Ridge regression

Normal regression: $\hat{\Theta} = (X^T X)^{-1} X^T Y$

$(p+1)\cdot n \quad n\cdot(p+1)$

Now, $rank(X) < p+1$, then $rank(X^T X) < (p+1)$

which means

$$X^T X \Longrightarrow \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1(p+1)} \\ 0 & a_{22} & \cdots & a_{2(p+1)} \\ \vdots & & \ddots & 0 \\ 0 & \cdots & & 0 \end{bmatrix}$$

Therefore $(X^T X)^{-1}$ does not exist

Solution: Add a λ on diagnol. to $X^T X$

$$X^T X + \lambda I \Rightarrow \begin{bmatrix} a_{11} & \cdots & a_{1(p+1)} \\ \vdots & \ddots & 0 \\ 0 & \cdots & 0 \end{bmatrix} + \lambda \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \vdots & 1 & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & 1 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} a_{11}+\lambda & \cdots & & a_{1(p+1)} \\ 0 & a_{22}+\lambda & \cdots & a_{2(p+1)} \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & \lambda \end{bmatrix}$$

$$rank(X^T X + \lambda I) = p+1$$

Therefore $(X^T X + \lambda I)^{-1}$ does exist.

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

How to understand ridge ?

Going backwards though the proof.

$$SSE = (Y - X\hat{\theta})^T (Y - X\hat{\theta}) + \lambda \theta^T \theta$$

$$\frac{\delta SSE}{\delta \theta} = \frac{\delta [(Y - X\hat{\theta})^T (Y - X\hat{\theta})] + \lambda \delta [\theta^T \theta]}{\delta \theta}$$

$$\Downarrow$$

$$\frac{\delta SSE}{\delta \theta} = -2X^T Y + 2 [X^T X + \lambda I] \hat{\theta}$$

加上λI
从下往上
推导

New cost Function $SSE = (Y - X\hat{\theta})^T (Y - X\hat{\theta}) + \lambda \theta^T \theta$

$$= (Y - X\hat{\theta})^T (Y - X\hat{\theta}) + \lambda [\theta_0 \cdots \theta_p] \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_p \end{bmatrix}$$

$$= (Y - X\hat{\theta})^T (Y - X\hat{\theta}) + \lambda \sum_{i=0}^{p} \theta_i^2$$

$$\hat{\theta} = \underset{\hat{\theta}}{\arg\min} \left[(Y - X\hat{\theta})^T (Y - X\hat{\theta}) + \lambda \sum_{i=0}^{p} \theta_i^2 \right]$$

We want to find a $\hat{\theta}$, such that SSE is minimal.

From Lagrange (拉格朗日优化). The Above could be write us the following

$$\hat{\theta}_{ridge} = \arg\min [(Y - X\hat{\theta})^T (Y - X\hat{\theta})]$$

subject to $\sum_{i=1}^{0} \theta_i^2 \leq t^2$

where $t^2 \uparrow$ then $\lambda \downarrow$

$\lambda \uparrow$ then $t^2 \downarrow$

Understand Ridge From graph.

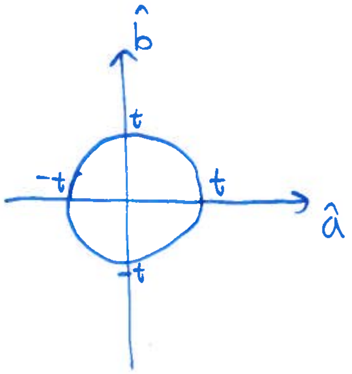
For 2-D Data sets

$$Y = aX + b \,, \quad SSE = \sum (y_i - \hat{a}x_i - \hat{b})^2$$

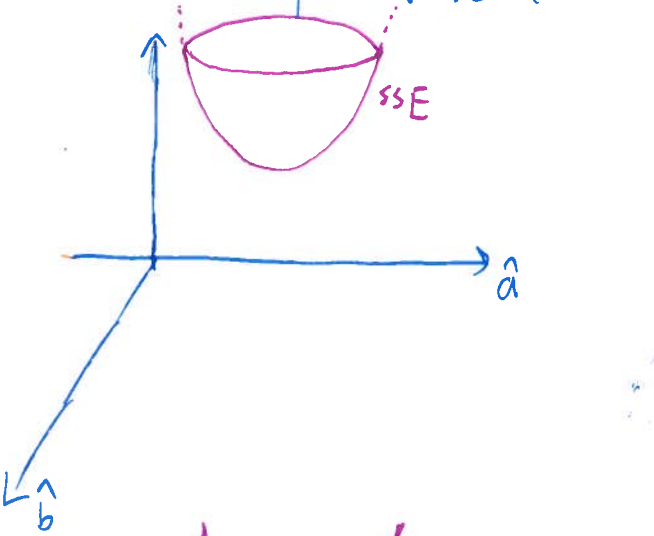
$$\hat{\Theta} = \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \underset{(\hat{a},\hat{b})}{\operatorname{argmin}} \sum (y_i - \hat{a}x_i - \hat{b})^2$$

