VB, EM and EP

Joris Mooij

December 16, 2013

1 VB

1.1 Inequality

Suppose $\boldsymbol{X}$ is an observed variable and $\boldsymbol{Z}$ a latent variable. Then we can write

$$\ln p(\boldsymbol{X}) = \mathcal{L}(q) + KL(q||p_{\boldsymbol{Z}|\boldsymbol{X}})$$

where

$$\mathcal{L}(q) = \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})}\, d\boldsymbol{Z}, \qquad KL(q||p_{\boldsymbol{Z}|\boldsymbol{X}}) = -\int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z} \mid \boldsymbol{X})}{q(\boldsymbol{Z})}\, d\boldsymbol{Z}$$

Then, because of the properties of the Kullback-Leibler divergence,

$$\ln p(\boldsymbol{X}) = \mathcal{L}(q) + KL(q||p_{\boldsymbol{Z}|\boldsymbol{X}}) \geq \mathcal{L}(q)$$

with equality if $q = p_{\boldsymbol{Z}|\boldsymbol{X}}$.

1.2 Variational Bayes

The VB approximation thus approximates the posterior $p_{\boldsymbol{Z}|\boldsymbol{X}}$ by

$$q^* = \arg\max_{q \in \mathcal{Q}} \mathcal{L}(q) = \arg\max_{q \in \mathcal{Q}} \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})}\, d\boldsymbol{Z}$$

and the evidence $\ln p(X) \approx \mathcal{L}(q^*)$.

As a special case, we take the family of distributions to factorize, i.e., we assume

$$q(\boldsymbol{Z}) = \prod_{i=1}^{M} q_i(\boldsymbol{Z}_i)$$

Then we get the VB update equation:

$$\begin{aligned} \ln q_i^*(\boldsymbol{Z}_i) &= \int \ln p(\boldsymbol{X}, \boldsymbol{Z}) q_{\setminus i}(\boldsymbol{Z})\, d\boldsymbol{Z}_{\setminus i} + \text{const.} \\ &= \mathbb{E}_{q_{\setminus i}} \ln p(\boldsymbol{X}, \boldsymbol{Z}) + \text{const.} \end{aligned} \tag{1}$$

where

$$q_{\backslash i}(\boldsymbol{Z}) = \prod_{\substack{j=1 \\ j \neq i}}^{M} q_j(\boldsymbol{Z}_j).$$

2 EM

In EM, we distinguish latent variables $\boldsymbol{\theta}$ over which we *optimize* from latent variables $\boldsymbol{Z}$ which we marginalize over. EM can be derived as a special case of VB, taking $q(\boldsymbol{Z}, \boldsymbol{\theta}) = q(\boldsymbol{Z})\delta(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$. However, this will lead to a lower bound on the evidence of $-\infty$ because of the delta function. So it pays off to treat the $\boldsymbol{\theta}$ variables in a slightly different way. We write

$$\ln p(\boldsymbol{X}, \boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q||p_{\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}})$$

where

$$\mathcal{L}(q, \boldsymbol{\theta}) = \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta})}{q(\boldsymbol{Z})} \, d\boldsymbol{Z}, \qquad KL(q||p_{\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}}) = -\int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta})}{q(\boldsymbol{Z})} \, d\boldsymbol{Z}$$

2.1 E-step

This is the maximization over $q(\boldsymbol{Z})$. The result depends on $\boldsymbol{\theta}$:

$$q^*_{\boldsymbol{\theta}}(\boldsymbol{Z}) = \arg\max_{q \in \mathcal{Q}} \mathcal{L}(q, \boldsymbol{\theta}) = \arg\max_{q \in \mathcal{Q}} \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta})}{q(\boldsymbol{Z})} \, d\boldsymbol{Z}$$

If we allow all possible distributions, then we simply obtain the posterior:

$$q^*_{\boldsymbol{\theta}}(\boldsymbol{Z}) = p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}).$$

2.2 M-step

This is the maximization over $\boldsymbol{\theta}$. The result depends on q:

$$\boldsymbol{\theta}^*_q = \arg\max_{\boldsymbol{\theta}} \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta})}{q(\boldsymbol{Z})} \, d\boldsymbol{Z}$$

If we take for $q = q^*_{\boldsymbol{\theta}}$ (the result of the standard E-step), then we get the standard formulation of the M-step:

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \int p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{\text{old}}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta})}{p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{\text{old}})} \, d\boldsymbol{Z}$$

2.3 Hybrid EM/VB

It should be obvious now how to derive hybrid versions of EM and VB. Simply take $\mathcal{Q}$ to be a strict subset of all possible distributions on $\boldsymbol{Z}$. Then the standard E-step will be replaced by the VB updates for $\boldsymbol{Z}$, and the M-step does not use the exact posterior $p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{\text{old}})$ but the current VB approximation $q(\boldsymbol{Z})$.

3 EP

Given again observed variables $\boldsymbol{X}$ and latents $\boldsymbol{Z}$, with joint probability

$$p(\boldsymbol{X}, \boldsymbol{Z}) = \prod_i f_i(\boldsymbol{Z}, \boldsymbol{X}).$$

We want to approximate the posterior

$$p(\boldsymbol{Z} \,|\, \boldsymbol{X}) = \frac{1}{p(\boldsymbol{X})} \prod_i f_i(\boldsymbol{Z}, \boldsymbol{X}) \approx \frac{1}{Z} \prod_i q_i(\boldsymbol{Z}) =: q(\boldsymbol{Z})$$

and model evidence

$$p(\boldsymbol{X}) = \int \prod_i f_i(\boldsymbol{Z}, \boldsymbol{X}) d\boldsymbol{Z} \approx \int \prod_i q_i(\boldsymbol{Z}) d\boldsymbol{Z}$$

The EP update for factor $q_i(\boldsymbol{Z})$ is obtained by considering:

$$q(\boldsymbol{Z})^{\text{new}} = \operatorname*{arg\,min}_{q \in \mathcal{Q}} KL\left(\frac{1}{Z_j} f_j(\boldsymbol{Z}, \boldsymbol{X}) q_{\setminus j}^{\text{old}}(\boldsymbol{Z}) \,\Big|\Big|\, q(\boldsymbol{Z}) \right)$$

where

$$q_{\setminus j}(\boldsymbol{Z}) := \prod_{i \neq j} q_i(\boldsymbol{Z}).$$

$$Z_j = \int f_j(\boldsymbol{Z}, \boldsymbol{X}) q^{\setminus j}(\boldsymbol{Z}) d\boldsymbol{Z}.$$

This optimization is easily done if all q_j are in an exponential family $\mathcal{Q}_j$, because then $\mathcal{Q}$ is also an exponential family. The new approximate factor $q_j(\boldsymbol{Z})$ is then given by:

$$q_j^{\text{new}}(\boldsymbol{Z}) = Z_j \frac{q^{\text{new}}(\boldsymbol{Z})}{q^{\setminus j}(\boldsymbol{Z})}$$

3.1 Moment matching

The solution is obtained by matching the moments of the sufficient statistics. Suppose

$$\mathcal{Q} = \{ h(\boldsymbol{Z}) g(\boldsymbol{\eta}) \exp\left(\boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{Z}) \right) : \boldsymbol{\eta} \}$$

Then

$$q^* = \operatorname*{arg\,min}_{q \in \mathcal{Q}} KL(p \,||\, q) \iff \mathbb{E}_{q^*}\big(\boldsymbol{u}(\boldsymbol{Z})\big) = \mathbb{E}_p\big(\boldsymbol{u}(\boldsymbol{Z})\big).$$