Regularization

What is regularization?

Regularization is a technique to reduce the overfitting of a model.

→ Overfitted models fails to generalize → the model performs well in training dataset, but doesn't perform well in test dataset.

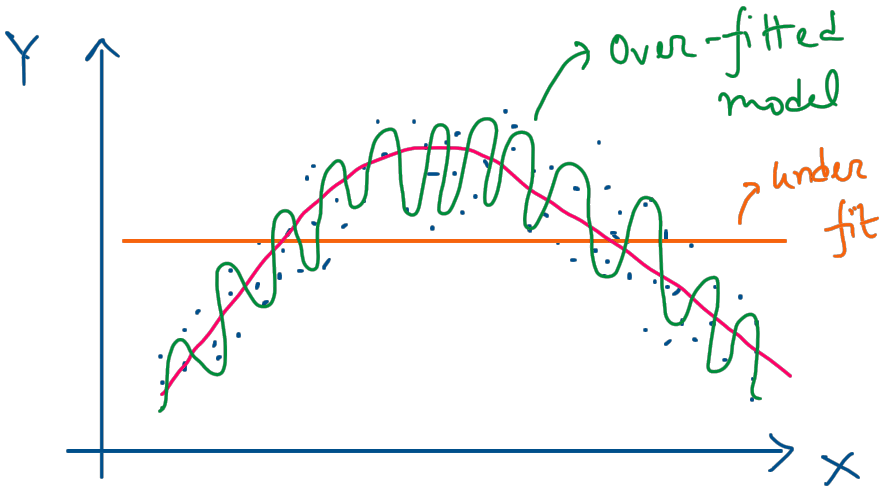
→ Regularization tries to reduce the overfitting by reducing the variance in the dataset.

What is the cause of overfitting?

→ Less number of training datapoints.

→ Too many features

→ Too complex model.



Linear Regression

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k$$

$\frac{\Delta \hat{y}}{\Delta x_j} = \theta_j$ } interpretation of model weights: change in response per unit change in feature.

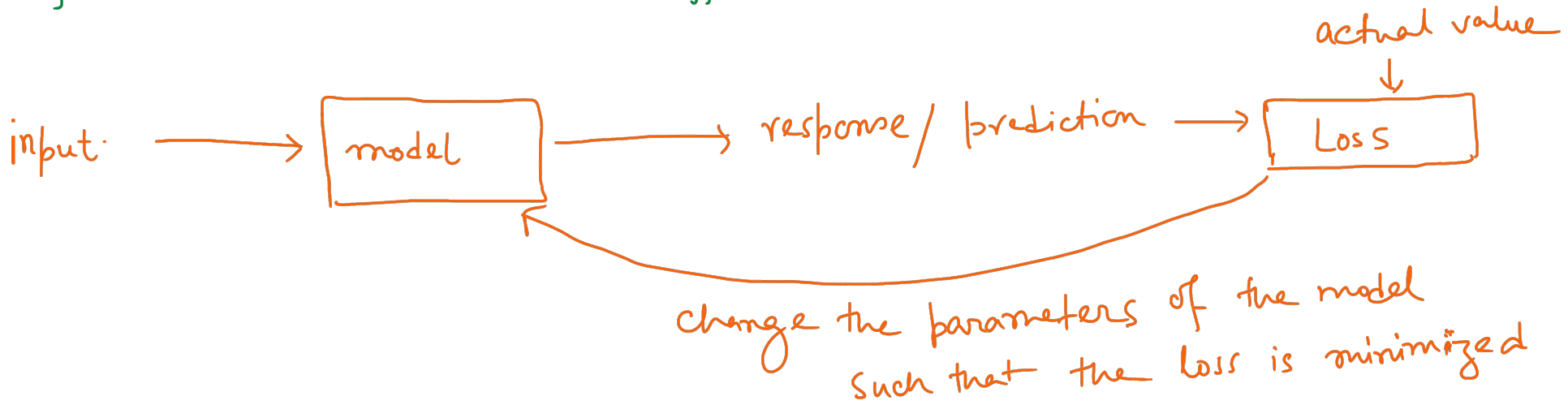
if θ_j is high → model response changes is high when the corresponding feature changes. | model is more susceptible to feature x_j

θ_j is small → change in model response is low when the corresponding feature changes. | model is less susceptible to feature x_j

Logistic regression: $\hat{y} = \text{sigmoid}(z)$ where $z = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k$

θ_j: $\frac{\Delta z}{\Delta x_j}$

Regularization: Is a way to limit the parameters (in case of a parametric model like linear or logistic regression) so that slight change in features (due to noise) doesn't affect the model response much.



We introduce a "penalty" term in the loss-function. When the model parameters are large the penalty will be large & when the model parameter is low the penalty will be low.

