Expectation Maximization

David Andrzejewski
andrzeje@cs.wisc.edu

February 11, 2010

1 Introduction

Expectation Maximization is a very general algorithm for doing maximum likelihood estimation of parameters in models which contain latent variables.

1.1 Definitions

$\mathbf{x}$ are the observed variables, $\mathbf{y}$ are the latent variables, and θ is the set of model parameters.

1.2 Maximum Likelihood

We want to maximize the marginal log likelihood of the observed data $\mathbf{x}$.

$$\log(P(\mathbf{x}|\theta)) = \log(\sum_{\mathbf{y}} P(\mathbf{x},\mathbf{y}|\theta))$$

1.3 Auxiliary distribution

We add an auxililary distribution $q(\mathbf{y})$.

$$\begin{aligned}\log(P(\mathbf{x}|\theta)) &= \log(\sum_{\mathbf{y}} P(\mathbf{x},\mathbf{y}|\theta)) \\ &= \log(\sum_{\mathbf{y}} q(\mathbf{y})\frac{P(\mathbf{x},\mathbf{y}|\theta)}{q(\mathbf{y})})\end{aligned}$$

1.4 Jensen's Inequality

Jensen's Inequality states that for a convex function $f(x)$

$$E[f(x)] \geq f(E[x]),$$

which follows from the convexity of the epigraph of a convex function. The reversed inequality holds for a concave function (such as log), which we apply to lower bound the marginal log likelihood.

2 The EM Algorithm

2.1 E-step

We begin by applying Jensen's Inequality to lower bound the marginal log-likelihood $\log(P(\mathbf{x}|\theta))$. We then use the linearity of expectation and log identities to further re-arrange this lower bound.

$$\begin{aligned}
\log(P(\mathbf{x}|\theta)) &= \log(\sum_{\mathbf{y}} q(\mathbf{y})\frac{P(\mathbf{x},\mathbf{y}|\theta)}{q(\mathbf{y})}) \\
&\geq E_q[\log(\frac{P(\mathbf{x},\mathbf{y}|\theta)}{q(\mathbf{y})}] \\
&\geq E_q[\log(\frac{P(\mathbf{y}|\mathbf{x},\theta)P(\mathbf{x}|\theta)}{q(\mathbf{y})}] \\
&\geq E_q[\log(P(\mathbf{x}|\theta)] - E_q[\log(\frac{q(\mathbf{y})}{P(\mathbf{y}|\mathbf{x},\theta)}] \\
&\geq E_q[\log(P(\mathbf{x}|\theta))] - KL(q(\mathbf{y})\|P(\mathbf{y}|\mathbf{x},\theta)) \\
&\geq \log(P(\mathbf{x}|\theta)) - KL(q(\mathbf{y})\|P(\mathbf{y}|\mathbf{x},\theta))
\end{aligned}$$

The E-step consists of maximizing the lower bound with respect to $q(\mathbf{y})$. Since the first term does not depend on $q(\mathbf{y})$, this simply means minimizing the KL-divergence term. The Gibbs inequality states that KL-divergence is non-negative, and in fact is zero only for identical distributions. Therefore we maximize with respect to $q(\mathbf{y})$ by setting $q(\mathbf{y}) = P(\mathbf{y}|\mathbf{x},\theta)$. Note that our lower bound is actually equal to the log marginal likelihood for this $q(\mathbf{y})$. Any increase in this lower bound during the M-step therefore *must* increase the marginal log-likelihood.

2.2 M-step

The M-step then consists of maximizing this lower bound with respect to the parameters θ. Returning to our lower bound

$$\begin{aligned}
\log(P(\mathbf{x}|\theta)) &\geq E_q[\log(\frac{P(\mathbf{x},\mathbf{y}|\theta)}{q(\mathbf{y})}] \\
&\geq E_q[\log(P(\mathbf{x},\mathbf{y}|\theta)] - E_q[\log(q(\mathbf{y}))] \\
&\geq E_q[\log(P(\mathbf{x},\mathbf{y}|\theta)] + H(q(\mathbf{y}))
\end{aligned}$$

Here the second entropy term does not depend on θ, so we simply set $\theta = \arg\max_\theta E_q[\log(P(\mathbf{x}, \mathbf{y}|\theta)]$. That is, we wish to marginalize the expectation of the complete-data log-likelihood with respect to our auxiliary distribution $q(\mathbf{y})$.

Hopefully, $P(\mathbf{x}, \mathbf{y}|\theta)$ has been chosen "nicely" in that this expected complete-data log-likelihood is easy to optimize. For example, in the Baum-Welch algorithm for Hidden Markov Models, the inter-state transitions are modeled as exponential family distributions (multinomial), and their sufficient statistics are the transition counts. Therefore the model is log-linear in the counts, and we can therefore simply maximize the expected complete-data log-likelihood $\log(P(\mathbf{x}, \mathbf{y}|\theta))$ with respect to the *expected* transition counts under $q(\mathbf{y})$.

2.3 The Big Picture

Essentially, we are doing coordinate ascent optimization of the marginal log-likelihood $\log(P(\mathbf{x}|\theta))$. We achieve this by alternating between setting $q(\mathbf{y})$ to get a tight lower bound (E-step), and then setting θ to increase this bound (M-step). Since we are guaranteed to increase $\log(P(\mathbf{x}|\theta))$ every iteration, we will eventually reach a local optimum.

3 Variants

3.1 Generalized EM

Generalized EM relaxes the requirement that the new θ *maximize* $E_q[\log(P(\mathbf{x}, \mathbf{y}|\theta)]$ during the M-step. As long as the new θ still *increases* the lower bound, the algorithm will still converge to a local optimum.

3.2 Variational EM

Variational EM relaxes the requirement that $q(\mathbf{y}) = P(\mathbf{y}|\mathbf{x}, \theta)$ during the E-step. It may be that the true posterior is intractable (e.g., Latent Dirichlet Allocation), so we use a simplified family of $q(\mathbf{y})$ distributions (e.g., fully factorized $q(\mathbf{y}) = \prod_i q(y_i)$) to approximate the true posterior distribution. The E-step then consists of finding the $q(\mathbf{y})$ in this restricted family which minimizes the KL-divergence with the true posterior $KL(q(\mathbf{y})\|P(\mathbf{y}|\mathbf{x}, \theta))$.

3.3 MCMC EM

We can think of the "output" of the E-step as being a distribution $q(\mathbf{y}) = P(\mathbf{y}|\mathbf{x}, \theta)$ which can be used to calculate (and optimize) $E_q[\log(P(\mathbf{x}, \mathbf{y}|\theta)]$ during the M-step. Even if we cannot calculate $P(\mathbf{y}|\mathbf{x}, \theta)$, we may still be able to optimize $E_q[\log(P(\mathbf{x}, \mathbf{y}|\theta)]$ by using Markov Chain Monte Carlo sampling methods to approximate expectations of interest. For example, if we have a model which is log-linear in some statistics (e.g., a multinomial distribution), we can approximate the expected counts by averaging over samples taken from $P(\mathbf{y}|\mathbf{x}, \theta)$.