PC8. Ecole Polytechnique. MAP 569. Machine Learning II.

Exercise 1 (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 $(X_1, \dots, X_n)$ taking values in $\mathbb{R}^p$. The K-means algorithm seeks to minimize over all partitions $C = (C_1, \dots, C_K)$ of $\{1, \dots, n\}$ 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

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

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

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

2. Check that the expected value of the criterion is

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

3. 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 $a \in C_k$, $\mu_a = m_k$?

Mixture model

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

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

5. In the special case where $\gamma_1^* = \ldots = \gamma_K^* = \gamma$, which partition $C = (C_1, \dots, C_K)$ minimizes $\mathbb{E}[\mathrm{crit}(C)]$ under the constraint for all $k \in [\![1, K]\!]$ and all $a \in C_k$, $v_a = \gamma$?

6. 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 $\|m_2 - m_3\|^2$?

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

8. 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

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

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

Exercise 2 (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:

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

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.

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 $\theta^{(t)}$:

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

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

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

Therefore, we straightforwardly have that the EM algorithm produces a non decreasing sequence of loglikelihoods $\left(\ell(Y; \theta^{(t)})\right)_t$.

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 $\Theta = [0, 1] \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^{d \times d}$.

2. Write the complete data loglikelihood.

3. Let $\theta^{(t)}$ be the current parameter estimate. Compute $\theta \mapsto Q(\theta, \theta^{(t)})$ (tips: use $\omega_t^i = \mathbb{P}_{\theta^{(t)}}(X_i = 1 | Y_i)$)

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