模式识别与机器学习

Pattern Recognition & Machine Learning

第三讲 逻辑回归

- 本节学习目标

  - ✓掌握线性回归及其模型求解方法
  - ✓理解贝叶斯线性回归
  - ✓掌握逻辑回归及其模型求解方法
  - ✓理解贝叶斯逻辑回归

（a）大熊猫

（b）小熊猫

（c）棕熊

例：通过熊猫食量估计来介绍如何使用线性回归建模数据。在一篇关于圈养大熊猫食竹量观察的文献中，记录了四只大熊猫的夜间食竹量，如下表所示

熊猫名称	性别	年龄/岁	体重/kg	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
莉莉	雌	10-11	102.5	2.8	3.3	2.6	3.5	2.7	4.9	1.3	1.7	1.9	1.6	2.5	3.9
青青	雌	3-4	82.5	3.4	3.7	3.7	3.9	4.1	5.7	1.6	2.1	2.4	2.7	3.3	4.1
金金	雄	22-23	128.0	1.9	2.5	1.7	2.1	2.2	4.5	1.1	1.5	1.2	1.7	1.7	2.1
平平	雄	9-10	82.0	4.2	4.4	4.1	4.6	4.5	6.9	3.2	3.5	3.4	3.4	3.7	4.5

晶晶 雌 4岁 100kg 各月份的食竹量？

目录

给定有N个样本的数据集$\mathcal{D} = \{y_i, x_{i1}, \cdots, x_{iD}\}_{i=1}^{N}$，线性回归模型假设因变量$y_i$与自变量$\mathbf{x_i}$（由$\{x_{i1}, \cdots, x_{iD}\}$构成的$D$维向量）间是线性关系。此关系通过回归系数$\boldsymbol{\beta}$构建，模型的形式如下：

$$y_i = \beta_0 1 + \beta_1 x_{i1} + \cdots + \beta_D x_{iD} = \mathbf{x_i^T}\boldsymbol{\beta}, i = 1, \cdots, N$$

自变量不一定是原始的数据特征，可以是原始特征的非线性函数。假设$\boldsymbol{\phi}(\mathbf{x_i})$表示对输入特征得变换函数，也称为基函数，那么线性函数可以更一般表示为

$$y_i = \boldsymbol{\phi}(\mathbf{x_i})^{\mathrm{T}}\boldsymbol{\beta}$$

常见基函数有三种：

1. 多项式基函数：

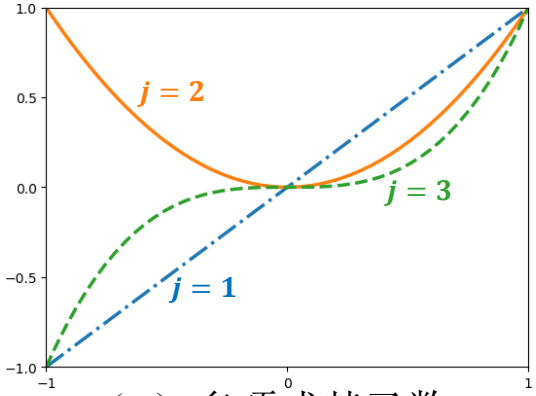
$$\phi_j(x) = x^j$$

2. 高斯基函数：

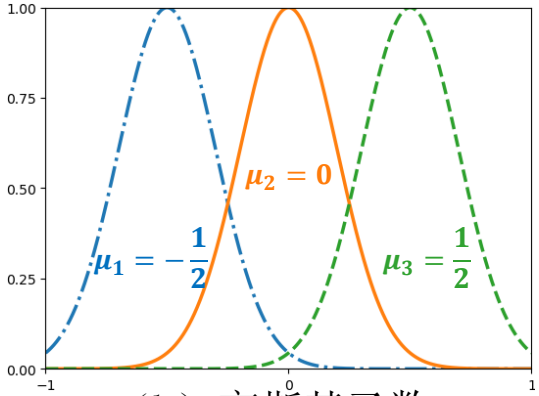
$$\phi_j(x) = \exp\{-\frac{(x-\mu_j)^2}{2s^2}\}$$

3. S形（sigmoidal）基函数：

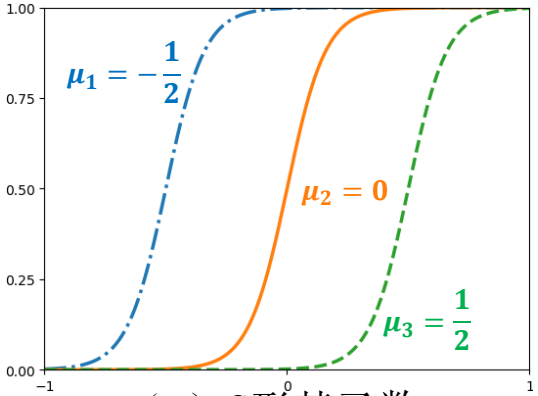
$$\phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right), \qquad \sigma(a) = \frac{1}{1+\exp(-a)}$$



（a）多项式基函数



（b）高斯基函数



（c）S形基函数

例：通过熊猫食量估计来介绍如何使用线性回归建模数据。在一篇关于圈养大熊猫食竹量观察的文献中，记录了四只大熊猫的夜间食竹量，如下表所示

熊猫名称	性别	年龄/岁	体重/kg	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
莉莉	雌	10-11	102.5	2.8	3.3	2.6	3.5	2.7	4.9	1.3	1.7	1.9	1.6	2.5	3.9
青青	雌	3-4	82.5	3.4	3.7	3.7	3.9	4.1	5.7	1.6	2.1	2.4	2.7	3.3	4.1
金金	雄	22-23	128.0	1.9	2.5	1.7	2.1	2.2	4.5	1.1	1.5	1.2	1.7	1.7	2.1
平平	雄	9-10	82.0	4.2	4.4	4.1	4.6	4.5	6.9	3.2	3.5	3.4	3.4	3.7	4.5

$$\mathbf{x} = (x_1, x_2, x_3, x_4)$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}$$

使用高斯随机噪声实现概率建模

观测输出被假设为确定性的线性回归再加上高斯随机噪声

$$y = f(\mathbf{x},\boldsymbol{\beta}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0,\sigma^2),$$

其中 $f(\mathbf{x},\boldsymbol{\beta}) = \phi(\mathbf{x}_i)^\top \boldsymbol{\beta}.$

根据概率的线性变换关系，可以得到每个观测数据的似然概率分布为

$$p(y|\mathbf{x},\boldsymbol{\beta},\sigma) = \mathcal{N}(y \mid f(\mathbf{x},\boldsymbol{\beta}),\sigma^2).$$

假设数据是独立同分布的，所有观测的似然概率分布为

$$p(\mathbf{y} \mid \mathrm{X},\boldsymbol{\beta},\sigma) = \prod_{i=1}^{N} \mathcal{N}(y_i \mid f(\mathbf{x}_i,\boldsymbol{\beta}),\sigma^2).$$

在确定了模型的概率表示之后，对于新的测试数据，可以使用输出变量的期望作为预测值

$$\mathbb{E}[y \mid \mathbf{x}_*] = \int y p(y \mid \mathbf{x}_*,\boldsymbol{\beta},\sigma) dy = f(\mathbf{x}_*,\boldsymbol{\beta}).$$

