[redacted]

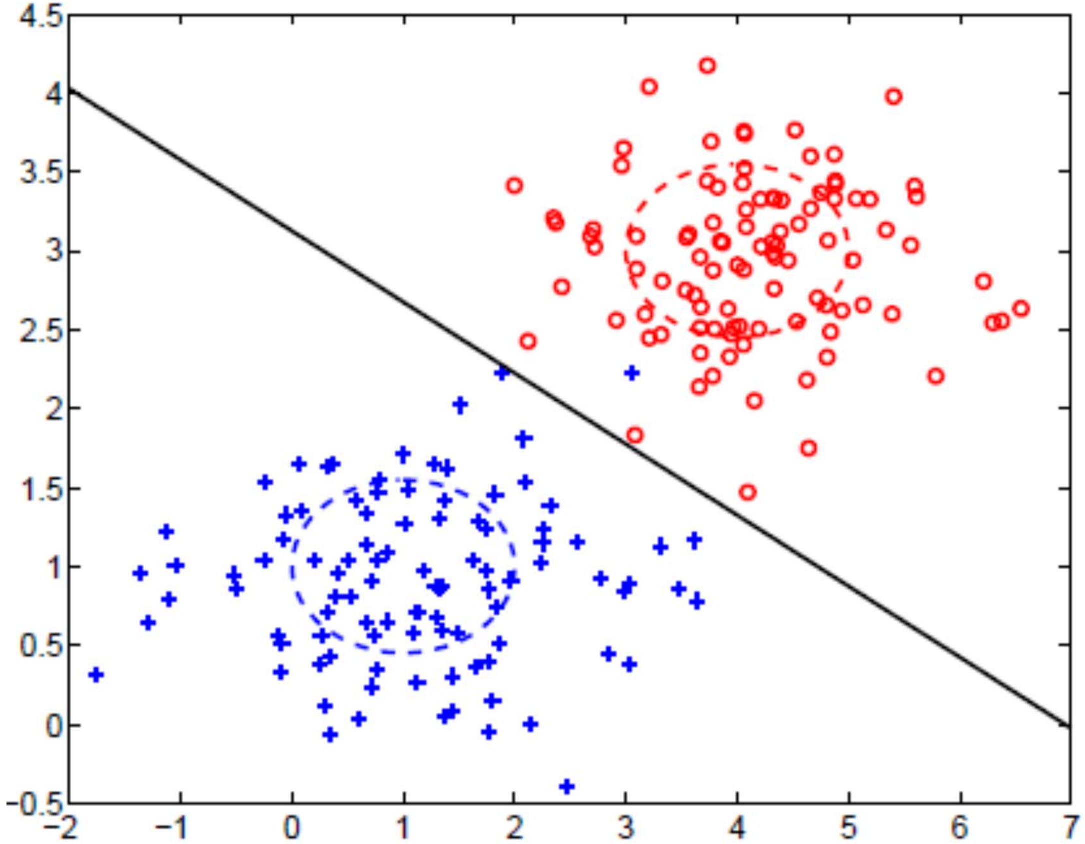
线性分类模型(2)

回顾

- 概率生成模型
  - 两类分类问题，类后验概率写成 x 线性函数的Logistic sigmoid函数
  - 多类分类问题，类后验概率写成 x 线性函数的 Softmax 变换
  - 类条件密度 $p(\mathbf{x} \mid c_k)$的参数使用最大似然估计及先验概率 $p(c_k)$，然后使用Bayes定理得到后验概率。
- 可选方法
  - 使用广义线性模型函数形式，直接使用最大似然确定函数的参数，如迭代再加权最小平方(iterative reweighted least squares, IRLS) 方法
- 直接方法
  - 最大化通过条件分布 $p(c_k \mid \mathbf{x})$ 定义的似然函数，表示判别式训练的形式。
  - 优点：一般只需确定少量自适应参数
  - 适应：类条件密度假设对真实分布近似很差的情形

判别式分类

- 如果只对分类决策感兴趣，为什么要在输入样本上建模呢?
- 对于给定样本，直接估计类标签的条件分布 $P(c_k \mid \mathbf{x}, \theta)$，其中：$\theta = \{\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \Sigma_0, \Sigma_1\}$



判别式分类

- 如果各个类别有相等的先验概率，那么给定样本 x 的标签 c_1 的后验概率为

$$P\left(c_1 \mid x, \theta\right) = \frac{p(x \mid \mu_1, \sigma_1^2)}{p(x \mid \mu_1, \sigma_1^2) + p(x \mid \mu_0, \sigma_0^2)}$$

$$= \frac{1}{1 + \dfrac{p(x \mid \mu_0, \sigma_0^2)}{p(x \mid \mu_1, \sigma_1^2)}}$$

$$= \frac{1}{1 + \exp\left\{-\log \dfrac{p(x \mid \mu_1, \sigma_1^2)}{p(x \mid \mu_0, \sigma_0^2)}\right\}}$$

其中：$\theta = \{\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \sigma_1^2, \sigma_2^2\}$

后验概率的形式

- 因为决策边界是线性的或二次型的，对某些参数 **w**，有

$$\log\frac{P(x\mid\boldsymbol{\mu}_1,\sigma_1^2)}{P(x\mid\boldsymbol{\mu}_0,\sigma_0^2)}=\begin{cases}w_0+w_1x & \text{if }\sigma_1^2=\sigma_0^2\\ w_0'+w_1'x+w_2'x^2 & \text{otherwise}\end{cases}$$

