⊛ Regularization *

Q. What is regularization and where should we use it?

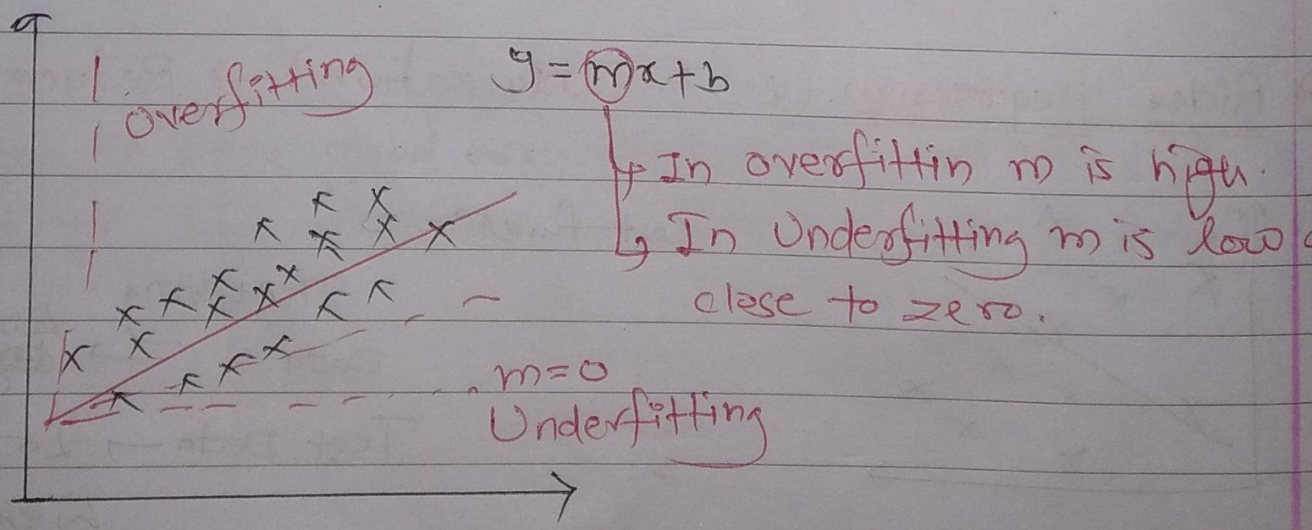
→ It is a technique that are used to calibrate machine Learning models in order to minimize the adjusted loss function and prevent overfitting or under-fitting.

⊛ Techniques to deal with overfitting

① Bagging ② Boosting ③ Regularization

⊛ Overfitting :-

It is phenomenon or criteria where ML model performing exceptionally well on training data, but not so well on the testing data. (this means we have high varience in our ~~dataset~~ model.)



Date: YOUVA

⋆ Ridge - Regularization :-

Loss funcn

$$L = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 + \lambda (\text{slope})^2 \longrightarrow (1)$$

Differentiate the above equ^n w.r.t. b.

Let's rewrite the eq^n

$$L = \sum_{i=1}^{n} (y_i - mx_i - b)^2 + \lambda (m)^2$$

$$\frac{\partial L}{\partial b} = \frac{\partial \Sigma (y_i - mx_i - b)^2}{\partial b} + \underbrace{\lambda (m)^2}_{\text{const w.r.t } b}$$

$$\Rightarrow 2\Sigma (y_i - mx_i - b)(-1) + 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i - m \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} b = 0$$

$$\Rightarrow n\bar{y} - mn\bar{x} - nb = 0$$

$$\Rightarrow \bar{y} - m\bar{x} - b = 0$$

$$\Rightarrow \boxed{b = \bar{y} - m\bar{x}}$$

$\bar{y}$ - mean of y
$\bar{x}$ - mean of x
m - slope.

Now, put this b in eq^n (1)

$$\therefore L = \sum_{i=1}^{n} (y_i - mx_i - \bar{y} + m\bar{x})^2 + \lambda m^2$$

