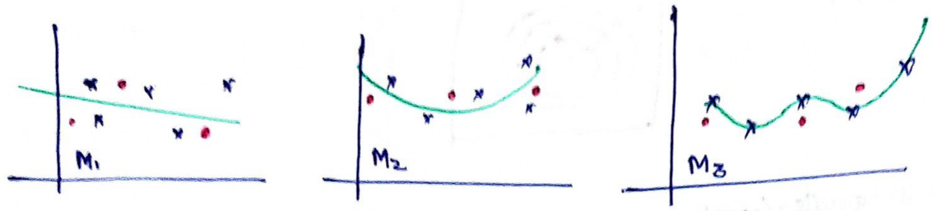
Bias variance Trade off

1158 Am Sat Apr



Bias

Inability of manchine learning model true capture the Relationship in the Training data

	M_1	M_2	M_3
Bias	High	mid	low
variance	low	Mid	H
	underfiting	Sweet spot	overfiting

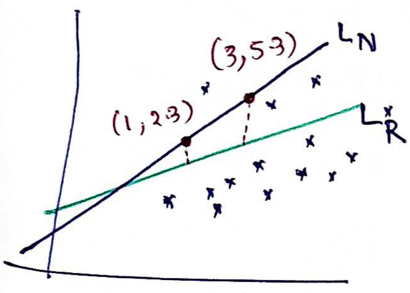
variance → train data error − test data error

Sweet spot →
1. Regularization
2. Bagging
3. Bossting

Ridge Regularization

P-1

1. Ridge (L2)
2. Lasso (L1)
3. Elastic Net



$Y = 1.5X + 0.8$

$Y = 0.9 + 1.5$

hyperparameter

$m \to \infty$ y ∝ x overfitting

$$J = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \lambda(m^2)$$

$$= \sum (Y_i - mx_i - b)^2 + \lambda(m^2)$$

$Y = mx + b$

$J = b$ → underfit

Loss L_N

$\lambda = 1$

line is pass through both point so zero

$0 + (1.5)^2$

$= 2.25$

loss L_R

$\lambda = 1$

$= (2.3 - 0.9 - 1.5)^2 + (5.3 - 0.9 - 1.5)^2 + 1(0.9)^2$

$= (0.1)^2 + (1.1)^2 + (0.9)^2$

$= 2.03$

we want to say our ML focus on other data bec. that is best fit line

? How can we do

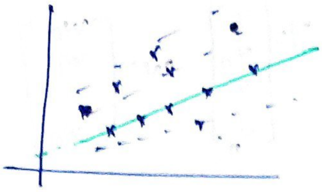
- we add some extra dulo
- calculate loss

Now ML can understand L_R have less loss so it choose L_R.

more axis.

$\lambda(m_1^2 + m_2^2 + \dots m_n^2)$

Ridge Regression



?

$y = mx + b$

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

no change

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

$\frac{\partial J}{\partial b} = 0$

$\frac{\partial J}{\partial m} = 0$

add m so here is changes.

after $\frac{\partial J}{\partial b} = 0$

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

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

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

$$= -\cancel{2}\sum (y_i - \bar{y} - m x_i + m\bar{x})(x_i - \bar{x}) + \cancel{2}\lambda m = 0$$

$$= \lambda m - \sum \left[(y_i - \bar{y}) - m(x_i - \bar{x})\right](x_i - \bar{x}) = 0$$

$$= \lambda m - \sum \left[(y_i - \bar{y})(x_i - \bar{x}) - m(x_i - \bar{x})^2\right]$$

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

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

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

Extra

simple LR

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

Ridge Regression for nData

$x_1 \quad x_2 \quad \cdots \quad x_n$

m rows $\downarrow$ $\quad w_1 \quad w_2 \quad \cdots \quad w_n$

$$J = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$= (XW - Y)^T (XW - Y)$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y \\ \vdots \\ y_n \end{bmatrix} \quad W = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & & & & \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

$$W^T W = [w_0 w_1\ w_2 \cdots w_n] \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix}$$

$$J = (XW - Y)^T (XW - Y) + \lambda \|W\|^2 \quad \{\lambda w_0^2 + \lambda w_1^2 \cdots \lambda w_n^2\}$$
$$= (XW - Y)^T (XW - Y) + \lambda W^T W$$
$$= (W^T X^T - Y^T)(XW - Y) + \lambda W^T W$$
$$= W^T X^T X W - W^T X^T Y - Y^T X W + Y^T Y + \lambda W^T W$$
$$= W^T \underbrace{X^T}_{\text{cont}} X W - 2 W^T X^T Y + Y^T Y + \lambda W^T W$$

