Notes on Bayesian Logistic Regression

Max Leung

May 7, 2023

1 Laplace Approximation

$$lnf(z) \approx lnf(z_0) + \overbrace{\nabla lnf(z_0)'}^{0'}(z-z_0) + \frac{1}{2}(z-z_0)' \overbrace{\nabla \nabla lnf(z_0)}^{-A}(z-z_0)$$

$$= lnf(z_0) - \frac{1}{2}(z-z_0)'A(z-z_0)$$

$$f(z) \approx exp(lnf(z_0) - \frac{1}{2}(z - z_0)'A(z - z_0))$$

$$= exp(lnf(z_0))exp(-\frac{1}{2}(z - z_0)'A(z - z_0))$$

$$= f(z_0)exp(-\frac{1}{2}(z - z_0)'A(z - z_0))$$

If we approximate f(.) by $N(z_0, A^{-1})$, we have

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

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

where $z_0 = arg \ max_z lnf(z)$ and $A = -\nabla \nabla lnf(z_0)$

2 Bayesian Logistic Regression (Laplace Approximation Approach)

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

$$lnp(\boldsymbol{w}|\boldsymbol{t},\boldsymbol{X}) = ln[\prod_{n=1}^{N} \sigma(\boldsymbol{w}'\boldsymbol{x}_{n})^{t_{n}} (1 - \sigma(\boldsymbol{w}'\boldsymbol{x}_{n}))^{1-t_{n}} N(\boldsymbol{w}|\boldsymbol{m}_{0},\boldsymbol{S}_{0})]$$

$$= \sum_{n=1}^{N} [t_{n}ln\sigma(\boldsymbol{w}'\boldsymbol{x}_{n}) + (1 - t_{n})ln(1 - \sigma(\boldsymbol{w}'\boldsymbol{x}_{n}))] + ln[N(\boldsymbol{w}|\boldsymbol{m}_{0},\boldsymbol{S}_{0})]$$

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

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

We can approximate p(w|t, 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 $w_{MAP} = arg \ max_{w} lnp(w|t, X)$

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

$$\frac{\partial \sum_{n=1}^{N} [t_n ln\sigma(\boldsymbol{w}'\boldsymbol{x}_n) + (1-t_n)ln(1-\sigma(\boldsymbol{w}'\boldsymbol{x}_n))] + ln[\frac{1}{(2\pi)^{D/2}|S_0|^{1/2}}] - \frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_0)'\boldsymbol{S}_0^{-1}(\boldsymbol{w}-\boldsymbol{m}_0)}{\partial \boldsymbol{w}}|_{\boldsymbol{w}_{MAP}} = \boldsymbol{0}$$

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

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}

$$S = -\nabla \nabla lnp(\mathbf{w}_{MAP}|\mathbf{t})$$

$$= -\nabla \nabla \{\sum_{n=1}^{N} [t_n ln\sigma(\mathbf{w}'\mathbf{x}_n) + (1 - t_n)ln(1 - \sigma(\mathbf{w}'\mathbf{x}_n))] + ln[\frac{1}{(2\pi)^{D/2}|S_0|^{1/2}}] - \frac{1}{2}(\mathbf{w} - \mathbf{m}_0)'S_0^{-1}(\mathbf{w} - \mathbf{m}_0)\}$$

$$= -(-\mathbf{X}'\mathbf{W}\mathbf{X} - \mathbf{S}_0^{-1})$$

$$= \mathbf{X}'\mathbf{W}\mathbf{X} + \mathbf{S}_0^{-1}$$

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

2.2 Special Case

if m_0 is chosen to be w_{MAP} then

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

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 S_0 is chosen to be close to O^{-1} . We have