Differentiat w.r.t. m & equat it to zero

$$\frac{\partial L}{\partial m} = 0 \Rightarrow \frac{\partial}{\partial m} \sum_{i=1}^{n} (y_i - mx_i - \bar{y} + m\bar{x})^2 + \lambda m^2 = 0$$

$$\Rightarrow 2 \sum_{i=1}^{n} (y_i - mx_i - \bar{y} + m\bar{x})(-x_i + \bar{x}) + 2m\lambda = 0$$

$$\Rightarrow -2\sum_{i=1}^{n}(y_i-\bar{y}-mx_i+m\bar{x})(x_i-\bar{x})+2\lambda m=0$$

$$\Rightarrow \lambda m-\sum_{i=1}^{n}(y_i-\bar{y}-mx_i+m\bar{x})(x_i-\bar{x})=0$$

$$\Rightarrow \lambda m-\sum_{i=1}^{n}[(y_i-\bar{y}-m(x_i-\bar{x})](x_i-\bar{x})=0$$

$$\Rightarrow \lambda m-\sum_{i=1}^{n}[(y_i-\bar{y})(x_i-\bar{x})-m(x_i-\bar{x})^2]=0$$

$$\Rightarrow \lambda m-\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})-m\sum_{i=1}^{n}(x_i-\bar{x})^2=0$$

$$\Rightarrow \lambda m+m\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})$$

$$\Rightarrow m\left[\lambda+\sum_{i=1}^{n}(x_i-\bar{x})^2\right]=\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})$$

$$\Rightarrow \boxed{m=\frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2+\lambda}}$$

⊛ if we ↑ λ then m will ↓

x_1	x_2	$\cdots$	x_n	y → (n+1)
↓	↓		↓	
w_1	w_2	$\cdots$	w_n	
m rows				

m values

$$L=\sum_{i=1}^{m}(y_i-\hat{y}_i)^2$$

$$=(Xw-y)^T(Xw-y)$$

$$y=\begin{bmatrix}-\\-\\-\\-\\-\\-\end{bmatrix} \quad w=\begin{bmatrix}w_0\\w_1\\w_2\\\vdots\\w_n\end{bmatrix}_{n\times 1} \quad X=\begin{bmatrix}1 & x_{11} & x_{12} & \cdots & x_{1n}\\1 & x_{21} & x_{22} & \cdots & x_{2n}\\1 & x_{31} & x_{32} & \cdots & x_{3n}\\\vdots & & & & \\1 & x_{m1} & x_{m2} & \cdots & x_{mn}\end{bmatrix}$$

$$L = (X\theta - Y)^T (X\theta - Y) + \lambda \|\theta\|^2$$

$$\lambda\theta_0^2 + \lambda\theta_1^2 + \lambda\theta_2^2 + \cdots + \lambda\theta_n^2$$
$$\lambda [\theta_0^2 + \theta_1^2 + \theta_2^2 + \cdots + \theta_n^2]$$

We can rewrite as $\theta^T\theta$

$$[\theta_0 \ \theta_1 \ \theta_2 \cdots \theta_n] \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix}$$

$$L = (X\theta - Y)^T (X\theta - Y) + \lambda \theta^T\theta$$

$$L = [(X\theta)^T - (Y)^T] (X\theta - Y) + \lambda\theta^T\theta \quad [\because (a-b)^T = a^T - b^T]$$

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

$$= (\theta^T X^T X\theta - \theta^T X^T Y - Y^T X\theta + Y^T Y + \lambda\theta^T\theta$$

(both same)

$$L = \underbrace{\theta^T X^T X\theta}_{B} - 2\underbrace{\theta^T X^T Y}_{B} + \underbrace{Y^T Y}_{\text{const}} + \lambda\theta^T\theta$$

$$\frac{dL}{d\theta} = 2X^T X\theta - 2X^T Y + 0 + 2\lambda\theta = 0$$

$\{$ $\because X^T B X \rightarrow 2BX$
$\because X^T B \rightarrow B$
$\because X^T X \rightarrow 2X$ $\}$ → vector derivative

$$\Rightarrow X^T X\theta + \lambda\theta = X^T Y$$
$$\Rightarrow (X^T X + \lambda I)\theta = X^T Y$$

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

In LR $w = (X^T X)^{-1} X^T y$ → $n \times 1$, $n \times 1$

⊛ Ridge with Gradient Descent :-

$$L = (Xw - y)^T (Xw - y) + \lambda \|w\|^2$$

$$L = (Xw - y)^T (Xw - y) + \lambda w^T w$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2n} \\ 1 & & & & & \\ \vdots & & & & & \\ 1 & x_{m1} & x_{m2} & x_{m3} & \cdots & x_{mn} \end{bmatrix}_{m \times n}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}_{m \times 1}, \quad w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}_{n \times 1}$$

