5-二元Logistic回归算法

前面我们讲述的线性回归算法要求因变量是连续变量，但很多情况下因变量是离散的，而非连续的。例如，预测下雨的概率，是下雨还是不下雨；预测一笔贷款业务的资产质量，包括正常、关注、次级、可疑、损失等。Logistic回归算法可以有效地解决这一问题，包括二元Logistic回归算法、多元Logistic回归算法等。当因变量只有两种取值，则使用二元Logistic回归算法来解决问题；当因变量有多种取值，则使用多元Logistic回归算法来解决问题。

目录

CONTENTS

PART 01

二元Logistic回归的基本原理

二元Logistic回归算法的基本原理

在线性回归算法中，我们假定因变量为连续定量变量，但在很多情况下，因变量只能取二值（0,1），比如是否满足某一特征等。因为一般回归分析要求因变量呈现正态分布，并且各组中具有相同的方差—协方差矩阵，所以直接用来为二值因变量进行回归估计是不恰当的。这时候就可以用到本节介绍的二元Logistic回归算法。

二元Logistic回归算法的基本原理是考虑因变量（0,1）发生的概率，用发生概率除以没有发生概率再取对数。通过这一变换改变了“回归方程左侧因变量估计值取值范围为0~1，而右侧取值范围是无穷大或者无穷小”这一取值区间的矛盾，也使得因变量和自变量之间呈线性关系。当然，正是由于这一变换，使得Logistic 回归自变量系数不同于一般回归分析自变量系数，而是模型中每个自变量概率比的概念。

二元Logistic回归算法的基本原理

Logistic回归系数的估计通常采用最大似然法，最大似然法的基本思想是先建立似然函数与对数似然函数，再通过使对数似然函数最大，求解相应的系数值，所得到的估计值称为系数的最大似然估计值。

Logistic 模型的公式如下：

$$\ln\frac{p}{1-p}=\alpha+X\beta+\varepsilon$$

其中，p 为发生的概率，$\alpha\begin{pmatrix}\alpha_1\\ \alpha_2\\ \vdots\\ \alpha_n\end{pmatrix}$ 为模型的截距项，$\beta\begin{pmatrix}\beta_1\\ \beta_2\\ \vdots\\ \beta_n\end{pmatrix}$ 为待估计系数，

$X=\begin{pmatrix}x_{11} & x_{12} & \cdots & x_{1k}\\ x_{21} & x_{22} & \cdots & x_{2k}\\ \vdots & \vdots & \ddots & \vdots\\ x_{n1} & x_{n2} & \cdots & x_{nk}\end{pmatrix}$ 为自变量，$\varepsilon\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{pmatrix}$ 为误差项。

