逻辑回归

本章主要内容

1. 从线性回归到逻辑回归
2. 逻辑回归的求解
3. 案例：用逻辑回归预测企业员工是否离职

本章主要内容

1. 从线性回归到逻辑回归
2. 逻辑回归的求解
3. 案例：用逻辑回归预测企业员工是否离职

从线性回归到逻辑回归

- 之前学习了线性回归，考虑一个问题，金融机构根据一个人的工资(x_1)、住房(x_2)、年龄(x_3)等特征预测放贷量y（借多少钱），这其实是一个多元线性回归问题，所以线性回归能够预测一个连续型的数值。
- 线性回归方程：$y = w_1x_1 + w_2x_2$+...+ w_nx_n+b
- 而现实中还有一类问题需要预测类别，比如，金融机构根据一个人的工资、住房、年龄等特征预测是/否给这个人放贷，这种预测类别的问题就要用到逻辑回归。
- 逻辑回归，Logistic Regression，逻辑回归是在线性回归的基础上，加入了Logistic函数，所以把这种回归称为Logistic（逻辑）回归，所以逻辑回归是一种广义的线性回归模型，主要用于解决分类问题。

Logistic函数

Logistic函数，也叫sigmoid函数，能够将一个实数映射到(0,1)。

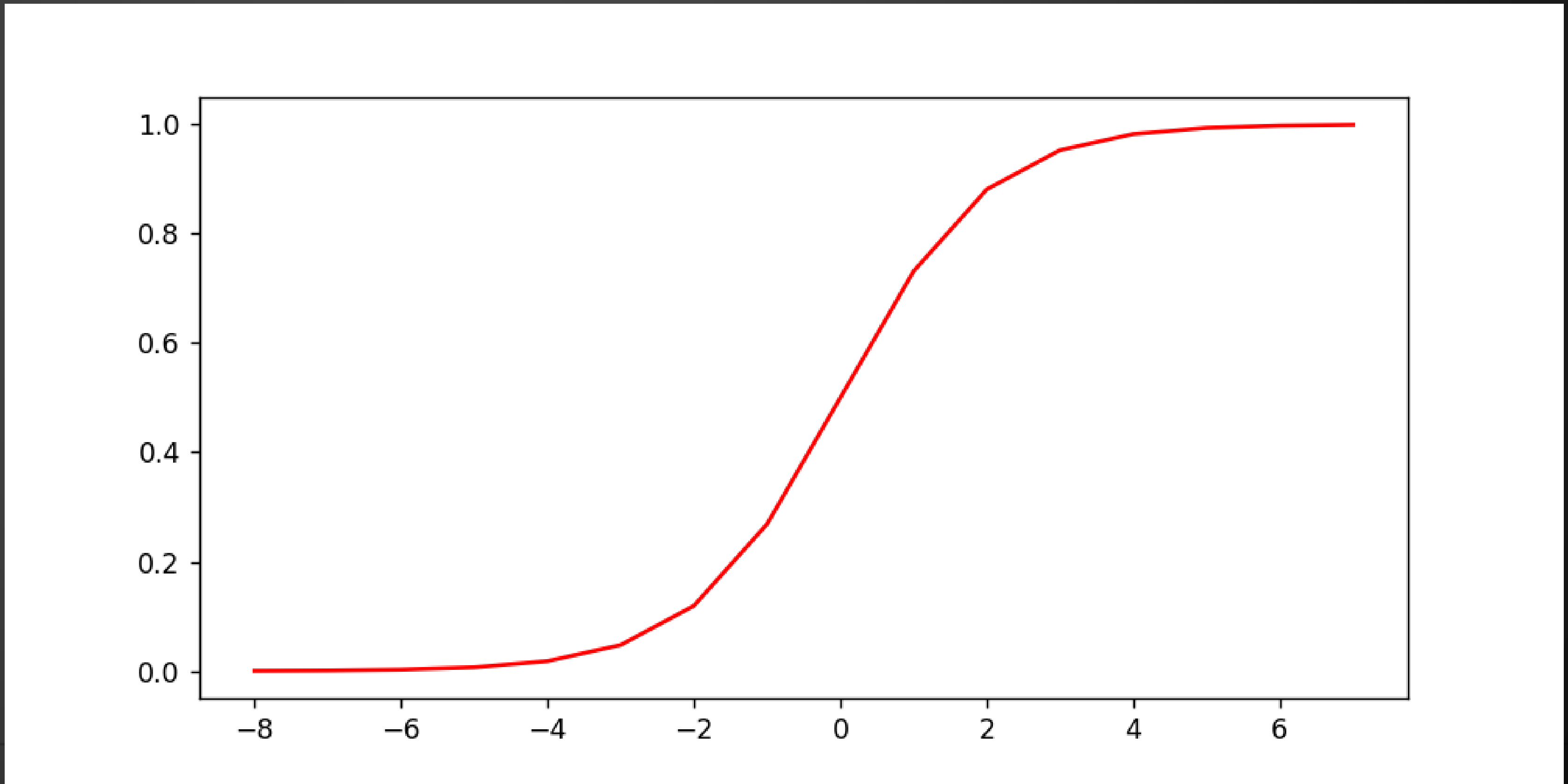
$$f(x)=\frac{1}{1+e^{-x}}, f(x)\in(0,1)$$

