

# TD : Supervisée

## 1 Exercice 1 : Compréhension de l'analyse discriminante

On effectue un analyse discriminante sur un jeu de données pour lequel on tente de prédire auquel de trois groupes les 80 observations appartiennent à partir de la valeur des variables  $X_1, \dots, X_5$ . Les valeurs moyennes des 5 variables dans chacun des trois groupes sont données dans le tableau qui suit.

| $k$ | $\bar{X}_{k1}$ | $\bar{X}_{k2}$ | $\bar{X}_{k2}$ | $\bar{X}_{k3}$ | $\bar{X}_{k4}$ |
|-----|----------------|----------------|----------------|----------------|----------------|
| 1   | -0.155         | 0.202          | -0.003         | -0.909         | -0.952         |
| 2   | 0.227          | 0.050          | 0.091          | 0.067          | 0.084          |
| 3   | 0.505          | -0.313         | -0.092         | 0.475          | 0.776          |

On obtient les matrices de sommes des carrées suivantes, desquelles trois valeurs ont été masquées et remplacées par AAAAA, BBBB et CCCCC.

$$S = \begin{pmatrix} 62.8029695 & \text{AAAAA} & -4.471622 & -0.3359447 & 10.369236 \\ \text{AAAAA} & 81.038586 & 26.437970 & -4.6206012 & 7.087506 \\ -4.4716215 & 26.437970 & 80.626818 & 12.6439431 & -6.670094 \\ -0.3359447 & -4.620601 & 12.643943 & 85.5572762 & 30.720944 \\ 10.3692360 & 7.087506 & -6.670094 & 30.7209436 & 88.442644 \end{pmatrix},$$

$$W = \begin{pmatrix} 56.401092 & 10.81546 & -3.671480 & -13.814659 & -6.382273 \\ 10.815459 & 76.90362 & 25.490769 & 5.371660 & 19.919442 \\ -3.671480 & 25.490770 & 80.197630 & 13.933547 & -4.655554 \\ -13.814659 & 5.371660 & 13.933547 & 56.704695 & \text{BBBBB} \\ -6.382273 & 19.919440 & -4.655554 & \text{BBBBB} & 44.590677 \end{pmatrix}$$

et

$$B = \begin{pmatrix} 6.4018771 & -4.934314 & -0.8001411 & 13.478714 & 16.751510 \\ -4.9343145 & 4.134966 & 0.9472010 & CCCCC & -12.83194 \\ -0.8001411 & 0.947201 & 0.4291877 & -1.289604 & -2.01454 \\ 13.4787145 & CCCCC & -1.2896044 & 28.852581 & 35.36415 \\ 16.7515092 & -12.831936 & -2.0145400 & 35.364146 & 43.85197 \end{pmatrix}.$$

De plus, on vous dit que les deux plus grandes valeurs propres de la matrice  $S^{-1}B$  sont 0.751 et 0.011 et que les vecteurs propres normés correspondants sont

$$a_1 = \begin{pmatrix} -0.40 \\ 0.37 \\ -0.09 \\ -0.44 \\ -0.71 \end{pmatrix}, \quad \text{et} \quad a_2 = \begin{pmatrix} 0.08 \\ -0.65 \\ -0.28 \\ -0.64 \\ 0.29 \end{pmatrix}$$

1. Quelles sont les valeurs de AAAAA, BBBBB et CCCCC ?
2. Quel est le pouvoir discriminant de la fonction discriminante de Fisher ?
3. Selon cette fonction discriminante, dans lequel des trois groupes classeriez-vous une observation dont les valeurs de  $(X_1, X_2, X_3, X_4, X_5)$  sont respectivement  $(-0.5, 0.5, 0, 1, 1)$  ?

## 2 Exercice 2 : Construction d'un arbre

On cherche à prédire  $Y$  à partir des variables  $X_1$ ,  $X_2$  et  $X_3$ . Toutes les variables sont binaires. On souhaite construire un arbre à partir des données, et en particulier choisir le premier embranchement. Voici les tableaux croisés entre  $Y$  et chacune des trois variables.

|           | $Y = 1$ | $Y = 0$ |
|-----------|---------|---------|
| $X_1 = 1$ | 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       |

|           | $Y = 1$ | $Y = 0$ |
|-----------|---------|---------|
| $X_3 = 0$ | 7       | 1       |

1. Calculer le taux d'erreur obtenu en créant un embranchement à partir de chacune des trois variables explicatives.
2. Calculer l'indice de Gini obtenu en créant un embranchement à partir de chacune des trois variables explicatives.
3. Calculer l'entropie obtenu en créant un embranchement à partir de chacune des trois variables explicatives.
4. Quelle variable choisir pour faire le premier embranchement ?
5. Calculer le coût de complexité de l'arbre obtenu, en utilisant le taux d'erreur comme mesure du coût (prendre  $\alpha = 0.5$ ).