The EM algorithm and variational inference

Aodong Li

February 3, 2018

This document only serves as personal thoughts or notes.

In this notes the derivation of the Expectation Maximization (EM) algorithm will be provided. The derivation utilizes Jensen's inequality for convex functions, which is also absorbed in variational methods for inference. These two algorithms both provide the lower bound of objective function.

Both them use the similar derivation. Is there any connection between them? Sure! The relation will be made explicit in the notes.

1 The EM Algorithm

The EM algorithm is an iterative method to find the maximum likelihood estimator. The algorithm utilizes the variational method by introducing hidden variables no matter whether they exist or not. So if the model does involve hidden variables, the EM algorithm naturally fit in.

There are two steps in each iteration: E-step and M-step. In E-step, the hidden variables are estimated given the observed ones and current estimated model parameters. In M-step, the likelihood function is maximized given current hidden variables.

Assume after n iterations, we have θ_n. We want to maximize the difference between two log-likelihoods,

$$\begin{aligned}L(\theta) - L(\theta_n) &= \ln P(X|\theta) - \ln P(X|\theta_n)\\ &= \ln \sum_z P(X|z,\theta)P(z|\theta) - \ln P(X|\theta_n).\end{aligned}$$

Take advantage of $P(z|X,\theta_n)$ and Jensen's inequality,

$$\begin{aligned}L(\theta) - L(\theta_n) &= \ln \sum_z P(z|X,\theta_n)\frac{P(X|z,\theta)P(z|\theta)}{P(z|X,\theta_n)} - \ln P(X|\theta_n)\\ &\geq \sum_z P(z|X,\theta_n) \ln \frac{P(X|z,\theta)P(z|\theta)}{P(z|X,\theta_n)} - \ln P(X|\theta_n)\\ &= \sum_z P(z|X,\theta_n) \ln \frac{P(X|z,\theta)P(z|\theta)}{P(z|X,\theta_n)P(X|\theta_n)}\end{aligned}$$

If $\theta = \theta_n$, this gives $L(\theta) - L(\theta_n) = 0 \geq 0$. So $\sum_z P(z|X, \theta_n) \ln \frac{P(X|z,\theta)P(z|\theta)}{P(z|X,\theta_n)P(X|\theta_n)} + L(\theta_n)$ is the lower bound of $L(\theta)$. If we try to maximize the lower bound, this also maximizes the objective log-likelihood function.

$$\begin{aligned}
\theta_{n+1} &= \max_\theta \arg \sum_z P(z|X, \theta_n) \ln \frac{P(X|z,\theta)P(z|\theta)}{P(z|X,\theta_n)P(X|\theta_n)} + L(\theta_n) \\
&= \max_\theta \arg \sum_z P(z|X, \theta_n) \ln P(X, z|\theta) \\
&= \max_\theta \arg \mathbb{E}_{z|X,\theta_n} \ln P(X, z|\theta).
\end{aligned}$$

So this gives rise to two steps. The E-step calculates the conditional expectation $\mathbb{E}_{z|X,\theta_n} \ln P(X, z|\theta)$ and the M-step maximizes the expectation.

Since every iteration we have $L(\theta_{n+1}) \geq L(\theta_n)$, the EM algorithm converges.

2 Variational Inference

Variational methods used in inference usually contain two main methods: sequential methods and block methods. The sequential methods transform the local conditional probability one by one (by introducing variational parameters) until the remaining graph is tractable with exact inference. On the other hand, the block methods first observe the tractable components or subgraphs in the original graph. Then use the set of approximate distributions to model the original distribution. This can be seen as an offline version of the sequential methods[JG99]. Note that the approximate distribution is parameterized by variational parameters.

Nowadays, the block methods are utilized more widely than the sequential methods due to the development of machine learning. We can directly model the approximate distributions by machine learning techniques and do not need to transform local distributions one by one to make sure to reach a tractable graph tightly.

For block methods, we replace the original inference graph (representing the conditional distribution $P(H|E)$) of interest with an approximate sub-graph (representing the approximate distribution $Q(H|E, \lambda)$), which is parameterized by variational variables λ. If there is obvious tractable structure, we can make a decision manually. If there is not, we can choose a sub-graph specified by λ implicitly by the introduction of machine learning techniques.

Formally, λ is chosen to minimize the KL divergence between Q and P. This operation is justified when looking at the probability of evidence. It means the choice of λ gives the tightest lower bound of the likelihood.

$$\begin{aligned}
\ln P(X) &= \ln \sum_z P(z, X) \\
&= \ln \sum_z Q(z|X) \frac{P(z, X)}{Q(z|X)} \\
&\geq \sum_z Q(z|X) \ln \frac{P(z, X)}{Q(z|X)}
\end{aligned}$$

It turns out that $\ln P(X) - \sum_z Q(z|X) \ln \frac{P(z,X)}{Q(z|X)} = KL(Q(z|X)||P(z|X))$. So when we minimize KL divergence, we are actually reaching the best lower bound of the probability of evidence.

3 The Relation Between Them

When variational methods used in parameter estimation, notice that the Evidence Lower Bound (ELBO) is a function of the approximate distribution Q and the unknown parameter θ,

$$ELBO(Q, \theta) = \sum_z Q(z|X) \ln \frac{P(z, X|\theta)}{Q(z|X)},$$

when $Q(z|X) = P(z|X)$, the ELBO restores the original form of $P(X|\theta)$.

This form encourages coordinate ascent as an optimization method.

$$\text{Step 1: } Q^{(k+1)} = \arg\max_Q ELBO(Q, \theta^{(k)})$$
$$\text{Step 2: } \theta^{(k+1)} = \arg\max_\theta ELBO(Q^{(k+1)}, \theta)$$

Now it can be seen the relation between the EM algorithm and the variational inference.

Let's *fix* the approximate distribution $Q(z|X)$ to be $P(z|X, \theta^{(k)})$, i.e., we do not transform the graph via any variational variables. Then the Step 1 has nothing to do with the maximization over Q, which is fixed. The ELBO now is a function of θ. We can safely omit the terms in ELBO that don't contain θ in the sense of optimization,

$$EL\tilde{B}O(\theta) = \sum_z P(z|X, \theta^{(k)}) \ln P(X, z|\theta).$$

At the end the Step 1 degenerates to compute the conditional expectation and the Step 2 do the same thing. This is essentially the traditional EM algorithm.

References

[JG99] MICHAEL I. JORDAN, ZOUBIN GHAHRAMANI, TOMMI S. JAAKKOLA and LAWRENCE K. SAUL, "An Introduction to Variational Methods for Graphical Models," *Machine Learning, 37*, 1999, pp. 183–233.