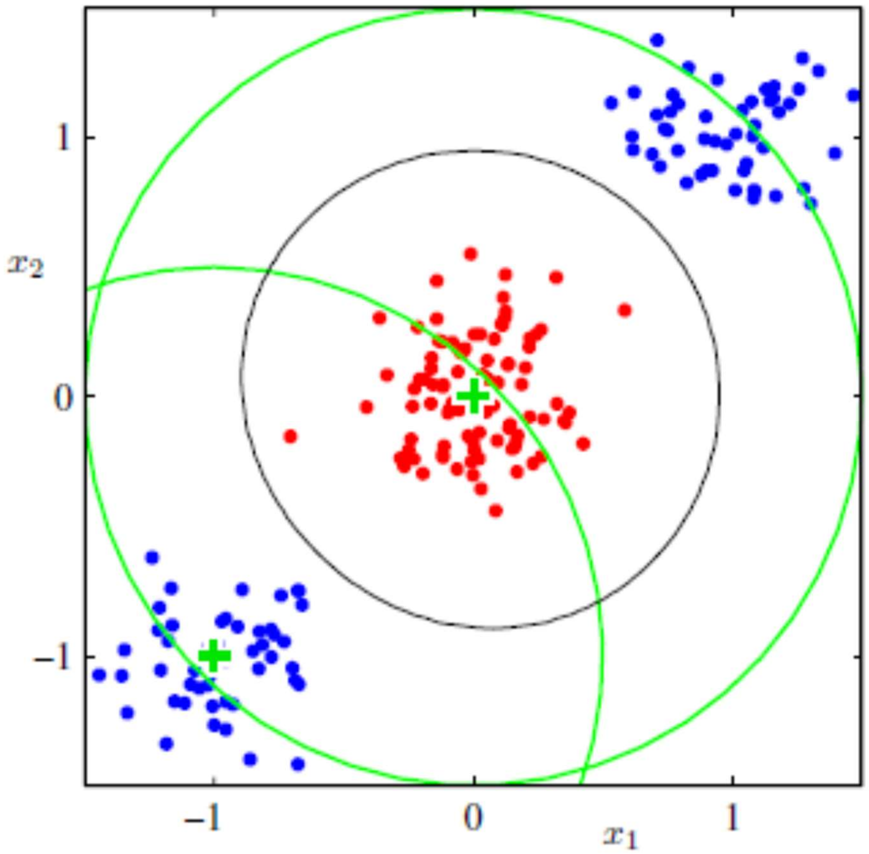
当$\sigma_1^2=\sigma_0^2$，有

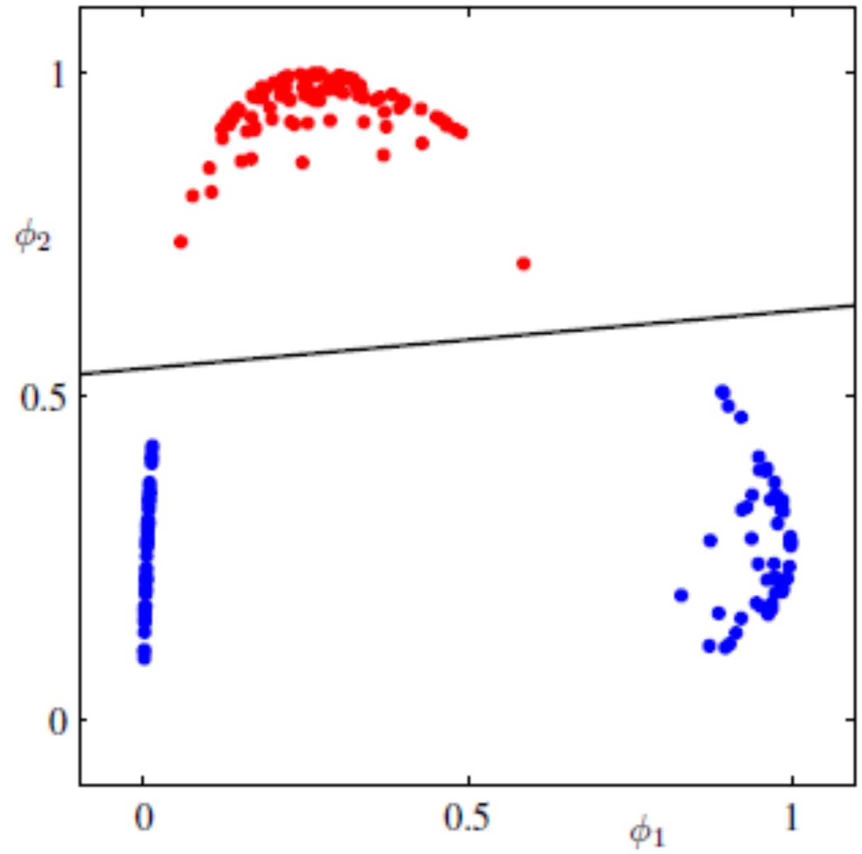
$$P\left(c_1\mid x,\theta\right)=\frac{1}{1+\exp\left\{-\log\dfrac{p\left(x\middle|\mu_1,\sigma_1^2\right)}{p\left(x\middle|\mu_0,\sigma_0^2\right)}\right\}}$$

$$=\frac{1}{1+\exp\left\{-\left(w_0+w_1x\right)\right\}}$$

固定基函数

- 分类模型
  - 在输入矢量 x 空间求解 ⇒ 应用于基函数 $\phi(\mathbf{x})$ 张开的非线性变换空间
  - 非线性决策边界 ⇒ 线性决策边界





固定基函数

- 在实际应用中
  - 类条件密度 $p(\mathbf{x}|c_k)$ 之间存在明显的重叠
  - 后验概率 $p(c_k|\mathbf{x})$ 在某些 x 处可能就是非 0 或非 1
  - 最优解：对后验概率准确建模，然后利用决策理论

- 非线性变换 $\phi(\mathbf{x})$
  - 不能除去类别重叠，甚至加大重叠
  - 合适选择 $\phi(\mathbf{x})$ 可以使后验概率建模更容易

逻辑回归模型

- 在两类问题中，类别 c_1 的后验概率通常简化为参数 w 的逻辑回归模型(logistic regression model)

