Lasso Regularization

L1 Penality

As we know that in Ridge regg.

$L = MSE + \text{Ridge term}$

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

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

For Lasso we add L1 Penality term as → $\lambda *$ Norm of (Coeff-cient Vector)

$\lambda \|w\|$

then After applying Lasso hole function becomes.

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

$\lambda \{|w_1| + |w_2| + |w_3| \dots\}$

- ↑ → Underfitting starts on increasing λ
- 0 → In this case it will behave like normal linear regg
- ↓ → starts with overfitting.

⇒ what is benifits of Lasso regularization of it's making some of feature's co-eff as zero?

If we are working on Higher Dimensional data

↳ there would be chance of overfitting → bcoz there would be some (value) of Co-eff. for each feature.

⇓

when we apply Lasso (L_1) then it will make Co-eff ⇒ 0 for the features which are not impor.

⇓

Nothing but this process is called Feature Selection

⇒ which one should we prefer? ridge or Lass ?

data

- High Dimensi → If few dimension is not imp → use Lass
- Low Dimensional → If all dimensions are imp → use Ridge.

Q1. How Co-eff are affected by increasing alpha value?

$\lambda \uparrow \longrightarrow$ Co-eff ↓

↓

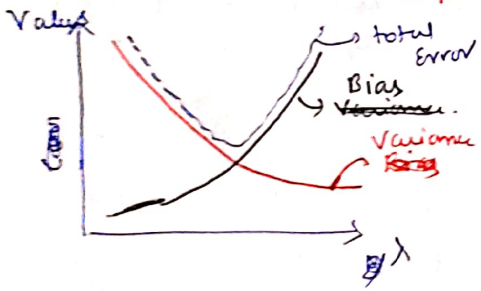
for very high value Co-eff becomes 0

for a certain λ value, it will make co-eff of non-important feature as 0.

⇓

that works as feature selection.

Q How λ in Lasso affect- in Bias & Variance tradeoff



Q Effect of Regularization on Loss Function:-

At certain value of lambda loss becomes minimum but on increasing λ minima of loss will shift towards higher values

Why Lasso creates sparsity?

↳ Meaning of sparsity?

↳ on increasing $\lambda \uparrow$ some of the feature's co-eff becomes 0

If we remove in order to predict y

Eqn: $y = mx + b$

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

where $\bar{y}$ = mean(y)

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

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

After applying ridge

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

Extra term is λ.

Now, we will understand what extra we will add in Lasso?

in Lasso

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

$m = ?$

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

in order to find minimum value of $m \rightarrow$ we have to find $\frac{\partial L}{\partial m}$

And we know that $\lambda|m|$ {mod} function is not differentiable at 0.

Let's take $m > 0$

$$\text{loss} = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \lambda|m| \Rightarrow \sum_{i=1}^{n}(y_i - mx_i - \bar{y} + m\bar{x})^2 + 2\lambda m$$

[∵ 2 is common multiplying with 2]

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

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

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

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

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

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

∴ for Lasso

↓ $m > 0$

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

↓ $m = 0$

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

↓ $m < 0$

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

Let take

$\sum(y_i-\bar{y})(x_i-\bar{x}) = yx$

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

then $m = \frac{yx-\lambda}{x^2}$, $yx = 100$, $x^2 = 50$

But we cannot take $m < 0$ in $m > 0$ case, in this case automatically for $\lambda = 150$,

$m = 2, 9/5, 1, 0, (-1), \ldots$

$\lambda = 0, 10, 50, 100, 150, \times$

it will go to this one return $m = 5$

when $m > 0$

$\lambda = 0, 10, 50, 100, \boxed{150}$ ⇒ $m = 2, 9/5, 1, 0, 5$

as we increase

on $\lambda = 150$ once, it crosses 0 by using formula for $m > 0$ it becomes worst than when we started and it stopped to 0

Similarly ⇓

when:

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

let's say to make $m < 0$ ⇒ we must have $\sum (y_i - \bar{y})(x_i - \bar{x})$ to be negative.

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

$\lambda = 0 \quad 50 \quad 100 \quad \boxed{150}$

$m = -2 \quad -1 \quad 0$

After doing the m become 1 and we are calculating for $m < 0$ so, it will not go beyond 0 so, it will stop.

why ridge regression do not create sparsity?

$$m = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x}) + \boxed{\lambda}}$$ → extra term.

∵ Here, λ is in denominator term so, to being $m = 0$, depends on Numerator value, and Numerator value do not go to zero (0) value. generally

where as

in Lasso, λ is in Numerator term, so to being $m = 0$, λ is in Numerator plays important role.

That's why sparsity is in Lass but not in Ridge.