$$\text{subject to} \quad \hat{a}^2 + \hat{b}^2 \le t^2$$

$$\hat{a}^2 + \hat{b}^2 = t^2$$

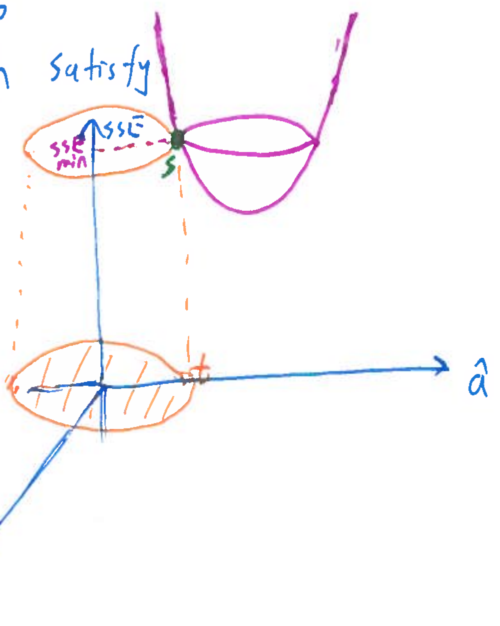


SSE could be proved to be a convex



We are looking for $(\hat{a}, \hat{b})$, which satisfy

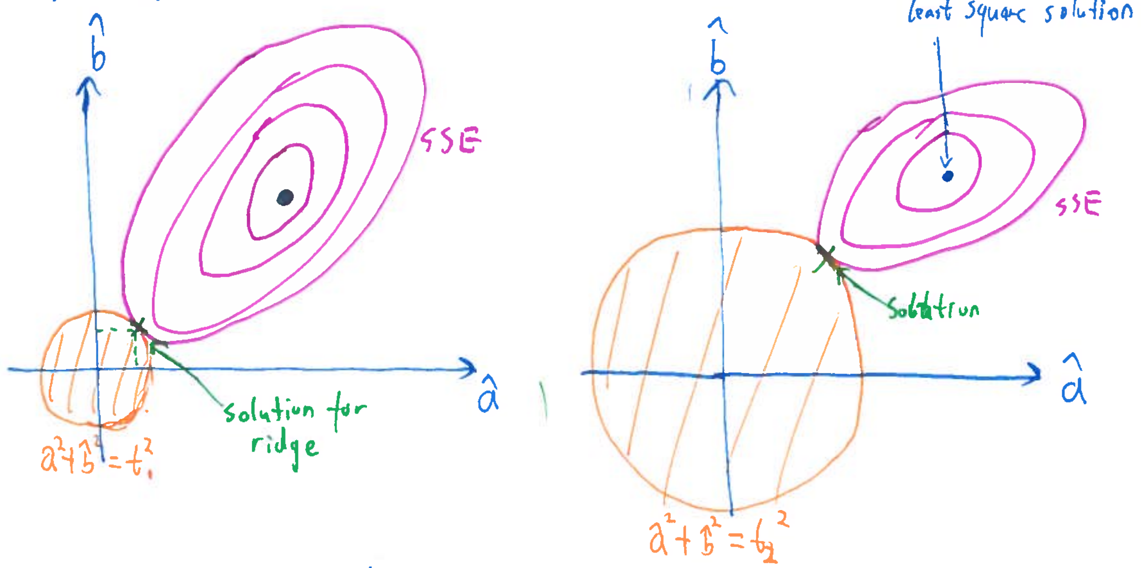
1. $\hat{a}^2 + \hat{b}^2 = t^2$
2. $SSE = \sum (y_i - \hat{a}x_i - \hat{b})^2$



We could find it in the graph.

point S is the solution

Ridge Regression Projection



$$\hat{\Theta}^{ridge} = \underset{\hat{\Theta}}{\operatorname{argmin}} (Y - X\hat{\Theta})^T (Y - X\hat{\Theta}) + \lambda \sum \Theta_i^2$$

$$\Updownarrow$$

$$\hat{\Theta}^{ridge} = \operatorname{argmin} (Y - X\hat{\Theta})^T (Y - X\hat{\Theta}) \quad \text{subject to } \sum \Theta_i^2 \leq t^2$$

Note, when $\lambda \uparrow$, $t^2 \downarrow$

when $\lambda \downarrow$, $t^2 \uparrow$

when t^2 decrease, λ increase

$\hat{a} \to 0$, $\hat{b} \to 0$

when t^2 increase, λ decrease

$[\hat{a}, \hat{b}]$ approach least square solution

Therefore, λ acts like a penalty.

