

### K means algorithm

Let  $n \geq 1$  and  $X_1, \dots, X_n$  in  $\mathbb{R}^d$ . The  $K$ -means algorithm aims at minimizing over all partitions  $G = (G_1, \dots, G_K)$  of  $\{1, \dots, n\}$  the criterion

$$\mathcal{L}(G) = \sum_{k=1}^K \sum_{a \in G_k} \|X_a - \bar{X}_{G_k}\|^2 \quad \text{with} \quad \bar{X}_{G_k} = \frac{1}{|G_k|} \sum_{a \in G_k} X_a .$$

1. Prove that

$$\mathcal{L}(G) = \sum_{k=1}^K \frac{1}{|G_k|} \sum_{a, b \in G_k} \langle X_a, X_a - X_b \rangle = \frac{1}{2} \sum_{k=1}^K \frac{1}{|G_k|} \sum_{a, b \in G_k} \|X_a - X_b\|^2 .$$

2. Assume now that the observations are independent. Write  $\mathbb{E}[X_a] = \mu_a \in \mathbb{R}^d$  so that  $X_a = \mu_a + \varepsilon_a$  with  $\varepsilon_1, \dots, \varepsilon_n$  centered and independent. Define  $v_a = \text{trace}(\mathbb{V}[X_a])$ . Prove that

$$\mathbb{E}[\mathcal{L}(G)] = \frac{1}{2} \sum_{k=1}^K \frac{1}{|G_k|} \sum_{a, b \in G_k} (\|\mu_a - \mu_b\|^2 + v_a + v_b) \mathbf{1}_{a \neq b} .$$

What is the value of  $\mathbb{E}[\mathcal{L}(G)]$  when all the within-group variables have the same mean?

3. We assume now that there exists a partition  $G^* = (G_1^*, \dots, G_K^*)$  such that there exist  $m_1, \dots, m_K \in \mathbb{R}^d$  and  $\gamma_1, \dots, \gamma_K > 0$  satisfying  $\mu_a = m_k$  and  $v_a = \gamma_k$  for all  $a \in G_k^*$  and  $k = 1, \dots, K$ . Compute  $\mathbb{E}[\mathcal{L}(G^*)]$ .
4. In the special case where  $\gamma_1 = \dots = \gamma_K = \gamma$ , which partition  $G = (G_1, \dots, G_K)$  minimizes  $\mathbb{E}[\mathcal{L}(G)]$ ?