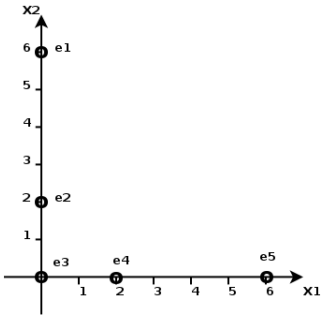
Université Paris Descartes
UFR de Mathématiques et Informatique

Classification Master 1 Ingénierie Mathématique
Examen partiel du 8 mars 2021 - Correction de l'exercice 1

Exercice 1

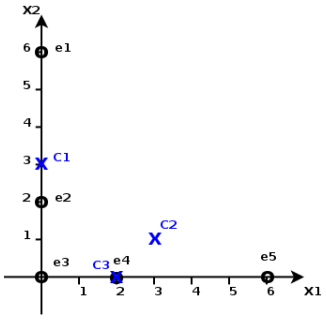
1. 

2. (a) partition initiale : $\{\{e_1, e_3\}, \{e_2, e_5\}, \{e_4\}\}$

   (b) étape 1 :

   i. calcul des centres :

$$\begin{cases} \Gamma_1^1 = \frac{X_1+X_3}{2} = (0,3), \\ \Gamma_2^1 = \frac{X_2+X_5}{2} = (3,1), \\ \Gamma_3^1 = X_4 = (2,0), \end{cases}$$



   ii. affectation :

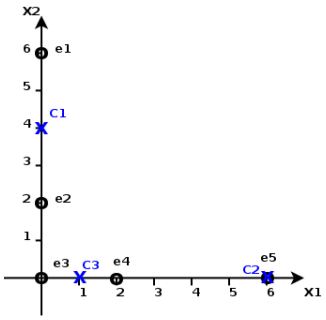
$$\begin{cases} e_1 \text{ et } e_2 \text{ sont plus proches de } \Gamma_1^1, \\ e_3 \text{ et } e_4 \text{ sont plus proches de } \Gamma_3^1, \\ e_5 \text{ est plus proche de } \Gamma_2^1. \end{cases}$$

   D'où la partition : $\{\{e_1, e_2\}, \{e_5\}, \{e_3, e_4\}\}$

   (c) étape 2 :

   i. calcul des centres :

$$\begin{cases} \Gamma_1^2 = \frac{X_1+X_2}{2} = (0,4), \\ \Gamma_2^2 = X_5 = (6,0), \\ \Gamma_3^2 = \frac{X_3+X_4}{2} = (1,0), \end{cases}$$



ii. affectation :

$$\begin{cases} e_1 \text{ et } e_2 \text{ sont plus proches de } \Gamma_1^2, \\ e_3 \text{ et } e_4 \text{ sont plus proches de } \Gamma_3^2, \\ e_5 \text{ est plus proche de } \Gamma_2^2. \end{cases}$$

D'où la partition : $\{\{e_1, e_2\}, \{e_5\}, \{e_3, e_4\}\}$. Cette partition est identique à la précédente, par conséquent l'algorithme est terminé.

3. Au vu de la position des points, il semble que la partition obtenue n'est pas optimale ; une partition plus cohérente serait $\{\{e_1\}, \{e_2, e_3, e_4\}, \{e_5\}\}$.
4. cf. pages 20 à 22 des slides.
5. Pour une partition $\mathcal{P} = \{C_1, \ldots, C_K\}$ l'inertie intra s'écrit

$$I_{intra}(\mathcal{P}) = \frac{1}{n} \sum_{k=1}^{K} \sum_{e_i \in C_k} \|X_i - \Gamma_k\|^2,$$

où Γ_k est le barycentre de la classe C_k.
Pour la partition initiale $\mathcal{P}_{init} = \{\{e_1, e_3\}, \{e_2, e_5\}, \{e_4\}\}$ on a $\Gamma_1 = (0, 3)$, $\Gamma_2 = (3, 1)$, $\Gamma_3 = (2, 0)$, et donc :

