13: Modal And Distributional Approximations

11/25/19

Variational Inference

Variational inference approximates an intractable posterior $p(\theta \mid y)$ with some chosen distribution $g(\theta \mid \phi)$ (e.g. multivariate normal).

Variational Inference

Variational inference approximates an intractable posterior $p(\theta \mid y)$ with some chosen distribution $g(\theta \mid \phi)$ (e.g. multivariate normal).

We will assume this approximating distribution factors into J components:

$$g(\theta \mid \phi) = \prod_{j=1}^{J} g_j(\theta_j \mid \phi_j) = g_j(\theta_j \mid \phi_j) g_{-j}(\theta_{-j} \mid \phi_{-j}).$$

We will find ϕ using an EM-like algorithm that minimizes Kullback-Leibler divergence.

Variational Inference

Kullback-Leibler divergence is "reversed" this time:

$$
\begin{aligned}
KL(g||p) &= -\int \log\left(\frac{p(\theta \mid y)}{g(\theta \mid \phi)}\right) g(\theta \mid \phi)\mathrm{d}\theta \\
&= -\int \log\left(\frac{p(\theta, y)}{g(\theta \mid \phi)}\right) g(\theta \mid \phi)\mathrm{d}\theta + \int \log p(y) g(\theta \mid \phi)\mathrm{d}\theta \\
&= \underbrace{-\int \log\left(\frac{p(\theta, y)}{g(\theta \mid \phi)}\right) g(\theta \mid \phi)\mathrm{d}\theta}_{\text{variational lower bound}} + \log p(y)
\end{aligned}
$$

The term that we maximize (minimize the negative) is called the **variational lower bound** aka the **evidence lower bound** (ELBO).

Variational Inference

Every iteration, we cycle through all the hyper-parameters $\phi_1, \ldots, \phi_J$, and change them until convergence is reached.

Variational Inference

Every iteration, we cycle through all the hyper-parameters $\phi_1, \ldots, \phi_J$, and change them until convergence is reached.

Looking at ϕ_j...

$$\int \log\left(\frac{p(\theta, y)}{g(\theta \mid \phi)}\right) g(\theta \mid \phi) \mathrm{d}\theta$$

$$= \iint \left[\log p(\theta, y) - \log g_j(\theta_j \mid \phi_j) - \log g_{-j}(\theta_{-j} \mid \phi_{-j})\right]$$

$$g_j(\theta_j \mid \phi_j) g_{-j}(\theta_{-j} \mid \phi_{-j}) \mathrm{d}\theta_j \mathrm{d}\theta_{-j}$$

$$= \int \left[\int \log p(\theta, y) g_{-j}(\theta_{-j} \mid \phi_{-j}) \mathrm{d}\theta_{-j}\right] g_j(\theta_j \mid \phi_j) \mathrm{d}\theta_j$$

$$- \int \log g_j(\theta_j \mid \phi_j) g_j(\theta_j \mid \phi_j) \mathrm{d}\theta_j - \int \log g_{-j}(\theta_{-j} \mid \phi_{-j}) g_{-j}(\theta_{-j} \mid \phi_{-j}) \mathrm{d}\theta_{-j}$$

$$= \int \log\left(\frac{\tilde{p}(\theta_j)}{g_j(\theta_j \mid \phi_j)}\right) g_j(\theta_j \mid \phi_j) \mathrm{d}\theta_j + \text{constant} \qquad (*)$$

Variational Inference

We think of $\tilde{p}(\theta_j)$ as an unnormalized density

$$\log \tilde{p}(\theta_j) = \int \log p(\theta, y) g_{-j}(\theta_{-j} \mid \phi_{-j}) \mathrm{d}\theta_{-j}$$

because usually

$$\begin{aligned}
\int \tilde{p}(\theta_j) \mathrm{d}\theta_j &= \int \exp\left[\int \log p(\theta, y) g_{-j}(\theta_{-j} \mid \phi_{-j}) \mathrm{d}\theta_{-j}\right] \mathrm{d}\theta_j \\
&\leq \int \exp\left[\log \int p(\theta, y) g_{-j}(\theta_{-j} \mid \phi_{-j}) \mathrm{d}\theta_{-j}\right] \mathrm{d}\theta_j \quad \text{(Jensen's)} \\
&= \iint p(\theta, y) g_{-j}(\theta_{-j} \mid \phi_{-j}) \mathrm{d}\theta_{-j} \mathrm{d}\theta_j \\
&< \infty
\end{aligned}$$

