MAP569 Machine Learning II

PC8: K-means, Expectation Maximization

K-means algorithm

The K-means algorithm is a procedure which aims at partitioning a data set into K distinct, non-overlapping clusters. Consider $n \geqslant 1$ observations $(X_1, \ldots, X_n)$ taking values in $\mathbb{R}^p$. The K-means algorithm seeks to minimize over all partitions $C = (C_1, \ldots, C_K)$ of $\{1, \ldots, n\}$ the following criterion

$$\mathrm{crit}(C) = \sum_{k=1}^{K} \frac{1}{2|C_k|} \sum_{a,b \in C_k} \|X_a - X_b\|^2 ,$$

where for all $1 \leqslant i \leqslant n$, $1 \leqslant k \leqslant K$, $i \in C_k$ if and only if X_i is in the k-th cluster.

Symmetrization

1. Establish that

$$\mathrm{crit}(C) = \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{a,b \in C_k} \langle X_a, X_a - X_b \rangle = \sum_{k=1}^{K} \sum_{a \in C_k} \|X_a - \bar{X}_{C_k}\|^2 ,$$

where

$$\bar{X}_{C_k} = \frac{1}{|C_k|} \sum_{b \in C_k} X_b .$$

Solution.

Note that

$$\begin{aligned}
\mathrm{crit}(C) &= \sum_{k=1}^{K} \frac{1}{2|C_k|} \sum_{a,b \in C_k} \|X_a - X_b\|^2 ,\\
&= \sum_{k=1}^{K} \frac{1}{2|C_k|} \sum_{a,b \in C_k} \langle X_a - X_b, X_a - X_b \rangle ,\\
&= \sum_{k=1}^{K} \frac{1}{2|C_k|} \left\{ \sum_{a,b \in C_k} \langle X_a - X_b, X_a \rangle + \langle X_b - X_a, X_b \rangle \right\} ,\\
&= \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{a,b \in C_k} \langle X_a - X_b, X_a \rangle .
\end{aligned}$$

which concludes the proof of the first inequality. For the second inequality, write

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

where

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

Thus,

$$\mathrm{crit}(C) = \sum_{k=1}^{K} \sum_{a\in C_k} \|X_a - \bar{X}_{C_k}\|^2\,.$$

□

Independent observations

Assume that the observations are random and independent. Write, for all $1 \leqslant a \leqslant n$, $\mathbb{E}[X_a] = \mu_a \in \mathbb{R}^p$ so that

$$X_a = \mu_a + \varepsilon_a\,,$$

with $(\varepsilon_1, \dots, \varepsilon_n)$ centered and independent random variables. For all $1 \leqslant a \leqslant n$, define

$$v_a = \mathrm{trace}(cov(X_a))\,.$$

1. Check that the expected value of the criterion is

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

Solution.

The expectation of $\mathrm{crit}(C)$ is given by

$$\mathbb{E}\left[\mathrm{crit}(C)\right] = \frac{1}{2}\sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{a,b\in C_k} \mathbb{E}\left[\|X_a - X_b\|^2\right]\,.$$

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

$$\begin{aligned}
\mathbb{E}\left[\|X_a - X_b\|^2\right] &= \mathbb{E}\left[\|\mu_a - \mu_b + \varepsilon_a - \varepsilon_b\|^2\right]\,,\\
&= \mathbb{E}\left[\|\mu_a - \mu_b\|^2\right] + \mathbb{E}\left[\|\varepsilon_a - \varepsilon_b\|^2\right] + 2\underbrace{\mathbb{E}\left[\langle \mu_a - \mu_b, \varepsilon_a - \varepsilon_b\rangle\right]}_{=0}\,,\\
&= \|\mu_a - \mu_b\|^2 + \mathbb{E}\left[\|\varepsilon_a\|^2\right] + \mathbb{E}\left[\|\varepsilon_b\|^2\right] + 2\underbrace{\mathbb{E}\left[\langle \varepsilon_a, \varepsilon_b\rangle\right]}_{=0}\,,
\end{aligned}$$

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

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

□

2. What is the value of $\mathbb{E}[\mathrm{crit}(C)]$ when for all $1 \leqslant k \leqslant K$, there exists $m_k \in \mathbb{R}^p$ such that for all $a \in C_k$, $\mu_a = m_k$?
Solution.

In this setting,

$$\mathbb{E}\left[\mathrm{crit}(C)\right] = \frac{1}{2}\sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{a,b\in C_k} (v_a + v_b)\, \mathbb{1}_{a\neq b}\,,$$

where

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

Consequently, if, for all $a\in C_k$, $\mu_a = m_k$,

$$\mathbb{E}\left[\mathrm{crit}(C)\right] = \sum_{k=1}^{K}\frac{|C_k|-1}{|C_k|}\sum_{a\in C_k}v_a\,.$$

□

Mixture model

Assume now that there exists a partition $C^* = (C_1^*,\ldots,C_K^*)$ such that there exist $m_1^*,\ldots,m_K^*$ in $\mathbb{R}^p$ and $\gamma_1^*,\ldots,\gamma_K^*$ in $\mathbb{R}_+^*$ satisfying $\mu_a = m_k^*$ and $v_a = \gamma_k^*$ for all $a\in C_k^*$ and $k=1,\ldots,K$. This section investigates under which condition the expected value of the K-means criterion is minimum at C^*.