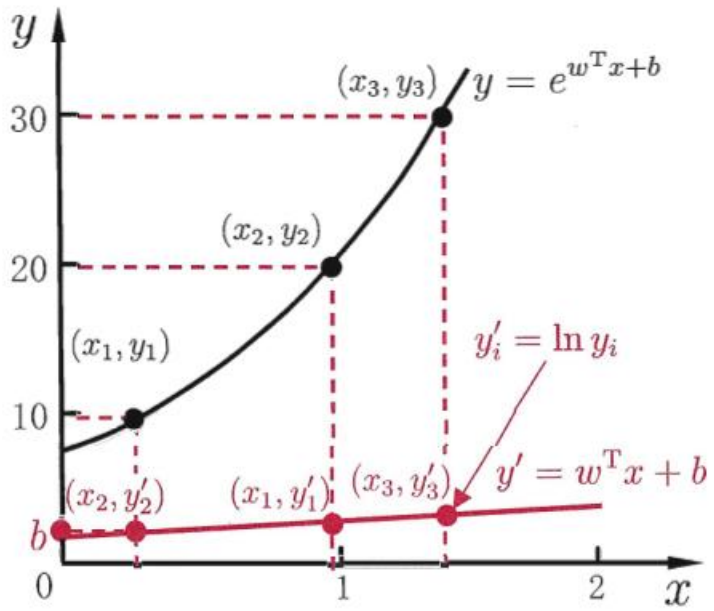
二元Logistic回归算法的基本原理

通过公式也可以看出，Logistic 模型实质上是建立了因变量发生的概率和自变量之间的关系，回归系数是模型中每个自变量概率比的概念。

所以，与线性回归算法不同的是，二元 Logistic 回归算法中所估计的参数不能被解释为特征变量对响应变量的边际效应，系数估计值 $\hat{\beta}_i$ 衡量的是因变量取 1 的概率会因自变量变化而如何变化，$\hat{\beta}_i$ 为正数表示自变量增加会引起因变量取 1 的概率提高、取 0 的概率降低，$\hat{\beta}_i$ 为负数则表示自变量增加会引起因变量取 0 的概率提高、取 1 的概率降低。

当然，二元 Logistic 回归算法也有自身的适用条件：一是因变量需为二分类的分类变量，自变量可以是连续变量或分类变量；二是残差和因变量都要服从二项分布；三是自变量和 Logistic 概率是线性关系；四是各样本观测值相互独立。

对数线性模型(逻辑回归)



$$y = \omega^T x + b$$

$$\ln y = \omega^T x + b$$

让$exp(\omega^T x + b)$逼近y

$$y = g^{-1}(\omega^T x + b)$$

$g(x)$称为link function

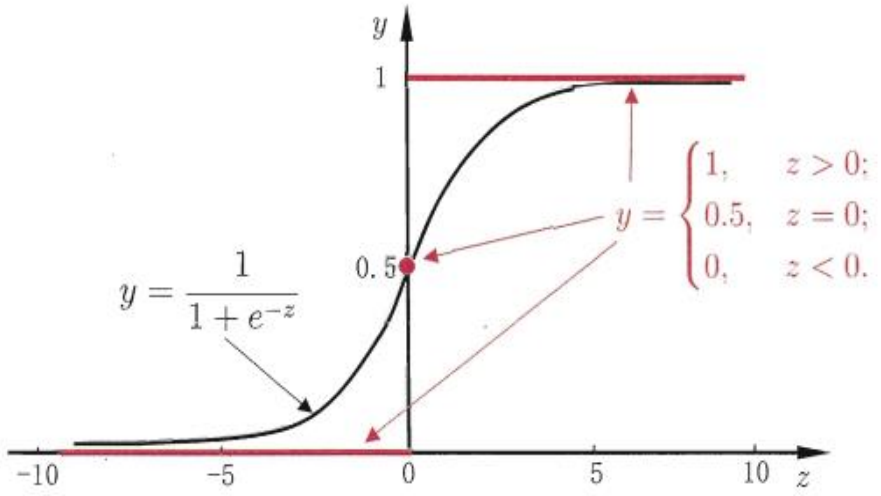
Sigmoid 函数

单位阶跃函数不连续，不能直接作用于 link function

$$y = \frac{1}{1 + e^{-z}}$$

$$y = \frac{1}{1 + e^{-(\omega^T x + b)}}$$

$$\ln\left(\frac{y}{1-y}\right) = \omega^T x + b$$



y视为样本x作为正例的可能性，则1-y便是其反例的可能性。二者的比值便被称为“几率”。

用线性回归模型的预测结果去逼近真是标记的对数几率。实际上是分类学习方法

对数几率回归参数优化

$$p(y=1|x)=\frac{\exp(\omega^T x+b)}{1+\exp(\omega^T x+b)}=h(x);$$

$$\ln\left(\frac{y}{1-y}\right)=\omega^T x+b$$

$$p(y=0|x)=\frac{1}{1+\exp(\omega^T x+b)}=1-h(x)$$

$$ln\frac{p(y=1|x)}{p(y=0|x)}=\omega^T x+b$$

综合

$$p(y|x,\omega)=(h(x))^y(1-h(x))^{1-y}$$

极大似然估计

$$\min J(\omega)=\frac{1}{m}\sum_{i=1}^{m}(-y^{(i)}\ln\left(h\left(x^{(i)}\right)\right)-\left(1-y^{(i)}\right)\ln(1-h\left(x^{(i)}\right)))$$

