计算机模式识别

----- 理论、技术与编程

主讲：图像处理与模式识别研究所
赵群飞

邮　箱：zhaoqf@sjtu.edu.cn
办公室：电院 2-441
电　话：13918191860

- 本章学习目标

  ✓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

  ✓了解贝叶斯逻辑回归

目录

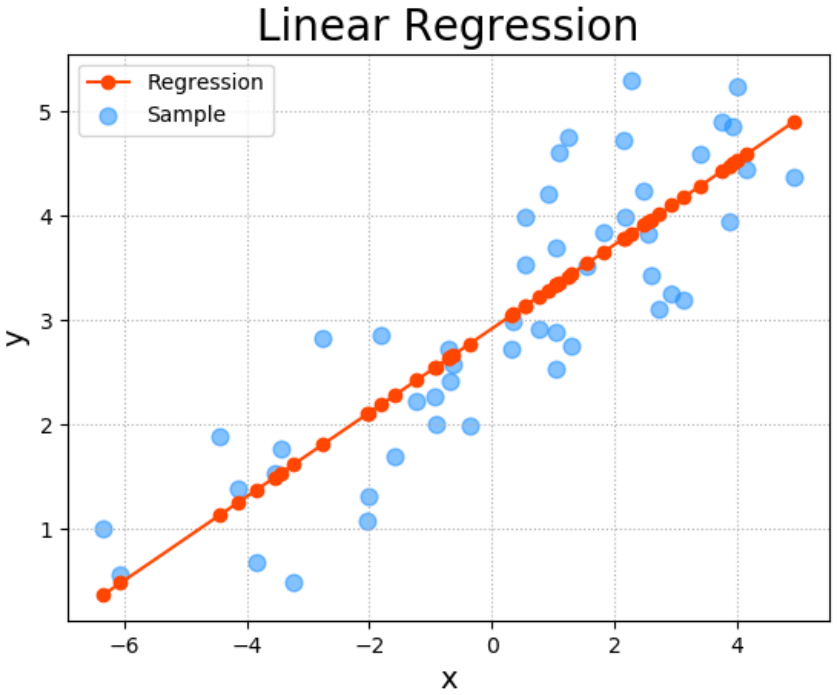
- 回归（Regression）与分类一样，也是机器学习的基本任务，属于监督学习的范畴，即通过训练有标注的样本来学习输入与输出的关系（学习模型），进而预测新样本的输出。
- 通常，样本数据的输出取值为有限个离散值时，任务称之为分类；输出取值为1或0时，任务称为二分类；当输出取值为连续值时，任务就称之为回归。

  如大熊猫、小熊猫，熊猫的竹饲料，股票、房价走势。

- 逻辑回归，从字面上看，好像是一种回归方法，其实是一种分类方法。
- 传统的逻辑回归用于处理二分类问题，但通过引入softmax函数就可处理多分类问题。
- 之所以称之为逻辑回归，主要是它由线性回归转变而来，通过逻辑函数（sigmoid函数）实现对输出函数的非线性转换，得到样本属于某一类别的概率，然后是由该概率值进行分类决策。

- 线性回归（Linear Regression）

是一种通过属性的线性组合来进行预测的**线性模型**，其目的是找到一条直线或者一个平面或者更高维的超平面，**使得预测值与真实值之间的误差最小化。**



- 给定数据集$\mathcal{D} = \{y_i, x_{i1}, \cdots, x_{iD}\}_{i=1}^{N}$，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。
  - N 代表训练集中样本的数量
  - D 代表特征的数量
  - x 代表特征/输入变量
  - y 代表目标变量/输出变量
  - (y,x) 代表训练集中的样本
  - $(y_i,\mathbf{x_i})$ 代表第i个观察样本

- 多元线性回归关系的模型形式如下：

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

令x_{i0}=1， $\mathbf{x_i^T}$=$(x_{i0}, x_{i1}, \cdots, x_{iD})$

$\boldsymbol{\beta}$为回归系数

- 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。只要回归系数$\boldsymbol{\beta}$是线性的，就认为是线性模型。
- 假设$\phi(\mathbf{x_i})$表示对输入特征的变换函数，也称为基函数，那么线性函数可以更一般表示为

