3F8: Inference
Full Technical Report

Author's Name

March 4, 2025

Abstract

This is the abstract.
Try for 1-2 sentences on each of: motive (what it's about), method (what was done), key results and conclusions (the main outcomes).

- Don't exceed 3 sentences on any one.
- Write this last, when you know what it should say!

1 Introduction

1. What is the problem and why is it interesting?
2. What novel follow-up will the rest of your report present?

2 Exercise a)

The normalisation of posterior probabilities is straightforward for conjugate Gaussian pairs, where the model evidence $Z = p(y|X) = \int p(y|X,\beta)p(\beta)d\beta$ can be directly calculated. However, non-conjugate distributions pose a significant challenge as the posterior shape is not predefined. Laplace approximation resolves this issue by fitting a Gaussian $q(\beta)$ at the posterior's local maximum, efficiently approximating the model evidence Z needed for Bayesian logistic regression.

2.1 Bayesian logistic regression

The posterior distribution of model weights β given the data y and X is defined by Bayes' Theorem as:

$$p(\beta|y,X) = \frac{p(y|X,\beta)p(\beta)}{p(y|X)}$$

where $p(y|X,\beta)$ is the likelihood, $p(\beta)$ is the prior distribution, and $p(y|X)$ is the model evidence or marginal likelihood. For model simplicity, the prior is assumed to be Gaussian with zero mean and variance σ_0^2, and the likelihood is chosen as Bernoulli with the logistic sigmoid function $\sigma(\beta^T\phi)$. Denoting $\mathbf{S}_0 = \sigma_0^2\mathbf{I}$, the Gaussian prior is formalised as $p(\beta) = \mathcal{N}(\beta|0,\mathbf{S}_0)$ to unify matrix shape and expressions.

2.2 Laplace approximation of the posterior distribution $p(\beta|y,X)$

The expression of the Laplace approximation $q(\beta)$ can be found using the truncated Taylor expansion. Around the mode β_0 where $\left.\frac{df(\beta)}{d\beta}\right|_{\beta=\beta_0} = 0$, monoticity of the logarithm function gives $\nabla \log f(\beta) = 0$, hence:

$$\log f(\beta) \simeq \log f(\beta_0) - \frac{1}{2}\mathbf{A}(\beta-\beta_0)^2, \quad \mathbf{A} = -\nabla^2 \log f(\beta)\big|_{\beta=\beta_0}$$

Restoring to exponential form yields a Gaussian distribution centered at β_0, which is identified using Maximum A Posteriori (MAP) estimation. With $\beta_0 = \beta_{MAP}$, the covariance of this Gaussian is defined by the Hessian matrix $\mathbf{S}_N^{-1} = -\nabla^2 \log f(\beta)\big|_{\beta=\beta_{MAP}}$. Standard Gaussian normalization method is then applied to yield the posterior approximation:

$$p(\beta|y, X) \approx q(\beta) = \mathcal{N}(\beta|\beta_0, \mathbf{S}_N^{-1}) = \frac{1}{(2\pi)^{\frac{N}{2}} \det \mathbf{S}_N^{-\frac{1}{2}}} \exp\{\frac{1}{2}(\beta - \beta_{MAP})^T \mathbf{S}_N^{-1}(\beta - \beta_{MAP})\}. \tag{1}$$

Using this expression, the Hessian matrix can be simplified as:

$$\mathbf{S}_N^{-1} = -\nabla^2 \log p(\beta|y, X) = \mathbf{S}_0 + \sum_{n=1}^{N} \sigma(\beta^T x_n)(1 - \sigma(\beta^T x_n)) x_n x_n^T \tag{2}$$

2.3 Approximated log model evidence

Using the approximated posterior distribution, the log model evidence can then be calculated as:

$$\begin{aligned} \log p(y|X) &= \log \int p(y|X, \beta) p(\beta) d\beta \\ &\approx \log p(y|X, \beta_{MAP}) + \log p(\beta_{MAP}) - \frac{1}{2} \log |\mathbf{S}_N^{-1}| + \frac{N}{2} \log 2\pi \end{aligned}$$

which is a combination of the log prior, log likelihood, and the negative log of the determinant of the inverse covariance matrix for the posterior.

