线性回归模型的矩阵表示法

- 最小二乘法（最小平方法）是基于误差为高斯分布（正态）的假设的来的。
- 在用线性回归做预测时，或多或少存在误差，因此引入误差变量epsilon，将误差看作是随机变量。
- 根据大数定理，样本数增多，误差会慢慢服从正态分布。

假设epsilon服从正态分布，针对每个样本得出条件概率：

线性回归的误差模型

$$y = w^T x + b + \epsilon \qquad \epsilon \sim N(0, \sigma^2)$$

针对于样本 (x_i, y_i)

$$p(y_i | x_i) \sim N(w^T x_i + b, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(w^T x_i + b - y_i)^2}{2\sigma^2}}$$

高斯噪声模型根据最大似然估计法求参数w,b:（优化最小二乘）

基于所有样本的最大似然 (Maximum Likelihood)

贪 心 科 技
GREEDY TECHNOLOGY

$$D = \{(x_1, y_1), \cdots, (x_N, y_N)\}$$

$$
\begin{aligned}
w^*, b^* &= \underset{w,b}{\arg\max} \prod_{i=1}^{N} p(y_i | x_i) = \underset{w,b}{\arg\max} \log\left[\prod_{i=1}^{N} p(y_i | x_i)\right] \\
&= \underset{w,b}{\arg\max} \sum_{i=1}^{N} \log p(y_i | x_i) = \underset{w,b}{\arg\max} \sum_{i=1}^{N} \log\left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(w^T x_i + b - y_i)^2}{2\sigma^2}\right)\right] \\
&= \underset{w,b}{\arg\max} \sum_{i=1}^{N} \cancel{-\log(\sqrt{2\pi}\sigma)} - \frac{(w^T x_i + b - y_i)^2}{2\sigma^2} \\
&= \underset{w,b}{\arg\max} \sum_{i=1}^{N} -(w^T x_i + b - y_i)^2 \\
&= \underset{w,b}{\arg\min} \sum_{i=1}^{N} (w^T x_i + b - y_i)^2
\end{aligned}
$$

$$\max \log p(D) \iff \min \sum_{i=1}^{N} (w^T x_i + b - y_i)^2$$

高斯误差模型下的最大似然　　　　线性回归的最小二乘

- 与w,b无关的参数忽略--在求导过程中为0

假设有N维向量$x_1, x_2, ..., x_N, x_i \in R^n$
观测值Y：$y_1, y_2, ..., y_N, y_i \in R^n$

定义线性方程：$y_i = w^T x_i, w \in R^n$
拟合误差：$e_i = y_i - w^T x_i$

假设误差符合标准正太分布：$e_i \sim N(0,1)$
即概率密度函数：$e_i \sim \frac{1}{\sqrt{2\pi}} e^{-\frac{e_i^2}{2}}$

似然函数

$$\text{似然函数}L = \ln[\frac{1}{\sqrt{2\pi}} e^{-\frac{e_1^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{e_2^2}{2}} \cdots, \frac{1}{\sqrt{2\pi}} e^{-\frac{e_N^2}{2}}]$$
$$= -N \ln \sqrt{2\pi} - \frac{1}{2}(e_1^2 + e_2^2 + \cdots + e_N^2)$$

最小化误差

最大化L等价于

最小化$(e_1^2 + e_2^2 + \cdots + e_N^2)$

$$J = \min (y_1 - w^T x_1)^2 + (y_2 - w^T x_2)^2 + \cdots + (y_N - w^T x_N)^2$$

求解权重最优解

$$J=\min(y_1-w^Tx_1)^2+(y_2-w^Tx_2)^2+\cdots+(y_N-w^Tx_N)^2$$

$$\frac{\partial J}{\partial w}=(y_1-wx_1{}^T)x_1+\cdots+(y_N-wx_N{}^T)x_N=0$$

$$w(\sum_{i=1}^{N}x_i{}^Tx_i)=\sum_{i=1}^{N}x_iy_i$$

$$w=(\sum_{i=1}^{N}x_iy_i)(\sum_{i=1}^{N}x_i{}^Tx_i)^{-1}$$

$$\text{对比之前}\ a=(x^Tx)^{-1}x^TY$$

最终得到的w和最小二乘估计得到的结果是一致的。

逻辑回归-分类问题:

0-1问题：是否违约，情感分析（开心与否），广告点击率，疾病分析

逻辑回归本身解决的是二分类问题:p(y=1|x)和p(y=0|x)

