Module 8

Taylor

2022-07-13

Schedule Information

- HW3 Due 7/17
- Project Updates:
  - I'm checking in with groups that request my services
  - Paper+Presentation due 8/10 (last day of semester)
- HW4 out soon
- HW5 will be due 8/10 as well
  - Less time to complete, so will be shorter

Big Picture for Today

1. The primary object of interest/inference is the **parameter posterior** $p(\theta \mid y)$
2. When we use conjugate priors, there is a formula for it.
3. In the last few weeks, we have also been drawing samples from it (with MCMC algos).
4. Today we will start try to find its mode, or to approximate the distribution with another distribution that is hopefully "close."

- These strategies are less ambitious ceteris paribus, but they might be more scalable or feasible for certain problems.

Big Picture for Today

Today will be more mathy.

I'll give you some examples next week.

Most of the math tricks are:

- properties of logarithms
- linearity of expectation

Finding Modes: The EM algorithm

It's an optimization algorithm–it finds the mode of $p(\theta \mid y)$ (aka $\underset{\theta}{\operatorname{argmax}}\, p(\theta \mid \mathbf{y})$)

It works on models that have latent/hidden variables: $\mathbf{x}$

At every iteration you do two things: the "E" step (stands for expectation), and then the "M"-step (stands for maximization)

Finding Modes: The EM algorithm

Caveats:

- it just finds a mode–ignores shape of posterior
- it only works on hidden/latent variable models
- it is sometimes given a different name when it is applied to a specific model
- $p(\theta^k \mid y)$ increases monotonically, but no guarantee your local mode is the global mode.
- frequentists use EM to find $\underset{\theta}{\operatorname{argmax}}\, p(\mathbf{y} \mid \theta)$ (i.e. maximum likelihood. . . no priors)

Finding Modes: The EM algorithm

Notation

- $\mathbf{y} := \{y_1, \ldots, y_n\}$: observed data
- $\mathbf{x}$: latent/hidden data
- θ all parameters

Complete-data likelihood

$$p(\mathbf{y}, \mathbf{x} \mid \theta) = p(\mathbf{y} \mid \mathbf{x}, \theta) p(\mathbf{x} \mid \theta)$$

Observed-data likelihood

$$p(\mathbf{y} \mid \theta) = \begin{cases} \int p(\mathbf{y} \mid \mathbf{x}, \theta) p(\mathbf{x} \mid \theta) \mathrm{d}\mathbf{x} & \text{continuous hidden variables} \\ \sum_{\mathbf{x}} p(\mathbf{y} \mid \mathbf{x}, \theta) p(\mathbf{x} \mid \theta) & \text{discrete hidden variables} \end{cases}$$

Finding Modes: The EM algorithm

Marginal posterior:

$$p(\theta \mid \mathbf{y}) \propto \underbrace{p(\mathbf{y} \mid \theta)}_{\text{can't evaluate this or its derivatives}} p(\theta)$$

The (high-dimensional!) joint posterior:

$$p(\theta, \mathbf{x} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \mathbf{x}, \theta) p(\mathbf{x} \mid \theta) p(\theta)$$

The conditional posterior:

$$p(\mathbf{x} \mid \theta, \mathbf{y})$$

Finding Modes: The EM algorithm

. . . currently at iteration $k - 1$. . .

Step 1 of 2: the "E"-step

$$Q(\theta \mid \theta^{k-1}) := \mathbf{E}_{\mathbf{x}|\theta^{k-1},\mathbf{y}} \left[\log p(\theta, \mathbf{x} \mid \mathbf{y})\right]$$

(model-specific pencil & paper derivations)

Finding Modes: The EM algorithm

. . . currently at iteration $k - 1$. . .

