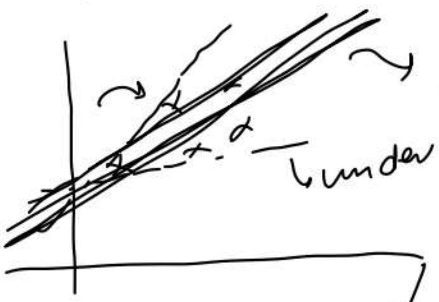
Ridge Regression

Wednesday, June 2, 2021 11:13 AM



bias ↑ variance ↓

$$y = mx + b$$

2D lines

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

m slope

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

m <

$\frac{\partial L}{\partial b} = 0 \rightarrow b$

$\frac{\partial L}{\partial m} = 0 \rightarrow m$

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

$\bar{y} \rightarrow$ y-mean

$\bar{x} \rightarrow$ x_mean

m → slope

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

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

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

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

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

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

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

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

$\lambda \rightarrow$ hyperparameter, alpha

$\lambda = 0 = \lambda = 10,100$

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

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

Ridge Regression for 2D data

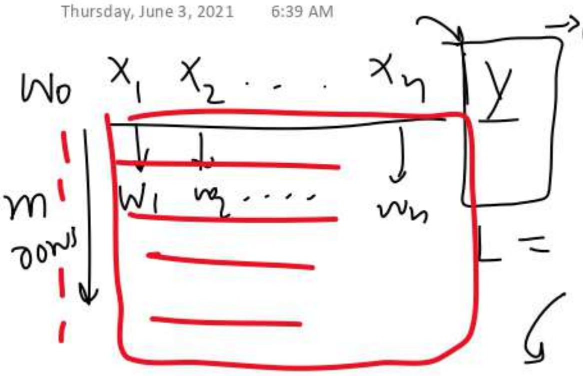
Thursday, June 3, 2021 6:38 AM

Code

Thursday, June 3, 2021 6:39 AM

Ridge Regression for nD data

Thursday, June 3, 2021 6:39 AM



$\rightarrow (n+1)$

$\rightarrow$ Ridge

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

$$= (XW - Y)^T (XW - Y)$$

m vals

$$Y = \begin{bmatrix} - \\ - \\ - \\ - \\ - \end{bmatrix} \quad W = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

$(n \times 1)$

Normal LR $\longrightarrow$ Ridge

$$L = \boxed{(XW - Y)^T (XW - Y) + \lambda \|W\|^2}$$

$\lambda w_0^2 + \lambda w_1^2 + \lambda w_2^2 + \ldots + \lambda w_n^2$

$$[w_0\ w_1\ w_2 \ldots w_n] \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix} \qquad \underline{W^T W} \longrightarrow \lambda(w_0^2 + w_1^2 + w_2^2 \ldots w_n^2)$$

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

$\dfrac{dL}{dW}$ $\qquad (ab)^T = b^T a^T$

$$L = \underline{(XW - Y)^T (XW - Y)} + \lambda W^T \underline{W}$$

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

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

$$= W^T X^T X W - \underbrace{W^T X^T Y - Y^T X W}_{1.} + Y^T Y + \lambda W^T W$$

$$L = w^T X^T X w - 2 w^T X^T y + y^T y + \lambda w^T w$$

$$\frac{dL}{dw} = \cancel{2} X^T X w - \cancel{2} X^T y + 0 + \cancel{2} \lambda w = 0$$

$$X^T X w + \lambda w = X^T y$$

$$(X^T X + \lambda I) w = X^T y$$

3 $(n \times n)$

$(n \times 1, n \times 1)$

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

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

$$w = (X^T X)^{-1} X^T y$$

Code

Thursday, June 3, 2021 6:39 AM

Ridge Regression using Gradient Descent

Friday, June 4, 2021 2:17 PM

Vector form loss

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

$$L = (XW - Y)^T (XW - Y) + \lambda \|W\|^2$$

$$\boxed{L = (XW - Y)^T (XW - Y) + \lambda W^T W}$$

$x_1\ x_2 \dots x_n\ Y$

$\to X$

m rows

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ 1 & x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ 1 & x_{m1} & x_{m2} & x_{m3} & \dots & x_{mn} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} \to (n+1)$$

