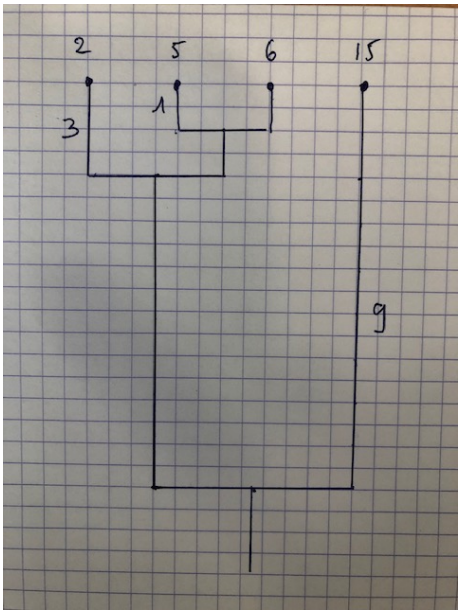
Partiel Analyse de Données

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. Puisque le classifieur Bayésien minimise la probabilité d'erreur de classification, dans quels cas peut-il être intéressant d'étudier d'autres classifieurs ?
   Le classifieur Bayésien nécessite de connaître les probabilités a priori des différentes classes et les densités de probabilité du vecteur d'observation conditionnellement à chaque classes. Lorsque ces quantités sont connues parfaitement, le classifieur Bayésien minimise la probabilité d'erreur de classification. En revanche, lorsque ces quantités sont inconnues ou partiellement connues, il est intéressant de considérer d'autres classifieurs comme les machines à vecteurs supports, les réseaux de neurones ou les arbres de régression.
2. On rappelle que la probabilité d'erreur de la règle du plus proche voisin notée P_1 vérifie l'inégalité suivante
   $$P^* \leq P_1 \leq P^* \left(2 - \frac{K}{K-1} P^*\right).$$
   Que désignent K et P^* dans cette inégalité ?
   K désigne le nombre de classes et P^* est la probabilité d'erreur du classifieur Bayésien.
3. Représenter l'arbre obtenu pour $\chi = \{2, 5, 6, 15\}$ avec la méthode de classification hiérarchique lorsqu'on utilise la distance entre groupes
   $$d(X_i, X_j) = \min_{x \in X_i, y \in X_j} d(x, y).$$
   Cet arbre est représenté ci-dessous.



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. On cherche à résoudre un problème de classification à 4 classes avec un réseau de neurones. Combien de noeuds de sortie choisiriez vous ? Quelle est la sortie désirée de ce réseau pour un élément de la première classe ?

   La solution la plus simple consiste à choisir autant de noeuds de sortie que de classes, i.e., ici 4 noeuds de sortie. La sortie désirée pour les éléments de la première classe est alors $(1, 0, 0, 0)$.

Exercice 1 : ACP et kppv

On compte les ordres de déplacements *pendule inversé, sauter* d'un drone par 5 utilisateurs. On obtient les données suivantes.

Utilisateur	Sauter	Pendule inversé
Ind. 1	0	2
Ind. 2	-2	-1
Ind. 3	1	0
Ind. 4	0	0
Ind. 5	1	-1

1. Ces ordres sont-ils corrélés ? Expliquer votre réponse.
2. Calculer le premier vecteur principal, de norme 1, de ces données.
3. Représenter sur un graphe, de la manière la plus précise, les données, l'axe principal et les données projetées.
4. Calculer les composantes principales 1D des données sur l'axe principal.
5. A partir des composantes principales 1D, calculer la matrice des distances euclidiennes entre les données projetées.
6. Appliquer, sur les composantes principales 1D, l'algorithme des k-plus proches voisins pour $k = 1$ en supposant que le seuil est égal à 1.1.