Step 2 of 2: the "M"-step

$$\theta^k := \operatorname*{argmax}_{\theta} Q(\theta \mid \theta^{k-1})$$

(can involve calculus and setting derivatives equal to 0, or some computational technique that circumvents pencil & paper math)

The EM algorithm: an example

```
##     eruptions  waiting
## 0       3.600       79
## 1       1.800       54
## 2       3.333       74
## 3       2.283       62
## 4       4.533       85
```

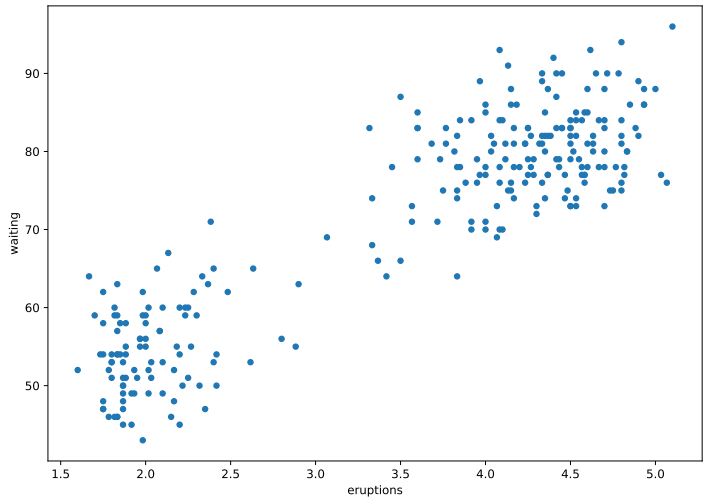

Figure 1: old faithful

More info on Jupyter notebook demo.

Variational Bayes

Benefits:

- Instead of finding $p(\theta \mid y)$, approximate it with some $g(\theta)$
- You can find credible intervals now, because it's more than just mode-finding.
- $g(\theta)$ will be optimal in the sense that it is the "closest" distribution to the true posterior
- $g(\theta)$ are usually "simple" (e.g. multivariate normal and all parameters are independent)
- This can be applied to models that don't have missing/latent variables.
- It can be much faster than MCMC, and sometimes it works when MCMC doesn't (e.g. models lots of parameters).

Variational Bayes

Caveats:

- To my knowledge, there are no convergence guarantees.
- You should run some posterior predictive checks after you get an approximate posterior.
- $q(\theta^k)$ improves monotonically, but no guarantee your local mode is the global mode.
- frequentists use something similar, so it can be confusing with terminology overlap.
- adding to that confusion, some presentations blur the distinction between parameters and latent variables

Variational Bayes

Similarities:

- It is kind of like Gibbs sampling in that, at every iteration, it cycles through all the components of the parameter vector. However, we're not drawing samples here.
- It is kind of like the EM algorithm in that there are expectations of log densities. However, we're doing more than finding a mode here.

Variational Bayes

Say $\theta = (\theta^1, \ldots, \theta^J)$. A standard simplifying assumption is

$$g(\theta) = g_1(\theta^1) \times \cdots \times g_J(\theta^J)$$

To be more precise

$$\begin{aligned} g(\theta \mid \phi) &= g_1(\theta^1 \mid \phi^1) \times \cdots \times g_j(\theta^j \mid \phi^j) \\ &= g_j(\theta^j \mid \phi^j) g_{-j}(\theta^{-j} \mid \phi^{-j}) \end{aligned}$$

Variational Bayes

We need something that tells us the "distance" between two **distributions**. We need **Kullback-Leibler divergence**

$$\mathrm{KL}(g||p) = -\mathbb{E}_g \left[\log \frac{p(\theta \mid y)}{g(\theta)} \right] \geq 0$$

When the functions in the numerator and denominator are equal at **every** value of θ, then the quantity is equal to 0.

