Partiel Analyse de Données

Documents autorisés : planches de cours, sujets de TD/TP, notes MANUSCRITES PERSONNELLES de cours/TD (PAS de PHOTOCOPIES), pas de calculatrice.
Durée : 1h30 (+30 min tiers temps)

1 Questions de cours

1 point par question

1. Qu'appelle-t-on méthode de classification supervisée ?
   Une méthode de classification est dite supervisée lorsqu'on connaît le nombre de classes et qu'on a un ensemble de données fournies par l'expert appartenant à chacune de ces classes constituant la base d'apprentissage. Le classifieur est alors construit en utilisant les données de cette base d'apprentissage (dites étiquetées car on connaît la classe de chaque vecteur de données). Des exemples de classifieurs supervisés vus en cours sont le classifieur Bayésien, les machines à vecteurs supports, les réseaux de neurones ou les arbres de régression.
2. Expliquer le principe de la validation croisée.
   Lorsque certains paramètres d'un classifieur doivent être réglés par l'utilisateur (comme le nombre de plus proches voisins dans la règle des plus proches voisins, il est habituel de chercher les paramètres qui minimisent une fonction de coût comme la probabilité d'erreur du classifieur. Pour ce faire, on découpe la base d'apprentissage en S sous-ensembles de tailles égales, on construit le classifieur à l'aide de $S-1$ de ces sous-ensembles et on détermine la fonction de coût avec le dernier de ces sous-ensembles. On répète cette opération S fois, en utilisant à chaque fois un sous-ensemble différent, et on moyenne les fonctions de coût obtenues. La valeur du vecteur paramètre minimisant cette fonction de coût moyenne est retenue pour construire le classifieur.
3. Expliquer ce que sont les vecteurs supports dans la méthode de classification SVM (support vector machines).
   La méthode de classification basée sur les machines à vecteurs supports (classifieur SVM) recherche l'hyperplan séparateur maximisant la marge de la base d'apprentissage. Cette marge est égale à la distance minimale entre les points de la base d'apprentissage et l'hyperplan séparateur. Les vecteurs de la base d'apprentissage situés à une distance de l'hyperplan égale à la marge de la base d'apprentissage sont appelés les vecteurs supports. Un vecteur support $\boldsymbol{x}_i$ vérifie l'équation $y_i(\boldsymbol{w}^T\boldsymbol{x}_i - b) = 1$ pour un hyperplan d'équation $\boldsymbol{w}^T\boldsymbol{x} - b) = 0$ où y_i est l'étiquette indiquant la classe de $\boldsymbol{x}_i$ ($y_i = 1$ si $\boldsymbol{x}_i$ appartient à la classe ω_1 et $y_i = -1$ si $\boldsymbol{x}_i$ appartient à la classe ω_2).
4. Dans quelle situation est-il intéressant d'utiliser un noyau dans la méthode de classification SVM ?
   Lorsque les données des différentes classes ne sont pas linéairement séparables, on peut utiliser l'astuce du noyau qui consiste à appliquer une transformation non linéaire ϕ à chaque vecteur

de la base d'apprentissage $\boldsymbol{x}_i$. Le problème de classification découlant des machines à vecteurs supports ne dépendant que des produits scalaires $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$, quand on utilise ce prétraitement non-linéaire il suffit de connaître les produits scalaires entre les vecteurs $\phi(\boldsymbol{x}_i)$ et $\phi(\boldsymbol{x}_i)$ qui s'expriment à l'aide d'un noyau κ tel que $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle$.

5. Donner l'expression mathématique d'une fonction sigmoide dérivable en tout point utilisée dans les réseaux de neurones.
   L'exemple donné en cours est
   $$f(x) = \frac{1}{1 + e^{-\alpha x}}.$$

2 Barbecue !

Avec l'arrivée des beaux jours, le barbecue est un repas très apprécié en soirée. Nous essayons dans ce problème de modéliser les ventes de brochettes sur deux mois entre le 30 avril et le 30 juin dans un petit supermarché.

Tout d'abord, on constate qu'au 30 avril ($t = 0$), 10 barquettes de brochettes ont été vendues. Au 30 mai ($t = 1$), seule 1 barquette a été vendue. On décide de modéliser les ventes par la fonction f suivante :

$$f(t) = a_0 t^2 + b_0$$

avec (a_0, b_0) des réels et t exprimé en mois.

1. Résoudre le système linéaire permettant de satisfaire les ventes observées.
2. On remarque qu'au 30 juin ($t = 2$), 16 barquettes de brochettes ont été vendues. Calculer l'erreur aux moindres carrés réalisée par cette modélisation.

