Expectation Maximization Algorithm

Introduction to Artificial Intelligence with Mathematics
Lecture Notes

Ganguk Hwang

Department of Mathematical Sciences
KAIST

Expectation Maximization Algorithm

The expectation–maximization (EM) algorithm is an *iterative* method to find the maximum likelihood or maximum a posteriori (MAP) estimates of the parameters in statistical models, where the model depends on unobserved latent variables.

The EM iteration alternates between performing an expectation step and a maximization step.

- Expectation (E) step: it derives a function for the expectation of the log-likelihood function evaluated using the distribution of the latent variables based on the current estimate for the parameters, and
- Maximization (M) step: it computes the values of the parameters maximizing the expected log-likelihood function obtained in the E step. The estimated parameters are then used to determine the distribution of the latent variables in the next E step.

The MLE based EM Algorithm

To explain the MLE based Expectation-Maximization (EM) algorithm, we consider a random variable of which distribution has parameter θ.

$$X \sim p(X;\theta).$$

We start with introducing a latent random variable, say, ϕ, that satisfies

$$p(X;\theta) = \int p(X,\phi;\theta)\ d\phi.$$

From Bayes' rule we have

$$p(X;\theta)p(\phi|X;\theta) = p(X,\phi;\theta).$$

Taking logarithm yields

$$\log p(X;\theta) = \log p(X,\phi;\theta) - \log p(\phi|X;\theta).$$

Let $\phi \sim q(\phi)$. Multiplying $q(\phi)$ on both sides and integraing both sides we obtain

$$\int q(\phi) \log p(X;\theta)\ d\phi = \int q(\phi) \log p(X,\phi;\theta)\ d\phi - \int q(\phi) \log p(\phi|X;\theta)\ d\phi$$

$$\log p(X;\theta) = \int q(\phi) \log p(X,\phi;\theta)\ d\phi - \int q(\phi) \log p(\phi|X;\theta)\ d\phi$$

Hence, we finally obtain

$$\begin{aligned}\log p(X;\theta) &= \int q(\phi) \log p(X,\phi;\theta)\ d\phi - \int q(\phi) \log q(\phi)\ d\phi \\ &\quad + \int q(\phi) \log q(\phi)\ d\phi - \int q(\phi) \log p(\phi|X;\theta)\ d\phi \\ &= \int q(\phi) \log \frac{p(X,\phi;\theta)}{q(\phi)}\ d\phi + \int q(\phi) \log \frac{q(\phi)}{p(\phi|X;\theta)}\ d\phi\end{aligned}$$

Let

$$\mathcal{L}(\theta) = \int q(\phi) \log \frac{p(X, \phi; \theta)}{q(\phi)} \, d\phi.$$

Then, the previous equation is rewritten by

$$\log p(X; \theta) = \mathcal{L}(\theta) + \mathsf{KL}(q(\phi) || p(\phi | X; \theta)).$$

Since $\mathsf{KL}(q(\phi) || p(\phi | X; \theta)) \geq 0$, we have

$$\log p(X; \theta) \geq \mathcal{L}(\theta).$$

Moreover, if $q(\phi) = p(\phi | X; \theta)$, then $\mathsf{KL}(q(\phi) || p(\phi | X; \theta)) = 0$ and hence

$$\log p(X; \theta) = \mathcal{L}(\theta) = \int q(\phi) \log \frac{p(X, \phi; \theta)}{q(\phi)} \, d\phi.$$

The MLE based EM Algorithm

For $t = 1, 2, \cdots$

Step 1. (E-step)

Set $q_t(\phi) = p(\phi|X; \theta_{t-1})$ and compute

$$\mathcal{L}_t(\theta) = \int q_t(\phi) \log p(X, \phi; \theta)\ d\phi - \int q_t(\phi) \log q_t(\phi)\ d\phi.$$

Step 2.(M-step)

Consider $\mathcal{L}_t(\theta)$ as a function of θ and find the optimal value θ_t that maximizes

$$\theta_t = \mathrm{argmax}_\theta\, \mathcal{L}_t(\theta).$$

Note that $\int q_t(\phi) \log q_t(\phi)\ d\phi$ is a constant with respect to θ.