梯度下降法（min J）：

(repeat:)$\theta_j=\theta_j-\alpha\frac{d}{d\theta_j}J(\theta)$

$\frac{d}{d\theta_j}J(\theta)=\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right)-y^{(i)}\right)x_j^{(i)}$

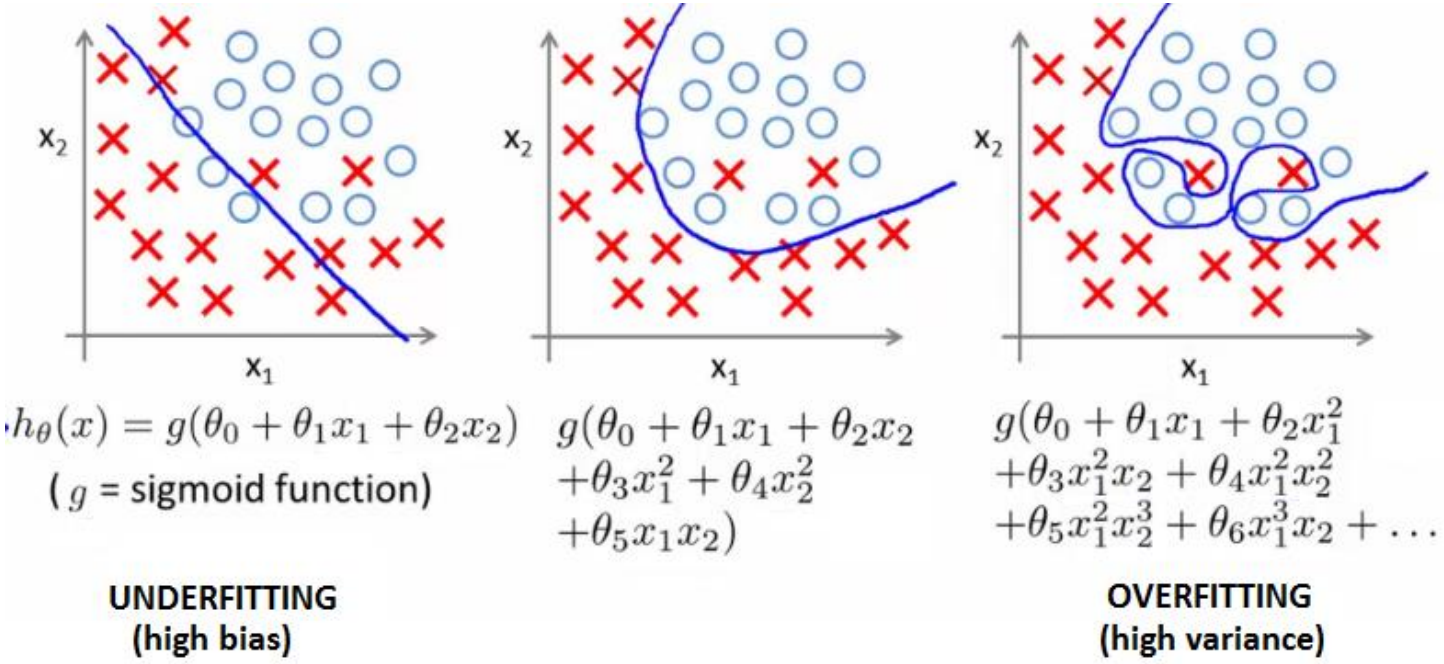
正则化

正则化的方法，就是给代价函数后面加个“惩罚项” 来降低它对数据的拟合能力

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(-y^{(i)}\ln\left(h\left(x^{(i)}\right)\right) - \left(1-y^{(i)}\right)\ln\left(1-h\left(x^{(i)}\right)\right)\right) + \lambda\sum_{j=1}^{n}\theta_j^2$$

n表示特征量的总数。

λ是正规化参数，决定了惩罚的量度。过高会欠拟合，过小无法解决过拟合



正则化的梯度下降

之前：

$$
\begin{aligned}
&Repeat\ \{ \\
&\quad \theta_j := \theta_j - \frac{\alpha}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)} \\
&\}
\end{aligned}
$$

