TD : Supervisée

EXERCICE 1 :

① D'après le cours, on a $S = B + W$.

Donc $AAAAA = 10.81546 + (-4.934314)$

$BBBBB = 30.720944 - 35.36415$

$CCCCC = -4.620601 - 5.37166$

② Par définition du pouvoir discriminant, c'est la plus grande valeur propre de la matrice $S^{-1}B$. L'énoncé donne $\lambda_1 = 0.751$.

③ La fonction discriminante de Fisher est $f(x) = a^T x$ où a est égale à a_1 le vecteur propre associé à la première valeur propre de $S^{-1}B$. Donc $f(x) = a_1^T x = -0.77$ (en prenant $x = (-0.5, 0.5, 0, 1, 1)^T$.

Notons, pour $k = 1,2,3$, $\tilde{X}_k = (\bar{X}_{k1}, \bar{X}_{k2}, \bar{X}_{k3}, \bar{X}_{k4}, \bar{X}_{k5})^T$

On a $f(\tilde{X}_1) = 1.21$

$f(\tilde{X}_2) = 0.21$

$f(\tilde{X}_3) = -1.10$

On classe x tel que $\underset{k}{\operatorname{argmin}} |f(x) - f(\tilde{X}_k)|$

On assigne donc l'individu x à la classe 3.

Exercice 2 :

On cherche à prédire Y à partir de X_1, X_2 et X_3. Les variables sont binaires $(0, 1)$. On cherche à créer le 1^{er} embranchement de l'arbre.

	$Y=1$	$Y=0$
$X_1=1$	5	5
$X_1=0$	5	5

	$Y=1$	$Y=0$
$X_2=1$	10	0
$X_2=0$	0	10

	$Y=1$	$Y=0$
$X_3=1$	3	9
$X_3=0$	7	1.

On peut construire les arbres si on coupe suivant X_1, X_2 ou X_3.

20 obs (R_0)
$X_1 = 1$ / \ $X_1 = 0$
10 obs (R_1) | 10 obs (R_2)
$5 \to Y=1$ | $5 \to Y=1$
$5 \to Y=0$ | $5 \to Y=0$

20 obs (R_0)
$X_2 = 1$ / \ $X_2 = 0$
10 obs (R_1) | 10 obs (R_2)
$10 \to Y=1$ | $10 \to Y=0$

20 obs (R_0)
$X_3 = 1$ / \ $X_3 = 0$
12 obs (R_1) | 8 obs (R_2)
$3 \to Y=1$ | $7 \to Y=1$
$9 \to Y=0$ | $1 \to Y=0$

On note $\hat{p}_{jk}$ la proportion d'obs. de la région R_j qui appartiennent à la classe k. Notons le cas où $Y=0$, la classe 0 et le cas où $Y=1$, la classe 1.

Pour tous les arbres, on a : $\hat{p}_{00} = \frac{1}{2}$ et $\hat{p}_{01} = \frac{1}{2}$.

$\hat{p}_{10} = \frac{5}{10} = \frac{1}{2}$

$\hat{p}_{11} = \frac{1}{2}$

$\hat{p}_{20} = \frac{1}{2}$

$\hat{p}_{21} = \frac{1}{2}$

$\hat{p}_{10} = \frac{0}{10} = 0$

$\hat{p}_{11} = \frac{10}{10} = 1$

$\hat{p}_{20} = \frac{10}{10} = 1$

$\hat{p}_{21} = \frac{0}{10} = 0$

$\hat{p}_{10} = \frac{9}{12} = \frac{3}{4}$

$\hat{p}_{11} = \frac{3}{12} = \frac{1}{4}$

$\hat{p}_{20} = \frac{1}{8}$

$\hat{p}_{21} = \frac{7}{8}$

À partir des valeurs des $\hat{p}_{jk}$, on peut calculer le taux d'erreur de classification, l'indice de Gini et l'entropie croisée.

Notons E_j le taux d'erreur de classification dans la région R_j.

G_j l'indice de Gini dans la région R_j.

D_j l'entropie croisée dans la région R_j.

