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)

Questions de cours

1. (1pt) Expliquer ce qu'est une matrice de confusion pour un classifieur travaillant en mode supervisé. Donner un exemple pour quatre classes en expliquant les différents termes de la matrice.
En mode supervisé, pour chaque vecteur de la base d'apprentissage, on a un label indiquant la classe de ce vecteur. La matrice de confusion indique le nombre ou le pourcentage de labels prédits par le classifieur pour chaque classe. Un exemple de matrice de confusion est donné ci-dessous

Classes	Ocean	Desert	Forest	Ice
Ocean	**100**	0.0	0.0	0.0
Desert	0.0	**96.0**	5.4	0.0
Forest	0.0	4.0	**92.8**	0.8
Ice	0.0	0.0	1.8	**99.2**

Il indique que 100% des vecteurs de la classe "Ocean" ont été prédits comme appartenant à la classe "Ocean", tandis que 96% des vecteurs de la classe "desert" ont été affectés à la classe "desert" et les 4% restants ont été affecté à la classe "Forest".

2. (2pt) Expliquer le principe du classifieur Bayésien. Que doit-on connaître pour mettre en oeuvre ce classifieur ? Expliquer avec soin votre réponse.
Le classifieur Bayésien affecte un vecteur $\boldsymbol{x}$ à la classe la plus probable a posteriori, c'est-à-dire à la classe ω_i telle que $P(\omega_i|\boldsymbol{x}) \geq P(\omega_k|\boldsymbol{x}), \forall k$. Pour mettre en oeuvre ce classifieur, il faut connaître les densités de probabilité de $\boldsymbol{x}$ conditionnellement à chaque classe, i.e., $p(\boldsymbol{x}|\omega_i)$ et les probabilité a priori de chaque classe $P(\omega_i)$, ce qui permet de calculer

$$P(\omega_i|\boldsymbol{x}) = \frac{p(\omega_i|\boldsymbol{x})P(\omega_i)}{p(\boldsymbol{x})}.$$

3. (1pt) Expliquer le principe de l'algorithme des k moyennes (k-means).
Cet algorithme suppose qu'on connait le nombre de classes K et qu'on a des représentants de

chaque classe notés initialement $\boldsymbol{g}_1$, $\boldsymbol{g}_2$, ..., $\boldsymbol{g}_K$. L'algorithme alterne des étapes de classification et de mise à jour des représentants. On commence par initialiser les barycentres des classes par leurs représentants et on affecte chaque point $\boldsymbol{x}_i$ à la classe ω_k à l'aide de la règle de la distance aux barycentres ($\boldsymbol{x}_i$ est affecté à la classe ω_k qui minimise $d^2(\boldsymbol{x}_i, \boldsymbol{g}_k)$). Après cette étape de classification, on met à jour les représentants en les remplaçant par les moyennes des vecteurs des différentes classes, et on recommences les deux étapes de classification et de mise à jour jusqu'à convergence (la convergence est atteinte lorsque les barycentres ne changent plus).

4. (1pt) Dans un problème de classification à deux classes ω_1 et ω_2, quelle est la valeur de l'indice de Gini d'un ensemble possédant 4 éléments de ω_1 et 6 éléments de ω_2 ? Comment utilise-t-on cet indice pour scinder une branche d'un arbre de classification en deux parties ?
Pour un problème de classification à deux classes ω_1 et ω_2, l'indice de Gini d'un ensemble $\mathcal{E}$ est défini par $G(\mathcal{E}) = 2p(1-p)$, où p est la proportion d'éléments d'une des deux classes. Dans l'exemple donné, $G(\mathcal{E}) = 2 \times \frac{4}{10} \times \frac{6}{10} = \frac{12}{25}$. Pour scinder une branche de l'arbre de classification en deux parties, on cherche la séparation qui rend les ensembles des deux branches de l'arbre les plus homogènes possibles, c'est-à-dire qui minimise l'indice de Gini après séparation défini par $P_L G_L + P_R G_R$, où G_L et G_R sont les indices de Gini des ensembles de gauche et de droite après séparation et P_L et P_R sont les proportions de ces ensembles.

Exercice 2 : Moindres carrés

La quantification de l'abondance d'une espèce dans son habitat est un problème essentiel en écologie. On considère généralement qu'une relation lie le nombre d'individus N à la surface S de l'habitat en question par la formule suivante :

$$N = aS^b$$

On dispose de k couples d'observations (N_i, S_i), obtenus en comptant les représentants d'espèces de fleurs dans des prairies des Pyrénées, et en mesurant la surface de ces prairies. On voudrait ainsi estimer la valeur des paramètres a et b à partir de ces observations.

