مهرداد برادران (99222020)

Neural network (AI First project)

Classification Fashion M-nist

Q1 - In machine learning and regression data analysis in modeling, we use coefficients to limit the estimates to make the data smaller or close to zero. One of the most used methods for this is regularization. This method is used to reduce errors and prevent Learning is used a lot.

In simpler terms, it can be said that by regularizing and normalizing regression coefficients with high values, it prevents overfitting.

In fact, it makes the model smaller by simplifying and reducing the variables. And for this reason, smaller and simpler models are less prone to overfitting. And they generalize better.

There are many regularization techniques including L1 and L2

Both of these models add a penalty to the loss function.

Because in neural networks with lower weights, a simpler model can be designed. In regularizing the matrix, it reduces the weights.

In this regularization, an L1 penalty is added to the loss function to change the weights. In fact, this L1 is the absolute value of the weight parameters.

$$\text{Cost} = \sum_{i=0}^{N}(y_i - \sum_{j=0}^{M} x_{ij}W_j)^2 + \lambda \sum_{j=0}^{M} |W_j|$$

Here, if the lambda is zero, it can make the coefficients with many large values zero. and make it under-fitting.

And if it is too much, it can be necessary to increase the weights to a very large amount, causing under-fitting. As a result, choosing the right lambda has a significant effect in preventing overfitting and proper fitting.

L1 regularization is preferred when we have many features and want few solutions.L1 regularization technique is called Lasso Regression too.

In L2 regularization, we add the square of the coefficients as a penalty to the loss function.

$$\text{Cost} = \underbrace{\sum_{i=0}^{N}\left(y_i - \sum_{j=0}^{M} x_{ij}W_j\right)^2}_{\text{Loss function}} + \underbrace{\lambda \sum_{j=0}^{M} W_j^2}_{\text{Regularization Term}}$$

As can be seen, like the example of L1 regularization, if the lambda coefficient is zero or very large, it cannot perform proper regularization and fitting.

So these two regularizations are both used to prevent overfitting, but the choice of lambda value and variables has a significant impact.

L2 has a great use in estimating the importance of predictions, as a result, it can limit or eliminate distant and more insignificant predictions by penalizing and limiting their coefficients.

L2 regularization techniques is called Ridge Regression.

In choosing between these two regularizations, one must check many conditions and choose one according to the type of problem, but in some ways L1 is stronger because it only takes the absolute value. But in L2, extra costs increase. And it increases exponentially.

but L1 Unable to learn complex data patterns. And it is Computationally inefficient over non-sparse conditions. while L2 Able to learn complex data patterns and it is Computationally efficient because of having analytical solutions.