantillon à la bonne matrice de variance covariance. Justifier brièvement.
2. On effectue une analyse discriminante linéaire sur **un de ces 4 échantillons**. On obtient les sorties R suivantes :

```
> lda(Y~.,data=df)

Prior probabilities of groups:
  0   1
0.5 0.5

Group means:
         X1          X2
0 -3.000000  0.00000000
1  3.000000  0.00000000

Coefficients of linear discriminants:
         LD1
V1  3.100000
V2 -2.800000
```

   On considère la règle de classification issue de cette analyse discriminante. Calculer la frontière de cette règle.
3. Sur quel échantillon de la figure 1 l'analyse discriminante linéaire a t-elle été ajustée ? Justifier.