线性分类模型（二）：logistic回归模型分析

原创 石头 机器学习算法那些事 2018-11-02

前言

上一篇文章介绍了线性判别模型，本文介绍线性生成模型——logistic回归模型。本文介绍logstic回归模型相关的知识，为了更好理解模型的决策边界函数，本文同时分析了多元变量的协方差对概率分布的影响。

目录

logistic回归模型的含义

我们把分类模型分成两个阶段，推断阶段和决策阶段，推断阶段对联合概率分布建模，然后归一化，得到后验概率。决策阶段确定每个新输入x的类别。

我们用推断阶段的方法来推导logistic回归模型，首先对类条件概率密度$p(\vec{x} \mid C_k)$和类先验概率分布$p(C_k)$建模，然后通过贝叶斯定理计算后验概率密度。

考虑二分类的情形，类别$C1$的后验概率密度：

$$P(C1|\vec{x})=\frac{P(\vec{x}|C1)P(C1)}{P(\vec{x})}$$

$$P(C1|\vec{x})=\frac{P(\vec{x}|C1)P(C1)}{P(\vec{x}|C1)P(C1)+P(\vec{x}|C2)P(C2)}$$

$$P(C1|\vec{x})=\frac{1}{1+\dfrac{P(\vec{x}|C2)P(C2)}{P(\vec{x}|C1)P(C1)}}$$

令$\ln\dfrac{P(\vec{x}|C1)P(C1)}{P(\vec{x}|C2)P(C2)}=a$

则：$P(C1|\vec{x})=\dfrac{1}{1+\mathrm{e}^{-a}}=\sigma(a)$

式中的$\sigma(a)$就是$\log istic$函数

因此,$\log istic$回归的值等于输入变量为x的条件下类别为$C1$的概率($P(C1|\vec{x})$

$$\sigma(a)=\frac{1}{1+\mathrm{e}^{-a}}$$

$$\mathrm{a}=\ln\frac{P(\vec{x}|C1)P(C1)}{P(\vec{x}|C2)P(C2)}$$

$$a=\ln\frac{P(\vec{x},C1)}{P(\vec{x},C2)}$$

(1)当$a\geq 0$时，$P(\vec{x},C1)\geq P(\vec{x},C2)$，$P(C1|\vec{x})\geq\dfrac{1}{2}$，分类结果为$C1$

(2)当$a<0$时，$P(\vec{x},C1)<P(\vec{x},C2)$，$P(C1|\vec{x})<\dfrac{1}{2}$，分类结果为$C2$

结论： logistic回归值表示所属类的后验概率，无论是二分类还是多分类，分类结果都是后验概率最大所对应的类。

logistic的决策边界函数分析

决策边界函数，简而言之，就是函数的两侧是不同的分类结果，如上篇文章所涉及的边界函数是直线，本节首先介绍多元变量高斯分布的概念，然后讨论logistic的决策边界函数。

多元变量高斯分布的协方差解析

多元变量的高斯分布公式：

$$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

其中，x是D维变量，Σ 是变量x的协方差矩阵，u是变量的均值。

令$\Delta^2=(\vec{x}-\vec{u})^T\Sigma^{-1}(\vec{x}-\vec{u})$ (1)

易知协方差矩阵Σ是对称矩阵

对称矩阵的性质之一是对称矩阵可以分解成两两正交的特征向量

即：$\Sigma\vec{u_i}=\lambda_i\vec{u_i}$ (2) 其中$\vec{u_i}$是特征向量，λ_i是特征值

$\vec{u_i}^T\vec{u_j}=I_{ij}$，当$i=j$时，$I_{ij}=1$；反之，$I_{ij}=0$

(2)式右乘$\vec{u_i}^{-1}$，得：

$$\Sigma=\sum_{i=1}^{D}\lambda_i\vec{u_i}\vec{u_i}^T \quad (3)$$

(3)式左右两边取逆，得：

$$\Sigma^{-1}=\sum_{i=1}^{D}\frac{1}{\lambda_i}\vec{u_i}\vec{u_i}^T \quad (4)$$

把（4）式带入（1）式得：

$$\Delta^2=(\vec{x}-\vec{u})^T\sum_{i=1}^{D}\frac{1}{\lambda_i}\vec{u_i}\vec{u_i}^T(\vec{x}-\vec{u})=\sum_{i=1}^{D}\frac{1}{\lambda_i}[(\vec{x}-\vec{u})^T\vec{u_i}][(\vec{u_i}^T(\vec{x}-\vec{u})]$$