基本性质：

当$x\to+\infty, f(x)\to 1$

当$x\to-\infty, f(x)\to 0$

当$x=0, f(x)=\frac{1}{2}$



从线性回归到逻辑回归

特征输入：X → 线性回归：回归方程 → 输出：y（数值型）

回归方程：$y = w_1x_1 + w_2x_2 + \ldots + w_nx_n + b$

特征输入：X → 线性回归：输出y（数值型） → 将y代入sigmoid函数 → 输出：$f(y)$ (0,1)，即概率

sigmoid函数：$f(x) = \frac{1}{1+e^{-x}}, f(x) \in (0,1)$

将y代入sigmoid函数：$f(y) = \frac{1}{1+e^{-y}}$，若$f(y) > 0.5$，标记为1，若$f(y) < 0.5$，则标记为0

本章主要内容

1. 从线性回归到逻辑回归
2. 逻辑回归的求解
3. 案例：用逻辑回归预测企业员工是否离职

逻辑回归的求解

1. 将线性回归转化为概率问题：将回归方程代入sigmoid函数
2. 求解第1步得到的概率问题：极大似然估计法
3. 求解极大似然估计法得到的优化问题：梯度下降

说明：求解过程感兴趣的可以看看，不感兴趣的可以直接跳过，因为实际工作中，一般利用工具SPSS、Python等工具实现逻辑回归，这些工具可以直接给出结果。

第1步：将线性回归转化为概率问题

多元线性回归方程：

$$y = w_1x_1 + w_2x_2 + ... + w_nx_n + b$$

为了方便后面的运算，将其写成下面的形式：

$$y = w_0x_0 + w_1x_1 + w_2x_2 + ... + w_nx_n$$

其中，$w_0 = b, x_0 = 1$

接着，将其写成向量的形式

$$\vec{x} = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \vec{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$$

将其写为矩阵的乘法。

$$\begin{bmatrix} w_0, w_1, w_2, \cdots, w_n \end{bmatrix} \times \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$y = w^T x$$

将线性回归函数$y = w^T x$代入sigmoid函数，得

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

前面已经得到

$$f(y)=\frac{1}{1+e^{-w^Tx}}$$

假设预测结果只有两个类别：1和0，设$y=1$的概率为p，$y=0$的概率为$1-p$

$$\begin{cases}P(y=1|x;w)=f(y)=\frac{1}{1+e^{-w^Tx}}=p\\P(y=0|x;w)=1-p\end{cases}$$

为了方便计算，将以上式子统一为以下形式：

$$P(y|x;w)=p^y(1-p)^{1-y}$$

在上式中，当$y=1$时，结果是p，当$y=0$时，结果是$1-p$。

第2步：求解第1步得到的概率问题

假设现在采集到了m个样本，其似然函数为：

$$L(w) = P = P(y_1|x_1;w)P(y_2|x_2;w)\cdots P(y_m|x_m;w) = \Pi_{i=1}^{m} p^{y_i}(1-p)^{1-y_i}$$

但是，由于相乘的式子不好计算，所以通过取对数，将乘法变成加法，得到对数似然函数

$$lnL(w) = ln(\Pi_{i=1}^{m} p^{y_i}(1-p)^{1-y_i}) = \sum_{i=1}^{m} lnp^{y_i}(1-p)^{1-y_i}$$

$$= \sum_{i=1}^{m}(y_i lnp + (1-y_i)ln(1-p))$$

其中，$p = \frac{1}{1+e^{-w^T x}}$

根据极大似然估计法，要求这个对数似然函数的最大值，为了应用梯度下降法，需要将这个函数做一下转化。

设$l(w) = -\frac{1}{m} lnL(w)$，此时将求解最大值问题转化为了求解最小值问题，所以可以利用梯度下降来求解这个最小值。

说明：除以m表示平均损失。
于是，有：

$$\begin{aligned} l(w) &= -\frac{1}{m} lnL(w) \\ &= -\frac{1}{m} \sum_{i=1}^{m} (y_i lnp + (1 - y_i) ln(1 - p)) \end{aligned}$$