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

Thus,

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

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

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

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

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

$$= \int \int \delta(a - \mathbf{w}' \mathbf{x}) \sigma(a) da \ q(\mathbf{w}) \partial \mathbf{w}$$

$$= \int \int \delta(a - \mathbf{w}' \mathbf{x}) \sigma(a) q(\mathbf{w}) \partial \mathbf{w} da$$

$$= \int \sigma(a) \underbrace{\int \delta(a - \mathbf{w}' \mathbf{x}) q(\mathbf{w}) \partial \mathbf{w}}_{p(a)} da$$

p(a)'s moments can be evaluated as

$$\mathbb{E}_{p}(a) = \int p(a)a \ da$$

$$= \int \int \delta(a - \mathbf{w}'\mathbf{x})q(\mathbf{w})\partial \mathbf{w}a \ da$$

$$= \int \int \delta(a - \mathbf{w}'\mathbf{x})q(\mathbf{w})a \ da\partial \mathbf{w}$$

$$= \int \int \delta(a - \mathbf{w}'\mathbf{x})a \ da \ q(\mathbf{w})\partial \mathbf{w}$$

$$= \int \delta(\mathbf{w}'\mathbf{x} - \mathbf{w}'\mathbf{x})\mathbf{w}'\mathbf{x}q(\mathbf{w})\partial \mathbf{w}$$

$$= \int \mathbf{w}'\mathbf{x}q(\mathbf{w})\partial \mathbf{w}$$

$$= \mathbb{E}_{q}(\mathbf{w}'\mathbf{x})$$

$$= \mathbb{E}_{q}(\mathbf{w})'\mathbf{x}$$

$$Var_{p}(a) = \mathbb{E}_{p}(a^{2}) - \mathbb{E}_{p}(a)^{2}$$

$$= \mathbb{E}_{p}(a^{2} - \mathbb{E}_{p}(a)^{2})$$

$$= \int p(a)(a^{2} - \mathbb{E}_{p}(a)^{2})da$$

$$= \int \int \delta(a - \mathbf{w}'\mathbf{x})q(\mathbf{w})\partial\mathbf{w}(a^{2} - \mathbb{E}_{p}(a)^{2})da$$

$$= \int \int \delta(a - \mathbf{w}'\mathbf{x})q(\mathbf{w})(a^{2} - \mathbb{E}_{p}(a)^{2})da\partial\mathbf{w}$$

$$= \int q(\mathbf{w}) \int \delta(a - \mathbf{w}'\mathbf{x})(a^{2} - \mathbb{E}_{p}(a)^{2})da\partial\mathbf{w}$$

$$= \int q(\mathbf{w})\delta(\mathbf{w}'\mathbf{x} - \mathbf{w}'\mathbf{x})((\mathbf{w}'\mathbf{x})^{2} - \mathbb{E}_{p}(\mathbf{w}'\mathbf{x})^{2})\partial\mathbf{w}$$

$$= \int q(\mathbf{w})((\mathbf{w}'\mathbf{x})^{2} - \mathbb{E}_{p}(\mathbf{w}'\mathbf{x})^{2})\partial\mathbf{w}$$

$$\approx \int q(\mathbf{w})((\mathbf{w}'\mathbf{x})^{2} - \mathbb{E}_{q}(\mathbf{w}'\mathbf{x})^{2})\partial\mathbf{w}$$

$$= \mathbb{E}_{q}((\mathbf{w}'\mathbf{x})^{2} - \mathbb{E}_{q}(\mathbf{w}'\mathbf{x})^{2})$$

$$= Var_{q}(\mathbf{x}'\mathbf{w})$$

$$= \mathbf{x}'Var_{q}(\mathbf{w})\mathbf{x}$$

$$ln[L(\boldsymbol{\theta})] = ln[p(\boldsymbol{X}|\boldsymbol{\theta})] = ln[\sum_{\boldsymbol{z}} p(\boldsymbol{X}, \boldsymbol{z}|\boldsymbol{\theta})] = ln[\sum_{\boldsymbol{z}} p(\boldsymbol{X}|\boldsymbol{z}, \boldsymbol{\theta})p(\boldsymbol{z}|\boldsymbol{\theta})]$$