"penalty" = f(model weights / parameters)

$$\text{Loss}(\Theta) = \frac{1}{m} \sum_{i=1}^{m} \text{error}\left(\hat{y}^{(i)}, y^{(i)}\right)$$

for linear regression: $$\text{Loss}(\Theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)}\right)^2$$

for logistic regression (binary): $$\text{Loss}(\Theta) = -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} \log\left(\hat{y}^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - \hat{y}^{(i)}\right) \right]$$

for softmax regression: $$\text{Loss}(\Theta) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{k} y_j^{(i)} \log\left(\hat{y}_j^{(i)}\right)$$

Regularized Loss function: $$L(\Theta, \alpha) = \text{Loss}(\Theta) + \boxed{\alpha \cdot (\text{penalty})}$$

α: regularization parameter

Regularization term.

L2 - regularization :

$$L(\Theta, \alpha) = \frac{1}{m}\sum_{i=1}^{m} \text{error}\left(\hat{y}^{(i)}, y^{(i)}\right) + \alpha\left(\frac{1}{2}\sum_{j=1}^{k}\theta_j^2\right)$$

penalty term : $\frac{1}{2}\sum_{j=1}^{k}\theta_j^2 = \frac{1}{2}\left(\theta_1^2 + \theta_2^2 + \theta_3^2 + \cdots + \theta_k^2\right)$

We don't use bias in the penalty. : In the penalty term we use only those parameters which are associated with the inputs of the model.

Regularization parameter (α) : How much penalty is injected to the regularized - loss function. (usually $\alpha \geq 0$)

if $\alpha = 0$: no-penalty (unregularized model)

α is high : high-penalty

L2 Regularized MSE loss:

$$L(\Theta, \alpha) = \frac{1}{2m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)} \right)^2 + \frac{\alpha}{2} \sum_{j=1}^{k} \theta_j^2$$

Ridge-regression (L2-regularized linear regression)

$$\frac{\partial L}{\partial \theta_j} = \begin{cases} \frac{1}{m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)} \right) & ; \text{for bias } \theta_0 \\ \frac{1}{m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)} \right) x_j^{(i)} + \alpha \theta_j & ; \text{for weight } \theta_j \end{cases}$$

$$\theta_j(t+1) \leftarrow \theta_j(t) - \eta \frac{\partial L}{\partial \theta_j} = \theta_j(t) - \eta \left[\frac{1}{m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)} \right) x_j^{(i)} + \alpha \theta_j(t) \right]$$

$$= (1 - \eta\alpha)\, \theta_j^{(t)} - \frac{\eta}{m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)} \right) x_j^{(i)}$$

L1- regularization

$$\vec{v} = [v_1, v_2, v_3, \ldots, v_k]^T$$

$$\text{L2-norm} = \|\vec{v}\|_2 = \sqrt{v_1^2 + v_2^2 + \cdots + v_k^2}$$

$$\text{L1-norm} = \|\vec{v}\|_1 = |v_1| + |v_2| + \cdots + |v_k|$$

$$L(\Theta, \alpha) = \frac{1}{m} \sum_{i=1}^{m} \text{Error}(\hat{y}^{(i)}, y^{(i)}) + \alpha \sum_{j=1}^{k} |\theta_j|$$

L1 regularized MSE Loss : $L(\theta, \alpha) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)})^2 + \alpha \sum_{j=1}^{k} |\theta_j|$

LASSO Regression

↓

Least Absolute Shrinkage and Selection Operator. → feature selection

Combination of L_1 & L_2:

$$L(\Theta, \alpha, \beta) = \frac{1}{m}\sum_{i=1}^{m}\left(\hat{y}^{(i)} - y^{(i)}\right)^2 + \alpha\left[\underbrace{\beta\sum_{j=1}^{k}|\theta_j|}_{L_1} + \underbrace{\frac{1-\beta}{2}\sum_{j=1}^{k}\theta_j^2}_{L_2}\right]$$

$\alpha \rightarrow$ regularization parameter ($\alpha = 0$ means no regularization)

$\beta \rightarrow L_1$ ratio $\rightarrow$ it controls the presence of L_1 & L_2 regularization in the penalty

$0 \leq \beta \leq 1 \rightarrow \beta = 0$: no L_1, only L_2

$\beta = 1$: only L_1, no L_2

Regularized MSE cost:

$$L(\Theta, \alpha, \beta) = \frac{1}{2m}\sum_{i=1}^{m}\left(\hat{y}^{(i)} - y^{(i)}\right)^2 + \alpha\left[\beta\sum_{j=1}^{k}|\theta_j| + \frac{1-\beta}{2}\sum_{j=1}^{k}\theta_j^2\right]$$

} Elastic-Net Regression