DS 102 Discussion 4

Monday, February 24, 2020

In this discussion, we'll continue to develop intuition and experience with how expectation-maximization (EM) allows us to fit model parameters by approximating maximum likelihood estimation. In particular, EM comes in handy when our models involve *latent variables*, or variables we never actually observe in the data, which are common when we try to model complex phenomena.

Consider the beta-binomial model:

$$Z \sim \text{Beta}(\alpha, \beta)$$
$$X \sim \text{Binomial}(n, Z)$$

where the integer n is considered fixed and known, and $\alpha, \beta > 0$ are the two parameters. You can think of the beta-binomial as randomly picking the bias Z of a coin, then flipping that coin n times and observing how many heads show up. In practice, it's often used to model data where a binomial would seem appropriate, but the data has higher variance than a vanilla binomial random variable. The randomness in picking p captures that increased variance.

Suppose we're interested in the distribution of batting averages in Major League Baseball, which we model as a beta distribution with unknown positive parameters α and β. That is, each player's true batting average is a value $X \in [0, 1]$ drawn from this distribution. However, we don't actually observe each player's true batting average. Instead, over the course of a season we observe X, the number of hits out of n total pitches.

The two steps of EM are motivated by two insights.

- *Expectation (E) step*: $q^{(t)} \leftarrow \mathbb{P}(Z \mid X, \alpha^{(t)}, \beta^{(t)})$. If you knew α, β, it'd be straightforward to compute $\mathbb{P}(Z \mid X, \alpha, \beta)$ (which we'll show). This can be interpreted as imputing the "missing values" of Z that you didn't observe.

- *Maximization (M) step*: $\alpha^{(t+1)}, \beta^{(t+1)} \leftarrow \text{argmax}_{\alpha,\beta>0} \mathbb{E}_{Z\sim q^{(t)}}[\log \mathbb{P}(X, Z \mid \alpha, \beta)]$. The insight behind this is that if you knew Z, it'd be straightforward to find the α, β that maximize $\mathbb{P}(X, Z \mid \alpha, \beta)$ (which we'll show).

1. For the E-step, derive the probability density function of the posterior $p(z \mid x, \alpha, \beta)$. Recall that the probability density function of the $\text{Beta}(\alpha, \beta)$ distribution is given by

$$p(z \mid \alpha, \beta) = \frac{z^{\alpha-1}(1-z)^{\beta-1}}{B(\alpha, \beta)}$$

   where $B(\alpha, \beta)$ is the normalizing constant.

   What fact about the beta and binomial distributions have we recovered?

Solution:

$$\begin{aligned} p(z \mid x, \alpha, \beta) &= \frac{p(z, x \mid \alpha, \beta)}{p(x \mid \alpha, \beta)} \\ &\propto p(z, x \mid \alpha, \beta) \end{aligned}$$

where $\propto$ indicates that since we consider the data x as fixed, these expressions are proportional to each other (related via a constant that is not a function of z). Ignoring the normalizing constant to deduce the form of the distribution is a common technique for deriving posteriors. Continuing,

$$\begin{aligned} p(z, x \mid \alpha, \beta) &= p(x \mid z, \alpha, \beta) p(z \mid \alpha, \beta) && (1) \\ &= \binom{n}{x} z^x (1-z)^{n-x} \cdot \frac{z^{\alpha-1}(1-z)^{\beta-1}}{B(\alpha, \beta)} && (2) \\ &\propto_z z^{x+\alpha-1}(1-z)^{n-x+\beta-1} && (3) \end{aligned}$$

where, again, we can ignore constants that are not functions of z to figure out the distributional form.

Note that this expression $z^{x+\alpha-1}(1-z)^{n-x+\beta-1}$ is the (unnormalized) probability density function of the $\mathrm{Beta}(x+\alpha, n-x+\beta)$ distribution. Therefore, we can conclude that in the E-step, we set $q^{(t)}$ to be the $\mathrm{Beta}(x+\alpha^{(t)}, n-x+\beta^{(t)})$ distribution.

As review, this derivation shows again that the beta and binomial distributions are *conjugate distributions*: if the prior is a beta distribution and the likelihood is a binomial distribution, the posterior is conveniently also a beta distribution (with different parameters), as opposed to some arbitrary distribution that would make life as a Bayesian difficult.