$$\begin{split} ln[L(\boldsymbol{\theta})] - ln[L(\boldsymbol{\theta}_n)] &= ln[\sum_{\boldsymbol{z}} p(\boldsymbol{X}|\boldsymbol{z}, \boldsymbol{\theta}) p(\boldsymbol{z}|\boldsymbol{\theta})] - ln[p(\boldsymbol{X}|\boldsymbol{\theta}_n)] \\ &= ln[\sum_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}_n) \frac{p(\boldsymbol{X}|\boldsymbol{z}, \boldsymbol{\theta}) p(\boldsymbol{z}|\boldsymbol{\theta})}{p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}_n)}] - ln[p(\boldsymbol{X}|\boldsymbol{\theta}_n)] \end{split}$$

Jensen's Inequality says $f(\sum_i \lambda_i x_i) \geq \sum_i \lambda_i f(x_i)$ if f is concave, $\sum_i \lambda_i = 1$ and $\lambda_i \geq 0$ for $\forall i$. As ln(.) is concave and $p(\boldsymbol{z}|\boldsymbol{X},\boldsymbol{\theta}_n)$ is density function, we have

$$\geq \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_{n}) ln[\frac{p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})}{p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_{n})}] - ln[p(\mathbf{X}|\boldsymbol{\theta}_{n})]$$

$$= \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_{n}) ln[\frac{p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})}{p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_{n})}] - ln[p(\mathbf{X}|\boldsymbol{\theta}_{n})] \cdot 1$$

$$= \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_{n}) ln[\frac{p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})}{p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_{n})}] - ln[p(\mathbf{X}|\boldsymbol{\theta}_{n})] \cdot \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_{n})$$

$$= \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_{n}) ln[\frac{p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})}{p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_{n})}] - \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_{n}) ln[p(\mathbf{X}|\boldsymbol{\theta}_{n})]$$

$$= \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_{n}) \{ ln[\frac{p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})}{p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_{n})}] - ln[p(\mathbf{X}|\boldsymbol{\theta}_{n})] \}$$

$$= \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_{n}) ln[\frac{p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})}{p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_{n})}]$$

Thus,

$$ln[L(\boldsymbol{\theta})] \geq ln[L(\boldsymbol{\theta}_n)] + \sum_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}_n) ln[\frac{p(\boldsymbol{X}|\boldsymbol{z}, \boldsymbol{\theta})p(\boldsymbol{z}|\boldsymbol{\theta})}{p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}_n)p(\boldsymbol{X}|\boldsymbol{\theta}_n)}] =: l(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$$

It can be shown that $ln[L(\boldsymbol{\theta})] = l(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$ when $\boldsymbol{\theta} = \boldsymbol{\theta}_n$

$$l(\boldsymbol{\theta}_n|\boldsymbol{\theta}_n) = ln[L(\boldsymbol{\theta}_n)] + \sum_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}_n) ln[\frac{p(\boldsymbol{X}|\boldsymbol{z}, \boldsymbol{\theta}_n) p(\boldsymbol{z}|\boldsymbol{\theta}_n)}{p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}_n) p(\boldsymbol{X}|\boldsymbol{\theta}_n)}]$$

$$= ln[L(\boldsymbol{\theta}_n)] + \sum_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}_n) ln[1]$$

$$= ln[L(\boldsymbol{\theta}_n)]$$

Thus, we know that log likelihood function wraps $l(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$ i.e., $ln[L(\boldsymbol{\theta})] \geq l(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$ and $l(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$ touches log likelihood function at $\boldsymbol{\theta}_n$ i.e., $l(\boldsymbol{\theta}_n|\boldsymbol{\theta}_n) = ln[L(\boldsymbol{\theta}_n)]$.