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

常见基函数有三种：

1. 多项式基函数：

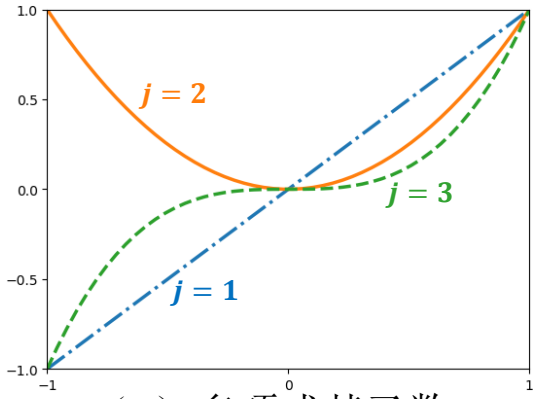
$$\phi_j(x) = x^j$$

2. 高斯基函数：

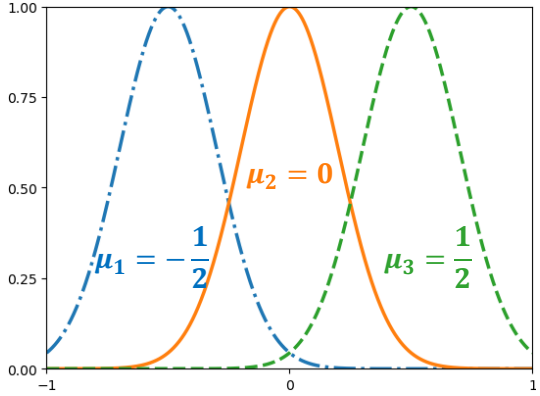
$$\phi_j(x) = \exp\{-\frac{(x-\mu_j)^2}{2s^2}\}$$

3. S形（sigmoidal）基函数：

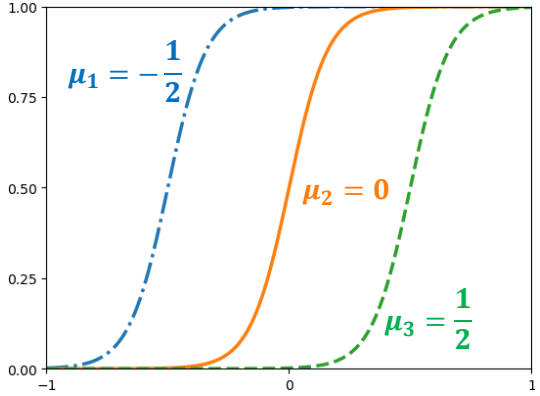
$$\phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right), \qquad \sigma(a) = \frac{1}{1+\exp(-a)}$$



（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

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

其中 $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\beta}) = \boldsymbol{\phi}(\mathbf{x_i})^{\mathrm{T}}\boldsymbol{\beta}$

- 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

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

- 假设数据是独立同分布的，所有观测的似然概率分布为

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

- 在确定了模型的概率表示之后，对于新的测试数据$\mathbf{x}_*$，可以使用输出变量的数学期望作为预测值：

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

• 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：

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

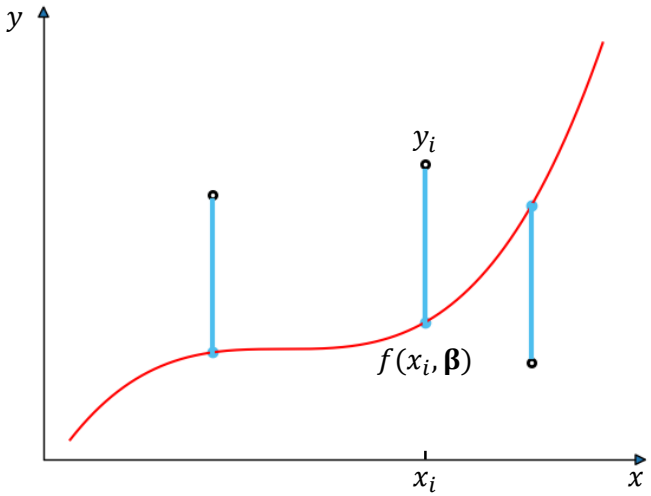


图3-2 误差几何意义示意图（图中纵向线段长度代表不同数据点的误差。）

- 最小二乘通过最小化平方误差和S开学习最优参数值：

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

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 $f(\mathbf{x}_i,\boldsymbol{\beta})=\phi(\mathbf{x}_i)^{\mathrm{T}}\boldsymbol{\beta}$

我们定义 $X=[\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_N]^{\mathrm{T}}$, $\mathbf{y}=[y_1,y_2,...,y_N]^{\mathrm{T}}$, $\Phi=[\phi(\mathbf{x}_1),\phi(\mathbf{x}_2),...,\phi(\mathbf{x}_n)]^{\mathrm{T}}$

那么模型在训练数据上的预测平方误差为

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

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

其中，$\mathbf{0}$表示元素为0的列向量。

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

上式分子 $d\left((\mathbf{y}-\Phi\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\Phi\boldsymbol{\beta})\right)$ 可利用向量微积分的运算法则进一步化简：

