第3章　逻辑回归

学习目标：

（1）掌握线性回归及其模型求解方法；

（2）理解贝叶斯线性回归；

（3）掌握逻辑回归及其模型求解方法；

（4）理解贝叶斯逻辑回归。

分类与回归是机器学习中两类基本的任务，二者都属于监督学习的范畴，都是通过训练有标注的样本来学习输入与输出之间的关系，进而预测新样本的输出。分类与回归任务可以简单地通过数据输出的类型进行区分：通常情况下，当数据的输出取值是有限个离散值时，该任务是分类任务，例如当输出数据的取值为1或0时，该任务是一个二分类任务；当数据的输出取值是连续值时，该任务是回归任务。一方面，我们熟知的图像分类可以是二分类任务，也可以是多类分类任务。例如，判断一张图片是否是大熊猫是一个二类分类问题，判断一张图片是大熊猫、小熊猫还是棕熊是一个三类分类问题。另一方面，利用某个事物的已知特征来预测其具有连续取值的其他信息是回归问题。例如，通过一些特征，包括熊猫的年龄、性别、体重等，来预测熊猫的食量是回归问题。

从字面来看，逻辑回归似乎是一种回归方法，但事实上是分类方法。传统的逻辑回归用于处理二分类问题，通过进一步引入softmax函数，也可以处理多类分类问题。逻辑回归的名字主要是因为它从线性回归转变而来，并且在线性回归中引入sigmoid函数（逻辑函数），实现对输出变量的非线性转换。逻辑回归可以给出样本属于某一类别的概率，该概率值可以被用于分类决策。例如，逻辑回归预测一张图片是大熊猫的概率为0.7，如果使用0.5作为

分类阈值，那么该图片就被判定为大熊猫。

由于逻辑回归与线性回归有着密切联系，下面首先介绍线性回归的相关方法，包括线性回归的确定性模型与概率模型，然后介绍逻辑回归的原理及贝叶斯逻辑回归。

3.1　线 性 回 归

线性回归通过拟合关于观测数据的线性方程来建模两个变量间的关系，其中一个变量是自变量，另一个是因变量，自变量与因变量都可以是多元变量。多元自变量对应事物的多个输入特征，多元因变量对应事物的多种输出信息。例如，可以通过一个线性回归模型来关联熊猫的性别、年龄、体重和食量，其中性别、年龄和体重作为自变量，食量作为因变量。

下面给出线性回归的原理与表示。给定有 N 个样本的数据集 $\mathcal{D}=\{y_i,x_{i1},x_{i2},\cdots,x_{iD}\}_{i=1}^{N}$，线性回归模型假设因变量 y_i 与自变量 $\boldsymbol{x}_i$（由 $\{x_{i1},x_{i2},\cdots,x_{iD}\}$ 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，模型的形式为：

$$y_i=\beta_0 1+\beta_1 x_{i1}+\cdots+\beta_D x_{iD}=\boldsymbol{x}_i^{\top}\boldsymbol{\beta},\quad i=1,2,\cdots,N \tag{3.1}$$

其中 $\boldsymbol{x}_i^{\top}\boldsymbol{\beta}$ 表示向量 $\boldsymbol{x}_i$ 和向量 $\boldsymbol{\beta}$ 之间的内积。通常我们会把一个常数 1 包含在自变量里，以得到简洁表示，也就是说，可以设 $x_{i0}=1,i=1,2,\cdots,N$。对应的 β_0 被称为截距。当 $D=1$，即 $\boldsymbol{x}_i$ 是标量且 y_i 也为简单的标量时，模型称为简单线性回归；当自变量 $\boldsymbol{x}_i$ 是向量时，模型称为多元线性回归。

另外，值得注意的是，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。只要模型关于参数向量 $\boldsymbol{\beta}$ 是线性的，模型就被认为是线性模型；如果模型关于参数是非线性的，则被认为是非线性模型。假设 $\phi(\boldsymbol{x}_i)$ 表示对输入特征的变换函数，也称为基函数，那么线性回归可以表示为更一般化的形式，即

$$y_i=\phi(\boldsymbol{x}_i)^{\top}\boldsymbol{\beta} \tag{3.2}$$

常见的基函数有 3 种，即

① 多项式基函数。

$$\phi_j(x)=x^j \tag{3.3}$$

② 高斯基函数。

$$\phi_j(x)=\exp\left[-\frac{(x-\mu_j)^2}{2s^2}\right] \tag{3.4}$$

③ S 形(sigmoidal)基函数。

$$\phi_j(x)=\sigma\left(\frac{x-\mu_j}{s}\right),\quad \sigma(a)=\frac{1}{1+\exp(-a)} \tag{3.5}$$

3 种基函数的示意图如图 3-1 所示。

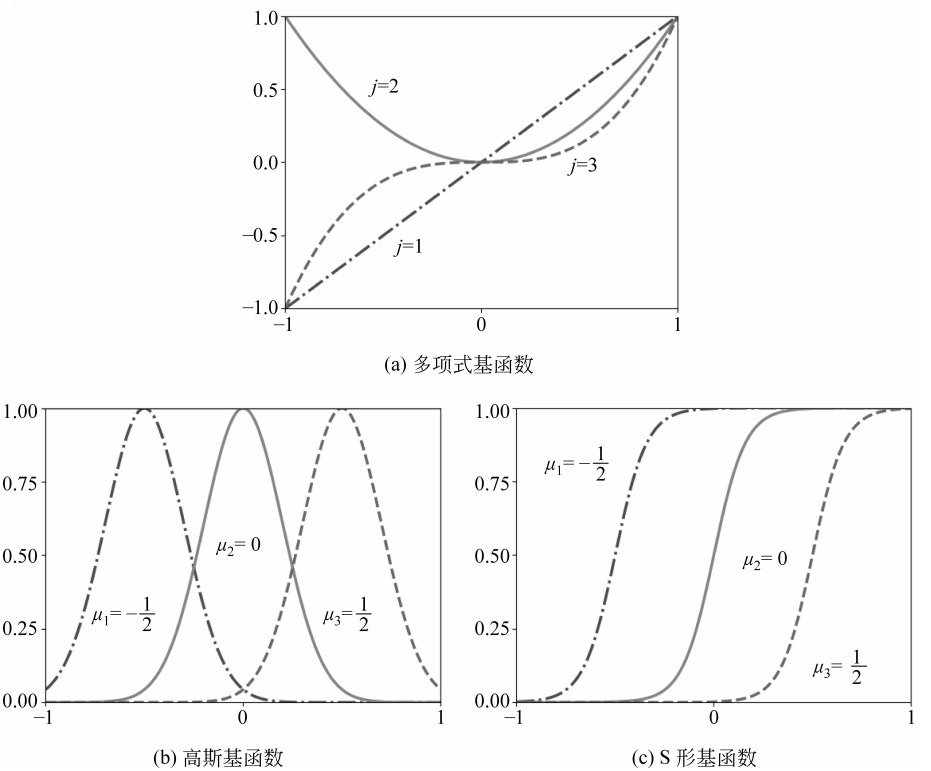


图 3-1　3 种基函数示意图

注：每张子图展示了使用不同参数产生的 3 条基函数曲线

得到回归模型后，可以通过公式(3.2)对新的测试数据进行预测，那么测试输入 $\boldsymbol{x}_*$ 的预测输出 y_* 表示为 $y_* = \phi(\boldsymbol{x}_*)^\top \boldsymbol{\beta}$。为得到该回归函数中参数的最优值，本节将介绍两种对线性回归模型的训练方法：最小二乘和正则化最小二乘。