$w_0, w_1, w_2, \cdots, w_n$ (parameters)

if we want to update

$$w_0 = w_0 - \eta \frac{\partial L}{\partial w_0}, \quad w_1 = w_1 - \eta \frac{\partial L}{\partial w_1}$$

$$\cdots, \quad w_n = w_n - \eta \frac{\partial L}{\partial w_n}$$

$$w_{new} = w_{old} - \eta \boxed{\frac{\Delta L}{\Delta w}} \text{ gradient} \quad \begin{bmatrix} \partial L/\partial w_0 \\ \partial L/\partial w_1 \\ \partial L/\partial w_2 \\ \vdots \\ \partial L/\partial w_n \end{bmatrix}$$

multiply $\frac{1}{2}$ through the loss and write it again

$$L = \frac{1}{2}(Xw - y)^T (Xw - y) + \frac{1}{2}\lambda w^T w$$

$$= \frac{1}{2}(W^T X^T - Y^T)(XW - Y) + \frac{1}{2}\lambda W^T W$$

~~$\frac{1}{2}(W^T X^T - Y^T)$~~

$$= \frac{1}{2}\left[W^T X^T XW - W^T X^T Y - Y^T XW + Y^T Y\right] + \frac{1}{2}\lambda W^T W$$

(same)

$$= \frac{1}{2}\left[W^T X^T XW - 2 W^T X^T Y + Y^T Y\right] + \frac{1}{2}\lambda W^T W$$

$$\frac{dL}{dW} = \frac{1}{2}\left[2X^T XW - 2X^T Y\right] + \frac{1}{2}2\lambda W$$

$$\frac{dL}{dW} = X^T XW - X^T Y + \lambda W = \frac{\partial L}{\partial W}$$

Starting pt $W = [W_0, W_1, W_2, \ldots, W_n]$

$$= [0, 1, 1, \ldots, 1]$$

epochs $W = W - \eta \frac{dL}{dW}$

$$\boxed{\frac{dL}{dW} = X^T XW - X^T Y + \lambda W}$$

⊛ 5 Key points Regarding Ridge Reg.

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \boxed{\lambda \|W\|^2}$$

$\rightarrow \lambda(W_1^2 + W_2^2 + \ldots + W_n^2)$

{ shrinkage coef

↳ Reduce overfitting.

Date:

(1) How the coef gets affected ? if $\lambda \uparrow$

$w_1, w_2, \ldots, w_n \rightarrow$ coef get closer to 0, but not will be 0.

(2) Higher values are impacted more.

$x_1, x_2, x_3 \uparrow$

↳ Normal LR

$w_1 = 1000 \quad w_2 = 10 \quad w_3 = 1$

In Ridge $\lambda \rightarrow \infty$

$w_1 = 1000 \downarrow$ Rapidly

$w_2 = 10 \downarrow$ slowly than w_1

$w_1 = 1 \downarrow$ slowly than w_1 and w_2

(3) Bias Varience Tradeoff

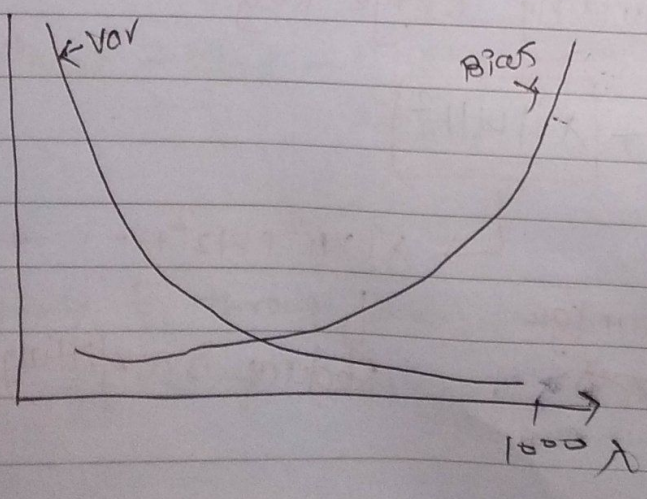
Bias & varience both depends on the value of (λ)

if $\lambda \rightarrow 0$ then,

Bias ↓ Overfitting Varience ↑

if $\lambda \rightarrow \infty$ then,

Bias ↑ Underfitting Varience ↓



④ Impact on the loss-function of λ ?

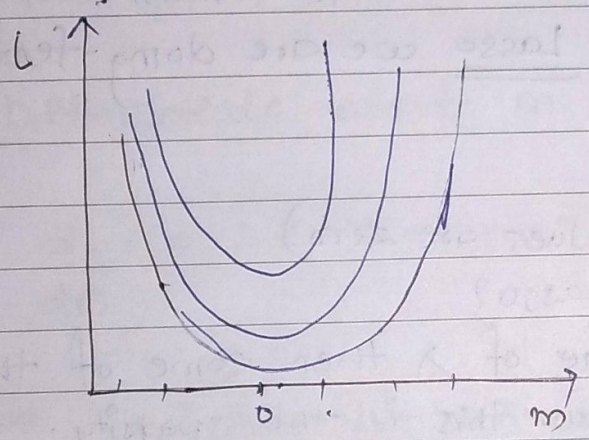
$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \|w\|^2$$

