中国高科集团
CHINA HI-TECH GROUP

Logistic回归

乐 · 教育 | 智 · 全球
Embrace education | Enlighten horizon

2021年12月

回顾

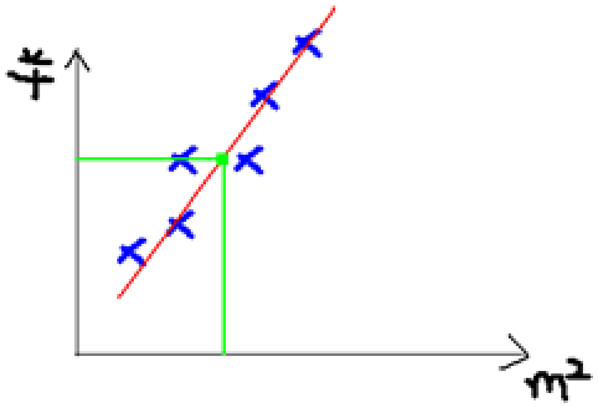
- KNN
- 决策树
- 贝叶斯分类器

- 是否有一种算法
  - 二分类、多类
  - 输出连续值
  - 预测
  - 优化算法

----回归

回归

面积 (m^2)	销售价钱 （万元）
123	250
150	320
87	160
102	220
…	…



$$h(x) = h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

回归

- 回归：广义线性模型（generalized linear model）
- 分类：根据因变量的不同
  - 连续：多重线性回归
  - 二项分布：logistic回归
  - poisson分布：poisson回归
  - 负二项分布：负二项回归

逻辑斯蒂分布

- Logistic distribution
- 设X是连续随机变量，X服从Logistic distribution，
- 分布函数：

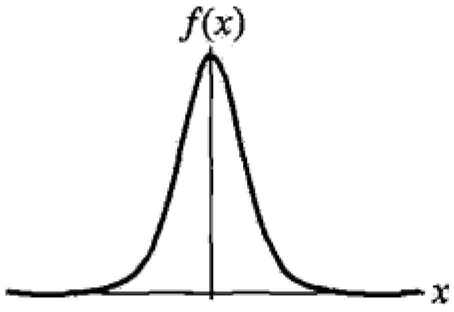
$$F(x)=P(X\leqslant x)=\frac{1}{1+\mathrm{e}^{-(x-\mu)/\gamma}}$$

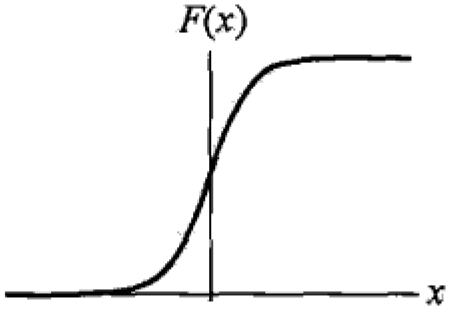
- 密度函数：

$$f(x)=F'(x)=\frac{\mathrm{e}^{-(x-\mu)/\gamma}}{\gamma(1+\mathrm{e}^{-(x-\mu)/\gamma})^2}$$