- 最小二乘与最大似然

给定有N个数据点$(\mathbf{x}_i, y_i)$的数据集，其中$\mathbf{x}_i$为自变量，y_i为因变量。模型函数具有形式$f(\mathbf{x}_i, \boldsymbol{\beta})$，其中$\boldsymbol{\beta}$保存了$D$个可调整的参数。最小二乘问题的**目标**为调整模型函数的参数来最好地拟合数据集。

模型对数据的拟合程度是通过其误差来测量的。**误差**定义为因变量的真实值和模型预测值之间的差：

$$e_i = y_i - f(\mathbf{x}_i, \boldsymbol{\beta}).$$

最小二乘通过最小化平方误差和S开学习最优参数值：

$$S = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left(y_i - f(\mathbf{x}_i, \boldsymbol{\beta})\right)^2.$$

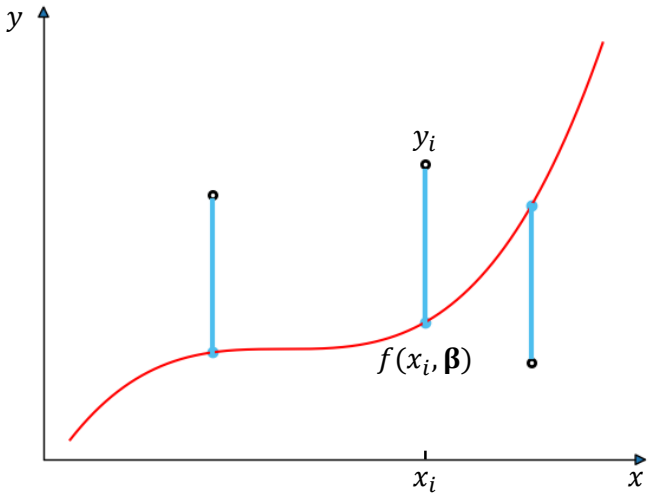


图3-2 误差几何意义示意图（图中纵向线段长度代表不同数据点的误差。）

$$S=\sum_{i=1}^{N}e_i^2=\sum_{i=1}^{N}\left(y_i-f(\mathbf{x}_i,\boldsymbol{\beta})\right)^2.$$

上式最小值可通过将对优化目标关于参数的导数设为0求解得到。如果分别考虑每一个参数，那么由于模型有D个参数，有D个梯度方程

$$\frac{\partial S}{\partial \beta_d}=0,\ d=1,2,...,D,$$

$$-2\sum_{i=1}^{N}\left(y_i-f(\mathbf{x}_i,\boldsymbol{\beta})\right)\frac{\partial f(\mathbf{x}_i,\boldsymbol{\beta})}{\partial \beta_d}=0,\ d=1,2,...,D.$$

以线性回归问题为例，具体介绍最小二乘法的解。

一般的线性回归模型表示为 $f(\mathbf{x}_i,\boldsymbol{\beta})=\phi(\mathbf{x}_i)^{\top}\boldsymbol{\beta}$，定义 $X=[\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_N]^{\top}$，$\mathbf{y}=[y_1,y_2,...,y_N]^{\top}$，$\Phi=[\phi(\mathbf{x}_1),\phi(\mathbf{x}_2),...,\phi(\mathbf{x}_n)]^{\top}$，那么模型在训练数据上的预测平方误差为

$$S=(\mathbf{y}-\Phi\boldsymbol{\beta})^{\top}(\mathbf{y}-\Phi\boldsymbol{\beta}).$$

$$\frac{dS}{d\boldsymbol{\beta}}=\frac{d\left((\mathbf{y}-\Phi\boldsymbol{\beta})^{\top}(\mathbf{y}-\Phi\boldsymbol{\beta})\right)}{d\boldsymbol{\beta}}=\mathbf{0}^{\top}.$$

其中，$\mathbf{0}$表示元素为0的列向量，$d\left((\mathbf{y}-\Phi\boldsymbol{\beta})^{\top}(\mathbf{y}-\Phi\boldsymbol{\beta})\right)$ 可利用向量微积分的运算法则（见附录C）进一步化简：

$$\begin{aligned}d\left((\mathbf{y}-\Phi\boldsymbol{\beta})^{\top}(\mathbf{y}-\Phi\boldsymbol{\beta})\right)&=\left(d(\mathbf{y}-\Phi\boldsymbol{\beta})^{\top}\right)(\mathbf{y}-\Phi\boldsymbol{\beta})+(\mathbf{y}-\Phi\boldsymbol{\beta})^{\top}d(\mathbf{y}-\Phi\boldsymbol{\beta})\\&=2(\mathbf{y}-\Phi\boldsymbol{\beta})^{\top}d(\mathbf{y}-\Phi\boldsymbol{\beta})\\&=-2(\mathbf{y}-\Phi\boldsymbol{\beta})^{\top}\Phi d\boldsymbol{\beta}\\&=2(\boldsymbol{\beta}^{\top}\Phi^{\top}\Phi-\mathbf{y}^{\top}\Phi)d\boldsymbol{\beta}.\end{aligned}$$

因此，得到$\boldsymbol{\beta}$最优解 $\hat{\boldsymbol{\beta}}_{ls}=(\Phi^{\top}\Phi)^{-1}\Phi^{\top}\mathbf{y}.$

当概率线性回归的似然假设为高斯分布时，其对数似然的表达式可以进一步推导得出

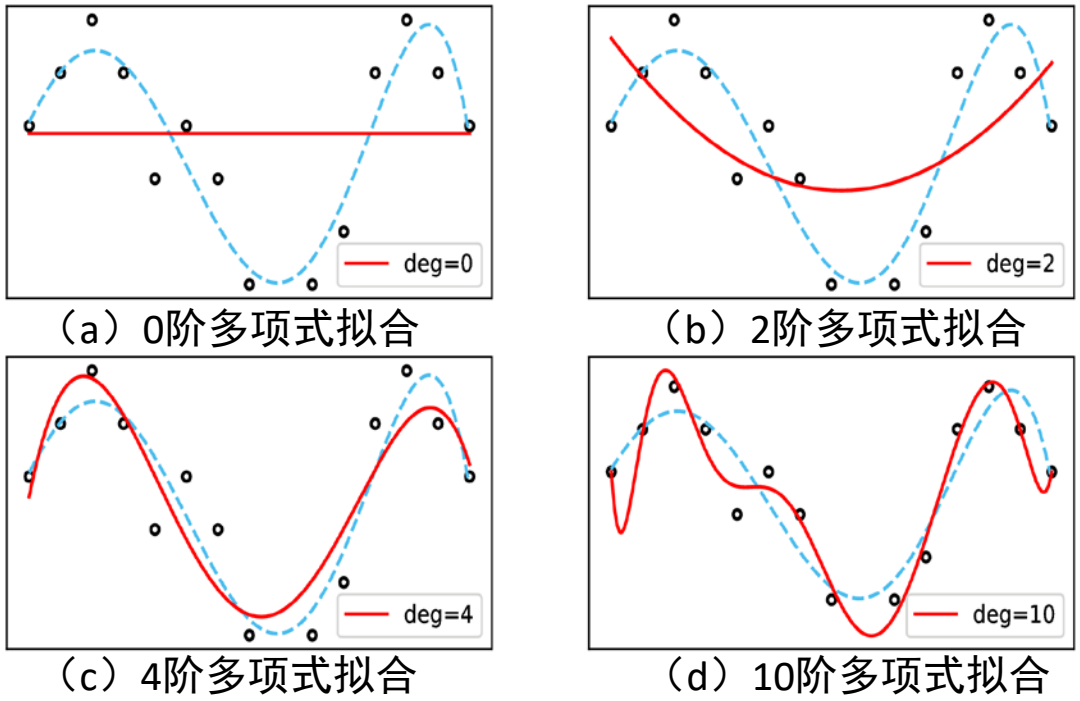
$$\ln p(\mathbf{y} \mid X, \boldsymbol{\beta}, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y_i - f(\mathbf{x}_i, \boldsymbol{\beta}) \right)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi).$$