I/P: X, Y | O/P: [m, b] — b is const

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \|w\|^2$$

Where $b = 0$ & slope $= m$

$$\therefore L = \sum_{i=1}^{n} (y_i - mx_i)^2 + \lambda m^2$$



$\lambda = 1$

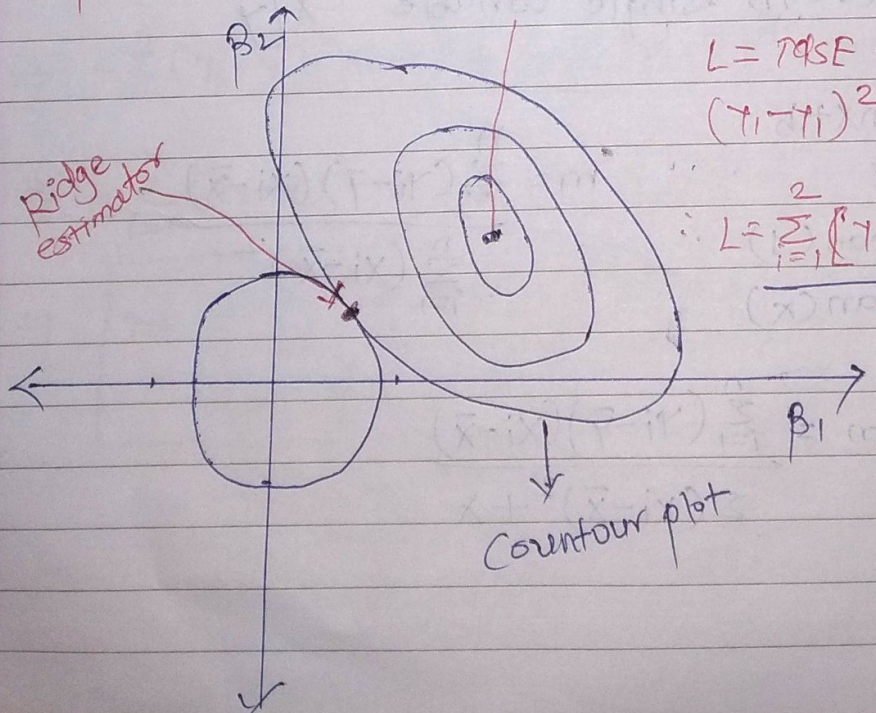
change λ

The graph will shrink and minima shift towards 0.

⑤ Why called Ridge.

we need to study

Hard constrain Ridge Reg.



$L = MSE + \lambda \|w\|^2$

$(y_i - \hat{y}_i)^2$ $\beta_0, \beta_1, \beta_2$

$$\therefore L = \sum_{i=1}^{2} [y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})]^2 + \lambda (\beta_1 + \beta_2)^2$$

MSE

→ slope

Date:

Practical Tip

Apply when more than 2 I/p variables.

⊛ Lasso Regression * (L1)

Loss function

$L = MSE + \lambda \|w\|$

$\hookrightarrow [|w_1| + |w_2| + \cdots + |w_n|]$

⊛ Which one is most prefarable among lasso and Ridge.

→ Lasso → Becase for high value of λ it will automatically bring the coef zero whose corresponding variable are less important. it will reduce the dimension of the data that with the help of Lasso we are doing feature selection

* Lasso sparsity (some values are zero)

⊛ What is sparsity in Lasso?

→ if we increase the value of λ then some of the coefficients will be zero. this is the sparsity.

* Why sparsity in Lasso and why not in Ridge?

→ Lets consider in single variable X/Y

$Y = mX + b$

$b = \bar{y} - m\bar{x}$

$\bar{y}$ - mean(y)

$\bar{x}$ - mean(x)

$$m = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Ridge - Case.

$$m = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\Sigma(x_i - \bar{x})^2 + \lambda}$$

Page No.:
Date:

Lasso case ⇒ $b = \bar{y} - m\bar{x}$

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda |m|$$

$$L = \Sigma (y_i - mx_i - b)^2 + \lambda |m| \quad [\because \hat{y}_i = x_i m + b$$

$$L = \sum_{i=1}^{n} (y_i - mx_i - \bar{y} + m\bar{x})^2 + \lambda |m|$$

$\{\because b = \bar{y} - m\bar{x}\}$

Lets consider $m > 0$

$\therefore |m| = m$

$$\therefore L = \sum_{i=1}^{n} (y_i - mx_i - \bar{y} + m\bar{x})^2 + 2\lambda m$$

