河北师范大学软件学院
Software College of Hebei Normal University

分类模型

第三部分 其它分类模型

LOGISTIC REGRESSION

张朝晖

2018-2019学年 20181009-1030

1.分类模块

产生式分类模型 A.贝叶斯分类模型

判别式分类模型 **线性分类模型**
- B. Fisher判别分类
- C. 感知器分类模型
- D. 大间隔分类模型（线性*SVM*）

非线性分类模型
- E. 核*SVM*（非线性*SVM*）
- F. 核Fisher判别分类
- G. 神经网络

其它分类模型
- *H*.KNN分类模型
- *I*.决策树分类模型
- *J*.*Logistic*回归
- *K*.*Softmax*回归

2.聚类模块
- *L*.K-均值聚类
- *M*.高斯混合聚类
- *N*.DBSCAN聚类
- *O*.层次聚类

3.回归模块
- *P*.KNN回归
- *Q*.回归树
- *R*.最小二乘线性回归
- *S*.岭回归
- *T*.LASSO回归

4.集成学习
- *U*.*Bagging*
- *V*.随机森林
- *W*.*Boosting*

5.特征工程
- *X*.主成分分析*(PCA)*
- …

6.评价模块 混淆矩阵（及其相关指标）、ROC曲线、交叉验证

主要内容

罗杰斯特回归(Logistic Regression)

--统计学习的经典分类方法。

将分类问题转化为有关概率的函数回归。

--二项罗杰斯特回归模型(面向两类问题)

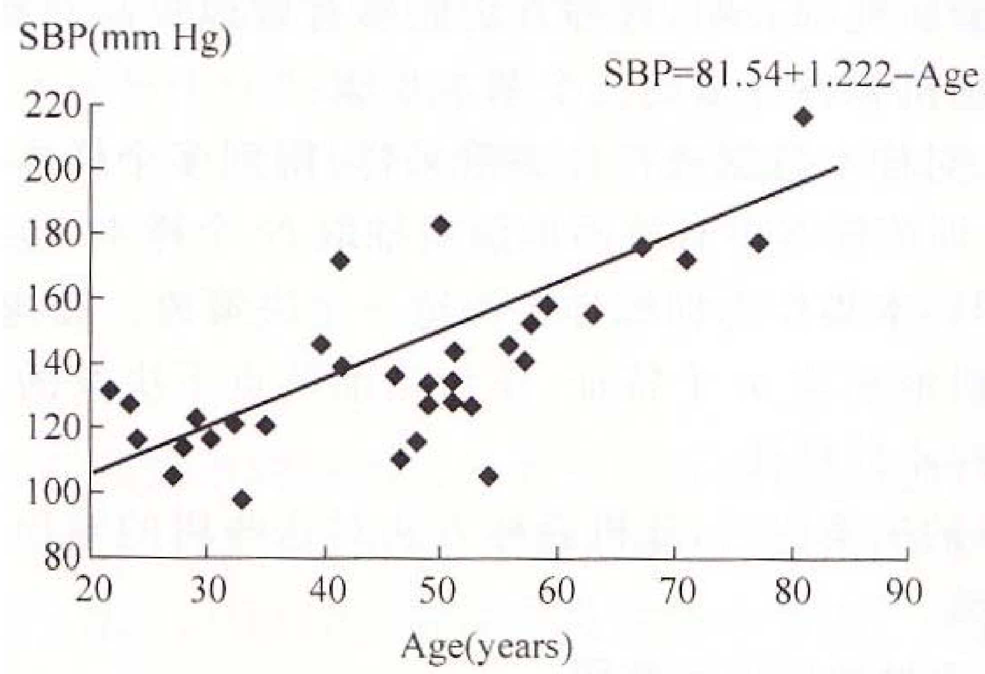
--多项罗杰斯特回归模型(面向多类问题)

