Notes on Bayesian Logistic Regression

Max Leung

April 19, 2023

1 Laplace Approximation

$$\begin{aligned}
lnf(\boldsymbol{z}) &\approx lnf(\boldsymbol{z}_0) + \overbrace{\nabla lnf(\boldsymbol{z}_0)'}^{\boldsymbol{0}'}(\boldsymbol{z}-\boldsymbol{z}_0) + \frac{1}{2}(\boldsymbol{z}-\boldsymbol{z}_0)'\overbrace{\nabla\nabla lnf(\boldsymbol{z}_0)}^{-\boldsymbol{A}}(\boldsymbol{z}-\boldsymbol{z}_0) \\
&= lnf(\boldsymbol{z}_0) - \frac{1}{2}(\boldsymbol{z}-\boldsymbol{z}_0)'\boldsymbol{A}(\boldsymbol{z}-\boldsymbol{z}_0)
\end{aligned}$$

$$\begin{aligned}
f(\boldsymbol{z}) &\approx exp(lnf(\boldsymbol{z}_0) - \frac{1}{2}(\boldsymbol{z}-\boldsymbol{z}_0)'\boldsymbol{A}(\boldsymbol{z}-\boldsymbol{z}_0)) \\
&= exp(lnf(\boldsymbol{z}_0))exp(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{z}_0)'\boldsymbol{A}(\boldsymbol{z}-\boldsymbol{z}_0)) \\
&= f(\boldsymbol{z}_0)exp(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{z}_0)'\boldsymbol{A}(\boldsymbol{z}-\boldsymbol{z}_0))
\end{aligned}$$

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

$$\begin{aligned}
&\approx \frac{1}{(2\pi)^{M/2}|\boldsymbol{A}|^{-1/2}}\underbrace{exp\{-\frac{1}{2}(\boldsymbol{z}_0-\boldsymbol{z}_0)'\boldsymbol{A}(\boldsymbol{z}_0-\boldsymbol{z}_0)\}}_{1}exp(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{z}_0)'\boldsymbol{A}(\boldsymbol{z}-\boldsymbol{z}_0)) \\
&= N(\boldsymbol{z}|\boldsymbol{z}_0, \boldsymbol{A}^{-1}) = q(\boldsymbol{z})
\end{aligned}$$

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

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

$$\begin{aligned}
p(\boldsymbol{w}|\boldsymbol{t},\boldsymbol{X}) &\propto p(\boldsymbol{t}|\boldsymbol{w},\boldsymbol{X})p(\boldsymbol{w}|\boldsymbol{X}) \\
&= \prod_{n=1}^{N}\sigma(\boldsymbol{w}'\boldsymbol{x}_n)^{t_n}(1-\sigma(\boldsymbol{w}'\boldsymbol{x}_n))^{1-t_n}N(\boldsymbol{w}|\boldsymbol{m}_0,\boldsymbol{S}_0)
\end{aligned}$$

which is not a well known joint density function

$$\begin{aligned}
lnp(\boldsymbol{w}|\boldsymbol{t},\boldsymbol{X}) &= ln[\prod_{n=1}^{N}\sigma(\boldsymbol{w}'\boldsymbol{x}_n)^{t_n}(1-\sigma(\boldsymbol{w}'\boldsymbol{x}_n))^{1-t_n}N(\boldsymbol{w}|\boldsymbol{m}_0,\boldsymbol{S}_0)] \\
&= \sum_{n=1}^{N}[t_n ln\sigma(\boldsymbol{w}'\boldsymbol{x}_n) + (1-t_n)ln(1-\sigma(\boldsymbol{w}'\boldsymbol{x}_n))] + ln[N(\boldsymbol{w}|\boldsymbol{m}_0,\boldsymbol{S}_0)] \\
&= \sum_{n=1}^{N}[t_n ln\sigma(\boldsymbol{w}'\boldsymbol{x}_n) + (1-t_n)ln(1-\sigma(\boldsymbol{w}'\boldsymbol{x}_n))] + ln[\frac{1}{(2\pi)^{D/2}|\boldsymbol{S}_0|^{1/2}}exp\{-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_0)'\boldsymbol{S}_0^{-1}(\boldsymbol{w}-\boldsymbol{m}_0)\}] \\
&= \sum_{n=1}^{N}[t_n ln\sigma(\boldsymbol{w}'\boldsymbol{x}_n) + (1-t_n)ln(1-\sigma(\boldsymbol{w}'\boldsymbol{x}_n))] + ln[\frac{1}{(2\pi)^{D/2}|\boldsymbol{S}_0|^{1/2}}] - \frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_0)'\boldsymbol{S}_0^{-1}(\boldsymbol{w}-\boldsymbol{m}_0)
\end{aligned}$$

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

$$\begin{aligned}
\frac{\partial lnp(\boldsymbol{w}|\boldsymbol{t},\boldsymbol{X})}{\partial \boldsymbol{w}}|_{\boldsymbol{w}_{MAP}} &= \boldsymbol{0} \\
\frac{\partial \sum_{n=1}^{N}[t_n ln\sigma(\boldsymbol{w}'\boldsymbol{x}_n) + (1-t_n)ln(1-\sigma(\boldsymbol{w}'\boldsymbol{x}_n))] + ln[\frac{1}{(2\pi)^{D/2}|\boldsymbol{S}_0|^{1/2}}] - \frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_0)'\boldsymbol{S}_0^{-1}(\boldsymbol{w}-\boldsymbol{m}_0)}{\partial \boldsymbol{w}}|_{\boldsymbol{w}_{MAP}} &= \boldsymbol{0} \\
\boldsymbol{X}'(\boldsymbol{t}-\boldsymbol{p}) - \boldsymbol{S}_0^{-1}(\boldsymbol{w}_{MAP}-\boldsymbol{m}_0) &= \boldsymbol{0}
\end{aligned}$$