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

因此，得到$\boldsymbol{\beta}$最优解

$$\hat{\boldsymbol{\beta}}_{ls}=(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}\mathbf{y}$$

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

$$\hat{\boldsymbol{\beta}}_{ml} = (\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}\mathbf{y}$$

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

- 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）

欠拟合的处理

1.**添加新特征**

当特征不足或者现有特征与样本标签的相关性不强时，模型容易出现欠拟合。通过挖掘组合特征等新的特征，往往能够取得更好的效果。

2.**增加模型复杂度**

简单模型的学习能力较差，通过增加模型的复杂度可以使模型拥有更强的拟合能力。例如，在线性模型中添加高次项，在神经网络模型中增加网络层数或神经元个数等。

3.**减小正则化系数**

正则化是用来防止过拟合的，但当模型出现欠拟合现象时，则需要有针对性地减小正则化系数。

过拟合的处理

1.获得更多的训练数据

使用更多的训练数据是解决过拟合问题最有效的手段，因为更多的样本能够让模型学习到更多更有效的特征，减小噪声的影响。

2.降维

即丢弃一些不能帮助我们正确预测的特征。可以是手工选择保留哪些特征，或者使用一些模型选择的算法来帮忙（例如PCA）。

3.正则化

正则化(regularization)的技术，保留所有的特征，但是减少参数的大小（magnitude），改善或者减少过拟合问题。

4.集成学习方法

集成学习是把多个模型集成在一起，来降低单一模型的过拟合风险。

- 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（$\|\boldsymbol{\beta}\|_{L_2}=\sqrt{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}}$）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

$$S'=\sum_{i=1}^{N}\left(y_i-f(\mathbf{x}_i,\boldsymbol{\beta})\right)^2+\lambda\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta},$$

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。
- 正则化也可使用L_1范数作为惩罚项：

$$\|\boldsymbol{\beta}\|_{L_1}=\sum_d|\beta_d|$$

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

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

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

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

因此，得到$\boldsymbol{\beta}$的最优解为

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

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

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

根据贝叶斯公式可以得出参数的对数后验分布是

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

Const 表示与$\boldsymbol{\beta}$无关的项。

目录

- 线性回归
- 贝叶斯线性回归
- 逻辑回归
- 贝叶斯逻辑回归（参考内容）

- 使用最大似然估计与最大后验估计得到的线性回归模型参数，通过数据集训练可得，有了模型就可估计任意新的数据点的输出值，它是给定数据时可能性最大的估计。
- 但是当数据较少或者不确定性较大时，把输出的预测值表示为一个可能值的分布更加合理。
- 贝叶斯线性回归就可以求得参数和输出值的分布。

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

$$\boldsymbol{y}_i = \boldsymbol{x}^T{}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_\mathbf{i}$$

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

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

即

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

- 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

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

需要进行贝叶斯估计的参数是$\boldsymbol{\beta}$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

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

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

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

给定参数$\boldsymbol{\beta}$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

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

由此可得，　$p(\boldsymbol{\beta}\,|\,\sigma^2,\mathbf{y},\mathbf{X})$ 是高斯分布　$\mathrm{N}(\boldsymbol{\beta}\,|\,\boldsymbol{\mu}_N,\sigma^2\Lambda_N^{-1})$，

$p(\sigma^2\,|\,\mathbf{y},\mathbf{X})$　是逆伽马分布　$\text{Inv-Gamma}(\sigma^2\,|\,a_N,b_N)$

其参数具体表示如下：

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

目录

- 逻辑回归，从字面上看，好像是一种回归方法，其实是一种分类方法。
- 传统的逻辑回归用于处理二分类问题，但通过引入softmax函数就可处理多分类问题。
- 之所以称之为逻辑回归，主要是它由线性回归转变而来，通过逻辑函数（sigmoid函数）实现对输出函数的非线性转换，得到样本属于某一类别的概率，然后是由该概率值进行分类决策。

逻辑回归就是在线性回归的基础上实现二分类和多分类

- 二类逻辑回归
  - 二分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“1”or“0”。
  - 逻辑回归模型本身只根据输入建立了输出概率的模型，并不进行分类，即模不是分类器。
  - 当然，如果选定一个阈值，概率大于该阈值的分为一类，否则分为另一类，从而相当于构建了一个分类器。
  - **逻辑回归模型使用逻辑函数将线性回归的返回值转换为区间 [0,1] 的概率。**

