Machine learning - Expectation Maximization

Yonghoon Dong

May 10, 2024

What is EM algorithm?

EM algorithm is a optimization algorithm designed to compute the MLE or MAP parameter estimate for probability models that have **missing data** and/or hidden variables (sometimes we call them **latent variables**).

Notation

From now on, let x_n be the visible data(fully observed data), z be the hidden data(latent variable).

EM algorithm procedure

EM algorithm has two main procedures which are E step (expectation step) and M step (maximization step) respectively.

1. E step : estimating the hidden variables (or missing values) during this step
2. M step : using the fully observed data to compute the MLE/MAP during this step

Why we have to use EM algorithm?

As we have seen so far, our goal is to maximize the log-likelihood of the observed data

$$\ell(\theta) = \sum_{n=1}^{N} \log p(x_n|\theta) \tag{1}$$

where ℓ is a log-likelihood

Caution

The Expectation-Maximization (EM) algorithm is typically used in unsupervised learning scenarios, particularly for estimating parameters of probabilistic models when you have missing data.

Why we have to use EM algorithm?

However when we use EM algorithm we have hidden variables. So it is impossible to optimize our log-likelihood function directly.

$$\ell(\theta) = \sum_{n=1}^{N} \log \left[\sum_{z \in Z} p(x_n, z|\theta) \right] \tag{2}$$

where ℓ is a log-likelihood

Explanation of Equation 2

Note that $p(x_n|\theta) = \sum_{z \in Z} p(x_n, z|\theta)$ by marginalization.

ELBO (Evidence Lower bound)

Consider a set of arbitary distributions $Q_n(z)$ over each hidden variables z.

$$\ell(\theta) = \sum_{n=1}^{N} \log \left[\sum_{z \in Z} Q_n(z) \frac{p(x_n, z|\theta)}{Q_n(z)} \right] \tag{3}$$

$$= \sum_{n=1}^{N} \log \left[\mathbb{E}_Z \left[\frac{p(x_n, Z|\theta)}{Q_n(Z)} \right] \right] \tag{4}$$

$$\geq \sum_{n=1}^{N} \mathbb{E}_Z \left[\log \frac{p(x_n, Z|\theta)}{Q_n(Z)} \right] = \sum_{n=1}^{N} \sum_{z \in Z} Q_n(z) \log \frac{p(x_n, z|\theta)}{Q_n(z)} \tag{5}$$

Explanation of Inequality 5

Note that log is a concave function. The inequality 5 is derived from the Jensen's Inequality.

ELBO (Evidence Lower bound)

Evidence Lower bound (ELBO) is defined by

$$\text{ELBO}(x_n; Q_n, \theta) = \sum_{z \in Z} Q_n(z) \log \frac{p(x_n, z|\theta)}{Q_n(z)} \tag{6}$$

Therefore,

$$\ell(\theta) \geq \sum_{n=1}^{N} \text{ELBO}(x_n; Q_n, \theta) \tag{7}$$

for all Q_n, θ, x_n.

Note

Optimizing the ELBO instead of log-likelihood is the basis of variational inference, which we will discuss when we study Variational autoencoder(VAE).

Expectation step (E step)

Consider the following equations

$$\text{ELBO}(x_n; Q_n, \theta) = \sum_{z \in Z} Q_n(z) \log \frac{p(x_n, z|\theta)}{Q_n(z)} \tag{8}$$

$$= \sum_{z \in Z} Q_n(z) \log \frac{p(z|x_n, \theta) p(x_n|\theta)}{Q_n(z)} \tag{9}$$

$$= \sum_{z \in Z} Q_n(z) \log p(x_n|\theta) + \sum_{z \in Z} Q_n(z) \log \frac{p(z|x_n, \theta)}{Q_n(z)} \tag{10}$$

$$= \sum_{z \in Z} Q_n(z) \log p(x_n|\theta) - \sum_{z \in Z} Q_n(z) \log \frac{Q_n(z)}{p(z|x_n, \theta)} \tag{11}$$

$$= \log p(x_n|\theta) - D_{KL}(Q_n(z)||p(z|x_n, \theta)) \tag{12}$$

Expectation step (E step)

Recall the following property of KL divergence.

Theorem (Non-negativity of KL divergence)

Let $A = \{x : p(x) > 0\}$ *be the support of* $p(x)$.

$$D_{KL}(p\|q) \geq 0 \text{ with equality iff } p = q$$

So we can maximize the ELBO with respect to Q_n. As you might guess, the optimal Q_n is

$$Q_n^*(z) = p(z|x_n, \theta) \tag{13}$$

Caution

Not to confuse that the variables involved in the optimization problem being solved in the E step are denoted as Q_n. In other words, it's about finding the optimal distribution among the selectable Q_n.