2. Now we derive the maximization step.

Solution: Given $q^{(t)}$ from the previous step, we first need to compute

$$\mathbb{E}_{Z \sim q^{(t)}}[\log \mathbb{P}(X, Z \mid \alpha, \beta)].$$

Note that we're not taking the expectation with respect to X, only with respect to Z. Recall from the previous part that

$$\begin{aligned} p(z, x \mid \alpha, \beta) &= \binom{n}{x} z^x (1-z)^{n-x} \cdot \frac{z^{\alpha-1}(1-z)^{\beta-1}}{B(\alpha, \beta)} \\ &\propto_{\alpha,\beta} \frac{1}{B(\alpha, \beta)} z^{x+\alpha-1}(1-z)^{n-x+\beta-1}. \end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}_{z\sim q^{(t)}}[\log p(z,x \mid \alpha,\beta)] \propto_{\alpha,\beta} \mathbb{E}_{z\sim q^{(t)}} & \left[\log\left(\frac{1}{B(\alpha,\beta)} z^{x+\alpha-1}(1-z)^{n-x+\beta-1}\right)\right] \\
&= \mathbb{E}_{z\sim q^{(t)}}[(x+\alpha-1)\log z + (n-x+\beta-1)\log(1-z) \\
&\quad - \log(B(\alpha,\beta))] \\
&= (x+\alpha-1)\mathbb{E}_{z\sim q^{(t)}}[\log z] + (n-x+\beta-1)\mathbb{E}_{z\sim q^{(t)}}[\log(1-z)] \\
&\quad - \log(B(\alpha,\beta)).
\end{aligned}$$

Since $q^{(t)}$ was set to be the $\text{Beta}(x+\alpha^{(t)}, n-x+\beta^{(t)})$ distribution, according to Wikipedia,

$$\mathbb{E}_{z\sim q^{(t)}}[\log z] = \psi(x+\alpha^{(t)}) - \psi(\alpha^{(t)}+n+\beta^{(t)}),$$

where $\psi(x)$ is known as the Digamma function:

$$\psi(x) = \frac{d}{dx}\log\Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}.$$

To compute $\mathbb{E}_{z\sim q^{(t)}}[\log(1-z)]$, note that if $z \sim \text{Beta}(\alpha,\beta)$, then $(1-z) \sim \text{Beta}(\beta,\alpha)$. Therefore, letting $q^{(t)\prime}$ be $\text{Beta}(n-x+\beta^{(t)}, x+\alpha^{(t)})$, we have

$$\mathbb{E}_{z\sim q^{(t)}}[\log(1-z)] = \mathbb{E}_{(1-z)\sim q^{(t)\prime}}[\log(1-z)] = \psi(n-x+\beta^{(t)}) - \psi(\alpha^{(t)}+n+\beta^{(t)}).$$

Combining these,

$$\begin{aligned}
&\mathbb{E}_{z\sim q^{(t)}}[\log p(z,x \mid \alpha,\beta)] \\
&\propto (x+\alpha-1)(\psi(x+\alpha^{(t)}) - \psi(\alpha^{(t)}+n+\beta^{(t)})) \\
&\quad + (n-x+\beta-1)(\psi(n-x+\beta^{(t)}) - \psi(\alpha^{(t)}+n+\beta^{(t)})) \\
&\quad - \log(B(\alpha,\beta)).
\end{aligned} \tag{4}$$

Important clarification: The distribution $q^{(t)}$ from the "E" step is considered to be fixed during the maximization, and we don't maximize over the parameters of $q^{(t)}$ during the "M" step. That is, when maximizing over α, β, we consider the previously computed $\alpha^{(t)}$ and $\beta^{(t)}$ that the $q^{(t)}$ distribution depends on to be constants.

Simplifying Equation (5) to only include terms that depend on α, β, we have

$$\begin{aligned}
&\mathbb{E}_{z\sim q^{(t)}}[\log p(z,x \mid \alpha,\beta)] \\
&\propto_{\alpha,\beta} \alpha(\psi(x+\alpha^{(t)}) - \psi(\alpha^{(t)}+n+\beta^{(t)})) \\
&\quad + \beta(\psi(n-x+\beta^{(t)}) - \psi(\alpha^{(t)}+n+\beta^{(t)})) \\
&\quad - \log(B(\alpha,\beta)).
\end{aligned} \tag{5}$$

To complete the maximization step, we want to find α, β that maximize $\mathbb{E}_{z\sim q^{(t)}}[\log p(z, x \mid \alpha, \beta)]$:

$$\alpha^{(t+1)}, \beta^{(t+1)} \leftarrow \operatorname*{argmax}_{\alpha,\beta>0} \mathbb{E}_{z\sim q^{(t)}}[\log p(z, x \mid \alpha, \beta)].$$

For this problem, it's not easy to solve for $\alpha^{(t+1)}, \beta^{(t+1)}$ in closed form. One way to do it would be to take the derivative of $\mathbb{E}_{z\sim q^{(t)}}[\log p(z, x \mid \alpha, \beta)]$ and find the values of α, β that set the derivative to 0. Maximizing $\mathbb{E}_{z\sim q^{(t)}}[\log p(z, x \mid \alpha, \beta)]$ can also be done approximately using numerical solvers (e.g. scipy.optimize). Once we have found $\alpha^{(t+1)}, \beta^{(t+1)}$ that (approximately) maximize $\mathbb{E}_{z\sim q^{(t)}}[\log p(z, x \mid \alpha, \beta)]$, we will have completed the "M" step.