1. (2pt) Formulez le problème d'estimation des paramètres a et b au sens des moindres carrés et donnez-en la formulation matricielle en explicitant les matrices $\boldsymbol{A}$ et $\boldsymbol{B}$ associées.
On pose $a' = \ln(a)$. La relation $N = aS^b$ peut se réécrire $\ln(N) = \ln(a) + b\ \ln(S)$, i.e. $\ln(N) = a' + b\ \ln(S)$. Le problème d'estimation de a et b au sens des moindres carrés s'écrit donc :

$$\min_{\boldsymbol{a'},\boldsymbol{b}\in\mathbb{R}^2} \frac{1}{k} \sum_{i=1}^{k} \Big(\ln(N_i) - a' - b\ \ln(S_i) \Big)^2$$

ce que l'on peut réécrire

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

en posant $\boldsymbol{\beta} = [a', b]^T$ et

$$\boldsymbol{A} = \begin{bmatrix} 1 & \ln(S_1) \\ 1 & \ln(S_2) \\ \dots & \dots \\ 1 & \ln(S_k) \end{bmatrix} \text{ et } \boldsymbol{B} = \begin{bmatrix} \ln(N_1) \\ \ln(N_2) \\ \dots \\ \ln(N_k) \end{bmatrix}$$

2. (1pt) Expliquez comment estimer les paramètres a et b à partir des matrices $\boldsymbol{A}$ et $\boldsymbol{B}$.
   Si $\boldsymbol{A}^T\boldsymbol{A}$ est inversible alors
   $$\widehat{\boldsymbol{\beta}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{B}.$$
   On en déduit $\widehat{a} = \exp(\widehat{\boldsymbol{\beta}}_1)$ et $\widehat{b} = \widehat{\boldsymbol{\beta}}_2$.

Exercice 3 : ACP

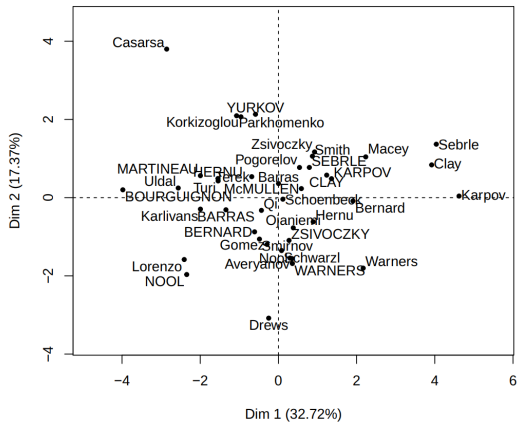
On dispose d'une base de données répertoriant les performances de plusieurs athlètes lors de deux épreuves : le décathlon des jeux olympiques d'Athènes (août 2004) et le Decastar organisé à Talence en Gironde (septembre 2004). Une partie des données est présentée dans le tableau ci-dessous :

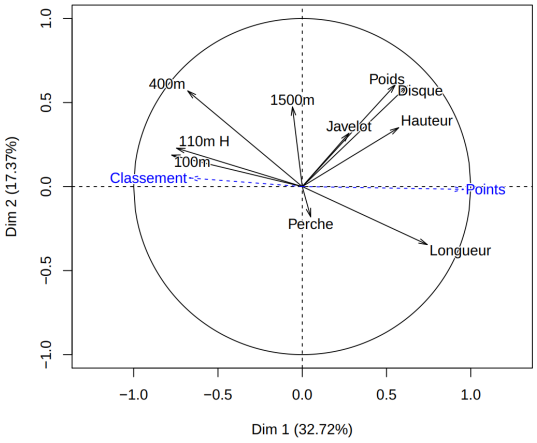
Athlète	100m	Longueur	Poids	Hauteur	400m	110m H	Disque	Perche	Javelot	1500m	Points
...											
A	10.87	7.38	13.07	1.88	48.51	14.01	40.11	5	51.53	274.21	7926
B	11.36	6.68	14.92	1.94	53.2	15.39	48.66	4.4	58.62	296.12	7404
C	10.5	7.81	15.93	2.09	46.81	13.97	51.65	4.6	55.54	278.11	8725
D	11.1	7.03	13.22	1.85	49.34	15.38	40.22	4.5	58.36	263.08	7592
...											