- 逻辑函数也称为Sigmoid函数

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

- 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。

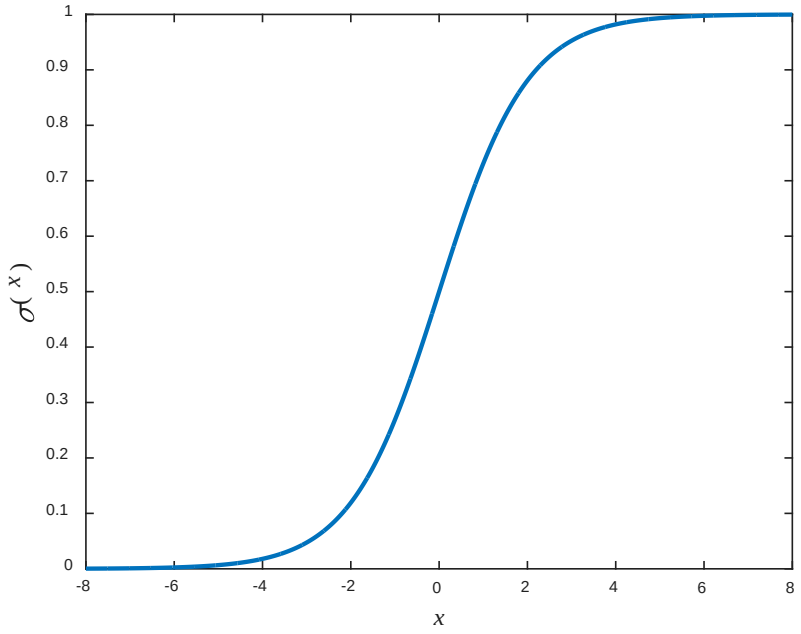


图3-5 逻辑函数示意图

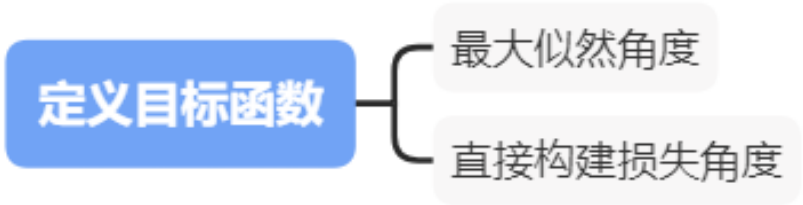
假设f是自变量$\mathbf{x}$的一个线性函数，即$f = \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}$。逻辑回归假设样本$\mathbf{x}$属于正类的概率为

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

那么，$\mathbf{x}$属于负类的概率为

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

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

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

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

- 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$q(y_i=0|\mathbf{x}_i)=1-y_i$。

那么，分布$q(y_i|\mathbf{x}_i)$和$p(y_i|\mathbf{x}_i)$的交叉熵是

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

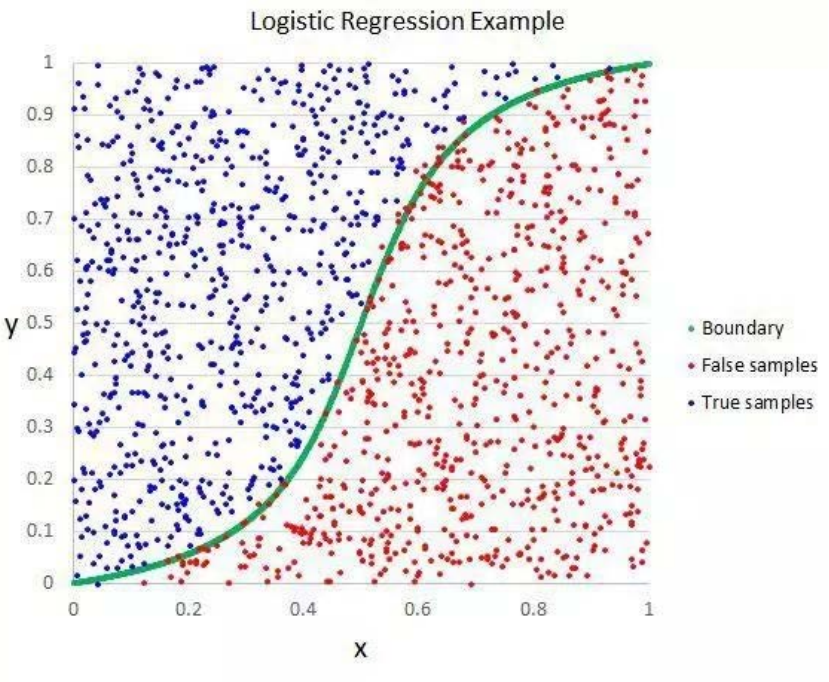
◆ 因此逻辑回归的交叉熵损失为

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

- 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$h_\theta(\mathbf{x})$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：