【示例】 下面通过估计熊猫食量的例子介绍如何使用线性回归建模数据。一篇关于圈养大熊猫食竹量观察的文献记录了 4 只大熊猫的夜间食竹量，如表 3-1 所示。

表 3-1　大熊猫平均夜间食竹量[1]　　(单位：kg)

熊猫名	性别	年龄/岁	体重	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

从表 3-1 中可以看出，食竹量与熊猫的性别、年龄、体重及月份都有关系，并且从数据可以简单地分析得到食竹量与月份有两段不同的规律。因此可以以 7 月为界限，分两段进行线性回归，每一段有 24 个训练样本。在该示例中，自变量是一个四维向量 $\boldsymbol{x}=(x_1,x_2,x_3,x_4)$，因变量是一个标量 y。假定用线性模型建模，即

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} \tag{3.6}$$

如果要进一步增强模型的灵活性，可以对某些自变量先进行非线性变换，得到新的自变量 $\tilde{\boldsymbol{x}}$，再进行线性回归建模。此时，虽然模型关于某些变量是非线性的，但是关于参数还是线性的。读者可以尝试对上述示例中的输入特征设置合适的非线性变换，然后进行线性回归建模。

正如前文中介绍的，使用概率模型是建模不确定性的有效方法。概率线性回归的一种实现方式是使用高斯随机噪声实现概率建模。具体来说，观测输出被假设为确定性的线性回归再加上一个高斯随机噪声，表示为

$$y = f(\boldsymbol{x}, \boldsymbol{\beta}) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{3.7}$$

其中

$$f(\boldsymbol{x}, \boldsymbol{\beta}) = \phi(\boldsymbol{x})^{\top} \boldsymbol{\beta} \tag{3.8}$$

根据概率分布的变换关系，可以得到每个观测数据的似然概率分布为

$$p(y \mid \boldsymbol{x}, \boldsymbol{\beta}, \sigma^2) = \mathcal{N}(y \mid f(\boldsymbol{x}, \boldsymbol{\beta}), \sigma^2) \tag{3.9}$$

处理实际问题时，模型通常假设数据是独立同分布的，所有观测 $\boldsymbol{y}$ 的似然概率分布表示为

$$p(\boldsymbol{y} \mid X, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{N} \mathcal{N}(y_i \mid f(\boldsymbol{x}_i, \boldsymbol{\beta}), \sigma^2) \tag{3.10}$$

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，计算表达公式为

$$\mathbb{E}[y \mid \boldsymbol{x}_*] = \int y p(y \mid \boldsymbol{x}_*, \boldsymbol{\beta}, \sigma^2) \mathrm{d}y = f(\boldsymbol{x}_*, \boldsymbol{\beta}) \tag{3.11}$$

关于如何确定模型中的参数值，下面将介绍对概率线性回归模型的最大似然估计和最大后验估计，并说明二者分别与最小二乘和正则化最小二乘之间的关系。

3.1.1 最小二乘与最大似然

最小二乘法(least square method)中“最小二乘”的意思是最小化误差的平方和，误差是指观测数据的真实输出值和由模型拟合的因变量值之间的差。下面分别给出最小二乘问题的描述，如何求解最小二乘问题，以及概率线性回归的最大似然估计。

(1) 最小二乘问题描述。

给定有 N 个数据点$(\boldsymbol{x}_i, y_i)$的数据集，其中 $\boldsymbol{x}_i$ 为自变量，y_i 为因变量。模型函数具有形式 $f(\boldsymbol{x}_i, \boldsymbol{\beta})$，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，即

$$e_i = y_i - f(\boldsymbol{x}_i, \boldsymbol{\beta}) \tag{3.12}$$

以曲线拟合为例，误差的几何意义如图 3-2 所示。最小二乘法通过最小化平方误差和 S 学习最优参数值，即

$$S = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - f(\boldsymbol{x}_i, \boldsymbol{\beta}))^2 \tag{3.13}$$

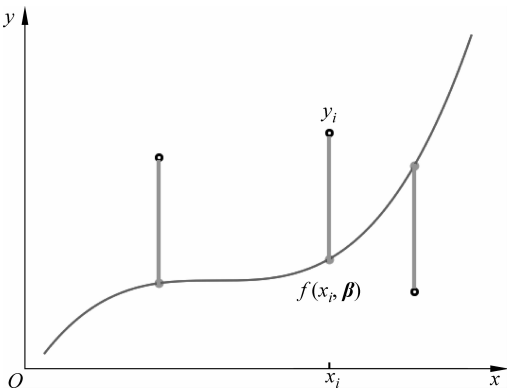


图 3-2　误差的几何意义示意图

注：图中纵向线段长度代表不同数据点的误差

(2) 求解最小二乘问题。

上述平方和的最小化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，就有 D 个梯度方程，即

$$\frac{\partial S}{\partial \beta_d} = 0, \quad d = 1, 2, \cdots, D \tag{3.14}$$

代入公式(3.13)可得

$$-2\sum_{i=1}^{N} (y_i - f(\boldsymbol{x}_i, \boldsymbol{\beta})) \frac{\partial f(\boldsymbol{x}_i, \boldsymbol{\beta})}{\partial \beta_d} = 0, \quad d = 1, 2, \cdots, D \tag{3.15}$$

