线性回归：最小二乘法

$D=\{(x_i,y_i)\}$ $x_i\in\mathbb{R}^p$ p维列向量 $y_i\in\mathbb{R}$ $i=1,2,\cdots,N$ N个样本点

数据集D可以表示为

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

p维向量

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

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

$$=\underbrace{(W^Tx_1-y_1\quad W^Tx_2-y_2\quad\cdots\quad W^Tx_N-y_N)}_{\text{行向量}}\underbrace{\begin{pmatrix}W^Tx_1-y_1\\ W^Tx_2-y_2\\ \vdots\\ W^Tx_N-y_N\end{pmatrix}}_{\text{列向量}}$$

$$=(W^Tx_1\ \ W^Tx_2\ \cdots\ W^Tx_N)-(y_1\ y_2\ \cdots\ y_N)\ (\text{列})$$

$$=W^T(x_1\ x_2\ \cdots\ x_N)-(y_1\ y_2\ \cdots\ y_N)\ (\text{列})$$

$$=(W^TX^T-Y^T)(XW-Y)$$

$$=W^TX^TXW-W^TXY-YXW^T+Y^TY$$

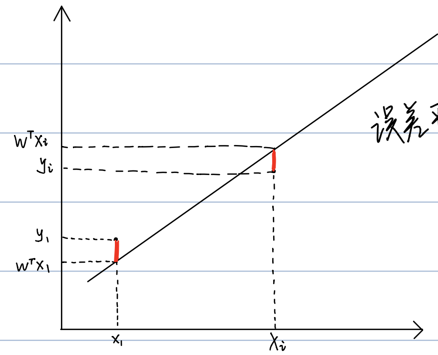
$$=W^TX^TXW-2W^TX^TY+Y^TY$$

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

求导数 $\frac{\partial}{\partial W}L(W)=2X^TXW-2X^TY$

令 $2X^TXW-2X^TY=0$ 得到 $W=(X^TX)^{-1}X^TY=X^{\dagger}Y$

最小二乘法几何意义



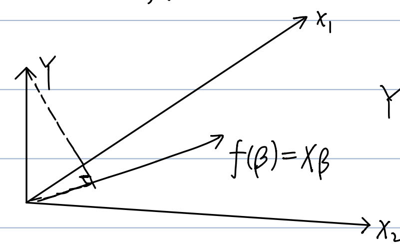
误差平均分散在N个样本上

由 $W=X^+Y$ 得到 $W\in\mathbb{R}^{p\times 1}$

$X^T\in\mathbb{R}^{p\times N}$ $Y\in\mathbb{R}^{N\times 1}$ $\beta\in\mathbb{R}^{p\times 1}$

误差分散在p个维度上

再假设 $f(w)=w^Tx=X\beta$



Y独立于X构成的p维空间

最小二乘法：找到真值Y在 $\underbrace{\text{p维子空间}}_{x_1,x_2,\cdots,x_N\text{的线性组合}}$ 的投影

$X^T\underbrace{(Y-X\beta)}_{\text{相当于法向量}}=0$ $\quad X^TY=X^TX\beta \quad$ 所以 $\beta=(X^TX)^{-1}X^TY=X^+Y$

最小二乘法概率视角

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

p维向量

最小二乘估计损失函数 $L(w)=\sum_{i=1}^{N}\|w^Tx_i-y_i\|_2^2$ $\qquad \hat{W}=(X^TX^{-1})X^TY$

假设 $y=f(w)+\varepsilon$（量化模型与真值的误差 ε，且要求 $\varepsilon\sim N(0,\sigma^2)$）

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

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

log-likelihood

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

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

$\hat{w} = \arg\max_w L(w) = \arg\min_w \sum_{i=1}^{N} (y_i - w^T x_i)^2 \longrightarrow$ 极大似然估计 ≡ 最小二乘

结论：least square estimation ≡ maximum likelihood estimation
当且仅当 noise 服从高斯分布

线性回归：正则化

线性回归引入正则化的原因：

① $\hat{W} = (X^T X)^{-1} X^T Y$ 难以得到解析解. if $N \gg p$

② 过拟合.（样本数太少）$\longrightarrow$
- 添加样本
- 特征选择
- 特征提取（降维 e.g.）
- 正则化（对参数空间 w 进行约束）

正则化 $\arg\min_w \underbrace{L(W)}_{\text{loss}} + \lambda \underbrace{P(w)}_{\text{penalty}}$

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

L2: Ridge, $P(w) = W^T W = \|W\|^2$

权值衰减

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

$$= \begin{pmatrix} w^T x_1 - y_1 & w^T x_2 - y_2 & \cdots & w^T x_N - y_N \end{pmatrix} \begin{pmatrix} w^T x_1 - y_1 \\ \vdots \\ w^T x_N - y_N \end{pmatrix} + \lambda\|w\|^2$$

$$= \left(w^T (x_1\ x_2\ \cdots\ x_N) - (y_1\ y_2 \cdots\ y_N) \right) \begin{pmatrix} w^T x_1 - y_1 \\ \vdots \\ w^T x_N - y_N \end{pmatrix} + \lambda\|w\|^2$$

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

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

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

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

令导数为0，$\hat{w} = \underbrace{(X^T X + \lambda I)^{-1}}_{\text{半定阵 对角阵}} X^T Y$

假设 $w \sim p(w) = N(0, \sigma_0^2)$，$y = f(w) + \varepsilon$

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

$$\hat{w} = \underset{w}{\arg\max}\ p(w|y) = \underset{w}{\arg\max}\ \underbrace{P(y|w)}_{\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y-w^Tx)^2}{2\sigma^2}\right)}\ \underbrace{P(w)}_{\frac{1}{\sqrt{2\pi}\sigma_0}\exp\left(-\frac{w^2}{2\sigma_0^2}\right)}$$

最大后验估计

$$= \underset{w}{\arg\max}\ \frac{1}{2\pi}\frac{1}{\sigma\sigma_0}\exp\left(-\frac{(y-w^Tx)^2}{2\sigma^2} - \frac{w^2}{2\sigma_0^2}\right)$$

$$= \underset{w}{\arg\max}\ -\frac{(y-w^Tx)^2}{2\sigma^2} - \frac{w^2}{2\sigma_0^2}$$

$$= \underset{w}{\arg\min}\ \underbrace{(y-w^Tx)^2 + \frac{\sigma^2}{\sigma_0^2} w^2}$$

Ridge Regression!

结论：① LSE $\iff$ MLE 极大似然估计

② Regularized LSE $\iff$ MAP 最大后验估计

} noise is Gaussian