Variational Bayes

Lloyd Elliott

November 18, 2010

Variational EM: Free energy reminder

Suppose $\mathbf{x}, p(\mathbf{x}|\mathbf{y}, \theta)$ are given. Any distribution $q(\mathbf{y})$ on $\mathbf{y}$ induces a lower bound on $\ell(\theta)$:

$$\ell(\theta) = \log \int p(\mathbf{x}, \mathbf{y}|\theta) \mathrm{d}\mathbf{y} \tag{1}$$

$$\geq \int q(\mathbf{y}) \log \frac{p(\mathbf{x}, \mathbf{y}|\theta)}{q(\mathbf{y})} \mathrm{d}\mathbf{y} = \mathcal{F}(q, \theta) \tag{2}$$

EM reminder

`E-step`:

$$q(\mathbf{y}) \leftarrow p(\mathbf{y}|\mathbf{x}, \theta) \tag{3}$$

`M-step`:

$$\theta \leftarrow \underset{\theta}{\text{argmax}}\, \mathbb{E}_{q(\mathbf{y})}[\log p(\mathbf{x}, \mathbf{y}|\theta)] \tag{4}$$

Variational Bayes: motivation

Problem: If $q(\mathbf{y})$ were high dimensional then both steps are intractable. We can solve this problem with variational methods wherein $q(\mathbf{y})$ is restricted to some family Ω.

- Parameterise q and use gradient descent for E-step.
- Assume q factors over $\mathbf{y}$ (mean field approximation).

Such assumptions lead to a VBEM algorithm.

Variational Bayes: VE step

`VE-step`:

$$q(\mathbf{y}) \leftarrow \underset{q(\mathbf{y}) \in \Omega}{\operatorname{argmin}} \, \mathrm{KL}(q(\mathbf{y}) || p(\mathbf{y}|\mathbf{x}, \theta)) \tag{5}$$

The free energy $\mathcal{F}(q, \theta)$ still increases after each E-step and M-step. So, VBEM converges. But the likelihood may not increase so VBEM may not find modes of $\ell(\theta)$.

If we assume $q(\mathbf{y})$ factors over $y_1, \ldots, y_n$ then the resulting VBEM is called the mean field approximation.

$$q(\mathbf{y}) = \prod_i \prod_j q_{ij}(y_{ij}) \tag{6}$$

Thus,

$$\mathcal{F}(q, \theta) = \sum_i \int \prod_j q_{ij}(y_{ij}) \log p(\mathbf{x}_i, \mathbf{y}_i|\theta) - \prod_j q_{ij}(y_{ij}) \log \prod_j q_{ij}(y_{ij}) \mathrm{d}\mathbf{y}_i, \tag{7}$$

$$= \sum_i \int \prod_j q_{ij}(y_{ij}) \log p(\mathbf{x}_i, \mathbf{y}_i|\theta) - \sum_i q_{ij}(y_{ij}) \log q_{ij}(y_{ij}) \mathrm{d}\mathbf{y}_i. \tag{8}$$

We solve for:

$$\frac{\delta}{\delta q_{ij}(\mathbf{y}_{ij})}\mathcal{F}(q,\theta) = 0 \tag{9}$$

whilst enforcing the constraints $\int q_{ij}(\mathbf{y}_{ij})\mathrm{d}\mathbf{y}_{ij} = 1$ using Lagrange multipliers.

$$0 = \frac{\delta}{\delta q_{ij}(\mathbf{y}_{ij})}\mathcal{F}(q,\theta) + \sum_{ij}\lambda_{ij}\left(1 - \int q_{ij}(\mathbf{y}_{ij})\mathrm{d}\mathbf{y}_{ij}\right) \tag{10}$$

$$= \int \prod_{j' \setminus j} q_{ij'}(\mathbf{y}_{ij'})p(\mathbf{x}_i, \mathbf{y}_i|\theta) - (\log q_{ij}(\mathbf{y}_{ij}) + 1) - \lambda_{ij}\mathrm{d}\mathbf{y}_i \tag{11}$$

$$\Leftrightarrow q_{ij}(\mathbf{y}_{ij}) \propto \exp\left(\int \prod_{j' \setminus j} q_{ij'}(\mathbf{y}_{ij'}) \log p(\mathbf{x}_i, \mathbf{y}_i|\theta)\mathrm{d}\mathbf{y}_{i \setminus j}\right). \tag{12}$$