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

- 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。



◆ **求得合适的参数后，对于新的测试样本可得其可能属于一类的概率，然后根据给定的阈值（如0.5），实施分类决策。**

- 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

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

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

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

- 多类逻辑回归

多类逻辑回归的似然函数为

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

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

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

$\boldsymbol{\theta}$通过最大似然或最小化交叉熵进行优化。优化目标中包含非线性函数，通常使用基于梯度的迭代优化。

目录

- 线性回归
- 贝叶斯线性回归
- 逻辑回归
- 贝叶斯逻辑回归（参考内容）

已知观测数据$X = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]^T$，$\mathbf{y} = [y_1, y_2, \cdots, y_N]^T$，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

$$p(y_* = 1 | \mathbf{x}_*) \approx \int \sigma(\boldsymbol{\theta}^T \mathbf{x}_*) q(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

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

$$p(\boldsymbol{\theta}) = \mathrm{N}\,(\boldsymbol{\theta}\,|\,\mathbf{m}_0, S_0).$$

则逻辑回归的似然为

$$p(y_* = 1|\,\mathbf{x}_*) \approx \int \sigma(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}_*) q(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

- 另一方面，预测分布 $p(y_* = 1|\,\mathbf{x}_*) \approx \int \sigma(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}_*) q(\boldsymbol{\theta}) d\boldsymbol{\theta}$ 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。 $p(\mathbf{y}\,|\,X, \boldsymbol{\theta}) = \prod_{i=1}^{N} p(y_i = 1|\,\mathbf{x}_i)^{y_i} \left(1 - p(y_i = 1|\,\mathbf{x}_i)\right)^{1-y_i}$.

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$\mathbf{\theta}_0$为均值的高斯分布$q(\mathbf{\theta})$，作为真实后验的近似分布：

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

其中，均值$\mathbf{\theta}_0$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$-\ln p(\mathbf{\theta}\,|\,X,\mathbf{y})$的Hessian矩阵（附录C）在$\mathbf{\theta}=\mathbf{\theta}_0$处的逆，即

$$S_N = \left(-\nabla\nabla \ln p(\mathbf{\theta}\,|\,X,\mathbf{y})\,|_{\mathbf{\theta}=\mathbf{\theta}_0}\right)^{-1}$$

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 $p(\boldsymbol{\theta}) = \mathrm{N}(\boldsymbol{\theta} \mid m_0, S_0)$，其中$m_0$和$S_0$是超参数。

后验分布 $p(\boldsymbol{\theta} \mid X, \mathbf{y}) \propto p(\boldsymbol{\theta}) p(\mathbf{y} \mid X, \boldsymbol{\theta})$ 。将先验概率 $p(\boldsymbol{\theta}) = \mathrm{N}(\boldsymbol{\theta} \mid \mathbf{m}_0, S_0)$ 和逻辑回归的似然函数

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

带入贝叶斯公式可得

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

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：

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

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

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

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

$$\sigma(\mathbf{\theta}^{\mathrm{T}}\mathbf{x}_*)\approx\int\delta(a-\sigma(\mathbf{\theta}^{\mathrm{T}}\mathbf{x}_*))\sigma(a)da$$

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

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

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

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

预测分布可以表示为

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

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

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

高斯分布和逆probit函数相乘后的积分还是一个逆probit函数，即

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

由此可获得最终预测概率为

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

其中$\mu_a = \theta_{\mathrm{map}}^{\mathrm{T}}\mathbf{x}_*$，$\sigma_a^2 = \mathbf{x}_*^{\mathrm{T}} S_N \mathbf{x}_*$。

对应于$p(y=1|\mathbf{x}_*)=0.5$的决策边界由$\mu_a = 0$给出。

1. 线性回归 实战编程案例

https://mp.weixin.qq.com/s?__biz=MzUzODYwMDAzNA==&mid=2247484999&idx=1&sn=5df94945ecd3a56d1191058a1bc2dd82&chksm=fad4714acda3f85cb7eada47469466eee74a6d8cb30e10b01435459876e7f9ce5dc5e3454531&scene=21#wechat_redirect

https://cloud.tencent.com/developer/article/1383880

https://cloud.tencent.com/developer/article/1389441?from=article.detail.1383880

2. 逻辑回归 实战编程案例

https://cloud.tencent.com/developer/article/1014959

https://cloud.tencent.com/developer/article/1506730

https://cloud.tencent.com/developer/article/1373192

3. 分类与回归 评估指标

https://cloud.tencent.com/developer/article/1356921

谢谢大家！