$\to W$

$w_0, w_1 \dots w_n$ (parameters)

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

$$W_{new} = W_{old} - \eta \boxed{\frac{\Delta L}{\Delta W}} \to \text{gradient} \begin{bmatrix} \frac{\partial L}{\partial w_0} \\ \partial L / \partial w_1 \\ \vdots \\ \partial L / \partial w_n \end{bmatrix}$$

L, W

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

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

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

$= \frac{1}{2} [W^T X^T X W \ldots]$

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

$2 W^T X^T y$

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

$X^T y$

$$= \boxed{X^T X W - \cancel{y^T X} + \lambda W} = \frac{dL}{dW} \quad \left(\frac{\partial L}{\partial W} \right)$$

$X^T y$

$$\begin{array}{cccc} w_0 & w_1 & \cdots & w_n \\ \end{array}$$

$$W = [0, 1 \ldots\ldots 1]_{nm} \quad \text{starting}$$

epochs

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

$W \rightarrow$ final answer

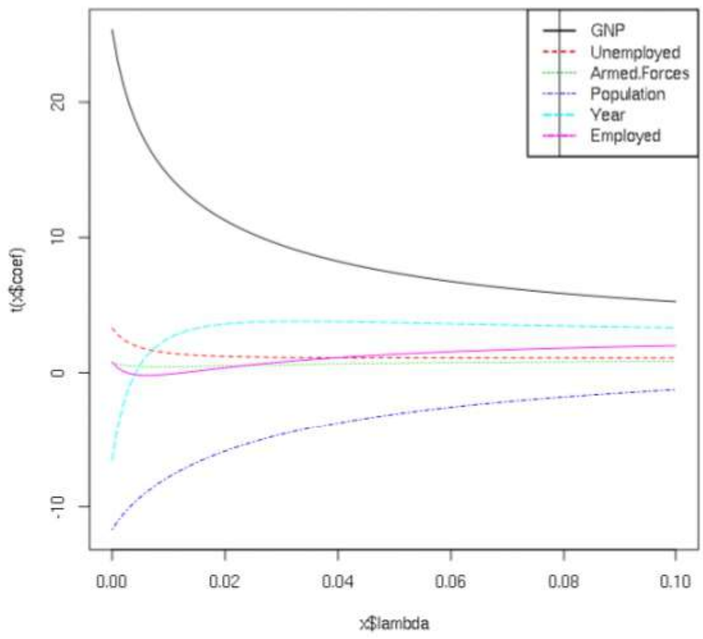
epoch times

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

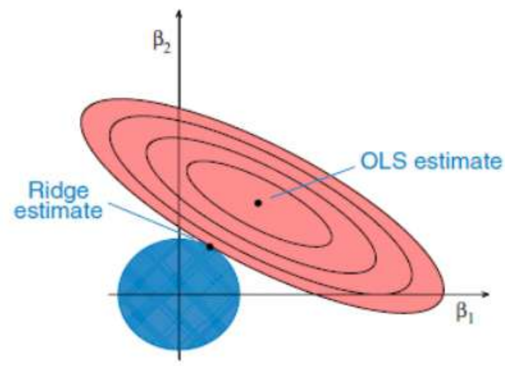
$$\boxed{\frac{dL}{dW} = X^T X W - X^T y + \lambda W}$$