公式(3.15)中的梯度方程适用于所有最小二乘问题。每个具体问题有特定的模型表达式和相应的偏导数。当然，求解公式(3.13)给出的最小平方误差和

也可以直接使用向量微积分的方法，直接对优化目标关于参数向量求导解得。

下面以线性回归问题为例，具体介绍最小二乘法的解。由公式(3.2)可知，一般化的线性回归模型表示为 $f(\boldsymbol{x}_i,\boldsymbol{\beta})=\phi(\boldsymbol{x}_i)^{\top}\boldsymbol{\beta}$。定义 $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N]^{\top}$，$\boldsymbol{y}=[y_1,y_2,\cdots,y_N]^{\top}$，$\boldsymbol{\Phi}=[\phi(\boldsymbol{x}_1),\phi(\boldsymbol{x}_2),\cdots,\phi(\boldsymbol{x}_N)]^{\top}$，那么模型在训练数据上的预测平方误差为

$$S=(\boldsymbol{y}-\boldsymbol{\Phi\beta})^{\top}(\boldsymbol{y}-\boldsymbol{\Phi\beta}) \tag{3.16}$$

根据公式(3.14)可以得到 $\boldsymbol{\beta}$ 的最优值满足

$$\frac{\mathrm{d}S}{\mathrm{d}\boldsymbol{\beta}}=\frac{\mathrm{d}((\boldsymbol{y}-\boldsymbol{\Phi\beta})^{\top}(\boldsymbol{y}-\boldsymbol{\Phi\beta}))}{d\boldsymbol{\beta}}=\boldsymbol{0}^{\top} \tag{3.17}$$

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

$$\begin{aligned}\mathrm{d}((\boldsymbol{y}-\boldsymbol{\Phi\beta})^{\top}(\boldsymbol{y}-\boldsymbol{\Phi\beta}))&=(\mathrm{d}(\boldsymbol{y}-\boldsymbol{\Phi\beta})^{\top})(\boldsymbol{y}-\boldsymbol{\Phi\beta})+(\boldsymbol{y}-\boldsymbol{\Phi\beta})^{\top}\mathrm{d}(\boldsymbol{y}-\boldsymbol{\Phi\beta})\\&=2(\boldsymbol{y}-\boldsymbol{\Phi\beta})^{\top}\mathrm{d}(\boldsymbol{y}-\boldsymbol{\Phi\beta})\\&=-2(\boldsymbol{y}-\boldsymbol{\Phi\beta})^{\top}\boldsymbol{\Phi}\mathrm{d}\boldsymbol{\beta}\\&=2(\boldsymbol{\beta}^{\top}\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}-\boldsymbol{y}^{\top}\boldsymbol{\Phi})\mathrm{d}\boldsymbol{\beta}\end{aligned} \tag{3.18}$$

因此，得到 $\boldsymbol{\beta}$ 的最优值为

$$\hat{\boldsymbol{\beta}}_{ls}=(\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\top}\boldsymbol{y} \tag{3.19}$$

确定性线性回归的优化准则通常使用损失函数来定义，最小二乘方法使用的是最小化平方误差和。概率线性回归的优化准则通常以最大似然为目标，这里给出其最大似然解，并说明与最小二乘之间的关系。

(3) 概率线性回归的最大似然估计。

当概率线性回归的似然假设为高斯分布时，如公式(3.10)，其对数似然的表达式可以进一步推导得出

$$\ln p(\boldsymbol{y}\mid X,\boldsymbol{\beta},\sigma^2)=-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-f(\boldsymbol{x}_i,\boldsymbol{\beta}))^2-\frac{N}{2}\ln\sigma^2-\frac{N}{2}\ln(2\pi) \tag{3.20}$$

最大化公式(3.20)可以获得参数 $\boldsymbol{\beta}$ 和 σ^2 的最大似然估计。二者的结果为

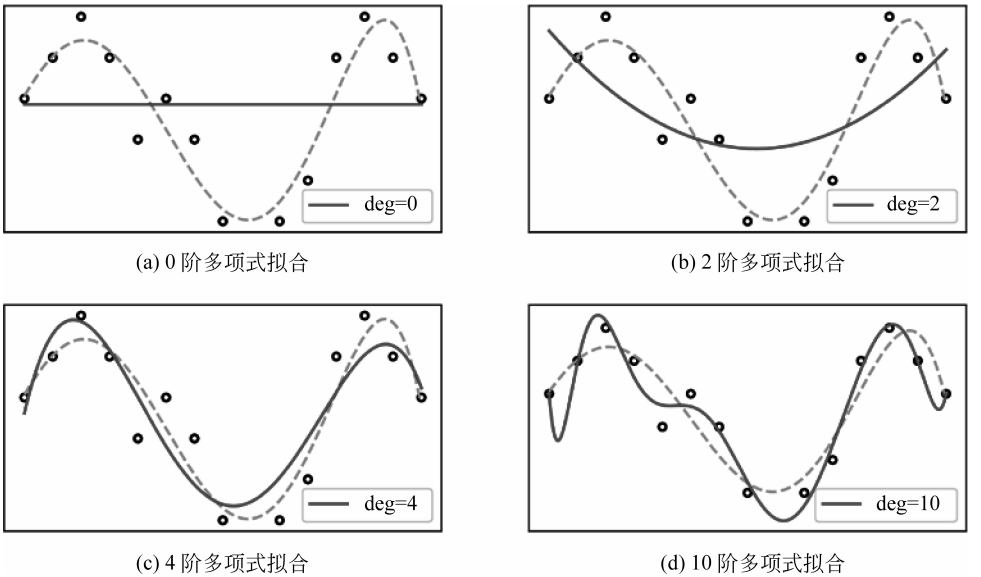
$$\hat{\boldsymbol{\beta}}_{\mathrm{m}\ell}=(\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\top}\boldsymbol{y} \tag{3.21}$$

$$\hat{\sigma}^{2}_{\mathrm{m}\ell}=\frac{1}{N}\sum_{i=1}^{N}(y_i-f(\boldsymbol{x}_i,\hat{\boldsymbol{\beta}}_{\mathrm{m}\ell}))^{2} \tag{3.22}$$

因此，可以看出当观测数据服从高斯分布时，线性回归参数 $\boldsymbol{\beta}$ 的最小二乘解和最大似然估计是等价的。

3.1.2　正则化最小二乘与最大后验

回归模型经常遇到数据过拟合(overfitting)问题，也就是模型在训练集上的拟合误差很小，但是在测试集上的误差很大。过拟合通常发生在数据量较少或模型的复杂度太高时。例如，在线性回归中，对数据特征引入多项式变换，回归系数的数量越多，则会导致曲线的波动越大，此时曲线容易对数据产生过拟合。图 3-3 给出了 4 种多项式拟合的效果。



(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　4 种不同的多项式拟合效果

注：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 $f(x)=\sum_{j=0}^{\mathrm{deg}}w_j x^j$，deg 表示多项式的阶数，4 张子图分别使用不同的阶数