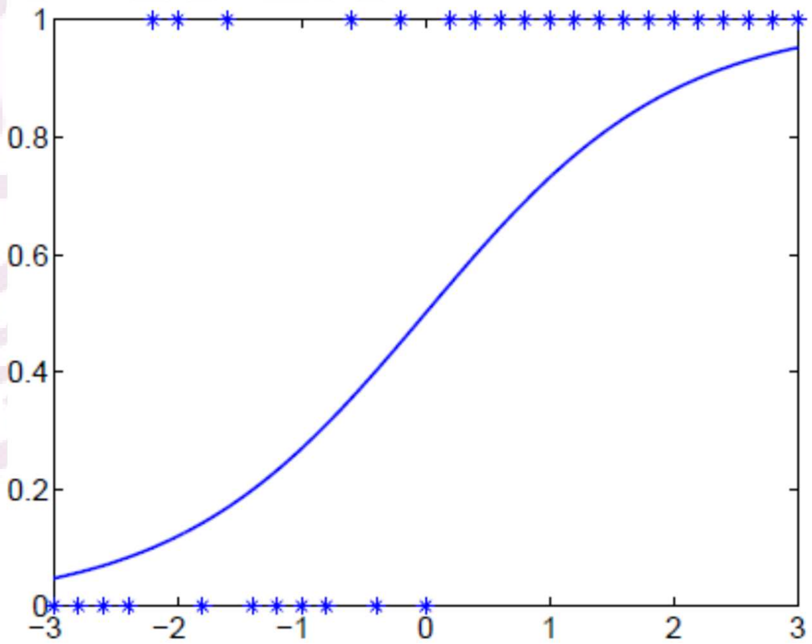
$$p\left(c_1 \middle| \phi, \mathbf{w}\right) = y(\phi) = \sigma\left(\mathbf{w}^{\mathrm{T}}\phi\right)$$

其中 $p(c_2 \mid \phi) = 1 - p(c_1 \mid \phi)$。

- Logistic sigmoid函数（挤压函数）

$$\sigma(z) = (1 + \exp(-z))^{-1}$$

将线性预测转换为概率。



逻辑回归模型

- 可以像线性回归模型一样，使用最大对数似然方法

$$l(\mathcal{D};\mathbf{w})=\sum_{i=1}^{n}\log p(c_1\mid\phi,\mathbf{w})$$

  来拟合逻辑回归模型

$$p(c_1\mid\phi,\mathbf{w})=\sigma(\mathbf{w}^{\mathrm{T}}\phi)$$

- 使用最大对数似然方法拟合高斯类条件密度分布和先验概率，需要 M(M+5)/2 + 1 个参数。
  - 2M 个均值参数，M(M+1)/2 个共享协方差矩阵参数和先验概率

- 提示：尽管可以将得到的参数与类条件均值和协方差矩阵有关，但是它们的数值与生成式方法中它们的数值会有很大不同。

最大似然确定逻辑回归模型参数

- 数据集 $\{\phi_n, t_n\}$，其中 $t_n \in \{0,1\}$ 且 $\phi_n = \phi(\mathbf{x}_n), n = 1,\ldots,N$
- 似然函数

$$p(\mathbf{t}\,|\,\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n}\{1-y_n\}^{1-t_n}$$

其中 $\mathbf{t} = \{t_1,\ldots,t_N\}^{\mathrm{T}}$ 且 $y_n = p(c_1\,|\,\phi_n)$ 。

- 交叉熵误差函数（负对数似然）

$$E(\mathbf{w}) = -\ln p(\mathbf{t}\,|\,\mathbf{w}) = -\sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}$$

其中 $y_n = \sigma(a_n)$ 且 $a_n = \mathbf{w}^{\mathrm{T}}\phi_n$ 。

最大似然确定逻辑回归模型参数

- 误差函数对矢量 w 的梯度

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n)\phi_n$$

其中 $d\sigma / da = \sigma(1-\sigma)$

- 迭代更新参数

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{\tau} - \eta \nabla E_n$$

其中： τ 是迭代次数， η 是学习率参数。

- 对线性可分数据集，最大似然解表现出严重的过拟合现象。
  - $\sigma = 0.5$ 等价于使用超平面 $\mathbf{w}^{\mathrm{T}}\phi = 0$ 分开两个类别，w 的幅值为无穷大
  - 在特征空间中，对应logistic sigmoid函数非常陡，每类 k 每个样本的后验概率 $p(c_k \mid \mathbf{x}) = 1$ 。
  - 解决办法：引入先验概率寻找 w 的MAP解，或添加正则项。

迭代再加权最小二乘法

- 最大似然解
  - 高斯噪声模型假设下，线性回归模型的最大似然解是闭合的
  - 非线性Logistic sigmoid函数使得逻辑回归模型最大似然解不是闭合的
- 逻辑回归模型的误差函数是凹的，具有唯一最小值
  - Newton-Raphson迭代优化方法使用局部二项式近似对数似然函数
- 参数更新公式

$$\mathbf{w}^{(\text{new})} = \mathbf{w}^{(\text{old})} - \mathbf{H}^{-1}\nabla E(\mathbf{w})$$

其中，H 是 Hessian 矩阵。

- 误差函数的梯度

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (\mathbf{w}^{\mathrm{T}}\phi_n - t_n)\phi_n = \Phi^{\mathrm{T}}\Phi\mathbf{w} - \Phi^{\mathrm{T}}\mathbf{t}$$

迭代再加权最小二乘法

- 误差函数的 Hessian 矩阵

$$\mathbf{H} = \nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} \phi_n \phi_n^{\mathrm{T}} = \Phi^{\mathrm{T}}\Phi$$

其中，Φ 是 N x M 维设计矩阵，第 n 行是 ϕ_n^{T}。

- Newton-Raphson更新公式

$$\mathbf{w}^{(\text{new})} = \mathbf{w}^{(\text{old})} - (\Phi^{\mathrm{T}}\Phi)^{-1}\{\Phi^{\mathrm{T}}\Phi\mathbf{w}^{(\text{old})} - \Phi\mathbf{t}\}$$

$$= (\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}\mathbf{t}$$

这是一个标准的最小二乘解。

迭代再加权最小二乘法

- 对于逻辑回归模型，将 Newton-Raphson 更新公式作用在交叉熵误差函数上。
- 交叉熵误差函数的梯度和 Hessian 矩阵

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n)\phi_n = \Phi^{\mathrm{T}}(\mathbf{y} - \mathbf{t})$$

$$\mathbf{H} = \nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} y_n(1 - y_n)\phi_n\phi_n^{\mathrm{T}} = \Phi^{\mathrm{T}}\mathbf{R}\Phi$$

