

K means algorithm

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

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

1. Prove that

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

By definition,

$$\begin{aligned} \mathcal{L}(G) &= \sum_{k=1}^K \sum_{a \in G_k} \|X_a - \bar{X}_{G_k}\|^2 \\ &= \sum_{k=1}^K \sum_{a \in G_k} \langle X_a - \frac{1}{|G_k|} \sum_{b \in G_k} X_b, X_a - \frac{1}{|G_k|} \sum_{c \in G_k} X_c \rangle \\ &= \sum_{k=1}^K \frac{1}{|G_k|^2} \sum_{a, b, c \in G_k} \langle X_a - X_b, X_a - X_c \rangle \\ &= \sum_{k=1}^K \frac{1}{|G_k|^2} \sum_{a, b, c \in G_k} \langle X_a - X_b, X_a \rangle - \sum_{k=1}^K \frac{1}{|G_k|^2} \sum_{a, b, c \in G_k} \langle X_a - X_b, X_c \rangle, \end{aligned}$$

where

$$\sum_{a, b, c \in G_k} \langle X_a - X_b, X_c \rangle = |G_k| \sum_{a, c \in G_k} \langle X_a, X_c \rangle - |G_k| \sum_{b, c \in G_k} \langle X_b, X_c \rangle = 0.$$

Thus,

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

For the second equality, note that

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

which concludes the proof.

2. Assume now that the observations are independent. Write $\mathbb{E}[X_a] = \mu_a \in \mathbb{R}^p$ 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?

The expectation of $\mathcal{L}(G)$ is given by

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

Let $a, b \in G_k, a \neq b$,

$$\begin{aligned} \mathbb{E}[\|X_a - X_b\|^2] &= \mathbb{E}[\|\mu_a - \mu_b + \varepsilon_a - \varepsilon_b\|^2] \\ &= \mathbb{E}[\|\mu_a - \mu_b\|^2] + \mathbb{E}[\|\varepsilon_a - \varepsilon_b\|^2] + 2\mathbb{E}[\langle \mu_a - \mu_b, \varepsilon_a - \varepsilon_b \rangle] \\ &= \|\mu_a - \mu_b\|^2 + \mathbb{E}[\|\varepsilon_a\|^2] + \mathbb{E}[\|\varepsilon_b\|^2] + 2\mathbb{E}[\langle \varepsilon_a, \varepsilon_b \rangle], \end{aligned}$$

since ε_a and ε_b are independent and centred. Finally, since for all $a \in G_k, \mathbb{E}[\|\varepsilon_a\|^2] = v_a$,

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

If all the within-group variables have the same mean, for all k , there exists μ_k such that, for all $a \in G_k, \mu_a = \mu_k$. Therefore,

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

where

$$\begin{aligned} \frac{1}{|G_k|} \sum_{a,b \in G_k} (v_a + v_b) \mathbf{1}_{a \neq b} &= \frac{1}{|G_k|} \left(\sum_{a,b \in G_k} (v_a + v_b) - \sum_{a,b \in G_k} (v_a + v_b) \mathbf{1}_{a=b} \right) \\ &= \frac{1}{|G_k|} \left(2|G_k| \sum_{a \in G_k} v_a - 2 \sum_{a \in G_k} v_a \right) \\ &= \frac{2(|G_k| - 1)}{|G_k|} \sum_{a \in G_k} v_a. \end{aligned}$$

Consequently, if, for all $a \in G_k, \mu_a = \mu_k$, we have

$$\mathbb{E}[\mathcal{L}(G)] = \sum_{k=1}^K \frac{|G_k| - 1}{|G_k|} \sum_{a \in G_k} v_a.$$

3. We assume now that there exists a partition $G^* = (G_1^*, \dots, G_K^*)$ such that within-group variables have the same mean and the same volume. More precisely, we assume that there exists $m_1, \dots, m_K \in \mathbb{R}^p$ and $\gamma_1, \dots, \gamma_K > 0$ such that $\mu_a = m_k$ and $v_a = \gamma_k$ for all $a \in G_k^*$ and $k = 1, \dots, K$. Compute $\mathbb{E}[\mathcal{L}(G^*)]$.

By definition of G^* ,

$$\begin{aligned}\mathbb{E}[\mathcal{L}(G^*)] &= \sum_{k=1}^K \frac{|G_k^*| - 1}{|G_k^*|} \sum_{a \in G_k^*} v_a \\ &= \sum_{k=1}^K \frac{|G_k^*| - 1}{|G_k^*|} |G_k^*| \gamma_k \\ &= \sum_{k=1}^K (|G_k^*| - 1) \gamma_k.\end{aligned}$$

4. In the special case where $\gamma_1 = \dots = \gamma_K = \gamma$, which partition $G = (G_1, \dots, G_K)$ minimizes $\mathbb{E}[\mathcal{L}(G)]$?

Assume that $\gamma_1 = \dots = \gamma_K = \gamma$. Then, for any partition G ,

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

In particular, for G^* we have

$$\mathbb{E}[\mathcal{L}(G^*)] = \gamma(n - K),$$

which leads to

$$\mathbb{E}[\mathcal{L}(G)] - \mathbb{E}[\mathcal{L}(G^*)] = \frac{1}{2} \sum_{k=1}^K \frac{1}{|G_k|} \sum_{a, b \in G_k} (\|\mu_a - \mu_b\|^2) \geq 0.$$

The minimum of $\mathbb{E}[\mathcal{L}(G)]$ is reached at $G = G^*$. To prove that this minimum is unique, choose G such that $\mathbb{E}[\mathcal{L}(G)] = \mathbb{E}[\mathcal{L}(G^*)]$. Then, for all k , and for all $a, b \in G_k$, $\mu_a = \mu_b$ which implies that $G = G^*$ (if all μ_k are different).