Bayesian Model Comparison

Xinghu Yao

September 25, 2018

1 Question

The Bayesian view of model comparison simply involves the use of probabilities to represent uncertainty in the choice of model, along with a consistent application of the sum and product rules of probability. Consider a data set $\mathcal{D}$ and a set of models $\{\mathcal{M}_i\}$ having parameters $\{\boldsymbol{\theta}_i\}$. For each model we define a likelihood function $p(\mathcal{D}|\boldsymbol{\theta}_i, \mathcal{M}_i)$. If we introduce a prior $p(\boldsymbol{\theta}_i|\mathcal{M}_i)$ for the various models. Try to approximate the distribution as follows using the Laplace Approximation.

$$p(\mathcal{D}) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \tag{1}$$

Note that in Eq. (1) each $\mathcal{M}_i$ is omitted to keep the notation uncluttered.

2 The Laplace Approximation

Laplace approximation is a simple but widely used framework which aims to find a Gaussian approximations to a probability density defined over a set of continuous variables. For a given distribution $p(\mathbf{z}) = f(\mathbf{z})/\mathbf{Z}$ defined over an M-dimensional space $\mathbf{z}$. At a stationary point $\mathbf{z}_0$ the gradient $\nabla f(\mathbf{z})$ will vanish. We therefor consider a Taylor expansion of $\ln f(\mathbf{z})$ centred on the mode $\mathbf{z}_0$ so that

$$\ln f(\mathbf{z}) \backsimeq \ln f(\mathbf{z}_0) - \frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^T\mathbf{A}(\mathbf{z}-\mathbf{z}_0) \tag{2}$$

Where the $M \times M$ Hessian matrix A is defined by

$$\mathbf{A} = -\nabla\nabla \ln f(\mathbf{z})|_{\mathbf{z}=\mathbf{z}_0} \tag{3}$$

and ∇ is the gradient operator. Taking the exponential of both sides we obtain

$$f(\mathbf{z}) \backsimeq f(\mathbf{z}_0)\exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^T\mathbf{A}(\mathbf{z}-\mathbf{z}_0)\right\} \tag{4}$$

The distribution $q(\mathbf{z})$ is proportional to $f(\mathbf{z})$ and the appropriate normalization coefficient can be found by using the standard result for a normalized multivariate Gaussian, giving

$$q(\mathbf{z}) = \frac{|\mathbf{A}|^{1/2}}{(2\pi)^{M/2}}\exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^T\mathbf{A}(\mathbf{z}-\mathbf{z}_0)\right\} = \mathcal{N}(\mathbf{z}|\mathbf{z}_0, \mathbf{A}^{-1}) \tag{5}$$

where $|\mathbf{A}|$ denotes the determinant of $\mathbf{A}$. This Gaussian distribution will be well defined provided its precision matrix, given by A, is positive definite, which implies that the stationary point $\mathbf{z}_0$ must be a local maximum, not a minimum of a saddle point. In fact, we can also obtain an approximation to the normalization Z. Because we have

$$q(\mathbf{z}) \backsimeq p(\mathbf{z}) = \frac{1}{Z}f(\mathbf{z}) \tag{6}$$

Thus, the approximation to the normalization constant Z have the following form

$$Z \backsimeq \frac{f(\mathbf{z})}{q(\mathbf{z})} = f(\mathbf{z}_0)\frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}} \tag{7}$$

3 Solution

Identifying $f(\boldsymbol{\theta}) = p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ and $Z = p(\mathcal{D})$. According to Eq. (7), we have

$$\begin{aligned} \ln p(\mathcal{D}) = \ln(Z) &\backsimeq \ln f(\mathbf{z}_0)\frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}} \\ &= \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) + \ln p(\boldsymbol{\theta}_{\mathrm{MAP}}) + \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}| \end{aligned} \tag{8}$$

where $\boldsymbol{\theta}_{\mathrm{MAP}}$ is the value of $\boldsymbol{\theta}$ at the mode of the posterior distribution, and $\mathbf{A}$ is the Hessian matrix of second derivatives of the negative log posterior

$$\mathbf{A} = -\nabla\nabla\ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}})p(\boldsymbol{\theta}_{\mathrm{MAP}}) = -\nabla\nabla\ln p(\boldsymbol{\theta}_{\mathrm{MAP}}|\mathcal{D}) \tag{9}$$

From Eq. (9), we have

$$\begin{aligned} \mathbf{A} &= -\nabla\nabla\ln p(\mathcal{D}|\theta_{\mathrm{MAP}})p(\theta_{MAP}) \\ &= \mathbf{H} - \nabla\nabla\ln p(\boldsymbol{\theta}_{\mathrm{MAP}}) \end{aligned} \tag{10}$$

and if $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}, \mathbf{V}_0)$, this becomes

$$\mathbf{A} = \mathbf{H} + \mathbf{V}_0^{-1}. \tag{11}$$

If we assume that the proir is broad or equivalently that the number of data points is large, we can neglect the term $\mathbf{V}_0^{-1}$ compared to $\mathbf{H}$. Using this result, Eq. (8) can be rewritten in the form

$$\ln p(\mathcal{D}) \backsimeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) - \frac{1}{2}(\boldsymbol{\theta}_{\mathrm{MAP}} - \mathbf{m})^T\mathbf{V}_0^{-1}(\boldsymbol{\theta}_{\mathrm{MAP}} - \mathbf{m}) - \frac{1}{2}\ln|\mathbf{H}| + \mathrm{const} \tag{12}$$

We now again invoke the broad prior assumption, allowing us to neglect the second term on the right hand side of Eq. (12). Since we assume i.i.d data, $\mathbf{H} = -\nabla\nabla\ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}})$ consists of a sum of terms, one term for each datum, and we can consider the following approximation:

$$\mathbf{H} = \sum_{n=1}^{N}\mathbf{H}_n = N\widehat{\mathbf{H}} \tag{13}$$

where $\mathbf{H}_n$ is the contribution from the n^{th} data point and

$$\widehat{\mathbf{H}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{H}_n \tag{14}$$

Combining this with the properties of the determinant, we have

$$\ln|\mathbf{H}| = \ln|N\widehat{\mathbf{H}}| = \ln\left(N^M|\widehat{\mathbf{H}}|\right) = M\ln N + \ln|\widehat{\mathbf{H}}| \tag{15}$$

Where M is the dimensionality of $\boldsymbol{\theta}$. Note that we are assuming that $\widehat{\mathbf{H}}$ has full rank M. Finally, using this result together Eq. (12), we obtain

$$\ln p(\mathcal{D}) \backsimeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) - \frac{1}{2}M\ln N \tag{16}$$