Apprentissage non-supervisé.

Exercice 1 *Distances et clusterings*

Question 1. Montrer que la distance de Manhattan est bien une mesure de distance.

Question 2. Dans le cours, des approches ont été données pour calculer la distance entre 2 clusters. On considère une mesure de distance d entre 2 exemples et deux groupes d'exemples $A = \{a_1, a_2, ..., a_{|A|}\}$ et $B = \{b_1, b_2, ..., b_{|B|}\}$, avec pour tout $i = 1, ..., |A|$, $a_i \in \mathbb{R}^p$ et pour tout $j = 1, ..., |B|$, $b_j \in \mathbb{R}^p$. Donner l'expression de la distance D entre A et B pour l'approche par centre de gravité ("centroid linkage").

Question 3. On considère la base d'apprentissage $\mathcal{X}$ de $[0, 10] \times [0, 10]$ contenant les 7 exemples suivants : $\mathbf{X} = \{(1, 2), (1, 4), (3, 4), (3, 5), (6, 2), (6, 5), (8, 3)\}$ (remarque : on considère que cette base est déjà normalisée). En détaillant les étapes et en expliquant les calculs réalisés et les regroupements effectués, appliquer sur $\mathbf{X}$ l'algorithme de classification hiérarchique, version ascendante, en utilisant la distance euclidienne et l'approche "centroid linkage". Donner le dendrogramme obtenu et interpréter le résultat.

Question 4. Quelles sont les différences entre l'algorithme des k-moyennes et l'algorithme des k-médoïdes ?

Question 5. On se place dans un espace de dimension 2 où chaque exemple est représenté par un couple de valeurs réelles. On note d_1 la distance entre 2 groupes d'exemples calculée par l'approche *centroid linkage*, et d_2 la distance calculée par l'approche *complete linkage*. Donner, en justifiant votre réponse et en illustrant par une représentation graphique, un exemple de 3 clusters A, B et C contenant chacun exactement 4 exemples et tels que : $d_1(A, B) > d_1(A, C)$ **et** $d_2(A, B) < d_2(A, C)$.

Question 6. On décide d'appliquer l'algorithme des k-moyennes avec $k = 2$ sur $\mathcal{X}$. On choisit les exemples $A = (1, 4)$ et $E = (6, 2)$ pour initialiser les centres des clusters. Représenter graphiquement $\mathcal{X}$ et la frontière de séparation des 2 clusters induits par A et E.

Question 7. Donner les coordonnées des 2 centres des deux clusters induits par A et E.

Question 8. Donner l'inertie intra-cluster des clusters obtenus à la question précédente et en déduire l'inertie globale de la partition.

Question 9. Au bout de combien d'étapes l'algorithme des k-moyennes converge-t-il pour cette initialisation par A et E ? Et quel est le résultat obtenu ?

Question 10. Soit `A` et `B` deux array numpy contenant des exemples d'apprentissage et possédant le même nombre de colonnes. Donner le code python **efficace** permettant de calculer la distance entre `A` et `B` en utilisant l'approche *simple linkage*.

Question 11. Rappeler la définition d'une partition $P = \{C_1, \ldots, C_K\}$ d'un ensemble $\mathcal{X}$ en K sous-ensembles.

Exercice 2 *Arbres de décision*

On considère la base d'apprentissage suivante :

Température	Saison	Pluie	Activité
élevée	printemps	non	pas de sortie
élevée	été	oui	sortie
basse	printemps	non	sortie
basse	printemps	oui	pas de sortie

Question 1. En utilisant l'algorithme de construction d'arbres de décision vu en cours, et en détaillant les étapes et calculs réalisés, construire un arbre de décision permettant de prédire l'activité connaissant les valeurs pour les 3 attributs de description (température, saison et pluie).
Indications : quelques valeurs de logarithme en base 2 : $\log(\frac{1}{2}) = -1$, $\log(\frac{1}{3}) = -1.58$, $\log(\frac{2}{3}) = -0.58$, $\log(\frac{1}{4}) = -2$ et $\log(\frac{3}{4}) = -0.42$.