The Laplace approximation applied to Bayesian logistic regression

There are multiple ways that we could try to fit a distribution with a Gaussian form. For example, we could try to match the mean and variance of the distribution. The Laplace approximation is another possible way to approximate a distribution with a Gaussian. It can be seen as an incremental improvement of the MAP approximation to Bayesian inference, and only requires some additional derivative computations.

In Bayesian logistic regression, we can only evaluate the posterior distribution up to a constant: we can evaluate the joint probability $p(\mathbf{w}, \mathcal{D})$, but not the normalizer $P(\mathcal{D})$. We match the shape of the posterior using $p(\mathbf{w}, \mathcal{D})$, and then the approximation can be used to approximate $P(\mathcal{D})$.

The Laplace approximation sets the mode of the Gaussian approximation to the mode of the posterior distribution, and matches the curvature of the log probability density at that location. We need to be able to evaluate first and second derivatives of $\log P(\mathbf{w}, \mathcal{D})$.

The rest of the note just fills in the details. We're not adding much to MacKay's textbook pp341–342, or Murphy's book p255. Although we try to go slightly more slowly and show some pictures of what can go wrong.

1 Matching the distributions

First of all we find the most probable setting of the parameters:

$$\mathbf{w}^* = \underset{\mathbf{w}}{\arg\max}\; p(\mathbf{w} \,|\, \mathcal{D}) = \underset{\mathbf{w}}{\arg\max}\; \log p(\mathbf{w}, \mathcal{D}). \tag{1}$$

The conditional probability on the left is what we intuitively want to optimize. The maximization on the right gives the same answer, but contains the term we will actually compute. Reminder: why do we take the log?[1]

[The website version of this note has a question here.]

We usually find the mode of the distribution by minimizing an 'energy', which is the negative log-probability of the distribution up to a constant. For a posterior distribution, we can define the energy as:

$$E(\mathbf{w}) = -\log p(\mathbf{w}, \mathcal{D}), \qquad \mathbf{w}^* = \underset{\mathbf{w}}{\arg\min}\, E(\mathbf{w}). \tag{2}$$

We minimize it as usual, using a gradient-based numerical optimizer.

The minimum of the energy is a turning point. For a scalar variable w the first derivative $\frac{\partial E}{\partial w}$ is zero and the second derivative gives the curvature of this turning point:

$$H = \left.\frac{\partial^2 E(w)}{\partial w^2}\right|_{w=w^*}. \tag{3}$$

The notation means that we evaluate the second derivative at the optimum, $w = w^*$. If H is large, the slope (the first derivative) changes rapidly from a steep descent to a steep ascent. We should approximate the distribution with a narrow Gaussian. Generalizing to multiple variables $\mathbf{w}$, we know $\nabla_{\mathbf{w}} E$ is zero at the optimum and we evaluate the *Hessian*, a matrix with elements:

$$H_{ij} = \left.\frac{\partial^2 E(\mathbf{w})}{\partial w_i \partial w_j}\right|_{\mathbf{w}=\mathbf{w}^*}. \tag{4}$$

[1] Because log is a monotonic transformation, maximizing the log of a function is equivalent to maximizing the original function. Often the log of a distribution is more convenient to work with, less prone to numerical problems, and closer to an ideal quadratic function that optimizers like.

This matrix tells us how sharply the distribution is peaked in different directions.

For comparison, we can find the optimum and curvature that we would get if our distribution were Gaussian. For a one-dimensional distribution, $\mathcal{N}(\mu, \sigma^2)$, the energy (the negative log-probability up to a constant) is:

$$E_{\mathcal{N}}(w) = \frac{(w-\mu)^2}{2\sigma^2}. \tag{5}$$

The minimum is $w^* = \mu$, and the second derivative $H = 1/\sigma^2$, implying the variance is $\sigma^2 = 1/H$. Generalizing to higher dimensions, for a Gaussian $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$, the energy is:

$$E_{\mathcal{N}}(\mathbf{w}) = \frac{1}{2}(\mathbf{w}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{w}-\boldsymbol{\mu}), \tag{6}$$

with $\mathbf{w}^* = \boldsymbol{\mu}$ and $H = \Sigma^{-1}$, implying the covariance is $\Sigma = H^{-1}$.

