Unsupervised learning

K means algorithm

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

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

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 $\varepsilon_1, \ldots, \varepsilon_n$ 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?

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 $\mathbb{E}[\mathcal{L}(G^*)]$.

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

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

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

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.
2. Let $\theta^{(t)}$ be the current parameter estimate. Compute $\theta \mapsto Q(\theta, \theta^{(t)})$.
3. Compute $\theta^{(t+1)}$.