Variational Inference

VI algorithm

For $j = 1, \ldots, J$ set ϕ_j so that $\log g_j(\theta_j \mid \phi_j)$ is equal to

$$\log \tilde{p}(\theta_j) = \int \log p(\theta, y) g_{-j}(\theta_{-j} \mid \phi_{-j}) \mathrm{d}\theta_{-j}$$

Variational Inference: educational testing example

When the parameters are $\alpha_1, \ldots, \alpha_8, \mu, \tau$, the log posterior is

$$\log p(\theta \mid y) = \text{constant} - \frac{1}{2}\sum_{j=1}^{8} \frac{(y_j - \alpha_j)^2}{\sigma_j^2} - 8\log\tau - \frac{1}{2}\frac{1}{\tau^2}\sum_{j=1}^{8}(\alpha_j - \mu)^2$$

and we assume

$$g(\alpha_1, \ldots, \alpha_8, \mu, \tau) = g(\alpha_1) \times \cdots \times g(\alpha_8) g(\mu) g(\tau).$$

Let's reparameterize τ as τ^2 and assume $g(\alpha_1), \ldots, g(\alpha_8) g(\mu)$ are all normal distributions, and $g(\tau^2)$ is an Inverse-Gamma.

Variational Inference: example

$$\begin{aligned}
&\log g(\alpha_j) \\
&\overset{\text{set}}{=} \log \tilde{p}(\alpha_j) \\
&= \int \log p(\theta, y) g_{-j}(\theta_{-j}) \mathrm{d}\theta_{-j} \\
&= -\frac{1}{2} \sum_{i=1}^{8} \frac{E_{-j}[(y_i - \alpha_i)^2]}{\sigma_i^2} - 8E_{-j}[\log \tau] - \frac{1}{2} E_{-j}\left[\frac{1}{\tau^2}\right] \sum_{i=1}^{8} E_{-j}[(\alpha_i - \mu)^2] + c \\
&= -\frac{1}{2} \frac{(y_j - \alpha_j)^2}{\sigma_j^2} - \frac{1}{2} E_{-j}\left[\frac{1}{\tau^2}\right] E_{-j}[(\alpha_j - \mu)^2] + c' \\
&= -\frac{1}{2} \frac{(y_j - \alpha_j)^2}{\sigma_j^2} - \frac{1}{2} E_{-j}\left[\frac{1}{\tau^2}\right] (\alpha_j^2 - 2\alpha_j E_{-j}[\mu])] + c''
\end{aligned}$$

We are using linearity, independence, the data aren't random, and we're grouping all the terms that don't involve α_j into the constant.

Variational Inference: example

For μ:

$$
\begin{aligned}
\log \tilde{p}(\mu) &= \int \log p(\theta, y) g_{-j}(\theta_{-j} \mid \phi_{-j}) \mathrm{d}\theta_{-j} \\
&= -\frac{1}{2} E_{-\mu} \left[\frac{1}{\tau^2} \sum_{j=1}^{8} (\alpha_j - \mu)^2 \right] + \text{constant} \\
&= -\frac{1}{2} E_{-\mu} \left[\frac{1}{\tau^2} \right] \sum_{j=1}^{8} \left(\mu^2 - 2\mu E_{-\mu}[\alpha_j] \right) + \text{constant} \\
&= -\frac{1}{2} E_{-\mu} \left[\frac{1}{\tau^2} \right] \left(8\mu^2 - 2\mu \sum_{j=1}^{8} E_{-\mu}[\alpha_j] \right) + \text{constant}
\end{aligned}
$$

So $g(\mu) =$...

