逻辑斯蒂回归

logistic regression

目录

CONTENTS

01

logistic回归

4 定 义 描 述

定义

当因变量的值是两个离散值时，我们称这是一个二分类模型。最基础二分类模型就是logistic 回归。

Simmons商店的案例

Simmons商店正在使用的一种直接邮寄广告的促销手段。他们计划印刷5000份昂贵的彩色商品目录，并且每份商品目录还赠送一张商家优惠券。因为商品目录价格昂贵，所以Simmons只愿意将商品目录寄送给那些最有可能使用优惠券并购买商品的顾客。

根据客人的年消费支出与是否拥有Simmons信用卡两个特征，来计算顾客使用优惠券的概率。表格中记录了部分去年消费者的特征与是否使用优惠券的数据。

Simmons样本数据

Customer	Spending	Card	Coupon	
	1	2.291	1	0
	2	3.215	1	0
	3	2.135	1	0
	4	3.924	0	0
	5	2.528	1	0
	6	2.473	0	1
	7	2.384	0	0

5 logistic回归方程

logistic回归方程由一个因变量和一个或一个以上的自变量组成

logistic回归方程

$$E(y)=\frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}$$

因变量的取值为1或0，那么$E(y)$给出的是在给定一组自变量值的情形下，有关$y=1$的概率。

logistic回中的$E(y)$被解释为概率

$$E(y)=P(y=1|x_1,x_2,\ldots,x_p)$$

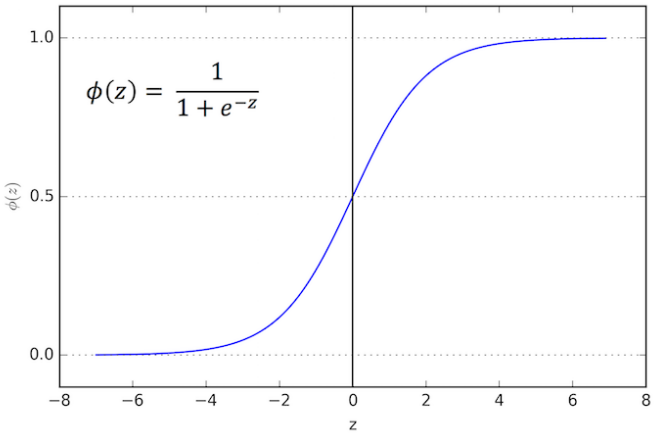
6 logistic回归方程

如果我们定义$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$，那么logistic回归方程可以写成

$$E(y) = \frac{e^z}{1 + e^z}$$

进一步转化，等号右边公式分子分母都除以e^z，可得：

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



- $E(y)$的取值在0-1之间，这从理论上保证了logistic适合做概率模型
- 随着z值的增大，输出的结果也就越大，越来越接近1
- 在S曲线的中间阶段，值变化的速度很快，但是在两端趋于平稳缓和

7 估计logistic回归方程

logistic回归的求解通常使用的是凸优化理论的梯度法和牛顿法，这超出了我们课程要学习的范围。但是我们可以使用PYTHON很方便的求解

估计的logistic回归方程

$$\hat{y} = P(y = 1|x_1, x_1, \dots, x_1)\text{的估计} = \frac{e^{b_0+b_1x_1+b_2x_2+\cdots+b_px_p}}{1 + e^{b_0+b_1x_1+b_2x_2+\cdots+b_px_p}}$$

给出了$y = 1$的概率的估计

8 案例求解

回到Simmons商店的例子。我们定义的变量如下：

$$y=\begin{cases}0，如果顾客没有使用了优惠券\\ 1，如果顾客使用了优惠券\end{cases}$$

$x_1=$ 在Simmons商店的年消费支出

$$x_2=\begin{cases}0，如果顾客没有信用卡\\ 1，如果顾客有信用卡\end{cases}$$

于是，我们有两个自变量的logistic回归方程为

$$E(y)=\frac{e^{\beta_0+\beta_1x_1+\beta_2x_2}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2}}$$

9 案例求解

```
import  statsmodels.api as sm
import pandas as pd

df = pd.read_csv("Simmons.csv")
x = df[['Spending','Card']]
y = df['Coupon']

x = sm.add_constant(x)
model  = sm.Logit(y,x).fit()
print(model.summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.604869
         Iterations 5
                         Results: Logit
================================================================
Model:              Logit            No. Iterations:   5.0000
Dependent Variable: Coupon           Pseudo R-squared: 0.101
Date:               2020-01-09 21:18 AIC:              126.9739
No. Observations:   100              BIC:              134.7894
Df Model:           2                Log-Likelihood:   -60.487
Df Residuals:       97               LL-Null:          -67.301
Converged:          1.0000           Scale:            1.0000
----------------------------------------------------------------
             Coef.   Std.Err.     z     P>|z|    [0.025   0.975]
----------------------------------------------------------------
const       -2.1464    0.5772  -3.7183  0.0002  -3.2778  -1.0150
Spending     0.3416    0.1287   2.6551  0.0079   0.0894   0.5938
Card         1.0987    0.4447   2.4707  0.0135   0.2271   1.9703
================================================================
```

利用PYTHON进行总体参数进行估计，于是得到logistic回归方程为：

$$\hat{y} = \frac{e^{-2.146+0.342x_1+1.099x_2}}{1+e^{-2.146+0.342x_1+1.099x_2}}$$

