A quick intro to variational Bayes

Miguel Lázaro-Gredilla
miguel@tsc.uc3m.es

Feb 2014

Machine Learning Group
http://www.tsc.uc3m.es/~miguel/MLG/

Contents

Bayes theorem

Given non-conjugate $p(y|\theta)$ and $p(\theta)$, it is hard to compute

$$p(\theta|\mathbf{y}) = \frac{p(\theta)\prod_n p(y_n|\theta)}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\theta)p(\theta)}{\int p(\mathbf{y}|\theta)p(\theta)d\theta}$$

Bayes theorem

Given non-conjugate $p(y|\theta)$ and $p(\theta)$, it is hard to compute

$$p(\theta|\mathbf{y}) = \frac{p(\theta)\prod_n p(y_n|\theta)}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\theta)p(\theta)}{\int p(\mathbf{y}|\theta)p(\theta)d\theta}$$

which is needed to compute

$$p(y_*|\mathbf{y}) = \int p(y_*|\theta)p(\theta|\mathbf{y})d\theta$$

$$\mathbb{E}[y_*|\mathbf{y}] = \int \mathbb{E}[y_*|\theta]p(\theta|\mathbf{y})d\theta$$

$$\mathbb{V}[y_*|\mathbf{y}] = \int \mathbb{V}[y_*|\theta]p(\theta|\mathbf{y})d\theta$$

The variational bound

$$\log p(\mathbf{y}) \geq \mathcal{F} = \log p(\mathbf{y}) - \mathrm{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y}))$$

The variational bound

$$\log p(\mathbf{y}) \geq \mathcal{F} = \log p(\mathbf{y}) - \mathrm{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y}))$$
$$= \int q(\boldsymbol{\theta}) \log p(\mathbf{y}) d\boldsymbol{\theta} + \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}|\mathbf{y})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}$$

The variational bound

$$
\begin{aligned}
\log p(\mathbf{y}) \geq \mathcal{F} &= \log p(\mathbf{y}) - \mathrm{KL}(q(\boldsymbol{\theta}) || p(\boldsymbol{\theta}|\mathbf{y})) \\
&= \int q(\boldsymbol{\theta}) \log p(\mathbf{y}) d\boldsymbol{\theta} + \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}|\mathbf{y})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\theta}) \log \frac{p(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}
\end{aligned}
$$

The variational bound

$$
\begin{aligned}
\log p(\mathbf{y}) \geq \mathcal{F} &= \log p(\mathbf{y}) - \mathrm{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y})) \\
&= \int q(\boldsymbol{\theta}) \log p(\mathbf{y}) d\boldsymbol{\theta} + \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}|\mathbf{y})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\theta}) \log \frac{p(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \text{ (direct from Jensen)} \\
&= \int q(\boldsymbol{\theta}) \log p(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta} - \mathrm{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}))
\end{aligned}
$$

The variational bound

$$
\begin{aligned}
\log p(\mathbf{y}) \geq \mathcal{F} &= \log p(\mathbf{y}) - \mathrm{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y})) \\
&= \int q(\boldsymbol{\theta}) \log p(\mathbf{y}) d\boldsymbol{\theta} + \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}|\mathbf{y})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\theta}) \log \frac{p(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \text{ (direct from Jensen)} \\
&= \int q(\boldsymbol{\theta}) \log p(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta} - \mathrm{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta})) \\
&= \langle \log p(\mathbf{y}|\boldsymbol{\theta}) \rangle_{q(\boldsymbol{\theta})} - \mathrm{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}))
\end{aligned}
$$

The variational bound

$$
\begin{aligned}
\log p(\mathbf{y}) \geq \mathcal{F} &= \log p(\mathbf{y}) - \mathrm{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y})) \\
&= \int q(\boldsymbol{\theta}) \log p(\mathbf{y}) d\boldsymbol{\theta} + \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}|\mathbf{y})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\theta}) \log \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \text{ (direct from Jensen)} \\
&= \int q(\boldsymbol{\theta}) \log p(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta} - \mathrm{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta})) \\
&= \langle \log p(\mathbf{y}|\boldsymbol{\theta}) \rangle_{q(\boldsymbol{\theta})} - \mathrm{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta})) \\
&= \langle \log p(\mathbf{y}, \boldsymbol{\theta}) \rangle_{q(\boldsymbol{\theta})} + \mathcal{H}(q(\boldsymbol{\theta}))
\end{aligned}
$$

The variational bound

$$
\begin{aligned}
\log p(\mathbf{y}) \geq \mathcal{F} &= \log p(\mathbf{y}) - \mathrm{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y})) \\
&= \int q(\boldsymbol{\theta}) \log p(\mathbf{y}) d\boldsymbol{\theta} + \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}|\mathbf{y})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\theta}) \log \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \text{ (direct from Jensen)} \\
&= \int q(\boldsymbol{\theta}) \log p(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta} - \mathrm{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta})) \\
&= \langle \log p(\mathbf{y}|\boldsymbol{\theta}) \rangle_{q(\boldsymbol{\theta})} - \mathrm{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta})) \\
&= \langle \log p(\mathbf{y}, \boldsymbol{\theta}) \rangle_{q(\boldsymbol{\theta})} + \mathcal{H}(q(\boldsymbol{\theta}))
\end{aligned}
$$