As $ln[L(\boldsymbol{\theta})] \geq l(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$, an increase in $l(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$ implies an increase in $ln[L(\boldsymbol{\theta})]$. $l(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$ stops increasing at its maximum. We pick $\boldsymbol{\theta}_{n+1} = arg \; max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$. This guarantees that $ln[L(\boldsymbol{\theta})]$ increases at every step. More formally,

$$\begin{split} \boldsymbol{\theta}_{n+1} &= arg \; max_{\boldsymbol{\theta}} \{ l(\boldsymbol{\theta}|\boldsymbol{\theta}_n) \} \\ &= arg \; max_{\boldsymbol{\theta}} \{ ln[L(\boldsymbol{\theta}_n)] + \sum_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}_n) ln[\frac{p(\boldsymbol{X}|\boldsymbol{z}, \boldsymbol{\theta})p(\boldsymbol{z}|\boldsymbol{\theta})}{p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}_n)p(\boldsymbol{X}|\boldsymbol{\theta}_n)}] \} \\ &= arg \; max_{\boldsymbol{\theta}} \{ \sum_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}_n) ln[\frac{p(\boldsymbol{X}|\boldsymbol{z}, \boldsymbol{\theta})p(\boldsymbol{z}|\boldsymbol{\theta})}{p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}_n)p(\boldsymbol{X}|\boldsymbol{\theta}_n)}] \} \end{split}$$

Thus, $\sum_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}_n) ln[\frac{p(\boldsymbol{X}|\boldsymbol{z}, \boldsymbol{\theta}_{n+1})p(\boldsymbol{z}|\boldsymbol{\theta}_{n+1})}{p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}_n)p(\boldsymbol{X}|\boldsymbol{\theta}_n)}] \ge \sum_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}_n) ln[\frac{p(\boldsymbol{X}|\boldsymbol{z}, \boldsymbol{\theta})p(\boldsymbol{z}|\boldsymbol{\theta}_n)}{p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}_n)p(\boldsymbol{X}|\boldsymbol{\theta}_n)}]$ for $\forall \boldsymbol{\theta}$ including $\boldsymbol{\theta}_n$

$$\sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n) ln[\frac{p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta}_{n+1})p(\mathbf{z}|\boldsymbol{\theta}_{n+1})}{p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n)p(\mathbf{X}|\boldsymbol{\theta}_n)}] \geq \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n) ln[\frac{p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta}_n)p(\mathbf{z}|\boldsymbol{\theta}_n)}{p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n)p(\mathbf{X}|\boldsymbol{\theta}_n)}]$$

Since $ln[L(\boldsymbol{\theta}_{n+1})] - ln[L(\boldsymbol{\theta}_n)] \ge \sum_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}_n) ln[\frac{p(\boldsymbol{X}|\boldsymbol{z}, \boldsymbol{\theta}_{n+1})p(\boldsymbol{z}|\boldsymbol{\theta}_{n+1})}{p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}_n)p(\boldsymbol{X}|\boldsymbol{\theta}_n)}]$ and $\sum_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}_n) ln[\frac{p(\boldsymbol{X}|\boldsymbol{z}, \boldsymbol{\theta}_n)p(\boldsymbol{z}|\boldsymbol{\theta}_n)}{p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}_n)p(\boldsymbol{X}|\boldsymbol{\theta}_n)}] = 0$, we have $ln[L(\boldsymbol{\theta}_{n+1})] - ln[L(\boldsymbol{\theta}_n)] \ge 0$

E step and M step can be shown as

$$\begin{split} &\boldsymbol{\theta}_{n+1} = arg \; max_{\boldsymbol{\theta}} \{ l(\boldsymbol{\theta}|\boldsymbol{\theta}_n) \} \\ &= arg \; max_{\boldsymbol{\theta}} \{ ln[L(\boldsymbol{\theta}_n)] + \sum_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}_n) ln[\frac{p(\boldsymbol{X}|\boldsymbol{z}, \boldsymbol{\theta})p(\boldsymbol{z}|\boldsymbol{\theta})}{p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}_n)p(\boldsymbol{X}|\boldsymbol{\theta}_n)}] \} \\ &= arg \; max_{\boldsymbol{\theta}} \{ ln[L(\boldsymbol{\theta}_n)] + \sum_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}_n) ln[p(\boldsymbol{X}|\boldsymbol{z}, \boldsymbol{\theta})p(\boldsymbol{z}|\boldsymbol{\theta})] - \sum_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}_n) ln[p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}_n)p(\boldsymbol{X}|\boldsymbol{\theta}_n)] \} \\ &= arg \; max_{\boldsymbol{\theta}} \{ \sum_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}_n) ln[p(\boldsymbol{X}|\boldsymbol{z}, \boldsymbol{\theta})p(\boldsymbol{z}|\boldsymbol{\theta})] \} \\ &= arg \; max_{\boldsymbol{\theta}} \{ \sum_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}_n) ln[p(\boldsymbol{X}, \boldsymbol{z}|\boldsymbol{\theta})] \} \\ &= arg \; max_{\boldsymbol{\theta}} \{ \mathbb{E}_{\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}_n} [ln[p(\boldsymbol{X}, \boldsymbol{z}|\boldsymbol{\theta})]] \} \end{split}$$

