18/07/23

ML Day-3

Agenda

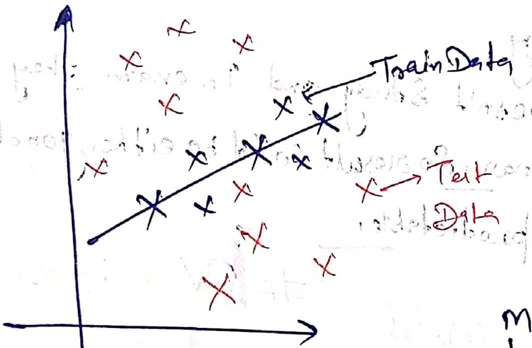
1. Ridge Regression
2. Lasso Regression
3. Elastic net Regression

---

① Ridge Regression (L2 Regularization) :→ Reduce Overfitting

(*) The main AIM of Ridge Regression is to Reduce Overfitting.

How does it Reduce Overfitting ?



Overfitting → Low Bias

Training Data → Accuracy ↑↑ → 95%

Test Data → Accuracy ↓↓ → 60%

↳ High Variance

$$\text{Cost } f^n = \underbrace{\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x)^{(i)} - y^{(i)}\right)^2}_{\text{MSE}} + \lambda \underbrace{\sum (\text{slope})^2}_{}$$

MSE; λ: Hyperparameter

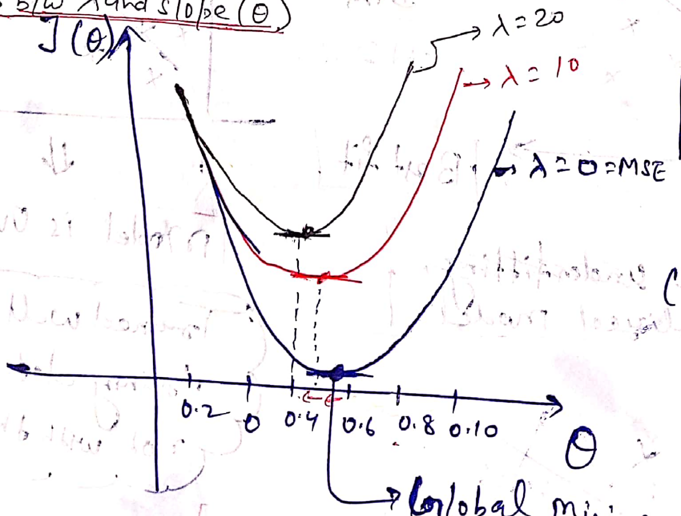
→ $(\theta_1)^2$, $h_\theta(x) = \theta_0 + \theta_1 x$

→ $(\theta_1^2 + \theta_2^2 + \theta_3^2 + \cdots \theta_n^2)$

$h\theta = \theta_0 + \theta_1 x_1 + \theta_2 x_2 \cdots$

(in Case of multi Regression when no. of independent Variables are more than one)

Relationship b/w λ and slope (θ)



(*) As we keep increasing the λ value the curve keep shifting towards left which means θ keep decreasing i.e slope keeps decreasing and Global minima keep shifting

Hence λ is inversely proportional to θ.

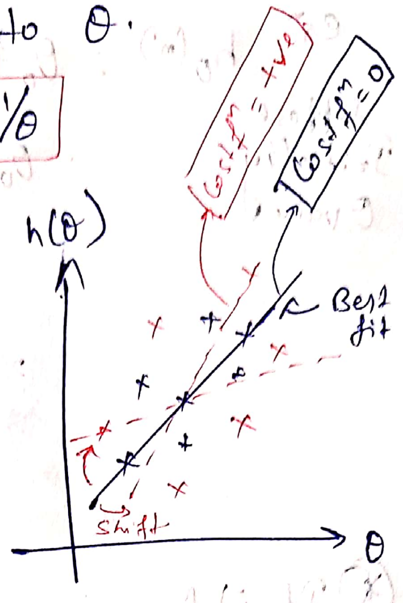
$\boxed{\lambda \uparrow\uparrow \quad \theta \downarrow\downarrow}$ $\boxed{\lambda \propto 1/\theta}$

In case of overfitting

MSE ≈ 0 ← Initially

Cost f^n = $0 + \lambda(\text{slope})^2$

= $\boxed{+ve}$ ↓↓↓ ← AIM



(*) Since cost f^n is +ve hence the best fit line will ~~either~~ move in either direction.

(*) Our main AIM ~~to~~ is to reduce the cost function, Hence the line can either move up or down by changing θ value and creating another best fit line.

