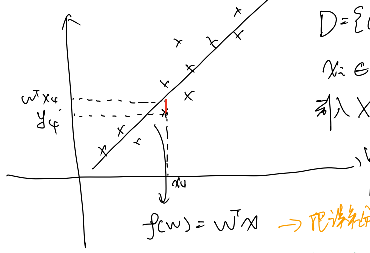
最小二乘法（矩阵表达，几何意义）



$D=\{(x_1,y_1),(x_2,y_2)\cdots(x_N,y_N)\}$

$x_i \in \mathbb{R}^p \quad y_i \in \mathbb{R} \quad i=1,2,\cdots N$ 默认为列向量

引入 $X=(x_1,x_2\cdots x_N)^T=\begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix}=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{pmatrix}_{N\times p}$

$Y=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}_{N\times 1}$

一般情况下 $N>p$

$f(w)=w^Tx$ → 把误差分散在N个点上

最小二乘估计 $L(w)=\sum_{i=1}^{N}\|\underset{1\times p}{w^T}\underset{p\times 1}{x_i}-\underset{1}{y_i}\|^2$

$=\sum_{i=1}^{N}(w^Tx_i-y_i)^2$

$=(w^Tx_1-y_1,\ w^Tx_2-y_2,\cdots w^Tx_N-y_N)\begin{pmatrix} w^Tx_1-y_1 \\ w^Tx_2-y_2 \\ \vdots \\ w^Tx_N-y_N \end{pmatrix}$

$(w^Tx_1-y_1,\ w^Tx_2-y_2,\cdots w^Tx_N-y_N)$

$=(w^Tx_1,w^Tx_2,\cdots w^Tx_N)-(y_1,y_2\cdots y_N)$

$=w^T(x_1,x_2\cdots x_N)-(y_1,y_2\cdots y_N)$

$=w^TX^T-Y^T$

$\begin{pmatrix} w^Tx_1-y_1 \\ \vdots \\ w^Tx_N-y_N \end{pmatrix}=(w^TX^T-Y^T)^T=Xw-Y$

$=(w^TX^T-Y^T)(Xw-Y)$

$=w^TX^TXw-\underset{1\times p}{w^T}\underset{p\times N}{X^T}\underset{N\times 1}{Y}-\underset{1\times N}{Y^T}\underset{N\times p}{X}\underset{p\times 1}{w}+Y^TY$ 每一项都是一维的 所以可以得到 $w^TX^TY=Y^TXw$

$=w^TX^TXw-2w^TX^TY+Y^TY$

$\hat{w}=\arg\min L(w)$

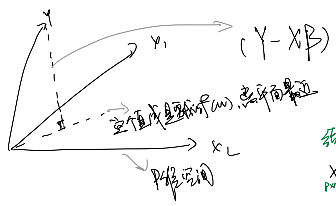
$\frac{\partial L(w)}{\partial w}=2X^TXw-2X^TY=0$ 矩阵/向量求导公式

$X^TXw=X^TY$

$$\boxed{w=(X^TX)^{-1}X^TY}$$

一般也会把 $(X^TX)^{-1}X^T$ 记作 "X^+" 伪逆

$f(w)=w^Tx=x^T\beta$ → 把误差分散在p个维度上



$X=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{pmatrix}_{N\times p}$ 从 X^T 角度来看，P个向量（维度）

结论 若 $\vec{a}\perp\vec{b}$ 则 $\vec{a}^T\cdot b=0$

$\underset{p\times N}{X^T}\cdot(\underset{N\times 1}{Y}-\underset{N\times p}{X}\underset{p\times 1}{\beta})=\underset{p\times 1}{0}$ → 这里0是p×1的0向量 $\begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$

$X^T\cdot Y=X^TX\beta$

$$\boxed{\beta=(X^TX)^{-1}X^TY}$$

从概率角度：最小二乘法 ⇔ noise为Gaussian 的MLE

$D=\{(x_1,y_1),(x_2,y_2),\cdots(x_N,y_N)\}\quad x_i\in\mathbb{R}^p\quad y_i\in\mathbb{R}\quad i=1,2\cdots N$

$X=(x_1,x_2\cdots x_N)^T=\begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix}=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{pmatrix}$

$Y=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$

最小二乘估计 $L(w)=\sum_{i=1}^{N}\|w^Tx_i-y_i\|^2$

$\hat{w}=\arg\min_w L(w)$

$\hat{w}=(X^TX)^{-1}X^TY$

设噪声 $\varepsilon\sim N(0,\sigma^2)$

$y=f(w)+\varepsilon=w^Tx+\varepsilon$

$y|x;w\sim N(w^Tx,\sigma^2)\ \rightarrow\ P(y|x;w)=\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{(y-w^Tx)^2}{2\sigma^2}\}$