Expectation step (E step)

Suppose $Q_n^*(z) = p(z|x_n, \theta)$ for all $n \in \{1, \ldots, N)$. In other words, by non-negativity of KL divergence,

$$D_{KL}(Q_n^*(z)||p(z|x_n, \theta)) = 0 \tag{14}$$

Therefore,

$$\begin{aligned} \text{ELBO}(x_n; Q_n^*, \theta) &= \log p(x_n|\theta) - D_{KL}(Q_n^*(z)||p(z|x_n, \theta)) && (15) \\ &= \log p(x_n|\theta) && (16) \end{aligned}$$

for all $n \in \{1, \ldots, N\}$

Expectation step (E step)

Recall that

$$\ell(\theta) = \sum_{n=1}^{N} \log p(x_n|\theta) \tag{17}$$

$$\geq \sum_{n=1}^{N} \text{ELBO}(x_n; Q_n, \theta) \tag{18}$$

for arbitrary Q_n

Expectation step (E step)

Take $Q_n = Q_n^*$ (This is possible because above inequality holds for arbitrary Q_n). Then

$$\ell(\theta) = \sum_{n=1}^{N} \log p(x_n|\theta) \tag{19}$$

$$\geq \sum_{n=1}^{N} \text{ELBO}(x_n; Q_n^*, \theta) \tag{20}$$

$$\geq \sum_{n=1}^{N} \log p(x_n|\theta) \tag{21}$$

Therefore, $\ell(\theta) = \sum_{n=1}^{N} \text{ELBO}(x_n; Q_n^*, \theta)$. So, by maximize ELBO with respect to Q_n, we can get a **tight lower bound** of our log-likelihood.

Maximization step (M step)

In the M step at iteration t, we need to maximize $\text{ELBO}(x_n; Q^*_{n,t}, \theta_t)$ with respect to θ_t for all $n \in \{1, \ldots, N\}$.

Caution

$Q^*_{n,t}$ are the distributions computed in the E step at iteration t that corresponds to n'th observed data (i.e. x_n).

Maximization step (M step)

Therefore, the objective in M step at iteration t is

$$\theta_{t+1} := \underset{\theta_t}{maximize} \sum_{n=1}^{N} \text{ELBO}(x_n; Q^*_{n,t}, \theta_t) \tag{22}$$

In other words,

$$\underset{\theta_t}{maximize} \sum_{n=1}^{N} \sum_{z \in Z} \left(Q^*_{n,t}(z) \log \frac{p(x_n, z|\theta_t)}{Q^*_{n,t}(z)} \right) \tag{23}$$

$$\Leftrightarrow \underset{\theta_t}{maximize} \sum_{n=1}^{N} \sum_{z \in Z} \left(Q^*_{n,t}(z) \log p(x_n, z|\theta_t) + \mathbb{H}(Q^*_{n,t}) \right) \tag{24}$$

where $\mathbb{H}(Q^*_{n,t})$ is a entorpy of $Q^*_{n,t}$

Maximization step (M step)

Since $\mathbb{H}(Q^*_{n,t})$ is constant with respect to θ_t, we can drop them in the M step. Therefore, the objective in M step at iteration t is

$$\theta_{t+1} := \underset{\theta_t}{maximize} \sum_{n=1}^{N} \sum_{z \in Z} Q^*_{n,t}(z) \log p(x_n, z|\theta_t) \tag{25}$$

Visualization of EM algorithm

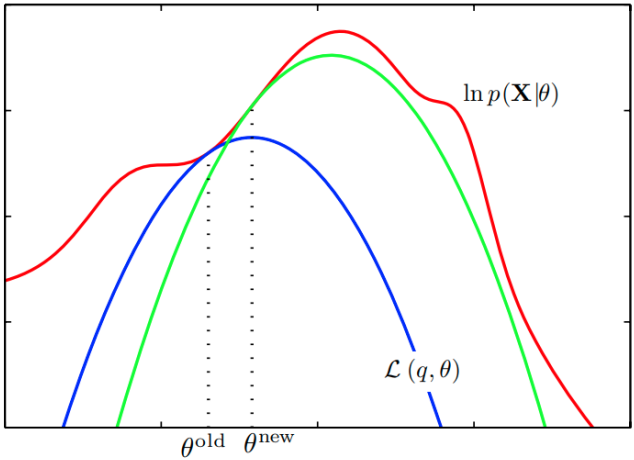

Figure: Visualization of EM algorithm