Linear Regression - 2

How to determine which features impact the model most during prediction?

Ans: The feature with the highest weight → most important feature

What does the -ve/+ve sign mean in the weights of the model?

- The -ve sign means → if feature value ↑ , $\hat{y}$ ↓
- The +ve sign means → if feature value ↑ , $\hat{y}$ ↑
- 0 weight means → no change in $\hat{y}$ as feature value changes

How to find optimal weights for Lin. Reg. ?

Ans: Gradient Descent → minimizes the Mean Squared error to reach global minima

How does the ML model update weights?

Ans: By finding the gradients of the weights w.r.t the loss function and by subtracting that gradient from the weights.

How to find the Gradients of Mean Square Error?

Ans: We define loss function as :

$$L(w, w_0) = \frac{1}{n}\sum_{i=1}^{n} (y_i - (w^T x_i + w_0))^2$$

On finding gradients for w, loss becomes:

$$\frac{\partial L(w,w_0)}{\partial w} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial (y_i - (w^T x_i + w_0))^2}{\partial w} = \frac{2}{n}\sum_{i=1}^{n} (y_i - (w^T x_i + w_0)) \frac{\partial (y_i - (w^T x_i + w_0))}{\partial w}$$

As we know,

$$\frac{d(uv+c+a)}{du} = v \text{ ,}$$

hence on simplifying, the equation becomes:

$$\frac{\partial L(w,w_0)}{\partial w} = \frac{2}{n}\sum_{i=1}^{n}(y_i - (w^T x_i + w_0))\,(-x_i)$$

Similarly gradient for $w_0$becomes:

$$\frac{\partial L(w,w_0)}{\partial w_0} = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i-(w^T x_i+w_0))^2}{\partial w_0} = \frac{2}{n}\sum_{i=1}^{n}(y_i - (w^T x_i + w_0))\,\frac{\partial(y_i-(w^T x_i+w_0))}{\partial w_0}$$

$$\frac{\partial L(w,w_0)}{\partial w_0} = \frac{2}{n}\sum_{i=1}^{n}(y_i - (w^T x_i + w_0))(-1)$$

Updating weights (w, w_0) with a learning rate α :

$$w = w - \alpha \times \frac{\partial l(w,w_0)}{\partial w}$$

$$w_0 = w_0 - \alpha \times \frac{\partial l(w,w_0)}{\partial w_0}$$

Why use a Learning Rate?

Ans: Learning Rate α → hyperparameter to control the rate at which Gradient Descent reaches global minima

What happens if a too-small value of Learning Rate(α) is used?

Ans: makes Gradient Descent reach the global minima **very slowly**

What happens if a too-large value of Learning Rate(α) is used?

Ans: may make the Gradient Descent **overshoot** the global minima