正则化之后：

$$
\begin{aligned}
&\text{Repeat}\ \{ \\
&\quad \theta_0 := \theta_0 - \alpha\ \frac{1}{m}\ \sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_0^{(i)} \\
&\quad \theta_j := \theta_j - \alpha\left[\left(\frac{1}{m}\ \sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}\right) + \frac{\lambda}{m}\theta_j\right] \qquad j \in \{1,2...n\} \\
&\}
\end{aligned}
$$

Logistic回归的优缺点

该模型依旧是线性的，只有当数据线性可分时（例如，数据可被某决策平面完全分离），这一算法才会有很好的表现。

优点：预测结果是界于0和1之间的概率，而算法也能通过正则化以避免过拟合。逻辑模型很容易通过随机梯度下降来更新数据模型。

缺点：只能处理二分类问题，必须线性可分。非线性特征需要转化。预测结果呈“S”型，因此从log(odds)向概率转化的过程是非线性的，在两端随着log(odds)值的变化，概率变化很小，边际值太小，slope太小，而中间概率的变化很大，很敏感。 导致很多区间的变量变化对目标概率的影响没有区分度，无法确定阈值。

PART 02

数据准备

数据准备

本节我们用于分析的数据是XX银行XX省分行的700个对公授信客户的信息数据。这700 个对公授信客户是以前曾获得贷款的客户，包括存量授信客户和已结清贷款客户。在数据文件中共有9个变量，“V1~V9”，分别代表“征信违约记录”“资产负债率”、“行业分类”、“实际控制人从业年限”、“企业经营年限”、“主营业务收入”、“利息保障倍数”、“银行负债”、“其他渠道负债”。由于客户信息数据既涉及客户隐私，也涉及商业机密，所以进行了适当的脱密处理，对于其中的部分数据也进行了必要的调整。

针对“V1征信违约记录”，分别用0、1来表示未违约、违约。

针对“V3行业分类”，分别用1、2、3、4、5来表示“制造业”“批发零售业”“建筑业、房地产与基础设施”“科教文卫”“农林牧渔业”。

数据准备

我们要研究的是对公授信客户违约的影响因素，或者说那些特征可以影响对公客户的信用状况，进而可以提出针对性的风险防控策略，所以把响应变量设定为“V1征信违约记录”，将其他变量作为特征变量，具体包括“V2资产负债率”、“V3行业分类”、“V4实际控制人从业年限”、“V5企业经营年限”、“V6主营业务收入”、“V7利息保障倍数”、“V8银行负债”、“V9其他渠道负债”。

数据准备

模型构建的基本思路

商业银行对公客户违约问题，本质上还是一种对客户的分类问题。基本逻辑是把客户是否守约作为响应变量，这一响应变量在测量方式上属于二分类变量，把客户分为“违约”和“守约”两类；把客户特征作为特征变量，客户特征包括客户的经营能力、盈利能力、偿债能力、发展潜力、现有负债及担保情况等等，这些特征变量既可以是生产经营指标、财务指标等连续变量，也可以是是否对外担保、是否存在历史违约记录等分类变量。

当然，这一概念可以扩展，比如针对单笔债项进行预测，响应变量可能是多分类的，比如按资产质量五级分类，正常、关注、次级、可疑、损失等，这就是前面所述的需要使用多元 Logistic 回归算法或其他算法解决的问题了。特征变量也可能会扩展到除客户资质之外的影响因子，比如针对贸易融资业务，因为需要考虑贸易背景的真实性、贸易融资的自偿性，那么除了考虑借款人，还应该充分考虑交易对手的资质、担保货品的特征、应收账款的特征、供应链整体运营状况等因子的影响。本例中为讲解方便，采用了“V2 资产负债率”等 8 个特征变量，实务中大家需根据实际业务情况及数据的可获得性、便利程度等因素灵活选取特征变量。