Turns inference into a maximization problem

ML-II can be performed simultaneously

No need for correct scaling on $p(\mathbf{y}|\boldsymbol{\theta})$ or $p(\boldsymbol{\theta})$

Posterior independence

If $q(\boldsymbol{\theta}) = q(\boldsymbol{\theta}_1)q(\boldsymbol{\theta}_2)$

$$\mathcal{F}(q(\boldsymbol{\theta}_1), q(\boldsymbol{\theta}_2)) = \langle \log p(\mathbf{y}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \rangle_{q(\boldsymbol{\theta}_1)q(\boldsymbol{\theta}_2)} + \mathcal{H}(q(\boldsymbol{\theta}_1)) + \mathcal{H}(q(\boldsymbol{\theta}_2))$$

Posterior independence

If $q(\boldsymbol{\theta}) = q(\boldsymbol{\theta}_1)q(\boldsymbol{\theta}_2)$

$$\mathcal{F}(q(\boldsymbol{\theta}_1), q(\boldsymbol{\theta}_2)) = \langle \log p(\mathbf{y}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \rangle_{q(\boldsymbol{\theta}_1)q(\boldsymbol{\theta}_2)} + \mathcal{H}(q(\boldsymbol{\theta}_1)) + \mathcal{H}(q(\boldsymbol{\theta}_2))$$

$$\log q^*(\boldsymbol{\theta}_2) = \underset{\log q(\boldsymbol{\theta}_2)}{\text{argmax}}\ \mathcal{F}(q(\boldsymbol{\theta}_1), q(\boldsymbol{\theta}_2))$$

Posterior independence

If $q(\boldsymbol{\theta}) = q(\boldsymbol{\theta}_1)q(\boldsymbol{\theta}_2)$

$$\mathcal{F}(q(\boldsymbol{\theta}_1), q(\boldsymbol{\theta}_2)) = \langle \log p(\mathbf{y}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \rangle_{q(\boldsymbol{\theta}_1)q(\boldsymbol{\theta}_2)} + \mathcal{H}(q(\boldsymbol{\theta}_1)) + \mathcal{H}(q(\boldsymbol{\theta}_2))$$

$$\log q^*(\boldsymbol{\theta}_2) = \underset{\log q(\boldsymbol{\theta}_2)}{\operatorname{argmax}} \mathcal{F}(q(\boldsymbol{\theta}_1), q(\boldsymbol{\theta}_2))$$

$$\overset{c}{=} \langle \log p(\mathbf{y}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \rangle_{q(\boldsymbol{\theta}_1)} \quad \text{(Jensen's reversal)}$$

Posterior independence

If $q(\boldsymbol{\theta}) = q(\boldsymbol{\theta}_1)q(\boldsymbol{\theta}_2)$

$$\mathcal{F}(q(\boldsymbol{\theta}_1), q(\boldsymbol{\theta}_2)) = \langle \log p(\mathbf{y}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \rangle_{q(\boldsymbol{\theta}_1)q(\boldsymbol{\theta}_2)} + \mathcal{H}(q(\boldsymbol{\theta}_1)) + \mathcal{H}(q(\boldsymbol{\theta}_2))$$

$$\log q^*(\boldsymbol{\theta}_2) = \underset{\log q(\boldsymbol{\theta}_2)}{\text{argmax}}\ \mathcal{F}(q(\boldsymbol{\theta}_1), q(\boldsymbol{\theta}_2))$$

$$\overset{c}{=} \langle \log p(\mathbf{y}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \rangle_{q(\boldsymbol{\theta}_1)} \quad \text{(Jensen's reversal)}$$

and simetrically

$$\log q^*(\boldsymbol{\theta}_1) \overset{c}{=} \langle \log p(\mathbf{y}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \rangle_{q(\boldsymbol{\theta}_2)}$$

Posterior independence

If $q(\boldsymbol{\theta}) = q(\boldsymbol{\theta}_1)q(\boldsymbol{\theta}_2)$

$$\mathcal{F}(q(\boldsymbol{\theta}_1), q(\boldsymbol{\theta}_2)) = \langle \log p(\mathbf{y}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \rangle_{q(\boldsymbol{\theta}_1)q(\boldsymbol{\theta}_2)} + \mathcal{H}(q(\boldsymbol{\theta}_1)) + \mathcal{H}(q(\boldsymbol{\theta}_2))$$

$$\log q^*(\boldsymbol{\theta}_2) = \underset{\log q(\boldsymbol{\theta}_2)}{\operatorname{argmax}} \mathcal{F}(q(\boldsymbol{\theta}_1), q(\boldsymbol{\theta}_2))$$

$$\stackrel{c}{=} \langle \log p(\mathbf{y}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \rangle_{q(\boldsymbol{\theta}_1)} \quad \text{(Jensen's reversal)}$$

and simetrically

$$\log q^*(\boldsymbol{\theta}_1) \stackrel{c}{=} \langle \log p(\mathbf{y}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \rangle_{q(\boldsymbol{\theta}_2)}$$