最大化上式可以获得参数$\boldsymbol{\beta}$和σ^2的最大似然估计

$$\hat{\boldsymbol{\beta}}_{\mathrm{m}\ell} = (\Phi^{\top}\Phi)^{-1}\Phi^{\top}\mathbf{y},$$

$$\hat{\sigma}^2_{\mathrm{m}\ell} = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - f(\mathbf{x}_i, \boldsymbol{\beta}_{m\ell}) \right)^2.$$

- 正则化最小二乘与最大后验



（a）0阶多项式拟合　（b）2阶多项式拟合

（c）4阶多项式拟合　（d）10阶多项式拟合

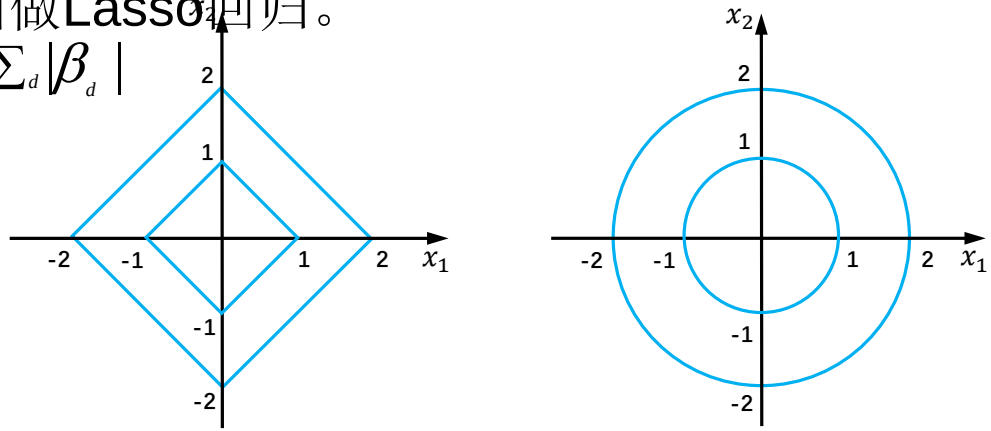
图3-3 四种不同的多项式的拟合效果

（图中小圆圈表示样本，虚线表示真实情况，实线表示拟合曲线，使用的多项式形式为 $f(x)=\sum_{j=0}^{\deg} w_j x^j$ ，deg表示多项式的阶数，四张子图分别使用不同的阶数）

对最小二乘进行正则化的方法叫做**正则化最小二乘**。
约束回归系数构成的向量的L_2范数的平方（$\|\boldsymbol{\beta}\|_{L_2}=\sqrt{\boldsymbol{\beta}^{\top}\boldsymbol{\beta}}$）不超过一个给定值。该约束相当于求解一个带有惩罚项（penalty term）$\lambda\|\boldsymbol{\beta}\|^2$的最小二乘的无约束最小化问题。此时，正则化最小二乘的优化目标为

$$S'=\sum_{i=1}^{N}\left(y_i-f(\mathbf{x}_i,\boldsymbol{\beta})\right)^2+\lambda\boldsymbol{\beta}^{\top}\boldsymbol{\beta},$$

其中λ是常数，可以通过模型选择的方法确定取值。使用L_2范数作为惩罚项的正则化最小二乘也叫做岭回归。使用L_1范数作为惩罚项的正则化最小二乘也叫做Lasso回归。
L_1范数：$\|\boldsymbol{\beta}\|_{L_1}=\sum_d|\beta_d|$



（a）L_1范数　　（b）L_2范数

（图中曲线表示二维空间中向量$\mathbf{x}=[x_1,x_2]^{\mathrm{T}}$的$L_1$范数$\left|\|\mathbf{x}\|\right|_{L_1}$和$l_2$范数$\left|\|\mathbf{x}\|\right|_{l_2}$的等高线。）

求解正则化最小二乘问题

对于使用L_2范数的正则化最小二乘，其最优解满足

$$\frac{dS}{d\boldsymbol{\beta}}=\frac{d\left((\mathbf{y}-\Phi\boldsymbol{\beta})^{\top}(\mathbf{y}-\Phi\boldsymbol{\beta})+\lambda\boldsymbol{\beta}^{\top}\boldsymbol{\beta}\right)}{d\boldsymbol{\beta}}=\mathbf{0}^{\top}$$

$d\left((\mathbf{y}-\Phi\boldsymbol{\beta})^{\top}(\mathbf{y}-\Phi\boldsymbol{\beta})\right)$可利用向量微积分的运算法则（见附录C）化简

$$\begin{aligned}d\left((\mathbf{y}-\Phi\boldsymbol{\beta})^{\top}(\mathbf{y}-\Phi\boldsymbol{\beta})+\lambda\boldsymbol{\beta}^{\top}\boldsymbol{\beta}\right)&=d(\mathbf{y}-\Phi\boldsymbol{\beta})^{\top}(\mathbf{y}-\Phi\boldsymbol{\beta})+(\mathbf{y}-\Phi\boldsymbol{\beta})^{\top}d(\mathbf{y}-\Phi\boldsymbol{\beta})\\&=2\left((\mathbf{y}-\Phi\boldsymbol{\beta})^{\top}d(\mathbf{y}-\Phi\boldsymbol{\beta})+\lambda\boldsymbol{\beta}^{\top}d\boldsymbol{\beta}\right)\\&=-2\left((\mathbf{y}-\Phi\boldsymbol{\beta})^{\top}\Phi-\lambda\boldsymbol{\beta}^{\top}\right)d\boldsymbol{\beta}\\&=2(\boldsymbol{\beta}^{\top}\Phi^{\top}\Phi-\mathbf{y}^{\top}\Phi+\lambda\boldsymbol{\beta}^{\top})d\boldsymbol{\beta}.\end{aligned}$$

因此，得到$\boldsymbol{\beta}$的最优解为

$$\hat{\boldsymbol{\beta}}_{r/s}=(\lambda\mathbf{I}+\Phi^{\top}\Phi)^{-1}\Phi^{\top}\mathbf{y}.$$

概率线性回归的最大后验估计

在高斯似然的模型中，通常使用高斯分布作为先验，这样得到的概率线性回归中参数的后验分布还是高斯分布。

一种简单常用的先验分布是

$$p(\boldsymbol{\beta} \mid \alpha) = \mathcal{N}(\boldsymbol{\beta} \mid \mathbf{0}, \alpha^{-1}\mathbf{I}).$$