1.回归问题的引入

线性回归、非线性回归

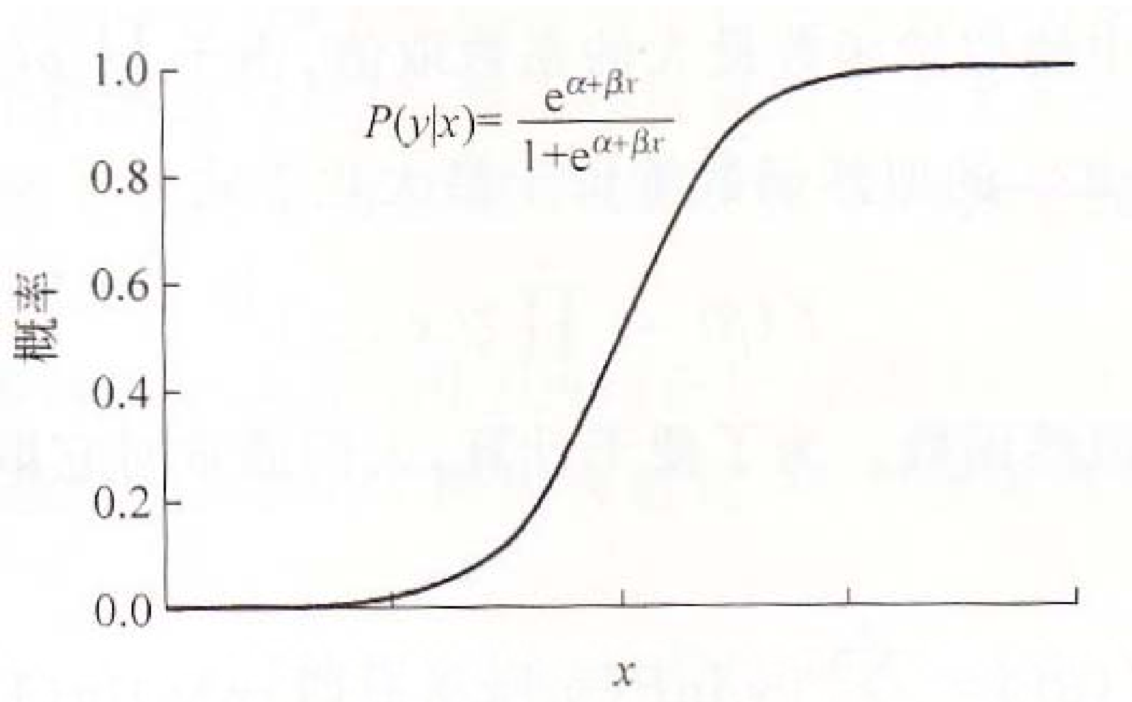
单变量回归、多元回归

例：基于最小二乘法的单变量、线性回归



回归分析的目的

- **定量分析**　基于观测数据，研究未知机理的问题
- **定性分析**　基于样本数据，研究特征与分类关系



2. Logistic回归(罗吉斯特回归)

因变量—离散型的分类变量(类别状态变量：两类、多类)发生结果的概率

自变量—分类结果的影响因素(样本特征)

Logistic 回归基于*logit*形式的对数几率模型，描述样本属于某类的可能性与样本特征之间的关系；以训练数据集估计*logit*函数中的参数。

从概率的角度，研究分类变量与样本特征之间关系。属于概率型、多元、非线性回归方法

二项Logistic回归

以**两类问题**为例。对于观测样本x,若其属于正类$(y=1)$的概率符合

Logistic函数, 则其**属于正类$(y=1)$与负类$(y=0)$概率之比("几率")**

$$\frac{P(y|\boldsymbol{x})}{1-P(y|\boldsymbol{x})}=e^{\beta_0+\boldsymbol{\beta}x} \qquad \left[记P(y=1|\boldsymbol{x})为P(y|\boldsymbol{x})\right]$$

注：一个事件的几率(odds)是指该事件发生的概率与该事件不发生的概率的比值。

对数几率

$$\boldsymbol{ln}\left(\frac{P(y|\boldsymbol{x})}{1-P(y|\boldsymbol{x})}\right)=\beta_0+\boldsymbol{\beta}\bullet\boldsymbol{x}$$

