

Feuille de TD

**Exercice 1. Formule de Huygens**

On considère un jeu de données  $\mathbf{X} = \{x_1, \dots, x_n\}$  avec  $x_i \in \mathbb{R}^p$ . Soit  $\mathcal{P}_K = \{\mathcal{C}_1, \dots, \mathcal{C}_K\}$  une partition de  $\{1, \dots, n\}$ . Soit  $m_k$  le centre de gravité associé à la classe  $\mathcal{C}_k$  pour la distance euclidienne usuelle  $\|\cdot\|$  de  $\mathbb{R}^p$  et  $c$  le centre de gravité de l'ensemble des points.

1. Donnez la définition de l'inertie totale  $\mathcal{I}$  associée à  $\mathbf{X}$ .
2. Donnez la définition de l'inertie intra-classe  $\mathcal{I}_{\text{intra}}$  et l'inertie inter-classe  $\mathcal{I}_{\text{inter}}$  associées à  $\mathcal{P}_K$  et  $\mathbf{X}$ .
3. En remarquant que  $x_i - c = x_i - m_k + m_k - c$ , montrez que  $\mathcal{I} = \mathcal{I}_{\text{intra}} + \mathcal{I}_{\text{inter}}$ .

**Exercice 2. Propriétés des Kmeans**

On considère un jeu de données  $\mathbf{X} = \{x_1, \dots, x_n\}$  avec  $x_i \in \mathbb{R}^p$ . A la  $t$ -ième itération de l'algorithme des Kmeans, on a

- une partition  $\mathcal{P}_K^{(t)} = \{\mathcal{C}_1^{(t)}, \dots, \mathcal{C}_K^{(t)}\}$  de  $\{1, \dots, n\}$  avec

$$\mathcal{C}_k^{(t)} = \left\{ i \in \{1, \dots, n\}; h^{(t)}(i) = k \right\} \text{ où } h^{(t)}(i) = \underset{k=1, \dots, K}{\operatorname{argmin}} \left\| x_i - c_k^{(t-1)} \right\|.$$

- des centres de gravité  $c_1^{(t)}, \dots, c_K^{(t)}$  pour la partition  $\mathcal{P}_K^{(t)}$  définis par

$$c_k^{(t)} = \frac{1}{|\mathcal{C}_k^{(t)}|} \sum_{i \in \mathcal{C}_k^{(t)}} x_i.$$

1. Rappelons que l'algorithme des Kmeans est un algorithme qui alterne les 2 étapes suivantes : 1). Mise à jour de l'allocation des individus à une classe; 2). Mise à jour des centroides.
  - (a) Définissez la fonction d'allocation  $h^{(t+1)}$  correspondant à l'étape 1 et la partition  $\mathcal{P}_K^{(t+1)}$ .
  - (b) Précisez le problème d'optimisation dont  $c_k^{(t+1)}$  est solution dans l'étape 2.
2. Montrons que l'inertie intra-classe décroît à chaque étape de l'algorithme des Kmeans.
  - (a) Donnez la définition de l'inertie intra-classe associée à la partition  $\mathcal{P}_K^{(t)}$ . On la note  $\mathcal{I}_{\text{intra}}(\mathcal{P}_K^{(t)})$  dans la suite.
  - (b) Montrez que

$$\mathcal{I}_{\text{intra}}(\mathcal{P}_K^{(t)}) \geq \sum_{i=1}^n \|x_i - c_{h^{(t+1)}(i)}^{(t)}\|^2 = \sum_{k=1}^K \sum_{i \in \mathcal{C}_k^{(t+1)}} \|x_i - c_k^{(t)}\|^2$$

