Regularization in Deep learning

Regularization is the process by which we solve the overfitting problem.

There are two types of regularization:

1. lasso regularization (L_1)
2. Ridge regularization (L_2)

let's do for regression loss function to understand:

$$\text{Cost function } (C) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

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

In regression, we find the best parameter valve using gradient descent. putting which the valve of cost should be minimum

In, Regularization we simply put penalty term to our loss function.

i.e $C = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \text{penalty term}$

Lasso regularization (L_1):

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

$$C = L + \frac{\lambda}{n} \sum_{i=1}^{K} |w_i|$$

fact: overfitting occurs due to weight valve getting very high number. so, we reduce it's valve using regularization to reduce overfitting.

Ridge regression (L_2):

$$\boxed{\begin{aligned} C &= \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \frac{\lambda}{n}\sum_{i=1}^{k}(w_i)^2 \\ C &= L + \frac{\lambda}{n}\sum_{i=1}^{k} w_i^2 \end{aligned}}$$

Suppose we have 10 weighs in our neural network then, applying L_2 regression will give:

$$C = L + \frac{\lambda}{n}\sum_{i=1}^{10}(w_i)^2$$

$$= L + \frac{\lambda}{n}(w_1^2 + w_2^2 + w_3^2 \ldots\ldots + w_{10}^2)$$

In ridge regularization, the weight decay and goes closer to zero but Never zero.

lamda (λ):

λ is a hyperparameter if we increase it we Solve the problem of overfitting.

> λ value increasing very much gives underfitting
> λ value decreasing verymuch gives overfitting

$\lambda = 0$ vayo vane

$$C = L + \cancel{\frac{\lambda}{n}}^{\,0}\sum_{i=1}^{k}(w_i)^2$$

$$C = L + 0$$

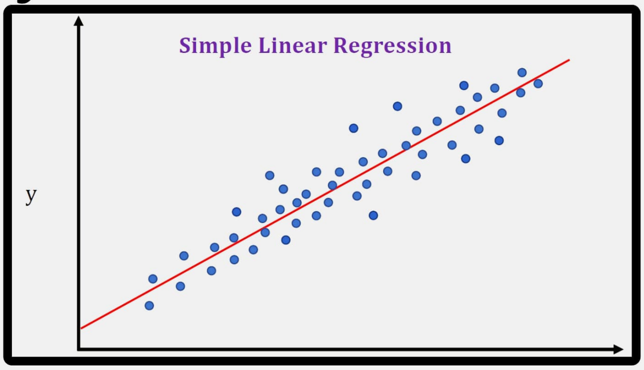
$$C = L \text{ (overfitting nai hunxa)}$$

over fitting increase as we increase the no of features like we can see in both figure below

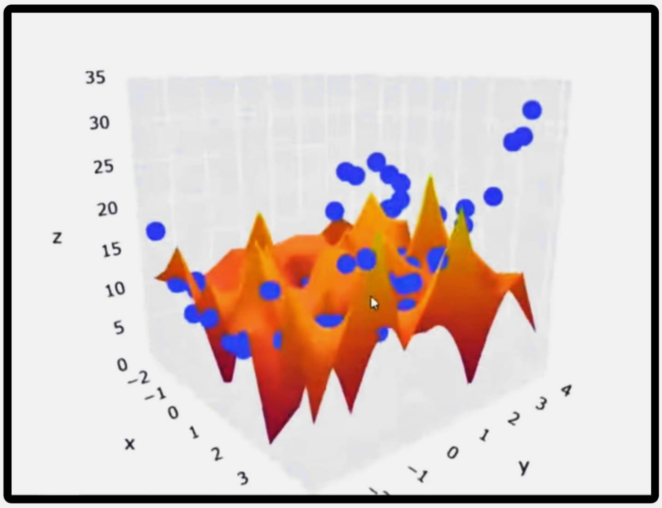
when we increase the features(x1,x2...) the line not only becomes line. It becomes curve of higher dimension

lasso regression make weight value zero which cause feature value also zero. It removes the less imp features and decrese feature number to prevent from overfit

y=wx+b



y=w1x1+w2x2+w3x3+w4x4+...+b



like lasso, ridge make the weight value very very low and the product of (w1 very very low* feature X) value also comes very low like 0.000001 which is near to zero which make that term negligible value. so like thus it also prevent from overfit .

example: **y=0.5*w1+0.4*w2+0.0001*w3+0.0001*w4+b**

y=0.5*w1+0.4*w2+b

dimension decreased so overfitting is also reduced

neglating it since that term value will be close to 0

Intution how does weight decrease in regularization.

let's See in ridge regression for easyness. since the differentation of modulus in lasso is abit lengthy one so, we will do intution of ridge.

As we know, we perform backpropogation and update weight. Suppose we are using Schostatic gradient descent.

$$W_n = w_0 - \alpha \underbrace{\boxed{\frac{d\,loss'}{d w_0}}}_{A}$$

loss' = loss + penalty

$$A = \frac{d\,loss'}{dw_0} = \frac{d\left(Loss + \frac{1}{2}\sum_{i=1}^{k}(w_i)^2\right)}{dw_0}$$

$$= \frac{d\,loss}{dw_0} + \frac{1}{2}\frac{d\,(w_0^2 + w_1^2 + w_2^2 \cdots)}{dw_0}$$

$$= \frac{d\,loss}{dw_0} + \frac{1}{2}\left(\frac{dw_0^2}{dw_0} + \frac{dw_1^2}{dw_0} + \frac{dw_2^2}{dw_0} + \cdots\right)$$

$$= \frac{d\,loss}{dw_0} + \frac{1}{2}\left(\frac{dw_0^2}{dw_0} + 0 + 0 + \cdots + 0\right)$$

$$= \frac{d\,loss}{dw_0} + \cancel{2} w_0 \times \frac{\lambda}{\cancel{2}}$$

$$\therefore\ W_n = w_0 - \alpha\left(\frac{d\,loss}{dw_0} + \lambda w_0\right)$$

$$= w_0 - \alpha\lambda w_0 - \alpha\frac{d\,loss}{dw_0}$$

$$= w_0(1 - \alpha\lambda) - \alpha\frac{d\,loss}{dw_0}$$

$$= \boxed{(1 - \alpha\lambda)w_0 - \alpha\frac{d\,loss}{dw_0}}$$

$$\boxed{w_n = (1 - \alpha\lambda) w_o - \alpha \left(\frac{d\,loss}{d\,w_o} \right)}$$

α alpha and λ (regularization factor) is small +ve number their product also gives small positive number Jaba yo small +ve number 1 bata ghatainxa hamle 1 vanda kaam valve pauxam.

for example: (let)

$\alpha = 0.01$

$\lambda = 50$

$(1 - \alpha\lambda) = (1 - 0.01 \times 50) = 0.5$

Jaba yo 0.5 ley w_o lai multiply garxam hamle w_o ko valve ajhai kaam aauxa paila ko vanda ne.

like:

$w_o = 50$ (let)

$\Rightarrow (1 - \alpha\lambda) w_o = 0.5 \times 50 = 25$

$$w_n = (1 - \alpha\lambda) w_o - \alpha \frac{d\,loss}{dw_o}$$

$$= (1 - 0.01 \times 50) \times 50 - 0.01 \times \frac{d\,loss}{dw_o}$$

$= 25 - 15$ (let's say we obtain this)

$= 10$

like this regularization decrease the weights.