 $\mathbb{E}_{\boldsymbol{z}|\boldsymbol{X},\boldsymbol{\theta}_n}$ is the E step. $arg\ max_{\boldsymbol{\theta}}$ is the M step.

4 Bayesian Logistic Regression (Approximate EM Approach)

4.1 Uniform Prior

Posterior density function is

$$p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}) \propto L(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X})U(\boldsymbol{\beta}|\boldsymbol{X})$$

= $L(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X}) \cdot constant$
 $\propto L(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X})$

Uniform prior is not a function of beta.

Thus, posterior mode is the same as maximum likelihood estimate. Log posterior density is

$$ln[p(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X})] \propto ln[L(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{X})]$$

Log likelihood function is approximated as negative weighted least squares function locally (see IRLS).

$$pprox - \sum_{i=1}^{N} w_i (z_i - \boldsymbol{x}_i' \boldsymbol{eta})^2$$

$$\propto -\frac{1}{2} \sum_{i=1}^{N} w_i (z_i - \boldsymbol{x}_i' \boldsymbol{eta})^2$$

where

$$\begin{aligned} z_i &= \pmb{x}_i' \widehat{\pmb{\beta}} + \frac{y_i - p_i}{w_i} \\ w_i &= p_i (1 - p_i) \\ p_i &= logistic(\pmb{x}_i' \widehat{\pmb{\beta}}) \end{aligned}$$
 $\widehat{\pmb{\beta}}$ is from last step

4.2 Independent Normal / Gaussian Prior

$$\begin{split} & ln[p(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{X})] \propto ln[L(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{X})N(\boldsymbol{\beta}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{X})] \\ & = ln[L(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{X})] + ln[N(\boldsymbol{\beta}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{X})] \\ & \approx -\frac{1}{2}\sum_{i=1}^{N}w_{i}(z_{i}-\boldsymbol{x}_{i}'\boldsymbol{\beta})^{2} - \frac{1}{2}\sum_{j=1}^{J}(\frac{(\beta_{j}-\mu_{j})^{2}}{\sigma_{j}^{2}} + ln(\sigma_{j}^{2})) + constant \\ & = -\frac{1}{2}\{\sum_{i=1}^{N}\frac{1}{w_{i}^{-1}}(z_{i}-\boldsymbol{x}_{i}'\boldsymbol{\beta})^{2} + \sum_{j=1}^{J}\frac{1}{\sigma_{j}^{2}}(\mu_{j}-1\cdot\boldsymbol{\beta}_{j})^{2} + \sum_{j=1}^{J}ln(\sigma_{j}^{2})\} + constant \\ & \approx -\frac{1}{2}\{\sum_{i=1}^{N}\frac{1}{w_{i}^{-1}}(z_{i}-\boldsymbol{x}_{i}'\boldsymbol{\beta})^{2} + \sum_{j=1}^{J}\frac{1}{\sigma_{j}^{2}}(\mu_{j}-\boldsymbol{e}_{j}'\boldsymbol{\beta})^{2}\} \qquad \qquad \boldsymbol{e}_{j}' \text{ is the } j \text{ row of identity matrix} \\ & = -\frac{1}{2}\{(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{W}(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{\mu}-\boldsymbol{I}_{J}\boldsymbol{\beta})'\boldsymbol{\Sigma}(\boldsymbol{\mu}-\boldsymbol{I}_{J}\boldsymbol{\beta})\} \\ & = -\frac{1}{2}(\boldsymbol{z}_{*}-\boldsymbol{X}_{*}\boldsymbol{\beta})'\boldsymbol{W}_{*}(\boldsymbol{z}_{*}-\boldsymbol{X}_{*}\boldsymbol{\beta}) \end{split}$$