根据贝叶斯公式可以得出参数的对数后验分布是

$$\ln p(\boldsymbol{\beta} \mid X, \mathbf{y}, \alpha, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - f(\mathbf{x}_i, \boldsymbol{\beta})\right)^2 - \frac{\alpha}{2}\boldsymbol{\beta}^{\top}\boldsymbol{\beta} + \text{const},$$

思考如何优化使用 L1范数进行正则化的最小二乘。

$$\frac{\partial \|\boldsymbol{\beta}\|_{L_1}}{\partial \boldsymbol{\beta}} = \text{sign}(\boldsymbol{\beta}) = \begin{cases} +1 & \beta_d > 0 \\ -1 & \beta_d < 0 \\ [-1,+1] & \beta_d = 0 \end{cases}$$

$$g(\boldsymbol{\beta}) \approx g(\boldsymbol{\beta}_k) + (\nabla g(\boldsymbol{\beta}_k))^\top (\boldsymbol{\beta} - \boldsymbol{\beta}_k) + \frac{a}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_k)^\top (\boldsymbol{\beta} - \boldsymbol{\beta}_k)$$

$$= \frac{a}{2}\left\| \boldsymbol{\beta} - \left(\boldsymbol{\beta}_k - \frac{1}{a}\nabla g(\boldsymbol{\beta}_k) \right) \right\|_2^2 + \text{const}$$

$$\boldsymbol{\beta}_{k+1} = \arg\min_{\boldsymbol{\beta}} \frac{a}{2}\left\| \boldsymbol{\beta} - \left(\boldsymbol{\beta}_k - \frac{1}{a}\nabla g(\boldsymbol{\beta}_k) \right) \right\|_2^2 + \lambda \|\boldsymbol{\beta}\|_{L_1}$$

$$\beta_{k+1}^d = \arg\min_{\beta^d} \frac{a}{2}\left(\beta^d - \left(\boldsymbol{\beta}_k - \frac{1}{a}\nabla g(\boldsymbol{\beta}_k) \right)^d \right)^2 + \lambda |\beta^d|$$

$$= \arg\min_{\beta^d} \frac{a}{2}\left(\beta^d - \left(\boldsymbol{\beta}_k - \frac{1}{a}\nabla g(\boldsymbol{\beta}_k) \right)^d \pm \frac{\lambda}{a} \right)^2 + \text{const}$$

$$= \begin{cases} \beta^d = \left(\boldsymbol{\beta}_k - \frac{1}{a}\nabla g(\boldsymbol{\beta}_k) \right)^d + \frac{\lambda}{a}, & \left(\boldsymbol{\beta}_k - \frac{1}{a}\nabla g(\boldsymbol{\beta}_k) \right)^d + \frac{\lambda}{a} < 0 \\ \beta^d = \left(\boldsymbol{\beta}_k - \frac{1}{a}\nabla g(\boldsymbol{\beta}_k) \right)^d - \frac{\lambda}{a}, & \left(\boldsymbol{\beta}_k - \frac{1}{a}\nabla g(\boldsymbol{\beta}_k) \right)^d - \frac{\lambda}{a} > 0 \\ \beta^d = 0, & -\frac{\lambda}{a} \le \left(\boldsymbol{\beta}_k - \frac{1}{a}\nabla g(\boldsymbol{\beta}_k) \right)^d \le \frac{\lambda}{a} \end{cases}$$

Python package怎么实现的?

torch.autograd?

目录

考虑一个标准的线性回归问题，对于$i = 1, \cdots, N$假设在给定自变量$\mathbf{x}_i$的情况下y_i如下产生

$$y_i = \mathbf{x}_i^{\top}\boldsymbol{\beta} + \epsilon_i,$$

其中$\boldsymbol{\beta}$是$D \times 1$维向量，ϵ_i是独立同分布的随机变量，并且$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. 定义$X = [x_1, x_2, \cdots, x_N]^{\mathrm{T}}$，$\mathbf{y} = [y_1, y_2, \cdots, y_N]^{\mathrm{T}}$，可以得到因变量$\mathbf{y}$的似然函数为

$$p(\mathbf{y} \mid X, \boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-\frac{N}{2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - X\boldsymbol{\beta})^{\top}(\mathbf{y} - X\boldsymbol{\beta})\right),$$

即 $\mathbf{y} \sim \mathcal{N}(X\boldsymbol{\beta}, \sigma^2\mathbf{I})$

如果先验分布和似然函数可以使得后验分布和先验分布具有相同的形式，那么就称先验分布与似然函数是**共轭**的，该先验叫做该似然函数的**共轭先验**。

给定模型的似然假设

$$p(\mathbf{y} \mid X, \boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-\frac{N}{2}} \exp\left(-\frac{1}{2\sigma^2} (\mathbf{y} - X\boldsymbol{\beta})^{\top} (\mathbf{y} - X\boldsymbol{\beta}) \right),$$

为了使得后验分布可以得到与先验分布相同的形式，这里假设参数$\boldsymbol{\beta}$和σ的联合先验为

$$p(\boldsymbol{\beta}, \sigma^2) = p(\sigma^2)\, p(\boldsymbol{\beta} \mid \sigma^2),$$

其中$p(\sigma^2)$是逆伽马分布$\text{Inv} - \text{Gamma}(a_0, b_0)$

$$p(\sigma^2) \propto (\sigma^2)^{-a_0 - 1} \exp(-\frac{b_0}{\sigma^2}),$$

而$p(\boldsymbol{\beta}|\sigma^2)$的条件先验密度服从正态分布$\mathcal{N}(\mu_0, \sigma^2 \Lambda_0^{-1})$，即

$$p(\boldsymbol{\beta} \mid \sigma^2) \propto (\sigma^2)^{-\frac{D}{2}} \exp(-\frac{1}{2\sigma^2} (\boldsymbol{\beta} - \boldsymbol{\mu}_0)^{\top} \Lambda_0 (\boldsymbol{\beta} - \boldsymbol{\mu}_0)).$$

给定β和σ的先验假设，根据贝叶斯公式，可以得到贝叶斯线性回归参数的后验分布为

$$p(\boldsymbol{\beta},\sigma^2\,|\,\mathbf{y},\mathbf{X}) = p(\boldsymbol{\beta}\,|\,\sigma^2,\mathbf{y},\mathbf{X})\,p(\sigma^2\,|\,\mathbf{y},\mathbf{X}) \propto p(\mathbf{y}\,|\,\mathbf{X},\boldsymbol{\beta},\sigma^2)\,p(\boldsymbol{\beta}\,|\,\sigma^2)\,p(\sigma^2).$$

由此可得，$p(\boldsymbol{\beta}\,|\,\sigma^2,\mathbf{y},\mathbf{X})$ 是高斯分布$\mathcal{N}(\boldsymbol{\beta}\,|\,\boldsymbol{\mu}_N,\sigma^2\Lambda_N^{-1})$，以及 $p(\sigma^2\,|\,\mathbf{y},\mathbf{X})$ 是逆伽马分布 Inv-Gamma$(\sigma^2\,|\,a_N,b_N)$，其参数具体表示如下：

$$\begin{cases} \Lambda_N = (\mathbf{X}^\top\mathbf{X} + \Lambda_0), \\ \boldsymbol{\mu}_N = (\Lambda_N)^{-1}(\mathbf{X}^\top\mathbf{y} + \Lambda_0\boldsymbol{\mu}_0), \\ a_N = a_0 + \dfrac{N}{2}, \\ b_N = b_0 + \dfrac{1}{2}(\mathbf{y}^\top\mathbf{y} + \boldsymbol{\mu}_0^\top\Lambda_0\boldsymbol{\mu}_0 - \boldsymbol{\mu}_N^\top\Lambda_N\boldsymbol{\mu}_N). \end{cases}$$

