4.5 Bayesian Logistic Regression

在该小结中，我们主要讨论两个问题：

- 我们为logistic回归模型的参数引入先验，并进而求参数的后验分布
- 基于得到的后验分布，求新数据的预测分布

由于logistic函数为非线性函数，以上两步的计算都比较困难，我们沿用4.4中的近似思想来寻找所求参数的近似解。

求后验分布(拉普拉斯近似)

为参数引入先验

$$p(w) = \mathrm{N}(w|m_0, S_0)$$

根据贝叶斯公式，我们有

$$p(w|t) \propto p(w)p(t|w)$$

则

$$\begin{aligned} lnp(w|t) = &-\frac{1}{2}(w-m_0)^T S_0^{-1}(w-m_0) \\ &+ \sum_{n=1}^{N}\{t_n ln y_n + (1-t_n)ln(1-y_n)\} + C \end{aligned}$$

在4.4节中，我们已求拉普拉斯近似的近似分布的协方差阵的逆，即精度阵

$$A = -\nabla\nabla lnf(z)|_{z=z_0}$$

所以

$$\begin{aligned} S_N^{-1} &= -\nabla\nabla lnp(w|t) \\ &= S_0^{-1} + \sum_{n=1}^{N} y_n(1-y_n)\phi_n\phi_n^T \end{aligned}$$

