Logistic回归模型

讲师：刘顺祥

课程目标

1. 理解Logistic回归模型的系数求解过程
2. 理解Logistic回归模型的系数含义
3. 熟系几个常见的模型评估方法
4. 掌握Logistic回归模型的应用实操

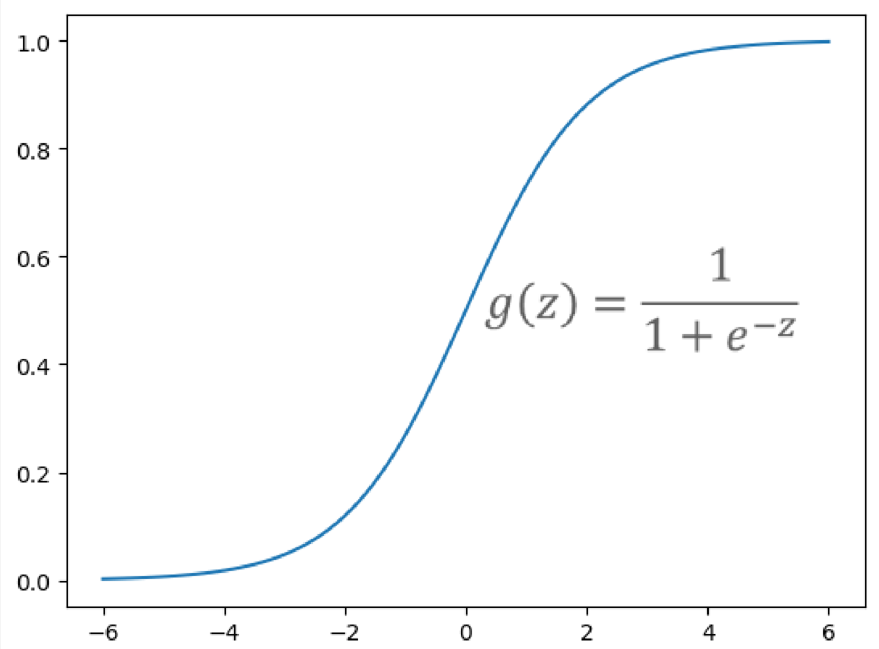
Logistic回归模型

假定线下回归模型为：$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$

则Logit变换为：$g(z) = \dfrac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}} = h_\beta(X)$

上式中的$h_\beta(X)$也被称为Logistic回归模型，它是将线性回归模型的预测值经过非线性的Logit函数转换为[0,1]之间的概率值。

Logistic回归模型



其中，$z\in(-\infty,+\infty)$。当z趋于正无穷大时，e^{-z}将趋于0，进而导致$g(z)$逼近于1；

相反，当z趋于负无穷大时，e^{-z}会趋于正无穷大，最终导致$g(z)$逼近于0；

当z=0时，e^{-z}=1，所以得到$g(z)$=0.5；

Logistic回归模型

模型变换

条件概率，y取值为1时的概率：$P(y=1|X;\beta)=h_\beta(X)=p$

条件概率，y取值为0时的概率：$P(y=0|X;\beta)=1-h_\beta(X)=1-p$

则两个概率的商为：$\frac{p}{1-p}=\frac{h_\beta(X)}{1-h_\beta(X)}$

$$=\left(\frac{\frac{1}{1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}}}{1-\frac{1}{1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}}}\right)$$

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

$$=e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}$$

Logistic回归模型

EDU
CSDN学院 IT实战派

参数求解过程

$$P(y=1|X;\beta)=h_\beta(X)=p$$

$$P(y=0|X;\beta)=1-h_\beta(X)=1-p$$

等价

$$P(y|X;\beta)=h_\beta(X)^y\times\left(1-h_\beta(X)\right)^{1-y}$$

构造似然函数

$$L(\beta)=P(\vec{y}|X;\beta)$$

$$=\prod_{i=1}^{n}P\left(y^{(i)}|x^{(i)};\beta\right)$$

$$=\prod_{i=1}^{n}h_\beta\left(x^{(i)}\right)^{y^{(i)}}\times\left(1-h_\beta\left(x^{(i)}\right)\right)^{1-y^{(i)}}$$