So for the mean field approximation the M-step is unchanged. The E-step is now:

`VE-step`:

$$q_{ij}(\mathbf{y}_{ij}) \leftarrow \frac{1}{Z_{ij}} \exp\left(\int \prod_{j' \setminus j} q_{ij'}(\mathbf{y}_{ij'}) \log p(\mathbf{x}_i, \mathbf{y}_i | \theta) \mathrm{d}\mathbf{y}_{i \setminus j} \right). \quad (13)$$

Note: these updates must be made sequentially to ensure that the likelihood decreases. They may be repeated until convergence or they may be performed once.

MAP learning

Suppose we have a prior distribution $p(\theta)$ on θ. Then,

$$\log p(\theta|x) \propto \log p(\theta) + \ell(\theta) \tag{14}$$

So the M-step can be augmented to approximate θ^{MAP}:
`M-step`:

$$\theta \leftarrow \operatorname*{argmax}_{\theta} \log p(\theta) + \sum_i \int q_i(\mathbf{y}_i) \log p(\mathbf{x}_i, \mathbf{y}_i|\theta) \tag{15}$$

In conjunction with an E or VE-step, this yields modes in $p(\theta|\mathbf{x})$.

Dangers of MAP (in general, not just EM/VEM)

Suppose we're given $\mathbf{x}$, $p(\mathbf{x}|\theta)$, $p_\theta(\theta)$, θ a continuous random variables. We're trying to compute the maximum *a posteriori*:

$$\theta^{\text{MAP}} = \operatorname*{argmax}_\theta p(\mathbf{x}|\theta)p(\theta) \tag{16}$$

If ν is any smooth monotonic function then we can write the pdf for $\phi = \nu(\theta)$ in terms of p_θ as follows:

$$p_\theta(\theta) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \int_\theta^{\theta+\varepsilon} p_\theta(\theta)\mathrm{d}\theta \tag{17}$$

$$p_\phi(\phi) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \int_{\nu^{-1}(\phi)}^{\nu^{-1}(\phi)+\varepsilon} p_\theta(\nu^{-1}(\phi))\mathrm{d}\phi \tag{18}$$

$$= \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \int_\theta^{\theta+\varepsilon} p_\theta(\theta)\nu'(\theta)\mathrm{d}\theta \tag{19}$$

$$= p_\theta(\theta)\nu'(\theta) \tag{20}$$

Dangers of MAP

So, we have $p_\theta(\theta)\nu'(\theta) = p_\phi(\phi)$. Suppose θ is supported on $[0, 1]$ and p_θ is smooth. Pick an arbitrary $\theta^\star$ in $[0, 1]$. Define:

$$\nu_r(\theta) = \begin{cases} r\theta & \theta \in [0, \theta^\star - r) \\ r(\theta^\star - r) + \frac{\alpha}{r}\theta & \theta \in [\theta^\star - r, \theta^\star + r) \\ r(\theta^\star - r) + \frac{\alpha}{r}(\theta^\star + r) + r\theta & \theta \in [\theta^\star + r, 1]. \end{cases} \tag{21}$$

Here, α is chosen so that $\nu_r(1) = 1$. Also note that ν_r is not smooth but we can interpolate with a smooth function without affecting the following result. As $r \to 0$, $p_\phi(\nu_r(\theta^\star)) \to \infty$ and if $\theta_0 \neq \theta^\star$, $p_\phi(\nu_r(\theta_0)) \to 0$.

- For a continuous distribution, reparameterization allows arbitrary MAPs.
- Not so with a discrete distribution.

Latent Dirichlet Allocation

Consider the following generative model:

$$\theta_i \sim \text{Dirichlet}(\alpha, \ldots, \alpha) \tag{22}$$

$$z_{ij} \sim \text{Discrete}(\theta_i) \tag{23}$$

$$w_{ij} \sim \text{Discrete}(\phi_{z_{ij}}) \tag{24}$$

Here, ϕ, α are hyperparameters, z_{ij} are latent variables and w_{ij} are observations (Blei, Ng, and Jordan, 2003).

$$p(\theta, z, w|\alpha, \phi) = \prod_i p(\theta_i|\alpha) \prod_j \theta_{iz_{ij}} \phi_{z_{ij} w_{ij}}. \tag{25}$$

Note that this model is exchangeable: probability is invariant under permutation of the observations and the coordinates of the latent variables.

