Exercise 1 L'éclat (de rire de) moi hyène 7 points

L'algorithme des K-moyennes est une procédure qui vise à partitionner un ensemble de données en K clusters distincts (d'intersection nulle). Considérons $n \geqslant 1$ observations $(X_1, \ldots, X_n)$ à valeurs dans $\mathbb{R}^p$. L'algorithme des K-moyennes cherche à minimiser sur toutes les partitions $C = (C_1, \ldots, C_K)$ de $\{1, \ldots, n\}$ le critère suivant :

$$\mathrm{J}(C) = \sum_{k=1}^{K} \frac{1}{2|C_k|} \sum_{a,b \in C_k} \|X_a - X_b\|^2 ,$$

où pour tout $1 \leqslant i \leqslant n$, $1 \leqslant k \leqslant K$, $i \in C_k$ si et seulement si X_i est dans le cluster k.

1. Symétrisation
   (a) Montrez que

$$\mathrm{J}(C) = \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{a,b \in C_k} \langle X_a, X_a - X_b \rangle = \sum_{k=1}^{K} \sum_{a \in C_k} \|X_a - \bar{X}_{C_k}\|^2 ,$$

   où

$$\bar{X}_{C_k} = \frac{1}{|C_k|} \sum_{b \in C_k} X_b .$$

2. Supposons que les observations soient aléatoires et indépendantes et que, pour tout $1 \leqslant a \leqslant n$, $\mathbb{E}[X_a] = \mu_a \in \mathbb{R}^p$ de sorte que $X_a = \mu_a + \varepsilon_a$, avec $(\varepsilon_1, \ldots, \varepsilon_n)$ des variables aléatoires centrées et indépendantes. Pour tout $1 \leqslant a \leqslant n$, notons $v_a = \mathrm{trace}(cov(X_a))$.
   (a) Vérifiez que l'espérance du critère est

$$\mathbb{E}[J(C)] = \frac{1}{2} \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{a,b \in C_k} \left(\|\mu_a - \mu_b\|^2 + v_a + v_b\right) \mathbb{1}_{a \neq b} .$$

   (b) Que devient cette valeur de $\mathbb{E}[J(C)]$ quand, pout tout $1 \leqslant k \leqslant K$, il exist $m_k \in \mathbb{R}^p$ tel que, pour tout $a \in C_k$, $\mu_a = m_k$?
3. Supposons maintenant qu'il existe une partition $C^* = (C_1^*, \ldots, C_K^*)$ telle qu'il existe $m_1^*, \ldots, m_K^*$ in $\mathbb{R}^p$ et $\gamma_1^*, \ldots, \gamma_K^*$ dans $\mathbb{R}_+^*$ vérifiant $\mu_a = m_k^*$ et $v_a = \gamma_k^*$ pour tout $a \in C_k^*$ et $k = 1, \ldots, K$. Nous allons maintenant étudier dans quelles conditions la valeur attendue du critère des K-moyenne est minimum pour C^*.
   (a) Quelle est dans ce cas la valeur de $\mathbb{E}[J(C^*)]$?
   (b) Dans le cas particulier où $\gamma_1^* = \ldots = \gamma_K^* = \gamma$, quelle est la partition $C = (C_1, \ldots, C_K)$ qui minimise $\mathbb{E}[J(C)]$? Sous quelle(s) hypothèse(s) cette solution est-elle unique ?
   (c) Supposons maintenant que C^* contienne $K = 3$ groupes de taille s (avec s pair),

$$m_1 = (1, 0, 0)^T, \quad m_2 = (0, 1, 0)^T, \quad m_3 = (0, 1 - \tau, \sqrt{1 - (1 - \tau)^2})^T,$$

   avec $\tau > 0$, et

$$\gamma_1 = \gamma_+, \quad \gamma_2 = \gamma_3 = \gamma_- .$$

   Quelle est alors la valeur de $\|m_2 - m_3\|^2$?
   (d) Calculer $\mathbb{E}[J(C^*)]$.
   (e) Définir C' obtenu en divisant C_1^* en deux groupes C_1', C_2' de même taille $\frac{s}{2}$ et en fusionnant C_2^* et C_3^* dans un seul et même groupe C_3' de taille $2s$. Vérifier que, dans ce cas,

$$\mathbb{E}[J(C')] = s(\gamma_+ + 2\gamma_- + \tau) - (2\gamma_+ + \gamma_-) .$$

   (f) Sous quelle(s) hypothèse(s) $\mathbb{E}[J(C^*)] < \mathbb{E}[J(C')]$?

Exercise 2 — L'épreuve du réel — 8 points

Le but de cet exercice est d'écrire la fonction en python, Julia, R ou matlab permettant de prédire une sortie associée à un vecteur X de 64 composants. Pour ce faire, nous allons utiliser les données présentes dans le fichier `digits` disponible dans l'API sklearn qui contient :

— observations (entrées) X

— étiquettes y

Pour accéder à ces données, vous pourrez utiliser les instructions suivantes :

```python
from sklearn.datasets import load_digits

X,y = load_digits(n_class=10, return_X_y=True)
```

```
%whos

Variable      Type        Data/Info
-----------------------------------
X             ndarray     1797x64: 115008 elems, type `float64`, 920064 bytes
    (898.5 kb)
load_digits   function    <function load_digits at 0x120f98048>
y             ndarray     1797: 1797 elems, type `int64`, 14376 bytes
```

1. Traitement des données en mode non supervisé (en ignorant les étiquettes y)
   a) Visualisez l'ensemble des données.
   b) Proposez un découpage non supervisée des données en groupes homogènes.
2. Traitement des données en mode supervisé (en utilisant les étiquettes y). Donner un programme (en python, R ou matlab) permettant de prédire si une entrée future x_f, un vecteur de 64 composants représente un chiffre pair ou impair. On insistera sur la méthodologie mise en œuvre et sur la mesure des performances.

Exercise 3 — Questions courtes — 5 points

1. Que'est-ce que la malédiction de la dimensionnalité ?
2. Quel est le rôle de la programmation quadratique dans le Lasso ?
3. Quel est le rôle du stacking dans les méthodes de type autoML ?
4. Quel est le rôle du graphe de proximité en l'apprentissage non supervisé ?
5. Expliquez le principe d'une attaque en machine learning.
