Design Matrix

$$A = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{m-1} \\ \vdots & x_2 & x_2^2 & \cdots & x_2^{m-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1 & x_n & x_n^2 & & x_n^{m-1} \end{bmatrix}_{n \times m}$$

Weight Vector

$$w = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_{m-1} \end{bmatrix}_{m \times 1}$$

target Vector

$$b = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1}$$

◎ LSE is we hope

Find a w to make

$Aw = b$

But A is not invertible (perhaps not)

→ We only find $\tilde{w}$

Such $Aw = b$ to $\min \|Aw - b\|^2$

$$\|Aw-b\|^2 = (Aw-b)^T(Aw-b)$$
$$= w^TA^TAw - 2\underline{w^TA^Tb} + b^Tb$$

☆ w^TA^Tb is "Scalar" not a Vetocr

So, $(w^TA^Tb)^T = b^TAw = w^TA^Tb$

$$\min_{w_{LSE}} \|Aw-b\|^2 \Rightarrow \frac{\partial}{\partial w}(w^TA^TAw) - \frac{\partial}{\partial w}(2^TA^Tb) + \frac{\partial}{\partial w}(b^Tb)$$
$$\Rightarrow 2A^TAw - 2A^Tb + 0$$
$$\Rightarrow A^TAw = A^Tb$$
$$\Rightarrow w_{LSE} = (A^TA)^{-1}A^Tb \ \#$$

But $w = (A^TA)^{-1}A^Tb \Rightarrow$ We only know $\det(A^TA)^{-1} \geq 0$

Can't 100% make sure full rank (invertable)

We must add regularized term "L_1 norm" or "L_2 norm"

To 100% make sure full rank (invertable)

We usually call rLSE,

We choose L2 norm $= \lambda\|w\|^2 \Rightarrow \frac{\partial}{\partial w} L_2 = 2\lambda w$

$\min_{w} \|Aw-b\|^2 \Rightarrow 2A^TAw - 2A^Tb + 2\lambda w = 0$

$\Rightarrow (A^TA+\lambda I)w = A^Tb \Rightarrow w_{rLSE} = (A^TA+\lambda I)^{-1}A^Tb$ #

while $\lambda > 0$, $A^TA+\lambda I$ 100% $\det > 0$, inertable #

Both LSE / rLSE we all need to calculate

Inverse Matirx, I will choose LU

Because only we ① slove I each row e_i

To slove $LUx_i = e_i$ ② Assemble together we

Let A^{-1} For example: $A = LU \Rightarrow A^{-1} = U^{-1}L^{-1}$

$A^{-1} = [x_1, x_2 \cdots x_n]$, x_i is a column vector

By $AA^{-1} = I \Rightarrow Ax_i = e_i$, $i = 1 \cdots n$ $\Rightarrow$ ① Solve y_i

$\Rightarrow Ax_i = e_i \Rightarrow LUx_i = e_i$, let $Ux_i = y_i$ ② Slove x_i

Summary
① Foward substitution $Ly_i = e_i$, Slove y_i
② Backward substitution $Ux_i = y_i$, slove x_i

$A^{-1} = [x_1, x_2 \cdots x_n]$ → We can get A^{-1} #

⑥ Steepest decent

The Update fomula of steepest decent

$$W_{t+1} = W_t - \eta \nabla f(W_t)$$

W_t is parameter of t steps
η is learning rate (Steps size)
$\nabla f(W_t)$ is W_t's gradient

Lipschitz Contiuns with $L > 0$ (L is Lipschitz constant)
This make sure the change will not so fierce

From $f(W_t)$ we expansion of f (f is Loss function) by Tayler

$$f(W_{t+1}) \leq f(W_t) + \nabla f(W_t)^T (W_{t+1} - W_t) + \frac{L}{2} \|W_{t+1} - W_t\|^2$$

$$= f(W_t) - \eta \|\nabla f(W_t)\|^2 + \frac{L\eta^2}{2} \|\nabla f(W_t)\|^2$$

$$= f(W_t) - \eta \left(1 - \frac{L\eta}{2}\right) \|\nabla f(W_t)\|^2$$

$\|W_{t+1} - W_t\|$ must small enough, which implies η also must small enough

while $\eta \leq \frac{L}{2}$, this is monotonically increasing (unless $\nabla f(W_t) = 0$)

$\Rightarrow$ Best learning rate : $\eta = \frac{1}{L}$

$\Rightarrow f(W_{t+1}) \leq f(W_t) - \frac{1}{2L} \|\nabla f(W_t)\|^2$ this is discription of convergance properties

Regularized term in L_1 norm

the gradient of it can be written

as sign funtion $sign(W_i) = \begin{cases} 1, \text{ if } W_i > 0 \\ -1, \text{ if } W_i < 0 \\ 0, \text{ if } W_i = 0 \end{cases}$

The gradient in total is $2A^T(A\vec{w}-b) + \lambda \cdot sign(\vec{w})$

---

⑥ Newton's method

Basic type:

Based Talyers expesion

$$f(x) \approx f(x_0) + \nabla f(x_0)^T (x - x_0) + \frac{1}{2}(x - x_0)^T H(x_0)(x - x_0)$$

$H(x_0) = \nabla^2 f(x_0)$ is Hessian Matrix

Newton's method optimization problem :

Same We wanna slove $\min_{w} f(w)$

We must find $\nabla f(w) = 0$

$$\nabla f(w_t) + \underline{H(t)}(w - w_t) = 0$$

Slove step : $w - w_t = -H(t)^{-1} \nabla f(w_t)$

Update fomula : $w_{t+1} = w_t - H(t)^{-1} \nabla f(w_t)$

For LSE problem

We already knew $f(w) = \|Aw - b\|^2$

$$\nabla f(w) = 2A^T(Aw - b)$$

$$\nabla^2 f(w) \triangleq H(w) = \underline{2A^T A} \quad (\text{it not related with } w\,!)$$

Apply Update fomula $w_{t+1} = w_t - (2A^T A)^{-1} \cdot 2A^T(Aw_t - b)$

$$\Rightarrow w_{t+1} = w_t - (A^T A)^{-1} A^T (Aw_t - b)$$

Special propeties of Newton's method

For any quadratic func (ie. LSE)

One step is enough!

$$
\begin{aligned}
W_1 &= W_0 - (A^TA)^{-1}A^T(AW_0 - b) \\
&= W_0 - (A^TA)^{-1}A^TAW_0 + (A^TA)^{-1}A^Tb \\
&= W_0 - W_0 + (A^TA)^{-1}A^Tb \\
&= (A^TA)^{-1}A^Tb
\end{aligned}
$$

Does it look familer ~

Because this is optimal Solution of LSE, as shown in previous

We calcuted #