所以近似分布$q(w)$的协方差S_N。前文中我们提到，拉普拉斯近似要求目标分布与近似分布拥有相同的极大值点，我们用w_{MAP}表示，则所求近似分布为

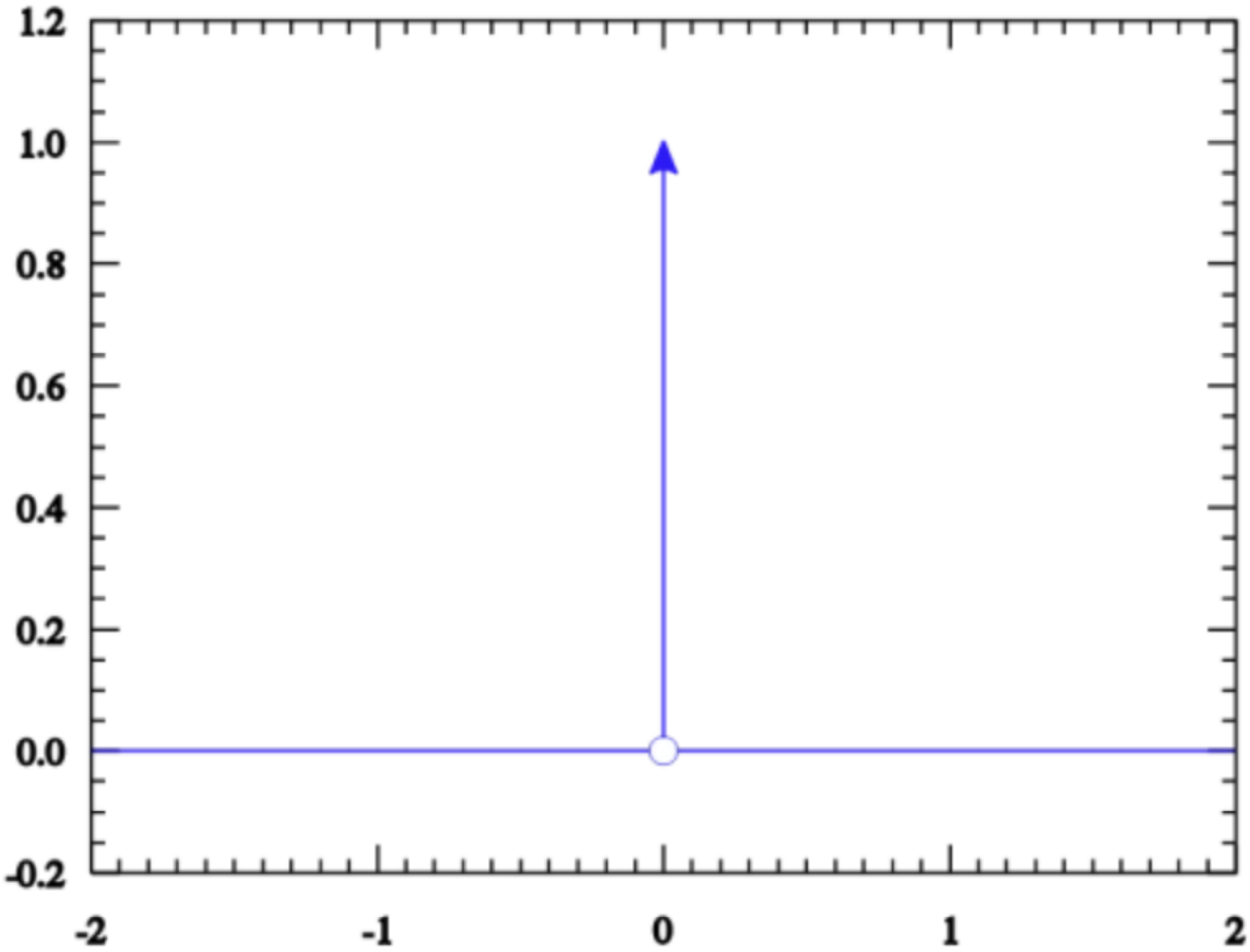
$$q(w) = \mathrm{N}(w|w_{MAP}, S_N)$$

求预测分布

δ函数

狄拉克函数，即信号学中的冲激函数，定义为：

$\delta(x) = \begin{cases} 0, x \neq 0 \\ \infty, x = 0 \end{cases}$ 且 $\int_{-\infty}^{+\infty}\delta(x)dx = 1$



选值作用

$$\int_{-\infty}^{+\infty} f(x)\delta(x)dx = f(0)$$

平移后有

$$\delta(x-a) = \begin{cases} 0, x \neq a \\ \infty, x = a \end{cases}, \qquad \int_{-\infty}^{+\infty} f(x)\delta(x-a)dx = f(a)$$

所以

$$\sigma(w^T\phi) = \int_{-\infty}^{\infty} \delta(a - w^T\phi)\sigma(a)da$$

当且仅当$a = w^T\phi$时，$\delta(a - w^T\phi)$的定义不为0

所以

$$\begin{aligned} \int \sigma(w^T\phi)q(w)dw &= \int \{\int \delta(a - w^T\phi)\sigma(a)da\}q(w)dw \\ &= \int\int \delta(a - w^T\phi)q(w)dw\sigma(a)da \\ &= \int \sigma(a)p(a)da \end{aligned}$$

因为$p(a)$可以看作$q(w)$的边缘分布，而$q(w)$本身是多维高斯分布，所以其边缘分布$p(a)$仍然是高斯分布，即$p(a)$服从高斯分布。

下面我们来求一下高斯分布$p(a)$的均值和协方差。

均值

- 根据定义我们有

$$
\begin{aligned}
\mu_a = E[a] &= \int p(a)ada \\
&= \int\int \delta(a - w^T\phi)q(w)dwada \\
&= \int\int \delta(a - w^T\phi)adaq(w)dw \\
&= \int w_T\phi q(w)dw \\
&= \int w^T q(w)dw\phi \\
&= w_{MAP}^T\phi
\end{aligned}
$$

不能用

$$
\begin{cases} p(a) = q(w) \\ a = w^T\phi \end{cases} \Rightarrow \int p(a)ada = \int q(w)w^T\phi dw^T\phi
$$

- 使用替换的方法

令$\frac{a}{\phi} = w_a$(这是不规范的，应该是$a\psi^T = w^T\phi\psi^T = w^T$，其中$\phi\psi^T = I$)

因为原本有

$$
p(a) = q(w_a)
$$

所以

$$
p(a) = q(\frac{a}{\phi})
$$

所以

$$
\mu_a = \int p(a)ada = \int q(\frac{a}{\phi}ada)
$$

令$t = \frac{a}{\phi}$，根据1.27式

$$
P_y(y) = P_x(x)|\frac{dx}{dy}| = P_x(g(y))|g'(y)| \qquad x = g(y)
$$

所以

$$
q(\frac{a}{\phi}) = q_t(t) = q_a(a)|\frac{da}{dt}| = q(a)\phi \qquad a = t\phi
$$

所以

$$
\mu_a = \int q(a)\phi ada = \int q(w)w^T\phi dw = w_{MAP}^T\phi
$$

协方差

$$
\begin{aligned}
\sigma_a^2 &= \int p(a)\{a - E[a]\}^2 da \\
&= \int \int \delta(a - w^T\phi) q(w) dw \{a - E[a]\}^2 da \\
&= \int \int \delta(a - w^T\phi) \{a - E[a]\}^2 da q(w) dw \\
&= \int \{w^T\phi - E[w^T\phi]\}^2 q(w) dw
\end{aligned}
$$

page219

We can evaluate $p(a)$ by noting that the delta function imposes a linear constraint on $\mathbf{w}$ and so forms a marginal distribution from the joint distribution $q(\mathbf{w})$ by integrating out all directions orthogonal to ϕ. Because $q(\mathbf{w})$ is Gaussian, we know from