Variational Inference: example

For τ (not τ^2):

$$\log \tilde{p}(\tau) = \int \log p(\theta, y) g_{-j}(\theta_{-j} \mid \phi_{-j}) \mathrm{d}\theta_{-j}$$
$$= -8 \log \tau - \frac{1}{2}\frac{1}{\tau^2} E_{-\tau}\left[\sum_{j=1}^{8} (\alpha_j - \mu)^2\right] + c$$

So $g(\tau) \propto \tau^{-8} \exp\left[-\frac{\sum_j E_{-\tau}[(\alpha_j - \mu)^2]}{2\tau^2}\right]$ which means

$$g(\tau^2) = (\tau^2)^{-(\frac{7}{2}+1)} \exp\left[-\frac{\sum_j E_{-\tau}[(\alpha_j - \mu)^2]}{2\tau^2}\right]$$

which is an InverseGamma$\left(\frac{7}{2}, \frac{\sum_j E_{-\tau}[(\alpha_j - \mu)^2]}{2}\right)$

Variational Inference: example

To complete this example, we need to derive:

- for $g(\alpha_j)$:
  1. $E_{-j}\left[\frac{1}{\tau^2}\right] = E_{\tau^2}\left[\frac{1}{\tau^2}\right]$,
  2. $E_{-j}[\mu] = E_{\mu}[\mu]$
- for $g(\mu)$:
  1. $E_{-\mu}[\alpha_j] = E_{\alpha_j}[\alpha_j]$,
  2. $E_{-j}\left[\frac{1}{\tau^2}\right] = E_{\tau^2}\left[\frac{1}{\tau^2}\right]$
- for $g(\tau^2)$:
  1. $\sum_j E_{-\tau}[(\alpha_j - \mu)^2] = \sum_j E_{\alpha_j,\mu}[(\alpha_j - \mu)^2]$,

Variational Inference: example

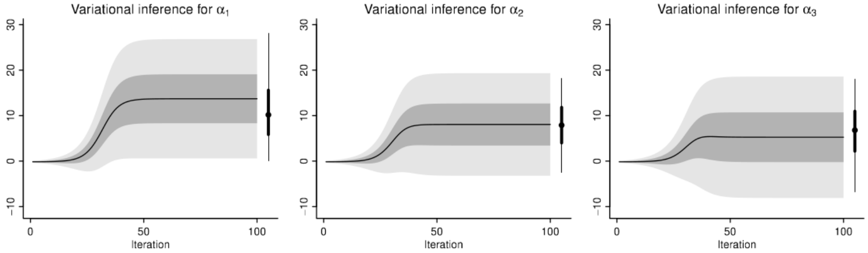

Figure 13.6 *Progress of inferences for the effects in schools A, B, and C, for 100 iterations of variational Bayes. The lines and shaded regions show the median, 50% interval, and 90% interval for the variational distribution. Shown to the right of each graph are the corresponding quantiles for the full Bayes inference as computed via simulation.*

Expectation Propagation: warmup

$p(x \mid \theta)$ is in the exponential family if it can be written as

$$h(x) \exp \left[\eta(\theta)' T(x) - A(\theta) \right]$$

Example:

$$N(\theta \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{1}{2\sigma^2}\theta^2 + \frac{\mu}{\sigma^2}\theta - \frac{\mu^2}{2\sigma^2} \right]$$

sufficient statistic: (θ^2, θ)
canonical/natural parameters: $(-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2})$

Expectation Propagation

Expectation Propogation is another deterministic iterative technique that approximates the posterior with a distribution that is in the exponential family.

1. $p(\theta \mid y) = f(\theta) = \prod_{i=0}^{n} f_i(\theta)$
2. $g(\theta) = \prod_{i=0}^{n} g_i(\theta)$

$f_0(\theta) = p(\theta)$, $f_1(\theta) = p(y_1 \mid \theta)$, ...

For more info: `https://arxiv.org/abs/1412.4869`

Expectation Propagation

The **cavity distribution** is

$$g_{-i}(\theta) \propto g(\theta)/g_i(\theta),$$

and the **tilted distribution** is

$$g_{-i}(\theta)f_i(\theta).$$

At each stage, we update $g_i(\theta)$ so that we “target” $g_{-i}(\theta)f_i(\theta)$ with $g(\theta)$.

Expectation Propagation

At each stage, we update $g_i(\theta)$ so that we "target" $g_{-i}(\theta)f_i(\theta)$ with $g(\theta)$.