通常情况下，模型的复杂度是相对的，它与数据量相关。如果训练数据足够多，高复杂度的模型可以很好地拟合数据；如果数据量较少，就需要一个相对简单的模型来拟合数据。在实际应用中，数据量的多少通常是无法改变的，建模者可以控制的是模型的设置，希望可以通过某种约束实现对数据较为合适的拟合。对于使用多项式变换的线性回归的例子，回归系数是关键参数，如果固定多项式的次数，控制回归系数的大小同样可以控制模型复杂度。

（1）正则化最小二乘。

由于回归系数越大模型波动越大，为了降低过拟合的风险，可以对回归系数进行约束。对最小二乘进行正则化的方法叫作正则化最小二乘。例如，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

$$S'=\sum_{i=1}^{N}(y_i-f(\boldsymbol{x}_i,\boldsymbol{\beta}))^2+\lambda\boldsymbol{\beta}^{\top}\boldsymbol{\beta} \tag{3.23}$$

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称为岭回归[2]。

正则化最小二乘不仅限于 L_2 范数，其他如 L_1 范数（$\|\boldsymbol{\beta}\|_{L_1}=\sum_d|\beta_d|$）等也是可行的，不同的正则化项具有不同的约束性质。例如，L_2 范数的惩罚项可以帮助模型避免过拟合，L_1 范数的惩罚项除了使得模型减轻过拟合以外，还能够得到较为稀疏的参数解。为了直观理解两种正则化方法，图 3-4 展示了这两种正则化项的等高线。

（2）求解正则化最小二乘问题。

求解正则化最小二乘与求解最小二乘是类似的，同样可以使用求导数的方法得到参数的闭式解。对于使用 L_2 范数的正则化最小二乘，其最优解满足

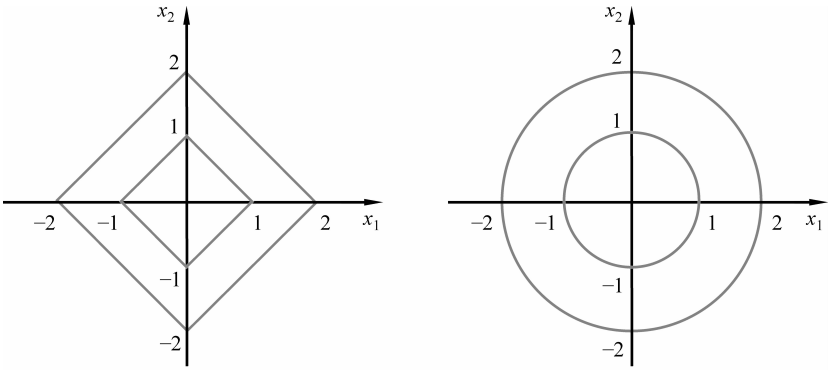


图 3-4　L_1 范数和 L_2 范数的等高线示意图

注：图中的曲线表示二维空间中的向量 $\boldsymbol{x}=[x_1,x_2]^\top$ 的 L_1 范数 $\|\boldsymbol{x}\|_{L_1}$ 和 L_2 范数 $\|\boldsymbol{x}\|_{L_2}$ 的等高线

$$\frac{\mathrm{d}S'}{\mathrm{d}\boldsymbol{\beta}}=\frac{\mathrm{d}((\boldsymbol{y}-\boldsymbol{\Phi\beta})^\top(\boldsymbol{y}-\boldsymbol{\Phi\beta})+\lambda\boldsymbol{\beta}^\top\boldsymbol{\beta})}{\mathrm{d}\boldsymbol{\beta}}=\boldsymbol{0}^\top \tag{3.24}$$

$\mathrm{d}((\boldsymbol{y}-\boldsymbol{\Phi\beta})^\top(\boldsymbol{y}-\boldsymbol{\Phi\beta}))$可利用向量微积分的运算法则(附录 C)进一步化简为

$$\begin{aligned}\mathrm{d}((\boldsymbol{y}-\boldsymbol{\Phi\beta})^\top(\boldsymbol{y}-\boldsymbol{\Phi\beta})+\lambda\boldsymbol{\beta}^\top\boldsymbol{\beta})&=2((\boldsymbol{y}-\boldsymbol{\Phi\beta})^\top\mathrm{d}(\boldsymbol{y}-\boldsymbol{\Phi\beta})+\lambda\boldsymbol{\beta}^\top\mathrm{d}\boldsymbol{\beta})\\&=-2((\boldsymbol{y}-\boldsymbol{\Phi\beta})^\top\boldsymbol{\Phi}-\lambda\boldsymbol{\beta}^\top)\mathrm{d}\boldsymbol{\beta}\\&=2(\boldsymbol{\beta}^\top\boldsymbol{\Phi}^\top\boldsymbol{\Phi}-\boldsymbol{y}^\top\boldsymbol{\Phi}+\lambda\boldsymbol{\beta}^\top)\mathrm{d}\boldsymbol{\beta}\end{aligned} \tag{3.25}$$

因此，得到 $\boldsymbol{\beta}$ 的最优值为

$$\hat{\boldsymbol{\beta}}_{\mathrm{rls}}=(\lambda\boldsymbol{I}+\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\boldsymbol{y} \tag{3.26}$$

(3) 概率线性回归的最大后验估计。

回顾 3.1.1 节，线性回归的似然假设为高斯分布时，如果使用最大后验估计来获取模型参数，需要假设参数的先验分布。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为

$$p(\boldsymbol{\beta}\mid\alpha)=\mathcal{N}(\boldsymbol{\beta}\mid\boldsymbol{0},\alpha^{-1}\boldsymbol{I}) \tag{3.27}$$

根据贝叶斯公式可以得出参数的对数后验分布为

$$\ln p(\boldsymbol{\beta} \mid X, \boldsymbol{y}, \alpha, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - f(\boldsymbol{x}_i, \boldsymbol{\beta}))^2 - \frac{\alpha}{2}\boldsymbol{\beta}^{\top}\boldsymbol{\beta} + \text{const} \tag{3.28}$$

其中 const 表示与 $\boldsymbol{\beta}$ 无关的项。将最大后验估计的优化目标(公式(3.28))与正则化最小二乘的目标(公式(3.23))对比可以发现,当 $\alpha = \lambda/\sigma^2$ 时,二者的解是等价的。因此,通过给模型参数增加先验并进行最大后验估计的方法同样可以达到减轻过拟合的效果。

3.2 贝叶斯线性回归

3.1 节介绍的使用最大似然估计与最大后验估计的线性回归模型,模型参数的值完全通过数据集训练得出。一旦得到 $\hat{\boldsymbol{\beta}}$,可以根据模型估计任意新数据点 $\boldsymbol{x}_*$ 的输出值:$\hat{y} = \phi(\boldsymbol{x}_*)^{\top}\hat{\boldsymbol{\beta}}$。它们得到的是参数的点估计,即给定数据时可能性最大的估计。但是,当数据集比较小或不确定性较大时,将估计表示为一个可能值的分布更加合理。这样得到的输出预测值将是一个分布。下面介绍可以给出参数与预测值分布的贝叶斯线性回归。