其中，R 是 N x N 维的对角矩阵 $R_{nn} = y_n(1 - y_n)$

- 使用性质 $0 < y_n < 1$，对于任意矢量 u，存在 $\mathbf{u}^{\mathrm{T}}\mathbf{H}\mathbf{u} > 0$，即 Hessian矩阵 H 是正定的。故，误差函数是凹的，具有唯一最小值。

迭代再加权最小二乘法

- 逻辑回归模型的 Newton-Raphson 更新公式

$$\mathbf{w}^{(\text{new})} = \mathbf{w}^{(\text{old})} - (\Phi^{\mathrm{T}}\mathbf{R}\Phi)^{-1}\Phi^{\mathrm{T}}(\mathbf{y}-\mathbf{t})$$
$$= (\Phi^{\mathrm{T}}\mathbf{R}\Phi)^{-1}\{\Phi^{\mathrm{T}}\mathbf{R}\Phi\mathbf{w}^{(\text{old})} - \Phi^{\mathrm{T}}(\mathbf{y}-\mathbf{t})\}$$
$$= (\Phi^{\mathrm{T}}\mathbf{R}\Phi)^{-1}\Phi^{\mathrm{T}}\mathbf{R}\mathbf{z}$$

其中，z 是一个 N 维矢量

$$\mathbf{z} = \Phi\mathbf{w}^{(\text{old})} - \mathbf{R}^{-1}(\mathbf{y}-\mathbf{t})$$

  - 更新公式是加权最小二乘问题的正规方程组
  - R 与参数矢量 w 有关，每次 w 更新都需要计算新值
- 由此得名，迭代再加权最小二乘法(iterative reweighted least squares, IRLS)

迭代再加权最小二乘法

- 从逻辑回归模型中 t 的均值和方差，可以将对角加权矩阵 R 解释为方差。

$$\mathbb{E}[t] = \sigma(\mathbf{x}) = y$$

$$\mathrm{var}[t] = \mathbb{E}[t^2] - \mathbb{E}[t]^2 = \sigma(\mathbf{x}) - \sigma(\mathbf{x})^2 = y(1-y)$$

$$\Leftrightarrow$$

$$R_{nn} = y_n(1-y_n)$$

其中，上述推导利用到性质：对 $t \in \{0,1\}$，存在 $t^2 = t$ 。

- IRLS 可以看作在变量 $a = \mathbf{w}^{\mathrm{T}}\phi$ 空间中的线性问题。

迭代再加权最小二乘法

❖ 矢量 z 的第 n 个分量

$$a_n(\mathbf{w}) \simeq a_n(\mathbf{w}^{(\text{old})}) + \left.\frac{da_n}{dy_n}\right|_{\mathbf{w}^{(\text{old})}} (t_n - y_n)$$

$$= \phi_n^{\mathrm{T}}\mathbf{w}^{(\text{old})} - \frac{(y_n - t_n)}{y_n(1-y_n)} = z_n$$

- 解释：在围绕当前操作点 $\mathbf{w}^{(\text{old})}$ 周围利用局部线性近似 logistic sigmoid 函数获得的空间中的有效目标值。

多类逻辑回归问题

- 对于最大的类分布，后验概率由特征变量线性函数的 Softmax 变换给出：

$$p(c_k \mid \phi) = y_k(\phi) = \frac{\exp(a_k)}{\sum_j \exp\left(a_j\right)}$$

其中，激活值为 $a_k = \mathbf{w}_k^{\mathrm{T}} \phi$。

- 使用最大似然分别确定类条件密度和先验概率，然后，使用 Bayes 定理得到对应的后验概率，由此隐含地确定参数 $\{\mathbf{w}_k\}$ 。

最大似然方法多类模型确定参数

- y_k 对激活值 a_j 的偏导数

$$\frac{\partial y_k}{\partial a_j} = y_k(I_{kj} - y_j)$$

其中，I_{kj} 是单位矩阵的元素。

- 似然函数

$$p(\mathbf{T} \mid \mathbf{w}_1, \ldots, \mathbf{w}_K) = \prod_{n=1}^{N} \prod_{k=1}^{K} p(c_k \mid \phi_n)^{t_{nk}} = \prod_{n=1}^{N} \prod_{k=1}^{K} y_{nk}^{t_{nk}}$$

其中 $y_{nk} = y_k(\phi_n)$，T 是 N x K 维目标变量矩阵，元素为 t_{nk}。

最大似然方法多类模型确定参数

- 负对数似然

$$E(\mathbf{w}_1,\ldots,\mathbf{w}_K) = -\ln p(\mathbf{T}\,|\,\mathbf{w}_1,\ldots,\mathbf{w}_K) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\ln y_{nk}$$

即，多类分类问题的交叉熵误差函数

- 误差函数对参数矢量 $\mathbf{w}_j$ 的梯度

$$\nabla_{\mathbf{w}_j} E(\mathbf{w}_1,\ldots,\mathbf{w}_K) = \sum_{n=1}^{N}(y_{nj}-t_{nj})\phi_n$$

其中 $\sum_k t_{nk} = 1$。

- 权矢量迭代更新公式

$$\mathbf{w}_j^{(\tau+1)} = \mathbf{w}_j^{\tau} - \eta\nabla_{\mathbf{w}_j} E_n$$

最大似然方法多类模型确定参数

❖ 多类问题 Newton-Raphson 更新公式中 Hessian 矩阵为

$$\mathbf{H} = \nabla_{\mathbf{w}_k} \nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\sum_{n=1}^{N} y_{nk}(I_{kj} - y_{nj}) \phi_n \phi_n^{\mathrm{T}}$$

