INP ENSEEIHT

# 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 ?
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 ?
3. Représenter l'arbre obtenu pour $\chi = \{2, 5, 6, 15\}$ avec la méthode de classification hiérarchique lorsqu'on utilise la distance entre groupes
$$d(X_i, X_j) = \min_{x \in X_i, y \in X_j} d(x, y).$$
4. Dans quelle situation est-il intéressant d'utiliser un noyau dans la méthode de classification SVM ?
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 ?

## 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.

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

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 :

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

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.