【我们可以通过注意δ函数对w施加线性约束来评估$p(a)$，从而通过积分出与ϕ正交的所有方向，从联合分布$q(w)$形成边缘分布】

$q(w)$为联合分布：

$$q(w_1, w_2, \cdots, w_M)$$

因为要求$w^T\phi = a$，上式可写作

$$w_{\setminus i}^T \phi_{\setminus i} + w_i \phi_i = a$$

其中

$$w_{\setminus i}^T \phi_{\setminus i} = 0 \qquad w_i \phi_i = a$$

w_i表示在$\{w_1, w_2, \cdots, w_M\}$中的一个、两个或多个维度，$\setminus i$表示在$\{w_1, w_2, \cdots, w_M\}$中除去$w_i$以外的所有维度。

所以

$$
\begin{aligned}
p(a) &= \int \delta(a - w^T\phi) q(w) dw_1 dw_2 \cdots dw_m \\
&= \int \delta(a - w_i\phi_i) q(w_i) dw_i \\
&= q(w_i)
\end{aligned}
$$

即$p(a)$相当于一个边缘分布$q(w_i)$，其中被边缘化掉的方向是和ϕ_i正交的方向

二分类问题

预测分布为

$$p(C_1|\phi, t) = \int p(C_1|\phi, w) \quad p(w|t) \quad dw$$

我们已求

$$p(C_1|\phi, t) = \int \sigma(a)\mathrm{N}(a|\mu_a, \sigma_a^2)da$$

二分类问题的预测分布可用logistic函数乘以一个高斯分布再做积分表示，即logistic函数和高斯的卷积。但由于logistic函数本身的非线性，这个计算会十分困难，沿用以上思路求其近似解。

probit函数经过一定缩放后，其函数图像几乎可以和logistic函数完全重合。

$$\Phi(\lambda a) \approx \sigma(a) \qquad \lambda = \sqrt{\frac{\pi}{8}}$$

则

$$\begin{aligned} p(C_1|\phi, t) &= \int \sigma(a)\mathrm{N}(a|\mu_a, \sigma_a^2)da \\ &\approx \int \Phi(\lambda a)\mathrm{N}(a|\mu_a, \sigma_a^2)da \\ &= \Phi(\frac{\mu}{(\lambda^{-2} + \sigma^2)^{1/2}}) \end{aligned}$$

Q1:在用probit函数近似logistic函数时，为什么要使$\lambda = \sqrt{\frac{\pi}{8}}$?

令probit函数和logistic函数在原点出的斜率相等，即二者导数的最大值相等

对于logistic函数

$$\sigma' = \sigma(1 - \sigma) \overset{\text{最大值}}{\Longrightarrow} \sigma = \frac{1}{2} \Longrightarrow \sigma'_{max} = \frac{1}{4}$$

对于probit函数

$$\Phi(\lambda a) = \int_{-\infty}^{+\infty} \mathrm{N}(\theta|0, 1)d\theta$$

根据变上限积分求导公式

$$\frac{d}{dx}\{\int_{\varphi(x)}^{\phi(x)} f(t)dt\} = f(\phi(x))\phi'(x) - f(\varphi(x))\varphi'(x)$$

所以

$$\begin{aligned} \frac{d\Phi(\lambda a)}{da} &= \frac{d}{d\lambda a}\int_{-\infty}^{\lambda a} \mathrm{N}(\theta|0, 1)d\theta \frac{d\lambda a}{da} \\ &= \mathrm{N}(\lambda a|0, 1)\lambda \end{aligned}$$

$\mathrm{N}(\lambda a|0, 1)$在$\lambda a = 0$处取得最大值，且有

$$\frac{d\Phi(\lambda a)}{da}|_{max} = \frac{1}{\sqrt{2\pi}}\lambda$$

所以

$$\frac{\lambda}{\sqrt{2\pi}} = \frac{1}{4} \Longrightarrow \lambda = \sqrt{\frac{\pi}{8}}$$

Q2:为什么probit函数和高斯分布的卷积，即$\int \Phi(\lambda a)\mathrm{N}(a|\mu_a, \sigma_a^2)da$仍然是一个probit函数，且结果为$\Phi(\frac{\mu}{(\lambda^{-2}+\sigma^2)^{1/2}})$?