$(a-b)^T = a^T - b^T$

$(ab)^T = b^T a^T$

matrix derivative

$$\frac{dJ}{dW} = \cancel{2} X^T X W - \cancel{2} X^T Y + 0 + \cancel{2} \lambda W = 0$$
$$\Rightarrow X^T X W + \lambda W = X^T Y$$
$$\Rightarrow W (X^T X + \lambda I) = X^T Y$$
$$\Rightarrow W = (X^T X + \lambda I)^{-1} (X^T Y)$$

$I_{(n+1) \times (n+1)}$

simple LR

$W = (X^T X)^{-1} X^T Y$

$$I = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

solver='cholesky'

Ridge Regression using Gradient Descent

$$J = \sum (y_i - \hat{y}_i)^2 + \lambda \|w\|^2$$

$$= (XW - Y)^T (XW - Y) + \lambda W^T W$$

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

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

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

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

$$= X^T X W - X^T Y + \lambda W$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & & & & \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$W = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix}_{n+1}$$

$$w_0 = w_0 - \eta \frac{\partial L}{\partial w_0}$$

$$w_1 = w_1 - \eta \frac{\partial L}{\partial w_1}$$

$$\vdots$$

$$w_n = w_n - \eta \frac{\partial L}{\partial w}$$

$$w_{new} = w_{old} - \eta \frac{\Delta L}{\Delta W}$$

P-4

$$\sum (y_i - \hat{y}_i)^2 + \underline{\lambda \|w\|^2}$$

$$\hookrightarrow \lambda(w_1^2 + w_2^2 - - - w_n^2)^2$$

(shrinkage coef)

1. Affect of coefficient in srink & trends to 0.

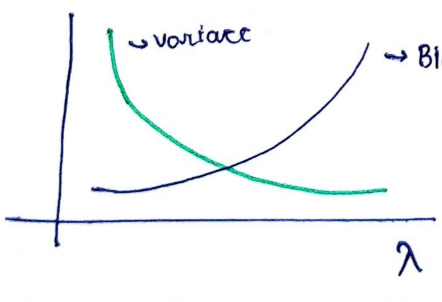
2. Higer values are impacted more

x_1	x_2	xw	Y
w_1	w_2	w_3	
1000	10	1	

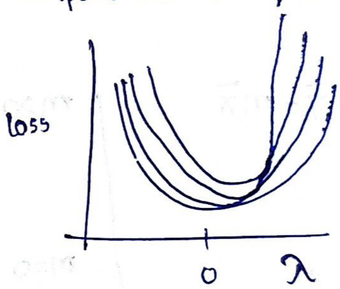
This srinks more

3. Bias variance trade off

$$\lambda \propto \frac{1}{Bias} \propto \frac{1}{variance}$$



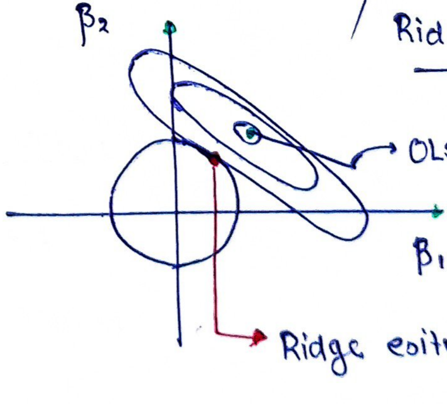
4. Impact on Loss func $\lambda \uparrow$ coef $\downarrow$



- it reduce & srink
- then shift towards to zero.