where

$$egin{aligned} oldsymbol{z}_* &= egin{pmatrix} oldsymbol{z} \ oldsymbol{X}_* &= egin{pmatrix} oldsymbol{X} \ oldsymbol{I}_J \end{pmatrix} \ oldsymbol{W}_* &= egin{pmatrix} oldsymbol{W} & oldsymbol{O} \ oldsymbol{O} & oldsymbol{\Sigma} \end{pmatrix} \end{aligned}$$

Thus, original IRLS algorithm can be applied with augmented data (μ and Σ from prior normal distribution) to estimate β

$$\begin{split} ln[N(\pmb{\beta}|\pmb{\mu},\pmb{\Sigma},\pmb{X})] &= ln[\prod_{j=1}^{J} \frac{1}{\sqrt{2\pi\sigma_{j}^{2}}} exp(-\frac{1}{2} \frac{(\beta_{j} - \mu_{j})^{2}}{\sigma_{j}^{2}})] \\ &= \sum_{j=1}^{J} ln[\frac{1}{\sqrt{2\pi\sigma_{j}^{2}}} exp(-\frac{1}{2} \frac{(\beta_{j} - \mu_{j})^{2}}{\sigma_{j}^{2}})] \\ &= \sum_{j=1}^{J} \{ ln[1] - ln[(2\pi\sigma_{j}^{2})^{1/2}] + ln[exp(-\frac{1}{2} \frac{(\beta_{j} - \mu_{j})^{2}}{\sigma_{j}^{2}})] \} \\ &= \sum_{j=1}^{J} \{ -\frac{1}{2} ln[2\pi\sigma_{j}^{2}] - \frac{1}{2} \frac{(\beta_{j} - \mu_{j})^{2}}{\sigma_{j}^{2}} \} \\ &= \sum_{j=1}^{J} \{ -\frac{1}{2} ln[2\pi] \} + \sum_{j=1}^{J} \{ -\frac{1}{2} ln[\sigma_{j}^{2}] - \frac{1}{2} \frac{(\beta_{j} - \mu_{j})^{2}}{\sigma_{j}^{2}} \} \\ &= constant + -\frac{1}{2} \sum_{j=1}^{J} \{ ln[\sigma_{j}^{2}] + \frac{(\beta_{j} - \mu_{j})^{2}}{\sigma_{j}^{2}} \} \end{split}$$

Independent Student-t Prior

$$ln[p(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X})] \propto ln[L(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{X})t(\boldsymbol{\beta}|\boldsymbol{X})]$$

Multivariate Student-t density is a product of Multivariate Normal's and Inverse Chi Squares' density after "integrate out"

$$\begin{split} &= ln[L(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X}) \int N(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{X}) Inv - \chi^{2}(\boldsymbol{\Sigma}|\boldsymbol{X}) \partial \boldsymbol{\Sigma}] \\ &\approx ln[L(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X})] + ln[N(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{X})] + ln[Inv - \chi^{2}(\boldsymbol{\Sigma}|\boldsymbol{X})] \\ &\approx -\frac{1}{2} \{ \sum_{i=1}^{N} \frac{1}{w_{i}^{-1}} (z_{i} - \boldsymbol{x}_{i}'\boldsymbol{\beta})^{2} + \sum_{j=1}^{J} \frac{1}{\sigma_{j}^{2}} (\mu_{j} - 1 \cdot \beta_{j})^{2} + \sum_{j=1}^{J} ln(\sigma_{j}^{2}) \} + constant - p(\sigma_{j}|v_{j}, s_{j}) \\ &= -\frac{1}{2} (\boldsymbol{z}_{*} - \boldsymbol{X}_{*}\boldsymbol{\beta})' \boldsymbol{W}_{*} (\boldsymbol{z}_{*} - \boldsymbol{X}_{*}\boldsymbol{\beta}) - \frac{1}{2} \sum_{i=1}^{J} ln(\sigma_{j}^{2}) + constant - p(\sigma_{j}|v_{j}, s_{j}) \end{split}$$