Logistic回归模型

参数求解过程

似然函数对数化

$$l(\beta) = log(L(\beta)) = log\left(\prod_{i=1}^{n} h_\beta(x^{(i)})^{y^{(i)}} \times \left(1 - h_\beta(x^{(i)})\right)^{1-y^{(i)}}\right)$$

$$= \sum_{i=1}^{n} log\left(h_\beta(x^{(i)})^{y^{(i)}} \times \left(1 - h_\beta(x^{(i)})\right)^{1-y^{(i)}}\right)$$

$$= \sum_{i=1}^{n} \left(y^{(i)} log\left(h_\beta(x^{(i)})\right) + (1 - y^{(i)}) log\left(1 - h_\beta(x^{(i)})\right)\right)$$

参数求解过程

梯度下降

$$J(\beta) = -l(\beta)$$

$$= -\sum_{i=1}^{n}\left(y^{(i)}log\left(h_\beta(x^{(i)})\right) + (1-y^{(i)})log\left(1-h_\beta(x^{(i)})\right)\right)$$

对每一个未知参数β_j做梯度下降

$$\beta_j := \beta_j - \alpha\frac{\partial J(\beta)}{\partial \beta_j}, \quad (j = 1,2,\ldots p)$$

其中，α为学习率，也称为参数β_j变化的步长，通常步长可以取0.1,0.05,0.01等。如果设置的α过小，会导致β_j变化微小，需要经过多次迭代，收敛速度过慢；但如果设置的α过大，就很难得到理想的β_j值，进而导致目标函数可能是局部最小。

Logistic回归模型

参数含义的解释

假设影响是否患癌的因素有性别和肿瘤两个变量，通过建模可以得到对应的系数β_1和β_2，则Logistic回归模型可以按照事件发生比的形式改写为：

$$odds = \frac{p}{1-p} = e^{\beta_0+\beta_1 Gender+\beta_2 Volum}$$

$$= e^{\beta_0} \times e^{\beta_1 Gender} \times e^{\beta_2 Volum}$$

Logistic回归模型

参数含义的解释

分别以性别变量和肿瘤体积变量为例，解释系数β_1和β_2的含义。假设性别中男用1表示，女用0表示，则：

$$\frac{odds_1}{odds_0} = \frac{e^{\beta_0} \times e^{\beta_1 \times 1} \times e^{\beta_2 Volum}}{e^{\beta_0} \times e^{\beta_1 \times 0} \times e^{\beta_2 Volum}} = e^{\beta_1}$$

所以，性别变量的发生比率为e^{β_1}，表示男性患癌的发生比约为女性患癌发生比的e^{β_1}倍。

参数含义的解释

对于连续型的自变量而言，参数解释类似，假设肿瘤体积为$Volum_0$，当肿瘤体积增加1个单位时，体积为$Volum_0+1$，则：

$$\frac{odds_{Volum_0+1}}{odds_{Volum_0}}=\frac{e^{\beta_0}\times e^{\beta_1 Gender}\times e^{\beta_2(Volum_0+1)}}{e^{\beta_0}\times e^{\beta_1 Gender}\times e^{\beta_2 Volum_0}}=e^{\beta_2}$$

所以，在其他变量不变的情况下，肿瘤体积每增加一个单位，将会使患癌发生比变化e^{β_2}倍。

混淆矩阵

	实际值			
预测值		良性--0	恶性--1	
	良性--0	A, True Negative	B, False Negtive	A+B, Predict Negtive
	恶性--1	C, False Positive	D, True Positive	C+D, Predict Positive
		A+C, Acture Negtive	B+D ,Acture Positive	