Pour tous les arbres, on a :

$$E_0 = 1 - \max_k \hat{p}_{0k} = \frac{1}{2}$$

$$G_0 = \sum_{k=0}^{\hat{1}} \hat{p}_{0k}(1 - \hat{p}_{0k}) = \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = \frac{1}{2}$$

$$D_0 = -\sum_{k=0}^{\hat{1}} \hat{p}_{0k} \log(\hat{p}_{0k}) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = \log 2$$

$E_1 = 1 - \frac{1}{2} = \frac{1}{2}$	$E_1 = 1 - 1 = 0$	$E_1 = 1 - \frac{3}{4} = \frac{1}{4}$
$E_2 = 1 - \frac{1}{2} = \frac{1}{2}$	$E_2 = 1 - 1 = 0$	$E_2 = 1 - \frac{7}{8} = \frac{1}{8}$
$G_1 = \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = \frac{1}{2}$	$G_1 = 0 \times (1-0) + 1 \times (1-1) = 0$	$G_1 = \frac{3}{4} \times \frac{1}{4} + \frac{1}{4} \times \frac{3}{4} = \frac{3}{8}$
$G_2 = \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = \frac{1}{2}$	$G_2 = 1 \times (1-1) + 0 \times (1-0) = 0$	$G_2 = \frac{1}{8} \times \frac{7}{8} + \frac{7}{8} \times \frac{1}{8} = \frac{7}{32}$
$D_1 = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = \log 2$	$D_1 =$ indéfini $(0 \times \infty)$	$D_1 = -\frac{3}{4}\log\left(\frac{3}{4}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right)$
$D_2 = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = \log 2$	$D_2 =$ indéfini $(0 \times \infty)$	$D_2 = -\frac{1}{8}\log\left(\frac{1}{8}\right) - \frac{7}{8}\log\left(\frac{7}{8}\right)$.

Pour choisir la variable suivant laquelle faire le 1er embranchement, on calcule le gain d'information : ΔE, ΔG ou ΔD.

Par exemple, avec l'indice de Gini :

$\Delta G = G_0 - \left(\frac{n_1}{n} G_1 + \frac{n_2}{n} G_2\right)$	$\Delta G = G_0 - \left(\frac{n_1}{n} G_1 + \frac{n_2}{n} G_2\right)$	$\Delta G = G_0 - \left(\frac{n_1}{n} G_1 + \frac{n_2}{n} G_2\right)$
$= \frac{1}{2} - \left(\frac{10}{20} \times \frac{1}{2} + \frac{10}{20} \times \frac{1}{2}\right)$	$= \frac{1}{2} - \left(\frac{10}{20} \times 0 + \frac{10}{20} \times 0\right)$	$= \frac{1}{2} - \left(\frac{12}{20} \times \frac{3}{8} + \frac{8}{20} \times \frac{7}{32}\right)$
$= 0 \rightarrow$ gain nul	$= \frac{1}{2} \rightarrow$ gain maximal.	$= 0.1875$

On choisit donc la variable qui maximise ΔG : ici la 2^{e} variable.

Notons $\hat{Y}$ la prédiction de Y à l'aide de l'arbre obtenu.

Le coût de complexité est $\frac{1}{m}\sum_{j=1}^{|T|} \mathbb{1}\{\hat{Y} \neq Y\} + \alpha |T|$.

Ici $|T| = 2$ (nb de feuille de l'arbre)

$\sum_{j=1}^{|T|} \mathbb{1}\{\hat{Y} \neq Y\} = 0$ (toutes les obs sont bien classées)

Donc $\frac{1}{m}\sum_{j=1}^{|T|} \mathbb{1}\{\hat{Y} \neq Y\} + \alpha |T| = 1$.

Un cas plus intéressant serait de prendre le 3^{e} arbre.

À partir du tableau croisé, les données pourrait ressemblées à ça :

X_3	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	0
Y	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
$\hat{Y}$	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1

Arbre : 20 obs

- $X_3 = 1$: 12 obs — $3 \to Y = 1$, $9 \to Y = 0$
- $X_3 = 0$: 8 obs — $7 \to Y = 1$, $1 \to Y = 0$

Prédiction : $\hat{Y} = 0$ (pour $X_3 = 1$), $\hat{Y} = 1$ (pour $X_3 = 0$)

On utilise la classe majoritaire dans une feuille pour faire la prédiction.

Sur les 20 observations, il y en aura donc 4 de mal classés, donc $\sum_{j=1}^{|T|} \mathbb{1}\{\hat{Y} \neq Y\} = 4$

et donc $\frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{\hat{Y} \neq Y\} + \alpha |T| = \frac{4}{20} + 0.5 \times 2 = \frac{6}{5}$