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

when $\lambda \uparrow$, $\hat{\Theta}$ approach 0.

when $\lambda \to 0$, $\hat{\Theta}$ approach the true least square solution $(X^T X)^{-1} X^T Y$

Lasso Regression

$$\hat{\Theta} = \text{argmin} \left[(Y - X\hat{\Theta})^T (Y - X\hat{\Theta}) \right] + \lambda \Sigma |\theta_i|$$

$$\Updownarrow$$

$$\hat{\Theta} = \underset{\hat{\Theta}}{\text{argmin}} \left[(Y - X\hat{\Theta})^T (Y - X\hat{\Theta}) \right]$$

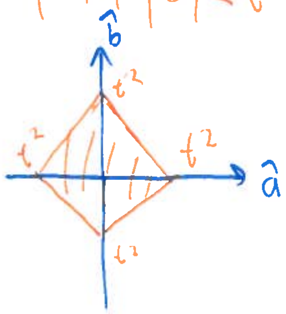
subject to $\Sigma |\theta_i| \le t^2$

when $\lambda \uparrow$, $t^2 \downarrow$

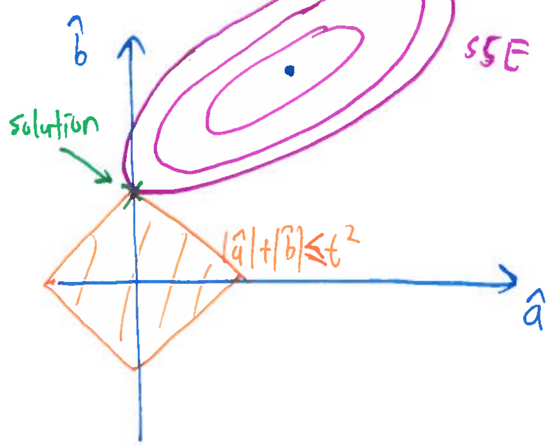
$\lambda \downarrow$, $t^2 \uparrow$

For 2-D

$|\hat{a}| + |\hat{b}| \le t^2$



SSE projections on 2-D



Therefore, for Lasso, SSE is very likely to touch the pointy point. which implies, some Θ_i in the solution are 0.

Solution of Lasso is not discussed here because it is kind of complicated

Derivation of coordinated descent for Lasso

Lasso cost function [coordinated descent] (安点式证法)

$$SSE(\theta) = SSE^{OLS}(\theta) + \lambda \|\theta\|$$

$$= \frac{1}{2}\sum_{i=1}^{n}\left[y_i - \sum_{j=0}^{p}\theta_j x_{ij}\right]^2 + \lambda\sum_{j=0}^{p}|\theta_j|$$

where do they come from

OLS: original_least_square

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \underset{i:[1,n]}{=} \begin{array}{c} j:[0,p] \\ \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & x_{ij} & \vdots \\ 1 & x_{n1} & \cdots & \cdots & x_{np} \end{bmatrix} \end{array} \times \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_p \end{bmatrix}$$

i: sample $\quad$ j: feature

For ith sample: $y_{pred} = \hat{\theta}_0 + x_{i1}\hat{\theta}_1 + x_{i2}\hat{\theta}_2 + \cdots x_{ip}\hat{\theta}_p$

error^2 for ith sample: $[y_{true} - y_{pred}]^2$

error^2 sum: $\sum (y_{true} - y_{pred})^2 = \sum_{i=1}^{n}\left[y_i - \sum_{j=0}^{p}\theta_j x_{ij}\right]^2$

Lasso penalty $\quad \lambda\|\theta\| = \lambda\sum_{j=0}^{p}|\theta_j|$

$$= \lambda\left[|\theta_0| + |\theta_1| + |\theta_2| + \cdots |\theta_p|\right]$$

---

we want to solve

$$\frac{\delta SSE(\theta)}{\delta\theta} = \frac{\delta\left[\frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p}\theta_j x_{ij}\right)^2 + \lambda\sum|\theta|\right]}{\delta\theta} = 0$$

OLS part:

$$\frac{\delta SSE^{OLS}(\theta)}{\delta\theta} = \frac{\delta}{\delta\theta_j}\left[\frac{1}{2}\sum_{i=1}^{n}\left(y_i-\sum_{j=0}^{p}\theta_j x_{ij}\right)^2\right]$$

$$= -\sum_{i=1}^{n} x_{ij}\left[y_i-\sum_{j=0}^{p}\theta_j x_{ij}\right]$$

$$= -\sum_{i=1}^{n} x_{ij}\left[y_i-\sum_{k\neq j}^{p}\theta_k x_{ik}-\theta_j x_{ij}\right]$$

$$= -\sum_{i=1}^{n} x_{ij}\left[y_i-\sum_{k\neq j}^{p}\theta_k x_{ik}\right]+\theta_j\sum_{i=1}^{n} x_{ij}^2$$