目录

逻辑回归使用逻辑函数和回归模型可以解决二类分类问题，其中逻辑函数的返回值用于表示二类分类问题中的正类或负类的概率。

假设f是自变量$\mathbf{x}$的一个线性函数，即$f = \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}$。逻辑回归假设样本$\mathbf{x}$属于正类的概率为

$$p(y=1|\mathbf{x}) = h_{\theta}(\mathbf{x}) = \sigma(\boldsymbol{\theta}^{\top}\mathbf{x}) = \frac{1}{1+\exp(-\boldsymbol{\theta}^{\top}\mathbf{x})},$$

那么，$\mathbf{x}$属于负类的概率为

$$p(y=0|\mathbf{x}) = 1 - p(y=1|\mathbf{x}) = 1 - h_{\theta}(\mathbf{x}) = \frac{1}{1+\exp(\boldsymbol{\theta}^{\top}\mathbf{x})}.$$

逻辑回归可以从两个角度定义目标函数：

最大似然角度

假设每一个样本的类标签是独立同分布的伯努利变量，伯努利变量取值为“1”和“0”的概率分别为

$$p(y=1|\mathbf{x})=h_{\theta}(\mathbf{x})=\sigma(\boldsymbol{\theta}^{\top}\mathbf{x})=\frac{1}{1+\exp(-\boldsymbol{\theta}^{\top}\mathbf{x})},$$

$$p(y=0|\mathbf{x})=1-p(y=1|\mathbf{x})=1-h_{\theta}(\mathbf{x})=\frac{1}{1+\exp(\boldsymbol{\theta}^{\top}\mathbf{x})}.$$

对于有二元标签的训练集$\{(\mathbf{x}_i,y_i):i=1,2,...,N\}$，$N$个独立样本的联合似然可以写成

$$p(\mathbf{y}|\boldsymbol{\theta})=\prod_{i=1}^{N}p(y_i=1|\mathbf{x}_i)^{y_i}(1-p(y_i=1|\mathbf{x}_i))^{(1-y_i)}.$$

最大化似然等价于最小化负对数似然，因此，最大似然得到的损失函数为

$$-\ln p(\mathbf{y}|\boldsymbol{\theta})=-\sum_{i=1}^{N}\left[y_i\ln p(y_i=1|\mathbf{x}_i)+(1-y_i)\ln(1-p(y_i=1|\mathbf{x}_i))\right].$$

构建损失函数角度

假设每个样本的真实分布为 $q(y_i|\mathbf{x}_i)$，那么 $q(y_i=1|\mathbf{x}_i)=y_i$，且 $q(y_i=0|\mathbf{x}_i)=1-y_i$。分布$q(y_i|\mathbf{x}_i)$和$p(y_i|\mathbf{x}_i)$的交叉熵是

$$H(q(y_i|\mathbf{x}_i),p(y_i|\mathbf{x}_i))=-\sum_{y_i}q(y_i|\mathbf{x}_i)\ln p(y_i|\mathbf{x}_i).$$

因此逻辑回归的交叉熵损失为

$$J(\boldsymbol{\theta})=\sum_{i=1}^{N}H(q(y_i|\mathbf{x}_i),p(y_i|\mathbf{x}_i))$$

$$=-\sum_{i=1}^{N}\left[y_i\ln h_{\boldsymbol{\theta}}(\mathbf{x}_i)+(1-y_i)\ln(1-h_{\boldsymbol{\theta}}(\mathbf{x}_i))\right].$$

注意分析损失函数的性质：分类器犯错时才能得到优化，与GAN饱和损失有关

无论从最大似然角度还是最小损失函数角度，二者得到的目标损失是一致的。可以通过最小化$\boldsymbol{J}(\boldsymbol{\theta})$来找到假设函数$h_{\boldsymbol{\theta}}(\mathbf{x})$中$\boldsymbol{\theta}$的最优值，从而学得分类器。使用梯度下降等方法优化$\boldsymbol{\theta}$需要计算$\boldsymbol{J}(\boldsymbol{\theta})$关于$\boldsymbol{\theta}$的梯度：

$$\nabla_{\theta}J(\boldsymbol{\theta})=\left(\frac{dJ(\boldsymbol{\theta})}{d\boldsymbol{\theta}}\right)^{\top}=\left(\frac{\sum_{i=1}^{N}\left((\sigma(\boldsymbol{\theta}^{\top}\mathbf{x}_i)-y_i)\mathbf{x}_i^{\top}\right)d\boldsymbol{\theta}}{d\boldsymbol{\theta}}\right)^{\top}=\sum_i\mathbf{x}_i(h_{\theta}(\mathbf{x}_i)-y_i),$$

- 多类逻辑回归

定义类别标签为$c \in \{1,2,\cdots,C\}$，每一个类别对应于一个回归函数

$$f_c(\mathbf{x}_i) = \boldsymbol{\theta}_c^{\top}\mathbf{x}_i,$$

其中$\boldsymbol{\theta}_c$是与类别c对应的回归系数，$\mathbf{x}_i$是第i个样本向量。经过softmax函数转换后得到样本属于某一类别的概率为

$$p(y_i = c) = \frac{\exp\{\boldsymbol{\theta}_c^{\top}\mathbf{x}_i\}}{\sum_{k=1}^{C}\exp\{\boldsymbol{\theta}_k^{\top}\mathbf{x}_i\}}.$$

多类逻辑回归的似然函数为

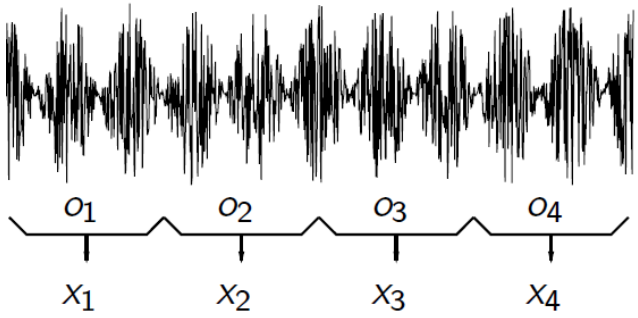
$$p(Y \mid \theta_1,\theta_2,\cdots,\theta_C) = \prod_{i=1}^{N}\prod_{c=1}^{C} p(y_i = c \mid \mathbf{x}_i)^{I(y_i=c)},$$

其中，$I(y_i = c)$在仅当$y_i = c$时函数值为1，其余为0。对应的负对数似然，也就是交叉熵损失为

$$-\ln p(Y \mid \theta_1,\theta_2,\ldots,\theta_K) = -\sum_{i=1}^{N}\sum_{c=1}^{C} I(y_i = c)\ln p(y_i = c \mid \mathbf{x}_i).$$

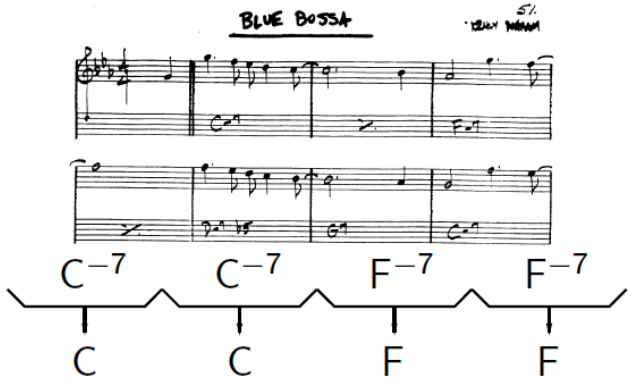
- 逻辑回归的最大熵解释
  - 取值区间为[0,1]的函数很多，为何选择sigmoid或者softmax呢？ （最大熵）
  - 以一个“从音符识别和弦”为例子讲解最大熵

Continuous features



real-valued: *e.g. MFCC, chroma, tempo...*

Discrete/categorical



symbolic: *e.g. note/chord played, key...*

Chords: ***C*7, *G*maj7, *F*min7, ...**

- 逻辑回归的最大熵解释（均匀选择未知）
  - Let's assume we observe the note **C**.

