Regularization: Add something extra to our machine learning model that tend to reduce overfitting

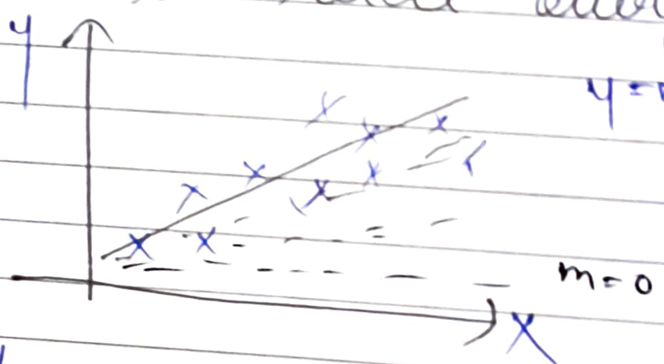
Use: When there is a chance of overfitting.

famous Regularization

Ridge (L2) Lasso (L1) Elastic Net (L1+L2)

GEOMETRIC INTUITATION

How to detect overfitting?

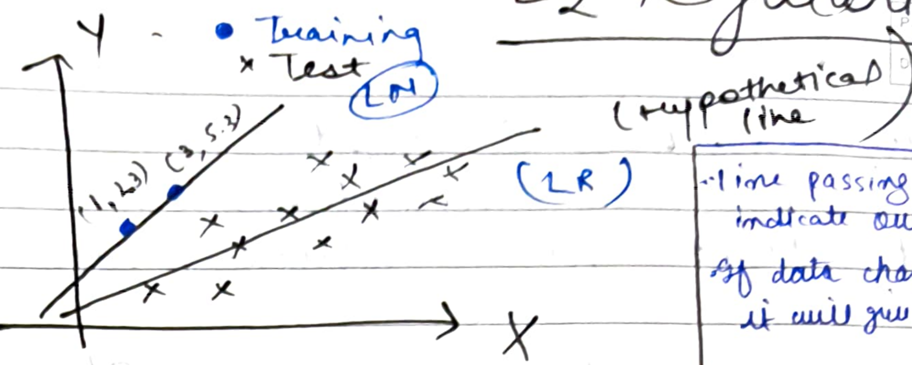


If m value too high means overfitting means only (mx) x is responsible for prediction of y.

If m is too low means underfitting

If you want to reduce overfitting then mathematically you have to decrease the slope (little bit)

L_2 Regularization



- line passing through 2 data indicate overfitting.
- If data changes (test data), it will give bad result

I have to convince my ML model to choose the line (Hypothetical) line

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

To convince / or change our best fit line, we add something to our loss function

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

λ → Hyper parameter; m → slope

Let the eq^n of I line is $\boxed{y = 1.5x + 0.8}$

II line $\boxed{y = 0.9x + 1.5}$

Let's calculate this loss fn (new) for both of the line

LOSS LM	LOSS LR
$\lambda = 1$	$\lambda = 1$
$0 + (1.5)^2$	$(2.3 - 0.9 - 1.5)^2 + (5.3 - 2.7 - 1.5)^2 + (0.9)^2$
$= \boxed{2.25}$	$= (0.1)^2 + (1.1)^2 + (0.9)^2$
	$= \boxed{2.03}$

$2.03 < 2.25$

Hence model will chose this line

M T W T F S S
Page No
Date
YOUVA

If more than 1 axis

$$\text{Loss } f^n + \lambda(m_1^2 + m_2^2 + \cdots m_n^2)$$

named as L2 norm (due to square)

5 KEY UNDERSTANDING
(Ridge)

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

↓ known as shrinkage coefficient.

1) How the coefficients get affected?

$\lambda = 0$, Simple Linear Regression

$\lambda \uparrow$, $m \downarrow$ or coef. ↓ (~~but~~ may get close to 0 but never 0)

2) Higher values are impacted more.

LR → $w_1 = 1000$, $w_2 = 10$, $w_1 = 1$

↓ higher value will be impacted more

M T W T F S

Page No.

Date

3 Bias Variance Tradeoff.

Bias Variance depend on the value of λ.
If model is overfitted, then try increasing λ. but not increase very much that may lead to underfitting.

4) Impact on Loss function.

What is the effect of increasing λ on Loss function?

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

let ~~m~~ be constant $(b = 0)$

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

$\lambda \uparrow$, loss function shrunk and shifted towards origin

5) Why call Ridge?

To understand this, there is a concept called Hard constraints
Ridge constraint.

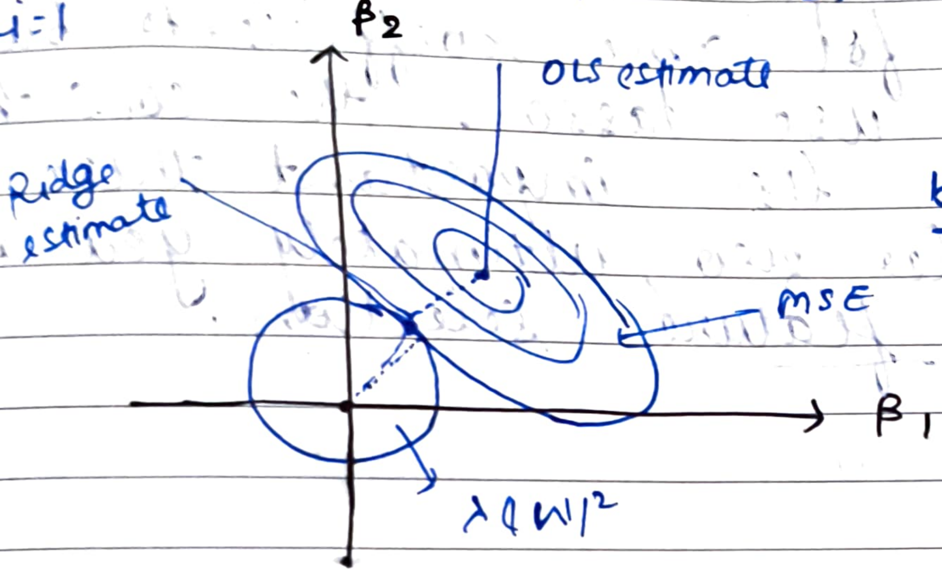
Let there be 2 coef & a intercept.

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

minimize both.

M T W T F S S
Page No
Date
YOUVA

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



both ridge