Note: Marginal likelihood for Bayesian models with a Gaussian-approximated posterior

M. Rule

June 22, 2017

I first learned this solution from Botond Cseke. I'm not sure where it originates; It is a corollary of Laplace's method for approximating integrals using a Gaussian distribution.

If I have a Bayesian statistical model with hyperparameters Θ, with a no closed-form posterior, how can I optimize Θ?

Consider a Bayesian statistical model with observed data $\mathbf{y} \in \mathcal{Y}$ and hidden (latent) variables $\mathbf{z} \in \mathcal{Z}$, which we infer. We have a prior on $\mathbf{z}$, $\Pr(\mathbf{z};\Theta)$, and a model for the probability of $\mathbf{y}$ given $\mathbf{z}$ (likelihood), $\Pr(\mathbf{y}|\mathbf{z};\Theta)$. The prior and likelihood are controlled by "hyperparameters" Θ, which we would like to estimate. Recall that Bayes theorem states:

$$\Pr(\mathbf{z}|\mathbf{y};\Theta) = \Pr(\mathbf{y}|\mathbf{z};\Theta)\frac{\Pr(\mathbf{z};\Theta)}{\Pr(\mathbf{y};\Theta)} \tag{1}$$

It is common for the posterior $\Pr(\mathbf{z}|\mathbf{y};\Theta)$ to lack a closed-form solution. In this case, one typically approximates the posterior with a more tractable distribution $Q(\mathbf{z}) \approx \Pr(\mathbf{z}|\mathbf{y};\Theta)$. Common ways of estimating $(\mathbf{z})$ include the Laplace approximation, variational Bayes, expectation propagation, and expectation maximization algorithms. The only approximating distribution in common use for high-dimensional $\mathbf{z}$ is the multivariate Gaussian (or some nonlinear transformation thereof), which succinctly captures joint statistics with limited computational overhead. Assume we have an inference procedure which returns the approximate posterior $Q(\mathbf{z}) = \mathcal{N}(\boldsymbol{\mu}_q, \Sigma_q)$.

We optimize the hyperparameters "Θ" of the prior kernel to maximize the marginal likelihood of the observations $\mathbf{y}$

$$\begin{aligned} \theta &\leftarrow \underset{\Theta}{\operatorname{argmax}} \Pr(\mathbf{y};\Theta) \\ \Pr(\mathbf{y};\Theta) &= \int_{\mathcal{Z}} \Pr(\mathbf{y},\mathbf{z};\Theta)\, d\mathbf{z} = \int_{\mathcal{Z}} \Pr(\mathbf{y}|\mathbf{z}) \Pr(\mathbf{z};\Theta)\, d\mathbf{z} \end{aligned} \tag{2}$$

Except in rare special cases, this integral does not have a closed form. However, we have already obtained a Gaussian approximation to the posterior distribution, $Q(\mathbf{z}) \approx \Pr(\mathbf{z}|\mathbf{y};\Theta)$. If we replace $\Pr(\mathbf{z}|\mathbf{y};\Theta)$ with our approximation $Q(\mathbf{z})$ in this equation, we can solve for (an approximation) of $\Pr(\mathbf{y};\Theta)$:

$$Q(\mathbf{z}) \approx \Pr(\mathbf{y}|\mathbf{z})\frac{\Pr(\mathbf{z};\Theta)}{\Pr(\mathbf{y};\Theta)} \Rightarrow \Pr(\mathbf{y};\Theta) \approx \Pr(\mathbf{y}|\mathbf{z})\frac{\Pr(\mathbf{z};\Theta)}{Q(\mathbf{z})} \tag{3}$$

Working in log-probability, and evaluating the expression at the (approximated) posterior mean $\mathbf{z} = \boldsymbol{\mu}_q$, we get

$$
\begin{aligned}
\ln \Pr(\mathbf{z}{=}\boldsymbol{\mu}_q;\Theta) &= -\tfrac{1}{2}\left\{\ln|2\pi\Sigma_z| + (\boldsymbol{\mu}_q-\boldsymbol{\mu}_z)^\top\Sigma_z^{-1}(\boldsymbol{\mu}_q-\boldsymbol{\mu}_z)\right\} \\
\ln Q(\mathbf{z}{=}\boldsymbol{\mu}_q) &= -\tfrac{1}{2}\left\{\ln|2\pi\Sigma_q| + (\boldsymbol{\mu}_q-\boldsymbol{\mu}_q)^\top\Sigma_q^{-1}(\boldsymbol{\mu}_q-\boldsymbol{\mu}_q)\right\} = -\tfrac{1}{2}\ln|2\pi\Sigma_q| \\
\ln \Pr(\mathbf{y};\Theta) &\approx \ln \Pr(\mathbf{y}|\mathbf{z}{=}\boldsymbol{\mu}_q;\Theta) + \ln \Pr(\mathbf{z}{=}\boldsymbol{\mu}_q;\Theta) - \ln Q(\mathbf{z}{=}\boldsymbol{\mu}_q) \\
&= \ln \Pr(\mathbf{y}|\mathbf{z}{=}\boldsymbol{\mu}_q;\Theta) - \tfrac{1}{2}\left\{\ln|\Sigma_q^{-1}\Sigma_z| + (\boldsymbol{\mu}_q-\boldsymbol{\mu}_z)^\top\Sigma_z^{-1}(\boldsymbol{\mu}_q-\boldsymbol{\mu}_z)\right\}
\end{aligned} \tag{4}
$$

.

This is quite tractable to compute.