令$y_i=(\vec{x}-\vec{u})^T\vec{u_i}$ (5)

$$\therefore \quad \Delta^2=\sum_{i=1}^{D}\frac{y_i^2}{\lambda_i} \quad (6)$$

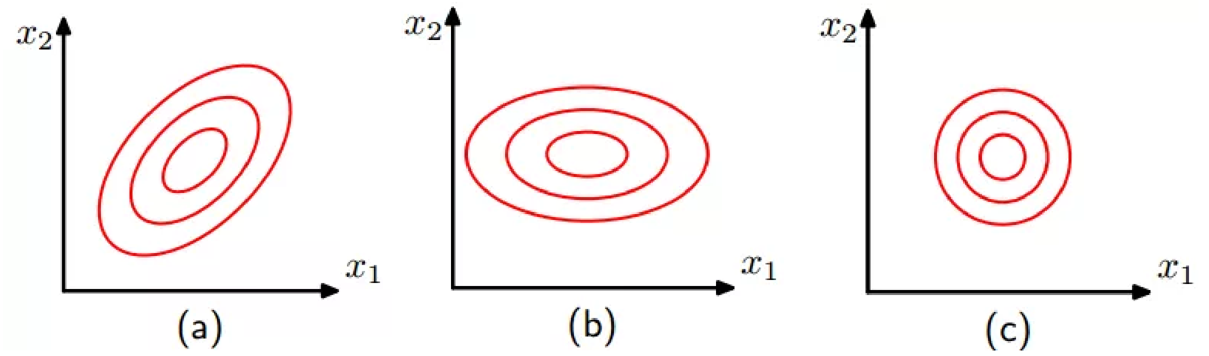
∴ y的特征分布表示x的分布：

$$p(y)=\prod_{i=1}^{D}\frac{1}{(2\pi\lambda)^{\frac{1}{2}}}\exp\{-\frac{y_j^2}{2\lambda_j}\} \quad (7)$$

由（5）式可知,$\{y_i\}$定义的坐标系是原坐标系平移和旋转生成的

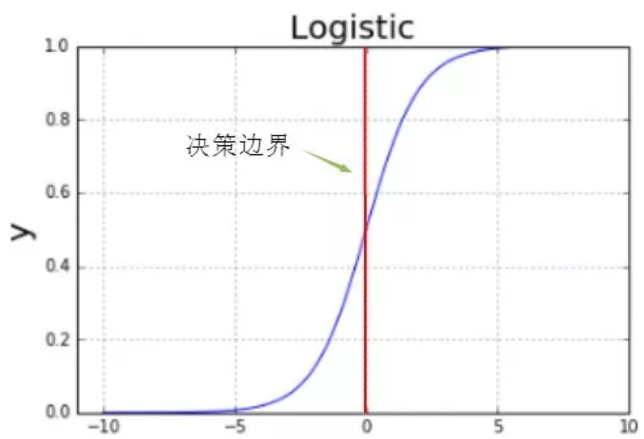
（7）式可知,$\{y_i\}$定义的曲线是椭圆，$\lambda_i^{\frac{1}{2}}$表示某一个维度的缩放因子

因此，**可定性的分析协方差的三种情况与分布图的关系，（a）图表示正常的协方差矩阵的高斯分布图：（b）图表示协方差矩阵是对角矩阵的高斯分布图；（c）图表示协方差矩阵是对角矩阵且对角元素都相等的高斯分布图。**



logistic的决策边界函数分析

logistic曲线如下图，红色直线（a=0）表示决策边界函数：



假设类条件概率密度是高斯分布，即P(x|Ck)，然后求解后验概率的表达式，即P(Ck|x)。由第一节可知logistic回归值就是所求的后验概率。

假设类条件概率密度的协方差相同，类条件概率密度为：

$$p(\boldsymbol{x} \mid \mathcal{C}_k) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_k)\right\}$$