→ The "matching" chord is among $\{\mathbf{C}maj, \mathbf{C}min, \mathbf{A}\flat maj, \mathbf{A}min, \mathbf{F}maj, \mathbf{F}min\}$.

In terms of statistics

$$P(\mathbf{C}maj) + P(\mathbf{C}min) + P(\mathbf{A}\flat maj) + P(\mathbf{A}min) + P(\mathbf{F}maj) + P(\mathbf{F}min) = 1.$$

- How to choose $P(\mathbf{C}maj), \cdots, P(\mathbf{F}min)$?
- Safe choice:

In terms of statistics

$$P(\mathbf{C}maj) = P(\mathbf{C}min) = \cdots = P(\mathbf{F}min) = \frac{1}{6}$$

- 逻辑回归的最大熵解释（如果有一些观测呢）

  - *The matching chord is* **C***maj or* **F***maj 30% of the time:*

$$\begin{aligned} P(\mathbf{C}maj) + P(\mathbf{F}maj) &= 3/10 \\ P(\mathbf{C}maj) + P(\mathbf{C}min) + \cdots + P(\mathbf{F}maj) + P(\mathbf{F}min) &= 1 \end{aligned}$$

  - Again many solutions... and a resonable choice is:

$$\begin{aligned} P(\mathbf{C}maj) = P(\mathbf{F}maj) &= 3/20 \\ P(\mathbf{C}min) = P(\mathbf{A}\flat maj) = P(\mathbf{A}min) = P(\mathbf{F}min) &= 7/40 \end{aligned}$$

- 逻辑回归的最大熵解释（运用观测的统计信息）
  - 思考：通常情况下Cmaj 或 Fmaj的频率是3/10是怎么得到的？（通过观测频次统计得出的）
  - 定义特征函数

$$f_i(o_i, y_i) = \begin{cases} 1 & \text{if } o_i = \mathbf{C} \text{ and } y_i = \mathbf{C}maj \\ 0 & \text{otherwise} \end{cases}$$

- 逻辑回归的最大熵解释（运用观测的统计信息）
  - 描述数据和模型的概率分布，分别计算特征函数在两种分布下的期望
  - The training sample can be described in terms of its **empirical** probability distribution $\tilde{p}(o, y)$:

$$\tilde{p}(o, y) \triangleq \frac{1}{N} \times \text{number of times that } (o, y) \text{ occurs in the sample}$$

  - $\tilde{\mathbb{E}}(f_j) \triangleq \sum_{o,y} \tilde{p}(o, y) f_j(o, y)$: expected value of f_j w.r.t $\tilde{p}(o, y)$.
  - $\mathbb{E}(f_j) \triangleq \sum_{o,y} p(o)p(y|o)f_j(o, y)$: expected v. of f_j w.r.t the **model** $p(o, y)$.

- 逻辑回归的最大熵解释（最大熵+观测约束）
  - 希望模型的概率分布能够刻画数据的概率分布
  - 约束二者关于特征函数的期望相同

Constraint equation

$$\mathbb{E}(f_j) = \tilde{\mathbb{E}}(f_j),\ i.e.$$

$$\sum_{o,y} p(o)p(y|o)f_j(o,y) = \sum_{o,y} \tilde{p}(o,y)f_j(o,y)$$

- 逻辑回归的最大熵解释（运用观测的统计信息）
  - 寻找条件熵最大的条件分布
  - 熵最大的离散分布是均匀分布

Maximum entropy criterion

$$p^*(y|o) = \underset{p(y|o)\in\mathcal{M}}{\operatorname{argmax}} H(y|o);$$

$$H(y|o) \stackrel{\Delta}{=} -\sum_{o,y} p(o)p(y|o)\log p(y|o): \text{ the } \textbf{conditional entropy}$$

- 逻辑回归的最大熵解释
  - 求解带约束的优化问题

Primal: $p^*(y|o) = \text{argmax}_{p(y|o)\in\mathcal{M}} H(y|o)$

Constraints: $\mathbb{E}(f_j) = \tilde{\mathbb{E}}(f_j)$ and $\sum_y p(y|o) = 1$

Lagrangian: $L(p, \boldsymbol{\lambda}) \triangleq H(y|o) + \lambda_0 \left(\sum_y p(y|o) - 1\right) + \sum_j \lambda_j \left(\mathbb{E}(f_j) - \tilde{\mathbb{E}}(f_j)\right)$

Equating the derivative of the Lagrangian with 0:

$$p_{\boldsymbol{\lambda}}(y|o) = \frac{1}{Z_\lambda(o)} \exp \sum_j \lambda_j f_j(o, y);$$

$Z_{\boldsymbol{\lambda}}(x) = \sum_y \exp\left(\sum_j \lambda_j f_j(o, y)\right)$

The solution is given by the dual optimal: $\boldsymbol{\lambda}^* = \text{argmax}_{\boldsymbol{\lambda}} L(p, \boldsymbol{\lambda})$.

- 逻辑回归的最大熵解释
  - 求解带约束的优化问题

Primal: $p^*(y|o) = \text{argmax}_{p(y|o)\in\mathcal{M}} H(y|o)$

Constraints: $\mathbb{E}(f_j) = \tilde{\mathbb{E}}(f_j)$ and $\sum_y p(y|o) = 1$

Lagrangian: $L(p, \boldsymbol{\lambda}) \triangleq H(y|o) + \lambda_0 \left(\sum_y p(y|o) - 1\right) + \sum_j \lambda_j \left(\mathbb{E}(f_j) - \tilde{\mathbb{E}}(f_j)\right)$

Equating the derivative of the Lagrangian with 0:

$$p_{\boldsymbol{\lambda}}(y|o) = \frac{1}{Z_\lambda(o)} \exp \sum_j \lambda_j f_j(o, y);$$

$Z_{\boldsymbol{\lambda}}(x) = \sum_y \exp\left(\sum_j \lambda_j f_j(o, y)\right)$

The solution is given by the dual optimal: $\boldsymbol{\lambda}^* = \text{argmax}_{\boldsymbol{\lambda}} L(p, \boldsymbol{\lambda})$.

