Derivation of the EM Algorithm

Gabriel Peyré

March 14, 2024

1 Variational reformulation of $-\log \sum$

For any vector u and for any probability vector p, one has thanks to Jensen inequality, since $-\log$ is convex

$$-\log(\sum_k u_k) = -\log(\sum_k p_k \frac{u_k}{p_k}) \leq -\sum_k p_k \log(\frac{u_k}{p_k}).$$

But actually, if one used the best $p = p^\star(u)$, one has an equality

$$-\log(\sum_k u_k) = \min_{p \geq 0, \sum_k p_k = 1} -\sum_k p_k \log(\frac{u_k}{p_k}) = \mathrm{KL}(p|u).$$

Indeed, this optimal $p^\star(u)$ is

$$p^\star(u) = \frac{u}{\sum_k u_k}.$$

2 MLE of mixtures reformulation

MLE problem minimizes the negative log-likelihood of a mixture

$$\min_{\theta,\pi} \mathcal{L}(\theta,\pi) := \sum_{i=1}^n -\log\left(\sum_{k=1}^K \pi_k f(x_i|\theta_k)\right) \tag{1}$$

We introduce probability weights $P_{i,\cdot}$ for each i, and using the variational formulation of $-\log \sum$ to obtain

$$\mathcal{L}(\theta,\pi) = \min_P \mathcal{G}(\theta,\pi,P) := -\sum_{i,k} P_{i,k} \log\left(\frac{\pi_k}{P_{i,k}} f(x_i|\theta_k)\right) = \mathrm{KL}(P|\tilde{P}),$$
$$\text{where} \quad \tilde{P}_{i,k} := \pi_k f(x_i|\theta_k).$$

The EM algorithm is an alternate minimization on the variables of the problem

$$\min_{P,\theta,\pi} \mathcal{G}(\theta,\pi,P)$$

This guarantees that $\mathcal{L}(\theta)$ is decaying through the iterations and if f is smooth and the functional is coercive (which is problematic for Gaussians!) then converging sub-sequences are guaranteed to converge to a stationary point.

E step. The E steps correspond, given the previous iterate θ, to minimizing with respect to P

$$\min_{P \in \mathbb{R}_+^{n\times K}} \{\mathcal{G}(\theta,\pi,P) = \mathrm{KL}(P|\tilde{P}) : \sum_k P_{i,k} = 1\} \quad \text{where} \quad \tilde{P}_{i,k} := \pi_k f(x_i|\theta_k),$$

which solution reads

$$P_{i,k} = \frac{\tilde{P}_{i,k}}{\sum_k \tilde{P}_{i,k}}.$$

M step. Then the M step corresponds to minimizing

$$\min_{\theta,\pi} \mathcal{G}(\theta,\pi,P)$$

For π, one solves

$$\min_{\pi} \{\sum_k \sum_{i=1}^n P_{i,k} \log(\pi_k/P_{i,k}) : \sum_k \pi_k = 1\}$$

which solution is

$$\pi_k = \frac{\sum_i P_{i,k}}{\sum_{i,\ell} P_{i,\ell}}$$

For θ, this splits independently over each k as a usual (non-mixtures) MLE where the points are weights by $P_{i,k}$

$$\min_{\theta_k} -\sum_k P_{i,k} \log(f(x_i|\theta_k)).$$

Gaussian case. In the Gaussian case, where

$$f(x|\Sigma,m) := \frac{1}{\sqrt{2\pi \det(\Sigma)}} \exp\left(-\frac{\langle \Sigma^{-1}(x-m), x-m\rangle}{2}\right)$$

one has

$$m_k = \sum_i P_{i,k} x_i \in \mathbb{R}^d \quad \text{and} \quad \Sigma_k = \sum_i P_{i,k}(x_i - m_k)^\top (x_i - m_k) \in \mathbb{R}^{d\times d}.$$