Bayesian Logistic Classification

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

March 12, 2022

Abstract

Abstract goes here

1 Introduction

2 Laplace approximation for logistic regression

The Laplace approximation finds a Gaussian, $q(\boldsymbol{x})$, that provides a good approximation to a probability distribution $p(\boldsymbol{x})$ near a local maximum of $p(\boldsymbol{x})$. This is useful when the desired distribution $p(\boldsymbol{x})$ is intractable and difficult to manipulate.

In the previous report, the maximum-likelihood estimate for the model weights was used in classification[1]. For Bayesian classification, a full posterior distribution of the weights is required:

$$p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) = \frac{p(\boldsymbol{X}|\boldsymbol{w}, \boldsymbol{y})p(\boldsymbol{w})}{p(\boldsymbol{X}|\boldsymbol{y})} \tag{1}$$

Calculating the normalising constant (also called the evidence) $p(\boldsymbol{X}|\boldsymbol{y})$ requires integrating a product of logistic functions, which is intractable. The Laplace approximation can be used to find an approximate posterior distribution and an approximate predictive distribution for the logistic classifier, while avoiding the difficult integral. The derivation of these results is found in Appendix A.

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

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

This gives the Laplace approximation of the posterior as:

$$p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) \approx \mathcal{N}(\boldsymbol{w}; \boldsymbol{w}_{\text{MAP}}, \boldsymbol{C}_N) \tag{3}$$

And the predictive distribution as:

$$p(y^* = 1|\boldsymbol{x}^*, \boldsymbol{y}, \boldsymbol{X}) \approx \sigma\left(\frac{\mu_p}{\sqrt{1 + \sigma_p^2\lambda^2}}\right) \tag{4}$$

where

$$\boldsymbol{C}_N = \left(\frac{1}{\sigma_0^2}\boldsymbol{I} + \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T\boldsymbol{\sigma}(1 - \boldsymbol{\sigma})\right)^{-1} \tag{5}$$

$$\boldsymbol{\sigma} = \sigma(\tilde{\boldsymbol{X}}^T\boldsymbol{w}) \tag{6}$$

$$\mu_p = \boldsymbol{x}^{*T}\boldsymbol{w}_{\text{MAP}} \tag{7}$$

$$\sigma_p^2 = \boldsymbol{x}^{*T}\boldsymbol{C}_N\boldsymbol{x}^* \tag{8}$$

$$\lambda = \sqrt{\frac{\pi}{8}} \tag{9}$$

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

A Deriving Bayesian Logistic Classification using the Laplace Approximation

A.1 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:

$$p(\boldsymbol{x}) := \frac{f(\boldsymbol{x})}{\int f(\boldsymbol{x}) d\boldsymbol{x}} = \frac{1}{K} f(\boldsymbol{x}) \tag{10}$$

In many cases, the integral $\int f(\boldsymbol{x}) d\boldsymbol{x}$ is intractable or 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. To find $q(\boldsymbol{x})$, we start with the truncated Taylor expansion of $\ln f(\boldsymbol{x})$ around a local maximum $\boldsymbol{x}_0$:

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

At a maximum of $f(\boldsymbol{x})$, $\nabla \ln f(\boldsymbol{x}) = 0$ as the logarithm is a monotonic function. Hence, close to $\boldsymbol{x}_0$:

$$\begin{aligned} \ln f(\boldsymbol{x}) &\approx \ln f(\boldsymbol{x}_0) + \frac{1}{2} (\boldsymbol{x} - \boldsymbol{x}_0)^T \nabla^2 \ln f(\boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_0} (\boldsymbol{x} - \boldsymbol{x}_0) \\ f(\boldsymbol{x}) &\approx f(\boldsymbol{x}_0) \exp\left(\frac{1}{2} (\boldsymbol{x} - \boldsymbol{x}_0)^T \nabla^2 \ln f(\boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_0} (\boldsymbol{x} - \boldsymbol{x}_0) \right) \end{aligned} \tag{11}$$