Therefore matching the minimum and curvature of the ‘energy’ (negative log-probability) to those of a Gaussian energy gives the Laplace approximation to the posterior distribution:

$$\boxed{p(\mathbf{w}\,|\,\mathcal{D}) \approx \mathcal{N}(\mathbf{w};\, \mathbf{w}^*, H^{-1})} \tag{7}$$

[The website version of this note has a question here.]

2 Approximating the normalizer Z

Evaluating our approximation for a D-dimensional distribution gives:

$$p(\mathbf{w}\,|\,\mathcal{D}) = \frac{p(\mathbf{w}, \mathcal{D})}{P(\mathcal{D})} \approx \mathcal{N}(\mathbf{w};\, \mathbf{w}^*, H^{-1}) = \frac{|H|^{1/2}}{(2\pi)^{D/2}} \exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{w}^*)^\top H(\mathbf{w}-\mathbf{w}^*)\right). \tag{8}$$

At the mode $\mathbf{w}^* = \mathbf{w}$, the exponential term disappears and we get:

$$\frac{p(\mathbf{w}^*, \mathcal{D})}{P(\mathcal{D})} \approx \frac{|H|^{1/2}}{(2\pi)^{D/2}}, \qquad \boxed{P(\mathcal{D}) \approx \frac{p(\mathbf{w}^*, \mathcal{D})(2\pi)^{D/2}}{|H|^{1/2}}}. \tag{9}$$

An equivalent expression is

$$\boxed{P(\mathcal{D}) \approx p(\mathbf{w}^*, \mathcal{D})\, |2\pi H^{-1}|^{1/2},} \tag{10}$$

where $|\cdot|$ means take the determinant of the matrix.

When some people say “the Laplace approximation”, they are referring to this approximation of the normalization $P(\mathcal{D})$, rather than the intermediate Gaussian approximation to the distribution.

3 Computing logistic regression predictions

Now we return to the question of how to make Bayesian predictions (all implicitly conditioned on a set of model choices $\mathcal{M}$):

$$\begin{aligned} P(y\,|\,\mathbf{x}, \mathcal{D}) &= \int p(y, \mathbf{w}\,|\,\mathbf{x}, \mathcal{D})\, \mathrm{d}\mathbf{w} && (11)\\ &= \int P(y\,|\,\mathbf{x}, \mathbf{w})\, p(\mathbf{w}\,|\,\mathcal{D})\, \mathrm{d}\mathbf{w}. && (12) \end{aligned}$$

We can approximate the posterior with a Gaussian, $p(\mathbf{w}\,|\,\mathcal{D}) \approx \mathcal{N}(\mathbf{w};\, \mathbf{w}^*, H^{-1})$, using the Laplace approximation (or variational methods, next week). Using this approximation, we still have an integral with no closed form solution:

$$P(y\!=\!1\,|\,\mathbf{x}, \mathcal{D}) \approx \int \sigma(\mathbf{w}^\top\mathbf{x})\,\mathcal{N}(\mathbf{w};\, \mathbf{w}^*, H^{-1})\;\mathrm{d}\mathbf{w} \tag{13}$$

$$= \mathbb{E}_{\mathcal{N}(\mathbf{w};\,\mathbf{w}^*, H^{-1})}\Big[\sigma(\mathbf{w}^\top\mathbf{x})\Big]. \tag{14}$$

However, this expectation can be simplified. Only the inner product $a = \mathbf{w}^\top\mathbf{x}$ matters, so we can take the average over this scalar quantity instead. The linear combination a is a linear combination of Gaussian beliefs, so our beliefs about it are also Gaussian. By now you should be able to show that

$$p(a) = \mathcal{N}(a;\; {\mathbf{w}^*}^\top\mathbf{x},\; \mathbf{x}^\top H^{-1}\mathbf{x}). \tag{15}$$

Therefore, the predictions given the approximate posterior, are given by a one-dimensional integral:

$$P(y\!=\!1\,|\,\mathbf{x}, \mathcal{D}) \approx \mathbb{E}_{\mathcal{N}(a;\; {\mathbf{w}^*}^\top\mathbf{x},\; \mathbf{x}^\top H^{-1}\mathbf{x})}\left[\sigma(a)\right] \tag{16}$$

$$= \int \sigma(a)\,\mathcal{N}(a;\; {\mathbf{w}^*}^\top\mathbf{x},\; \mathbf{x}^\top H^{-1}\mathbf{x})\;\mathrm{d}a. \tag{17}$$

One-dimensional integrals can be computed numerically to high precision.

