Examen d'analyse de données - Durée 1h30

Les documents de cours (transparents de cours, sujets de TP, TD, CTD et notes manuscrites) sont autorisés. Les trois exercices sont indépendants.

Exercice 1 : Classification bayésienne : 7.5 points

Soient deux classes C_1 et C_2 équiprobables dans $\mathbb{R}^2$, les observations de ces deux classes suivent une loi Gaussienne avec pour paramètres respectifs : $\mathbf{m_1}$ et $\mathbf{m_2}$ (vecteurs moyennes) et Σ_1 et Σ_2 (matrices de variance-covariance).

On rappelle l'expression de la fonction de densité de probabilité conditionnelle multivariée pour une classe C_i (ici, i vaut 1 ou 2), appelée vraisemblance, dans $\mathbb{R}^d$ (ici $d = 2$) :

$$p(\mathbf{x}|C_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\Big(-\frac{1}{2}(\mathbf{x}-\mathbf{m_i})^t\,\Sigma_i^{-1}\,(\mathbf{x}-\mathbf{m_i})\Big)$$

Questions

1. En posant $\mathbf{x} = [x_1, x_2]^t, m_1 = [m_{11}, m_{12}]^t, m_2 = [m_{21}, m_{22}]^t, \Sigma^{-1} = \begin{bmatrix} a & 0 \\ 0 & d \end{bmatrix}$ et en considérant que **les matrices de covariances sont égales :** $\Sigma_1 = \Sigma_2 = \Sigma$ **et lorsque les classes sont équiprobables**, montrer que la frontière de décision entre les 2 classes C_1 et C_2 est donnée par l'équation suivante :
$$(m_2 - m_1)^t\Sigma^{-1}\mathbf{x} + C = 0$$
où C est une constante.
Donner l'expression de C en fonction de $\mathbf{m_i}$ et Σ.

On considère le jeu de données suivant dont les moyennes, matrices de covariance et probabilités *a priori* de chaque classe sont : $m_1 = [0, 2]^t, \Sigma_1 = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}, P(C_1) = 0.5, m_2 = [0, 0]^t, \Sigma_2 = \Sigma_1$, $P(C_2) = 0.5$.

2. Soient les deux points $p_1 = [3, -2]^t, p_2 = [3, 2]^t$ à classer. Pour chacun de ces points, calculer les valeurs des log-vraisemblances pour les deux classes et donner la classe attribuée.

3. Calculer l'équation de la frontière de décision. Commenter.

Correction :

1. En prenant le logarithme de la vraisemblance, on trouve :
$$(x - m_1)^t\Sigma^{-1}(x - m_1) = (x - m_2)^t\Sigma^{-1}(x - m_2)$$
En développant on trouve,
$$a(x_1 - m_{11})^2 + d(x_2 - m_{12})^2 = a(x_1 - m_{21})^2 + d(x_2 - m_{22})^2$$
$$\begin{aligned} a[x_1^2 - 2x_1m_{11} + m_{11}^2 - x_1^2 + 2x_1m_{21} + m_{21}^2] + d[x_2^2 - 2x_2m_{22} + m_{22}^2 - x_2^2 + 2x_1m_{12} + m_{12}^2] &= 0 \\ a[2x_1(m_{11} - m_{21}) + m_{21}^2 - m_{11}^2] + d[2x_2(m_{12} - m_{22}) + m_{22}^2 - m_{12}^2] &= 0 \\ 2[m_{11} - m_{21}\ \ m_{12} - m_{22}]\begin{bmatrix} a & 0 \\ 0 & d \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + a[(m_{21}^2 - m_{11}^2) + d(m_{22}^2 - m_{12}^2)] &= 0 \end{aligned}$$

Donc par identification, $C = \frac{a}{2}(m_{21}^2 - m_{11}^2) + \frac{d}{2}(m_{22}^2 - m_{12}^2)$.
(Ronan $\rightarrow$ Sandrine : je crois que c'est $C = \frac{a}{2}(m_{11}^2 - m_{21}^2) + \frac{d}{2}(m_{12}^2 - m_{22}^2)$)
Sous forme vectorisée, $C = \frac{1}{2}[m_1 - m_2]^t \Sigma^{-1} [m_1 + m_2]$.