贝叶斯线性回归(Bayesian linear regression)将贝叶斯框架应用到线性回归中,回归系数 $\boldsymbol{\beta}$ 被假设为有一特定先验分布的随机变量,此先验分布可以影响回归系数的解。另外,贝叶斯参数估计不是给出回归系数的最佳单点估计,而是给出完整的后验分布,这种方式描述了估计量的不确定性。

考虑一个标准的线性回归问题,对于 $i = 1, 2, \cdots, N$,假设在给定自变量 $\boldsymbol{x}_i$ 的情况下,y_i 产生的公式为

$$y_i = \boldsymbol{x}_i^{\top}\boldsymbol{\beta} + \epsilon_i \tag{3.29}$$

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 $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N]^{\top}$,$\boldsymbol{y} = [y_1, y_2, \cdots, y_N]^{\top}$,可以得到因变量 $\boldsymbol{y}$ 的似然函

数为

$$p(\boldsymbol{y} \mid \boldsymbol{X},\boldsymbol{\beta},\sigma^2) \propto (\sigma^2)^{-\frac{N}{2}} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^{\top}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right] \tag{3.30}$$

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

在贝叶斯方法中，参数的先验概率分布为模型提供了额外信息。先验可以根据领域知识和已知信息采取不同的函数形式。确定似然函数后，对于一个任意的先验分布，后验分布不一定存在解析形式。这里讨论可以使后验分布被解析地推导出来的情况，常用的方法是设置似然函数的共轭先验。下面给出共轭先验的定义。

如果先验分布和似然函数可以使后验分布和先验分布具有相同的形式，就称先验分布与似然函数是共轭的，该先验称为该似然函数的共轭先验。共轭的好处是让后验分布与先验分布具有相同的形式，从而便于求解。

以上文介绍的线性回归为例，给定模型的似然假设公式(3.30)，需要进行贝叶斯估计的参数包括 $\boldsymbol{\beta}$ 和 σ^2。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^2 的联合先验为

$$p(\boldsymbol{\beta},\sigma^2) = p(\sigma^2)p(\boldsymbol{\beta} \mid \sigma^2) \tag{3.31}$$

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

$$p(\sigma^2) \propto (\sigma^2)^{-a_0-1} \exp\left(-\frac{b_0}{\sigma^2}\right) \tag{3.32}$$

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

$$p(\boldsymbol{\beta} \mid \sigma^2) \propto \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)^{\top}\Lambda_0(\boldsymbol{\beta}-\boldsymbol{\mu}_0)\right] \tag{3.33}$$

给定 $\boldsymbol{\beta}$ 和 σ^2 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

$$\begin{aligned} p(\boldsymbol{\beta},\sigma^2 \mid \boldsymbol{y},\boldsymbol{X}) &= p(\boldsymbol{\beta} \mid \sigma^2,\boldsymbol{y},\boldsymbol{X})p(\sigma^2 \mid \boldsymbol{y},\boldsymbol{X}) \propto \\ & p(\boldsymbol{y} \mid \boldsymbol{X},\boldsymbol{\beta},\sigma^2)p(\boldsymbol{\beta} \mid \sigma^2)p(\sigma^2) \end{aligned} \tag{3.34}$$

将公式(3.30)、(3.32)和(3.33)代入(3.34)，可得 $p(\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X})$是高斯分布

$\mathcal{N}(\boldsymbol{\beta}|\boldsymbol{\mu}_N,\sigma^2\boldsymbol{\Lambda}_N^{-1})$，以及 $p(\sigma^2|\boldsymbol{y},\boldsymbol{X})$是逆伽马分布 Inv-Gamma$(\sigma^2|a_N,b_N)$，其参数的具体表示为

$$\begin{cases}\boldsymbol{\Lambda}_N=(\boldsymbol{X}^\top\boldsymbol{X}+\boldsymbol{\Lambda}_0)\\ \boldsymbol{\mu}_N=(\boldsymbol{\Lambda}_N)^{-1}(\boldsymbol{X}^\top\boldsymbol{y}+\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0)\\ a_N=a_0+\dfrac{N}{2}\\ b_N=b_0+\dfrac{1}{2}(\boldsymbol{y}^\top\boldsymbol{y}+\boldsymbol{\mu}_0^\top\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0-\boldsymbol{\mu}_N^\top\boldsymbol{\Lambda}_N\boldsymbol{\mu}_N)\end{cases}\tag{3.35}$$

3.3 逻辑回归

前文介绍了使用线性函数预测连续取值的变量，这类问题称为回归问题。很多时候，也需要预测离散取值变量，例如判断一张图像属于哪个目标类别，这变成了分类问题。逻辑回归在线性回归的基础上实现了二类和多类分类。

逻辑回归(logistic regression)[3]模型是一种常用的分类算法，也可以认为是一种因变量为离散值的回归模型。逻辑回归可以处理二类分类和多类分类问题。在二类逻辑回归中，因变量只有两种取值，例如“0”或“1”。在多类逻辑回归中，因变量有两种以上的离散取值。

本节首先介绍二类逻辑回归，然后介绍多类逻辑回归。

3.3.1 二类逻辑回归

二类逻辑回归模型使用一个或多个自变量(特征)来估计因变量取值的概率。输出通常被编码为“0”或“1”。模型本身根据输入仅仅建模了输出的概率，并不执行分类，即模型本身并不是一个分类器。当然，通常可以使用此模型构造一个分类器，例如，选择一个阈值，将概率大于此阈值的输入分为一类，小于此阈值的分为另一类。逻辑回归模型使用逻辑函数(logistic

function)，将线性回归的返回值转换为区间[0,1]内的值，用于表示自变量属于某个类别的概率，即因变量取值为“0”或“1”的概率。

逻辑函数也称为 sigmoid 函数，输入可以是任意实数 $x(x\in\mathbb{R})$，输出的值属于区间[0,1]。逻辑函数 $\sigma(x)$的表达式为