A：表示正确预测负例的样本个数，用TN表示。

B：表示预测为负例但实际为正例的个数，用FN表示。

C：表示预测为正例但实际为负例的个数，用FP表示。

D：表示正确预测正例的样本个数，用TP表示。

准确率：表示正确预测的正负例样本数与所有样本数量的比值，即(A+D)/(A+B+C+D)。

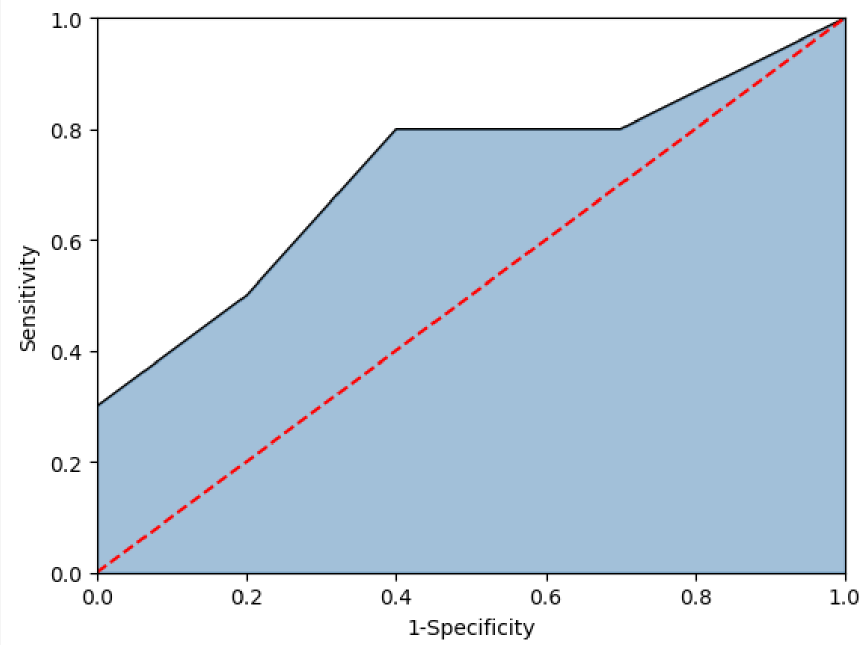
正例覆盖率：表示正确预测的正例数在实际正例数中的比例，即D/(B+D) 。

负例覆盖率：表示正确预测的负例数在实际负例数中的比例，即A/(A+C)。

正例命中率：表示正确预测的正例数在预测正例数中的比例，即D/(C+D)，

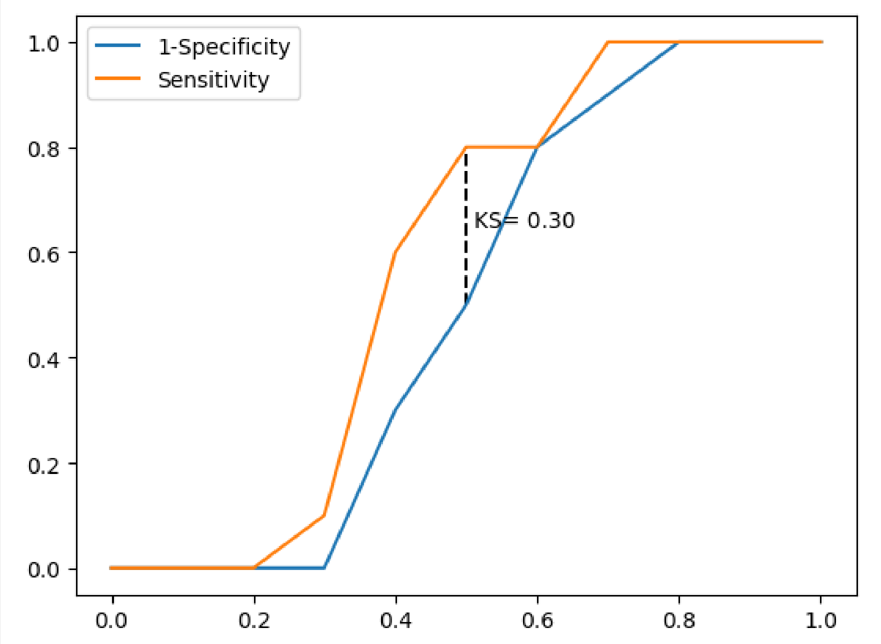
模型评估

ROC曲线



图中的红色线为参考线，即在不使用模型的情况下，Sensitivity 和 1-Specificity 之比恒等于 1。通常绘制ROC曲线，不仅仅是得到左侧的图形，更重要的是计算折线下的面积，即图中的阴影部分，这个面积称为AUC。在做模型评估时，希望AUC的值越大越好，通常情况下，当AUC在0.8以上时，模型就基本可以接受了。