$$\begin{aligned} I_{intra}(\mathcal{P}_{init}) &= \frac{1}{5}\left(\|X_1 - \Gamma_1\|^2 + \|X_3 - \Gamma_1\|^2 + \|X_2 - \Gamma_2\|^2 + \|X_5 - \Gamma_2\|^2 + \|X_4 - \Gamma_3\|^2\right) \\ &= \frac{1}{5}\left(\|(0,6) - (0,3)\|^2 + \|(0,0) - (0,3)\|^2\right. \\ &\quad \left. + \|(0,2) - (3,1)\|^2 + \|(6,0) - (3,1)\|^2 + \|(2,0) - (2,0)\|^2\right) \\ &= \frac{1}{5}(3^2 + 3^2 + (3^2 + 1^2) + (3^1 + 1^2) + 0) = \frac{38}{5} = 7.6 \end{aligned}$$

Pour la partition finale $\mathcal{P}_{init} = \{\{e_1, e_2\}, \{e_5\}, \{e_3, e_4\}\}$ on a $\Gamma_1 = (0, 4)$, $\Gamma_2 = (6, 0)$, $\Gamma_3 = (1, 0)$, et donc :

$$\begin{aligned} I_{intra}(\mathcal{P}_{init}) &= \frac{1}{5}\left(\|X_1 - \Gamma_1\|^2 + \|X_2 - \Gamma_1\|^2 + \|X_5 - \Gamma_2\|^2 + \|X_3 - \Gamma_3\|^2 + \|X_4 - \Gamma_3\|^2\right) \\ &= \frac{1}{5}\left(\|(0,6) - (0,4)\|^2 + \|(0,2) - (0,4)\|^2\right. \\ &\quad \left. + \|(6,0) - (6,0)\|^2 + \|(0,0) - (1,0)\|^2 + \|(2,0) - (1,0)\|^2\right) \\ &= \frac{1}{5}(2^2 + 2^2 + 0 + 1^2 + 1^2) = \frac{10}{5} = 2 \end{aligned}$$

Pour la partition "optimale" $\mathcal{P}_{opt} = \{\{e_1\}, \{e_2, e_3, e_4\}, \{e_5\}\}$ on a $\Gamma_1 = (0, 6)$, $\Gamma_2 = (2/3, 2/3)$,

$\Gamma_3 = (6, 0)$, et donc :

$$\begin{aligned} I_{intra}(\mathcal{P}_{opt}) &= \frac{1}{5}\left(\|X_1 - \Gamma_1\|^2 + \|X_2 - \Gamma_2\|^2 + \|X_3 - \Gamma_2\|^2 + \|X_4 - \Gamma_2\|^2 + \|X_5 - \Gamma_3\|^2\right) \\ &= \frac{1}{5}\left(\|(0,6) - (0,6)\|^2 + \|(0,2) - (2/3, 2/3)\|^2\right. \\ &\quad \left. + \|(0,0) - (2/3,2/3)\|^2 + \|(2,0) - (2/3,2/3)\|^2 + \|(6,0) - (6,0)\|^2\right) \\ &= \frac{1}{5}\left(0 + \frac{2^2+4^2}{3^2} + \frac{2^2+2^2}{3^2} + \frac{4^2+2^2}{3^2} + 0\right) \\ &= \frac{20+8+20}{5\times 9} = \frac{48}{5\times 9} = \frac{16}{15} \simeq 1.1 \end{aligned}$$

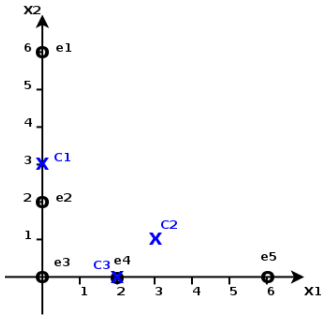
On peut remarquer que d'une part l'inertie intra classe finale est inférieure à l'inertie intra classe initiale, ce qui était attendue puisqu'on sait que l'inertie intra classe décroit au cours des itérations de l'algorithme.

D'autre part l'inertie intra classe de la partition "optimale" est bien inférieure à l'inertie intra classe pour la partition finale, conformément à l'intuition. Ceci signifie que l'algorithme des K-moyennes n'a ici pas convergé vers le minimum global de l'énergie.

