## Linear Regression

2021年4月26日 20:24

### P9. Least Squre Method and it's Geometic Meaning

$$X=\begin{pmatrix}x_1^T\\x_2^T\\\vdots\\x_N^T\end{pmatrix}=\begin{pmatrix}x_{11}&x_{12}&\cdots&x_{1p}\\x_{21}&x_{22}&\cdots&x_{2p}\\\vdots&\vdots&&\vdots\\x_{N1}&x_{N2}&\cdots&x_{Np}\end{pmatrix}_{N\times p}\quad Y=\begin{pmatrix}y_1\\y_2\\\vdots\\y_N\end{pmatrix}_{N\times 1}\quad w=\begin{pmatrix}w_1\\w_2\\\vdots\\w_p\end{pmatrix}_{p\times 1}\quad \underset{1\times p\ \ p\times 1}{f(w)=w^Tx} \to \text{each data}\quad L(w)=\sum_{i=1}^{N}(w^Tx_i-y_i)^2$$

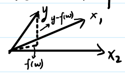
$$L(w)=(w^Tx_1-y_1\ \cdots\ w^Tx_N-y_N)\begin{pmatrix}w^Tx_1-y_1\\\vdots\\w^Tx_N-y_N\end{pmatrix}=\underbrace{[\underset{1\times N}{w^T}\underset{N\times p}{(x_1 x_2\cdots x_N)}-\underset{p\times 1}{(y_1 y_2\cdots y_N)}]}_{\Delta}\Delta^T\quad \Delta=w^TX^T-Y^T$$

$\therefore L(w)=(w^TX^T-Y^T)(Xw-Y)=w^TX^TXw-w^TX^TY-Y^TXw-Y^TY$ $\quad w^TX^TY=(Y^TXw)^T$ are constants, so they are equal

$\therefore L(w)=w^TX^TXw-2w^TX^TY-Y^TY \quad \frac{d}{dw}L(w)=2X^TXw-2X^TY=0$

$2X^TXw=2X^TY \quad w=(X^TX)^{-1}X^TY \quad (\because X^T=X^{-1}\ \because X[(X^TX)^{-1}X^T]X=X \therefore (X^TX)^{-1}X^T$ is called pseudoinverse or generalized inverse (伪逆或广义逆))

Geometric Meaning: ① Splitting the error to each data. LS minimizes the sum of the distance between obsevered values and theoretical values




② Splitting the error to each dimensions of attributes $\quad f(w)=w^Tx=Xw$

All dimensions form a vector space: $\underset{N\times 1}{f(w)}=\underset{N\times p\ \ p\times 1}{Xw}$

$Y$ doesn't lie in this vector space unless all data fit. LS is to find the closest $f(w)$ to $Y$ in the vector space

The $f(w)$ that minimizes $Y-f(w)$ $\quad Y-f(w)$ is perpendicular to all $X$ $\therefore X^T(Y-Xw)=0 \therefore X^TY=X^TXw \quad w=(X^TX)^{-1}X^TY \leftarrow$ analytic solution

numerical solution $\Rightarrow$ GD

### P10. Probabilistic Perspective LS

When the noise obeys Gaussian distribution $\quad$ LSE (Least Square Estimation) $\Longleftrightarrow$ MLE (Maximum Likeihood Estimation) $\quad$ LSE: $\hat{w}=\arg\min L(w) \quad L(w)=\sum_{i=1}^{N}\|w^Tx_i-y_i\|^2$

MLE: $\quad y=w^Tx+\varepsilon \quad \varepsilon\sim N(0,\sigma^2) \quad y|x,w\sim(w^Tx,\sigma^2) \quad P(y|x,w)=\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{(y-w^Tx)^2}{2\sigma^2}\}$

$l(w)=\log P(y|x,w)=\sum_{i=1}^{N}\left[\log\frac{1}{\sqrt{2\pi}\sigma}-\frac{1}{2\sigma^2}(y-w^Tx)^2\right] \quad \hat{w}=\arg\max_w l(w)$

$\hat{w}=\arg\min_w(y-w^Tx)^2=LSE$

### P11. Regularization – Ridge Regression – Frequentists (频率学派)

$X_{N\times p}$, $N$ samples, $P$ attributes Generally $N\gg P$, otherwise it may be over-fitted

The $(X^TX)^{-1}$ in $\hat{w}=(X^TX)^{-1}X^TY$ may not exist ① Add data ② Feature Selection/Extraction

③ Regularization: $\hat{w}=\arg\min_w[L(w)+\lambda P(w)]$ → penalty (惩罚项)

$L_1$: Lasso, $P(w)=\|w\|$ $\qquad L_2$: Ridge, 岭回归, $P(w)=\|w\|^2$

$J(w)=(w^TX^T-Y^T)(Xw-Y)-\lambda w^Tw=w^T(X^TX+\lambda I)w-2w^TX^TY+Y^TY$

$\hat{w}=\arg\min_w[J(w)] \quad \frac{\partial}{\partial w}J(w)=2(X^TX+\lambda I)w-2X^TY=0 \quad \hat{w}=(X^TX+\lambda I)^{-1}X^TY$

### P12. Regularization – Ridge Regression – Bayesians (贝叶斯学派)

Regulazation: As above ①

Bayesian Perspective: $y=w^Tx+\varepsilon \quad$ Let $\varepsilon\sim N(0,\sigma^2)$, $y\sim N(w^Tx,\sigma^2)$, $w\sim N(0,\sigma_0^2)$

$P(w|y)=\frac{P(y|w)P(w)}{P(y)} \quad$ Maximum a Posteriori Estimation, MAP: $\hat{w}=\arg\max_w[P(y|w)P(w)]$

$p(y|w)=\sum_{i=1}^{N}\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{(y_i-w^Tx_i)^2}{2\sigma^2}\} \quad P(w)=\frac{1}{\sqrt{2\pi}\sigma_0}\exp\{-\frac{\|w\|^2}{2\sigma_0^2}\}$

$\hat{w}=\arg\max_w\{\log[P(w|y)p(y)]\}=\arg\max_w\left\{\log\frac{1}{\sqrt{2\pi}\sigma}\frac{1}{\sqrt{2\pi}\sigma_0}-\left[\sum_{i=1}^{N}\frac{(y_i-w^Tx_i)^2}{2\sigma^2}+\frac{\|w\|^2}{2\sigma_0^2}\right]\right\}$

$=\arg\min_w\left[\underbrace{\sum_{i=1}^{N}(y_i-w^Tx_i)^2}_{\text{Loss}}+\underbrace{\frac{\sigma^2}{\sigma_0^2}}_{\lambda}\underbrace{\|w\|^2}_{\text{penalty}}\right]\Longleftrightarrow$ ①

### Supplement to P11, P12

LSE $\Longleftrightarrow$ MLE (When the noise obeys Gaussian distribution)

Regularized (正则化) LSE $\Longleftrightarrow$ MAP (When the noise and the priori obey the Gaussian distribution)