Pas de terme quadratique dans cette expression donc la frontière de décision est une droite.

Autre expression pour C : en prenant le logarithme de la vraisemblance, on trouve :

$$(x - m_1)^t \Sigma^{-1} (x - m_1) = (x - m_2)^t \Sigma^{-1} (x - m_2)$$

En développant on trouve,

$$x^t \Sigma^{-1} x - 2m_1^t \Sigma^{-1} x + m_1^t \Sigma^{-1} m_1 = x^t \Sigma^{-1} x - 2m_2^t \Sigma^{-1} x + m_2^t \Sigma^{-1} m_2$$

et

$$2(m_2 - m_1)^t \Sigma^{-1} x + m_1^t \Sigma^{-1} m_1 - m_2^t \Sigma^{-1} m_2 = 0$$

d'où

$$C = \frac{1}{2}(m_1^t \Sigma^{-1} m_1 - m_2^t \Sigma^{-1} m_2).$$

2. Pour le point p_1,

$$p(p_1|C_1) = -log(2\pi) - \frac{1}{2}log(4) - \frac{1}{2}[3 \ -4] \begin{bmatrix} \frac{1}{4} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ -4 \end{bmatrix} = -log(2\pi) - \frac{1}{2}log(4) - \frac{1}{2}[\frac{9}{4} + 16]$$

De même pour la classe C_2,

$$p(p_1|C_2) = -log(2\pi) - \frac{1}{2}log(4) - \frac{1}{2}[3 \ -2] \begin{bmatrix} \frac{1}{4} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ -2 \end{bmatrix} = -log(2\pi) - \frac{1}{2}log(4) - \frac{1}{2}[\frac{9}{4} + 4]$$

Donc d'après la règle de Bayes avec équiprobabilité des classes, $p(p_1|C_1) < p(p_1|C_2)$ donc p_1 est classé C_2.

Pour le point p_2,

$$p(p_2|C_1) = -log(2\pi) - \frac{1}{2}log(4) - \frac{1}{2}[3 \ 0] \begin{bmatrix} \frac{1}{4} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 0 \end{bmatrix} = -log(2\pi) - \frac{1}{2}log(4) - \frac{1}{2}[\frac{9}{4}]$$

De même pour la classe C_2,

$$p(p_2|C_2) = -log(2\pi) - \frac{1}{2}log(4) - \frac{1}{2}[3 \ 2] \begin{bmatrix} \frac{1}{4} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} = -log(2\pi) - \frac{1}{2}log(4) - \frac{1}{2}[\frac{9}{4} + 4]$$

Donc d'après la règle de Bayes avec équiprobabilité des classes, $p(p_2|C_2) < p(p_2|C_1)$ donc p_2 est classé C_1.

3. En reprenant la question 1, on obtient :

$$[0 \ -2] \begin{bmatrix} \frac{1}{4} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \frac{1}{2}[0 \ 2] \begin{bmatrix} \frac{1}{4} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 2 \end{bmatrix} = 0$$

Donc $-2x_2 + \frac{1}{2}4 = 0$, l'équation de la frontière est une droite horizontale d'équation $x_2 = 1$ (pas besoin de x_1 pour faire la classification), médiatrice du segment $[m_1, m_2]$.

Exercice 2 : Modélisation d'une réaction chimique par la méthode des moindres carrés : 5 points

Dans une réaction chimique, on souhaite modéliser l'évolution de la concentration d'un réactif en fonction du temps. On a mesuré expérimentalement :

Temps (s)	7	18	27	56
Concentration	32	28	25	18

Dans la suite, on notera $(T_i)_{1\leq i\leq 4}$ la suite des temps considérés et $(C_i)_{1\leq i\leq 4}$ la suite des concentrations mesurées.
Nous souhaitons effectuer une modélisation de la réaction par une réaction chimique à l'ordre 1, c'est-à-dire, si $C(t)$ désigne la concentration en fonction du temps :

$$\frac{dC(t)}{dt} = -\lambda C(t) \tag{1}$$

pour λ représente la constante de réaction. Elle admet pour solution $C(t) = C_0 e^{-\lambda t}$ où C_0 représente la concentration initiale. On souhaite estimer les paramètres réels C_0 et λ.

