逻辑回归

水果分类的例子

根据水果的属性，判断该水果的种类。

mass: 水果重量
width: 水果的宽度
height: 水果的高度
color_score: 水果的颜色数值，范围0-1
fruit_name：水果类别

前19个样本是苹果
后19个样本是橙子
用这38个样本预测后四个样本对应的水果种类。

逻辑回归logistic regression

类型	模型	Y的特点	例子
线性回归	OLS、GLS（最小二乘）	连续数值型变量	GDP、产量、收入
0-1回归	logistic回归	二值变量（0-1）	是否违约、是否得病
定序回归	probit定序回归	定序变量	等级评定（优良差）
计数回归	泊松回归（泊松分布）	计数变量	每分钟车流量
生存回归	Cox等比例风险回归	生存变量（截断数据）	企业、产品的寿命

对于因变量为分类变量的情况，我们可以使用逻辑回归进行处理。
把y看成事件发生的概率，y>0.5表示发生；y<0.5表示不发生

线性概率模型

线性概率模型（Linear Probability Model，简记LPM）

直接用原来的回归模型进行回归。

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \mu_i$$

写成向量乘积形式：$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i \ (i = 1, 2, \cdots, n)$

内生性问题：y_i只能取0或者1（回归系数估计出来不一致且有偏）

$$u_i = \begin{cases} 1 - \boldsymbol{x}_i'\boldsymbol{\beta} & , y_i = 1 \\ -\boldsymbol{x}_i'\boldsymbol{\beta} & , y_i = 0 \end{cases}$$

显然$cov(x_i, u_i) \neq 0$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki}$$

预测值却可能出现$\hat{y}_i > 1$或者$\hat{y}_i < 0$的不现实情况

两点分布（伯努利分布）

事件	1	0
概率	p	1-p

在给定$\boldsymbol{x}$的情况下，考虑y的两点分布概率

$$\begin{cases} P(y = 1|\boldsymbol{x}) = F(\boldsymbol{x}, \boldsymbol{\beta}) \\ P(y = 0|\boldsymbol{x}) = 1 - F(\boldsymbol{x}, \boldsymbol{\beta}) \end{cases}$$ 注：一般$F(\boldsymbol{x}, \boldsymbol{\beta}) = F(\boldsymbol{x}_i'\boldsymbol{\beta})$

$F(\boldsymbol{x}, \boldsymbol{\beta})$称为连接函数($link\ function$),它将解释变量$x$和被解释变量$y$连接起来。

我们只需要保证$F(\boldsymbol{x}, \boldsymbol{\beta})$是定义在$[0, 1]$上的函数，就能保证$0 \leq \hat{y} \leq 1$

因为$E(y|\boldsymbol{x}) = 1 \times P(y = 1|\boldsymbol{x}) + 0 \times P(y = 0|\boldsymbol{x}) = P(y = 1|\boldsymbol{x})$

所以我们可以将$\hat{y}$可以理解为‘$y = 1$’发生的概率。

$$F(\boldsymbol{x},\boldsymbol{\beta})=S(\boldsymbol{x}_i'\boldsymbol{\beta})=\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}$$

非线性模型，使用极大似然估计方法（MLE）进行估计

$$\begin{cases}P(y=1|\boldsymbol{x})=S(\boldsymbol{x}_i'\boldsymbol{\beta})\\P(y=0|\boldsymbol{x})=1-S(\boldsymbol{x}_i'\boldsymbol{\beta})\end{cases}\Rightarrow f(y_i|\boldsymbol{x}_i,\boldsymbol{\beta})=\begin{cases}S(\boldsymbol{x}_i'\boldsymbol{\beta}) & ,y_i=1\\1-S(\boldsymbol{x}_i'\boldsymbol{\beta}), & y_i=0\end{cases}$$

写成更加紧凑的形式：

$$f(y_i|\boldsymbol{x}_i,\boldsymbol{\beta})=[S(\boldsymbol{x}_i'\boldsymbol{\beta})]^{y_i}[1-S(\boldsymbol{x}_i'\boldsymbol{\beta})]^{1-y_i}$$