即：样本$\boldsymbol{x}$是正类的对数几率是关于$\boldsymbol{x}$的线性函数。

其中 $$\frac{P(y|\boldsymbol{x})}{1-P(y|\boldsymbol{x})}为P(y|\boldsymbol{x})的\boldsymbol{logit}函数$$

***Logistic*模型**--上述类别状态y与特征向量$\boldsymbol{x}$的关系模型

多元$\boldsymbol{logit}$函数

$$\boldsymbol{logit}(\boldsymbol{x}) = \boldsymbol{ln}\left(\frac{P(y|\boldsymbol{x})}{1-P(y|\boldsymbol{x})}\right) = \beta_0 + \boldsymbol{\beta}^T\boldsymbol{x} = \beta_0 + \sum_{i=1}^{m}\beta_i x_i$$

样本$\boldsymbol{x}$属于$y=1$类的概率

$$P(y|\boldsymbol{x}) = \frac{e^{\beta_0+\boldsymbol{\beta}^T\boldsymbol{x}}}{1+e^{\beta_0+\boldsymbol{\beta}^T\boldsymbol{x}}} = \frac{e^{\beta_0+\sum_{i=1}^{m}\beta_i x_i}}{1+e^{\beta_0+\sum_{i=1}^{m}\beta_i x_i}}$$

样本$\boldsymbol{x}$属于$y=0$类的概率

$$1-P(y|\boldsymbol{x}) = \frac{1}{1+e^{\beta_0+\boldsymbol{\beta}^T\boldsymbol{x}}} = \frac{1}{1+e^{\beta_0+\sum_{i=1}^{m}\beta_i x_i}}$$

3. 二项*Logistic*回归模型的参数估计(估计β参数)

对于**两类问题**，设训练样本集$\{(\boldsymbol{x}_i, y_i) \mid i=1,\ldots,n\}$

样本彼此独立，类别标号$y_i \in \{1,0\}$

ω_j类样本数$n_j, j=1,2$ $\qquad n_1+n_2=n$

样本(x_i, y_i)出现概率

$$P(\boldsymbol{x}_i, y_i) = \left[P(y_i \mid \boldsymbol{x}_i)\right]^{y_i}\left[1-P(y_i \mid \boldsymbol{x}_i)\right]^{1-y_i} p(\boldsymbol{x}_i) \overset{def}{=} \xi(\boldsymbol{x}_i, y_i) p(\boldsymbol{x}_i)$$

n个独立样本出现的似然函数

$$l = \prod_{i=1}^{n} P(\boldsymbol{x}_i, y_i) = \prod_{i=1}^{n} \xi(\boldsymbol{x}_i, y_i) \prod_{i=1}^{n} p(\boldsymbol{x}_i)$$

其中 $\prod_{i=1}^{n} p(\boldsymbol{x}_i)$与*Logistic*模型参数无关

$$\boldsymbol{ln}\left(\frac{P(y\mid\boldsymbol{x})}{1-P(y\mid\boldsymbol{x})}\right)=\beta_0+\boldsymbol{\beta}^T\boldsymbol{x}\qquad P(y\mid\boldsymbol{x})=\frac{e^{\beta_0+\boldsymbol{\beta}^T\boldsymbol{x}}}{1+e^{\beta_0+\boldsymbol{\beta}^T\boldsymbol{x}}}\qquad 1-P(y\mid\boldsymbol{x})=\frac{1}{1+e^{\beta_0+\boldsymbol{\beta}^T\boldsymbol{x}}}$$

$$\max\left(l=\prod_{i=1}^{n}P(\boldsymbol{x}_i,y_i)\right)\quad\Leftrightarrow\max\left(l'(\beta_0,\beta)=\prod_{i=1}^{n}\xi(\boldsymbol{x}_i,y_i)\right)$$