- (c) Déduisez-en que  $\left(\mathcal{I}_{\text{intra}}(\mathcal{P}_K^{(t)})\right)_{t \in \mathbb{N}}$  est une suite décroissante.
3. En remarquant que le nombre de partitions possibles est fini, prouvez que la suite  $\left(\mathcal{I}_{\text{intra}}(\mathcal{P}_K^{(t)})\right)_{t \in \mathbb{N}}$  converge et atteint sa limite.
4. Déduisez-en que la partition se stabilise en un nombre fini d'étape :

$$\exists t_0; \forall t \geq t_0 \mathcal{P}_K^{(t)} = \mathcal{P}_K^{(t_0)}.$$

### Exercice 3. Mesure d'agrégation de Ward

On considère un jeu de données  $\mathbf{X} = \{x_1, \dots, x_n\}$  avec  $x_i \in \mathbb{R}^p$ . Soit  $\mathcal{P}_K = \{\mathcal{C}_1, \dots, \mathcal{C}_K\}$  une partition de  $\{1, \dots, n\}$ . Soit  $m_k$  le centre de gravité associé à la classe  $\mathcal{C}_k$  pour la distance euclidienne usuelle  $\|\cdot\|$  de  $\mathbb{R}^p$  et  $c$  le centre de gravité de l'ensemble des points. L'objectif est de montrer que si on fusionne les classes  $\mathcal{C}_k$  et  $\mathcal{C}_{k'}$  en  $\mathcal{C}_{k \cup k'}$  pour  $(k, k') \in \{1, \dots, K\}^2$ ,  $k \neq k'$ , alors l'inertie inter-classe diminue de

$$D(\mathcal{C}_k, \mathcal{C}_{k'}) := \frac{n_k n_{k'}}{n_k + n_{k'}} \|m_k - m_{k'}\|^2$$

où  $n_k$  est le cardinal de  $\mathcal{C}_k$  et  $m_k = \frac{1}{n_k} \sum_{i \in \mathcal{C}_k} x_i$ . On note  $\mathcal{P}_{K-1}$  la partition obtenue après la fusion de  $\mathcal{C}_k$  et  $\mathcal{C}_{k'}$ .

1. On note  $\mathcal{I}_{\text{inter}}(\mathcal{P}_K)$  l'inertie inter-classe associée à la partition  $\mathcal{P}_K$ . Montrez que

$$\frac{\{\mathcal{I}_{\text{inter}}(\mathcal{P}_K) - \mathcal{I}_{\text{inter}}(\mathcal{P}_{K-1})\}}{n_k + n_{k'}} = v_k \|m_k - c\|^2 + v_{k'} \|m_{k'} - c\|^2 - \|m_{k \cup k'}\|^2$$

$$\text{avec } v_k = \frac{n_k}{n_k + n_{k'}} \text{ et } v_{k'} = \frac{n_{k'}}{n_k + n_{k'}}.$$

2. Exprimez  $m_{k \cup k'}$  en fonction de  $m_k$  et  $m_{k'}$ .
3. Montrez que  $\|m_{k \cup k'} - c\|^2 = v_k^2 \|m_k - c\|^2 + v_{k'}^2 \|m_{k'} - c\|^2 + 2v_k v_{k'} \langle m_k - c, m_{k'} - c \rangle$
4. Trouvez une relation entre  $\|m_k - m_{k'}\|^2$  et  $\langle m_k - c, m_{k'} - c \rangle$ .
5. Concluez.
6. A l'aide des calculs précédents, montrez que l'on peut utiliser la formule de Lance et Williams pour remettre à jour les mesures d'agrégation de Ward à chaque étape car

$$D(\mathcal{C}_u, \mathcal{C}_{k \cup k'}) = \alpha D(\mathcal{C}_u, \mathcal{C}_k) + \beta D(\mathcal{C}_u, \mathcal{C}_{k'}) + \gamma D(\mathcal{C}_k, \mathcal{C}_{k'})$$

avec

$$\alpha = \frac{n_u + n_k}{n_u + n_k + n_{k'}}, \beta = \frac{n_u + n_{k'}}{n_u + n_k + n_{k'}} \text{ et } \gamma = -\frac{n_u}{n_u + n_k + n_{k'}}.$$

#### Exercice 4. Propriété de l'algorithme EM

On reprend les notations du cours : soit  $\mathbf{x} = (x_1, \dots, x_n)$  les observations,  $\mathbf{z}$  le vecteur des labels,  $\mathcal{L}(\mathbf{x}|\theta)$  la logvraisemblance et  $\mathcal{Q}(\theta|\theta^{(r)}) = \mathbb{E} [\ln(f(\mathbf{z}|\theta)|\mathbf{x}, \theta^{(r)})]$ .