我们可以对新用户进行预估：

假设一位顾客去年消费2000美元并且拥有信用卡，我们将 $x_1 = 2$，$x_2 = 1$代入公式，得到：

$$\hat{y} = \frac{e^{-1.642}}{1+e^{-1.642}} = 0.1882$$

结论：对于这类特定的顾客，他们使用优惠券的概率大约为0.19

02

解释logistic回归

11 机 会 比 率

有利于一个事件发生的机会比：事件将要发生的概率与该事件将不会发生的概率比。
在自变量的一组特定值已知时，有利于事件 y = 1 发生的机会比计算公式：

$$\text{机会比} = \frac{P(y=1|x_1, x_2, \dots, x_p)}{P(y=0|x_1, x_2, \dots, x_p)} = \frac{P(y=1|x_1, x_2, \dots, x_p)}{1 - P(y=1|x_1, x_2, \dots, x_p)}$$

机会比率：度量了当一组自变量中只有一个自变量增加了一个单位时对机会比的影响。

机会比率

$$\text{机会比率} = \frac{odds_1}{odds_0}$$

12 机 会 比 率

将去年消费2000美元并且拥有信用卡的顾客，即$x_1 = 2$，x_2 =1，使用优惠券的的机会比，与去年消费2000美元但是没有信用卡的顾客，即$x_1 = 2$，$x_2 = 0$，使用优惠券的的机会比进行比较。我们感兴趣的是解释自变量x_2增加一个单位的影响。

$$odds_1 = \frac{P(y = 1|x_1 = 2, x_2 = 1)}{1 - P(y = 1|x_1 = 2, x_2 = 1)} = \frac{0.4102}{1 - 0.4102} = 0.6956$$

$$odds_0 = \frac{P(y = 1|x_1 = 2, x_2 = 0)}{1 - P(y = 1|x_1 = 2, x_2 = 0)} = \frac{0.1881}{1 - 0.1881} = 0.2318$$

估计的机会比率是

$$\frac{0.6956}{0.2318} = 3.00$$

13 机会比率解释

估计的机会比率是

$$\frac{0.6956}{0.2318} = 3.00$$

解释：

- 去年消费支出为2000美元并且拥有Simmons信用卡的顾客使用优惠券的机会比，是去年消费支出为2000美元但没有Simmons信用卡的顾客使用优惠券的机会比的3倍
- **其它的自变量取何值对于计算某一自变量的机会比率没有任何影响**，即拥有Simmons信用卡的顾客使用优惠券的机会比，是没有拥有Simmons信用卡的顾客使用优惠券的机会比的3倍

14 连续值机会比率

连续型自变量的估计的机会比率

考虑的问题是：当自变量增加一个或者超过一个单位时机会比的变化情况。例如，去年消费支出为5000美元的顾客使用优惠券的估计的机会比，是去年消费支出为2000美元的顾客使用优惠券的估计的机会比的多少倍呢？

15 连续值机会比率

在一个变量的机会比率和它所对应的回归系数之间存在一个唯一的关系。在logistic回归方程中，每一个自变量都能表示如下形式：

$$\textbf{机会比率} = e^{\beta_i}$$

x_2的估计的机会比率是：

$$e^{b_2} = e^{1.099} = 3.00$$

同样的，x_1的估计的机会比率是：

$$e^{b_1} = e^{0.342} = 1.407$$

16 连续值机会比率

当一个自变量变化一个单位，而其他所有的自变量都保持不变时，一个自变量的机会比率描述了该自变量机会比的变化。

回到案例中

去年消费支出为5000（x_1=5）美元的顾客使用优惠券的估计的机会比，与去年消费支出为2000（x_1=2）美元的顾客使用优惠券的估计的机会比进行进行比较。这种情形下，$c = 5 - 3 = 2$，对应的机会比率是：

$$e^{cb_1} = e^{3*0.342} = \mathbf{2.79}$$

结论

- 去年消费支出为5000美元的顾客使用优惠券的估计的机会比，是去年消费支出为2000美元的顾客使用优惠券的估计的机会比的多少2.79倍。
- 换句话说，对于一个去年消费支持增加3000美元的顾客而言，使用优惠券的估计的机会比率是2.79

总结logistic回归

18 回 归 总 结

回归定义

当因变量的值是两个离散值时，我们称这是一个二分类模型。最基础二分类模型就是logistic 回归

回归方程

$$E(y) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}$$

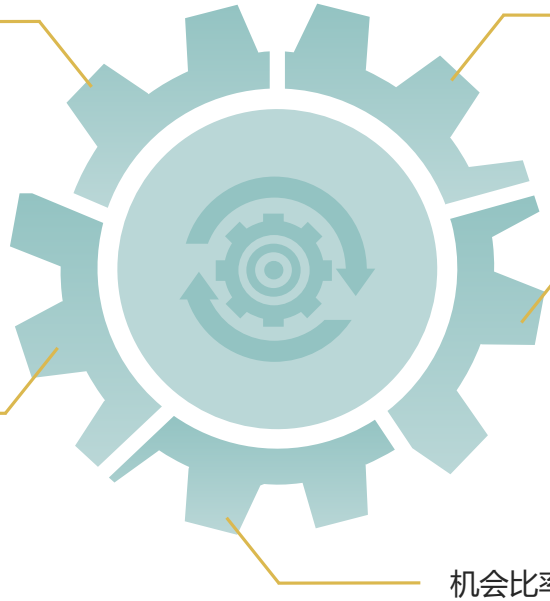
概率解释

$$E(y) = P(y = 1|x_1, x_1, \ldots, x_1)$$

统计模型

这节课涉及讲解的是统计意义上的logistic回归。它其实还在更深入的机器学习领域有着更广泛的应用！

参考SOTON《数据挖掘课程》



机会比率

机会比率 $= e^{\beta_i}$

度量了当一组自变量中只有一个自变量增加了一个单位时对机会比的影响