Notice that

$$\frac{\text{target}}{\text{"proposal"}} = \frac{g_{-i}(\theta)f_i(\theta)}{g(\theta)} = \frac{f_i(\theta)}{g_i(\theta)}.$$

However, we cannot ignore the cavity distribution in each "site" update. This is because we choose $g(\theta)$ so that its **moments match** those of $g_{-i}(\theta)f_i(\theta)$. This is like choosing $g_i(\theta)$ to approximate $f_i(\theta)$ **in the context of** $g_{-i}(\theta)$.

Expectation Propagation

If $g(\theta) = \text{Normal}(\mu, \Sigma)$, for each i we change μ and Σ by solving

$$\mu \stackrel{\text{set}}{=} E_{\text{tilted } i}[\theta]$$

and

$$\Sigma \stackrel{\text{set}}{=} \text{Var}_{\text{tilted } i}[\theta]$$

where $E_{\text{tilted } i}[\theta] = \int \theta g_{-i}(\theta) f_i(\theta) \text{d}\theta$ and
$\text{Var}_{\text{tilted } i}[\theta] = \int (\theta - \mu)(\theta - \mu)' g_{-i}(\theta) f_i(\theta) \text{d}\theta$.

The hard part is integrating.

Expectation Propagation: example

Let θ be a vector of regression parameters for a logistic regression:

$$\begin{aligned} p(\theta \mid y) &\propto \prod_{i=1}^{n} p(y_i \mid \theta) p(\theta) \\ &= \prod_{i=0}^{n} f_i(\theta) \\ &= f_0(\theta) \prod_{i=1}^{n} [\text{invlogit}(X_i'\theta)]^{y_i} [1 - \text{invlogit}(X_i'\theta)]^{m_i - y_i} \end{aligned}$$

and choose $g(\theta)$ to be Normal(μ, Σ)

Expectation Propagation: example

We choose $g(\theta)$ to be Normal(μ, Σ):

$$
\begin{aligned}
g(\theta) &\propto \prod_{i=0}^{n} \exp\left[-\frac{1}{2}(\theta-\mu_i)'\Sigma_i^{-1}(\theta-\mu_i)\right] \\
&\propto \exp\left[-\frac{1}{2}\sum_{i=0}^{n}\left(\theta'\Sigma_i^{-1}\theta - 2\mu_i'\Sigma_i^{-1}\theta\right)\right] \\
&= \exp\left[-\frac{1}{2}\left(\theta'\left[\underbrace{\sum_{i=0}^{n}\Sigma_i^{-1}}_{\Sigma^{-1}}\right]\theta - 2\left[\underbrace{\sum_{i=0}^{n}\mu_i'\Sigma_i^{-1}}_{\text{natural param 2}}\right]\theta\right)\right]
\end{aligned}
$$

Algorithmically, μ, Σ change at each iteration.

Expectation Propagation: example

$g(\theta)$ is Normal(μ, Σ):

$$g(\theta) \propto \exp\left[-\frac{1}{2}\left(\theta'\left[\underbrace{\sum_{i=0}^{n}\Sigma_i^{-1}}_{\Sigma^{-1}}\right]\theta - 2\left[\underbrace{\sum_{i=0}^{n}\mu_i'\Sigma_i^{-1}}_{\Sigma^{-1}\mu}\right]\theta\right)\right]$$

Step 1: determine cavity distribution. $g_{-i}(\theta) = \text{Normal}(\mu_{-i}, \Sigma_{-i})$ where

$$\Sigma_{-i}^{-1} = \Sigma^{-1} - \Sigma_i^{-1}$$

and

$$\Sigma_{-i}^{-1}\mu_{-i} = \Sigma^{-1}\mu - \Sigma_i^{-1}\mu_i$$

Expectation Propagation: example

Step 2: find cavity distribution for $\eta = X_i'\theta$.

Because any linear transformation of normals is normal and because $g_{-i}(\theta) = \text{Normal}(\mu_{-i}, \Sigma_{-i})$:

$$g_{-i}(\eta) = \text{Normal}(M_{-i}, V_{-i})$$

where $M_{-i} = X_i'\mu_{-i}$ and $V_{-i} = X_i'\Sigma_{-i}X_i$.