对 $l'(\beta)=\prod_{i=1}^{n}\xi(\boldsymbol{x}_i,y_i)$ 两边取对数:

$$L'(\beta_0,\boldsymbol{\beta})=\ln\left(l'(\beta_0,\boldsymbol{\beta})\right)=\sum_{i=1}^{n}\ln\left(\xi(\boldsymbol{x}_i,y_i)\right)$$

$$=\sum_{i=1}^{n}\left\{y_i\ln P(y_i\mid\boldsymbol{x}_i)+(1-y_i)\ln\left(1-P(y_i\mid\boldsymbol{x}_i)\right)\right\}$$

$$=\sum_{i=1}^{n}\left\{y_i\ln\frac{P(y_i\mid\boldsymbol{x}_i)}{1-P(y_i\mid\boldsymbol{x}_i)}+\ln\left(1-P(y_i\mid\boldsymbol{x}_i)\right)\right\}$$

$$=\sum_{i=1}^{n}\left\{y_i\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{x}_i\right)-\ln\left(1+e^{\beta_0+\boldsymbol{\beta}^T\boldsymbol{x}_i}\right)\right\}$$

$$L'(\beta_0,\beta)=\sum_{i=1}^{n}\left\{y_i\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{x}_i\right)-\ln\left(1+e^{\beta_0+\boldsymbol{\beta}^T\boldsymbol{x}_i}\right)\right\}$$

$$\begin{cases}\dfrac{\partial L'(\beta_0,\beta)}{\partial\boldsymbol{\beta}}=0\\[2ex]\dfrac{\partial L'(\beta_0,\beta)}{\partial\beta_0}=0\end{cases}\Rightarrow\begin{cases}\displaystyle\sum_{i=1}^{n}\boldsymbol{x}_i\left(y_i-\frac{e^{\beta_0+\boldsymbol{\beta}^T\boldsymbol{x}_i}}{1+e^{\beta_0+\boldsymbol{\beta}^T\boldsymbol{x}_i}}\right)=0\\[2ex]\displaystyle\sum_{i=1}^{n}\left(y_i-\frac{e^{\beta_0+\boldsymbol{\beta}^T\boldsymbol{x}_i}}{1+e^{\beta_0+\boldsymbol{\beta}^T\boldsymbol{x}_i}}\right)=0\end{cases}$$

迭代法解上述方程组$(m+1$个方程$)$，得参数$(\beta_0,\boldsymbol{\beta})$

--或直接采用梯度下降法。

4. 基于*Logistic*回归模型的分类决策

$$\begin{cases}\text{若 } \boldsymbol{logit}(\boldsymbol{x})=\boldsymbol{ln}\left(\dfrac{P(y|\boldsymbol{x})}{1-P(y|\boldsymbol{x})}\right)>0,\text{则}\boldsymbol{x}\in\omega_1\\[2ex]\text{若 } \boldsymbol{logit}(\boldsymbol{x})=\boldsymbol{ln}\left(\dfrac{P(y|\boldsymbol{x})}{1-P(y|\boldsymbol{x})}\right)<0,\text{则}\boldsymbol{x}\in\omega_2\end{cases}$$

多项Logistic回归
(SoftMax)

$$\log\left(\frac{P\left(y=\omega_j \mid \boldsymbol{x}\right)}{P\left(y=\omega_C \mid \boldsymbol{x}\right)}\right)=\beta_{j0}+\boldsymbol{\beta}_j^{T}\boldsymbol{x} \qquad \boldsymbol{j = 1,\ldots,C-1}$$

$$\Rightarrow P\left(y=\omega_j \mid \boldsymbol{x}\right)=P\left(y=\omega_C \mid \boldsymbol{x}\right)e^{\beta_{j0}+\boldsymbol{\beta}_j^{T}\boldsymbol{x}} \qquad \boldsymbol{j = 1,\ldots,C-1}$$

$$\Rightarrow 1=P\left(y=\omega_C \mid \boldsymbol{x}\right)+\sum_{j=1}^{C-1}P\left(y=\omega_j \mid \boldsymbol{x}\right)=P\left(y=\omega_C \mid \boldsymbol{x}\right)\left(1+\sum_{j=1}^{C-1}e^{\beta_{j0}+\boldsymbol{\beta}_j^{T}\boldsymbol{x}}\right)$$

