EM Algorithm: an informal tutorial

Tao Hu
Microsoft.com
tahu@microsoft.com

August 17, 2018

Abstract

1 Introduction

The expectation-maximization (EM) algorithm introduced by Dempster et al [1] in 1977 is a very general method to solve maximum likelihood estimation with hidden state problems. In this informal report, we review the theory behind EM.

2 Problem setting

Let Y a random variable with probability density function (pdf) $p(y|\theta)$, where θ is an unknown parameters vector. Let $y = y_{1..N}$ be N observations of Y, we aim at maximizing the likelihood function $p(y|\theta)$ or $\arg\min_\theta p(y|\theta)$. We can solve this by stochastic gradient descent (SGD) optimization. However if Y depends on both hidden variables Z and parameters θ such as in Gaussian mixture distribution, it becomes intractable to directly calculate $p(y|\theta) = \int_x p(y, x|\theta)dx$ because $x = x_{1..N}$, where N could be millions of samples. The EM algorithm iteratively solves the problem by assuming one of x and θ are known.

3 EM algorithm

3.1 Intuition

Since $\arg\min_\theta p(y|\theta) = \arg\min_\theta \ln p(y|\theta)$, we can maximize $\ln p(y|\theta)$ instead. For any distribution with pdf $q(z)$, the equation 1 is alway true.

$$
\begin{aligned}
\ln(p(y)) &= \int_z q(z) \ln p(y) dz && p(y) \text{ is constant given } z \\
&= \int_z q(z) \ln \frac{p(y,z)}{p(z|y)} dz && p(y) = p(y,z)/p(z|y) \\
&= \int_z q(z) \ln \frac{p(y,z)/q(z)}{p(z|y)/q(z)} dz && \text{both are divided by } q(z) \\
&= \int_z q(z) \ln \frac{p(y,z)}{q(z)} dz - \int_z q(z) \ln \frac{p(z|y)}{q(z)} dz && (1)
\end{aligned}
$$

Let evidence of low bound (ELBO) as

$$
\mathcal{L}(q, p(y,z)) = \int_z q(z) \ln \frac{p(y,z)}{q(z)} dz \tag{2}
$$

Since KL divergence $\mathcal{KL}(q \parallel p(z|y)) = -\int_z q(z) \ln(\frac{p(z|y)}{q(z)}) dz$, so equation 1 can be written as

$$
\ln p(y) = \mathcal{L}(q, p(y,z)) + \mathcal{KL}(q \parallel p(z|y)) \tag{3}
$$

From equation 3, we have two conclusions. First, since y are observations and $\ln p(y)$ is constant, minimizing KL divergence (alway positive) between any distribution $q(z)$ and posterior distribution $p(z|y)$ is equal to maximize ELBO. Second, the best distribution $q(z)$ approximating $p(z|y)$ is the posterior itself, where KL divergence is 0 and ELBO $\mathcal{L}$ has the maximum value $\ln p(y)$.

Introducing parameters θ, from equation 3, we have

$$
\ln p(y|\theta) = \mathcal{L}(q(z), p(y,z|\theta)) + \mathcal{KL}(q \parallel p(z|y,\theta)) \tag{4}
$$

If parameters θ are fixed, the best $q(z)$ is $p(z|y,\theta)$. This is the key idea in the E-step of EM algorithm.

If $q(z)$ is close enough to the true posterior $p(z|y,\theta)$, then KL divergence is close to 0. To find $\arg\min_\theta \ln p(y|\theta)$, we can approximately find the best θ to maximize $\mathcal{L}$ instead. In ideal case, if $q(z)$ is true posterior $p(z|y,\theta)$, then $\mathcal{L}$ is the same as $\ln p(y|\theta$. This is the key idea in the M-step of EM algorithm.

If we combine those two steps iteratively, we get the EM algorithm.

1. E-Step: given θ^t at time t, let $q^t(z)$ equal to $p(z|y,\theta^t)$
2. M-Step: given $q^t(z)$ at E-Step, find $\theta^{t+1} = \arg\max_\theta \mathcal{L}(q^t(z), p(y,z|\theta))$

3.2 Proof

To prove the EM algorithm will converge to the local minimum, we need to show $\ln p(y|\theta^{t+1})$ is no less than $\ln p(y|\theta^t)$.

Lemma 3.1. $\ln p(y|\theta^{t+1}) >= \ln p(y|\theta^t)$

Proof.

$$
\begin{aligned}
\ln p(y|\theta^{t+1}) &= \mathcal{L}(q^t(z), p(y,z|\theta^{t+1})) + \mathcal{KL}(q^t(z) \parallel p(z|y,\theta^{t+1})) && \text{according to 4} \\
&\geq \mathcal{L}(q^t(z), p(y,z|\theta^{t+1})) && \text{since } \mathcal{KL}(q^t \parallel p(z|y,\theta^{t+1})) \geq 0 \\
&\geq \mathcal{L}(q^t(z), p(y,z|\theta^{t})) && \text{in M-step, } \theta^{t+1} \text{ is the best} \\
&= \mathcal{L}(q^t(z), p(y,z|\theta^{t})) + \mathcal{KL}(q^t(z) \parallel p(z|y,\theta^{t})) && \text{in E-Step, } q^t(z) = p(z|y,\theta^t) \\
&= \ln p(y|\theta^t) && \text{according to 4}
\end{aligned}
$$

□

3.3 Improve

Notice in M-step, to maximize $\mathcal{L}$, we don't need to calculate its own entropy. We can further simplify it.