(*) W.r.t Linear Regression we know that there is a possibility that we may get an overfitted line but if we use Ridge regression it will never lead to overfitting.

Q. If at all the Model doesn't perform well even after Ridge Regression then what to do?

Sol:

Cost f^n = MSE $+ \lambda\Sigma(\text{slope})^2$

= $0 + \lambda\Sigma(\text{slope})^2$

Change the λ

Can't have -ve value bcz of squaring

(*) Even after performing lasso if at all it doesn't perform well w.r.t Test data then change the λ (hyperparameter) value and try to find the best λ parameter

In case of Ridge Regression $\boxed{\theta \neq 0}$ θ value can never become 0. why?

Sol:

eqn $h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$

Suppose Initially θ value

$= \theta_0 + 0.85x_1 + 0.82x_2 + 1.5x_3$

⇓ After Applying Ridge Regression θ value will decrease,

$= \theta_0 + 0.85x_1 + 0.70x_2 + 0.98x_3$

⇓

0

(If the Coefficient become '0' means the feature going get deleted. Hence it can never become 0.)

(*) If it becomes '0' means feature will get deleted. Hence it will never become 0 but can reach near 0 may be ≈ $\boxed{0.22}$ or sth.

(*) We finally we can see that how "Ridge Regression" is reducing overfitting by adding $[\lambda \sum (slope)^2]$ to MSE and which can never become 0. and if cost fn is +ve means it will adjust the best fit line to gets the best Model

(2) Lasso Regression (L1 regularization)

(*) We use it mainly for reduce the feature in short we can say for feature selection.

How does it works?

$$\text{Cost } f^n = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \lambda \sum_{i=1}^{m} |slope|$$

Eqn

$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$ least correlated feature

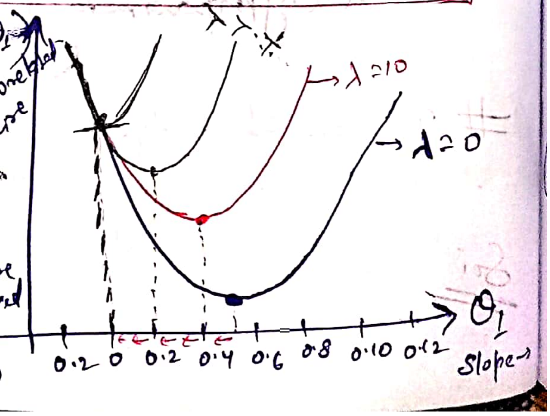
$= \theta_0 + \underline{0.54x_1} + 0.23x_2 + \boxed{0.10x_3}$

⇓ feature removed

0

independent feature = 3

This means if there is one unit movement in Y-axis than there is 0.54 movement in x-axis.



(*) Initially $\lambda = 0$

then $\lambda = 10$

$\vdots$

$\boxed{\lambda \uparrow \; \theta \downarrow}$

At last $\boxed{\theta \to 0}$

When we keep increasing the λ value. ~~Then after a point~~ our local minima will keep shifting and at one point it will get stuck at 0. After that even if you increase the λ value still it will be 0 only. Hence in short we can say that some of the coefficient will become 0.

(*) Hence Automatically the feature which is least correlated is getting removed just by changing the λ value.

(*) Hence here we can say that x_1 and x_2 are the most important feature and x_3 is not at all important and you are simply wasting your time on this.

Q: If the dataset is having outlier which one will it use Ridge or Lasso?

Sol^n: Lasso Regression $\xrightarrow{\text{use}}$ In case of outlier

(Since it is least correlated to O/P and will get deleted automatically)

(*) Ridge Regression $\xrightarrow{\text{use}}$ In case of Overfitting

(3) Elastic Net [L_1 and L_2 Norm]

(*) Basically Elastic Net is combination of both Ridge and Lasso Regression. i.e L_2 and L_1 regularization.

$$\boxed{\text{Cost } f^n = \underbrace{\frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x)^{(i)} - y^{(i)}\right)^2}_{\text{MSE}} + \underbrace{\boxed{\lambda_1 \sum_{i=1}^{m} (\text{slope})^2}}_{\text{Ridge}} + \underbrace{\boxed{\lambda_2 \sum |\text{slope}|}}_{\text{Lasso}}}$$

MSE $\Downarrow$ can be MAE $\Downarrow$ can be RMSE

(*) Hence it is best to use Elastic Net so we need not to worry about either Overfitting or outliers.