Bayesian Logistic Classification

3F8: Inference Coursework
Theo Brown
Selwyn College, University of Cambridge

February 25, 2022

Abstract

Abstract goes here

1 Introduction

2 Laplace approximation

To define a probability distribution function (PDF) $p(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^N$ it is necessary to integrate a function to find the normalising constant K:

$$p(\boldsymbol{x}) := \frac{f(\boldsymbol{x})}{\int f(\boldsymbol{x}) d\boldsymbol{x}} = \frac{1}{K} f(\boldsymbol{x}) \tag{1}$$

In many cases, the integral $\int f(\boldsymbol{x}) d\boldsymbol{x}$ is intractable and does not have a closed-form solution.

The Laplace approximation finds a Gaussian, $q(\boldsymbol{x})$, that provides a good approximation to $p(\boldsymbol{x})$ near a local maximum of $p(\boldsymbol{x})$. As $q(\boldsymbol{x})$ is Gaussian, its normalising constant is easy to find, so the problem of solving $\int f(\boldsymbol{x}) d\boldsymbol{x}$ is avoided.

Initially, consider the truncated Taylor expansion of $\ln f(\boldsymbol{x})$ around a local maximum $\boldsymbol{x}_0$:

$$\ln f(\boldsymbol{x}) \approx \ln f(\boldsymbol{x}_0) + \nabla \ln f(\boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_0} (\boldsymbol{x} - \boldsymbol{x}_0) + \frac{1}{2} (\boldsymbol{x} - \boldsymbol{x}_0)^T \nabla^2 \ln f(\boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_0} (\boldsymbol{x} - \boldsymbol{x}_0) \tag{2}$$

At a maximum of $f(\boldsymbol{x})$, $\nabla \ln f(\boldsymbol{x}) = 0$ as the logarithm is a monotonic function. Hence:

$$\begin{aligned} \ln f(\boldsymbol{x}) &\approx \ln f(\boldsymbol{x}_0) + \frac{1}{2} (\boldsymbol{x} - \boldsymbol{x}_0)^T \nabla^2 \ln f(\boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_0} (\boldsymbol{x} - \boldsymbol{x}_0) \\ f(\boldsymbol{x}) &\approx f(\boldsymbol{x}_0) \exp\left(\frac{1}{2} (\boldsymbol{x} - \boldsymbol{x}_0)^T \nabla^2 \ln f(\boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_0} (\boldsymbol{x} - \boldsymbol{x}_0) \right) \end{aligned} \tag{3}$$

Equation 3 is of the form of an un-normalised Gaussian. Let $\boldsymbol{P} = -\nabla^2 \ln f(\boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_0}$, and normalise Equation 3 to get an approximation for $p(\boldsymbol{x})$:

$$\begin{aligned} p(\boldsymbol{x}) &\approx \frac{1}{(2\pi)^{\frac{M}{2}} \det \boldsymbol{P}^{-\frac{1}{2}}} \exp\left(\frac{1}{2} (\boldsymbol{x} - \boldsymbol{x}_0)^T \boldsymbol{P} (\boldsymbol{x} - \boldsymbol{x}_0) \right) \\ &= \mathcal{N}(\boldsymbol{x}; \boldsymbol{x}_0, \boldsymbol{P}^{-1}) \end{aligned} \tag{4}$$

For this to hold, $\boldsymbol{P}$ must be positive definite, i.e. $\boldsymbol{x}_0$ must be a maximum.

This method can be used instead to find an approximate value of the normalising constant K. Substituting Equation 3 into the definition of K:

$$\begin{aligned} K = \int f(\boldsymbol{x}) d\boldsymbol{x} &\approx f(\boldsymbol{x}_0) \int \exp\left(\frac{1}{2} (\boldsymbol{x} - \boldsymbol{x}_0)^T \boldsymbol{P} (\boldsymbol{x} - \boldsymbol{x}_0) \right) d\boldsymbol{x} \\ &= \frac{(2\pi)^{\frac{N}{2}}}{\det \boldsymbol{P}^{\frac{1}{2}}} f(\boldsymbol{x}_0) \end{aligned} \tag{5}$$

3 Recovering Bayesian logistic regression

3.1 Posterior distribution

In the previous report, the maximum-likelihood estimate for the model weights was used in classification[1]. It is desirable to have a full posterior distribution of the weights:

$$p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) = \frac{p(\boldsymbol{X}|\boldsymbol{w}, \boldsymbol{y})p(\boldsymbol{w})}{p(\boldsymbol{X}|\boldsymbol{y})} \tag{6}$$

The problem is that calculating Equation 6 requires normalising a product of logistic functions, which is difficult. The Laplace approximation can be used to calculate the normalising constant $p(\boldsymbol{X}|\boldsymbol{y})$, also called the model evidence, and calculate an approximate posterior distribution.

Given that the posterior Laplace approximation will be Gaussian, it is sensible to choose a Gaussian prior:

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{m}_0, \boldsymbol{C}_0) \tag{7}$$

In order to apply the Laplace approximation, we need the location of a maximum in the posterior. Using the results for the log-likelihood of the logistic model and the log of Equation 7, the log-posterior is:

$$\begin{aligned}\ln p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) = &\sum_{n=1}^{N} y_n \log \sigma(\boldsymbol{w}^T \tilde{\boldsymbol{x}}_n) + (1 - y_n) \log \sigma(-\boldsymbol{w}^T \tilde{\boldsymbol{x}}_n) \\ &+ \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_0)^T \boldsymbol{C}_0^{-1} (\boldsymbol{w} - \boldsymbol{w}_0) \\ &+ \text{const}\end{aligned} \tag{8}$$

The value of $\boldsymbol{w}$ at the maximum, $\boldsymbol{w}_{\text{MAP}}$, can be found by setting the derivative of Equation 8 to zero. As outlined in Section 2 $\boldsymbol{w}_{\text{MAP}}$ will be the mean of $q(\boldsymbol{w})$.

Using Equation 4, the covariance matrix of $q(\boldsymbol{w})$ can be found:

$$\begin{aligned}\boldsymbol{C}_N^{-1} &= -\nabla^2 \ln p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y})\big|_{\boldsymbol{w}=\boldsymbol{w}_{\text{MAP}}} \\ &= \boldsymbol{C}_0^{-1} + \sum_{n=1}^{N} \sigma(\boldsymbol{w}^T \tilde{\boldsymbol{x}}_N)\sigma(-\boldsymbol{w}^T \tilde{\boldsymbol{x}}_N)\tilde{\boldsymbol{x}}_N \tilde{\boldsymbol{x}}_N^T\end{aligned} \tag{9}$$

Hence, using Equation 4 and Equation 5:

$$p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) \approx \mathcal{N}(\boldsymbol{w}; \boldsymbol{w}_{\text{MAP}}, \boldsymbol{C}_N) \tag{10}$$

$$p(\boldsymbol{X}|\boldsymbol{y}) = \frac{(2\pi)^{\frac{N}{2}}}{\det \boldsymbol{C}_N^{-\frac{1}{2}}} p(\boldsymbol{X}|\boldsymbol{w} = \boldsymbol{w}_{\text{MAP}}, \boldsymbol{y}) p(\boldsymbol{w} = \boldsymbol{w}_{\text{MAP}}) \tag{11}$$

3.2 Predictive distribution

3.3 Application

In this application, the prior on the model weights is chosen to be:

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; 0, \boldsymbol{I}\sigma_0^2) \tag{12}$$

[1] For the model definition, and definitions of $\boldsymbol{X}$, $\boldsymbol{y}$, $\tilde{\boldsymbol{x}}$, etc, see the previous report.