Variational assumption for LDA

Intractability arises from interaction between θ and z. Applying vB theory we parameterise the joint distribution on $q(\theta, z)$ by:

$$q(\theta, z) = \prod_i q(\theta_i|\gamma_i) \prod_j q(z_{ij}|\upsilon_{ij}), \tag{26}$$

$$q(z_{ij} = k|\upsilon_{ij}) = \upsilon_{ijk}, \tag{27}$$

$$q(\theta_i|\gamma_i) \propto \prod_k \theta_{ik}^{\gamma_{ik}-1}. \tag{28}$$

So υ_{ij} parameterises a discrete distribution and γ_i parameterises a Dirichlet distribution.

VBEM for LDA

This yields the following VBEM algorithm:
`VE-step`

$$\upsilon_{ijk} \leftarrow \phi_{kw_{ij}} \left(\Psi(\gamma_{ik}) - \Psi \left(\sum_{k'} \gamma_{ik'} \right) \right) \tag{29}$$

$$\gamma_{ik} \leftarrow \alpha + \sum_{j} \upsilon_{ijk}. \tag{30}$$

Here, Ψ is the Digamma function: $\Psi(x) = \frac{\mathrm{d}}{\mathrm{dx}} \log \Gamma(x)$
`M-step`

$$\phi_{k\ell} \leftarrow \sum_{i} \sum_{j} \upsilon_{ijk} \delta_{\ell}(w_{ij}). \tag{31}$$

Note that the updates for υ and ϕ are unnormalized.

VE-step for LDA

Recall that the VE-step is found by computing:

$$q(\theta, z|\upsilon, \gamma) \leftarrow \underset{q(\theta,z|\upsilon,\gamma)}{\text{argmin}} \ \text{KL}(q(\theta, z|\upsilon, \gamma)||p(\theta, z|w, \phi, \alpha)). \tag{32}$$

$$\text{KL}(q(\theta, z|\upsilon, \gamma)||p(\theta, z|w, \phi, \alpha)) \tag{33}$$

$$= \mathbb{E}_q \left[\log \frac{\prod_i q(\theta_i|\gamma_i) \prod_j q(z_{ij}|\upsilon_{ij})}{\prod_i p(\theta_i|\alpha) \prod_j p(z_{ij}|\theta, \phi, w_{ij})} \right] - \mathbb{E}_q[\log p(w|\alpha, \phi)] \tag{34}$$

$$= \sum_i \mathbb{E}_q[\log q(\theta_i|\gamma_i)] - \mathbb{E}_q[\log p(\theta_i|\alpha)] \tag{35}$$

$$+ \sum_j \mathbb{E}_q[\log q(z_{ij}|\upsilon)] - \mathbb{E}_q[\log p(z_{ij}|\theta, \phi, w_{ij})]. \tag{36}$$

(37)

Computing expected values: 1

$$
\begin{aligned}
\mathbb{E}_q[\log q(\theta_i|\gamma_i)] &= \int q(\theta_i|\gamma_i) \log q(\theta_i|\gamma_i) \mathrm{d}\theta_i, \\
&= \int \frac{1}{B(\gamma_i)} \prod_k \theta_{ik}^{\gamma_{ik}-1} \log \left(\frac{1}{B(\gamma_i)} \prod_k \theta_{ik}^{\gamma_{ik}-1} \right) \mathrm{d}\theta_i, \\
&= -\log B(\gamma_i) + \frac{1}{B(\gamma_i)} \sum_k (\gamma_{ik} - 1) \qquad (38) \\
&\quad \cdot \int_0^\infty \theta_{ik}^{\gamma_{ik}-1} \log \theta_{ik} \mathrm{d}\theta_{ik}, \\
&= -\log B(\gamma_i) + \sum_k (\gamma_{ik} - 1) \left(\Psi(\gamma_{ik}) - \Psi \left(\sum_{k'} \gamma_{ik'} \right) \right).
\end{aligned}
$$

