Unsupervised learning

1 K-means algorithm

Let $n \geqslant 1$ 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 $\{1, \dots, p\}$ the criterion

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

1. Prove that

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

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 $\mathbb{E}[X_a] = \mu_a \in \mathbb{R}^d$ so that $X_a = \mu_a + \varepsilon_a$ with $\varepsilon_1, \ldots, \varepsilon_n$ centered and independent. Define $v_a = \mathrm{trace}(\mathbb{V}[X_a])$. Prove that

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

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*

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

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

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

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

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

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,*

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

where

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

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

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

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

By definition of G^*,

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

4. In the special case where there exists $\gamma > 0$ such that $v_i = \gamma$ for all $i \in \{1,\dots,n\}$, which partition $G = (G_1,\dots,G_K)$ minimizes $\mathbb{E}[\mathcal{L}(G)]$?

Assume that $v_a = \gamma$ for all a. Then, for any partition G,

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

In particular, for G^* we have $\mathbb{E}\left[\mathcal{L}(G^*)\right] = \gamma(n-K)$. Therefore, the minimum of $\mathbb{E}\left[\mathcal{L}(G)\right]$ is reached at $G = G^*$. To prove that this minimum is unique when all μ_k are different, choose G such that $\mathbb{E}\left[\mathcal{L}(G)\right] = \mathbb{E}\left[\mathcal{L}(G^*)\right]$. Then, for all k, and for all $a,b \in G_k$, $\mu_a = \mu_b$ which implies that $G = G^*$.

2 EM algorithm (bonus)

In the case where we are interested in estimating unknown parameters $\theta \in \mathbb{R}^m$ characterizing a model with missing data, the Expectation Maximization (EM) algorithm (Dempster et al. 1977) can be used when the joint distribution of the missing data X and the observed data Y is explicit. For all $\theta \in \mathbb{R}^m$, let p_θ be the probability density function of (X,Y) when the model is parameterized by θ with respect to a given reference measure μ. The EM algorithm aims at computing iteratively an approximation of the maximum likelihood estimator which maximizes the observed data loglikelihood:

$$\ell(\theta; Y) = \log p_\theta(Y) = \log \int f_\theta(x, Y)\mu(\mathrm{d}x)\,.$$

As this quantity cannot be computed explicitly in general cases, the EM algorithm finds the maximum likelihood estimator by iteratively maximizing the expected complete data loglikelihood. Start with an inital value $\theta^{(0)}$ and let $\theta^{(t)}$ be the estimate at the t-th iteration for $t \geqslant 0$, then the next iteration of EM is decomposed into two steps.

1. **E step**. Compute the expectation of the complete data loglikelihood, with respect to the conditional distribution of the missing data given the observed data parameterized by $\theta^{(t)}$:

$$Q(\theta, \theta^{(t)}) = \mathbb{E}_{\theta^{(t)}}\left[\log p_\theta(X,Y)|Y\right]\,.$$

2. **M step**. Determine $\theta^{(t+1)}$ by maximizing the function Q:

$$\theta^{(t+1)} \in \mathrm{argmax}_\theta Q(\theta, \theta^{(t)}) \,.$$

1. Prove the following crucial property motivates the EM algorithm. For all $\theta, \theta^{(t)}$,

$$\ell(Y;\theta) - \ell(Y;\theta^{(t)}) \geqslant Q(\theta,\theta^{(t)}) - Q(\theta^{(t)},\theta^{(t)}) \,.$$

This may be proved by noting that

$$\ell(Y;\theta) = \log\left(\frac{p_\theta(X,Y)}{p_\theta(X|Y)}\right) \,.$$

Considering the conditional expectation of both terms given Y when the parameter value is $\theta^{(t)}$ yields

$$\ell(Y;\theta) = Q(\theta,\theta^{(t)}) - \mathbb{E}_{\theta^{(t)}}[\log p_\theta(X|Y)|Y] \,.$$

Then,

$$\ell(Y;\theta) - \ell(Y;\theta^{(t)}) = Q(\theta,\theta^{(t)}) - Q(\theta^{(t)},\theta^{(t)}) + H(\theta,\theta^{(t)}) - H(\theta^{(t)},\theta^{(t)}) \,,$$

where

$$H(\theta,\theta^{(t)}) = -\mathbb{E}_{\theta^{(t)}}[\log p_\theta(X|Y)|Y] \,.$$

The proof is completed by noting that

$$H(\theta,\theta^{(t)}) - H(\theta^{(t)},\theta^{(t)}) \geqslant 0 \,,$$

as this difference if a Kullback-Leibler divergence.

