Applied Probabilistic Programming & Bayesian Machine Learning

§3-3: Normal Approximations to the Joint Posterior

MIT IAP 2017 Max W. Shen

Bayesian inference of the full posterior distribution is much easier with Stan than with traditional methods, but Stan still takes effort and time. MLE and MAP estimates can be obtained quickly through numerical optimization, conveniently available in common R and Python packages.

However, if a quick approximation of the full posterior distributions is what we'd like the most, then we can turn to normal approximations to the posterior.

This approach works best when there are just a few parameters and the distributions are unimodal and roughly symmetric. It finds a normal approximation to the joint posterior distribution over all parameters.

1 Approach

The general approach is to find the posterior mode $\hat{\theta}$ of our parameters θ, then use a truncated Taylor expansion centered at the posterior mode to approximate our full posterior distribution. We will work the log posterior, $\log p(\theta|y)$, and use a Taylor expansion truncated at the 2nd power - this will give us a normal distribution, since the normal distribution takes the exponential of a quadratic function by the Gaussian probability density function:

$$\begin{aligned} p(x|\mu,\sigma^2) &= \frac{1}{\sqrt{2\sigma^2\pi}}\exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big) \\ p(x|\mu,\sigma^2) &\propto \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big) \\ \log p(x|\mu,\sigma^2) &\propto -\frac{(x-\mu)^2}{2\sigma^2} \end{aligned} \tag{1}$$

Now, our Taylor expansion for $\log p(\theta|y)$ centered at $\hat{\theta}$ is:

$$\log p(\theta|y) \approx \log p(\hat{\theta}|y) + \cancel{(\theta-\hat{\theta})\Big(\frac{d}{d\theta}\log p(\theta|y)\Big)_{\theta=\hat{\theta}}} + \frac{1}{2}(\theta-\hat{\theta})^T\Big(\frac{d^2}{d\theta^2}\log p(\theta|y)\Big)_{\theta=\hat{\theta}}(\theta-\hat{\theta}) + ... \tag{2}$$

The first derivative at the posterior mode is 0, by definition, so we can simplify to:

$$\log p(\theta|y) \approx \cancel{\log p(\hat{\theta}|y)} + \frac{1}{2}(\theta-\hat{\theta})^T\Big(\frac{d^2}{d\theta^2}\log p(\theta|y)\Big)_{\theta=\hat{\theta}}(\theta-\hat{\theta}) + ... \tag{3}$$

Next, we note that we are interested in approximating the function $\log p(\theta|y)$, which means our first term, $\log p(\hat{\theta}|y)$ is a constant, so we can ignore it. We also ignore higher-order terms, leaving us with:

$$\log p(\theta|y) \approx \frac{1}{2}(\theta - \hat{\theta})^T \Big(\frac{d^2}{d\theta^2} \log p(\theta|y) \Big)_{\theta=\hat{\theta}} (\theta - \hat{\theta}) \tag{4}$$

Comparing to equation 1, we can rewrite 4 as:

$$p(\theta|y) \approx Normal\Big(\text{mean} = \hat{\theta}, \text{variance} = \Big(- \frac{d^2}{d\hat{\theta}^2} \log p(\hat{\theta}|y) \Big)^{-1} \Big) \tag{5}$$

This is our normal approximation to the posterior. If θ is a vector of size N, then our variance is an N-by-N matrix.

2 In Practice

In practice, if you can analytically write down the posterior distribution, then:

1. Take the first derivative with respect to each parameter to find the posterior mode (at the critical point) for each parameter. This gives you the mean for your normal approximation.
2. Take the second derivative at the posterior mode points to get your matrix of second derivatives, and then invert this matrix.

This approach can work better when applied to a few parameters, marginalizing out the remaining parameters, instead of approximating the full joint distribution.

In practice, the major limitation of this approach is that posterior distributions can only be analytically written for simple models. In the future, it seems reasonable that automatic differentiation could be used to automatically, quickly, calculate a normal approximation the joint posterior distribution.

3 Bibliography and Additional Resources

Chapter 4 of Bayesian Data Analysis 3 discusses normal approximations to the posterior in greater depth.

The Stan reference is available here: http://mc-stan.org/documentation/