Notes on Bayesian Logistic Regression

Max Leung

April 11, 2023

1 Laplace Approximation

$$ln f(\boldsymbol{z}) \approx ln f(\boldsymbol{z}_0) + \overbrace{\nabla ln f(\boldsymbol{z}_0)'}^{\boldsymbol{0}'}(\boldsymbol{z} - \boldsymbol{z}_0) + \frac{1}{2}(\boldsymbol{z} - \boldsymbol{z}_0)'\overbrace{\nabla\nabla ln f(\boldsymbol{z}_0)}^{-\boldsymbol{A}}(\boldsymbol{z} - \boldsymbol{z}_0)$$

$$= ln f(\boldsymbol{z}_0) - \frac{1}{2}(\boldsymbol{z} - \boldsymbol{z}_0)'\boldsymbol{A}(\boldsymbol{z} - \boldsymbol{z}_0)$$

$$f(\boldsymbol{z}) \approx exp(ln f(\boldsymbol{z}_0) - \frac{1}{2}(\boldsymbol{z} - \boldsymbol{z}_0)'\boldsymbol{A}(\boldsymbol{z} - \boldsymbol{z}_0))$$

$$= exp(ln f(\boldsymbol{z}_0)) exp(-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{z}_0)'\boldsymbol{A}(\boldsymbol{z} - \boldsymbol{z}_0))$$

$$= f(\boldsymbol{z}_0) exp(-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{z}_0)'\boldsymbol{A}(\boldsymbol{z} - \boldsymbol{z}_0))$$

If we approximate $f(.)$ by $N(\boldsymbol{z}_0, \boldsymbol{A}^{-1})$, we have

$$\approx \frac{1}{(2\pi)^{M/2}|\boldsymbol{A}|^{-1/2}} \underbrace{exp\{-\frac{1}{2}(\boldsymbol{z}_0 - \boldsymbol{z}_0)'\boldsymbol{A}(\boldsymbol{z}_0 - \boldsymbol{z}_0)\}}_{1} exp(-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{z}_0)'\boldsymbol{A}(\boldsymbol{z} - \boldsymbol{z}_0))$$

$$= N(\boldsymbol{z}|\boldsymbol{z}_0, \boldsymbol{A}^{-1}) = q(\boldsymbol{z})$$

where $\boldsymbol{z}_0 = arg\ max_{\boldsymbol{z}} ln f(\boldsymbol{z})$ and $\boldsymbol{A} = -\nabla\nabla ln f(\boldsymbol{z}_0)$

2 Bayesian Logistic Regression

2.1 Posterior Distribution of Parameters

Assume we have prior density $p(\boldsymbol{w}|\boldsymbol{X}) = p(\boldsymbol{w}) = N(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0)$
Likelihood function is $p(\boldsymbol{t}|\boldsymbol{w}, \boldsymbol{X}) = \prod_{n=1}^{N} p(C_1|\boldsymbol{x}_n; \boldsymbol{w})^{t_n}(1 - p(C_1|\boldsymbol{x}_n; \boldsymbol{w}))^{1-t_n}$ where $p(C_1|\boldsymbol{x}_n; \boldsymbol{w}) = \sigma(\boldsymbol{w}'\boldsymbol{x}_n)$
Posterior density is

$$p(\boldsymbol{w}|\boldsymbol{t}, \boldsymbol{X}) \propto p(\boldsymbol{t}|\boldsymbol{w}, \boldsymbol{X})p(\boldsymbol{w}|\boldsymbol{X})$$

$$= \prod_{n=1}^{N} \sigma(\boldsymbol{w}'\boldsymbol{x}_n)^{t_n}(1 - \sigma(\boldsymbol{w}'\boldsymbol{x}_n))^{1-t_n} N(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0)$$

which is not a well known joint density function

$$ln p(\boldsymbol{w}|\boldsymbol{t}, \boldsymbol{X}) = ln[\prod_{n=1}^{N} \sigma(\boldsymbol{w}'\boldsymbol{x}_n)^{t_n}(1 - \sigma(\boldsymbol{w}'\boldsymbol{x}_n))^{1-t_n} N(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0)]$$

$$= \sum_{n=1}^{N}[t_n ln\sigma(\boldsymbol{w}'\boldsymbol{x}_n) + (1 - t_n)ln(1 - \sigma(\boldsymbol{w}'\boldsymbol{x}_n))] + ln[N(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0)]$$

$$= \sum_{n=1}^{N}[t_n ln\sigma(\boldsymbol{w}'\boldsymbol{x}_n) + (1 - t_n)ln(1 - \sigma(\boldsymbol{w}'\boldsymbol{x}_n))] + ln[\frac{1}{(2\pi)^{D/2}|\boldsymbol{S}_0|^{1/2}}exp\{-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{m}_0)'\boldsymbol{S}_0^{-1}(\boldsymbol{w} - \boldsymbol{m}_0)\}]$$

$$= \sum_{n=1}^{N}[t_n ln\sigma(\boldsymbol{w}'\boldsymbol{x}_n) + (1 - t_n)ln(1 - \sigma(\boldsymbol{w}'\boldsymbol{x}_n))] + ln[\frac{1}{(2\pi)^{D/2}|\boldsymbol{S}_0|^{1/2}}] - \frac{1}{2}(\boldsymbol{w} - \boldsymbol{m}_0)'\boldsymbol{S}_0^{-1}(\boldsymbol{w} - \boldsymbol{m}_0)$$