由第一节的推导公式可得后验概率为：

$$P(C_k \mid x) = \sigma(w_k^T x + w_{k0})$$

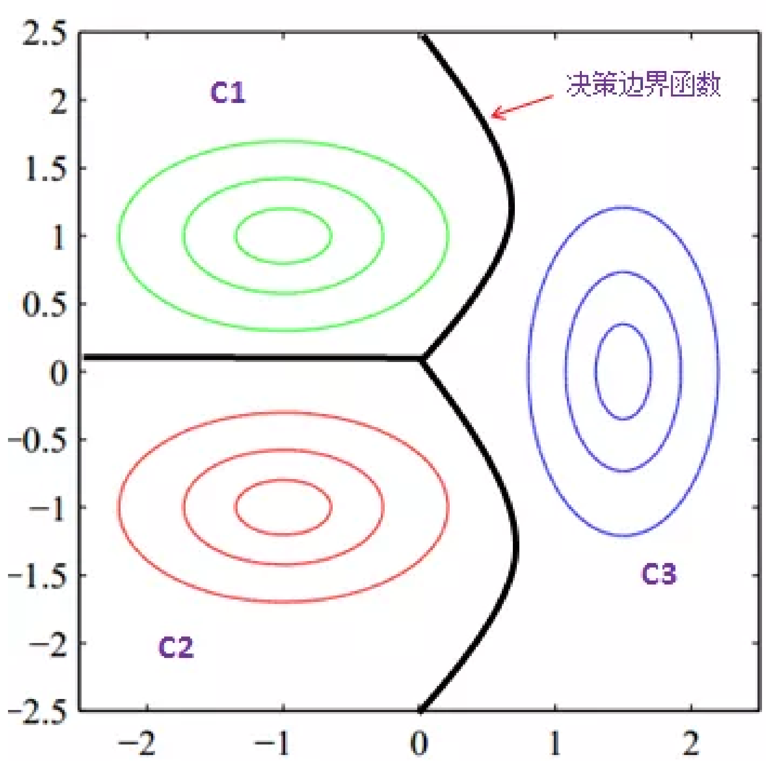
其中：

$$\boldsymbol{w}_k = \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k$$

$$w_{k0} = -\frac{1}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \ln p(\mathcal{C}_k)$$

由后验概率（$P(C_k|x)$）的表达式可知可知，当类条件的协方差矩阵相等时，决策边界函数是随x线性变化的直线。

结论：如下图，若两类的条件概率密度的协方差相同时（如C1和C2的协方差相同），则决策边界函数是直线；若两类的条件概率密度的协方差不相同时（如C1和C3，C2和C3），则决策边界函数是曲线。判断协方差矩阵是否相同可以根据分布图形形状是否相同来判断，如C1和C2的协方差相同，C3和C1、C2的协方差不相同，协方差如何影响多元变量分布可参考上一小节。



假设类条件概率密度符合高斯分布且具有相同的协方差矩阵，则决策边界函数是一条直线；若类条件概率密度符合更一般的指数分布且缩放参数s相同，决策边界函数仍是一条直线。

logistic模型的参数最优化

logistic模型损失函数

logistic回归模型的含义是后验概率分布，因此可以从概率的角度去设计损失函数。

考虑两分类情况，假设有N个训练样本，logistic模型是$h_\theta(x)$

$h_\theta(x)$表示后验概率$y=1$的概率，则$1-h_\theta(x)$表示$y=0$的概率

变量y_i取值1或0，且分别代表模型$h_\theta(x)$和$1-h_\theta(x)$

因此，似然函数$L(\theta)$：

$$L(\theta)=\prod_{i=1}^{N}(h_\theta(x)^{y_i})(1-h_\theta(x)^{y_i})$$

损失函数$J(\theta)$:

$$J(\theta)=-L(\theta)$$

$$J(\theta)=-\prod_{i=1}^{N}(h_\theta(x)^{y_i})(1-h_\theta(x)^{y_i})$$

logistic模型的参数最优化

损失函数最小化等价于模型参数的最优化，如下图：

$$J(\theta)=-\prod_{i=1}^{N}(h_\theta(x)^{y_i})(1-h_\theta(x)^{1-y_i})$$

$$(J(\theta))_{\min}=\ln(J(\theta))\min$$

$$\ln(J(\theta))=-\prod_{i=1}^{N}(y_i\ln(h_\theta(x)+(1-y_i)\ln(1-h_\theta(x)))$$