用 MLE

$$L(w) = \log P(Y|X;w) = \log \prod_{i=1}^{N} P(y_i|x_i;w) = \sum_{i=1}^{N} \log P(y_i|x_i;w)$$

(log likelihood; 假定同方差(i.i.d)可以写成连乘)

$$= \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} + \log \exp\left\{-\frac{(y_i - w^T x_i)^2}{2\sigma^2}\right\}$$

$$= \sum_{i=1}^{N} \left(\log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2}(y_i - w^T x)^2\right)$$

$$\hat{w} = \arg\max_w L(w)$$

$$= \arg\max_w \sum -\frac{1}{2\sigma^2}(y_i - w^T x_i)^2$$ (Σ log 1/√2πσ 可以直接消去(不含 w))

$$= \arg\min_w \sum \frac{1}{2\sigma^2}(y_i - w^T x_i)^2$$

最小二乘估计为 $L(w) = \sum_{i=1}^{N} \|w^T x_i - y_i\|^2$ $\hat{w} = \arg\min_w L(w)$ 两者完全一致

正则化

Loss function $L(w) = \sum_{i=1}^{N} \|w^T x_i - y_i\|^2$

$$w^T = (WX)^{-1} W^T Y$$

过拟合 → ① 加数据 ② 降维(特征选择/特征提取 PCA) ③ 正则化.

正则化框架 $\arg\min_w [L(w) + \lambda P(w)]$ (loss; penalty)

L1: LASSO $P(w) = \|w\|$

L2: Ridge $P(w) = \|w\|_2 = w^T w$ (权值衰减)

$$J(w) = \sum_{i=1}^{N} \|w^T x_i - y_i\|^2 + \lambda w^T w$$ (矩阵形式)

$$= (w^T X^T - Y^T)(Xw - Y) + \lambda w^T w$$

$$= (w^T X^T X w - 2 w^T X^T Y + Y^T Y) + \lambda w^T w$$

$$= w^T (X^T X + \lambda I) w - 2 w^T X^T Y + Y^T Y$$

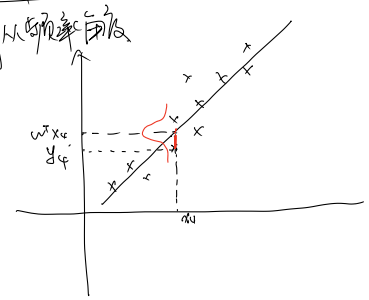
$$\hat{w} = \arg\min_w J(w)$$

$$\frac{\partial J(w)}{\partial w} = 2(X^T X + \lambda I) w - 2 X^T Y = 0$$

$$\hat{w} = (X^T X + \lambda I)^{-1} X^T Y$$

从概率角度看 Ridge Regression (岭回归)

之前从频率角度



这是从贝叶斯角度来看.

$w \sim N(0, \sigma_0^2)$ 参数w的先验概率

$p(w|y) = \frac{p(y|w)p(w)}{p(y)}$

$p(y|w) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\vec{w}x)^2}{2\sigma^2}\right)$

$p(w) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{\|w\|^2}{2\sigma_0^2}\right)$

$p(y|w)\cdot p(w) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \frac{1}{\sqrt{2\pi}\sigma_0} \cdot \exp\left\{-\frac{(y-\vec{w}x)^2}{2\sigma^2} - \frac{\|w\|^2}{2\sigma_0^2}\right\}$

MAP: 最大后验估计

$\hat{w} = \arg\max_w p(w|y)$

$= \arg\max_w p(y|w)p(w)$ $p(y)$常量

$= \arg\max_w \log[p(y|w)p(w)]$ 这是为了计算方便常做的

$= \arg\max_w \log\left(\frac{1}{\sqrt{2\pi}\sigma}\cdot\frac{1}{\sqrt{2\pi}\sigma_0}\right) + \log\exp\left\{-\frac{(y-\vec{w}x)^2}{2\sigma^2} - \frac{\|w\|^2}{2\sigma_0^2}\right\}$

$= \arg\min_w \left(\frac{(y-\vec{w}x)^2}{2\sigma^2} + \frac{\|w\|^2}{2\sigma_0^2}\right)$

$= \arg\min_w \left((y-\vec{w}x)^2 + \frac{\sigma^2}{\sigma_0^2}\|w\|^2\right)$

$= \arg\min_w \sum_{i=1}^{N} \left((y_i - \vec{w}^T x_i)^2 + \frac{\sigma^2}{\sigma_0^2}\|w\|^2\right)$ 也可以加上λ

loss function ↓ penalty 乘 λ

Regularized LSE ⇔ MAP (noise Gaussian Distribution, 先验也是GP)