建立逻辑回归模型要考虑定义域与值域的匹配问题，y与x的对应关系中，p(y|x)的取值只可能落在（0，1）里

逻辑函数:

$$y=\frac{1}{1+e^{-x}}$$

定义域(x): (x): $(-\infty, +\infty)$

值域(y): (y): y: $(0, 1)$

采用线性回归建模的思想来处理逻辑回归的映射关系:

对于特征向量x和二分类标签y，我们可以定义如下的条件概率：

$$p(y=1|x,w,b)=\frac{1}{1+e^{-(w^Tx+b)}}$$

$$p(y=0|x,w,b)=\frac{e^{-(w^Tx+b)}}{1+e^{-(w^Tx+b)}}$$

证明逻辑回归是线性分类器：

逻辑回归模型是线性分类器

$$p(y=1|x,w)=\frac{1}{1+e^{(w^Tx+b)}}$$

$$p(y=0|x,w)=\frac{e^{-(w^Tx+b)}}{1+e^{-(wx+b)}}$$

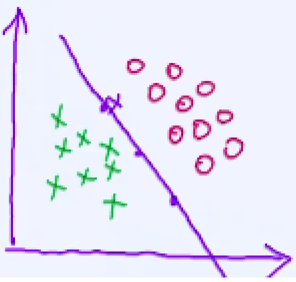
给定任意点 x，我们有

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

$$\Leftrightarrow \quad 1 = e^{-(w^Tx+b)} \quad \log$$

$$\Leftrightarrow \quad 0 = -(w^Tx+b)$$

$$\Leftrightarrow \quad w^Tx+b = 0 \Leftarrow \text{线性}$$



采用条件独立(conditional independence)的假设，在这个假设的前提下，可以把条件概率$p(x_1,\dots,x_n|w)$ 分解成$p(x_1|w)\dots p(x_n|w)$，构造极大似然函数。

目标函数（Objective Function）

假设我们拥有数据集：$D=\{(x_i,y_i)\}_{i=1}^{n} \quad x_i\in R^d, \quad y_i\in\{0, 1\}$

而且我们已经定义了：

$$p(y|x,w,b)=p(y=1|x,w,b)^y[1-p(y=1|x,w,b)]^{1-y}$$

我们需要最大化的目标函数为（也叫做最大似然函数）：

$$\hat{w}_{MLE}, \hat{b}_{MLE}=argmax_{w,b}\prod_{i=1}^{n}p(y_i|x_i,w,b)$$

将两个条件概率公式合并：当y=1时，后面那一项不起任何作用也就变成了第一个条件概率； 当y=0时，前面那一项不起作用，也就等同于第二个条件概率。

目标函数

给定条件概率：$p(y|x,w,b)=p(y=1|x,w,b)^y[1-p(y=1|x,w,b)]^{1-y}$

$$argmin_{w,b}-\sum_{i=1}^{n}\log p(y_i|x_i,w,b)$$

$$\log(a^x\cdot b^y)=\log(a^x)+\log(b^y)=x\cdot\log a+y\cdot\log b$$

$$=\underset{w,b}{argmin}-\sum_{i=1}^{n}\log\left\{p(y_i=1|x_i,w,b)^{y_i}\cdot[1-p(y_i=1|x_i,w,b)]^{1-y_i}\right\}$$

$$=\underset{w,b}{argmin}-\sum_{i=1}^{n}y_i\log p(y_i=1|x_i,w,b)+(1-y_i)\log[1-p(y_i=1|x_i,w,b)]$$

求解目标函数的最值：

- 导数为0：对逻辑回归不适用，导数表达式难以写出
- 迭代方法如梯度下降：不断更新参数，与lr参数相关。

梯度下降法：

方法2：梯度下降法

求使得f(w)值最小的参数w

初始化：w^1

for t = 1,2,...:

$$w^{t+1} = w^t - \eta \nabla f(w^t)$$

$$w^* = -\frac{5}{8} = -0.625$$

例子：求解函数 $f(w) = 4w^2 + 5w + 1$ 的最优解

$w^1=0$, $\eta=0.1$, $f'(w)=8w+5$

$$w^2 = w^1 - 0.1\cdot(8\cdot 0+5) = 0-0.1\cdot 5 = -0.5$$

$$w^3 = w^2 - 0.1(8\cdot(-0.5)+5) = -0.5-0.1\cdot 1 = -0.6$$

$$w^4 = w^3 - 0.1(8\cdot(-0.6)+5) = -0.6-0.1\cdot 0.2 = -0.62$$

$$w^5 = w^4 - 0.1(8\cdot(-0.62)+5) = -0.62-0.1\cdot 0.04 \approx -0.625$$

对于逻辑函数的求导

$$\sigma(x) = \frac{1}{1+e^{-x}} \qquad \sigma'(x) = ?$$

$$\sigma'(x) = \frac{(1)'\cdot(1+e^{-x}) - 1\cdot(1+e^{-x})'}{(1+e^{-x})^2} = \frac{e^{-x}}{(1+e^{-x})^2}$$

$$= \frac{1}{1+e^{-x}}\cdot\frac{e^{-x}}{1+e^{-x}} = \frac{1}{1+e^{-x}}\cdot\left[1-\frac{1}{1+e^{-x}}\right]$$

$$= \sigma(x)\cdot[1-\sigma(x)]$$

求解w参数：

逻辑回归的梯度下降法—求解w

$$p(y=1|x,w,b)=\frac{1}{1+e^{-(w^Tx+b)}}=\sigma(w^Tx+b)$$

$$argmin_{w,b}-\sum_{i=1}^{n}y_i\log p(y_i=1|x_i,w,b)+(1-y_i)\log(1-p(y_i=1|x_i,w,b))]$$

$$L=\underset{w,b}{argmin}-\sum_{i=1}^{n}y_i\log\sigma(w^Tx_i+b)+(1-y_i)\log[1-\sigma(w^Tx_i+b)]$$

$$\boxed{\frac{\partial L}{\partial w}}=-\sum_{i=1}^{n}y_i\cdot\frac{\cancel{\sigma(w^Tx_i+b)}[1-\sigma(w^Tx_i+b)]x_i}{\cancel{\sigma(w^Tx_i+b)}}+(1-y_i)\cdot\frac{-\sigma(w^Tx_i+b)\cancel{[1-\sigma(w^Tx_i+b)]}\cdot x_i}{\cancel{1-\sigma(w^Tx_i+b)}}$$

$$=-\sum_{i=1}^{n}y_i[1-\sigma(w^Tx_i+b)]x_i+(y_i-1)\sigma(w^Tx_i+b)x_i$$

$$=\sum_{i=1}^{n}[\sigma(w^Tx_i+b)-y_i]x_i$$

求解b参数：

逻辑回归的梯度下降法—求解b

$$p(y=1|x,w,b)=\frac{1}{1+e^{-(w^Tx+b)}}$$

$$argmin_{w,b}-\sum_{i=1}^{n}y_i\log p(y_i=1|x_i,w,b)+(1-y_i)\log(1-p(y_i=1|x_i,w,b))]$$

$$=\underset{w,b}{argmin}\ \underbrace{-\sum_{i=1}^{n}y_i\log\sigma(w^Tx_i+b)+(1-y_i)\log[1-\sigma(w^Tx_i+b)]}_{L}$$

$$\boxed{\frac{\partial L}{\partial b}}=-\sum_{i=1}^{n}y_i\cdot\frac{\cancel{\sigma(w^Tx_i+b)}[1-\sigma(w^Tx_i+b)]}{\cancel{\sigma(w^Tx_i+b)}}+(1-y_i)\cdot\frac{-\sigma(w^Tx_i+b)\cancel{[1-\sigma(w^Tx_i+b)]}}{\cancel{1-\sigma(w^Tx_i+b)}}$$

$$=\sum_{i=1}^{n}(\sigma(w^Tx_i+b)-y_i)$$

迭代更新：

w' = w - lr * f'(w)

b' = b - lr * f'(b)

f'(w,b)为损失函数

逻辑回归的梯度下降法

初始化：w^1，b^1

for $t = 1, 2 \ldots$

$$w^{t+1} = w^t - \eta_t \sum_{i=1}^{n} (\sigma(w^T x_i + b) - y_i) x_i$$

$$b^{t+1} = b^t - \eta_t \sum_{i=1}^{n} (\sigma(w^T x_i + b) - y_i)$$

判断收敛：

- 相邻时间段计算当前的损失函数，损失函数变化很小或者不变即收敛
- 相邻时间段计算当前参数的值w,b,参数值变化很小或者不变即收敛