We can approximate $p(\boldsymbol{w}|\boldsymbol{t}, \boldsymbol{X})$ by Laplace Approximation. As a result, our posterior follows multivariate normal distribution.

$$p(\boldsymbol{w}|\boldsymbol{t}, \boldsymbol{X}) \approx q(\boldsymbol{w}) = N(\boldsymbol{w}|\boldsymbol{w}_{MAP}, \boldsymbol{S}^{-1})$$

where $\boldsymbol{w}_{MAP} = arg\ max_{\boldsymbol{w}} lnp(\boldsymbol{w}|\boldsymbol{t}, \boldsymbol{X})$

$$\frac{\partial lnp(\boldsymbol{w}|\boldsymbol{t}, \boldsymbol{X})}{\partial \boldsymbol{w}}\Big|_{\boldsymbol{w}_{MAP}} = \boldsymbol{0}$$

$$\frac{\partial \sum_{n=1}^{N}[t_n ln\sigma(\boldsymbol{w}'\boldsymbol{x}_n) + (1 - t_n)ln(1 - \sigma(\boldsymbol{w}'\boldsymbol{x}_n))] + ln[\frac{1}{(2\pi)^{D/2}|\boldsymbol{S}_0|^{1/2}}] - \frac{1}{2}(\boldsymbol{w} - \boldsymbol{m}_0)'\boldsymbol{S}_0^{-1}(\boldsymbol{w} - \boldsymbol{m}_0)}{\partial \boldsymbol{w}}\Big|_{\boldsymbol{w}_{MAP}} = \boldsymbol{0}$$

$$\boldsymbol{X}'(\boldsymbol{t} - \boldsymbol{p}) - \boldsymbol{S}_0^{-1}(\boldsymbol{w}_{MAP} - \boldsymbol{m}_0) = \boldsymbol{0}$$

where $\boldsymbol{p} = (\sigma(\boldsymbol{w}'_{MAP}\boldsymbol{x}_1), \cdots, \sigma(\boldsymbol{w}'_{MAP}\boldsymbol{x}_D))'$

There is no closed form solution for $\boldsymbol{w}_{MAP}$

$$\boldsymbol{S} = -\nabla\nabla lnp(\boldsymbol{w}_{MAP}|\boldsymbol{t})$$
$$= -\nabla\nabla\{\sum_{n=1}^{N}[t_n ln\sigma(\boldsymbol{w}'\boldsymbol{x}_n) + (1 - t_n)ln(1 - \sigma(\boldsymbol{w}'\boldsymbol{x}_n))] + ln[\frac{1}{(2\pi)^{D/2}|\boldsymbol{S}_0|^{1/2}}] - \frac{1}{2}(\boldsymbol{w} - \boldsymbol{m}_0)'\boldsymbol{S}_0^{-1}(\boldsymbol{w} - \boldsymbol{m}_0)\}$$
$$= -(-\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} - \boldsymbol{S}_0^{-1})$$
$$= \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} + \boldsymbol{S}_0^{-1}$$

where $(\boldsymbol{W})_{ii} = \sigma(\boldsymbol{w}'_{MAP}\boldsymbol{x}_i)(1 - \sigma(\boldsymbol{w}'_{MAP}\boldsymbol{x}_i))$ and $(\boldsymbol{W})_{ij} = 0$ for $i \neq j$

2.2 Special Case

if $\boldsymbol{m}_0$ is chosen to be $\boldsymbol{w}_{MAP}$ then

$$\boldsymbol{X}'(\boldsymbol{t} - \boldsymbol{p}) - \boldsymbol{S}_0^{-1}(\boldsymbol{w}_{MAP} - \boldsymbol{w}_{MAP}) = \boldsymbol{0}$$
$$\boldsymbol{X}'(\boldsymbol{t} - \boldsymbol{p}) = \boldsymbol{0} \qquad\qquad \text{same as FOC of MLE}$$

Thus, $\boldsymbol{w}_{MAP} = \boldsymbol{w}_{MLE}$ in such case, which can be found by Iterated Reweighted Least Squares (IRLS) algorithm.

Additionally, if $\boldsymbol{S}_0$ is chosen to be close to $\boldsymbol{O}^{-1}$. We have

$$\boldsymbol{S} \approx \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} + (\boldsymbol{O}^{-1})^{-1}$$
$$= \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$$

Thus,

$$\boldsymbol{S}^{-1} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1} = -(-\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1} = -(\nabla\nabla p(\boldsymbol{t}|\boldsymbol{w}_{MLE}, \boldsymbol{X}))^{-1} = (\underbrace{-\nabla\nabla p(\boldsymbol{t}|\boldsymbol{w}_{MLE}, \boldsymbol{X})}_{\boldsymbol{I}(\boldsymbol{w}_{MLE})})^{-1}$$

$\boldsymbol{I}(\boldsymbol{w}_{MLE})^{-1}$ is the asymptotic variance of $\sqrt{D}(\boldsymbol{w}_{MLE} - \boldsymbol{w}_{TRUE})$

Thus, we have $p(\boldsymbol{w}|\boldsymbol{t}) \approx N(\boldsymbol{w}|\boldsymbol{w}_{MLE}, \boldsymbol{I}(\boldsymbol{w}_{MLE})^{-1})$

3 Reference

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. New York :Springer.