6-多元Logistic回归算法

当响应变量为离散变量且有多种取值时，则使用多元Logistic回归算法来解决问题。

目录

CONTENTS

PART 01

多元Logistic回归的基本原理

多元Logistic回归算法的基本原理

多元Logistic回归算法本质上是二元Logistic回归算法的拓展，用于响应变量取多个单值时的情形，如偏好选择、考核等级等。多元Logistic回归分析的基本原理同样是考虑响应变量（0,1）发生的概率，用发生概率除以没有发生概率再取对数。回归自变量系数也是模型中每个自变量概率比的概念，回归系数的估计同样采用迭代最大似然法。

多元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。

多元 Logistic 回归算法的公式为：

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

其中，p 为事件发生的概率，$\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix}$ 为模型的截距项，$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}$ 为自变量系数，

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

PART 02

数据准备

数据准备

本节我们以“数据6.1”为例进行讲解。“数据6.1”记录的是某商业银行全体员工收入档次（V1）（1为高收入，2为中收入，3为低收入）、工作年限（V2）、绩效考核得分（V3）违规操作积分（V4）和职称情况（V5）（1为高级职称，2为中级职称，3为初级职称）数据。

下面以收入档次（V1）为响应变量，以工作年限（V2）、绩效考核得分（V3）和违规操作积分（V4）为特征变量，构建多元Logistic回归算法模型。

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

描述性分析及图形绘制

描述性分析及图形绘制

描述性分析

绘制条形图

绘制箱线图

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

建立多元Logistic回归模型

建立多元Logistic回归算法模型

一、模型估计

二、模型性能分析

示例

参阅教材内容

PART 06

习　题

习题

继续使用“数据6.1”数据文件，以职称情况（V5）为响应变量，以工作年限（V2）、绩效考核得分（V3）和违规操作积分（V4）为特征变量，构建多元Logistic回归算法模型。

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）针对连续变量，计算平均值、标准差、最大值、最小值、四分位数等统计指标；

（2）绘制职称情况（V5）变量的条形图。

（3）绘制职称情况（V5）与工作年限（V2）、绩效考核得分（V3）和违规操作积分（V4）的箱线图

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、建立多元Logistic回归算法模型。

（1）开展模型估计；

（2）开展模型性能分析。

感谢聆听

THANKS