Variational Bayes

Finding the best g is equivalent to finding the best defining hyperparameters ϕ. We seek

$$\phi^* := \underset{\phi}{\operatorname{argmin}} \, \mathsf{KL}\left[g(\theta \mid \phi) || p(\theta \mid \phi)\right]$$

Variational Bayes

Let's play with $\mathrm{KL}(g||p)$ a little:

$$
\begin{aligned}
\mathrm{KL}(g||p) &= -\mathbb{E}_g\left[\log \frac{p(\theta \mid y)}{g(\theta)}\right] \\
&= -\mathbb{E}_g\left[\log p(\theta \mid y)\right] + \mathbb{E}_g\left[\log g(\theta)\right] \\
&= -\mathbb{E}_g\left[\log p(y, \theta)\right] + \mathbb{E}_g\left[\log p(y)\right] + \mathbb{E}_g\left[\log g(\theta)\right] \\
&= -\mathbb{E}_g\left[\log \frac{p(y, \theta)}{g(\theta)}\right] + \log p(y)
\end{aligned}
$$

$p(y)$ doesn't depend on θ, so we can just drop that term.

FWIW: the thing that's left over (without the negative sign) is called the **variational lower bound** or the **evidence lower bound (ELBO)**

Variational Bayes

Summarizing:

- want to find a g that approximates the posterior
- do this by minimizing $\mathrm{KL}(g||p)$
- equivalent to minimizing $-\mathbb{E}_g\left[\log \frac{p(y,\theta)}{g(\theta)}\right]$
- equivalent to maximizing ELBO $:= \mathbb{E}_g\left[\log \frac{p(y,\theta)}{g(\theta)}\right]$

Why did we assume that g factors, though?

Coordinate Ascent Variational Inference

Looking at ϕ_j only

$$-\mathbb{E}_g\left[\log \frac{p(y,\theta)}{g(\theta)}\right] = -\int \log\left(\frac{\tilde{p}(\theta_j)}{g_j(\theta_j \mid \phi_j)}\right) g_j(\theta_j \mid \phi_j)\mathrm{d}\theta_j + \text{constant} \qquad (*)$$

where

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

to minimize set $\tilde{p}(\theta_j) = g_j(\theta_j \mid \phi_j)$.

Coordinate Ascent Variational Inference

Longer version:

$$
\begin{aligned}
\mathbb{E}_g\left[\log \frac{p(y,\theta)}{g(\theta)}\right] &= \int \log\left(\frac{p(\theta,y)}{g(\theta \mid \phi)}\right) g(\theta \mid \phi)\mathrm{d}\theta \\
&= \iint \left[\log p(\theta,y) - \log g_j(\theta_j \mid \phi_j) - \log g_{-j}(\theta_{-j} \mid \phi_{-j})\right] \\
&\qquad g_j(\theta_j \mid \phi_j) g_{-j}(\theta_{-j} \mid \phi_{-j})\mathrm{d}\theta_j \mathrm{d}\theta_{-j} \\
&= \int \left[\int \log p(\theta,y) g_{-j}(\theta_{-j} \mid \phi_{-j})\mathrm{d}\theta_{-j}\right] g_j(\theta_j \mid \phi_j)\mathrm{d}\theta_j \\
&- \int \log g_j(\theta_j \mid \phi_j) g_j(\theta_j \mid \phi_j)\mathrm{d}\theta_j \\
&- \int \log g_{-j}(\theta_{-j} \mid \phi_{-j}) g_{-j}(\theta_{-j} \mid \phi_{-j})\mathrm{d}\theta_{-j} \\
&= \int \log\left(\frac{\tilde{p}(\theta_j)}{g_j(\theta_j \mid \phi_j)}\right) g_j(\theta_j \mid \phi_j)\mathrm{d}\theta_j + \text{constant}
\end{aligned}
$$

(*)

Coordinate Ascent Variational Inference

Coordinate Ascent Variational Inference algorithm

Do the following many times:
for $j = 1, \ldots, J$ set ϕ_j so that $g_j(\theta_j \mid \phi_j) := \tilde{p}(\theta_j)$

Difficult to derive. But after you do, it's easy to code.

Variational Bayes

Next class:

- more examples
- a recent "more automatic" version
- using it in `PYMC`