1. Calcul de Σ après avoir calculé X, l'individu moyen et X_c
   - $X = \begin{bmatrix} 0 & 2 \\ -2 & -1 \\ 1 & 0 \\ 0 & 0 \\ 1 & -1 \end{bmatrix}$
   - individu moyen : $\bar{x} = [0\ 0]^T$.
   - $X_c = X$
     0.5 point (X centré calculé)
   - $\Sigma = \frac{1}{5} X_c^T X_c = \frac{1}{5} \begin{bmatrix} 6 & 1 \\ 1 & 6 \end{bmatrix}$
     1 point (formule + résultat)
   - Donc la corrélation est $\frac{1}{5}$ et il y a une dépendance (faible).
     0.5 point
2. Calcul des composantes principales

   $det(5\Sigma - \lambda I_5) = (6-\lambda)^2 - 1^2 = (5-\lambda-1)(6-\lambda+1) = (5-\lambda)(7-\lambda)$

   Donc $\lambda_1 = \frac{7}{5}$ et $\lambda_2 = 1$.

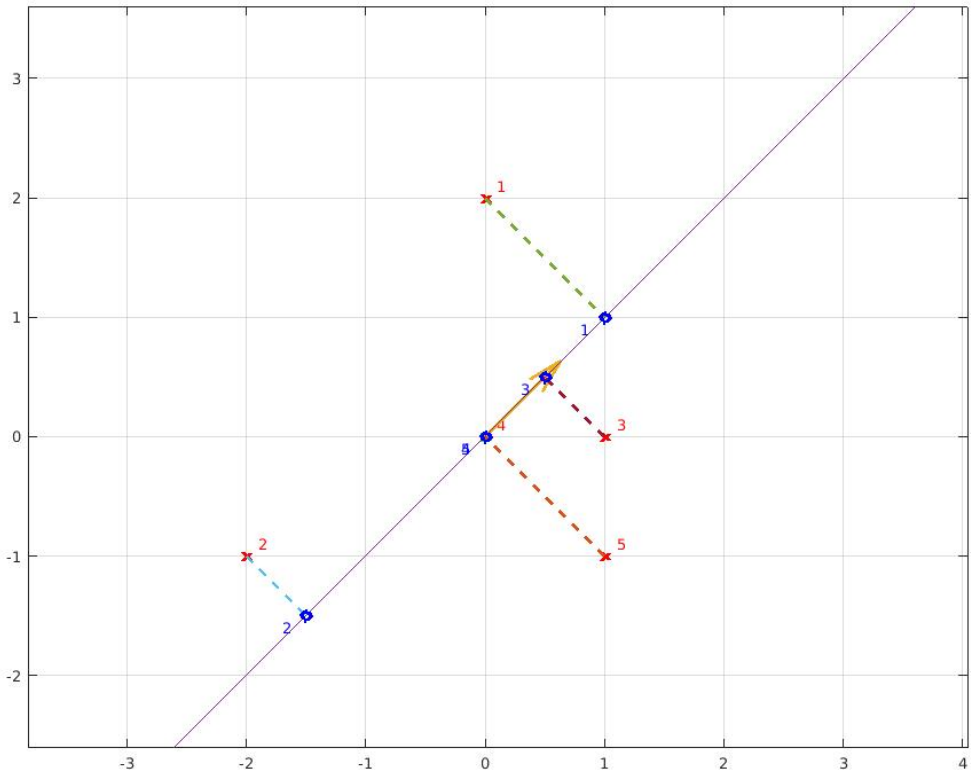
   Un vecteur propre associé à $\lambda_1 = \frac{7}{5}$ est tel que $\Sigma V = \frac{7}{5} V$.

On obtient un vecteur propre $V = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ qui vérifie cette équation

et si on le normalise, on pose $V_1 = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}$, vecteur principal

1 point pour les valeurs propres + 1 point pour le vecteur

3. graphe



1 point si complet

4. composantes principales

$$C_1 = X_c * V_1 = \begin{bmatrix} \sqrt{2} \\ -3\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \\ 0 \\ 0 \end{bmatrix}$$

1 point (formule + résultat)

5. tableau des distances

$$\begin{bmatrix} D & 1 & 2 & 3 & 4 & 5 \\ 1 & 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & \sqrt{2} & \sqrt{2} \\ 2 & & 0 & 2\sqrt{2} & 3\frac{\sqrt{2}}{2} & 3\frac{\sqrt{2}}{2} \\ 3 & & & 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ 4 & & & & 0 & 0 \\ 5 & & & & & 0 \end{bmatrix}$$