$$\Rightarrow\begin{cases}P\left(y=\omega_C \mid \boldsymbol{x}\right)=\dfrac{1}{1+\sum_{s=1}^{C-1}e^{\beta_{s0}+\boldsymbol{\beta}_s^{T}\boldsymbol{x}}}\\[2ex] P\left(y=\omega_j \mid \boldsymbol{x}\right)=\dfrac{e^{\beta_{j0}+\boldsymbol{\beta}_j^{T}\boldsymbol{x}}}{1+\sum_{s=1}^{C-1}e^{\beta_{s0}+\boldsymbol{\beta}_s^{T}\boldsymbol{x}}} \qquad \boldsymbol{j = 1,\ldots,C-1}\end{cases}$$

多项*Logistic*回归模型的参数估计(估计β参数)

对于**多类问题**，设训练样本集$\left\{\left(\boldsymbol{x}_i, y_i\right) \mid i=1,\ldots,n\right\}$

样本彼此独立，类别标号$y_i \in \left\{\omega_1,\ldots,\omega_C\right\}$

ω_i类样本数n_i，　$\sum_{i=1}^{C} n_i = n$

样本$\left(\boldsymbol{x}_i, y_i\right)$出现概率

$$P\left(\boldsymbol{x}_i, y_i\right) = \left[\prod_{k=1}^{C}\left[P\left(y_i \mid \boldsymbol{x}_i\right)\right]^{I\left(y_i=\omega_k\right)}\right] p\left(\boldsymbol{x}_i\right) \overset{def}{=} \xi\left(\boldsymbol{x}_i, y_i\right) p\left(\boldsymbol{x}_i\right)$$

n个独立样本出现的似然函数

$$l = \prod_{i=1}^{n} P\left(\boldsymbol{x}_i, y_i\right) = \prod_{i=1}^{n} \xi\left(\boldsymbol{x}_i, y_i\right) \prod_{i=1}^{n} p\left(\boldsymbol{x}_i\right)$$

$\prod_{i=1}^{n} p\left(\boldsymbol{x}_i\right)$与*Logistic*模型参数无关

$$\max\left(l = \prod_{i=1}^{n} P\left(\boldsymbol{x}_i, y_i\right) \right) \Leftrightarrow \max\left(l'\left(\beta_0, \beta\right) = \prod_{i=1}^{n} \xi\left(\boldsymbol{x}_i, y_i\right) \right)$$

对 $l'\left(\beta\right) = \prod_{i=1}^{n} \xi\left(\boldsymbol{x}_i, y_i\right) = \prod_{i=1}^{C} \prod_{r=1}^{n_i} \left[P\left(\boldsymbol{x}_{ir} \mid \omega_i\right) \right]$

两边取对数:

$$L'\left(\beta_0, \boldsymbol{\beta}\right) = \ln\left(l'\left(\beta_0, \boldsymbol{\beta}\right) \right) = \sum_{i=1}^{C} \sum_{r=1}^{n_i} \ln\left[P\left(\boldsymbol{x}_{ir} \mid \omega_i\right) \right]$$

基于多类*Logistic*回归模型的分类决策

若$\beta_{j0}+\boldsymbol{\beta}_j^T\boldsymbol{x}=\max\limits_{k\in\{1,2,\ldots,C-1\}}\left\{\beta_{k0}+\boldsymbol{\beta}_k^T\boldsymbol{x}\right\}>0,$

则将$\boldsymbol{x}$判断为ω_j类；

否则，将$\boldsymbol{x}$判断为ω_C类

小结

将分类问题转化为概率的回归估计问题

基于样本数据,定性研究样本特征与分类变量关系

属于概率型、多元、非线性回归方法

思考：

1. 给定观测样本x，该样本是不同类别的后验概率＝?
2. 如何估计该样本关于各类别的几率、对数几率?
3. 如何针对给定的观测样本x，进行类别决策?
4. 以两类别logistic regression模型学习为例，该模型学习所构造的目标函数？模型的求解方式？多类别呢?