- 与两类问题一样，多类逻辑回归模型的Hessian矩阵是正定的，误差函数是凹的，具有唯一最小值。

Probit 回归

- 一般求解步骤
  - 使用指数族函数描述类条件分布
  - 类后验概率就是对特征变量线性函数的 logistic (或 softmax) 变换结果
- 问题提出
  - 得到类后验概率的形式并不都是简单的
  - 希望探索其它类型的离散概率模型
- 两类问题的广义线性模型

$$p(t=1\mid a)=f(a)$$

其中：$a=\mathbf{w}^{\mathrm{T}}\phi$，$f(\cdot)$ 是激活函数。

Probit 回归

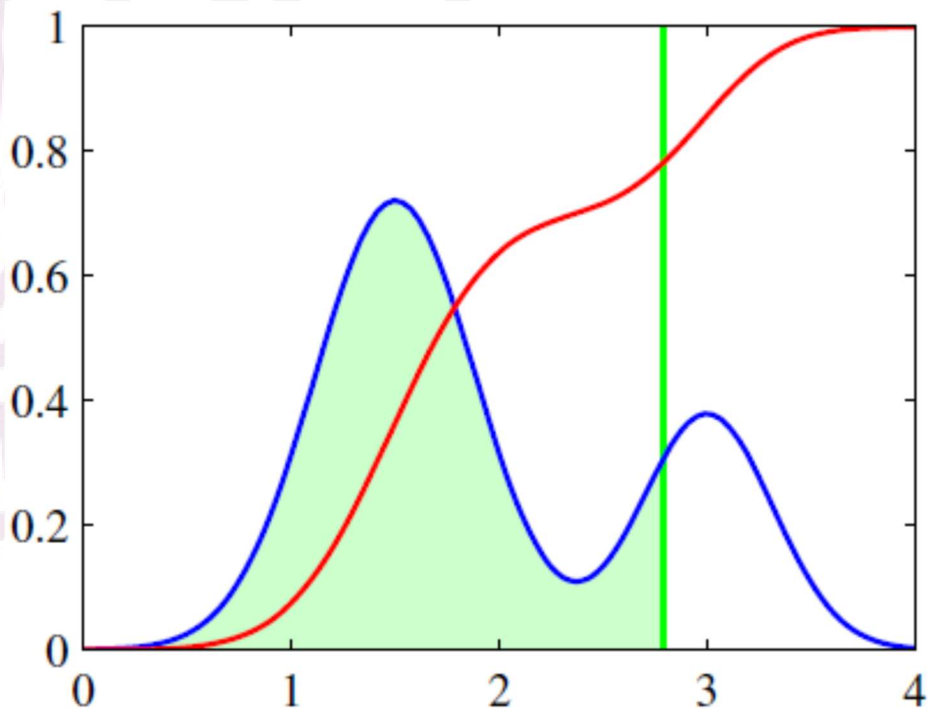
- 噪声阈值模型
  - 对每个输入 ϕ_n，计算 $a_n = \mathbf{w}^{\mathrm{T}}\phi_n$，然后，目标值为

$$t_n = \begin{cases} 1 & \text{if } a_n \geq \theta \\ 0 & \text{otherwise} \end{cases}$$

  - 如果 θ 的取值满足概率密度 $p(\theta)$，那么对应的激活函数将有累计分布给出：

$$f(a) = \int_{-\infty}^{a} p(\theta)d\theta$$

    - 蓝色曲线：概率密度 $p(\theta)$
    - 红色曲线：累计分布函数 $f(a)$
    - 蓝色曲线任一点取值是红色曲线相同点的斜率
    - 红色曲线任一点取值等于蓝色曲线绿色阴影的面积



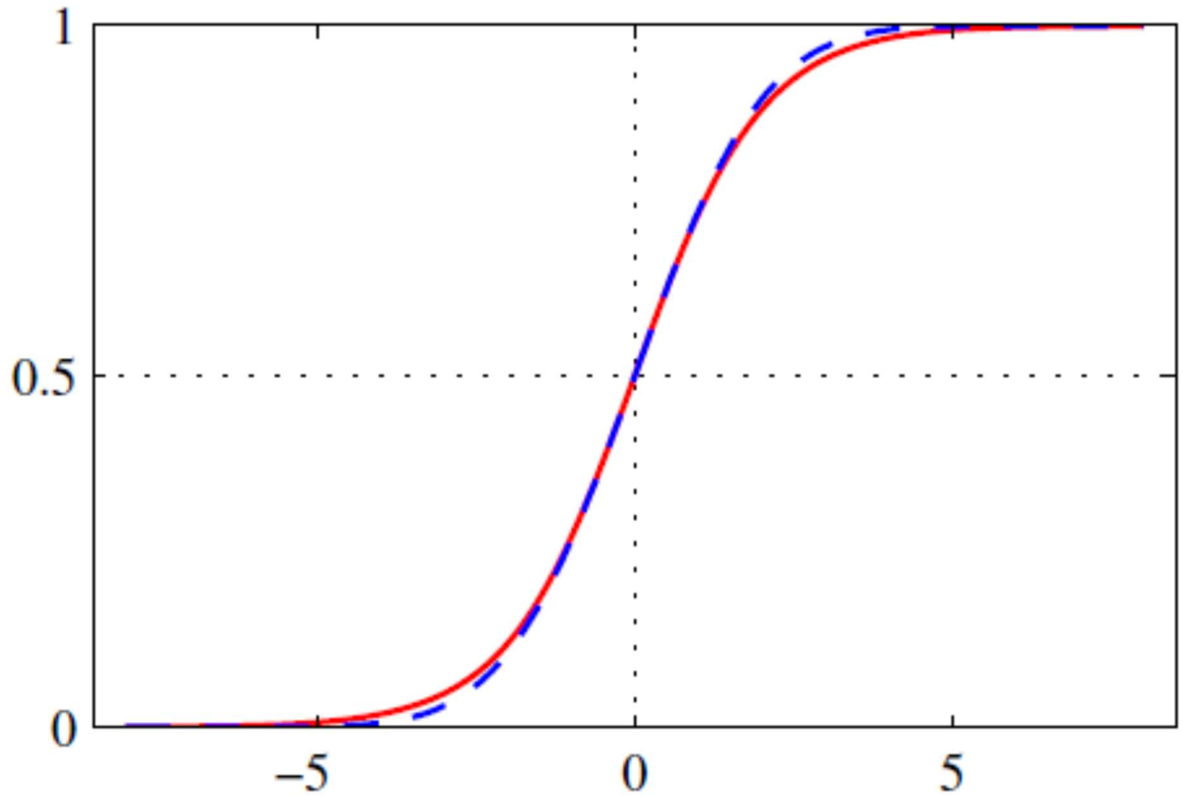
Probit 回归

❖ 如果 $p(\theta)$ 是零均值单位方差高斯函数，则累积分布函数为

$$\Phi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta \mid 0,1)\, d\theta$$