Ce tableau de données compte au total 41 lignes, correspondant aux performances de 41 athlètes (certains athlètes ont participé aux 2 épreuves et apparaissent donc 2 fois ; les participants au Decastar ont leur nom écrit en lettres majuscules). En plus des résultats aux 10 épreuves du décathlon, on dispose d'une colonne supplémentaire précisant le nombre de points obtenus par l'athlète lors de l'épreuve.

1. (1pt) Quelles sont les dimensions de la matrice $\boldsymbol{X}$ des données du problème, et de la matrice de variance-covariance $\boldsymbol{\Sigma}$ associée ?
   La problème compte 41 individus et 10 variables, la matrice $\boldsymbol{X}$ est donc de dimension 41×10.
   La matrice de variance-covariance $\boldsymbol{\Sigma}$ associée est de dimension 10×10.
2. (1pt) Expliquez pourquoi il est important de centrer-réduire les données de cette base.
   Les variables ont des unités différentes (mètres ou secondes) et d'ordres de grandeurs différents. Centrer-réduire les données permet de ne pas accorder arbitrairement trop d'importance à une variable dont l'ordre de grandeur serait plus grand que les autres.

Voici une représentation des données projetées sur les 2 premiers axes ainsi que le cercle de corrélation des variables :





3. (1pt) Certaines flèches du cercle de corrélation (par exemple, 100m et 110m H) pointent dans une direction très proche. Expliquez ce que cela signifie. Les flèches représentent les variables du problème. Le cosinus de l'angle entre deux flèches est égal à la corrélation entre les 2 variables correspondantes. Ainsi, les variables 100m et 110m H semblent fortement corrélées, ce qui signifie qu'un athlète qui performe bien au 100m tend à bien performer également au 110m H. Attention cependant, pour pouvoir conclure ceci avec certitude, il faut que les variables correspondantes soient bien représentées dans le cercle des corrélations, ce qu'on peut lire en regardant la longueur des flèches (plus elles sont proches du cercle, mieux les variables sont représentées). On peut ainsi affirmer avec plus de certitude que les variables Poids et Disque sont corrélées.

4. (1pt) La variable additionnelle "Points" a une corrélation de 0.92 avec l'axe 1 de l'ACP et de -0.03 avec l'axe 2. Représentez la variable sur le cercle des corrélations.
voir figure.

5. (1pt) D'après vous, quelle est l'information portée par l'axe 1 de l'ACP ?
L'axe 1 étant très fortement corrélé à la variable Points, on peut en conclure que cet axe représente la performance globale des athlètes lors du décathlon.

6. (1pt) Au vu du graphe des individus, diriez-vous que les meilleures performances ont été réalisées lors du Decastar ou des Jeux Olympiques (on rappelle que les données du Decastar sont représentées en lettres majuscules et que celles des jeux olympiques le sont en lettres minuscules) ? Justifiez votre réponse en vous appuyant sur le graphique.
L'axe 1 étant très fortement corrélé à la variable Points, on en déduit que les athlètes les plus à droite sur le graphe des individus sont ceux ayant le mieux performé. Considèrons des athlètes ayant participé (et bien performé) aux deux décathlons : Sebrle, Clay et Karpov. Ces 3 athlètes ont une valeur sur l'axe 1 supérieure à 4 pour leur performance des JO d'Athènes, contre à peine 2 pour leur Decastar. Cela signifie qu'ils ont beaucoup mieux performé aux Jeux Olympiques d'Athènes !

7. (1pt) Les athlètes A, B, C et D ont été anonymisés : il s'agissait de Lorenzo, Karpov, Casarsa et Drews. En vous appuyant sur le graphe des individus et le cercle des corrélations, associez chaque athlète à la ligne correspondante dans le tableau (A, B, C ou D).
D'après leur coordonnées sur l'axe 1, Karpov est celui qui a obtenu le plus de points au global, il est donc très certainement l'athlète C. Vient ensuite Drews, qui est donc probablement l'athlète

A. Pour les deux derniers, Casarsa doit être l'athlète avec le plus haut temps sur 1500m, il est donc très probablement l'athlète B. Drews est donc l'athlète D.