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 $\boldsymbol{w}_{MAP}$

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

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

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

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

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

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

2.3 Predictive Distribution

Let $\widetilde{T}$ be the predicted target / dependent variable. Predictive distribution is

$$\begin{aligned}
p(\widetilde{T}=1|\boldsymbol{t},\boldsymbol{X}) &= \int p(\widetilde{T}=1,\boldsymbol{w}|\boldsymbol{t},\boldsymbol{X})\partial\boldsymbol{w} \\
&= \int p(\widetilde{T}=1|\boldsymbol{w},\boldsymbol{t},\boldsymbol{X})p(\boldsymbol{w}|\boldsymbol{t},\boldsymbol{X})\partial\boldsymbol{w} \\
&\approx \int \sigma(\boldsymbol{w}'\boldsymbol{x})q(\boldsymbol{w})\partial\boldsymbol{w}
\end{aligned}$$

Note that $\int \delta(a - \boldsymbol{w}'\boldsymbol{x})\sigma(a)da = \delta(\boldsymbol{w}'\boldsymbol{x} - \boldsymbol{w}'\boldsymbol{x})\sigma(\boldsymbol{w}'\boldsymbol{x}) = 1 \cdot \sigma(\boldsymbol{w}'\boldsymbol{x})$ as $\delta(0) = 1$ and $\delta(a) = 0$ for $\forall a \neq 0$

$$\begin{aligned}
&= \int\int \delta(a - \boldsymbol{w}'\boldsymbol{x})\sigma(a)da\ q(\boldsymbol{w})\partial\boldsymbol{w} \\
&= \int\int \delta(a - \boldsymbol{w}'\boldsymbol{x})\sigma(a)q(\boldsymbol{w})\partial\boldsymbol{w}da \\
&= \int \sigma(a) \underbrace{\int \delta(a - \boldsymbol{w}'\boldsymbol{x})q(\boldsymbol{w})\partial\boldsymbol{w}}_{p(a)} da
\end{aligned}$$

$p(a)$'s moments can be evaluated as

$$\begin{aligned}
\mathbb{E}_p(a) &= \int p(a)a\ da \\
&= \int\int \delta(a - \boldsymbol{w}'\boldsymbol{x})q(\boldsymbol{w})\partial\boldsymbol{w}a\ da \\
&= \int\int \delta(a - \boldsymbol{w}'\boldsymbol{x})q(\boldsymbol{w})a\ da\partial\boldsymbol{w} \\
&= \int\int \delta(a - \boldsymbol{w}'\boldsymbol{x})a\ da\ q(\boldsymbol{w})\partial\boldsymbol{w} \\
&= \int \delta(\boldsymbol{w}'\boldsymbol{x} - \boldsymbol{w}'\boldsymbol{x})\boldsymbol{w}'\boldsymbol{x}q(\boldsymbol{w})\partial\boldsymbol{w} \\
&= \int \boldsymbol{w}'\boldsymbol{x}q(\boldsymbol{w})\partial\boldsymbol{w} \\
&= \mathbb{E}_q(\boldsymbol{w}'\boldsymbol{x}) \\
&= \mathbb{E}_q(\boldsymbol{w})'\boldsymbol{x}
\end{aligned}$$

$$\begin{aligned}
Var_p(a) &= \mathbb{E}_p(a^2) - \mathbb{E}_p(a)^2 \\
&= \mathbb{E}_p(a^2 - \mathbb{E}_p(a)^2) \\
&= \int p(a)(a^2 - \mathbb{E}_p(a)^2)da \\
&= \int\int \delta(a - \boldsymbol{w}'\boldsymbol{x})q(\boldsymbol{w})\partial\boldsymbol{w}(a^2 - \mathbb{E}_p(a)^2)da \\
&= \int\int \delta(a - \boldsymbol{w}'\boldsymbol{x})q(\boldsymbol{w})(a^2 - \mathbb{E}_p(a)^2)da\partial\boldsymbol{w} \\
&= \int q(\boldsymbol{w}) \int \delta(a - \boldsymbol{w}'\boldsymbol{x})(a^2 - \mathbb{E}_p(a)^2)da\partial\boldsymbol{w} \\
&= \int q(\boldsymbol{w})\delta(\boldsymbol{w}'\boldsymbol{x} - \boldsymbol{w}'\boldsymbol{x})((\boldsymbol{w}'\boldsymbol{x})^2 - \mathbb{E}_p(\boldsymbol{w}'\boldsymbol{x})^2)\partial\boldsymbol{w} \\
&= \int q(\boldsymbol{w})((\boldsymbol{w}'\boldsymbol{x})^2 - \mathbb{E}_p(\boldsymbol{w}'\boldsymbol{x})^2)\partial\boldsymbol{w} \\
&\approx \int q(\boldsymbol{w})((\boldsymbol{w}'\boldsymbol{x})^2 - \mathbb{E}_q(\boldsymbol{w}'\boldsymbol{x})^2)\partial\boldsymbol{w} \\
&= \mathbb{E}_q((\boldsymbol{w}'\boldsymbol{x})^2 - \mathbb{E}_q(\boldsymbol{w}'\boldsymbol{x})^2) \\
&= Var_q(\boldsymbol{x}'\boldsymbol{w}) \\
&= \boldsymbol{x}'Var_q(\boldsymbol{w})\boldsymbol{x}
\end{aligned}$$

3 Reference

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