N.B. On n'a pas prouvé ici que la partition "optimale" correspond bien au minimum global de l'inertie intra.

6. L'algorithme de MacQueen a l'avantage de converger souvent plus rapidement que l'algorithme de Forgy (en terme de nombre d'itérations). Son désavantage est qu'il y a plus de calculs à faire à chaque itération, car il faut mettre à jour les centres après chaque réaffectation. Cependant on peut remarquer qu'il n'y a besoin que de mettre à jour qu'au plus deux centres à chaque affectation d'une observation e_i (le centre de l'ancienne classe de e_i et le centre de sa nouvelle classe), et que même ces deux mises à jour peuvent se faire plus rapidement qu'un calcul complet de barycentre. Au final l'algorithme est à peine plus coûteux en temps de calcul par itération, et comme il converge plus rapidement, il est plus efficace.
7. On reprend l'algorithme :
   (a) partition initiale : $\{\{e_1, e_3\}, \{e_2, e_5\}, \{e_4\}\}$
   (b) calcul initial des centres :

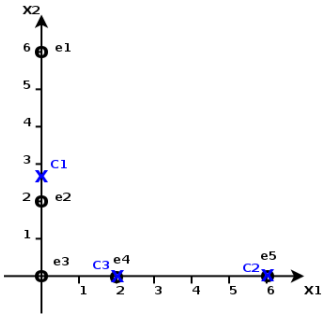
$$\begin{cases} \Gamma_1 = \frac{X_1+X_3}{2} = (0,3), \\ \Gamma_2 = \frac{X_2+X_5}{2} = (3,1), \\ \Gamma_3 = X_4 = (2,0), \end{cases}$$



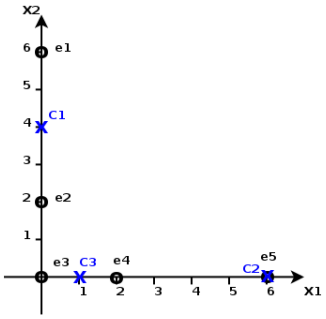
   (c) étape 1 : On ré-affecte successivement $e_1, e_2, \ldots, e_5$ en mettant à jour les centres quand c'est nécessaire.

— e_1 est plus proche de Γ_1, donc il reste dans C_1

— e_2 est plus proche de Γ_1, donc on l'affecte à C_1 et non plus à C_2. On a donc à présent $C_1 = \{e_1, e_2, e_3\}$, donc $\Gamma_1 = (0, 8/3)$ et $C_2 = \{e_5\}$, donc $\Gamma_2 = (6, 0)$.



— e_3 est plus proche de Γ_3, donc on l'affecte à C_3 et non plus à C_1. On a donc à présent $C_1 = \{e_1, e_2\}$, donc $\Gamma_1 = (0, 4)$ et $C_3 = \{e_3, e_4\}$, donc $\Gamma_3 = (1, 0)$.



— e_4 est plus proche de Γ_3, donc il reste dans C_3,

— e_5 est plus proche de Γ_2, donc il reste dans C_2

Après cette première étape la partition courante est donc $\{\{e_1, e_2\}, \{e_5\}, \{e_3, e_4\}\}$

(d) étape 2 : On ré-affecte successivement $e_1, e_2, \ldots, e_5$ en mettant à jour les centres quand c'est nécessaire.

— e_1 est plus proche de Γ_1, donc il reste dans C_1,

— e_2 est plus proche de Γ_1, donc il reste dans C_1,

— e_3 est plus proche de Γ_3, donc il reste dans C_3,

— e_4 est plus proche de Γ_3, donc il reste dans C_3,

— e_5 est plus proche de Γ_2, donc il reste dans C_2.

On voit donc que rien n'a changé et donc l'algorithme s'arrête.

La partition finale obtenue est donc $\{\{e_1, e_2\}, \{e_5\}, \{e_3, e_4\}\}$, c'est-à-dire la même que pour l'algorithme de Forgy, et on a obtenu également cette partition après 1 itération. Il n'y a donc pas vraiment de différence entre les deux algorithmes sur cet exemple jouet. Il faudrait des données de taille plus grande pour observer des différences.