Let $Q(\theta^t, \theta) = \int_z q^t(z) \ln p(y, z|\theta) dz$, and $H(q(z)) = -\int_z q(z) \ln q(z)$. We have

$$
\begin{aligned}
\mathcal{L}(q^t(z), p(y,z|\theta)) &= \int_z q^t(z) \ln p(y,z|\theta) dz + -\int_z q^t(z) \ln q^t(z) dz \\
&= Q(\theta^t, \theta) + H(q^t(z))
\end{aligned} \tag{5}
$$

Since entropy $H(q^t(z))$ doesn't depend on θ, we have $\arg\min_\theta \mathcal{L}(q^t(z), p(y,z|\theta)) = \arg\min_\theta Q(\theta^t, \theta))$. We can maximize $Q(\theta^t, \theta)$ instead of ELBO $\mathcal{L}$ in M-step.

3.4 Practical assumptions

For large N samples, where $y = y_{1..N}$ and $z = z_{1..N}$, without any constraint on joint distribution $p(y_{1..N}, z_{1..N}|\theta)$, it is not easy to calculate $p(z_{1..N}|y_{1..N}|\theta)$ and $Q(\theta^t, \theta)$. Such restrictions on the joint distribution are generally presented by probabilistic graphical models. One of common restriction is assuming each y_i is identical independent distribution (iid) conditional on its corresponding latent variable z_i.

With conditional idd assumption, we can simplify E-step as the following

$$
\begin{aligned}
p(z_{1..N}|y_{1..N}, \theta) &= p(y_{1..N}, z_{1..N}|\theta)/p(y_{1..N}|\theta) \\
&\propto \prod_{i=1}^{N} p(z_i, y_i|\theta) \\
&= \prod_{i=1}^{N} p(z_i|y_i|\theta) * p(y_i|\theta) \\
&\propto \prod_{i=1}^{N} p(z_i|y_i, \theta)
\end{aligned} \tag{6}
$$

In equation 6, we know normalizer must be 1 when integrating both sides over $z_{1..N}$. Thus, we have $p(z_{1..N}|y_{1..N}, \theta) = \prod_{i=1}^{N} p(z_i|y_i, \theta)$. So in E-step, we just need to calculate each observation marginal distribution $p(z_i|y_i, \theta)$, which dramatically simplify the calculation.

The equation 7 says instead of calculating expectation by join distribution, we can use marginal distribution to calculate expectation under decomposition.

$$
\begin{aligned}
\int_{z_1}\int_{z_2} q(z_1, z_2) * (f_1(z_1) + f_2(z_2)) dz_1 dz_2 &= \int_{z_1}\int_{z_2} q(z_1, z_2) * f(z_1) dz_1 dz_2 + \int_{z_1}\int_{z_2} q(z_1, z_2) * f_2(z_2) dz_1 dz_2 \\
&= \int_{z_1} f(z_1)(\int_{z_2} q(z_1, z_2) dz_2) dz_1 + \int_{z_2} f_2(z_2)(\int_{z_1} q(z_1, z_2) dz_1) dz_2 \\
&= \int_{z_1} q(z_1) * f_1(z_1) dz_1 + \int_{z_2} q(z_2) * f_2(z_2) dz_2
\end{aligned}
\tag{7}
$$

With equation 7, we can see how to simplify in calculation $Q(\theta^t, \theta)$.

$$
\begin{aligned}
Q(\theta^t, \theta) &= \int_z q^t(z) \ln p(y, z|\theta) dz \\
&= \int_z q^t(z) \ln(\prod_{i=1}^{N} p(y_i, z_i|\theta)) dz \\
&= \int_z q^t(z) \sum_{i=1}^{N} \ln p(y_i, z_i|\theta) dz \\
&= \int_{z_1} \dots \int_{z_N} q^t(z_{1..N}) \sum_{i=1}^{N} \ln p(y_i, z_i|\theta) dz_1 \dots dz_N \\
&= \sum_{i=1}^{N} \int_{z_i} q^t(z_i) \ln p(y_i, z_i|\theta) dz
\end{aligned}
\tag{8}
$$

The last step is by repeatedly applying equation 7. We also notice $q^t(z) = p(z_i|y_i, \theta^t)$. By assuming each observation (with its corresponding latent variables) is idd, we can dramatically simplifying computation.

4 Gaussian mixture example

References

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1 (1977), pp. 1–38. ISSN: 00359246. URL: http://www.jstor.org/stable/2984875.