其中，$p = \frac{1}{1+e^{-w^T x}}$

第3步：求解极大似然法得到的优化问题

上一步得到，

$$l(w)=-\frac{1}{m}\sum_{i=1}^{m}(y_i lnp+(1-y_i)ln(1-p))$$

其中，$p=\frac{1}{1+e^{-w^Tx}}$

根据梯度下降法，需要计算函数$l(w)$的梯度，即求l关于$w_j, j=0,1,2,\cdots,n$的偏导数。

$$\frac{\partial l(w)}{\partial w_j}=-\frac{1}{m}\sum_{i=1}^{m}(y_i\frac{1}{p}\frac{\partial p}{\partial w_j}+(-1)(1-y_i)\frac{1}{1-p}\frac{\partial p}{\partial w_j})$$

其中，$j=0,1,2,\cdots,n$

上式中含有p关于参数w_j的导数，所以还需要求$\frac{\partial p}{\partial w_j}$

下面求$\frac{\partial p}{\partial w_j}$，$p=\frac{1}{1+e^{-w^Tx}}$

$$\begin{aligned}\frac{\partial p}{\partial w_j} &= (-1)\times\frac{1}{(1+e^{-w^Tx})^2}\times e^{-w^Tx}\times(-x_j)\\ &= \frac{1}{(1+e^{-w^Tx})^2}\times e^{-w^Tx}\times x_j\\ &= \frac{1}{1+e^{-w^Tx}}\times\frac{e^{-w^Tx}}{1+e^{-w^Tx}}\times x_j\\ &= p(1-p)x_j\end{aligned}$$

所以，$\frac{\partial p}{\partial w_j}=p(1-p)x_j$

前面已经得到,

$$\frac{\partial l(w)}{\partial w_j} = -\frac{1}{m}\sum_{i=1}^{m}(y_i\frac{1}{p}\frac{\partial p}{\partial w_j} + (-1)(1-y_i)\frac{1}{1-p}\frac{\partial p}{\partial w_j}), \quad \frac{\partial p}{\partial w_j} = p(1-p)x_j$$

于是，接着之前的计算：

$$\begin{aligned}\frac{\partial l(w)}{\partial w_j} &= -\frac{1}{m}\sum_{i=1}^{m}\{y_i\frac{1}{p}p(1-p)x_j + (-1)(1-y_i)\frac{1}{1-p}p(1-p)x_j\} \\ &= -\frac{1}{m}\sum_{i=1}^{m}\{y_i(1-p)x_j^{(i)} - (1-y_i)px_j\} \\ &= -\frac{1}{m}\sum_{i=1}^{m}(y_i - p)x_j\end{aligned}$$

将$p=\frac{1}{1+e^{-w^Tx}}$代回，得

$$\frac{\partial l(w)}{\partial w_j}=-\frac{1}{m}\sum_{i=1}^{m}(y_i-\frac{1}{1+e^{-w^Tx}})x_j$$

接下来，给参数设置一个初始值，然后通过不断更新参数使函数$l(w)$的值减小。

参数更新的表达式为：

$$w_j=w_j-\alpha\frac{\partial l(w)}{\partial w_j}=w_j+\alpha\frac{1}{m}\sum_{i=1}^{m}(y_i-\frac{1}{1+e^{-w^Tx}})x_j$$

接下来通过计算工具，例如Python进行多次迭代。

说明：统计学中，一般用SPSS来做逻辑回归，只需要将数据输入SPSS，SPSS会自动计算，并给出分析结果。

本章主要内容

1. 从线性回归到逻辑回归
2. 逻辑回归的求解
3. 案例：用逻辑回归预测企业员工是否离职

案例：用逻辑回归预测企业员工是否离职

- 某公司需要根据员工的一些数据预测员工是否会离职。
- 样本数据：一份CSV文件，共有14999个样本。9个特征变量，1个类别变量。

序号	字段名称	中文名称	字段描述
1	satisfaction_level	对公司的满意度	数值型，0-1
2	last_evaluation	最近一次考核分数	数值型，0-1
3	number_project	项目数	数值型，个数
4	average_montly_hours	平均每月工作时长	数值型，小时数
5	time_spend_company	工作年限	数值型，年
6	Work_accident	是否有过工作事故	类别，0：没有，1：有过
7	**left**	**是否离职**	**类别标签y，0：未离职，1：离职**
8	promotion_last_5years	过去5年是否晋升	类别，0：没有，1：有
9	sales	岗位类别	字符型，10种岗位类别
10	salary	薪资水平	字符型，三级：高、中、低

案例：用逻辑回归预测企业员工是否离职

主要步骤：

1. 用SPSS读取数据，并调整每个字段的数据类型
2. 数据预处理：数据去重
3. 逻辑回归建模：SPSS菜单【分析】-【回归】-【二元Logistic】
4. 解释分析结果