- 逻辑回归的最大熵解释
  - 得到解析解

Maxent model:

$$p(y=k|o)=\frac{1}{Z_\lambda(o)}\exp\left(\sum_j \lambda_{jk} f_j(o,y)\right);$$

$$Z_\lambda(o)=\sum_y \exp\left(\sum_j \lambda_{jk} f_j(o,y)\right).$$

Logistic regression model:

$$\begin{aligned} p(y=k|\mathbf{x}) &= \frac{\exp(w_{k0}+\mathbf{w}_k^T\mathbf{x})}{1+\sum_{l=1}^{K-1}\exp(w_{l0}+\mathbf{w}_l^T\mathbf{x})} \\ &= \frac{\exp(w'_{k0}+\mathbf{w}_k'^T\mathbf{x})}{\sum_{l=1}^{K}\exp(w'_{l0}+\mathbf{w}_l'^T\mathbf{x})} \\ &= \frac{1}{Z_{\mathbf{w}}(\mathbf{x})}\exp(w'_{k0}+\mathbf{w}_k'^T\mathbf{x}). \end{aligned}$$

- 逻辑回归的最大熵解释
  - 对照逻辑回归表达式

Maxent model:

$$p(y=k|o)=\frac{1}{Z_{\lambda}(o)}\exp\sum_{j}\lambda_{kj}f_j(o,y)$$

Logistic regression model:

$$p(y=k|\mathbf{x})=\frac{1}{Z_{\mathbf{w}}(\mathbf{x})}\exp(w'_{k0}+\mathbf{w}_k^{'T}\mathbf{x})$$

Using:

- feature-function: $f_j(o,y)=x_j$; $f_0(o,y)=1$ and $\mathbf{x}=(x_1,\cdots,x_j,\cdots,x_D)^T$;
- $w'_{k0}+\mathbf{w}_k^{'T}\mathbf{x}=\sum_{j=0}^{D}w'_{kj}f_j(o,y)$;

- 逻辑回归的最大熵解释
  - 对照逻辑回归表达式

Maxent model:

$$p(y=k|o)=\frac{1}{Z_\lambda(o)}\exp\sum_j \lambda_{kj}f_j(o,y)$$

Logistic regression model:

- When $K=2$

$$\begin{aligned} P(\mathcal{C}_1|\mathbf{x}) &= p=\frac{1}{1+\exp-(w_{10}+\mathbf{w}_1^T\mathbf{x})} \\ P(\mathcal{C}_2|\mathbf{x}) &= 1-p \end{aligned}$$

Using:

- feature-function: $f_j(o,y)=x_j$; $f_0(o,y)=1$ and $\mathbf{x}=(x_1,\cdots,x_j,\cdots,x_D)^T$;
- $w'_{k0}+\mathbf{w}_k^{'T}\mathbf{x}=\sum_{j=0}^{D} w'_{kj}f_j(o,y)$;

目录

已知观测数据$X=[\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N]^{\mathrm{T}}$，$\mathbf{y}=[y_1,y_2,\cdots,y_N]^{\mathrm{T}}$，逻辑回归假设的似然概率使得后验分布$p(\boldsymbol{\theta}|X,\mathbf{y})$难以有解析表达，因此通常使用其他典型分布 $q(\boldsymbol{\theta})$ 来近似后验分布。在预测时，即便使用了近似分布，对新样本 $\mathbf{x}_*$ 的预测分布 $p(y_*=1|\mathbf{x}_*)\approx\int\sigma(\boldsymbol{\theta}^\top\mathbf{x}_*)q(\boldsymbol{\theta})d\boldsymbol{\theta}$ 的估计仍然是难解。

- 一方面，后验分布$p(\boldsymbol{\theta}|y)$等于先验乘以似然再进行归一化。其中先验通常假设为 $p(\boldsymbol{\theta})=\mathcal{N}(\boldsymbol{\theta}|\mathbf{m}_0,S_0)$. 逻辑回归的似然为

$$p(\mathbf{y}\mid X,\boldsymbol{\theta})=\prod_{i=1}^{N}p(y_i=1|\mathbf{x}_i)^{y_i}\left(1-p(y_i=1|\mathbf{x}_i)\right)^{1-y_i}.$$

- 另一方面，预测分布 $p(y_*=1|\mathbf{x}_*)\approx\int\sigma(\boldsymbol{\theta}^\top\mathbf{x}_*)q(\boldsymbol{\theta})d\boldsymbol{\theta}$ 需要关于sigmoid函数和高斯分布的乘积求积分，其精确求解也十分困难，可通过将sigmoid函数用逆probit函数近似得到其近似解。

拉普拉斯近似

对后验分布的拉普拉斯近似是通过数值优化算法得到一个以$\boldsymbol{\theta}_0$为均值的高斯分布$q(\boldsymbol{\theta})$，作为真实后验的近似分布：

$$q(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{D/2}\,|S_N|^{1/2}} \exp\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top S_N^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\} = \mathcal{N}(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}_0, S_N).$$

其中，均值$\boldsymbol{\theta}_0$是真实后验分布的最大值对应的参数，协方差矩阵是负对数真实后验分布$-\ln p(\boldsymbol{\theta}\,|\,X,\mathbf{y})$的Hessian矩阵（附录C）在$\boldsymbol{\theta}=\boldsymbol{\theta}_0$处的逆，即$S_N = \left(-\nabla\nabla \ln p(\boldsymbol{\theta}\,|\,X,\mathbf{y})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right)^{-1}$。

$$\frac{df(\mathbf{z})}{d\mathbf{z}}\Big|_{\mathbf{z}=\mathbf{z}_0} = 0,$$

$$\ln f(\mathbf{z}) = \ln f(\mathbf{z}_0) - \frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^\top A(\mathbf{z}-\mathbf{z}_0) + R_3,$$

$$A = -\frac{d^2 \ln f(\mathbf{z})}{d\mathbf{z}^2}\Big|_{\mathbf{z}=\mathbf{z}_0}.$$

$$f(\mathbf{z}) \approx f(\mathbf{z}_0)\exp\{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^\top A(\mathbf{z}-\mathbf{z}_0)\}.$$

$$q(\mathbf{z}) = \frac{|A|^{\frac{1}{2}}}{2\pi^{\frac{M}{2}}}\exp\{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^\top A(\mathbf{z}-\mathbf{z}_0)\},$$

均值$\boldsymbol{\theta}_0$和协方差矩阵S_N的具体计算过程

已知参数服从高斯先验$p(\boldsymbol{\theta})=\mathcal{N}(\boldsymbol{\theta}\,|\,m_0,S_0)$，其中$m_0$和$S_0$是超参数。后验分布$p(\boldsymbol{\theta}\,|\,X,\mathbf{y})\propto p(\boldsymbol{\theta})p(\mathbf{y}\,|\,X,\boldsymbol{\theta})$。将先验概率$p(\boldsymbol{\theta})=\mathcal{N}(\boldsymbol{\theta}\,|\,\mathbf{m}_0,S_0)$和逻辑回归的似然函数

$$p(\mathbf{y}\,|\,X,\boldsymbol{\theta})=\prod_{i=1}^{N}p(y_i=1\,|\,\mathbf{x}_i)^{y_i}\left(1-p(y_i=1\,|\,\mathbf{x}_i)\right)^{1-y_i}.$$