KS曲线



图中的两条折线分别代表各分位点下的正例覆盖率和1-负例覆盖率，通过两条曲线很难对模型的好坏做评估，一般会选用最大的KS值作为衡量指标。KS的计算公式为：KS= Sensitivity-(1- Specificity)= Sensitivity+ Specificity-1。对于KS值而言，也是希望越大越好，通常情况下，当KS值大于0.4时，模型基本可以接受。

Logistic回归模型

函数说明

LogisticRegression(tol=0.0001, fit_intercept=True,class_weight=None, max_iter=100)

tol：用于指定模型跌倒收敛的阈值

fit_intercept：bool类型参数，是否拟合模型的截距项，默认为True

class_weight：用于指定因变量类别的权重，如果为字典，则通过字典的形式{class_label:weight}传递每个类别的权重；如果为字符串'balanced'，则每个分类的权重与实际样本中的比例成反比，当各分类存在严重不平衡时，设置为'balanced'会比较好；如果为None，则表示每个分类的权重相等

max_iter：指定模型求解过程中的最大迭代次数， 默认为100

Logistic回归模型

代码演示

```
# 导入第三方模块
import pandas as pd
import numpy as np
from sklearn import linear_model

# 读取数据
sports = pd.read_csv(r'C:\Users\Administrator\Desktop\Run or Walk.csv')

# 利用训练集建模
sklearn_logistic = linear_model.LogisticRegression()
sklearn_logistic.fit(X_train, y_train)
# 返回模型的各个参数
print(sklearn_logistic.intercept_, sklearn_logistic.coef_)

out:
[ 4.35613952] [[ 0.48533325  6.86221041 -2.44611637 -0.01344578 -0.1607943   0.13360777]]
```

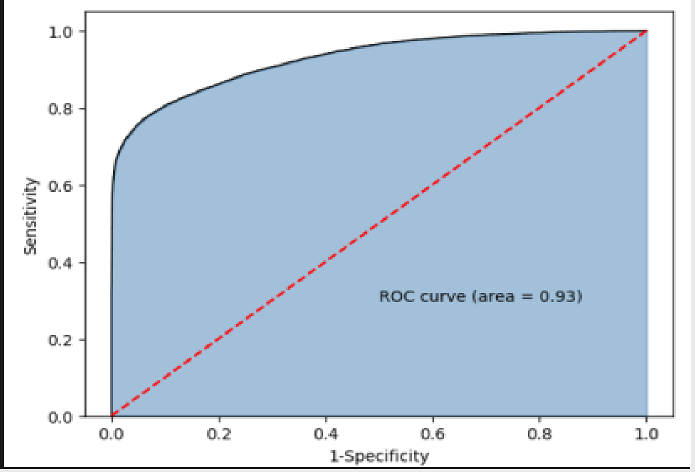
代码演示

```
# 导入第三方模块
from sklearn import metrics

# 混淆矩阵
cm = metrics.confusion_matrix(y_test, sklearn_predict, labels = [0,1])
cm

out:
array([[9971, 1120],
       [2150, 8906]], dtype=int64)
```

模型评估

代码演示

```
# y得分为模型预测正例的概率
y_score = sklearn_logistic.predict_proba(X_test)[:,1]
# 计算不同阈值下，fpr和tpr的组合值，其中fpr表示1-Specificity，tpr表示Sensitivity
fpr,tpr,threshold = metrics.roc_curve(y_test, y_score)

# 绘制面积图
plt.stackplot(fpr, tpr, color='steelblue', alpha = 0.5, edgecolor = 'black')
# 添加ROC曲线的轮廓
plt.plot(fpr, tpr, color='black', lw = 1)
# 添加对角线
plt.plot([0,1],[0,1], color = 'red', linestyle = '--')
# 显示图形
plt.show()
```



模型评估

代码演示

```
# 调用自定义函数，绘制K-S曲线
plot_ks(y_test = y_test, y_score = y_score, positive_flag = 1)
```