Differentiate w.r.t. m & equat to 0.

$$\frac{dL}{dm} = 2\Sigma (y_i - mx_i - \bar{y} + m\bar{x})(-x_i + \bar{x}) + 2\lambda = 0$$

$$\Rightarrow -2 \sum_{i=1}^{n} [(y_i - \bar{y}) - m(x_i - \bar{x})](x_i - \bar{x}) + 2\lambda = 0$$

$$\Rightarrow -\sum_{i=1}^{n} [(y_i - \bar{y})(x_i - \bar{x}) - m(x_i - \bar{x})^2] + \lambda = 0$$

$$\Rightarrow -\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) + m \sum_{i=1}^{n} (x_i - \bar{x})^2 + \lambda = 0$$

$$\Rightarrow m \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) - \lambda$$

$$\Rightarrow \boxed{m = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) - \lambda}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

Date:

Lasso

for $m>0$

$$m = \frac{\Sigma(y_i-\bar{y})(x_i-\bar{x}) - \lambda}{\Sigma(x_i-\bar{x})^2}$$

for $m=0$

$$m = \frac{\Sigma(y_i-\bar{y})(x_i-\bar{x})}{\Sigma(x_i-\bar{x})^2}$$

for $m<0$

$$m = \frac{\Sigma(y_i-\bar{y})(x_i-\bar{x}) + \lambda}{\Sigma(x_i-\bar{x})^2}$$

for $m>0$ & $\lambda>0$

$$m = \frac{yx - \lambda}{x^2} \quad \begin{cases} yx = 100 \\ x^2 = 50 \end{cases}$$

$$m = \frac{100-\lambda}{50}$$

$\lambda=0$	$\lambda=10$	$\lambda=50$	$\lambda=100$
$m=2$	$m=\frac{9}{5}$	$m=1$	$m=0$

if $\lambda>100 \Rightarrow m=-1$

Now we need to use $m<0$ formula. lets take $\lambda=150$

$$\therefore m = \frac{yx+\lambda}{x^2} = \frac{100+150}{50} = \frac{250}{50} = 5$$

2, 9/5, 1, 0, -1, 5.

-ve

Now, $m<0$ $\quad \lambda>0$

$$m = \frac{\Sigma(y_i-\bar{y})(x_i-\bar{x}) + \lambda}{\Sigma(x_i-\bar{x})^2}$$

$$m = \frac{-100+\lambda}{50}$$

$\lambda=0$	$\lambda=50$	$\lambda=100$	$\lambda=150$
$m=-2$	$m=-1$	$m=0$	$m=1$

$$m = \frac{\Sigma(y_i-\bar{y})(x_i-\bar{x}) - 1}{\Sigma(x_i-\bar{x})}$$

Ridge

$$m = \frac{\Sigma(y_i-\bar{y})(x_i-\bar{x})}{\Sigma(x_i-\bar{x})^2 + \lambda} \qquad \lambda \to \uparrow$$

→ This is denom term.

In order to becom $m=0$ num should be zero. But Here this is not possible. that is why

In Ridge coeff become close to zero, but not exactly 0. whereas, In Lasso the λ is numerators term that is why in Lasso coeff value will be zero.

* Elasticnet Regression *

* Its a combination of Ridge and Lasso.

$$L = \sum_{i=1}^{n}(y_i - \hat{y})^2 + \lambda \|w\|^2$$

Ridge

$w_1^2 + w_2^2 + \dots + w_n^2$

(Overfitting Reduce)

$\lambda \uparrow (\infty)$ $w \downarrow (0)$ but not 0

L2 norm

Use where All col are imp.

$$L = \sum_{i=1}^{n}(y_i - \bar{y})^2 + \lambda \|w\|$$

Lasso

$|w_1| + |w_2| + \dots + |w_n|$

$\lambda \uparrow$ $w \to 0$

feature selection

L1 norm

Use where some of cols imp.

* if we have big data and there are 1000 of cols in this case we cannot predict which cols are imp. so we use Elasticnet Reg here.

$$L = \sum_{i=1}^{n}(y_i - \hat{y}_i) + a\|w\|^2 + b\|w\|$$

λ, μ

In sklearn $\lambda = a + b$ μ_ratio $= \frac{a}{a+b}$

By default $\lambda = 1$, μ_ratio $= 0.5$

$\mu = a/\lambda$

$a = \mu \times \lambda$

$b = \lambda - a$

* if there is multicolinearity in input cols then we should use Elasticnet Reg

eg. x_1 (hight) / x_2 (weight)