Linear Regression

模型

预测函数: $h_\theta(X^i)=\Theta X^i$，其中$\Theta=(\theta_1,\theta_2,\dots,\theta_d,b);X^i=(x_1^i,x_2^i,\dots,x_d^i,1)$，这是一个线性函数。

预测误差可以写成：$y^i=h_\Theta(X^i)+\epsilon$，假设$\epsilon$满足高斯分布，则$\epsilon$的似然函数

$$L(\mu,\sigma,\theta)=\prod_{i=1}^{m}\frac{1}{\sqrt{2\pi}\sigma}exp(-\frac{(y^i-h_\theta(X^i)-\mu)^2}{2\sigma})$$

对数似然函数

$$\ell(\mu,\sigma,\theta)=-\frac{m}{2}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(y^i-h_\theta(X^i)-\mu)^2$$

我们希望每个误差函数的高斯分布的中心都正好处于预测的线性函数上，即$\mu\to 0$。

$$\ell(\mu=0,\sigma,\theta)=-\frac{m}{2}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(y^i-h_\theta(X^i))^2$$

可见，对于一定分布的ϵ，最大化似然函数等价于最小化损失函数

$$J(\Theta)=\frac{1}{2}\sum_{i=1}^{m}(y^i-h_\theta(X^i))^2=\frac{1}{2}(y-X\Theta)^T(y-X\Theta)$$

算法实现

线性模型都属于凸优化问题，可以使用梯度下降法，牛顿法。

梯度下降法

基本思想：使得参数沿着负梯度的方向前进，$\Theta_{t+1}=\Theta_t-\alpha\frac{\partial J(\Theta_t)}{\partial\Theta_t}$，t表示迭代第t次。

梯度

$$\frac{\partial J(\Theta_t)}{\partial\Theta_t}=\sum_{k=1}^{d+1}\frac{\partial J(\theta_k^t)}{\partial\theta_k^t}=\frac{1}{2}\sum_{k=1}^{d+1}\frac{\partial}{\partial\theta_k^t}\left[\sum_{i=1}^{m}(y^i-\sum_{j=1}^{d+1}x_{ij}\theta_j^t)^2\right]=X^T(X\Theta_t-y)$$

所以，第t+1次迭代的结果，$\Theta_{t+1}=\Theta_t-\alpha X^T(X\Theta_t-y)$。数学的解析解容易得到，

$$\Theta=(X^TX)^{-1}X^Ty;\quad h_\theta(X^i)=(X^TX)^{-1}X^TyX^i$$

算法优化

常见做法是正则化，使得算法尽量简单，使得$(X^TX)^{-1}$更容易计算。

岭优化（ridge regression）

损失函数

$$J(\Theta)=\frac{1}{2}(y-X\Theta)^T(y-X\Theta);\ s.t.||\Theta||_2^2\leqslant t$$

拉格朗日乘数法

$$J(\Theta)=\frac{1}{2}(y-X\Theta)^T(y-X\Theta)+\alpha(||\Theta||_2^2-t),\quad s.t.\alpha\geqslant 0$$

求梯度

$$\frac{\partial J(\Theta)}{\partial\Theta}=X^T(X\Theta-y)+2\alpha\Theta$$

使得梯度等于0，容易得到解析解

$$\Theta=(X^TX+\alpha I)^{-1}X^Ty;\quad I=\begin{pmatrix}0&0\\0&I_d\end{pmatrix}$$

LASSO regression

损失函数

$$J(\Theta)=\frac{1}{2}(y-X\Theta)^T(y-X\Theta);\ s.t.||\Theta||\leqslant t$$

优点：减低Θ的维度。缺点：该函数优化无法使用最小二乘法求解。