Gelman et al. (2008, 2013) propose an approximate EM algorithm, in which IRLS and EM algorithm are combined and altered, to estimate β

- 1. Use last $\boldsymbol{\beta}$ or initial value to get \boldsymbol{z} and \boldsymbol{W} . Use last σ_j^2 or initial value s_j^2 to get $\boldsymbol{\Sigma}$ 2. Now we have \boldsymbol{z}_* , \boldsymbol{W}_* and \boldsymbol{X}_* . We minimize $(\boldsymbol{z}_* \boldsymbol{X}_* \boldsymbol{\beta})' \boldsymbol{W}_* (\boldsymbol{z}_* \boldsymbol{X}_* \boldsymbol{\beta})$ to get new $\boldsymbol{\beta}^+$ i.e.,

$$\beta^+ = (X'_*W_*X_*)^{-1}(X'_*W_*z_*)$$

We also get $\widehat{Var}(\beta^+)$

3. Approximate E-step: First note that

$$\mathbb{E}[(\beta_j - \mu_j)^2] = Var(\beta_j - \mu_j) + (\mathbb{E}[\beta_j - \mu_j])^2 \approx \widehat{Var}(\beta_j^+) + (\beta_j^+ - \mu_j)^2$$

Expected value of log posterior density is

$$\mathbb{E}(\ln[p(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X})]) \approx \sum_{i=1}^{J} \frac{1}{\sigma_{j}^{2}} \mathbb{E}[(\mu_{j} - \beta_{j})^{2}] + \mathbb{E}[-\frac{1}{2} \{\sum_{i=1}^{N} \frac{1}{w_{i}^{-1}} (z_{i} - \boldsymbol{x}_{i}'\boldsymbol{\beta})^{2} + \sum_{i=1}^{J} \ln(\sigma_{j}^{2})\} + constant - p(\sigma_{j}|v_{j},s_{j})]$$

Substitute $\mathbb{E}[(\mu_j - \beta_j)^2]$ with $\widehat{Var}(\beta_j^+) + (\beta_j^+ - \mu_j)^2$

4. M-step: Maximize approximate expected log posterior density $\mathbb{E}(ln[p(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X})])$ with respect to σ_j^2 and get

$$\sigma_j^2 = \frac{\widehat{Var}(\beta_j^+) + (\beta_j^+ - \mu_j)^2 + v_j s_j^2}{1 + v_i}$$

5. Repeat step 1 to 4 until convergence of β

4.3.1 Observation

Without step 3 and 4, the algorithm is the same as that for independent normal prior.

Without step 3 and 4 and data augmentation, the algorithm is the same as IRLS.

Normal and uniform distribution is special cases of Student-t distribution by setting degree of freedom v_j and scale s_j to infinity, respectively. When $v_j = 1$, Student-t distribution is the same as Cauchy distribution.

Gelman et al. (2008) suggest default values of hyper-parameters $\mu_j = 0$, $s_j = 2.5$ and $v_j = 1$ (So, it is Cauchy default prior) for non-intercept.

5 Reference

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

Gelman, A., Jakulin, A., Pittau, M. G., & Su., Y.-S. (2008). A Weakly Informative Default Prior Distribution For Logistic And Other Regression Models. Annals of Applied Statistics, 2(4), 1360-1383.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian Data Analysis (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b16018

Borman, S. (2006). The Expectation Maximization Algorithm. A Short Tutorial