利用梯度下降法求最优解，学习速率α：

$$\theta=\theta-\alpha\frac{\partial J(\theta)}{\partial\theta}$$

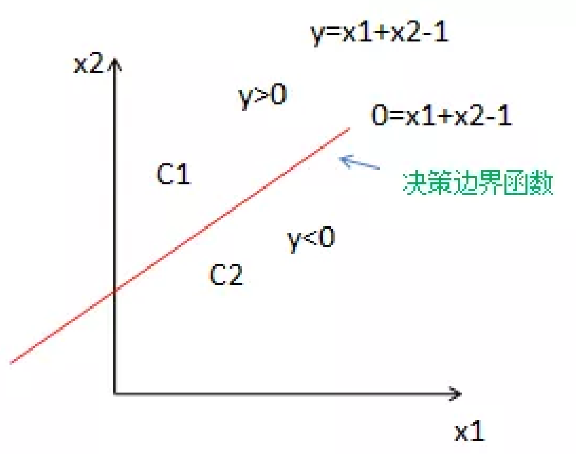
具体求法在本文不展开，只给出算法思想。

为了避免过拟合问题，则在原来的损失函数增加正则项，然后利用梯度下降法求最优解，这里也不展开。

logistic模型与感知机模型的比较

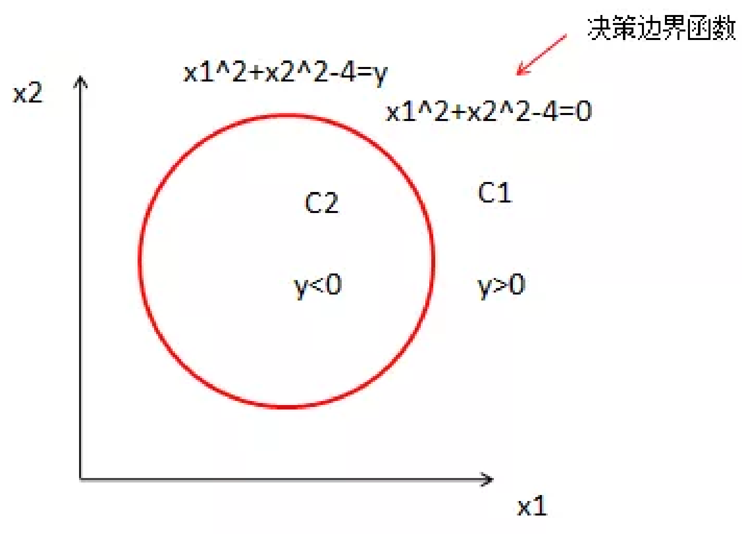
logistic模型与感知机模型的相同点

由第二节分析可知，假设类条件概率分布的协方差相同，则logistic模型的决策边界函数是随x线性变化的直线，因此，感知机模型与logistic模型的分类策略一样，即决策边界函数是一样的。如下图。



感知机模型： 当点落在直线上方，y>0，则分类结果C1；反之为C2。
logistic模型：当点落在直线上方，y>0，则后验概率P(C1|X)>0.5，分类结果C1；反之为C2。

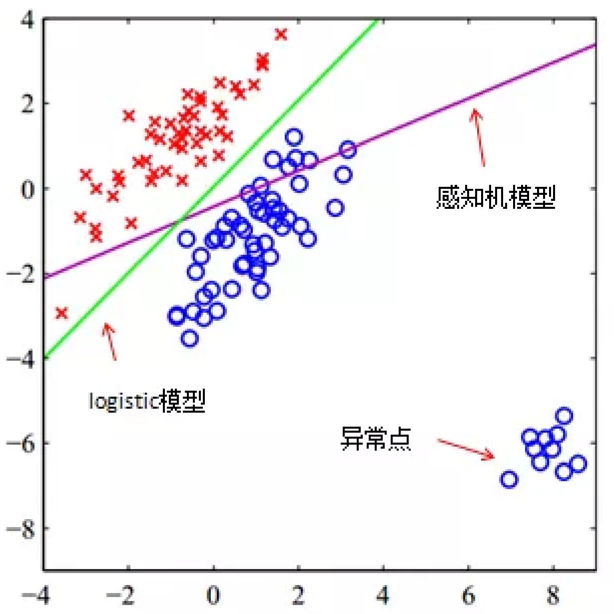
考虑到对输入变量x进行非线性变换 $\theta(x)$ ，感知机和logistic模型的分类策略仍一样，决策边界函数相同，如下图：



感知机模型： 当点落在圆外，y>0，则分类结果C1；反之为C2。
logistic模型：当点落在圆外，y>0，则后验概率P(C1|X)>0.5，分类结果C1；反之为C2。

logistic模型与感知机模型的异同点

（1）logistic回归模型限制值的范围在0~1，感知机模型对值范围没有限制，因此logistic模型相比感知机模型，对异常点有更强的鲁棒性。如下图，当有异常数据时，logistic模型要好于感知机模型。



（2）感知机模型用误分类点到超平面的距离衡量损失函数，而logistic模型则从概率角度去衡量损失函数。

总结

logistic回归的含义是后验概率分布，用概率的角度去设计似然函数，logistic模型相比于感知机模型对异常数据具有更好的鲁棒性。

参考：

Christopher M.Bishop <<Pattern Reconition and Machine Learning>>

推荐阅读文章

线性分类模型（一）：浅谈判别模型分析

深入理解线性回归算法（三）：浅谈贝叶斯线性回归

深入理解线性回归算法（二）：正则项的详细分析

深入理解线性回归算法（一）

线性回归：不能忽视的三个问题

浅谈频率学派和贝叶斯学派

浅谈先验分布和后验分布