1. What is the value of $\mathbb{E}[\mathrm{crit}(C^*)]$?
Solution.

By the previous question,

$$\mathbb{E}\left[\mathrm{crit}(C)\right] = \sum_{k=1}^{K}\frac{|C_k|-1}{|C_k|}\sum_{a\in C_k}v_a = \sum_{k=1}^{K}\frac{|C_k|-1}{|C_k|}|C_k|\gamma_k = \sum_{k=1}^{K}(|C_k|-1)\gamma_k\,.$$

□

2. In the special case where $\gamma_1^* = \ldots = \gamma_K^* = \gamma$, which partition $C=(C_1,\ldots,C_K)$ minimizes $\mathbb{E}[\mathrm{crit}(C)]$ under the constraint $\gamma_1 = \ldots = \gamma_K = \gamma$?
Solution.

For any partition C,

$$\mathbb{E}\left[\mathrm{crit}(C)\right] = \frac{1}{2}\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{a,b\in C_k}\left(\|\mu_a-\mu_b\|^2\right) + \frac{1}{2}\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{a,b\in C_k}(v_a+v_b)\,\mathbb{1}_{a\neq b}\,,$$
$$= \frac{1}{2}\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{a,b\in C_k}\left(\|\mu_a-\mu_b\|^2\right) + \sum_{k}\frac{|C_k|-1}{|C_k|}\sum_{a\in C_k}\gamma\,,$$
$$= \frac{1}{2}\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{a,b\in C_k}\left(\|\mu_a-\mu_b\|^2\right) + \gamma(n-K)\,.$$

In particular, for C^*,

$$\mathbb{E}\left[\mathrm{crit}(C^*)\right] = \gamma(n-K)\,,$$

which leads to

$$\mathbb{E}\left[\mathrm{crit}(C)\right] - \mathbb{E}\left[\mathrm{crit}(C^*)\right] = \frac{1}{2}\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{a,b\in C_k}\left(\|\mu_a-\mu_b\|^2\right) \geqslant 0\,.$$

The minimum of $\mathbb{E}\left[\mathrm{crit}(C)\right]$ is reached at $C=C^*$. To prove that this minimum is unique under the constraint, choose C such that $\mathbb{E}\left[\mathrm{crit}(C)\right] = \mathbb{E}\left[\mathrm{crit}(C^*)\right]$. Then, for all k, and for all $a,b\in C_k$, $\mu_a=\mu_b$ which implies that $C=C^*$ (if all μ_k are different). □

3. Assume now that C^* contains $K=3$ groups of size s (with s even),

$$m_1 = (1,0,0)^T,\quad m_2 = (0,1,0)^T,\quad m_3 = (0,1-\tau,\sqrt{1-(1-\tau)^2})^T,$$

with $\tau > 0$, and

$$\gamma_1 = \gamma_+, \quad \gamma_2 = \gamma_3 = \gamma_- .$$

What is the value of $\|\mu_2 - \mu_3\|^2$?
Solution.

Simple algebra leads to $\|\mu_2 - \mu_3\|^2 = 2\tau$. □

4. Compute $\mathbb{E}[\mathrm{crit}(C^*)]$.
Solution.

By question 1,

$$\mathbb{E}\left[\mathrm{crit}(C^*)\right] = \sum_{k=1}^{K} (|C_k| - 1)\gamma_k = (s-1)(\gamma_+ + 2\gamma_-) .$$

□

5. Define C' obtained by splitting C_1^* into two groups C_1', C_2' of equal size $s/2$ and by merging C_2^* and C_3^* into a single group C_3' of size $2s$. Check that

$$\mathbb{E}[\mathrm{crit}(C')] = s(\gamma_+ + 2\gamma_- + \tau) - (2\gamma_+ + \gamma_-) .$$

Solution.

Write

$$\begin{aligned}
\mathbb{E}\left[\mathrm{crit}(C')\right] &= \frac{1}{2}\sum_{k=1}^{3}\frac{1}{|C_k'|}\sum_{a,b\in C_k'}\left(\|\mu_a - \mu_b\|^2 + v_a + v_b\right)\mathbb{1}_{a\neq b}\,, \\
&= 2\left(\frac{1}{2}\frac{1}{s/2}\sum_{a,b\in C_k'}(2\gamma_+)\,\mathbb{1}_{a\neq b}\right) + \frac{1}{4s}\sum_{a,b\in C_k'}\|\mu_a - \mu_b\|^2 + \frac{1}{4s}\sum_{a,b\in C_k'}(2\gamma_-)\,\mathbb{1}_{a\neq b}\,, \\
&= 2\gamma_+(\frac{s}{2} - 1) + \frac{2s^2}{4s}\|\mu_2 - \mu_3\|^2 + \gamma_-(2s-1)\,, \\
&= 2\gamma_+(\frac{s}{2} - 1) + \tau s + \gamma_-(2s-1)\,.
\end{aligned}$$

□

6. Under which assumption $\mathbb{E}[\mathrm{crit}(C^*)] < \mathbb{E}[\mathrm{crit}(C')]$?
Solution.

According to question 4 and 5,

$$\begin{aligned}
\mathbb{E}\left[\mathrm{crit}(C^*)\right] < \mathbb{E}\left[\mathrm{crit}(C')\right] &\Leftrightarrow (s-1)(\gamma_+ + 2\gamma_-) < 2\gamma_+(\frac{s}{2} - 1) + \tau s + \gamma_-(2s-1)\,, \\
&\Leftrightarrow \gamma_+ - \gamma_- < s\tau\,, \\
&\Leftrightarrow \|\mu_2 - \mu_3\|^2 > 2\left(\frac{\gamma_+ - \gamma_-}{s}\right) .
\end{aligned}$$

□

Expectation Maximization algorithm

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 $\theta, \theta^{(t)}$,

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

Solution.

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

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

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

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

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.
Solution.

The complete data loglikelihood is given by

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

□

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

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

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

□

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

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

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

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

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

□