Analysis of The Algorithm

With the algorithm we see that $\log p(X;\theta_{t-1}) \le \log p(X;\theta_t)$ which is shown as follows:

$$\begin{aligned}
\log p(X;\theta_{t-1}) &= \mathcal{L}_t(\theta_{t-1}) + \mathsf{KL}(q_t(\phi)||p(\phi|X;\theta_{t-1})) \\
&= \mathcal{L}_t(\theta_{t-1}) \\
&\le \mathcal{L}_t(\theta_t) \\
&\le \mathcal{L}_t(\theta_t) + \mathsf{KL}(q_t(\phi)||p(\phi|X;\theta_t)) \\
&= \log p(X;\theta_t).
\end{aligned}$$

To apply the EM algorithm, it is required to know $p(\phi|X;\theta)$ explicitly. While $p(\phi|X;\theta)$ is in general much easier to infer than $p(X;\theta)$, in many interesting problems this is not possible and thus the EM algorithm is not applicable.

The MAP based EM Algorithm

To explain the MAP based Expectation-Maximization (EM) algorithm, we consider the following problem.

$$X \sim p(X|\theta), \qquad \theta \sim p(\theta).$$

We start with introducing a latent random variable, say, ϕ, that satisfies

$$p(X, \theta) = \int p(X, \theta, \phi)\ d\phi.$$

From Bayes' rule we have

$$p(X, \theta) p(\phi|X, \theta) = p(X, \theta, \phi).$$

Taking logarithm yields

$$\log p(X, \theta) = \log p(X, \theta, \phi) - \log p(\phi|X, \theta).$$

Let $\phi \sim q(\phi)$. Multiplying $q(\phi)$ on both sides and integraing both sides we obtain

$$\int q(\phi) \log p(X, \theta) \ d\phi = \int q(\phi) \log p(X, \theta, \phi) \ d\phi - \int q(\phi) \log p(\phi | X, \theta) \ d\phi$$

$$\log p(X, \theta) = \int q(\phi) \log p(X, \theta, \phi) \ d\phi - \int q(\phi) \log p(\phi | X, \theta) \ d\phi$$

Hence, we finally obtain

$$\begin{aligned} \log p(X, \theta) &= \int q(\phi) \log p(X, \theta, \phi) \ d\phi - \int q(\phi) \log q(\phi) \ d\phi \\ &\quad + \int q(\phi) \log q(\phi) \ d\phi - \int q(\phi) \log p(\phi | X, \theta) \ d\phi \\ &= \int q(\phi) \log \frac{p(X, \theta, \phi)}{q(\phi)} \ d\phi + \int q(\phi) \log \frac{q(\phi)}{p(\phi | X, \theta)} \ d\phi \end{aligned}$$

Let

$$\mathcal{L}(\theta) = \int q(\phi) \log \frac{p(X, \theta, \phi)}{q(\phi)} \, d\phi.$$

Then, the previous equation is rewritten by

$$\log p(X, \theta) = \mathcal{L}(\theta) + \mathsf{KL}(q(\phi) || p(\phi | X, \theta)).$$

Since $\mathsf{KL}(q(\phi) || p(\phi | X, \theta)) \geq 0$, we have

$$\log p(X, \theta) \geq \mathcal{L}(\theta).$$

Moreover, if $q(\phi) = p(\phi | X, \theta)$, then $\mathsf{KL}(q(\phi) || p(\phi | X, \theta)) = 0$ and hence

$$\log p(X, \theta) = \mathcal{L}(\theta) = \int q(\phi) \log \frac{p(X, \theta, \phi)}{q(\phi)} \, d\phi.$$

The MAP based EM Algorithm

For $t = 1, 2, \cdots$

Step 1. (E-step)

Set $q_t(\phi) = p(\phi|X, \theta_{t-1})$ and compute

$$\mathcal{L}_t(\theta) = \int q_t(\phi) \log p(X, \theta, \phi) \; d\phi - \int q_t(\phi) \log q_t(\phi) \; d\phi.$$

Step 2.(M-step)

Consider $\mathcal{L}_t(\theta)$ as a function of θ and find the optimal value θ_t that maximizes

$$\theta_t = \mathrm{argmax}_\theta \, \mathcal{L}_t(\theta).$$

Note that $\int q_t(\phi) \log q_t(\phi) \; d\phi$ is a constant with respect to θ.