Notes

Friday, June 4, 2021 4:47 PM



Why is it called ridge



5 Key Understandings

Saturday, June 5, 2021 4:20 PM

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

$\rightarrow \lambda(w_1^2 + w_2^2 + \dots + w_n^2)^2$

$\rightarrow$ Overfitting ↓

$\left\{ \begin{array}{l} \text{Shrinkage} \\ \text{coef} \end{array} \right.$

1. How the coefficients get affected?

Saturday, June 5, 2021 4:20 PM

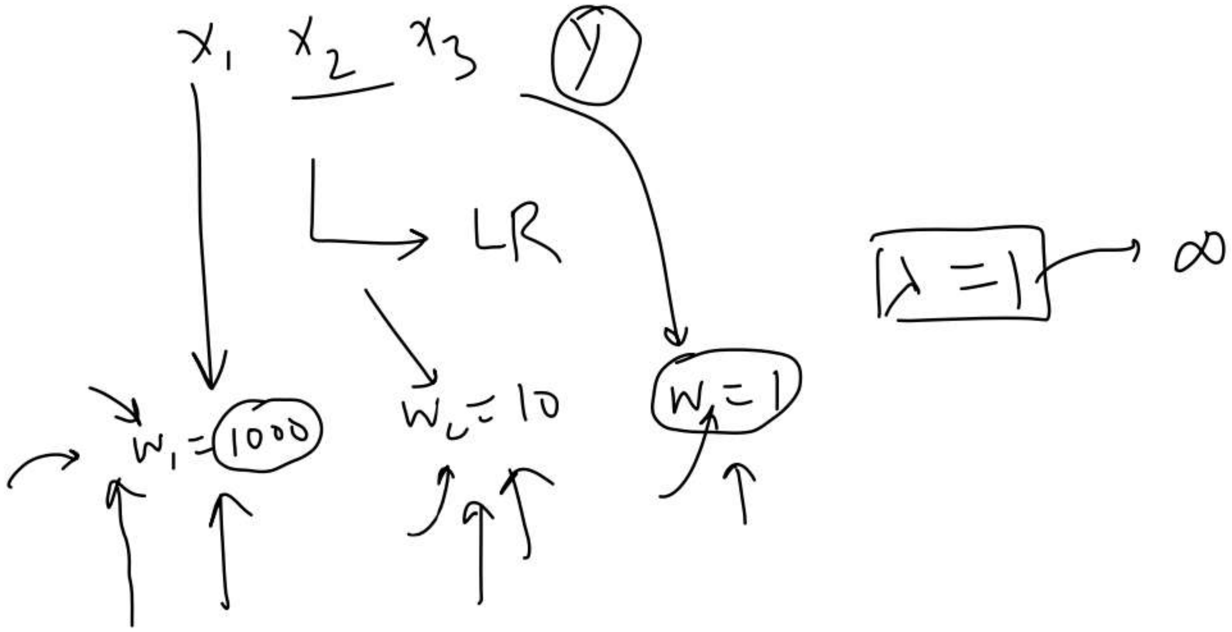
$\lambda \uparrow$ $\qquad \lambda \to 0 \to \infty$