Comme ce modèle n'est pas optimal, on décide d'utiliser une autre fonction g pour modéliser les ventes :

$$g(t) = a\, t^3 + b\, t + c \tag{1}$$

où (a, b, c) sont des coefficients réels.

3. En posant $\beta = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$, écrivez matriciellement le problème d'estimation aux moindres carrés à résoudre à partir des données de l'énoncé à $t = \{0, 1, 2\}$, c'est-à-dire définissez $\mathbf{A} \in \mathbb{R}^{3\times 3}$ et $\mathbf{B} \in \mathbb{R}^3$ tels que le problème aux moindres carrés puisse s'écrire :

$$\min_{\beta \in \mathbb{R}^3} \frac{1}{2} \|\mathbf{A}\,\beta - \mathbf{B}\|^2 \tag{2}$$

4. Dans le cas général d'un problème aux moindres carrés où la matrice $\mathbf{A} \in \mathbb{R}^{m\times n}$, avec $m > n$, donnez la solution théorique du problème (**??**) sans la calculer explicitement.

1. Résolution du système linéaire :
$$\begin{cases} f(0) = b = 10 \\ f(1) = a + b = 1 \end{cases}$$
On obtient $a = -9$ et $b = 10$. Donc $f(t) = -9t^2 + 10$.
2 points
2. Avec la nouvelle donnée, $Err = (16 - f(2))^2 = (16 - (-36 + 10))^2 = 42^2$.
1 point
3. $g(t) = \begin{bmatrix} t^3 & t & 1 \end{bmatrix} \beta$ donc avec les données, on obtient : $A = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & 1 \\ 8 & 2 & 1 \end{bmatrix}$ et $B = \begin{bmatrix} 10 \\ 1 \\ 16 \end{bmatrix}$.
2 points
4. $\beta = (A^T A)^{-1}.A^T b$ ou $\beta = A^+.b$ en mentionnant que A^+ est la pseudo-inverse de A.
1 point

3 Après le barbecue, un peu de cardio

La Fréquence Cardiaque Maximum, notée FCM, est un paramètre essentiel pour permettre au coureur de fond d'élaborer des plans d'entraînement efficaces. Cette fréquence peut se mesurer, soit en laboratoire sur tapis roulant, soit sur le terrain à l'aide d'un cardio-fréquencemètre. Une première étude a été faite auprès de 5 hommes s'entraînant régulièrement (2 à 4 fois par semaine), et participant à de petites compétitions. On a mesuré leur fréquence cardiaque maximum. On souhaite étudier une relation éventuelle entre l'âge d'un individu et sa fréquence cardiaque maximum.

	Âge	FCM
Homme 1	26	178
Homme 2	28	176
Homme 3	30	182
Homme 4	32	180
Homme 5	34	184

1. Calculer la matrice de variance-covariance Σ. Y a-t-il une dépendance entre l'âge et la fréquence cardiaque maximum ? Expliquer votre réponse.
2. Calculer les coordonnées du premier axe principal de ce jeu de données.
3. Représenter sur un même graphe les données, l'axe principal et les composantes principales.
4. Pouvez vous prédire la fréquence cardiaque maximum d'une personne de 38 ans ? Expliquer votre démarche et donner une valeur.

1. — $X = \begin{bmatrix} 26 & 178 \\ 28 & 176 \\ 30 & 182 \\ 32 & 180 \\ 34 & 184 \end{bmatrix}$
0.5 point

- individu moyen : $g = [30\ 180]^T$.
  0.5 point
- $X_c = \begin{bmatrix} -4 & -2 \\ -2 & -4 \\ 0 & 2 \\ 2 & 0 \\ 4 & 4 \end{bmatrix}$.
  0.5 point
- $\Sigma = \frac{1}{5}\begin{bmatrix} 40 & 32 \\ 32 & 40 \end{bmatrix}$
  1 point
- Donc la covariance est égale à $\frac{32}{5}$, ce qui implique que les deux variables sont dépendantes.
  0.5 point

2. Calcul des composantes principales $det(5\Sigma - \lambda I_5) = (40-\lambda)^2 - 32^2 = (8-\lambda)(72-\lambda)$ Donc $\lambda_1 = 72$ et $\lambda_2 = 8$. Soit $X_1 = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ vecteur propre associé à $\lambda_1 = 72$ tel que $\Sigma X_1 = 72 X_1$. On obtient $X_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ De même $X_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ vecteur propre associé à $\lambda_2 = 8$.
   2 points pour les valeurs propres + 1 point pour le vecteur
3. graphe
   2 points si complet
4. Prédiction de la fréquence cardiaque max pour une personne de 38 ans : y=38+150=188
   1 point