- μ为位置参数，γ大于0为形状参数，
  (μ,1/2)中心对称

$$F(-x+\mu)-\frac{1}{2}=-F(x-\mu)+\frac{1}{2}$$

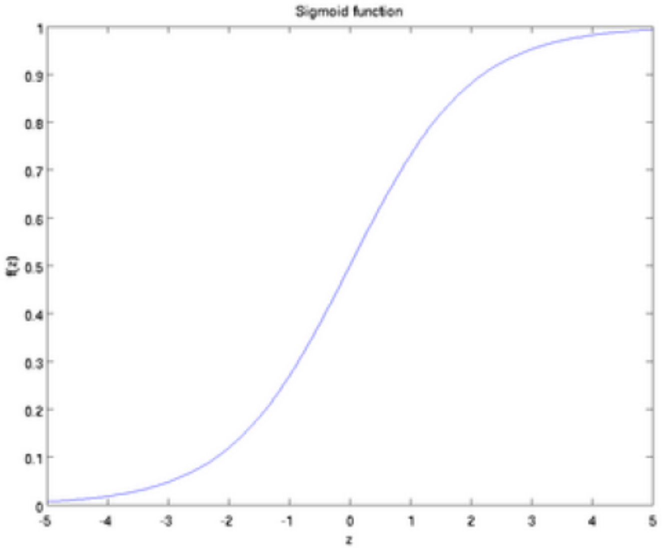




- Sigmoid：

$$f(z) = \frac{1}{1 + \exp(-z)}.$$

$$f'(z) = f(z)(1 - f(z))$$

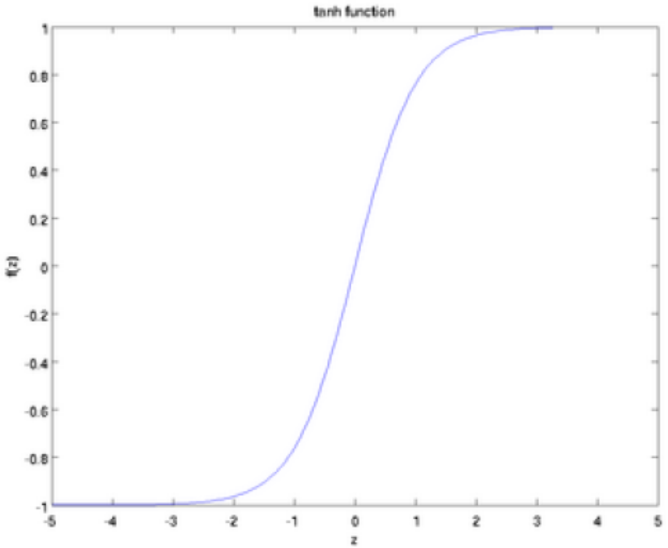


$[0, 1]$

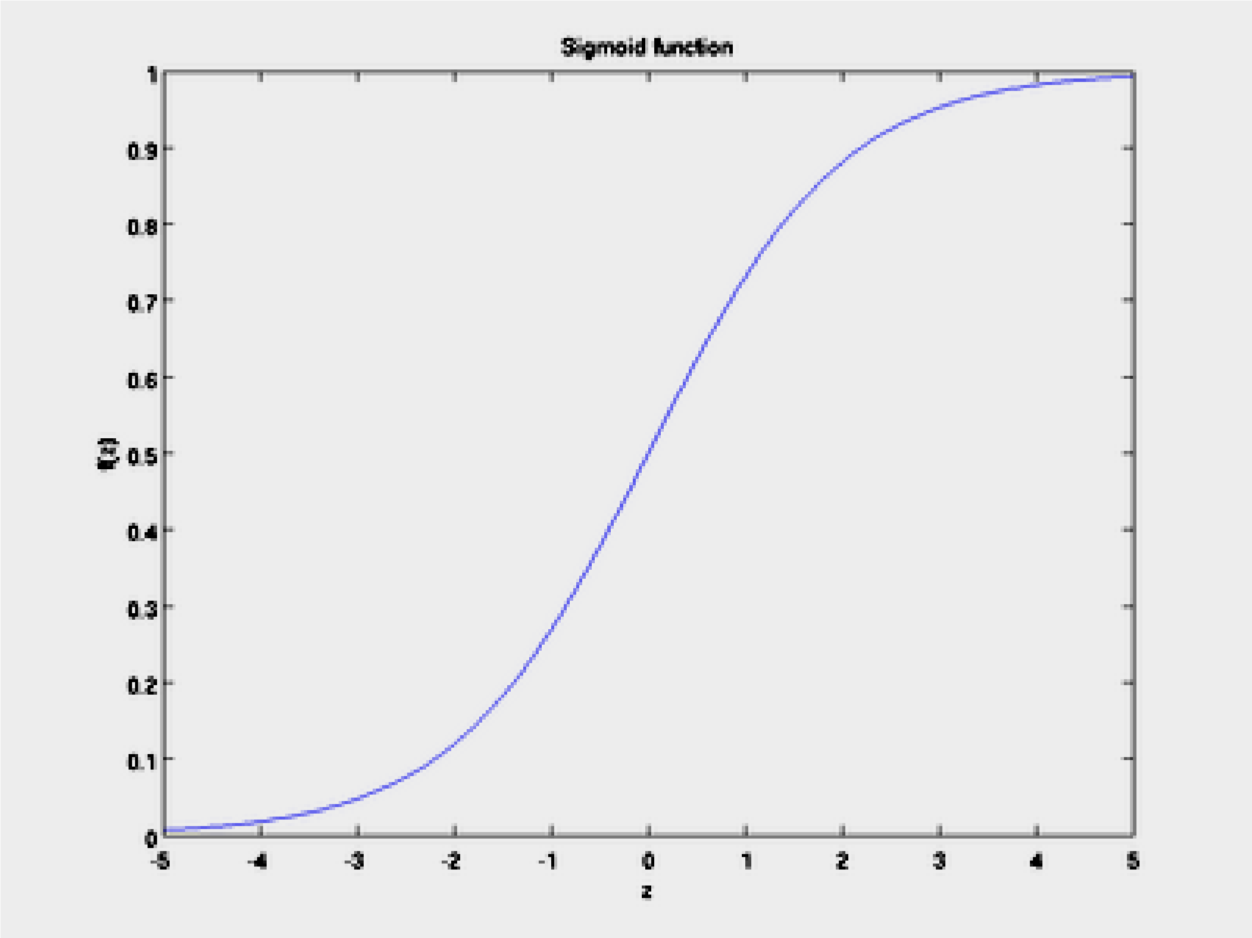
- 双曲正切函数（tanh）

$$f(z) = \tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

$$f'(z) = 1 - (f(z))^2$$



$[-1, 1]$



中国高科集团
CHINA HI-TECH GROUP

Sigmoid function:

$$h_{W,b}(x) = f(W^T x) = f(\textstyle\sum_{i=1}^{3} W_i x_i + b), \quad f(z) = \frac{1}{1+\exp(-z)}.$$

```
def sigmoid(inX):
    return 1.0/(1+exp(-inX))
```

二项逻辑斯蒂回归

- Binomial logistic regression model
  - 由条件概率P(Y|X)表示的分类模型
  - 形式化为logistic distribution
  - X取实数，Y取值1,0

$$P(Y=1|x)=\frac{\exp(w\cdot x+b)}{1+\exp(w\cdot x+b)}$$

$$P(Y=0|x)=\frac{1}{1+\exp(w\cdot x+b)}$$

$$P(Y=1|x)=\frac{\exp(w\cdot x)}{1+\exp(w\cdot x)}$$

$$P(Y=0|x)=\frac{1}{1+\exp(w\cdot x)}$$

$$w=(w^{(1)},w^{(2)},\cdots,w^{(n)},b)^{\mathrm{T}}$$

$$x=(x^{(1)},x^{(2)},\cdots,x^{(n)},1)^{\mathrm{T}}$$

二项逻辑斯蒂回归

- 事件的几率odds：事件发生与事件不发生的概率之比为

$$\frac{p}{1-p}$$

- 称为事件的发生比(the odds of experiencing an event),
- 对数几率：

$$\operatorname{logit}(p)=\log\frac{p}{1-p}$$

- 对逻辑斯蒂回归：

$$\log\frac{P(Y=1|x)}{1-P(Y=1|x)}=w\cdot x$$

似然函数

- *logistic*分类器是由一组权值系数组成的，最关键的问题就是如何获取这组权值，通过极大似然函数估计获得，并且$Y \sim f(x;w)$
- 似然函数是统计模型中参数的函数。给定输出x时，关于参数θ的似然函数L(θ|x)（在数值上）等于给定参数θ后变量X的概率：$L(\theta|x)=P(X=x|\theta)$
- 似然函数的重要性不是它的取值，而是当参数变化时概率密度函数到底是变大还是变小。
- 极大似然函数：似然函数取得最大值表示相应的参数能够使得统计模型最为合理

似然函数

- 那么对于上述*m*个观测事件，设

$$P(Y=1|x)=\pi(x)\,,\quad P(Y=0|x)=1-\pi(x)$$

- 其联合概率密度函数，即似然函数为：

$$\prod_{i=1}^{N}[\pi(x_i)]^{y_i}[1-\pi(x_i)]^{1-y_i}$$

- 目标：求出使这一似然函数的值最大的参数估，$w_1,w_2,\ldots,w_n$，使得L(w)取得 最大值。
- 对L(w)取对数：

模型参数估计

- 对数似然函数

$$
\begin{aligned}
L(w) &= \sum_{i=1}^{N}[y_i \log \pi(x_i) + (1-y_i)\log(1-\pi(x_i))] \\
&= \sum_{i=1}^{N}\left[y_i \log \frac{\pi(x_i)}{1-\pi(x_i)} + \log(1-\pi(x_i)) \right] \\
&= \sum_{i=1}^{N}[y_i(w \cdot x_i) - \log(1+\exp(w \cdot x_i))]
\end{aligned}
$$

- 对L(w)求极大值，得到w的估计值。
- 通常采用梯度下降法及拟牛顿法，学到的模型：

$$P(Y=1 \mid x) = \frac{\exp(\hat{w} \cdot x)}{1+\exp(\hat{w} \cdot x)} \qquad P(Y=0 \mid x) = \frac{1}{1+\exp(\hat{w} \cdot x)}$$

多项logistic回归

- 设Y的取值集合为

$$\{1,2,\cdots,K\}$$

- 多项logistic回归模型

$$P(Y=k\mid x)=\frac{\exp(w_k\cdot x)}{1+\sum_{k=1}^{K-1}\exp(w_k\cdot x)},\quad k=1,2,\cdots,K-1$$

$$P(Y=K\mid x)=\frac{1}{1+\sum_{k=1}^{K-1}\exp(w_k\cdot x)}$$