Questions

1. Justifier que $\lambda > 0$.
2. Ecrivez matriciellement le problème aux moindres carrés linéaire à résoudre (MCO) permettant d'estimer les paramètres (C_0, λ), c'est-à-dire définissez $\beta \in \mathbb{R}^2$, $A \in \mathbb{R}^{4\times 2}$ et $B \in \mathbb{R}^4$ tels que $\hat{\beta}_{OLS}$ soit la solution du problème suivant :
$$\min_{\beta\in\mathbb{R}^2} \|A\beta - B\|^2.$$
3. Donner la solution analytique de ce problème (on ne demande pas de calculer la solution numérique).

Correction

1. Il s'agit d'une équation différentielle du 1er ordre dont la solution est $C(t) = C_0 exp(-\lambda t)$. La fonction exponentielle est strictement croissante. Or la concentration mesurée décroît strictement en fonction du temps. Donc il faut prendre $\lambda > 0$.
2. On passe au logarithme la solution et on obtient ainsi :
$$\forall i \in \{1, .., 4\}, ln(C_i) = ln(C_0) - \lambda T_i$$
que l'on peut écrire sous forme matricielle :
$$\begin{bmatrix} \ln(C_1) \\ \vdots \\ \ln(C_4) \end{bmatrix} = \begin{bmatrix} -T_1 & 1 \\ \vdots & \vdots \\ -T_4 & 1 \end{bmatrix} \begin{bmatrix} \lambda \\ \ln C_0 \end{bmatrix}$$
donc on peut se ramener à un problème aux moindres carrés linéaires, en posant $\beta = \begin{bmatrix} \lambda \\ \ln(C_0) \end{bmatrix}$
$$\min_{\beta\in\mathbb{R}^2} \|A\beta - b\|^2$$
avec $A = \begin{bmatrix} -7 & 1 \\ -18 & 1 \\ -27 & 1 \\ -56 & 1 \end{bmatrix}$ et $b = \begin{bmatrix} \ln(32) \\ \ln(28) \\ \ln(25) \\ \ln(18) \end{bmatrix}$
3. $$\hat{\beta} = (A^T A)^{-1} A^T b = A^+ b$$

Exercice 3 : Rugby ! - 7.5 points

On cherche à construire un arbre de décision permettant de décider si une équipe de rugby (par exemple, le Stade Toulousain) va gagner ou perdre le prochain match. Une base d'apprentissage a été construite en considérant les données suivantes qui récapitulent les conditions qui accompagnent les succès et les échecs de cette équipe de rugby.

Match à domicile	Ciel	Match précédent gagné ?	Match gagné ?
oui	Soleil	oui	oui
oui	Pluie	non	non
oui	Soleil	non	oui
non	Couvert	oui	oui
non	Pluie	oui	oui
non	Soleil	non	non

Questions

1. Déterminer l'indice de Gini associé à cette base d'apprentissage vis-à-vis des deux classes "Match gagné" et "Match perdu". **2 points**

2. Déterminer la variation de l'indice de Gini lorsqu'on coupe les données à l'aide des variables "Match à domicile", "Ciel" et "Match précédent gagné ?" (**1.5 point par variable**). En déduire la variable qui sera utilisée au premier niveau de l'arbre de décision. (**1 point**)

Correction

1. Indice de Gini de la base (ici c'est "Match gagné ?"), $i \in \{1, 2\}$ pour $\{oui,\ non\}$
n=nbre d'occurences totales (ici $n = 6$) et n_i= nbre d'occurences "oui" ou "non".