带入贝叶斯公式可得

$$\ln p(\boldsymbol{\theta}\,|\,X,\mathbf{y})=-\frac{1}{2}(\boldsymbol{\theta}-m_0)^{\top}S_0^{-1}(\boldsymbol{\theta}-m_0)$$
$$+\sum_{i=1}^{N}\left[y_i\ln p(y_i=1\,|\,\mathbf{x}_i,\boldsymbol{\theta})+(1-y_i)\ln(1-p(y_i=1\,|\,\mathbf{x}_i,\boldsymbol{\theta}))\right]+\text{const.}$$

最大化该对数后验分布$\ln p(\boldsymbol{\theta}\mid X,\mathbf{y})$可以得到参数的最大后验估计 $\boldsymbol{\theta}_{map}$ ，作为近似分布$q(\boldsymbol{\theta})$的均值。$-\ln p(\boldsymbol{\theta}\mid X,\mathbf{y})$的Hessian矩阵计算如下：

$$
\begin{aligned}
H &= -\nabla\nabla \ln p(\boldsymbol{\theta}\mid X,\mathbf{y}) = \frac{d^2 \ln p(\boldsymbol{\theta}\mid X,\mathbf{y})}{d\boldsymbol{\theta} d\boldsymbol{\theta}^{\top}} \\
&= \frac{d\mathrm{Tr}[(\boldsymbol{\theta}-\mathbf{m}_0)^{\top} S_0^{-1} d\boldsymbol{\theta}] - \left(d\sum_{i=1}^{N}\left((y_i-\sigma(\boldsymbol{\theta}^{\top}\mathbf{x}_i))\mathbf{x}_i^{\top} d\boldsymbol{\theta}\right)\right)}{d\boldsymbol{\theta} d\boldsymbol{\theta}^{\top}} \\
&= \frac{\mathrm{Tr}[S_0^{-1} d\boldsymbol{\theta} d\boldsymbol{\theta}^{\top}] + \mathrm{Tr}\left[\sum_{i=1}^{N}\sigma(\boldsymbol{\theta}^{\top}\mathbf{x}_i)(1-\sigma(\boldsymbol{\theta}^{\top}\mathbf{x}_i))\mathbf{x}_i\mathbf{x}_i^{\top} d\boldsymbol{\theta} d\boldsymbol{\theta}^{\top}\right]}{d\boldsymbol{\theta} d\boldsymbol{\theta}^{\top}} \\
&= S_0^{-1} + \sum_{i=1}^{N} p(y_i=1\mid \mathbf{x}_i,\boldsymbol{\theta})(1-p(y_i=1\mid \mathbf{x}_i,\boldsymbol{\theta}))\mathbf{x}_i\mathbf{x}_i^{\top}.
\end{aligned}
$$

逆probit函数近似

在得到近似后验分布后，对于给定的新特征向量$\mathbf{x}_*$，其属于类别“1”的预测分布可以通过似然关于后验$p(\boldsymbol{\theta}|X,\mathbf{y})$的积分得到，即

$$\begin{aligned} p(y_*=1|\mathbf{x}_*) &= \int p(y_*=1,\boldsymbol{\theta}|\mathbf{x}_*)d\boldsymbol{\theta} \\ &= \int p(y_*=1|\mathbf{x}_*,\boldsymbol{\theta})\,p(\boldsymbol{\theta}|X,\mathbf{y})d\boldsymbol{\theta} \\ &\approx \int \sigma(\boldsymbol{\theta}^\top\mathbf{x}_*)q(\boldsymbol{\theta})d\boldsymbol{\theta}. \end{aligned}$$

由于函数$\sigma(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}_*)$仅通过$\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}_*$的值依赖于$\boldsymbol{\theta}$，因此定义新的变量$a=\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}_*$，并引入Dirac delta函数$\delta(\cdot)$，得到$\sigma(\boldsymbol{\theta}^\top\mathbf{x}_*)\approx\int\delta(a-\sigma(\boldsymbol{\theta}^\top\mathbf{x}_*))\sigma(a)da$

$$\begin{aligned} \int\sigma(\boldsymbol{\theta}^\top\mathbf{x}_*)q(\boldsymbol{\theta})d\boldsymbol{\theta} &= \int\left(\int\delta(a-\sigma(\boldsymbol{\theta}^\top\mathbf{x}_*))\sigma(a)da\right)q(\boldsymbol{\theta})d\boldsymbol{\theta} \\ &= \int\sigma(a)\int\delta(a-\sigma(\boldsymbol{\theta}^\top\mathbf{x}_*))q(\boldsymbol{\theta})d\boldsymbol{\theta}da, \end{aligned}$$

其中，$\int\delta(a-\sigma(\boldsymbol{\theta}^\top\mathbf{x}_*))q(\boldsymbol{\theta})d\boldsymbol{\theta}$是关于$a$的函数，并且可验证为是一个高斯概率分布，记为$p(a)=\mathcal{N}(a|\mu_a,\sigma_a^2)$，其中均值和方差分别为

$$\mu_a=\mathbb{E}[a]=\int p(a)da=\int q(\boldsymbol{\theta})\boldsymbol{\theta}^\top\mathbf{x}_*d\boldsymbol{\theta}=\boldsymbol{\theta}_{map}^\top\mathbf{x}_*,$$

$$\sigma_a^2=var[a]=\int p(a)a^2da-\mathbb{E}[a]^2=\int q(\boldsymbol{\theta})(\boldsymbol{\theta}^\top\mathbf{x}_*)^2d\boldsymbol{\theta}-(\boldsymbol{\theta}_{map}^\top\mathbf{x}_*)^2=\mathbf{x}_*^\top S_N\mathbf{x}_*.$$

预测分布可以表示为

$$p(y=1|\mathbf{x}_*)=\int\sigma(a)p(a)da=\int\sigma(a)\mathcal{N}(a|\mu_a,\sigma_a^2)da.$$

上式关于sigmoid和Gaussian的积分是不可解的，通常使用逆probit函数来替代sigmoid函数。定义标准高斯分布的累积分布函数，即逆probit函数为

$$\Phi(a)=\int_{-\infty}^{a}\mathcal{N}(w|0,1)dw.$$

高斯分布和逆probit函数相乘后的积分还是一个逆probit函数，即

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

由此可获得最终预测概率为

$$p(y=1|\mathbf{x}_*)=\int\sigma(a)\mathcal{N}(a|\mu_a,\sigma_a^2)da\approx\sigma\left(\mu_a/\left(1+\pi\sigma_a^2/8\right)^{1/2}\right),$$

其中$\mu_a=\theta_{\text{map}}^{\mathrm{T}}\mathbf{x}_*$，$\sigma_a^2=\mathbf{x}_*^{\mathrm{T}}S_N\mathbf{x}_*$。对应于$p(y=1|\mathbf{x}_*)=0.5$的决策边界由$\mu_a=0$给出。

1. 王雄清, 刘安全, 陈仁武. 圈养大熊猫全年食竹量观察[J]. 四川动物, 1989, 8(4): 18-18.

2. Draper N R, Nostrand R C V. Ridge Regression and James-Stein Estimation: Review and Comments[J]. Technometrics, 1979, 21(4): 451–466.

3. Cox D R. The Regression Analysis of Binary Sequences[J]. Journal of the Royal Statistical Society. Series B (Methodological), 1958, 20(2): 215–242.

4. Bishop C M. Pattern Recognition and Machine Learning[M]. New York: Springer, 2006.