13: Modal And Distributional Approximations

Taylor

11/20/19

Approximations using mixture distributions

The book has some discussions about approximating a posterior distribution in a direct way (including finding the mode/modes) that we don't address:

1. approximating multimodal distributions with normal mixtures
2. approximating multimodal distributions with t mixtures

The EM Algorithm

The **expectation-maximization algorithm** finds the argument that maximizes the marginal posterior. It's useful in situations where there is missing data in a model (e.g. hierarchical models, factor models, hidden markov models, state space models, etc.).

It folows the following steps

1. replace missing values by their expectations given the guessed parameters,
2. estimate parameters assuming the missing data are equal to their estimated values,
3. re-estimate the missing values assuming the new parameter estimates are correct,
4. re-estimate parameters,

and so forth, iterating until convergence.

The EM Algorithm

Call $\theta = (\gamma, \phi)$. You're interested in the mode of $p(\phi \mid y)$. Typically, γ is "hidden data."

$$\log p(\phi \mid y) = \log \frac{p(\gamma, \phi \mid y)}{p(\gamma \mid \phi, y)} = \log \underbrace{p(\gamma, \phi \mid y)}_{\text{joint posterior}} - \log \underbrace{p(\gamma \mid \phi, y)}_{\text{conditional posterior}}$$

The EM Algorithm

Call $\theta = (\gamma, \phi)$. You're interested in the mode of $p(\phi \mid y)$. Typically, γ is "hidden data."

$$\log p(\phi \mid y) = \log \frac{p(\gamma, \phi \mid y)}{p(\gamma \mid \phi, y)} = \log \underbrace{p(\gamma, \phi \mid y)}_{\text{joint posterior}} - \log \quad \underbrace{p(\gamma \mid \phi, y)}_{\text{conditional posterior}}$$

taking expectations on both sides with respect to $p(\gamma \mid \phi^{\text{old}}, y)$ yields:

$$\log p(\phi \mid y) = E\left[\log p(\gamma, \phi \mid y) \mid \phi^{\text{old}}, y\right] - E\left[\log p(\gamma \mid \phi, y) \mid \phi^{\text{old}}, y\right]$$

The EM Algorithm

We iteratively use the middle term in
$\log p(\phi \mid y) = E\left[\log p(\gamma, \phi \mid y) \mid \phi^{\text{old}}, y\right] - E\left[\log p(\gamma \mid \phi, y) \mid \phi^{\text{old}}, y\right]$.

The Q quantity in the "E" step

$$Q(\phi \mid \phi^{\text{old}}) = E\left[\log p(\gamma, \phi \mid y) \mid \phi^{\text{old}}, y\right]$$

The EM Algorithm

We iteratively use the middle term in
$\log p(\phi \mid y) = E\left[\log p(\gamma, \phi \mid y) \mid \phi^{\text{old}}, y\right] - E\left[\log p(\gamma \mid \phi, y) \mid \phi^{\text{old}}, y\right]$.

The Q quantity in the "E" step

$$Q(\phi \mid \phi^{\text{old}}) = E\left[\log p(\gamma, \phi \mid y) \mid \phi^{\text{old}}, y\right]$$

The EM algorithm

Repeat the following until convergence:

1. E-step: calculate $Q(\phi \mid \phi^{\text{old}})$
2. M-step: replace ϕ^{old} with $\arg\max Q(\phi \mid \phi^{\text{old}})$

The EM Algorithm

If we follow this strategy, $\log p(\phi \mid y)$ increases at every iteration:

$$\begin{aligned}
\log p(\phi \mid y) &= E\left[\log p(\gamma, \phi \mid y) \mid \phi^{\text{old}}, y\right] - E\left[\log p(\gamma \mid \phi, y) \mid \phi^{\text{old}}, y\right] \\
&= Q(\phi \mid \phi^{\text{old}}) - E\left[\log p(\gamma \mid \phi, y) \mid \phi^{\text{old}}, y\right] && \text{(defn. Q)} \\
&\geq Q(\phi \mid \phi^{\text{old}}) - E\left[\log p(\gamma \mid \phi^{\text{old}}, y) \mid \phi^{\text{old}}, y\right] && \text{(HW)}
\end{aligned}$$