Analysis of the Algorithm

With the algorithm we see that $\log p(X, \theta_{t-1}) \leq \log p(X, \theta_t)$ which is shown as follows:

$$
\begin{aligned}
\log p(X, \theta_{t-1}) &= \mathcal{L}_t(\theta_{t-1}) + \mathsf{KL}(q_t(\phi) || p(\phi | X, \theta_{t-1})) \\
&= \mathcal{L}_t(\theta_{t-1}) \\
&\leq \mathcal{L}_t(\theta_t) \\
&\leq \mathcal{L}_t(\theta_t) + \mathsf{KL}(q_t(\phi) || p(\phi | X, \theta_t)) \\
&= \log p(X, \theta_t).
\end{aligned}
$$

Variational EM Algorithm

In Variational Expectation Maximization, we approximate the posterior probability with a simple model that comes from the mean field approximation. That is, we assume that latent variables are independent, so that their joint pdf is given by

$$q(\phi) = \prod_i q(\phi_i).$$

Even though we use the independent approximation, it allows us to update the pdf of each latent variable separately and has been successful in many interesting problems.

$$
\begin{aligned}
\mathcal{L}(\theta) &= \int q(\phi) \log \left(\frac{p(X, \phi; \theta)}{q(\phi)} \right) \, d\phi \\
&= \int \prod_i q(\phi_i) \log p(X, \phi; \theta) \, d\phi - \sum_i \int q(\phi_i) \log q(\phi_i) \, d\phi_i \\
&= \int q(\phi_j) \int \left(\prod_{i \neq j} q(\phi_i) \log p(X, \phi; \theta) \right) \prod_{i \neq j} d\phi_i d\phi_j \\
&\qquad - \int q(\phi_j) \log q(\phi_j) \, d\phi_j - \sum_{i \neq j} \int q(\phi_i) \log q(\phi_i) \, d\phi_i \\
&= \int q(\phi_j) \log \left(\frac{\exp E[\log p(X, \phi; \theta)]_{i \neq j}}{q(\phi_j)} \right) \, d\phi_j \\
&\qquad - \sum_{i \neq j} \int q(\phi_i) \log q(\phi_i) \, d\phi_i
\end{aligned}
$$

$$
\begin{aligned}
&= \int q(\phi_j) \log \left(\frac{\tilde{p}_{i \neq j}}{q(\phi_j)} \right) - \sum_{i \neq j} \int q(\phi_i) \log q(\phi_i) \; d\phi_i + c \\
&= -\mathsf{KL}(q(\phi_j) || \tilde{p}_{i \neq j}) - \sum_{i \neq j} \int q(\phi_i) \log q(\phi_i) \; d\phi_i + c.
\end{aligned}
$$

Here, since $\exp E[\log p(X, \phi; \theta)]_{i \neq j}$ is not a proper pdf, the constant c is added.

Since $\mathsf{KL}(\cdot||\cdot) \geq 0$, $\mathcal{L}(\theta)$ is maximized when

$$
q(\phi_j) = \frac{1}{Z} \exp E[\log p(X, \phi; \theta)]_{i \neq j}.
$$

The Variational EM Algorithm

Step 1: E-step

Compute $q^*(\phi_j) = \frac{1}{Z}\exp(E[\log p(X,\phi;\theta)]_{i\neq j})$ and let

$$q^{new}(\phi) = \prod_i q^*(\phi_i)$$

Step 2: M-step

$$\theta^{new} = \mathrm{argmax}_\theta\, \mathcal{L}(\theta)$$

where

$$\mathcal{L}(\theta) = \int q^{new}(\phi)\log\left(\frac{p(x,\phi;\theta)}{q^{new}(\phi)}\right)\, d\phi$$