5. why called Ridge

Hard constrain
Ridge construction



Lasso Regression

D056

6:25pm
Sat Apr 26

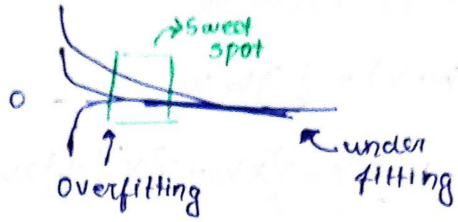
$J = MSE + \lambda|w|$

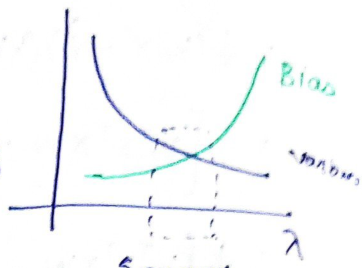
$\hookrightarrow \lambda[\,|w_1| + |w_2| + |w_3| \cdots + |w_n|\,]$

absolute

$\lambda\uparrow$ coeficient will zero.. in Lasso

- Benifits less imp. feature is zero.
- Apply the feature selection
- dimionenlity reduction.





Sparsity

Simple

$x/y \rightarrow y = mx + b$

$b = \bar{y} - m\bar{x}$ $\qquad m = \dfrac{\sum (y_i - \bar{y}_i)(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$

Ridge

$b = \bar{y} - m\bar{x}$ $\qquad m = \dfrac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2 + \lambda}$

Lasso

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

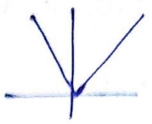
$m > 0 \qquad m = \dfrac{\sum (y_i - \bar{y})(x_i - \bar{x}) - \lambda}{\sum (x_i - \bar{x})^2}$

$m = 0 \qquad m = \dfrac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$

$m < 0 \qquad m = \dfrac{\sum (y_i - \bar{y})(x_i - \bar{x}) + \lambda}{\sum (x_i - \bar{x})^2}$

to make simple

$$\frac{\partial J}{\partial m} = \sum (y_i - mx_i - \bar{y} + m\bar{x})^2 + 2\lambda|m|$$



$|m| \begin{cases} m>0 & +m \\ m<0 & -m \end{cases}$

$$= \sum (y_i - mx_i - \bar{y} + m\bar{x})^2 + 2\lambda m$$

$$= 2\sum (y_i - mx_i - \bar{y} + m\bar{x})(-x_i + \bar{x}) + 2\lambda$$

$$= -\cancel{2}\sum \left[(y_i - \bar{y}) - m(x_i - \bar{x})\right](x_i - \bar{x}) + \cancel{2}\lambda$$

$$= -\sum \left[(y_i - \bar{y})(x_i - \bar{x})\right] + m\sum (x_i - \bar{x})^2 + \lambda$$

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

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

?

- why at the end all values are zero ?
- why not values are -ve ?
- why sparsity ?

because Ridge λ is deu so ≈ 0.
lasso λ is neu so $= 0$.

let $m > 0$ $\quad \frac{K - \lambda}{K'}$ $\qquad K = \sum (y_i - \bar{y})(x_i - \bar{x})$

$K' = \sum (x_i - \bar{x})^2$

$K = 100$

$K' = 50$

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

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

Then it use

$\frac{K + \lambda}{K'}$

$\downarrow$

$\frac{100 + 100}{50} =$

$m = 4$

$m \longrightarrow$

$2 \quad \frac{9}{5} \quad 1 \quad 0 \quad -1$
$\downarrow$
4

we want decease the m
so why use
we stop after zero.

D057

732pm
sat AM 26

Elastic Net Regression

Ridge	$J = MSE + \lambda\|\|W\|\|^2$	all I/P in imp.
Lasso	$J = MSE + \lambda\|W\|$	use feature selection.
EN	$J = MSE + a\|\|W\|\|^2 + b\|W\|$	multicollinarity

$$J = \sum (Y_i - \bar{Y}_i)^2 + a\|\|W\|\|^2 + b\|\|W\|\|$$

default

$\lambda = 1$

$l1 = 0.5$

$a = 0.5$
$b = 0.5$

if $l1 = 0.9$,
90% Ridge 10% Lasso

In sklearn

$\lambda = a + b$

$l1\_ratio = \frac{a}{a+b}$?

$a = l1 \times \lambda$

$b = \lambda - a$