Exercise 1

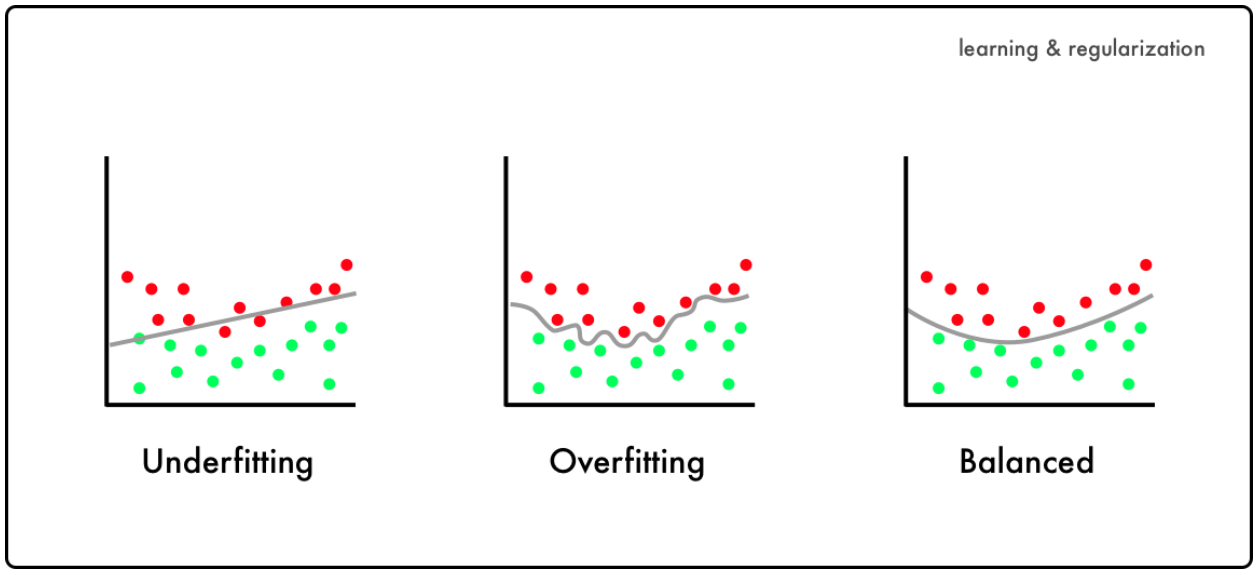
Mahdieh Sajedi Pour, Nima Taheri

November 7, 2022

Describe in detail how L1 regularization differs from L2 regularization, and which one do you prefer?
· In addition, please describe intuitively how each affects the model weights.

Why Regularization?
Regularization has some methods that are used to prevent overfitting.

What is Overfitting?
The graph below is an overfitting example, the model is trying to cover and fit the whole data exactly. This model would work perfectly on training data, but it would fail to predict unseen data. So, we need something to stop this.



1) L1 Regularization (Lasso):

L1 method reduces the weight parameter on which is applied to near zero. Now, if the features of the input data are close to zero, then we can say that applying the L1 norm creates a "sparse L1 norm" model. In this model, just some of weights are non-zero. So, not all of features of the data are involved and used in the learning process of algorithm. So, it is deduced that L1 regularization has the ability of feature selection. And it does

this by assigning zero weights to unimportant features and weights.
In L1 Regularization, the penalty value that we add to the cost function is the sum of the absolute value of all weights, which is multiplied by a constant so that the impact of this norm can be controlled. In the picture below, the term that is added to the cost function is the L1 penalty.

$$J(w)_{L1} = \sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)} \right)^2 + \alpha \|w\|_1$$

where

$$\|w\|_1 = \sum_{j=1}^{m} |w_j|$$

Choosing alpha

The higher the alpha, the less complex the model is and the error due to the not overfitted model is reduced. On the other hand, alphas that are too high increase the error due to bias. Therefore, it is important to choose an optimal alpha.

2) L2 Regularization (Ridge):

For L2 method, the penalty term in the cost function is obtained by the sum of the squares of all the weights.

L2 forces the weights of the learning model to remain small, but it does not set them to zero. So, using L2 does not lead to simple model.

L2 does not perform well in the presence of outlier data in the used dataset. Thats because at the outlier points, the prediction error of the model becomes very large, and with the L2 penalty, the model weights will be smaller. The model that uses L2 will have a better answer when all the features of the input data have an effect on its prediction

target and also the weights inside the model are initialized approximately equally.

$$J(w)_{L2} = \sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)} \right)^2 + \alpha \|w\|_2^2$$

where

$$\|w\|_2^2 = \sum_{j=1}^{m} w_j^2$$

Conclusion:

L2 regularization shrinks all weights by the same proportion but does not remove any, it only reduces them to values near 0. L1 regularization, on the other hand, can reduce some weights to 0 and does variable selection. It better to use L2 when we have outliers. It is because L2 takes the square of the weights, so the cost of outliers increases exponentially. L1 takes the absolute values of the weights, and the cost increases linearly.