$\boxed{w_1, w_2 \ldots\ldots w_n}$ $\qquad \downarrow 0 \longrightarrow 0$

2. Higher Values are impacted more

Saturday, June 5, 2021 4:21 PM

Never reaches 0



3. Bias Variance Tradeoff

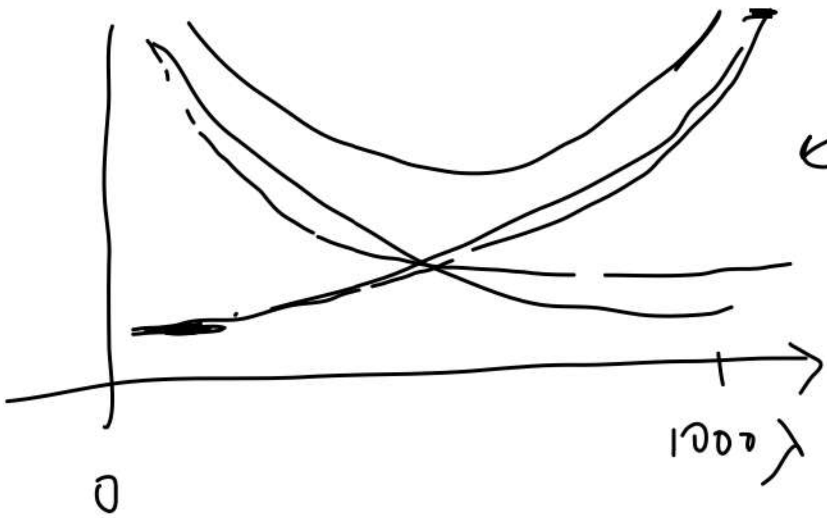
Saturday, June 5, 2021 4:21 PM

Blas Varianu

λ ↓ ↑

Bias ↓ overfit Varianu ↑

Bias ↑ undufitting Valainu ↓



← graph

4. Impact on the Loss Function

Saturday, June 5, 2021 4:21 PM

$\lambda \rightarrow$

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

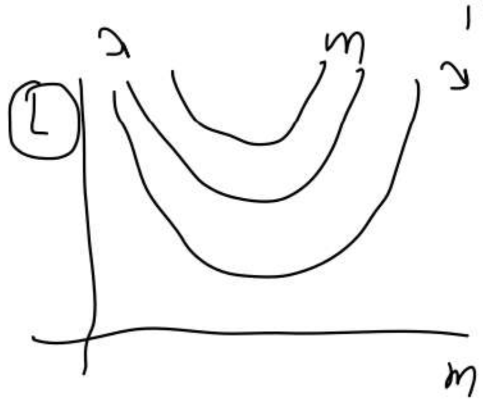
$x, y \rightarrow \boxed{m, b}$ → b is constant

m

$b = 0$

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

$b = -2.29$

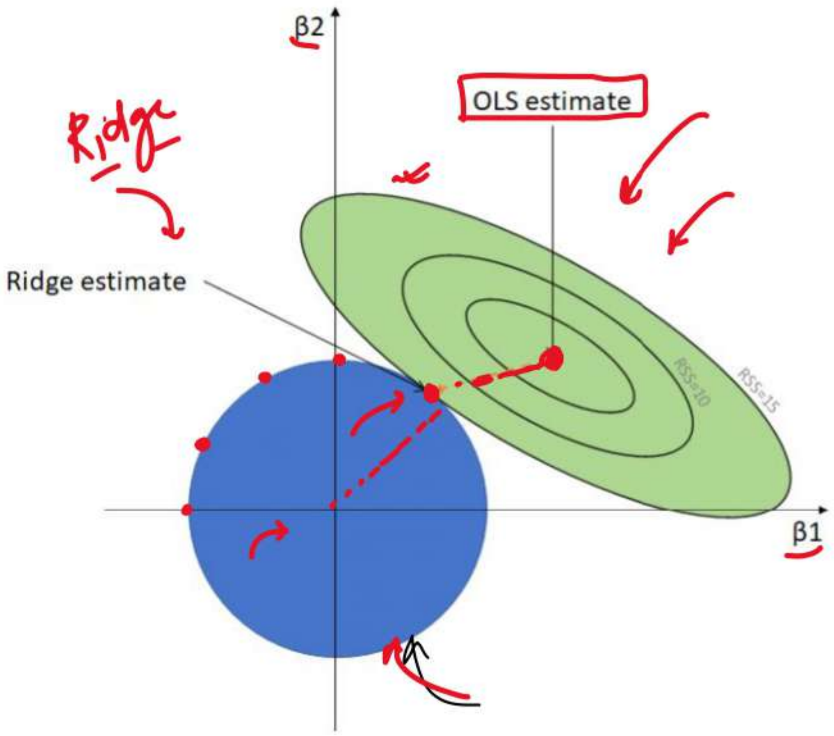


$\lambda = 1$

$\lambda =$

5. Why called Ridge

Saturday, June 5, 2021 4:22 PM



Hard constraint

Ridge constraint

2 coef β_1 β_2 β_0

$L = \underbrace{MSE} + \underbrace{\lambda \|w\|^2}$

Contour

$(y_1 - \hat{y}_1)^2$

$\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}))^2$

MSE

$\boxed{\lambda(\beta_1^2 + \beta_2^2)}$

Practical Tip

Monday, June 7, 2021 1:20 PM

Use ridge when there are more than 2 input cols

Ridge

$x_1 \; x_2 \; x_3 \ldots$

$\boxed{\text{coeff's}}$

$\boxed{x_1 \quad y}$

$>= 2$