2.4 Approximated predictive distribution

For the binary classification problem concerned in this exercise, the predictive distribution is then obtained by marginilising the approximated posterior probability obtained in Equation 1. For category $\mathcal{C}_1$, given a new feature vector $\phi(x)$:

$$p(\mathcal{C}_1|\phi, y, X) = \int p(\mathcal{C}_1|\phi, \beta) p(\beta|y, X) d\beta \simeq \int \sigma(\beta^T \phi) q(\beta) d\beta \tag{3}$$

Simplifying Equation 3 using the sifting property of Dirac delta function:

$$p(\mathcal{C}_1|\phi, y, X) = \int \sigma(\beta^T \phi) \mathcal{N}(\beta^T \phi | \beta_{MAP}^T \phi, \phi^T \mathbf{S}_N \phi) = \int \sigma(\beta^T \phi) \mathcal{N}(\mu_{pred}, \sigma_{pred}^2) d\beta$$

so $\mu_{pred} = \beta_{MAP}^T \phi$, $\sigma_{pred}^2 = \phi^T \mathbf{S}_N \phi$. This can be further simplified by approximating the logistic sigmoid with prodit function $\Phi(\lambda x)$ with the scale factor $\lambda^2 = \pi/8$:

$$p(\mathcal{C}_1|\phi, y, X) = \sigma(\kappa(\phi^T S_N \phi) \beta^T \phi | \beta_{MAP}^T \phi), \quad \kappa(\sigma^2) = (1 + \pi\sigma^2/8)^{-1/2} \tag{4}$$

3 Exercise b)

[Describe the new gradient form, the python code and any specific implementation details]

```
#
# Python code to be included
#
```

4 Exercise c)

[Include plots in Figure 1 and describe how the results differ from each other]

Figure 1: Plots showing data and contour lines for the predictive distribution generated by the Laplace approximation (left) and the MAP solution (right).

Avg. Train ll	**Avg. Test ll**
-0.220	-0.293

Table 1: Log-likelihoods for MAP solution.

Avg. Train ll	**Avg. Test ll**
-0.260	-0.318

Table 2: Log-likelihoods for Laplace approximation.

5 Exercise d)

[Include results in Tables 1, 2, 3 and 4 and explain the results obtained and any findings]

6 Exercise e)

[describe your grid search approach, the python code, the grid points chosen, the heat map plot from Figure 2 and the best hyper-parameter values obtained via grid search]

```
#
# Python code to be included
#
```

7 Exercise f)

[Describe the visualisation of the predictions in Figure 3 and the results in Tables 5 and 6. How do they compare to the ones obtained in previous exercises?]

8 Conclusions

1. Draw together the most important results and their consequences.
2. List any reservations or limitations.

		$\hat{y}$ 0	$\hat{y}$ 1
y	0	0.949	0.051
	1	0.059	0.941

Table 3: Conf. matrix for for MAP solution.

		$\hat{y}$ 0	$\hat{y}$ 1
y	0	0.949	0.051
	1	0.059	0.941

Table 4: Conf. matrix for Laplace approximation.

Figure 2: Heat map plot of the the approximation of the model evidence obtained in the grid search.

Figure 3: Visualisation of the contours of the class predictive probabilities for Laplace approximation after hyper-parameter tuning by maximising the model evidence.

Avg. Train ll	**Avg. Test ll**
-	-

Table 5: Average training and test log-likelihoods for Laplace approximation after hyper-parameter tuning by maximising the model evidence.

		$\hat{y}$	
		0	1
y	0	-	-
	1	-	-

Table 6: Confusion matrix for Laplace approximation after hyper-parameter tuning by maximising the model evidence.