Equation 11 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 11 to get an approximation for $p(\boldsymbol{x})$:

$$\begin{aligned} p(\boldsymbol{x}) &\approx \frac{1}{(2\pi)^{\frac{M}{2}} \det \boldsymbol{P}^{-\frac{1}{2}}} \exp\left(\frac{1}{2} (\boldsymbol{x} - \boldsymbol{x}_0)^T \boldsymbol{P} (\boldsymbol{x} - \boldsymbol{x}_0) \right) \\ &= \mathcal{N}(\boldsymbol{x}; \boldsymbol{x}_0, \boldsymbol{P}^{-1}) \end{aligned} \tag{12}$$

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 11 into the definition of K:

$$\begin{aligned} K = \int f(\boldsymbol{x}) d\boldsymbol{x} &\approx f(\boldsymbol{x}_0) \int \exp\left(\frac{1}{2} (\boldsymbol{x} - \boldsymbol{x}_0)^T \boldsymbol{P} (\boldsymbol{x} - \boldsymbol{x}_0) \right) d\boldsymbol{x} \\ &= \frac{(2\pi)^{\frac{N}{2}}}{\det \boldsymbol{P}^{\frac{1}{2}}} f(\boldsymbol{x}_0) \end{aligned} \tag{13}$$

A.2 Bayesian logistic regression

A.2.1 Posterior distribution

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

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{m}_0, \boldsymbol{C}_0) \tag{14}$$

In order to apply the Laplace approximation, we need the location of a maximum in the posterior. Using the results for the likelihood of the logistic model in the previous report and the log of Equation 14, the log-posterior of the model weights is:

$$\begin{aligned} \ln p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) = &\sum_{n=1}^{N} y_n \log \sigma(\boldsymbol{w}^T \tilde{\boldsymbol{x}}_n) + (1 - y_n) \log \sigma(-\boldsymbol{w}^T \tilde{\boldsymbol{x}}_n) \\ &+ \frac{1}{2} (\boldsymbol{w} - \boldsymbol{w}_0)^T \boldsymbol{C}_0^{-1} (\boldsymbol{w} - \boldsymbol{w}_0) \\ &+ \text{const} \end{aligned} \tag{15}$$

The value of $\boldsymbol{w}$ at the maximum, $\boldsymbol{w}_{\text{MAP}}$, can be found by setting the derivative of Equation 15 to zero. The mean of $q(\boldsymbol{w})$ will be set to $\boldsymbol{w}_{\text{MAP}}$.

Using Equation 12, the covariance matrix of $q(\boldsymbol{w})$ can be found:

$$\begin{aligned}\boldsymbol{C}_N^{-1} &= -\nabla^2 \ln p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y})\big|_{\boldsymbol{w}=\boldsymbol{w}_{\text{MAP}}} \\ &= \boldsymbol{C}_0^{-1} + \sum_{n=1}^{N} \sigma(\boldsymbol{w}^T\tilde{\boldsymbol{x}}_n)\sigma(-\boldsymbol{w}^T\tilde{\boldsymbol{x}}_n)\tilde{\boldsymbol{x}}_n\tilde{\boldsymbol{x}}_n^T \end{aligned} \tag{16}$$

Defining $\boldsymbol{\sigma} = \sigma(\tilde{\boldsymbol{X}}^T\boldsymbol{w})$, this can be written in vector form as:

$$\boldsymbol{C}_N^{-1} = \boldsymbol{C}_0^{-1} + \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T\boldsymbol{\sigma}(1-\boldsymbol{\sigma}) \tag{17}$$

Hence, using Equation 12, the posterior distribution is:

$$p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) \approx \mathcal{N}(\boldsymbol{w}; \boldsymbol{w}_{\text{MAP}}, \boldsymbol{C}_N) \tag{18}$$