$$Gini(Jouer) = \sum_{i=1}^{2} \frac{n_i}{n}(1 - \frac{n_i}{n}) = 1 - \sum_{i=1}^{2}(\frac{n_i}{n})^2 = 1 - (\frac{4}{6})^2 - (\frac{2}{6})^2 = \frac{4}{9}.$$

2. (a) Indice de Gini de la variable "Ciel": 3 sous-ensembles
$\frac{n_{se}}{n} = p_i$=proportion du sous-ensemble dans la variable
   i. sous-ensemble "Soleil" : $i \in \{1, 2\}$ pour $\{oui, non\}$, $n_{se_S} = 3$:
$$Gini(Ciel = Soleil) = 1 - (\frac{2}{3})^2 - (\frac{1}{3})^2 = \frac{4}{9}$$
   ii. sous-ensemble "Couvert" : $i \in \{1, 2\}$ pour $\{oui, non\}$, $n_{se_C} = 1$:
$$Gini(Ciel = couvert) = 1 - (\frac{1}{1})^2 - (\frac{0}{1})^2 = 0$$
   iii. sous-ensemble "Pluie" : $n_{se_p} = 2$
$$Gini(Ciel = Pluie) = 1 - (\frac{1}{2})^2 - (\frac{2}{2})^2 = \frac{1}{2}$$
$$Gini(Ciel) = \frac{n_{se_S}}{n} Gini(Soleil) + \frac{n_{se_c}}{n} Gini(Couvert) + \frac{n_{se_p}}{n} Gini(pluie) = \frac{3}{6} * \frac{4}{9} + \frac{1}{6} * 0 + \frac{2}{6} * \frac{1}{2} = \frac{7}{18}$$
   (b) Indice de Gini de la variable "Match) domicile" : 2 sous-ensembles
   i. sous-ensemble "oui" : $i \in \{1, 2\}$ pour $\{oui, non\}$, $n_{se_C} = 3$:
$$Gini(Matchdomicile = oui) = 1 - (\frac{2}{3})^2 - (\frac{1}{3})^2 = \frac{4}{9}$$

ii. sous-ensemble "non" : $n_{se_f} = 3$

$$Gini(Matchdomicile = non) = 1 - (\frac{2}{3})^2 - (\frac{1}{3})^2 = \frac{4}{9}$$

$$Gini(Matchdomicile) = \frac{n_{se_C}}{n} Gini(oui) + \frac{n_{se_f}}{n} Gini(non) = \frac{3}{6} * \frac{4}{9} + \frac{3}{6} * \frac{4}{9} = \frac{4}{9}$$

(c) Indice de Gini de "Match précédent gagné ?" : 2 sous-ensembles

i. sous ensemble "oui" : $i \in \{1, 2\}$ pour $\{oui, non\}$, $n_{se_f} = 3$:

$$Gini(Match\ prec\ gagne = oui) = 1 - (\frac{3}{3})^2 - (\frac{0}{3})^2 = 0$$

ii. sous-ensemble "non" : $n_{se_F} = 3$

$$Gini(Match\ prec\ gagne = non) = 1 - (\frac{2}{3})^2 - (\frac{1}{3})^2 = \frac{4}{9}$$

$$Gini(Match\ prec\ gagne) = \frac{n_{se_f}}{n} Gini(oui) + \frac{n_{se_F}}{n} Gini(non) = \frac{3}{6} * 0 + \frac{3}{6} * \frac{4}{9} = \frac{2}{9}.$$

Pour connaître la première variable utilisée au premier niveau de l'arbre CART, on maximise le gain défini par :

$$Gain(Variable) = Gini(base) - Gini(variable)$$

(a) $Gain(ciel) = \frac{4}{9} - \frac{7}{18} = \frac{1}{18}$

(b) $Gain(Match\ domicile) = \frac{4}{9} - \frac{4}{9} = 0$

(c) $Gain(Match\ prec\ gagne) = \frac{4}{9} - \frac{2}{9} = \frac{2}{9}$

Le gain est maximal pour la variable "Match précédent gagné" qui sera utilisée au premier niveau de l'arbre.