载入分析所需要的模块和函数

在进行分析之前，我们首先载入分析所需要的模块和函数，读取数据集并进行观察。

示例

参阅教材内容

数据读取及观察

大家首先需要将本书提供的数据文件放入安装python的默认路径位置，并从相应位置进行读取。

示例

参阅教材内容

PART 03

描述性分析

描述性分析

本节我们针对各变量开展描述性分析。针对连续变量，通常使用计算平均值、标准差、最大值、最小值、四分位数等统计指标的方式来进行描述性分析；针对分类变量，通常使用交叉表的方式开展分析。

交叉表分析是描述统计的一种，分析特色是将数据按照行变量、列变量进行描述统计。比如我们要针对体检结果分析高血脂和高血压情况，则可以使用交叉表分析方法将高血脂作为行变量、高血压作为列变量（当然，行列变量也可以互换），对所有被体检者生成二维交叉表格描述统计分析。

示例

参阅教材内容

PART 04

数据处理

区分分类特征和连续特征并进行处理

首先定义一个函数data_encoding()，该函数的作用是可以区分分类特征和连续特征，并对分类特征设置虚拟变量，对连续特征进行标准化处理。

示例

参阅教材内容

将样本示例全集分割为训练样本和测试样本

前面章节中我们反复提及，机器学习的主要目的是为了进行预测，为了避免模型出现“过拟合”导致泛化能力不足，需要将样本示例全集分割为训练样本和测试样本进行机器学习。

示例

参阅教材内容

PART 05

建立二元Logistic回归模型

使用statsmodels建立二元Logistic回归算法模型

一、模型估计

二、计算训练误差

三、计算测试误差

示例

参阅教材内容

使用sklearn建立二元Logistic回归算法模型

示例

参阅教材内容

特征变量重要性水平分析

在机器学习中，很多时候需要评价特征变量的重要性，或者说，在众多的特征变量中，哪些变量的贡献度较大，对于整个机器学习模型来说更加重要？

对于二元Logistic回归算法模型，其特征变量重要性水平体现为模型中回归方程的系数，在对各个变量进行标准化、有效消除变量量纲之间差距的前提下，特征变量系数的绝对值越大，其对于响应变量预测整体结果的影响就越大。或者说，特征变量重要性水平分析本质上是回归系数的一种直观化、图形化展示。

示例

参阅教材内容

绘制ROC曲线，计算AUC值

前面章节中我们讲到，ROC曲线和AUC值也是评价分类监督学习性能的重要度量指标。

示例

参阅教材内容

计算科恩kappa得分

示例

参阅教材内容

PART 06

习　题

习题

习题部分我们用于分析的数据是数据5.2。

1、载入分析所需要的库和模块

2、数据读取及观察。

3、描述性分析。

（1）针对数据集中各变量计算平均值、标准差、最大值、最小值、四分位数等统计指标，针对连续变量的结果进行解读；

（2）按照V1变量的取值分组对其他变量开展描述性分析；

（3）针对分类变量“V1是否购买本次推广产品”“V3年收入水平”，使用交叉表的方式开展分析。

习题

4、数据处理。

（1）区分分类特征和连续特征并进行处理，对分类特征设置虚拟变量，对连续特征进行标准化处理；

（2）将样本示例全集分割为训练样本和测试样本，测试样本占比为30%，设定随机数种子为123，以保证随机抽样的结果可重复。

5、使用statsmodels建立二元Logistic回归算法模型。

（1）开展模型估计；

（2）计算训练误差；

（3）计算测试误差.

习题

6、使用sklearn建立二元Logistic回归算法模型

7、开展特征变量重要性水平分析

8、绘制ROC曲线，计算AUC值

9、计算科恩kappa得分。

感谢聆听

THANKS