L1 vs L2

Kuba Jerzmanowski

Sunday 7th July, 2024

1 Introduction

This is a quick note on why L1 drives many weights to zero and L2 makes weights uniformly small. I assume understanding of L1 and L2 regularization. I found it hard to find a good explanation for the effects of the two different penalties to the loss function, henc this.

2 L1 Regularization

In L1, our loss function for one example looks like this (assuming Mean Squared Error):

$$J(w) = (y - xw)^2 + \lambda \cdot |w| \tag{1}$$

2.1 Derivatives

- Derivative of MSE part: $\frac{\partial}{\partial w}(y - xw)^2 = -2x(y - xw)$
- Derivative of L1 part: $\lambda \cdot \text{sign}(w)$

Where $\text{sign}(w)$ is:

$$\text{sign}(w) = \begin{cases} 1 & \text{if } w > 0 \\ -1 & \text{if } w < 0 \\ 0 & \text{if } w = 0 \end{cases}$$

2.2 Full Gradient Example

Let's assume the following values for simplicity:

- $x = 1$
- $y = 2$
- $\lambda = 0.5$
- $\alpha = 0.1$
- Initial weight $w = -0.1$

2.2.1 Calculate MSE Gradient

$$\frac{\partial}{\partial w}(y-xw)^2 = -2x(y-xw) = -2{\cdot}1{\cdot}(2-1{\cdot}-0.1) = -2{\cdot}(2+0.1) = -2{\cdot}2.1 = -4.2 \tag{2}$$

2.2.2 Calculate L1 Gradient

$$\frac{\partial}{\partial w}\lambda|w| = 0.5 \cdot \text{sign}(-0.1) = 0.5 \cdot (-1) = -0.5 \tag{3}$$

2.2.3 Full Gradient

$$\frac{\partial J}{\partial w} = -4.2 - 0.5 = -4.7 \tag{4}$$

2.2.4 Weight Update

$$w \leftarrow w - \alpha\frac{\partial J}{\partial w} = -0.1 - 0.1 \cdot (-4.7) = -0.1 + 0.47 = 0.37 \tag{5}$$

Let's do this a few more times:

- MSE gradient = -2*1(2-1*0.37) = 3.4
- L1 gradient = 0.5
- Full gradient = 3.4 + 0.5 = 3.9
- $w \leftarrow w - \alpha\frac{\partial J}{\partial w} = 0.37 - 0.1(3.9) = -0.02$

The important thing to realize is that L1 regularization continuously pushes the weight towards zero. Why? Well, the L1 gradient ($\lambda \cdot \text{sign}(w)$) remains significant (either λ or $-\lambda$), while the MSE gradient becomes smaller, so you will have a non-zero loss for weight as long as it's not zero.

Note: The λ strikes the balance between the fitting of the data (MSE) and the parameter regularization. As λ goes up, more params go to zero.

3 Comparison with L2 Regularization

If you think about how L2 works, you now see why there is a difference.

L2 loss function:

$$\text{Loss} = \text{MSE loss} + \lambda \cdot w^2 \tag{6}$$

The derivative of w^2 is $2w$, so the penalty will be bigger if w is big and smaller if w is small. This is as opposed to L1 where the penalty is constant as long as w is not zero.

Last Modified: Sunday 7$^{\text{th}}$ July, 2024