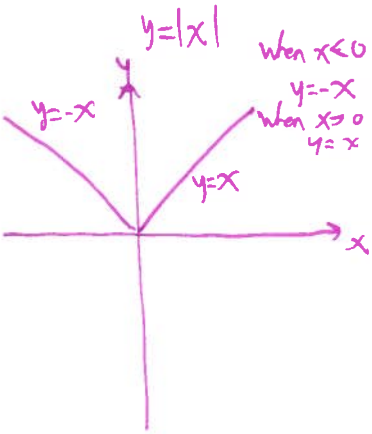
$$\triangleq -\rho_j+\theta_j z_j$$

where we define ρ_j and z_j, which the z_j is

Lasso part.

$$\lambda\sum_{j=0}^{p}|\theta_j| = \lambda|\theta_j| + \lambda\sum_{k\neq j}^{n}|\theta_k|$$

$$\frac{\delta[\lambda\Sigma|\theta|]}{\delta\theta_j} = \frac{\delta[\lambda|\theta_j|]}{\delta\theta_j} + \frac{\delta\left[\lambda\sum_{k\neq j}^{n}|\theta_k|\right]}{\delta\theta_j}$$



$$= \frac{\delta[\lambda|\theta_j|]}{\delta\theta_j} + 0$$

$$= \begin{cases} -\lambda & \text{if } \theta_j<0 \\ [-\lambda,\lambda] & \text{if } \theta_j=0 \\ \lambda & \text{if } \theta_j>0 \end{cases}$$

Putting OLS and Lasso together

$$\frac{\delta\, SSE(\theta)}{\theta_j} = -p_j + \theta_j z_j + \frac{\delta_\theta \lambda \|\theta_j\|}{\delta \theta_j} = 0$$

$$0 = \begin{cases} -p_j + \theta_j z_j - \lambda & \text{if } \theta_j < 0 \\ [-p_j - \lambda, -p_j + \lambda] & \text{if } \theta_j = 0 \\ -p_j + \theta_j z_j + \lambda & \text{if } \theta_j > 0 \end{cases}$$

note: $\theta \in [-p_j - \lambda, -p_j + \lambda]$

$-\lambda \leq p_j \leq \lambda.$

final solution

$$\begin{cases} \theta_j = \dfrac{p_j + \lambda}{z_j} & \text{for } p_j < -\lambda \\ \theta_j = 0 & \text{for } -\lambda \leq p_j \leq \lambda \\ \theta_j = \dfrac{p_j - \lambda}{z_j} & \text{for } p_j > \lambda. \end{cases}$$

where $p_j = \sum_{i=1}^{n} x_{ij} \left[y_i - \sum_{k \neq j}^{p} \theta_k x_{ik} \right]$

$$z_j = \sum_{i=1}^{n} x_{ij}^2$$

How to do Lasso using python

Do it using step-wise way.

Initial $\theta = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$ or $\theta = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$

$\lambda = 0.01$

def solution(ρ_j, λ):

$$\theta_j = \begin{cases} \rho_j + \lambda, & \rho_j < -\lambda \\ 0 & -\lambda \le \rho_j \le \lambda \\ \rho_j - \lambda, & \rho_j > \lambda \end{cases}$$

for _ in [1, 50000]:

for each col j in X:

$$X_{-j} = X[:, j] = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix}$$

$$Y_{pred} = X \cdot \theta = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix}$$

$$\rho_j = \sum_{i=1}^{n} x_{ij} \left[y_i - \sum_{k \neq j}^{p} \theta_k x_{ik} \right]$$

$$= [x_{1j} \ x_{2j} \cdots x_{nj}] \left\{ \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} \hat{y}_1 \\ \vdots \\ 0 \text{ (when } k=j) \\ \vdots \\ \hat{y}_n \end{bmatrix} \right\}$$

$$= [x_{1j} \cdots x_{nj}] \left\{ \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix} + \begin{bmatrix} 0 \\ \vdots \\ \hat{y}_j \\ \vdots \\ 0 \end{bmatrix} \right\}$$

$$= [x_{1j} \cdots x_{nj}] \left\{ \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix} + \theta_j \begin{bmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{bmatrix} \right\}$$

$$= X_{-j}^T \cdot (y - \hat{y} + \theta[j] \cdot X_{-j})$$

$\theta[j]$ = Solution(ρ_j, λ)