$$\sigma(x)=\frac{e^x}{e^x+1}=\frac{1}{1+e^{-x}} \tag{3.36}$$

其函数曲线如图 3-5 所示。它是一个 S 形曲线，在横坐标取值远离 0 时，纵坐标的值趋近 0 或 1。

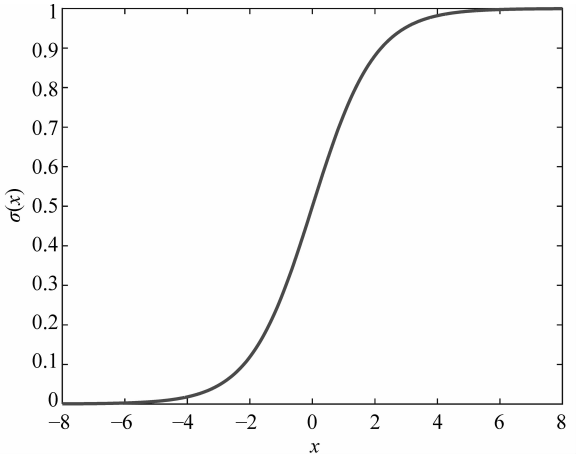


图 3-5　逻辑函数示意图

逻辑回归使用逻辑函数和回归模型可以解决二分类问题，其中逻辑函数的返回值用于表示二分类问题中的正类或负类的概率。假设 f 是自变量 $\boldsymbol{x}$ 的一个线性函数，即 $f=\boldsymbol{\theta}^\top\boldsymbol{x}$。逻辑回归假设样本 $\boldsymbol{x}$ 属于正类的概率为

$$p(y=1\mid\boldsymbol{x})=h_{\boldsymbol{\theta}}(\boldsymbol{x})=\sigma(\boldsymbol{\theta}^\top\boldsymbol{x})=\frac{1}{1+\exp(-\boldsymbol{\theta}^\top\boldsymbol{x})} \tag{3.37}$$

那么，$\boldsymbol{x}$ 属于负类的概率为

$$p(y=0\mid\boldsymbol{x})=1-p(y=1\mid\boldsymbol{x})=1-h_{\boldsymbol{\theta}}(\boldsymbol{x})=\frac{1}{1+\exp(\boldsymbol{\theta}^\top\boldsymbol{x})} \tag{3.38}$$

逻辑回归可以从两个角度定义目标函数，一种是从最大似然的角度，一种是从直接构建损失的角度。逻辑回归的优化目标是学习得到合适的参数值，使得概率 $p(y=1|\boldsymbol{x})=h_{\boldsymbol{\theta}}(\boldsymbol{x})$ 在当 $\boldsymbol{x}$ 属于“1”类时值比较大，且 $p(y=0|\boldsymbol{x})=1-h_{\boldsymbol{\theta}}(\boldsymbol{x})$ 在当 $\boldsymbol{x}$ 属于“0”类时值比较大。

从最大似然的角度分析，假设每一个样本的类标签都是独立同分布的伯努利变量，伯努利变量取值为 1 和 0 的概率分别为公式(3.37)和公式(3.38)。对于有标签的训练集 $\{(\boldsymbol{x}_i,y_i):i=1,2,\cdots,N\}$，$N$ 个独立样本的联合似然可以写成

$$p(\boldsymbol{y}\mid\boldsymbol{\theta})=\prod_{i=1}^{N}p(y_i=1\mid\boldsymbol{x}_i)^{y_i}(1-p(y_i=1\mid\boldsymbol{x}_i))^{(1-y_i)} \tag{3.39}$$

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

$$\begin{aligned}-\ln p(\boldsymbol{y}\mid\boldsymbol{\theta})=-\sum_{i=1}^{N}[&y_i\ln p(y_i=1\mid\boldsymbol{x}_i)+\\&(1-y_i)\ln(1-p(y_i=1\mid\boldsymbol{x}_i))]\end{aligned} \tag{3.40}$$

从构建损失函数的角度分析，逻辑回归使用真实概率分布与模型概率分布的交叉熵损失来直接定义训练集 $\{(\boldsymbol{x}_i,y_i):i=1,2,\cdots,N\}$ 的损失函数。假设每个样本的真实分布为 $q(y_i|\boldsymbol{x}_i)$，那么，$q(y_i=1|\boldsymbol{x}_i)=y_i$，且 $q(y_i=0|\boldsymbol{x}_i)=1-y_i$。分布 $q(y_i|\boldsymbol{x}_i)$ 和 $p(y_i|\boldsymbol{x}_i)$ 的交叉熵为

$$H(q(y_i\mid\boldsymbol{x}_i),p(y_i\mid\boldsymbol{x}_i))=-\sum_{y_i}q(y_i\mid\boldsymbol{x}_i)\ln p(y_i\mid\boldsymbol{x}_i) \tag{3.41}$$

因此，逻辑回归的交叉熵损失为

$$\begin{aligned}J(\boldsymbol{\theta})&=\sum_{i=1}^{N}H[q(y_i\mid\boldsymbol{x}_i),p(y_i\mid\boldsymbol{x}_i)]\\&=-\sum_{i=1}^{N}[y_i\ln h_{\boldsymbol{\theta}}(\boldsymbol{x}_i)+(1-y_i)\ln(1-h_{\boldsymbol{\theta}}(\boldsymbol{x}_i))]\end{aligned} \tag{3.42}$$

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 $J(\boldsymbol{\theta})$ 找到假设函数 $h_{\boldsymbol{\theta}}(\boldsymbol{x})$ 中 $\boldsymbol{\theta}$ 的最优值，从而学得分类器。关于该目标的优化得不到闭式解[4]，因此常用基于梯度的迭代优化方

法，例如一阶梯度下降或基于二阶梯度的牛顿法等。使用梯度下降等方法优化 $\boldsymbol{\theta}$，需要计算 $J(\boldsymbol{\theta})$ 关于 $\boldsymbol{\theta}$ 的梯度，计算公式为

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \left(\frac{\mathrm{d}J(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}}\right)^{\top} = \left(\frac{\sum_{i=1}^{N}\left[(\sigma(\boldsymbol{\theta}^{\top}\boldsymbol{x}_i) - y_i)\boldsymbol{x}_i^{\top}\right]\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}\boldsymbol{\theta}}\right)^{\top}$$

$$= \sum_i \boldsymbol{x}_i\left[h_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i\right] \tag{3.43}$$

其中，梯度的运算过程利用了 sigmoid 函数的导数性质：$\mathrm{d}\sigma(x) = \sigma(x)(1-\sigma(x))\mathrm{d}x$。使用牛顿法进行优化，则还需要计算 Hessian 矩阵。

得到合适的参数值后，对于新的测试样本 $\boldsymbol{x}_*$，如果 $p(y=1|\boldsymbol{x}_*) > p(y=0|\boldsymbol{x}_*)$，那么将此样本标记为“1”类，否则标记为“0”类。相应的决策函数为：如果 $p(y=1|\boldsymbol{x}_*) > 0.5$，那么 $y_* = 1$。通常情况下，选择 0.5 作为阈值进行决策，在很多实际应用中，也可以根据特定的情况选择不同的阈值。例如，如果对正例的判别查准率要求高，可以选择大于 0.5 的值作为阈值；如果对正例的查全率要求高，可以选择小于 0.5 的值作为阈值。