In the following, $X = (X_1,\ldots,X_n)$ and $Y = (Y_1,\ldots,Y_n)$ where $\{(X_i,Y_i)\}_{1\leqslant i\leqslant n}$ are i.i.d. in $\{-1,1\}\times\mathbb{R}^d$. For $k\in\{-1,1\}$, write $\pi_k = \mathbb{P}(X_1 = k)$. Assume that, conditionally on the event $\{X_1 = k\}$, Y_1 has a Gaussian distribution with mean $\mu_k\in\mathbb{R}^d$ and covariance matrix $\Sigma\in\mathbb{R}^{d\times d}$. In this case, the parameter $\theta = (\pi_1,\mu_1,\mu_{-1},\Sigma)$ belongs to the set $\Theta = [0,1]\times\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^{d\times d}$.

1. Write the complete data loglikelihood.

The complete data loglikelihood is given by

$$\begin{aligned}
\log p_\theta(X,Y) &= -\frac{nd}{2}\log(2\pi) + \sum_{i=1}^n \sum_{k\in\{-1,1\}} \mathbb{1}_{X_i=k}\left(\log\pi_k - \frac{\log\det\Sigma}{2} - \frac{1}{2}(Y_i-\mu_k)^\top\Sigma^{-1}(Y_i-\mu_k)\right) \,, \\
&= -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log\det\Sigma + \left(\sum_{i=1}^n \mathbb{1}_{X_i=1}\right)\log\pi_1 + \left(\sum_{i=1}^n \mathbb{1}_{X_i=-1}\right)\log(1-\pi_1) \\
&\quad - \frac{1}{2}\sum_{i=1}^n \mathbb{1}_{X_i=1}(Y_i-\mu_1)^\top\Sigma^{-1}(Y_i-\mu_1) - \frac{1}{2}\sum_{i=1}^n \mathbb{1}_{X_i=-1}(Y_i-\mu_{-1})^\top\Sigma^{-1}(Y_i-\mu_{-1}) \,.
\end{aligned}$$

2. Let $\theta^{(t)}$ be the current parameter estimate. Compute $\theta\mapsto Q(\theta,\theta^{(t)})$.

Write $\omega_t^i = \mathbb{P}_{\theta^{(t)}}(X_i = 1|Y_i)$. The intermediate quantity of the EM algorithm is given by

$$\begin{aligned}
Q(\theta,\theta^{(t)}) &= -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log\det\Sigma + \left(\sum_{i=1}^n \omega_t^i\right)\log\pi_1 + \sum_{i=1}^n\left(1-\omega_t^i\right)\log(1-\pi_1) \\
&\quad - \frac{1}{2}\sum_{i=1}^n \omega_t^i (Y_i-\mu_1)^T\Sigma^{-1}(Y_i-\mu_1) - \frac{1}{2}\sum_{i=1}^n (1-\omega_t^i)(Y_i-\mu_{-1})^T\Sigma^{-1}(Y_i-\mu_{-1}) \,.
\end{aligned}$$

3. Compute $\theta^{(t+1)}$.

The gradient of $Q(\theta, \theta^{(t)})$ with respect to θ is therefore given by

$$\begin{aligned}
\frac{\partial Q(\theta,\theta^{(t)})}{\partial \pi_1} &= \frac{\sum_{i=1}^n \omega_t^i}{\pi_1} - \frac{n - \sum_{i=1}^n \omega_t^i}{1-\pi_1}\,, \\
\frac{\partial Q(\theta,\theta^{(t)})}{\partial \mu_1} &= \sum_{i=1}^n \omega_t^i \left(2\Sigma^{-1}Y_i - 2\Sigma^{-1}\mu_1\right)\,, \\
\frac{\partial Q(\theta,\theta^{(t)})}{\partial \mu_{-1}} &= \sum_{i=1}^n (1-\omega_t^i) \left(2\Sigma^{-1}Y_i - 2\Sigma^{-1}\mu_{-1}\right)\,, \\
\frac{\partial Q(\theta,\theta^{(t)})}{\partial \Sigma^{-1}} &= \frac{n}{2}\Sigma - \frac{1}{2}\sum_{i=1}^n \omega_t^i \left(Y_i-\mu_1\right)\left(Y_i-\mu_1\right)^\top - \frac{1}{2}\sum_{i=1}^n (1-\omega_t^i)\left(Y_i-\mu_{-1}\right)\left(Y_i-\mu_{-1}\right)^\top\,.
\end{aligned}$$

Then, $\theta^{(t+1)}$ is defined as the only parameter such that all these equations are set to 0. It is given by

$$\begin{aligned}
\widehat{\pi}_1^{(t+1)} &= \frac{1}{n}\sum_{i=1}^n \omega_t^i\,, \\
\widehat{\mu}_1^{(t+1)} &= \frac{1}{\sum_{i=1}^n \omega_t^i}\sum_{i=1}^n \omega_t^i Y_i\,, \\
\widehat{\Sigma}^{(t+1)} &= \frac{1}{n}\sum_{i=1}^n \omega_t^i \left(Y_i-\mu_1\right)\left(Y_i-\mu_1\right)^\top + \frac{1}{n}\sum_{i=1}^n (1-\omega_t^i)\left(Y_i-\mu_{-1}\right)\left(Y_i-\mu_{-1}\right)^\top\,.
\end{aligned}$$