Artificial Neural Network

First assignment solutions

Mobin Nesari
Dr. Saeed Reza Kheradpisheh
November 4, 2022

Question 1:

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.

Answer:

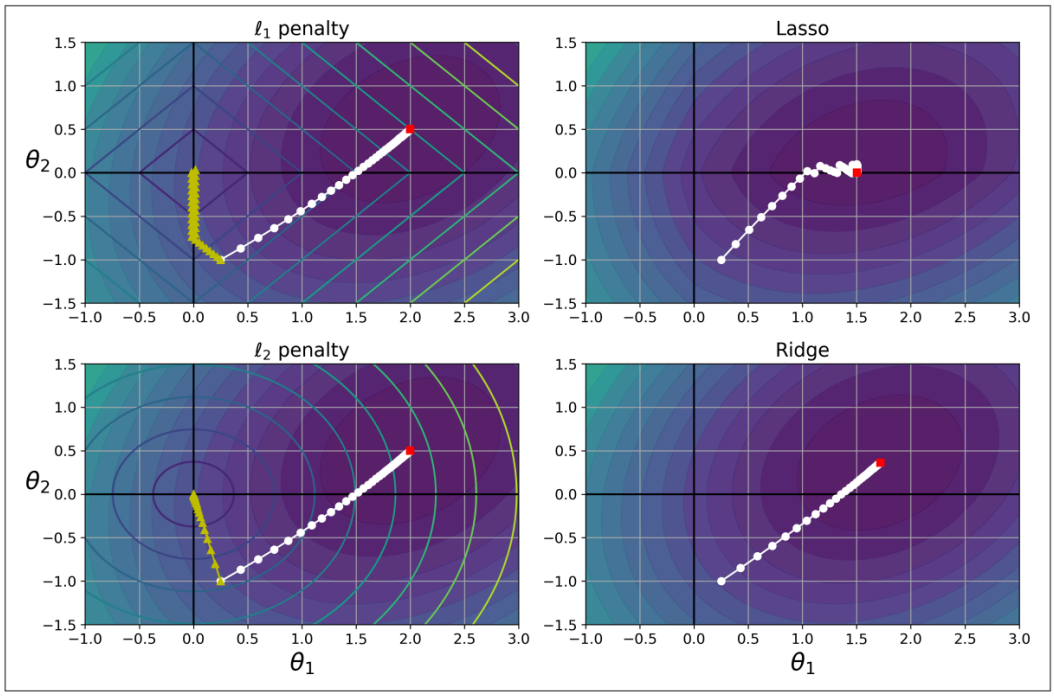
First, we shall define each regularization and then see their differences in action.

L2 regularization or Ridge Regression is a regularized version of linear regression. The regularization term which is equal to $\alpha \sum_{i=1}^{n} w_i^2$ is added to cost function. It simply uses l_2 norm. This term force model to not only fit on data but also keep its weights minimum. Keeping weights as minimum as possible will make model as small as possible. The hyper parameter α indicates how much you want to regularize your model. For instance, if $\alpha = 0$ then Ridge Regression will be equal to Linear Regression. If α is very large, then all weights end up very close to zero and the result is a flat line going through the data's mean. Ridge Regression cost function equals $J(w) = MSE(w) + \alpha \frac{1}{2} \sum_{i=1}^{n} w_i^2$. An important fact that worth to mention is we only regularize weights between 1 and n and we don't manipulate bias term which is w_0. Ridge Regression also has a closed-form solution which equals $\hat{w} = (X^T X + \alpha A)^{-}1 \quad X^T \quad y$ which A is $(n+1) \times (n+1)$ identity matrix which one addition zero on top left.

Lasso Regression(Least Absolute Shrinkage and Selection Operator Regression) is another regularization method which exactly like Ridge Regression, it append an extra term to cost function. This method uses l_1 norm instead of l_2 norm in regularization term. Lasso Regression cost function equals $J(w) = MSE(w) + \alpha \sum_{i=1}^{n} |w_i|$. An important attribute of Lasso Regression which distinguish Lasso Regression from Ridge and other methods is it tends to eliminate the weights of the least important features. This means that Lasso Regression automatically performs features selection and outputs a sparse model. The only disadvantage of l_1 regularization method is that it is not derivative in $w_i = 0$. As you know, cost function should be derivative because we use it in back propagation and update weights. We can solve this problem by using subgradiant vector equation:

$$g(w, J) = \nabla_w MSE(w) + \alpha \begin{pmatrix} \text{sign}(w_1) \\ \text{sign}(w_2) \\ \vdots \\ \text{sign}(w_n) \end{pmatrix} \text{ where } \text{sign}(w_i) = \begin{cases} -1 & \text{if } w_i < 0 \\ 0 & \text{if } w_i = 0 \\ 1 & \text{if } w_i > 1 \end{cases}.$$

Now let's compare these two methods and see which method shall we use in our works.



l_1 vs l_2 regularization (From Hands on Machine Learning book)

On top-left figure, background contours(ellipses) show an unregularized MSE cost function ($\alpha = 0$) and white circles show batch gradient descent path in cost function. The foreground contours(diamonds) represent l_1 penalty and the yellow triangles show BGD path when ($\alpha \to \infty$). The top-right plot, the contour represent same cost function plus l_1 penalty with $\alpha = 0.5$. It is obvious that global optimum is nearer to $\theta = 0$ than unregularized cost function, but the weights do not get fully eliminated.

In conclusion, there are two main differences between Ridge and Lasso. First, the gradients get smaller as the parameters approach the global optimum, so Gradient Descent naturally slows down, which helps convergence (as there is no bouncing around). Second, the optimal parameters get closer and closer to the origin when you increase α, but they never get eliminated entirely.