称为 probit 函数。

- 红色曲线：logistic sigmoid $\sigma(a)$
- 蓝色虚线：scaled probit function $\Phi(\lambda a)$, $\lambda^2 = \pi/8$



Probit 回归

- 使用更一般高斯分布不会改变模型，原因是等价于线性系数 w 的缩放。
- 定义：erf 函数

$$\mathrm{erf}(a)=\frac{2}{\sqrt{\pi}}\int_0^a \exp(-\theta^2/2)d\theta$$

与 probit 函数的关系：

$$\Phi(a)=\frac{1}{2}\left\{1+\frac{1}{\sqrt{2}}\mathrm{erf}\left(a\right)\right\}$$

- 基于 probit 激活函数的广义线性模型称为 probit 回归。
  - 使用早期讨论的思路，可以使用最大似然求解这个模型的参数。

Probit 回归

- 对于离群点，probit 回归模型比 logistic sigmoid模型具有更好的鲁棒性。
  - 对于 $x \to \infty$，logistic sigmoid 渐进衰减尾部相似于 $\exp(-x)$；而 probit 激活函数的衰减相似于 $\exp(-x^2)$。
- 通过引入目标值 t 翻转到错误数值的概率 ε，很容易将错误标签的影响纳入概率模型。
  - 对数据点 x，目标值分布为

$$
\begin{aligned}
p\left(t\middle|\mathbf{x}\right) &= (1-\varepsilon)\sigma(\mathbf{x}) + \varepsilon(1-\sigma(\mathbf{x})) \\
&= \varepsilon + (1-2\varepsilon)\sigma(\mathbf{x})
\end{aligned}
$$

其中 $\sigma(\mathbf{x})$ 是输入矢量 x 的激活函数。

正则连接函数

- 正则连接函数(canonical link function)：目标变量是指数族条件分布假设，以及激活函数的相应选择。
- 目标变量的条件分布

$$p(t\mid\eta,s)=\frac{1}{s}h\left(\frac{t}{s}\right)g(\eta)\exp\left\{\frac{\eta t}{s}\right\}$$

- 目标变量的条件均值

$$y\equiv\mathbb{E}[t\mid\eta]=-s\frac{d}{d\eta}\ln g(\eta)$$

因此， y 和 η 一定相关，这种关系表示为 $\eta=\psi(y)$

正则连接函数

- 广义线性模型(generalized linear model)：输入（或特征）变量线性组合的非线性函数

$$y = f(\mathbf{w}^{\mathrm{T}}\phi)$$

其中 $f(\cdot)$ 是机器学习中的激活函数(activation function)，$f^{-1}(\cdot)$ 称为统计学中的连接函数(link function)。

- 模型的似然函数（ η 的函数）

$$\ln p(\mathbf{t}\,|\,\eta, s) = \sum_{n=1}^{N} \ln p\left(t_n \,|\, \eta, s\right) = \sum_{n=1}^{N} \left\{ \ln g\left(\eta_n\right) + \frac{\eta_n t_n}{s} \right\} + \text{const}$$

假设所有观测量共享通常的比例参数（对应于实例高斯分布的噪声方差）， s 独立于 n 。

正则连接函数

- 对模型参数 w 的对数似然导数为

$$\nabla_{\mathbf{w}} \ln p(\mathbf{t}|\eta, s) = \sum_{n=1}^{N} \left\{ \frac{d}{d\eta_n} \ln g(\eta_n) + \frac{t_n}{s} \right\} \frac{d\eta_n}{dy_n} \frac{dy_n}{da_n} \nabla a_n$$

$$= \sum_{n=1}^{N} \frac{1}{s} \{t_n - y_n\} \psi'(y_n) f'(a_n) \phi_n$$

其中 $a_n = \mathbf{w}^{\mathrm{T}} \phi_n$。

- 选择连接函数为

$$f^{-1}(y) = \psi(y)$$

有 $f(\psi(y)) = y$ 且 $f'(\psi)\psi'(y) = 1$。也因为 $a = f^{-1}(y)$，有 $a = \psi$ 和 $f'(a)\psi'(y) = 1$。

