# Exercises on clustering

INSA Toulouse
Olivier Roustant & Joris Guerin

## Exercise 1: Hierarchical clustering and Ward dissimilarity

Let $\mathcal{C} = \{x_1, \dots, x_n\}$ be a dataset of $\mathbb{R}^d$. Assume that these data are split in 2 classes $\mathcal{C}_1, \mathcal{C}_2$ of sizes $n_1, n_2$. Denote by $g$ the global centroid and by $g_1, g_2$ the class centroids:

$$g = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad g_j = \frac{1}{n_j}\sum_{i \in \mathcal{C}_j} x_i \quad (j = 1, 2).$$

Denote by $\|.\|$ the usual Euclidean norm in $\mathbb{R}^d$. Define the inertia of a set of points as the unnormalized variance:

$$I(\mathcal{C}) = \sum_{i=1}^{n} \|x_i - g\|^2, \qquad I(\mathcal{C}_j) = \sum_{i \in \mathcal{C}_j} \|x_i - g_j\|^2 \quad (j = 1, 2).$$

1. In this question, you prove the formula of *inertia decomposition.*
   - By writing $x_i - g = (x_i - g_j) + (g_j - g)$, show that (one form of Huygens formula)
   $$\sum_{i \in \mathcal{C}_j} \|x_i - g\|^2 = I(\mathcal{C}_j) + n_j \|g_j - g\|^2.$$
   - Deduce that
   $$I(\mathcal{C}) = \sum_{j=1}^{2} I(\mathcal{C}_j) + \sum_{j=1}^{2} n_j \|g_j - g\|^2.$$

   Explain why the first term is called *within-class inertia*, and the second one *between-class inertia.* Is this formula true for $K$ classes?

2. In this question, you obtain a simple expression of the between-class inertia, involving the distance between centroids.
   - What property directly gives the formula: $g = \frac{n_1}{n} g_1 + \frac{n_2}{n} g_2$ ?
   - Deduce that $\|g_1 - g\| = \frac{n_2}{n}\|g_1 - g_2\|$ and $\|g_2 - g\| = \frac{n_1}{n}\|g_1 - g_2\|$.
   - Finally, prove the formula for the between-class inertia:
   $$\sum_{j=1}^{2} n_j \|g_j - g\|^2 = \frac{n_1 n_2}{n_1 + n_2}\|g_1 - g_2\|^2.$$

3. Explain why the between-class inertia is a dissimilarity, called *Ward dissimilarity.*

4. Illustrate through an example the difference between average and ward linkage.

## Exercise 2: Convergence of $k$-means algorithm.

Let $\{x_1, \dots, x_n\}$ be a dataset of $\mathbb{R}^d$. At step $t$ of $k$-means algorithm, we have:

- a partition $\mathcal{A}^{(t)}$ in $K$ classes, i.e. a function $\mathcal{A}^{(t)} : \{1, \dots, n\} \to \{1, \dots, K\}$ which allocates a class to each individual. The corresponding classes are $\mathcal{C}_j^{(t)} = \{i \in \{1, \dots, n\}$ such that $\mathcal{A}^{(t)}(i) = j\}$ (for $j = 1, \dots, K$).
- centers $c_1^{(t)}, \dots, c_K^{(t)}$, equal to the class centroids $\mathcal{C}(t)_1, \dots, \mathcal{C}(t)_K$ defined by $\mathcal{A}^{(t)}$.

Recall that $k$-means is a 2 stage procedure:

1. **(Allocation update)**. Choose the new allocation as the closest centroid obtained at previous step (choose one at random in case of equality):

$$\text{For } i = 1, \dots, n : \qquad \mathcal{A}^{(t+1)}(i) = \text{argmin}_{j=1,\dots,K} \|x_i - c_j^{(t)}\|.$$

2. **(Centroid update)**. Compute the centroid of the new class, defined by the new allocation obtained at stage 1:

$$\text{For } j = 1, \dots, K : \qquad c_j^{(t+1)} = \text{argmin}_{c \in \mathbb{R}^n} \sum_{i \in \mathcal{C}_j^{(t+1)}} \|x_i - c\|^2.$$

You are going to prove that **$k$-means stops after a finite number of iterations**: there exists $t_0 \in \mathbb{N}$ such that for all $t \geq t_0$,

$$\text{For all } i = 1, \dots, n, \text{ and } j = 1, \dots, K : \qquad \mathcal{A}^{(t+1)}(i) = \mathcal{A}^{(t)}(i), \quad c_j^{(t+1)} = c_j^{(t)}.$$

For that, we consider the within-class inertia associated to $\mathcal{A}^{(t)}$:

$$J(\mathcal{A}^{(t)}) = \sum_{j=1}^{K} \sum_{i \in \mathcal{C}_j^{(t)}} \|x_i - c_j^{(t)}\|^2 = \sum_{i=1}^{n} \|x_i - c_{\mathcal{A}^{(t)}(i)}^{(t)}\|^2.$$

1. Explain why $c_j^{(t+1)}$ defined in the 'centroid update' step is indeed the centroid of the updated class.
2. In this question you prove that the *within-class inertia decreases*.
   - Explain why $J(\mathcal{A}^{(t)}) \geq \sum_{i=1}^{n} \|x_i - c_{\mathcal{A}^{(t+1)}(i)}^{(t)}\|^2 = \sum_{j=1}^{K} \sum_{i \in \mathcal{C}_j^{(t+1)}} \|x_i - c_j^{(t)}\|^2$.
   - Deduce that $J(\mathcal{A}^{(t)}) \geq \sum_{j=1}^{K} \sum_{i \in \mathcal{C}_j^{(t+1)}} \|x_i - c_j^{(t+1)}\|^2 = J(\mathcal{A}^{(t+1)})$.
3. Prove that the sequence $J(\mathcal{A}^{(t)})$ is converging, and that the limit is reached. *Hint: What can say about all possible partitions $\mathcal{A}^{(t)}$?*
4. Deduce from the 'centroid update' step that, after some $t^*$, for all $j = 1, \dots, K$: $c_j^{(t+1)} = c_j^{(t)}$. Finally, explain why, after $t^* + 1$ we have for all $i = 1, \dots, K$: $\mathcal{A}^{(t+1)}(i) = \mathcal{A}^{(t)}(i)$.
5. Is the limit equal to the global minimum of the within-class inertia?