1 point

6. — 1, le point le plus proche est 3 et $D(1,3) = \frac{\sqrt{2}}{2} \approx 0.707 < 1.1$ donc il y a une classe $C_1 = \{1,3\}$;
   — 2, les point les + proches sont 4 et 5 avec $D(2,4) = D(2,5) = 3\frac{\sqrt{2}}{2} \approx 2.1 > 1.1$ donc il y a une seconde classe $C_2 = \{2\}$;
   — 3, les points les + proches sont 1, 4 et 5 à égale distance de 3 ($\frac{\sqrt{2}}{2} < 1.1$). Comme on fait l'algo du 1-ppv, si on considère le point de numéro le plus petit c'est 1 donc C_1 reste inchangé $C_1 = \{1,3\}$;
   — 4, le point le + proche est 3 donc C_1 devient $\{1,3,4\}$;
   — 5, le point le + proche est 4 donc C_1 devient $\{1,3,4,5\}$.
   On obtient deux classes $C_1 = \{1,3,4,5\}$ et $C_2 = \{2\}$.
   1 point (raisonnement (même si rajout des points 4 et 5 quand on examine 3 + résultat)

Exercice 2 : Soldes !

A l'approche des soldes, on considère les ventes de serviettes de plage chaque jour à partir de la dernière semaine de juin. Tout d'abord, on constate que le premier jour, soit le lundi 21 juin, seules deux serviettes ont été vendues. Au 4^e jour, 10 serviettes ont été vendues. On décide de modéliser les ventes par la fonction f suivante :

$$f(t) = a\sqrt{t} + bt$$

avec (a,b) des réels et $t > 0$ exprimé en jours.

1. Résoudre le système linéaire permettant de satisfaire les ventes observées.
   deux équations :

   $t_1 = 1, f(t_1) = y_1 = 2, a + b = 2$
   $t_2 = 4, f(t_2) = y_2 = 10, 2a + 4b = 10$

   regroupées en un système

   $$\begin{pmatrix} 1 & 1 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 2 \\ 10 \end{pmatrix}$$

   solution :

   $$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} -1 \\ 3 \end{pmatrix}$$

   2 points (système + solution)

2. On remarque que le 9^e jour, 18 serviettes ont été vendues. Calculer l'erreur aux moindres carrés réalisée par cette modélisation.

   $t_3 = 9, \quad \widetilde{y_3} = 18$

   $E = \sqrt{(y_1 - f(t_1))^2 + (y_2 - f(t_2))^2 + (\widetilde{y_3} - f(t_3))^2}$

   $E = \sqrt{0 + 0 + (18 - (-1 * \sqrt{9} + 3 * 9))^2}$

   $E = 6$

   1 point (6 ou 36)

Comme ce modèle n'est pas optimal, on décide de proposer la fonction g suivante pour modéliser les ventes :

$$g(t) = a\sqrt{t} + bt + c$$

avec (a, b, c) des réels.

3. En posant $\beta = [a \quad b \quad c]^T$, écrire sous forme matricielle le problème aux moindres carrés à résoudre à partir des données de l'énoncé c'est-à-dire définir $A \in \mathbb{R}^{3\times3}$ et $B \in \mathbb{R}^3$ tels que :

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

   En suivant la notation du CTD 2, on cherche C et D tels que

$$g(t, \beta) = C(t)\beta + D(t)$$

$$g(t, \beta) = \begin{bmatrix}\sqrt{t} & t & 1\end{bmatrix}\beta + 0$$

   donc $C(t) = \begin{bmatrix}\sqrt{t} & t & 1\end{bmatrix}$ et $D(t) = 0$

$$A = \begin{bmatrix}C(1)\\C(4)\\C(9)\end{bmatrix} = \begin{bmatrix}1 & 1 & 1\\2 & 4 & 1\\3 & 9 & 1\end{bmatrix}$$

   et

$$B = \begin{bmatrix}\widetilde{y_1} - D(t_1)\\\widetilde{y_2} - D(t_2)\\\widetilde{y_3} - D(t_3)\end{bmatrix} = \begin{bmatrix}2\\10\\18\end{bmatrix}$$

   2 points (construction + résultat)

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 (1) sans la calculer explicitement.

   $\beta = A^+B$ avec A^+, pseudo-inverse de A, $A^+ = (A^TA)^{-1}A^T$

   1 point (expression de B et de A^+)