取对数：$\ln f(y_i|\boldsymbol{x}_i,\boldsymbol{\beta})=y_i[S(\boldsymbol{x}_i'\boldsymbol{\beta})]+(1-y_i)[1-S(\boldsymbol{x}_i'\boldsymbol{\beta})]$

样本的对数似然函数：$\ln L(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{x})=\sum_{i=1}^{n}y_i\ln[S(\boldsymbol{x}_i'\boldsymbol{\beta})]+\sum_{i=1}^{n}(1-y_i)[1-S(\boldsymbol{x}_i'\boldsymbol{\beta}$

可以使用数值方法（梯度下降）求解这个非线性最大化的问题。

逻辑回归的推导：https://www.bilibili.com/video/av44798895/?p=45
极大使然估计：大家可参考概率论与数理统计的教材，或搜索相应视频学习

在给定$\boldsymbol{x}$的情况下，考虑y的两点分布概率

$$\begin{cases}P(y=1|\boldsymbol{x})=F(\boldsymbol{x},\boldsymbol{\beta})\\P(y=0|\boldsymbol{x})=1-F(\boldsymbol{x},\boldsymbol{\beta})\end{cases}$$

因为$E(y|\boldsymbol{x})=1\times P(y=1|\boldsymbol{x})+0\times P(y=0|\boldsymbol{x})=P(y=1|\boldsymbol{x})$

所以我们可以将$\hat{y}$可以理解为‘$y=1$’发生的概率。

$$\hat{y}_i=P(y_i=1|\boldsymbol{x})=S(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})=\frac{\exp(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})}{1+\exp(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})}=\frac{e^{\hat{\beta}_0+\hat{\beta}_1x_{1i}+\hat{\beta}_2x_{2i}+\cdots+\hat{\beta}_kx_{ki}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1x_{1i}+\hat{\beta}_2x_{2i}+\cdots+\hat{\beta}_kx_{ki}}}$$

如果$\hat{y}_i\geq 0.5$，则认为其预测的$y=1$；否则则认为其预测的$y=0$

如何确定合适的模型

把数据分为训练组和测试组，用训练组的数据来估计出模型，再用测试组的数据来进行测试。（训练组和测试组的比例一般设置为80%和20%）

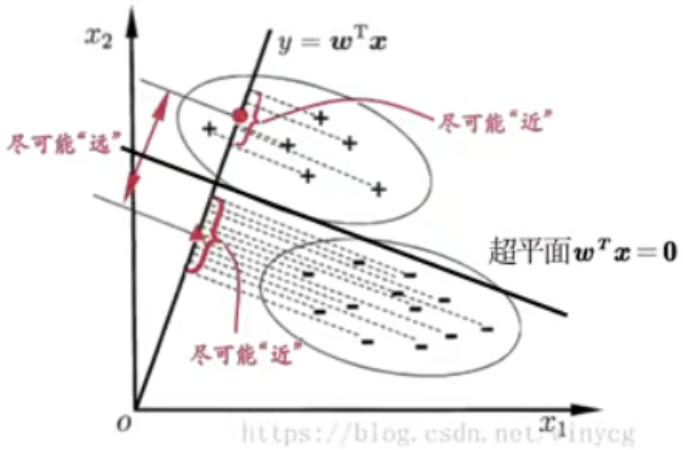
已知分类结果的水果ID为1-38，前19个为苹果，后19个为橙子。
每类水果中随机抽出3个ID作为测试组，剩下的16个ID作为训练组。
（比如：17-19、36-38这六个样本作为测试组）
比较设置不同的自变量后的模型对于测试组的预测效果。

（注意：为了消除偶然性的影响，可以对上述步骤多重复几次，最终对每个模型求一个平均的准确率，这个步骤称为交叉验证。）

Fisher线性判别分析

Fisher线性判别分析

LDA(Linear Discriminant Analysis)是一种经典的线性判别方法，又称Fisher判别分析。该方法思想比较简单:给定训练集样例，设法将样例投影到一维的直线上，使得同类样例的投影点尽可能接近和密集，异类投影点尽可能远离。



详细证明和求解步骤：https://www.bilibili.com/video/av33101528/?p=3