For an arbitrary factorization

$$\log q^*(\boldsymbol{\theta}_i) \stackrel{c}{=} \langle \log p(\mathbf{y}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n) \rangle_{q(\setminus \boldsymbol{\theta}_i)}$$

Example: Univariate Gaussian mixture (I/II)

Model with K univariate components:

$$p(y_n|\mathbf{m}, \mathbf{v}, z_n) = \mathcal{N}(y_n|m_{z_n}, v_{z_n}); \quad p(z_n) = \text{cat}(\gamma^{(1)}, \ldots, \gamma^{(K)})$$

Posterior is within independent categorical family:

$$q(z_n) = \text{cat}(\alpha_n^{(1)}, \ldots, \alpha_n^{(K)})$$

Example: Univariate Gaussian mixture (I/II)

Model with K univariate components:

$$p(y_n|\mathbf{m},\mathbf{v},z_n) = \mathcal{N}(y_n|m_{z_n},v_{z_n}); \quad p(z_n) = \text{cat}(\gamma^{(1)},\ldots,\gamma^{(K)})$$

Posterior is within independent categorical family:

$$q(z_n) = \text{cat}(\alpha_n^{(1)},\ldots,\alpha_n^{(K)})$$

Variational bound:

$$p(\mathbf{y}|\mathbf{m},\mathbf{v},\gamma) \geq \mathcal{F}(\mathbf{m},\mathbf{v},\gamma,\{\alpha_n^{(k)}\})$$
$$= \sum_n \left(\left\langle -\frac{(y_n - m_{z_n})^2}{2v_{z_n}} - \frac{1}{2}\log(2\pi v_{z_n}) \right\rangle_{q(z_n)} - \text{KL}(q(z_n)||p(z_n)) \right)$$

Example: Univariate Gaussian mixture (II/II)

$$
\begin{aligned}
p(\mathbf{y}|\mathbf{m},\mathbf{v},\boldsymbol{\gamma}) &\geq \mathcal{F}(\mathbf{m},\mathbf{v},\boldsymbol{\gamma},\{\alpha_n^{(k)}\}) \\
&= \sum_{n,k} \alpha_n^{(k)} \left(-\frac{(y_n - m_k)^2}{2v_k} - \frac{1}{2}\log(2\pi v_k) \right) - \sum_{n,k} \alpha_n^{(k)} \log \frac{\alpha_n^{(k)}}{\gamma^{(k)}}
\end{aligned}
$$

Example: Univariate Gaussian mixture (II/II)

$$p(\mathbf{y}|\mathbf{m},\mathbf{v},\boldsymbol{\gamma}) \geq \mathcal{F}(\mathbf{m},\mathbf{v},\boldsymbol{\gamma},\{\alpha_n^{(k)}\})$$
$$= \sum_{n,k} \alpha_n^{(k)} \left(-\frac{(y_n - m_k)^2}{2v_k} - \frac{1}{2}\log(2\pi v_k) \right) - \sum_{n,k} \alpha_n^{(k)} \log \frac{\alpha_n^{(k)}}{\gamma^{(k)}}$$

Maximum is at

$$\alpha_n^{(k)} \propto \gamma^{(k)} \exp\left(-\frac{(y_n - m_k)^2}{2v_k} - \frac{1}{2}\log(2\pi v_k) \right)$$
$$\gamma^{(k)} = \frac{1}{N}\sum_n \alpha_n^{(k)}$$
$$m_k = \sum_n \alpha_n^{(k)} y_n / \sum_n \alpha_n^{(k)}$$
$$v_k = \sum_n \alpha_n^{(k)} (y_n - m_k)^2 / \sum_n \alpha_n^{(k)}$$

Example: Univariate Gaussian mixture (II/II)

$$p(\mathbf{y}|\mathbf{m},\mathbf{v},\boldsymbol{\gamma}) \geq \mathcal{F}(\mathbf{m},\mathbf{v},\boldsymbol{\gamma},\{\alpha_n^{(k)}\})$$
$$= \sum_{n,k} \alpha_n^{(k)} \left(-\frac{(y_n - m_k)^2}{2v_k} - \frac{1}{2}\log(2\pi v_k) \right) - \sum_{n,k} \alpha_n^{(k)} \log \frac{\alpha_n^{(k)}}{\gamma^{(k)}}$$

Maximum is at

$$\alpha_n^{(k)} \propto \gamma^{(k)} \exp\left(-\frac{(y_n - m_k)^2}{2v_k} - \frac{1}{2}\log(2\pi v_k) \right)$$
$$\gamma^{(k)} = \frac{1}{N}\sum_n \alpha_n^{(k)}$$
$$m_k = \sum_n \alpha_n^{(k)} y_n / \sum_n \alpha_n^{(k)}$$
$$v_k = \sum_n \alpha_n^{(k)} (y_n - m_k)^2 / \sum_n \alpha_n^{(k)}$$

EM is just a particular case of VB (bound is tight after “E”)

I would call it MM... (wrt posterior/hyperparam)