Expectation Propagation: example

Step 3: define the unnormalized tilted distribution

$$g_{-i}(\eta)f_i(\eta) = g_{-i}(\eta)\text{Binomial}(m_i, \text{invlogit}(\eta)).$$

and find its expectations numerically with the Gauss-Kronrod quadrature method:

$$\begin{aligned} E_k &= \int_{-\infty}^{\infty} \eta^k g_{-i}(\eta) f_i(\eta) \mathrm{d}\eta \\ &\approx \int_{M_{-i}-\delta\sqrt{V_{-i}}}^{M_{-i}+\delta\sqrt{V_{-i}}} \eta^k g_{-i}(\eta) f_i(\eta) \mathrm{d}\eta \end{aligned}$$

for $k = 0, 1, 2$ and δ is some large number (e.g. 10). Finally compute $M = E_1/E_0$ and $V = E_2/E_0 - (E_1/E_0)^2$ and set $g(\eta) = \text{Normal}(M, V)$.

Expectation Propagation: example

In the previous steps we found $g(\eta) = \text{Normal}(M, V)$ and $g_{-i}(\eta) = \text{Normal}(M_{-i}, V_{-i})$.

Step 4: find $g_i(\eta) = \text{Normal}(M_i, V_i)$:

$$
\begin{aligned}
g_i(\eta) &= g(\eta)/g_{-i}(\eta) \\
&\propto \frac{\exp\left[-\frac{1}{2V}\eta^2 + \frac{M}{V}\eta\right]}{\exp\left[-\frac{1}{2V_{-i}}\eta^2 + \frac{M_{-i}}{V_{-i}}\eta\right]} \\
&= \exp\left[-\frac{1}{2}\left(\underbrace{\frac{1}{V} - \frac{1}{V_{-i}}}_{\frac{1}{V_i}}\right)\eta^2 + \left(\underbrace{\frac{M}{V} - \frac{M_{-i}}{V_{-i}}}_{\frac{M_i}{V_i}}\right)\eta\right]
\end{aligned}
$$

Expectation Propagation: example

Step 5: find $g_i(\theta)$.

$$g_i(\theta) = \text{Normal}(\mu_i, \Sigma_i)$$

where

$$\Sigma_i^{-1}\mu_i = X_i \frac{M_i}{V_i}$$

and

$$\Sigma_i^{-1} = X_i \frac{1}{V_i} X_i'$$

Expectation Propagation: example

Step 6: find $g(\theta) \propto g_i(\theta) g_{-i}(\theta)$.

$$g(\theta) \propto \exp\left[-\frac{1}{2}\left(\theta'\left[\underbrace{\sum_{i=0}^{n}\Sigma_i^{-1}}_{\Sigma^{-1}}\right]\theta - 2\left[\underbrace{\sum_{i=0}^{n}\mu_i'\Sigma_i^{-1}}_{\Sigma^{-1}\mu}\right]\theta\right)\right]$$

$$\Sigma^{-1}\mu = \overbrace{\Sigma_{-i}^{-1}\mu_{-i}}^{\text{from step 1}} + \overbrace{\Sigma_i^{-1}\mu_i}^{\text{from step 5}}$$

and

$$\Sigma^{-1} = \underbrace{\Sigma_{-i}^{-1}}_{\text{from step 1}} + \underbrace{\Sigma_i^{-1}}_{\text{from step 5}}$$

Discussions of Variational Bayes and Expectation Propagation

- EM algorithm for finding posterior mode of $p(\phi \mid y)$ can be thought of as a special case of variational Bayes by treating (a) $g(\phi)$ is a point mass (b) $g(\gamma)$ is of the form $g(\gamma \mid \phi, y)$ but unconstrained to any family of distribution.
- Moment matching of EP corresponds to minimizing the Kullback-Leibler divergence from the tilted distribution and to the new approximated distribution $g(\theta)$. But no guarantee that the KL divergence between the posterior distribution and the approximated distribution is decreasing.

Discussions of Variational Bayes and Expectation Propagation

- Both VI and EP have more generalized form by revising the KL divergence.
- Both VI and EP can serve as the proposal distribution and be incorporated into importance sampling/resampling framework.