正则连接函数

- 误差函数的梯度化简为

$$\nabla \ln E(\mathbf{w}) = \frac{1}{s}\sum_{n=1}^{N}\{y_n - t_n\}\phi_n$$

  - 对于高斯模型 $s=\beta^{-1}$，而对于 logistic 模型 $s=1$。

Laplace 近似法

- 假设，单个连续变量 z 的分布为

$$p(z)=\frac{1}{Z}f(z)$$

其中 $Z=\int f(z)$ 是归一化系数。

- 目标：寻找以分布 $p(z)$ 的众数为中心的高斯近似 $q(z)$
- $p(z)$ 的众数：点 z_0 满足

$$\left.\frac{df(z)}{dz}\right|_{z=z_0}=0$$

- 性质：随机变量高斯函数的对数函数是二次函数

Laplace 近似法

- $\ln f(z)$ 在众数 z_0 上的 Taylor 展开

$$\ln f(z) \simeq \ln f(z_0) - \frac{1}{2} A(z - z_0)^2$$

其中

$$A = -\frac{d^2}{dz^2} \ln f(z) \bigg|_{z=z_0}$$

  - 因为 z_0 是分布的局部最大值，故展开中没有一次项。

- 表示为指数形式

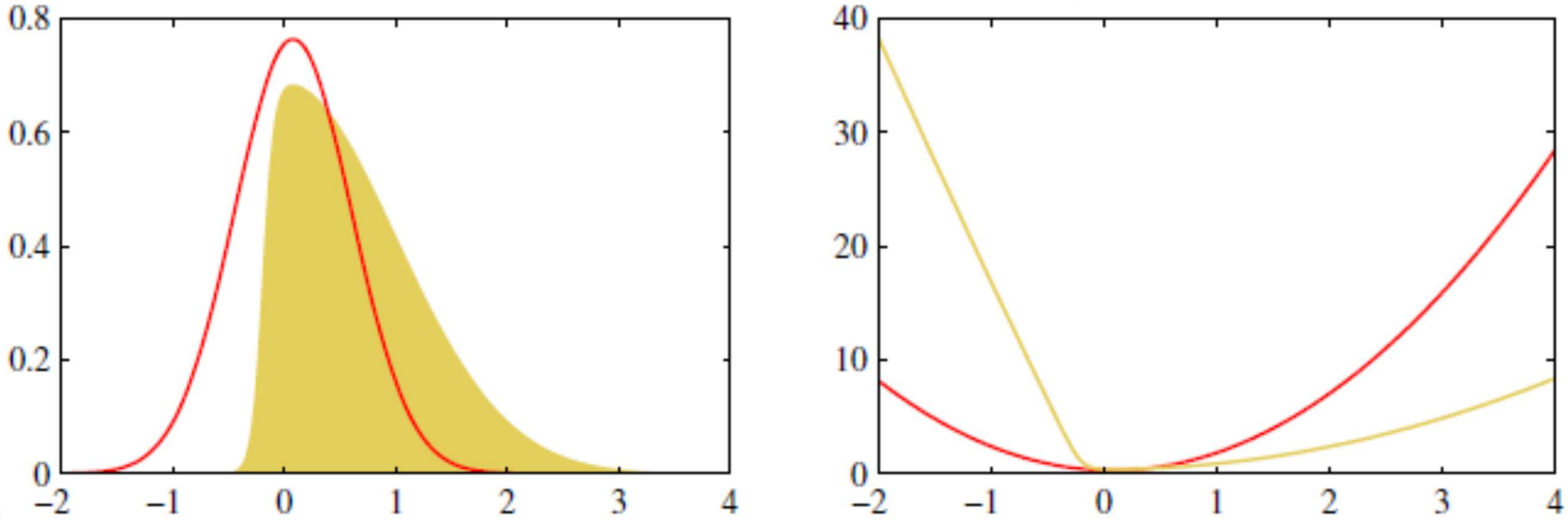
$$f(z) \simeq f(z_0) \exp\left\{ -\frac{A}{2}(z - z_0)^2 \right\}$$

Laplace 近似法

- 规范化分布

$$q(z)=\left(\frac{A}{2\pi}\right)^{1/2}\exp\left\{-\frac{A}{2}(z-z_0)^2\right\}$$

- 分布示意图
  - 分布 $p(z)\propto\exp(-z^2/2)\sigma(20z+4)$
  - 左图：黄色曲线为 $p(z)$，红色为在众数 z_0 的Laplace近似
  - 右图：左边曲线对应的负对数曲线



Laplace 近似法

- 定义在 M 维空间 z 上的分布 $p(\mathbf{z}) = f(\mathbf{z})/Z$，驻点 $\mathbf{z}_0$ 周围展开为

$$\ln f(\mathbf{z}) \simeq \ln f(\mathbf{z}_0) - \frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^{\mathrm{T}} \mathbf{A}(\mathbf{z}-\mathbf{z}_0)$$

其中，M x M 维的 Hessian 矩阵 A 定义为

$$\mathbf{A} = -\nabla\nabla \ln f(\mathbf{z})\big|_{\mathbf{z}=\mathbf{z}_0}$$

- 两边取指数

$$f(\mathbf{z}) \simeq f(\mathbf{z}_0)\exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^{\mathrm{T}} \mathbf{A}(\mathbf{z}-\mathbf{z}_0)\right\}$$

Laplace 近似法

- 近似分布

$$q(\mathbf{z}) = \frac{|\mathbf{A}|^{1/2}}{(2\pi)^{M/2}} \exp\left\{ -\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^{\mathrm{T}} \mathbf{A} (\mathbf{z}-\mathbf{z}_0) \right\} = \mathcal{N}\left(\mathbf{z} \middle| \mathbf{z}_0, \mathbf{A}^{-1}\right)$$

其中 |A| 表示矩阵 A 的行列式。

- A 是正定的
- 驻点 $\mathbf{z}_0$ 必须是局部最大值点，不是最小值点和鞍点

- 步骤
  - 寻找众数
    - 可通过数值优化算法发现
    - 针对实践中的多模分布，根据考虑的众数不同会有不同的近似
  - 在众数上计算 Hessian 矩阵

Laplace 近似法

- 特点
  - 在相对大量数据点的情形下，是最有用的。
  - 因为基于高斯分布，所以只能直接应用在实数变量上。
  - 由于纯粹基于变量某个特定值的真实分布上，所以缺乏全局视角。

模型比较

- 归一化常数

$$
\begin{aligned}
Z &= \int f(\mathbf{z})\,d\mathbf{z} \\
&\simeq f(\mathbf{z}_0)\int \exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^{\mathrm{T}}\mathbf{A}(\mathbf{z}-\mathbf{z}_0)\right\}d\mathbf{z} \\
&= f(\mathbf{z}_0)\frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}}
\end{aligned}
$$

- 问题：
  - 数据集合$\mathcal{D}$和参数为$\{\theta_i\}$的一组模型$\{\mathcal{M}_i\}$
  - 对于每个模型，定义似然函数 $p(\mathcal{D}|\theta_i,\mathcal{M}_i)$
  - 引入参数的先验概率 $p(\theta_i|\mathcal{M}_i)$
  - 针对各个模型，计算模型证据(model evidence) $p(\mathcal{D}|\mathcal{M}_i)$

模型比较

❖ 应用 Bayes 定理，得到模型证据(省略条件 $\mathcal{M}_i$)

$$p(\mathcal{D}) = \int p(\mathcal{D}\mid\theta)\,p(\theta)\,d\theta$$

❖ 应用归一化常数近似，有

- $f(\theta) = p(\mathcal{D}\mid\theta)\,p(\theta)$ 和 $Z = p(\mathcal{D})$

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\theta_{\mathrm{MAP}}) + \underbrace{\ln p(\theta_{\mathrm{MAP}}) + \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}|}_{\text{Occam factor}}$$

其中 θ_{MAP} 是在后验概率众数处的 θ 值，A 是负对数后验概率二阶导数的 Hessian 矩阵

$$\mathbf{A} = -\nabla\nabla \ln p(\mathcal{D}|\theta_{\mathrm{MAP}})\,p(\theta_{\mathrm{MAP}}) = -\nabla\nabla \ln p(\theta_{\mathrm{MAP}}|\mathcal{D})$$