3.3.2　多类逻辑回归

多类逻辑回归（multinomial logistic regression）的基本原理与二类逻辑回归类似，差别在于多类逻辑回归中因变量 y_i 的取值可以大于两个，一个 C 类逻辑回归的因变量可以在 $1 \sim C$ 取任意一个整数。多类逻辑回归使用 softmax 实现从实数到类别概率的转换。

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，即

$$f_c(\boldsymbol{x}_i) = \boldsymbol{\theta}_c^{\top}\boldsymbol{x}_i \tag{3.44}$$

其中 $\boldsymbol{\theta}_c$ 是与类别 c 对应的回归系数，$\boldsymbol{x}_i$ 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

$$p(y_i = c) = \frac{\exp(\boldsymbol{\theta}_c^{\top}\boldsymbol{x}_i)}{\sum_{k=1}^{C}\exp(\boldsymbol{\theta}_k^{\top}\boldsymbol{x}_i)} \tag{3.45}$$

根据公式(3.45),样本被分为概率最大的那一类。每个向量 $\boldsymbol{\theta}_c$ 中未知的参数可以通过最大似然或最小化交叉熵进行优化。多类逻辑回归的似然函数为

$$p(\boldsymbol{y} \mid \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_C) = \prod_{i=1}^{N} \prod_{c=1}^{C} p(y_i = c \mid \boldsymbol{x}_i)^{I(y_i = c)} \tag{3.46}$$

其中,$I(y_i = c)$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

$$-\ln p(\boldsymbol{y} \mid \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_K) = -\sum_{i=1}^{N} \sum_{c=1}^{C} I(y_i = c) \ln p(y_i = c \mid \boldsymbol{x}_i) \tag{3.47}$$

与二类逻辑回归类似,由于优化目标中包含非线性函数,通常得不到闭式解,因此常用的方法是基于梯度的迭代优化。此外,无论是二类逻辑回归还是多类逻辑回归,使用最大后验估计或最小化带惩罚项的交叉熵损失可以防止模型过拟合。

3.4 贝叶斯逻辑回归

本节以两分类为例介绍贝叶斯逻辑回归(Bayesian logistic regression)。逻辑回归是一种判别式概率线性分类器 $p(y=1 \mid \boldsymbol{x}, \boldsymbol{\theta}) = \sigma(\boldsymbol{\theta}^{\top} \boldsymbol{x})$。贝叶斯逻辑回归通过贝叶斯参数估计学习参数的后验分布,并且利用该分布进行预测。

已知观测数据 $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N]^{\top}$,$\boldsymbol{y} = [y_1, y_2, \cdots, y_N]^{\top}$,逻辑回归使用的似然分布导致后验分布 $p(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{y})$ 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。预测时,即便使用了近似分布,对新样本 $\boldsymbol{x}_*$ 的预测分布 $p(y_* = 1 \mid \boldsymbol{x}_*) \approx \int \sigma(\boldsymbol{\theta}^{\top} \boldsymbol{x}_*) q(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$ 的估计仍然是难解的。因此,贝叶斯逻辑回归通常使用近似求解方法。

一方面,后验分布 $p(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{y})$等于先验乘以似然,再进行归一化。其中先验通常假设为

$$p(\boldsymbol{\theta})=\mathcal{N}(\boldsymbol{\theta}\mid \boldsymbol{m}_0,\boldsymbol{S}_0) \tag{3.48}$$

逻辑回归的似然为

$$p(\boldsymbol{y}\mid X,\boldsymbol{\theta})=\prod_{i=1}^{N} p\left(y_i=1\mid \boldsymbol{x}_i\right)^{y_i}\left(1-p(y_i=1\mid \boldsymbol{x}_i)\right)^{1-y_i} \tag{3.49}$$

对逻辑回归中的后验分布进行精确求解非常困难，这时可以通过使用拉普拉斯近似得到近似的高斯后验分布 $q(\boldsymbol{\theta})$。

另一方面，预测分布 $p(y_*=1|\boldsymbol{x}_*)\approx\int\sigma(\boldsymbol{\theta}^{\top}\boldsymbol{x}_*)q(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ 需要关于 sigmoid 函数和高斯分布的乘积求积分，其精确求解也是十分困难的，可通过将 sigmoid 函数用逆 probit 函数近似得到其近似解[4]。下面对这两方面的近似进行详细介绍。

(1) 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为

$$q(\boldsymbol{\theta})=\frac{1}{(2\pi)^{D/2}\mid \boldsymbol{S}_N\mid^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\top}\boldsymbol{S}_N^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right]=\mathcal{N}(\boldsymbol{\theta}\mid \boldsymbol{\theta}_0,\boldsymbol{S}_N) \tag{3.50}$$

其中，均值 $\boldsymbol{\theta}_0$ 是真实后验分布的最大值对应的变量值，协方差矩阵是负对数真实后验分布 $-\ln p(\boldsymbol{\theta}|X,\boldsymbol{y})$ 的 Hessian 矩阵(附录 C)在 $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ 处的逆，即 $\boldsymbol{S}_N=(-\nabla\nabla\ln p(\boldsymbol{\theta}|X,\boldsymbol{y})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0})^{-1}$。下面来看均值 $\boldsymbol{\theta}_0$ 和协方差矩阵 $\boldsymbol{S}_N$ 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 $p(\boldsymbol{\theta}|X,\boldsymbol{y})\propto p(\boldsymbol{\theta})p(\boldsymbol{y}|X,\boldsymbol{\theta})$。将先验概率(公式(3.48))和逻辑回归的似然函数(公式(3.49))代入贝叶斯公式可得

$$\ln p(\boldsymbol{\theta}\mid X,\boldsymbol{y})=-\frac{1}{2}(\boldsymbol{\theta}-m_0)^{\top}\boldsymbol{S}_0^{-1}(\boldsymbol{\theta}-m_0)+\sum_{i=1}^{N}\left[y_i\ln p(y_i=1\mid \boldsymbol{x}_i,\boldsymbol{\theta})+(1-y_i)\ln(1-p(y_i=1\mid \boldsymbol{x}_i,\boldsymbol{\theta}))\right]+\text{const} \tag{3.51}$$