L'objectif de cet exercice est de démontrer que la logvraisemblance croît à chaque étape de l'algorithme EM :  $\mathcal{L}(\mathbf{x}|\theta^{(r)}) \leq \mathcal{L}(\mathbf{x}|\theta^{(r+1)})$

1. Quelle relation existe entre  $\mathcal{L}(\mathbf{x}|\theta)$ ,  $\mathcal{Q}(\theta|\theta^{(r)})$  et  $H(\theta|\theta^{(r)}) = \mathbb{E} [\ln(f(\mathbf{z}|\mathbf{x}, \theta)|\mathbf{x}, \theta^{(r)})]$ .
2. Justifiez que  $\mathcal{Q}(\theta^{(r+1)}|\theta^{(r)}) \geq \mathcal{Q}(\theta^{(r)}|\theta^{(r)})$ .
3. Montrez que  $\forall \theta \in \Theta, H(\theta|\theta^{(r)}) \leq H(\theta^{(r)}|\theta^{(r)})$ .

*Indication: on pourra utiliser l'inégalité de Jensen:*

*Si  $\Phi : \mathbb{R} \rightarrow \mathbb{R}$  est une fonction convexe et si  $f$  est une fonction borélienne telle que  $f$  et  $\Phi \circ f$  sont intégrables par rapport à une mesure de probabilité  $\mu$  alors  $\Phi(\int f d\mu) \leq \int \Phi(f) d\mu$ .*

4. Concluez.

#### Exercice 5. Lien entre K-means et CEM

On considère un ensemble d'observations  $\{x_1, \dots, x_n\}$  avec  $x_i \in \mathbb{R}^p$ .

1. On suppose que ces observations sont des réalisations d'un  $n$ -échantillon distribué selon un mélange gaussien à  $K$  composantes avec des proportions toutes identiques, des matrices de covariances de la forme  $\sigma^2 I_p$  ( $I_p$  étant la matrice identité de  $\mathcal{M}_p(\mathbb{R})$ ) et des vecteurs moyenne  $\mu_k$ . Ecrivez la densité de distribution des données.
2. On note  $\mathbf{z} = (z_1, \dots, z_n)$  le vecteur des labels. D'après la question 1, donnez l'expression de la logvraisemblance observée et de la logvraisemblance complétée des données.
3. On propose d'estimer le vecteur des paramètres  $\theta = (\mu_1, \dots, \mu_K, \sigma^2)$  en utilisant l'algorithme CEM. Décrivez les trois étapes de cet algorithme.
4. On s'intéresse au lien existant entre la procédure étudiée dans la question 3 et l'algorithme des K-means. Rappelez les étapes de l'algorithme des K-means et concluez.

#### Exercice 6. Algorithme EM pour mélange gaussien multidimensionnel

On considère que les observations  $\{x_1, \dots, x_n\}$  avec  $x_i \in \mathbb{R}^p$  sont des réalisations d'un  $n$ -échantillon distribué selon un mélange gaussien multidimensionnel

$$x \in \mathbb{R}^p \mapsto \sum_{k=1}^K \pi_k \Phi(x|\mu_k, \Sigma_k)$$

où  $\Phi(\cdot|\mu_k, \Sigma_k)$  est la densité de la loi gaussienne multidimensionnelle  $\mathcal{N}_p(\mu_k, \Sigma_k)$  de vecteur moyenne  $\mu_k$  et de matrice de covariance  $\Sigma_k$ . On ne fait aucune hypothèse sur la forme des matrices de covariance et les proportions  $\pi_k$  sont libres.

Décrivez les étapes de l'algorithme EM pour estimer le vecteur des paramètres.

*Indications:*

- on pourra remarquer que si  $x \in \mathbb{R}^p$  et  $A \in \mathcal{M}_p(\mathbb{R})$  alors  $x'Ax = \text{Tr}(Axx')$ .
- on pourra utiliser la propriété suivante:  
Soit  $B \in \mathcal{M}_p(\mathbb{R})$  une matrice symétrique définie positive et  $\alpha > 0$ . Alors la matrice minimisant  $M \in \mathcal{M}_p(\mathbb{R}) \mapsto \text{Tr}(BM^{-1}) + \alpha \ln(|M|)$  est  $M = \frac{B}{\alpha}$ .