Date 23-02-24

Ridge Regression

~~Ridge~~ Regularisation:

it is the technique in which we induce or add a information in the ML model. so that we can we reduce overfitting

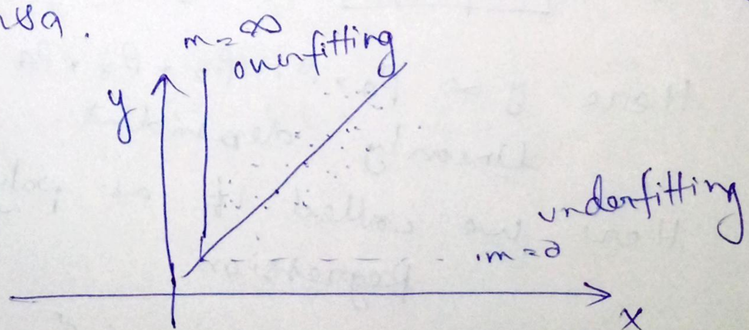
Regularisation Techniques:

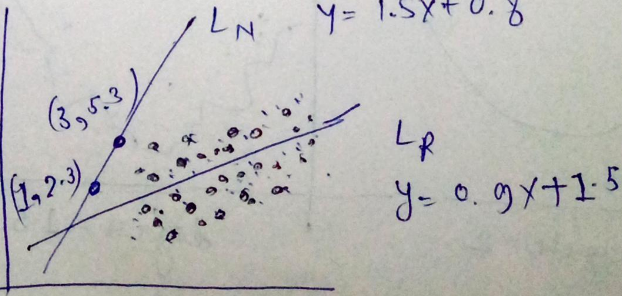
1. Ridge (L2) – L2 Norm
2. LASSO (L1) – L1 Norm
3. ELastic Net (combination of L1 and L2)

Overfitting:

Here the ML model exceptionally well on training data but not so well in testing data. it means the model variance is high.

in the linear regression a overfitted model in which $y = mx + b$, here the m is very high and vice-versa.





• train
• test

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

minimize it

in Regularisation

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

λ = hyperparameter (value from 0 to ∞)

m = slope

loss L_M	loss L_R
$\lambda = 1$ $L = 0 + (1.5)^2$ $L = 2.25$	$\lambda = 1$ $(2.3 - 0.9 - 1.5)^2 + (5.3 - 2.7 - 1.5)^2 + (0.9)^2$ $L = 2.03$

in 2D:

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

in 3D:

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

differentiant w.r.t b and m example

(i) $\frac{\partial L}{\partial b} = 0$ (ii) $\frac{\partial L}{\partial m} = 0$

$\downarrow$ get a b $\downarrow$ get a m

(i) Here $b = \bar{y} - m\bar{x}$

~~[illegible]~~ $\bar{y}$ = y का mean $\bar{x}$ = x ~~का~~ mean,

m = slope

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

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

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

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

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

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

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

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

in Ridge regression

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

in simple linear regression.

Here λ(alpha) is hyperparameter

if $\lambda = 0$ (i) == (ii)

Ridge Regression for nD Data:

let us take an example.

$x_1, x_2 \dots x_n \mid y$

with m rows. and weight are $w_1, w_2 \dots w_n$

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

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

Y = output column. (m values)

$W = [w_0, w_1, \dots w_n]$

X = input column

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

* Now the loss function:

$$L = (XW - Y)^T (XW - Y) \text{ — for simple LR}$$

$$L = (XW - Y)^T (XW - Y) + \lambda \|W\|^2 \text{ — for RR}$$

shrinkage coefficients

$$\therefore \lambda \|W\|^2 \equiv \lambda w_0^2 + \lambda w_1^2 + \lambda w_2^2 + \dots \lambda w_n^2$$

or

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

$$\frac{\partial L}{\partial W} = \quad L = [(XW)^T - (Y)^T] (XW - Y) + \lambda W^T W$$

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

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

$$L = W^T X^T X W - 2 W^T X^T Y + Y^T Y + \lambda W^T W$$

$$\frac{\partial L}{\partial W} = 2X^TXW - 2X^TY + 0 + 2\lambda W = 0$$

$$X^TXW - \lambda W = X^TY$$

$$(X^TX + \lambda I)W = X^TY$$

$$\boxed{W = (X^TX + \lambda I)^{-1} \cdot X^TY}$$

Ridge Regression using Gradient Descent:

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

in vector format

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

or

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

$$X = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1n} \\ 1 & X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \vdots & & \\ 1 & X_{m1} & X_{m2} & \cdots & X_{mn} \end{bmatrix} \quad y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{bmatrix} \quad W = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$$

$w_0, w_1, w_2 \ldots w_n$ (parameters)

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

or

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

here $\frac{\Delta L}{\Delta w} \rightarrow$ gradient

$$\frac{\Delta L}{\Delta w} = \begin{bmatrix} \frac{\partial L}{\partial w_0} \\ \frac{\partial L}{\partial w_1} \\ \vdots \\ \frac{\partial L}{\partial w_m} \end{bmatrix}$$

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

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

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

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

$$\frac{\partial L}{\partial W} = \frac{1}{\cancel{2}}\left[\cancel{2}X^T X W - \cancel{2}\underset{X^T Y}{\cancel{\quad}} + 0\right] + \frac{1}{\cancel{2}} \times \cancel{2}\lambda W$$

$$\frac{\partial L}{\partial W} = X^T X W - X^T Y + \lambda W = \frac{\Delta L}{\Delta W}$$

$$W = [w_0, w_1 \ldots w_n]$$ — starting

in epochs $$w = w - \eta \frac{dL}{dW}$$