根据定义

$$\Phi(a)=\int_{-\infty}^{a}\mathrm{N}(x|0,1)dx\Longrightarrow x\sim\mathrm{N}(x|0,1)$$

所以

$$P(X\le a)=\int_{-\infty}^{a}\mathrm{N}(x|0,1)dx=\Phi(a)$$

若$x\sim\mathrm{N}(x|\mu,\sigma^2)$，因为

$$\frac{x-\mu}{\sigma}\sim\mathrm{N}(x|0,1)$$

所以有

$$\begin{aligned}P(X\le a)&=P(X-\mu\le a-\mu)\\&=P(\frac{x-\mu}{\sigma}\le\frac{a-\mu}{\sigma})\\&=P(z\le\frac{a-\mu}{\sigma})\\&=\Phi(\frac{a-\mu}{\sigma})\end{aligned}$$

其中$P(z)=\mathrm{N}(z|0,1)$

因为有

$$\Phi(\lambda a)=\Phi(\frac{a-0}{\lambda^{-1}})\Longrightarrow P(X)\sim\mathrm{N}(X|0,\lambda^{-2})$$

所以

$$\begin{aligned}\int\Phi(\lambda a)\mathrm{N}(x|\mu,\sigma^2)da&=\int P_X(X\le a)f_Y(a)da\\&=\int P_X(X\le Y|Y=a)f_Y(Y=a)dY\\&=P_X(X\le Y)\text{（全概率）}\\&=P_X(X-Y\le 0)\end{aligned}$$

因为$X\sim\mathrm{N}(X|0,\lambda^{-2})$，$Y\sim\mathrm{N}(Y|\mu,\sigma^2)$，所以$z=X-Y$相当于$X$和$-Y$的卷积

$$z\sim\mathrm{N}(z|\mu_z,\sigma_z^2)$$

其中$\begin{cases}\mu_z=\mu_X-\mu_Y=0-\mu=-\mu\\\sigma_z^2=\sigma_X^2+\sigma_Y^2=\lambda^{-2}+\sigma^2\end{cases}$

所以有

$$\begin{aligned}\int\Phi(\lambda a)\mathrm{N}(a|\mu,\sigma^2)da&=P_z(z\le 0)\\&=P_z(z+\mu\le\mu)\\&=P_z(\frac{z+\mu}{\sqrt{\lambda^{-2}+\sigma^2}}\le\frac{\mu}{\sqrt{\lambda^{-2}+\sigma^2}})\\&=P_t(t\le\frac{\mu}{\sqrt{\lambda^{-2}+\sigma^2}})\\&=\Phi(\frac{\mu}{(\lambda^{-2}+\sigma^2)^{1/2}})\end{aligned}$$

其中$P(t)=\mathrm{N}(t|0,1)$

- 全概率公式

$$P(B) = \sum_{n=1}^{N} P(B|A_i)P(A_i)$$

- 概率密度的全概率公式

$$f_X(x) = \sum_{i=1}^{n} f(x|A_i)P(A_i)$$

- 连续型变量的全概率分布

$$P(A) = \int_{-\infty}^{+\infty} P(A|X = x)f_X(x)dx$$

以上，我们用probit函数去近似logistic函数，同样的，我们可以在probit函数的基础上用一个logistic函数去近似，得到结果

$$\begin{aligned}
p(C_1|\phi, t) &= \int \sigma(a)\mathrm{N}(a|\mu_a, \sigma_a^2)da \\
&\approx \int \Phi(\lambda a)\mathrm{N}(a|\mu_a, \sigma_a^2)da \\
&= \Phi(\frac{\mu}{(\lambda^{-2} + \sigma^2)^{1/2}}) \\
&\approx \sigma(k(\sigma^2)\mu)
\end{aligned}$$

已知

$$\Phi(\sqrt{\frac{\pi}{8}}x) \approx \sigma(x)$$

所以

$$\Phi(cx) = \Phi(\frac{\sqrt{2\pi}}{4}\{c\frac{4}{\sqrt{2\pi}}x\}) \approx \sigma(c\frac{4}{\sqrt{2\pi}}x)$$

令$c = \frac{1}{(\lambda^{-2}+\sigma^2)^{1/2}}$，所以

$$\begin{aligned}
c\frac{4}{\sqrt{2\pi}} &= \frac{1}{(\lambda^{-2} + \sigma^2)^{1/2}}\frac{4}{\sqrt{2\pi}} \\
&= (1 + \frac{\pi}{8}\sigma^2)^{-1/2} \\
&= k(\sigma^2) > 0
\end{aligned}$$

所以

$$\Phi(cx) \approx \sigma(k(\sigma^2)x)$$

所以

$$\Phi(c\mu) \approx \sigma(k(\sigma^2)\mu)$$

即

$$p(C_1|\phi, t) = \sigma(k(\sigma^2)\mu)$$