And using Equation 13, the normalizing constant is:

$$p(\boldsymbol{X}|\boldsymbol{y}) = \frac{(2\pi)^{\frac{N}{2}}}{\det \boldsymbol{C}_N^{-\frac{1}{2}}} p(\boldsymbol{X}|\boldsymbol{w}=\boldsymbol{w}_{\text{MAP}}, \boldsymbol{y})p(\boldsymbol{w}=\boldsymbol{w}_{\text{MAP}}) \tag{19}$$

A.3 Predictive distribution

The predictive distribution can also be approximated using the Laplace method:

$$\begin{aligned} p(y^*=1|\boldsymbol{x}^*, \boldsymbol{y}, \boldsymbol{X}) &= \int p(y^*=1|\boldsymbol{x}^*, \boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X})d\boldsymbol{w} \\ &\approx \int \sigma(\boldsymbol{w}^T\boldsymbol{x}^*)q(\boldsymbol{w})d\boldsymbol{w} \end{aligned} \tag{20}$$

Using the sifting property of the delta function:

$$\sigma(\boldsymbol{w}^T\boldsymbol{x}) = \int \delta(a - \boldsymbol{w}^T\boldsymbol{x})\sigma(a)da \tag{21}$$

Hence:

$$\begin{aligned} p(y^*=1|\boldsymbol{x}^*, \boldsymbol{y}, \boldsymbol{X}) &\approx \int\int \delta(a - \boldsymbol{w}^T\boldsymbol{x}^*)\sigma(a)q(\boldsymbol{w})d\boldsymbol{w}da \\ &= \int \sigma(a) \int \delta(a - \boldsymbol{x}^{*T}\boldsymbol{w})q(\boldsymbol{w})d\boldsymbol{w}da \end{aligned}$$

The inner integral applies a linear constraint to $q(\boldsymbol{w})$, as the argument of the delta function is 0 unless $a = \boldsymbol{x}^{*T}\boldsymbol{w}$. Hence, the approximate predictive distribution is:

$$\begin{aligned} p(y^*=1|\boldsymbol{x}^*, \boldsymbol{y}, \boldsymbol{X}) &\approx \int \sigma(a)\mathcal{N}(a; \boldsymbol{x}^{*T}\boldsymbol{w}_{\text{MAP}}, \boldsymbol{x}^{*T}\boldsymbol{C}_N\boldsymbol{x}^*)da \\ &= \int \sigma(a)\mathcal{N}(a; \mu_p, \sigma_p^2)da \end{aligned} \tag{22}$$

Where

$$\mu_p = \boldsymbol{x}^{*T}\boldsymbol{w}_{\text{MAP}} \tag{23}$$

$$\sigma_p^2 = \boldsymbol{x}^{*T}\boldsymbol{C}_N\boldsymbol{x}^* \tag{24}$$

This integral cannot be expressed analytically, so another approximation is required. The logistic function can be approximated well by a probit function scaled such that the gradient of the two functions at the origin are equal. It can be shown that this gives:

$$\sigma(x) \approx \Phi^{-1}\left(\sqrt{\frac{\pi}{8}}x\right) = \Phi^{-1}(\lambda x) \tag{25}$$

Substituting into Equation 22 and evaluating using properties of the probit function gives:

$$\begin{aligned} p(y^* = 1|\boldsymbol{x}^*, \boldsymbol{y}, \boldsymbol{X}) &\approx \int \Phi^{-1}(\lambda x)\mathcal{N}(a; \mu_p, \sigma_p)da \\ &= \Phi^{-1}\left(\frac{\mu_p}{\sqrt{\lambda^{-2} + \sigma_p^2}}\right) \end{aligned} \tag{26}$$

Using Equation 25, this can be converted back into a logistic function:

$$p(y^* = 1|\boldsymbol{x}^*, \boldsymbol{y}, \boldsymbol{X}) \approx \sigma\left(\frac{\mu_p}{\sqrt{1 + \sigma_p^2\lambda^2}}\right) \tag{27}$$