Computing expected values: 2

Similarly,

$$\mathbb{E}_q[\log p(\theta_i|\alpha)] = -\log B(\alpha) + \sum_k (\alpha - 1)\left(\Psi(\gamma_{ik}) - \Psi\left(\sum_{k'} \gamma_{ik'}\right)\right)$$
$$+ \sum_i \sum_j \sum_k \upsilon_{ijk}\left(\Psi(\gamma_{ik}) - \Psi\left(\sum_{k'} \gamma_{ik'}\right)\right)$$

Computing expected values: 3 & 4

$$\mathbb{E}_q[\log q(z_{ij}|\upsilon)] = \int q(z_{ij}|\upsilon) \log q(z_{ij}|\upsilon) \mathrm{d}z_{ij}, \tag{39}$$

$$= \sum_k \upsilon_{ijk} \log \upsilon_{ijk}. \tag{40}$$

Finally,

$$\mathbb{E}_q[\log p(z_{ij}|\theta, \phi, w_{ij})] = \int q(z_{ij}|\upsilon) \log p(z_{ij}|\theta, \phi, w_{ij}) \mathrm{d}z_{ij}, \tag{41}$$

$$= \sum_k \upsilon_{ijk} \log p(w_i j|z_{ij} = k, \phi) p(z_{ij} = k|\theta_i) \tag{42}$$

$$= \sum_k \upsilon_{ijk} \left(\log \phi_{z_{ij} w_{ij}} + \log \theta_{i z_{ij}}\right) \tag{43}$$

Finding VE-step with Lagrange multipliers

Now we add Lagrange multipliers to enforce $\sum_k \upsilon_{ijk} = 1$ and take derivatives and set them to zero:

$$L = \mathrm{KL}(q(\theta, z|\upsilon, \gamma)||p(\theta, z|w, \phi, \alpha)) + \lambda_{ij}\left(\sum_k \upsilon_{ik} - 1\right).$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\upsilon_{ik}} L =& \frac{\mathrm{d}}{\mathrm{d}\upsilon_{ik}} \upsilon_{ijk}\left(\Psi(\gamma_{ik}) - \Psi\left(\sum_{k'} \gamma_{ik}\right)\right) + \upsilon_{ijk} \log \phi_{kw_{ij}} \\ & - \upsilon_{ijk} \log \upsilon_{ijk} + \lambda_{ij}\left(\sum_k \upsilon_{ijk} - 1\right), \\ =& \Psi(\gamma_k) - \Psi\left(\sum_{k'} \gamma_{k'}\right) + \log \phi_{kw_{ij}} - \log \upsilon_{ijk} - 1 + \lambda_{ij}.\end{aligned}$$

Collecting υ_{ijk} yields:

$$\upsilon_{ijk} \leftarrow \phi_{kw_{ij}}\left(\Psi(\gamma_{ik}) - \Psi\left(\sum_{k'} \gamma_{ik'}\right)\right)$$

The M-step

A similar argument finds the VE-step for γ_{ik}. The M-step is:

$$\phi \leftarrow \underset{\phi}{\operatorname{argmax}}\, \mathbb{E}_q[\log p(w, z, \theta|\phi, \alpha)]. \tag{44}$$

$$L = \mathbb{E}_q[\log p(w, z, \theta|\phi, \alpha)] + \sum_k \lambda_k \left(\sum_\ell \phi_{k\ell} - 1 \right), \tag{45}$$

$$= \sum_i \sum_j \sum_k \sum_\ell \upsilon_{ijk} \delta_\ell(w_{ij}) \log \phi_{k w_{ij}} + \sum_k \lambda_k \left(\sum_\ell \phi_{k\ell} - 1 \right). \tag{46}$$

$$\Rightarrow \phi_{k\ell} \propto \sum_i \sum_j \upsilon_{ijk} \delta_\ell(w_{ij}) \tag{47}$$

Collapsed vB for LDA

In collapsed vB for LDA, we assume only that the z are independent (rather than z, θ all independent as in Blei et al. 2003). We marginalize out ϕ, θ during the E-step (Teh et al. 2006).

$$q(z, \theta, \phi) = q(\theta, \phi | z) \prod_{ij} q(z_{ij} | \upsilon_{ij}) \tag{48}$$

Other applications of VB

`http://www.gatsby.ucl.ac.uk/vbayes/`

- VB mixture of factor analyzers, VBHMM, VBLDS (Beal, 2003)
- 'An introduction to variational methods for graphical models' (Jordan et al. 1999)