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

$$\begin{aligned}\boldsymbol{H} &= -\nabla\nabla \ln p(\boldsymbol{\theta} \mid \boldsymbol{X},\boldsymbol{y}) = \frac{\mathrm{d}^2 \ln p(\boldsymbol{\theta} \mid \boldsymbol{X},\boldsymbol{y})}{\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\theta}^{\top}} \\ &= \frac{\mathrm{d}\mathrm{Tr}\left[(\boldsymbol{\theta}-\boldsymbol{m}_0)^{\top}\boldsymbol{S}_0^{-1}\mathrm{d}\boldsymbol{\theta}\right] - \left(\mathrm{d}\sum_{i=1}^{N}((y_i-\sigma(\boldsymbol{\theta}^{\top}\boldsymbol{x}_i))\boldsymbol{x}_i^{\top}\mathrm{d}\boldsymbol{\theta})\right)}{\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\theta}^{\top}} \\ &= \frac{\mathrm{Tr}\left[\boldsymbol{S}_0^{-1}\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\theta}^{\top}\right] + \mathrm{Tr}\left[\sum_{i=1}^{N}\sigma(\boldsymbol{\theta}^{\top}\boldsymbol{x}_i)(1-\sigma(\boldsymbol{\theta}^{\top}\boldsymbol{x}_i))\boldsymbol{x}_i\boldsymbol{x}_i^{\top}\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\theta}^{\top}\right]}{\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\theta}^{\top}} \\ &= \boldsymbol{S}_0^{-1} + \sum_{i=1}^{N} p(y_i=1 \mid \boldsymbol{x}_i,\boldsymbol{\theta})(1-p(y_i=1 \mid x_i,\boldsymbol{\theta}))\boldsymbol{x}_i\boldsymbol{x}_i^{\top} \end{aligned} \tag{3.52}$$

其中运算过程利用了 sigmoid 函数的导数性质：$\mathrm{d}\sigma(x)=\sigma(x)(1-\sigma(x))\mathrm{d}x$。得到 $\boldsymbol{H}$ 之后，根据 $\boldsymbol{S}_N=(\boldsymbol{H}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\text{map}}})^{-1}$得到近似分布的协方差矩阵 $\boldsymbol{S}_N$，可以得到后验分布的高斯近似 $q(\boldsymbol{\theta})=\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{map}},\boldsymbol{S}_N)$。

(2) 逆 probit 函数近似。

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

$$\begin{aligned} p(y_*=1 \mid \boldsymbol{x}_*) &= \int p(y_*=1,\boldsymbol{\theta} \mid x_*)\mathrm{d}\boldsymbol{\theta} \\ &= \int p(y_*=1 \mid \boldsymbol{x}_*,\boldsymbol{\theta})p(\boldsymbol{\theta} \mid \boldsymbol{X},\boldsymbol{y})\mathrm{d}\boldsymbol{\theta} \\ &\approx \int \sigma(\boldsymbol{\theta}^{\top}\boldsymbol{x}_*)q(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \end{aligned} \tag{3.53}$$

属于类别“0”的概率为

$$p(y_*=0 \mid \boldsymbol{x}_*)=1-p(y_*=1 \mid \boldsymbol{x}_*) \tag{3.54}$$

下面对公式(3.53)作进一步化简，由于函数 $\sigma(\boldsymbol{\theta}^{\top}\boldsymbol{x}_*)$ 仅通过 $\boldsymbol{\theta}^{\top}\boldsymbol{x}_*$ 的值依赖于 $\boldsymbol{\theta}$，因此定义新的变量 $a=\boldsymbol{\theta}^{\top}\boldsymbol{x}_*$，并引入 Dirac delta 函数 $\delta(\cdot)$，可以得

到 $\sigma(\boldsymbol{\theta}^{\top}\boldsymbol{x}_*)\approx\int\delta(a-\boldsymbol{\theta}^{\top}\boldsymbol{x}_*)\sigma(a)\mathrm{d}a$。因此，公式(3.53)的结果可以表示为

$$\begin{aligned}\int\sigma(\boldsymbol{\theta}^{\top}\boldsymbol{x}_*)q(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}&=\int\left(\int\delta(a-\boldsymbol{\theta}^{\top}\boldsymbol{x}_*)\sigma(a)\mathrm{d}a\right)q(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\\&=\int\sigma(a)\int\delta(a-\boldsymbol{\theta}^{\top}\boldsymbol{x}_*)q(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\mathrm{d}a\end{aligned}\tag{3.55}$$

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

$$\mu_a=\mathbb{E}[a]=\int p(a)a\mathrm{d}a=\int q(\boldsymbol{\theta})\boldsymbol{\theta}^{\top}\boldsymbol{x}_*\mathrm{d}\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathrm{map}}^{\top}\boldsymbol{x}_*\tag{3.56}$$

$$\begin{aligned}\sigma_a^2&=\mathrm{var}[a]=\int p(a)a^2\mathrm{d}a-\mathbb{E}[a]^2=\int q(\boldsymbol{\theta})(\boldsymbol{\theta}^{\top}\boldsymbol{x}_*)^2\mathrm{d}\boldsymbol{\theta}-(\boldsymbol{\theta}_{\mathrm{map}}^{\top}\boldsymbol{x}_*)^2\\&=\boldsymbol{x}_*^{\top}\mathrm{S}_N\boldsymbol{x}_*\end{aligned}\tag{3.57}$$

预测分布可以表示为

$$p(y=1\mid\boldsymbol{x}_*)=\int\sigma(a)p(a)\mathrm{d}a=\int\sigma(a)\,\mathcal{N}(a\mid\mu_a,\sigma_a^2)\mathrm{d}a\tag{3.58}$$

注意，在公式(3.58)的积分中，关于 sigmoid 和 Gaussian 乘积的积分是不可解的，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为

$$\Phi(a)=\int_{-\infty}^{a}\mathcal{N}(w\mid0,1)\mathrm{d}w\tag{3.59}$$

该函数也称为逆 probit 函数。由于累积分布函数的值域是(0,1)，因此可用逆 probit 函数来近似 sigmoid 函数。为使二者尽可能一致，需对逆 probit 函数的自变量进行放缩，即使用 $\Phi(\lambda a)$ 来近似 $\sigma(a)$，并且 λ 通常设置为 $\lambda=\sqrt{\pi/8}$，此时两者在原点具有相同的斜率(即导数相同)[4]。高斯分布和逆 probit 函数相乘后的积分还是一个逆 probit 函数，即

$$\int\Phi(\lambda a)\,\mathcal{N}(a\mid\mu_a,\sigma_a^2)\mathrm{d}a=\Phi\left(\frac{\mu_a}{(\lambda^{-2}+\sigma_a^2)^{1/2}}\right)\tag{3.60}$$

将公式(3.60)应用到公式(3.58)中，可以获得最终的预测概率为