Bishop p. 220 and Murphy Section 8.4.4.2 review a further approximation, which is quicker to evaluate and provides an interpretable closed-form expression:

$$P(y\!=\!1\,|\,\mathbf{x}, \mathcal{D}) \approx \sigma(\kappa\, {\mathbf{w}^*}^\top\mathbf{x}), \qquad \kappa = \frac{1}{\sqrt{1 + \frac{\pi}{8}\mathbf{x}^\top H^{-1}\mathbf{x}}}. \tag{18}$$

Under this approximation, the predictions use the most probable or MAP weights. However, the activation is scaled down (with κ) when the activation is uncertain, so that predictions will be less confident far from the data (as they should be).

4 Is the Laplace approximation reasonable?

If we think that the Energy is well-behaved and sharply peaked around the mode of the distribution, we might think that we can approximate it with a Taylor series. In one dimension we write

$$E(w^* + \delta) \approx E(w^*) \;+\; \frac{\partial E}{\partial w}\bigg|_{w^*}\delta \;+\; \frac{1}{2}\,\frac{\partial^2 E}{\partial w^2}\bigg|_{w^*}\delta^2 \tag{19}$$

$$\approx E(w^*) \;+\; \frac{1}{2}H\delta^2, \tag{20}$$

where the second term disappears because $\frac{\partial E}{\partial w}$ is zero at the optimum. In multiple dimensions this Taylor approximation generalizes to:

$$E(\mathbf{w}^* + \boldsymbol{\delta}) \approx E(\mathbf{w}^*) + \tfrac{1}{2}\boldsymbol{\delta}^\top H\boldsymbol{\delta}. \tag{21}$$

A quadratic energy (negative log-probability) implies a Gaussian distribution. The distribution is close to the Gaussian fit when the Taylor series is accurate.

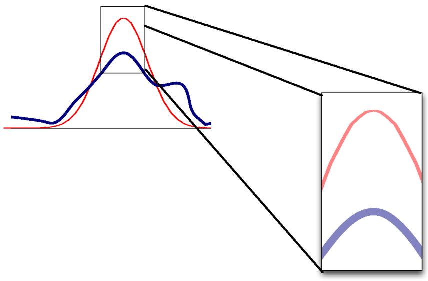
For models with a fixed number of identifiable parameters, the posterior becomes tightly peaked in the limit of large datasets. Then the Taylor expansion of the log-posterior doesn't need to be extrapolated far and will be accurate. Search term for more information: "Bayesian central limit theorem".

5 The Laplace approximation doesn’t always work well!

Despite the theory above, it is easy for the Laplace approximation to go wrong.

In high dimensions, there are many directions in parameter space where there might only be a small number of informative datapoints. Then the posterior could look like the first asymmetrical example in this note.

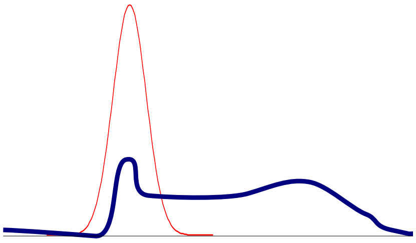
If the mode and curvature are matched, but the distribution is otherwise non-Gaussian, then the value of the densities won’t match[2].



As a result, the approximation of $P(\mathcal{D})$ will be poor.

[The website version of this note has a question here.]

One way for a distribution to be non-Gaussian is to be multi-modal. The posterior of logistic regression only has one mode, but the posterior for neural networks will be multimodal. Even if capturing one mode is reasonable, an optimizer could get stuck in bad local optima.



In models with many parameters, the posterior will often be flat in some direction, where parameters trade off each other to give similar predictions. When there is zero curvature in some direction, the Hessian isn’t positive definite and we can’t get a meaningful approximation.

6 Further Reading

Bishop covers the Laplace approximation and application to Bayesian logistic regression in Sections 4.4 and 4.5.

Or read Murphy Sections 8.4 to 8.4.4 inclusive. You can skip 8.4.2 on BIC.

Similar material is covered by MacKay, Ch. 41, pp492–503, and Ch. 27, pp341–342.

The Laplace approximation was used in some of the earliest Bayesian neural networks although — as presented here — it’s now rarely used. However, the idea does occur in recent work, such as on continual learning (Kirkpatrick et al., Google Deepmind, 2017) and a more sophisticated variant is used by the popular statistical package, R-INLA.

[2] The final two figures in this note come from previous MLPR course notes, by one of Amos Storkey, Chris Williams, or Charles Sutton.