The EM Algorithm

If we follow this strategy, $\log p(\phi \mid y)$ increases at every iteration:

$$
\begin{aligned}
\log p(\phi \mid y) &= E\left[\log p(\gamma, \phi \mid y) \mid \phi^{\text{old}}, y\right] - E\left[\log p(\gamma \mid \phi, y) \mid \phi^{\text{old}}, y\right] \\
&= Q(\phi \mid \phi^{\text{old}}) - E\left[\log p(\gamma \mid \phi, y) \mid \phi^{\text{old}}, y\right] && \text{(defn. Q)} \\
&\geq Q(\phi \mid \phi^{\text{old}}) - E\left[\log p(\gamma \mid \phi^{\text{old}}, y) \mid \phi^{\text{old}}, y\right] && \text{(HW)}
\end{aligned}
$$

So

$$
\begin{aligned}
&\log p(\phi^{\text{new}} \mid y) - \log p(\phi^{\text{old}} \mid y) \\
&= \log p(\phi^{\text{new}} \mid y) - \left\{ Q(\phi^{\text{old}} \mid \phi^{\text{old}}) - E\left[\log p(\gamma \mid \phi^{\text{old}}, y) \mid \phi^{\text{old}}, y\right] \right\} \\
&\geq Q(\phi^{\text{new}} \mid \phi^{\text{old}}) - E\left[\log p(\gamma \mid \phi^{\text{old}}, y) \mid \phi^{\text{old}}, y\right] \\
&\qquad - \left\{ Q(\phi^{\text{old}} \mid \phi^{\text{old}}) - E\left[\log p(\gamma \mid \phi^{\text{old}}, y) \mid \phi^{\text{old}}, y\right] \right\} \\
&= Q(\phi^{\text{new}} \mid \phi^{\text{old}}) - Q(\phi^{\text{old}} \mid \phi^{\text{old}})
\end{aligned}
$$

The EM Algorithm

Notes:

1. The EM algo isn't inherently Bayesian. It can also be used to accomplish maximum likelihood estimation.
2. The expectation of $\log p(\gamma, \phi \mid y)$ is usually easy to compute because it is a sum, and might only depend on sufficient statistics
3. The EM algorithm is parameterization independent
4. The *Generalized* EM (GEM) just increases Q instead of maximizing it.The book describes many generalizations including GEM.
5. You might find multiple modes if you start from multiple starting points (using mixture approximations afterwards)
6. if you can, debug by printing $\log p(\phi^i \mid y)$ at every iteration and make sure it increases monotonically

Marginal and conditional posterior density approximations

We can approximate $p(\phi \mid y)$ using normal, where the mode is found by the EM algorithm and we can numerically calculate the variance at the mode.
In other cases, it is not possible to construct an approximation to $p(\phi \mid y)$ using this method. We are going to discuss another possibility.

Marginal and conditional posterior density approximations

The approximation is constructed as follows:

$$p_{\text{approx}}(\phi \mid y) = \frac{p(\gamma, \phi \mid y)}{p_{\text{approx}}(\gamma \mid \phi, y)}$$

$p_{\text{approx}}(\gamma \mid \phi, y)$ is an analytic approximation to $p(\gamma \mid \phi, y)$.

- If $p(\gamma \mid \phi, y)$ itself is analytic, any γ should not affect the distribution on the left hand side.
- If the analytic approximation of the conditional distribution is not exact, we need to pick γ. The suggestion is use the center of the approximate distribution $p_{\text{approx}}(\gamma \mid \phi, y)$, denote it as $\hat{\gamma}(\phi)$.

If $p_{\text{approx}}(\gamma \mid \phi, y)$ is normal with mean $\hat{\gamma}(\phi)$ and variance $V_{\gamma}(\phi)$, $p_{\text{approx}}(\phi \mid y) \propto p(\hat{\gamma}(\phi), \phi \mid y)|V_{\gamma}(\phi)|^{1/2}$.

Marginal and conditional posterior density approximations

Since $p(\phi \mid y) = \int p(\gamma, \phi \mid y) d\gamma$. We can also use the above trick and importance sampling to approximate $p(\phi \mid y)$.

$$p(\phi \mid y) = \int \frac{p(\gamma, \phi \mid y)}{p_{\text{approx}}(\gamma \mid \phi, y)} p_{\text{approx}}(\gamma \mid \phi, y) d\gamma$$

Sample $\gamma^1, \ldots, \gamma^S$ from $p_{\text{approx}}(\gamma \mid \phi, y)$, then
$p(\phi \mid y) = (1/S) \sum_{s=1}^{S} [p(\gamma^s, \phi \mid y) / p_{\text{approx}}(\gamma^s \mid \phi, y)]$