- 右边第一项：使用优化参数计算的对数似然
- 右边后三项：惩罚模型复杂度的“奥坎姆因子”

模型比较

- 贝叶斯信息准则(Bayesian Information Criterion, BIC)
  - 假设：参数的高斯先验概率是宽阔的，且 Hessian 矩阵是满秩的

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\theta_{\mathrm{MAP}}) - \frac{1}{2} M \ln N$$

其中：N 是数据点数，M 是参数 θ 的数目

- 特点
  - 容易计算，但可能导致错误结果
  - 在实践中，因为参数不是唯一确定的，所以 Hessian 矩阵是满秩的假设通常无效。
  - 带有“奥坎姆因子”的近似计算会更加准确

贝叶斯逻辑回归

- 对逻辑回归的准确贝叶斯推理是棘手的(intractable)
  - 将 Laplace 近似应用在贝叶斯逻辑回归问题
- 高斯先验概率

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$$

其中 $\mathbf{m}_0$ 和 $\mathbf{S}_0$ 是固定超参数。

- 后验概率

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

其中 $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$。

贝叶斯逻辑回归

❖ 两边取对数

$$\ln p(\mathbf{w}|\mathbf{t}) = -\frac{1}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0)$$

$$+\sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} + \text{const}$$

其中:

- $y_n = \sigma(\mathbf{w}^{\mathrm{T}}\phi_n)$
- 似然函数

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n}\{1-y_n\}^{1-t_n}$$

贝叶斯逻辑回归

- 贝叶斯逻辑回归
  - 最大化后验概率分布得到最大后验概率解 $\mathbf{w}_{\text{MAP}}$，做为高斯分布的均值
  - 协方差矩阵

$$\mathbf{S}_N = -\nabla\nabla \ln p(\mathbf{w}|\mathbf{t}) = \mathbf{S}_0^{-1} + \sum_{n=1}^{N} y_n(1-y_n)\phi_n\phi_n^{\mathrm{T}}$$

  - 后验概率高斯近似

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{\text{MAP}}, \mathbf{S}_N)$$

  - 根据分布边缘化结果，做出预测。

贝叶斯逻辑回归

- 已知新特征矢量 $\phi(\mathbf{x})$，类别 c_1 的预测分布为

$$p(c_1|\phi,\mathbf{t})=\int p(c_1|\phi,\mathbf{w})\,p(\mathbf{w}|\mathbf{t})\,d\mathbf{w}\simeq\int\sigma(\mathbf{w}^{\mathrm{T}}\phi)\,q(\mathbf{w})\,d\mathbf{w}$$

对应类别c_2 的概率为 $p(c_2|\phi,\mathbf{t})=1-p(c_1|\phi,\mathbf{t})$

- 函数 $\sigma(\mathbf{w}^{\mathrm{T}}\phi)$ 与 w 的关系表现在 ϕ 的投影上，故

$$\sigma(\mathbf{w}^{\mathrm{T}}\phi)=\int\delta(a-\mathbf{w}^{\mathrm{T}}\phi)\,\sigma(a)\,da$$

其中： $a=\mathbf{w}^{\mathrm{T}}\phi$，$\delta(\cdot)$表示 Dirac 函数

- 获得

$$\int\sigma(\mathbf{w}^{\mathrm{T}}\phi)\,q(\mathbf{w})\,d\mathbf{w}=\int\sigma(a)\,p(a)\,da$$

其中

$$p(a)=\int\delta(a-\mathbf{w}^{\mathrm{T}}\phi)\,q(\mathbf{w})\,d\mathbf{w}$$

贝叶斯逻辑回归

- $p(a)$ 计算
  - Delta函数对 w 施加线性约束
  - $q(\mathbf{w})$ 的边缘分布通过所有与 ϕ 正交方向上积分获得
  - $q(\mathbf{w})$ 是高斯分布，所以边缘分布也是高斯分布
- 均值和方差

$$\mu_a = \mathbb{E}[a] = \int p(a)a\,da = \int q(\mathbf{w})\mathbf{w}^{\mathrm{T}}\phi\,d\mathbf{w} = \mathbf{w}_{\mathrm{MAP}}^{\mathrm{T}}\phi$$

$$\sigma_a^2 = \mathrm{var}[a] = \int p(a)\left\{a^2 - \mathbb{E}[a]^2\right\}da$$

$$= \int q(\mathbf{w})\left\{\left(\mathbf{w}^{\mathrm{T}}\phi\right)^2 - \left(\mathbf{m}_N^{\mathrm{T}}\phi\right)^2\right\}d\mathbf{w} = \phi^{\mathrm{T}}\mathbf{S}_N\phi$$

- 预测分布的变分近似
  - 积分表示为高斯函数和 logistic sigmoid函数的卷积，不能解析计算。

$$p(c_1|\mathbf{t}) = \int \sigma(a)p(a)\,da = \int \sigma(a)\mathcal{N}\left(a|\mu_a,\sigma_a^2\right)da$$

贝叶斯逻辑回归

- 使用 probit 函数 $\Phi(\lambda a)$ 近似 $\sigma(a)$，其中 $\lambda^2 = \pi/8$。
- Probit 函数与高斯函数的卷积可以有解析表达

$$\int \Phi(\lambda a)\mathcal{N}\left(a \middle| \mu_a, \sigma_a^2\right) da = \Phi\left(\frac{\mu_a}{\left(\lambda^{-2} + \sigma_a^2\right)^{1/2}}\right)$$

- Logistic sigmoid与高斯函数卷积的近似

$$\int \sigma(a)\mathcal{N}\left(a \middle| \mu_a, \sigma_a^2\right) da \simeq \sigma\left(\kappa\left(\sigma_a^2\right)\mu_a\right)$$

其中

$$\kappa\left(\sigma_a^2\right) = \left(1 + \pi\sigma_a^2/8\right)^{-1/2}$$

- 预测分布近似

$$p\left(c_1 \middle| \phi, \mathbf{t}\right) = \sigma\left(\kappa\left(\sigma_a^2\right)\mu_a\right)$$

诚信 创新 实践