Exercice 4 : Machines à vecteurs supports

On considère un problème de classification supervisée à deux classes avec une base d'apprentissage $\mathcal{B} = \{(\boldsymbol{x}_i, y_i), i = 1, ..., n\}$ avec $\boldsymbol{x}_i \in \mathbb{R}^d$ et $y_i \in \{-1, +1\}$. Le classifieur SVM cherche une fonction de décision de la forme $f(\boldsymbol{x}) = \text{sign}\left[\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) + b\right]$ avec $b \in \mathbb{R}$. Le vecteur $\boldsymbol{w}$ s'obtient par résolution du problème

$$\left\{ \begin{array}{l} \min\limits_{\substack{\boldsymbol{w} \in \mathbb{R}^n, b \in \mathbb{R} \\ (\xi_1, ..., \xi_n) \in \mathbb{R}^n}} \left\{ \frac{1}{2} \|\boldsymbol{w}\|^2 + C \sum\limits_{i=1}^{n} \xi_i \right\} \\ \textbf{sous les contraintes} \left\{ \begin{array}{ll} y_i \left(\boldsymbol{w}^T \boldsymbol{x}_i - b\right) \geq 1 - \xi_i & \forall i \in \{1, \ldots, n\} \\ \xi_i \geq 0 & \forall i \in \{1, \ldots, n\} \end{array} \right. \end{array} \right.$$

1. Expliquer l'intérêt du terme $C \sum\limits_{i=1}^{n} \xi_i$.
   Lorsque la variable ξ_i est nulle, le vecteur $\boldsymbol{x}_i$ satisfait la contrainte $y_i \left(\boldsymbol{w}^T \boldsymbol{x}_i - b\right) \geq 1$. Lorsque $\xi_i > 0$, le vecteur $\boldsymbol{x}_i$ ne satisfait pas la contrainte. L'intérêt d'introduire le terme $C \sum\limits_{i=1}^{n} \xi_i$ dans le critère à minimiser est d'éviter d'avoir trop de vecteurs $\boldsymbol{x}_i$ qui ne vérifient pas la contrainte $y_i \left(\boldsymbol{w}^T \boldsymbol{x}_i - b\right) \geq 1$.
2. Comment choisit-on la valeur de C ?
   Lorsque C a une valeur faible, on autorise beaucoup de vecteurs à ne pas respecter la contrainte $y_i \left(\boldsymbol{w}^T \boldsymbol{x}_i - b\right) \geq 1$. Inversement, lorsque C a une forte valeur, beaucoup de vecteurs $\boldsymbol{x}_i$ satisfont la contrainte et donc la solution est proche de celle obtenue sans l'introduction des variables ξ_i. Pour choisir C, on peut donc procéder par validation croisée, c'est-à-dire tester plusieurs valeurs de C et garder la valeur qui donne les meilleurs résultats.
3. La solution du problème ci-dessus est $\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \phi(\boldsymbol{x}_i)$ où les réels α_i sont positifs ou nuls. Expliquer pour quels vecteurs on a $\alpha_i > 0$ et $\alpha_i = 0$.
   Les vecteurs $\boldsymbol{x}_i$ satisfaisant $\alpha_i > 0$ sont les vecteur supports. Ce sont les vecteurs qui respectent la contrainte $y_i \left(\boldsymbol{w}^T \boldsymbol{x}_i - b\right) = 1$, i.e., les vecteurs situés du bon coté de la frontière séparatrice et qui minimisent la distance à cette frontière. Les vecteurs $\boldsymbol{x}_i$ qui ne sont pas vecteurs supports vérifient $\alpha_i = 0$ et ne participent donc pas au calcul de $\boldsymbol{w}$.
4. Donner la règle de classification lorsqu'on utilise un noyau κ tel que $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle$.

   On affecte $\boldsymbol{x}$ à la classe $+1$ si

   $$\sum_{x_i \text{vecteur support}} \alpha_i y_i \kappa(\boldsymbol{x}_i, \boldsymbol{x}) - b \geq 0,$$

   avec $b = \frac{1}{2}\left(\boldsymbol{w}^T \boldsymbol{x}^+ + \boldsymbol{w}^T \boldsymbol{x}^-\right)$, où $\boldsymbol{x}^+$ (resp. $\boldsymbol{x}^-$) est un vecteur support de la classe $\{+1\}$ (resp. de la classe $\{-1\}$).

5. Donner un exemple de noyau vu en cours.
   En cours on a essentiellement vu le noyau gaussien défini par $k(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\sigma^2}\right)$ mais nous avons également évoqué le noyau polynomial défini par $k(\boldsymbol{x}, \boldsymbol{y}) = (\langle \boldsymbol{x}, \boldsymbol{y} \rangle + c)^q$ et le noyau de Mahalanobis défini par $k(\boldsymbol{x}, \boldsymbol{y}) = \exp\left[-(\boldsymbol{x}-\boldsymbol{y})^T \boldsymbol{\Sigma} (\boldsymbol{x}-\boldsymbol{y})\right]$.