How regularization helps in feature selection. Suppose you have the following data set:

|  |  |  |  |
| --- | --- | --- | --- |
| Feature1 | Feature2 | Feature3 | Result |
| 1 | 3 | 5 | 10 |
| 1 | 3 | 5 | 10 |
| 1 | 3 | 6 | 12 |
| 1 | 3 | 6 | 12 |
| 1 | 3 | 7 | 14 |

A correct but very odd hypothesis would be to say that the result can be calculated as follows:

3000 \* Feature1 – 1000 \* Feature2 + 2 \* Feature3

This model results in absolutely no errors at all. Something about it “feels” wrong though. Why would you assign weights to features 1 and 2? When you can just double feature 3 and be done? Especially, why are the weights for Features 1 and 3 so high? If the objective of a machine learning algorithm is to minimize the sum of squared error, there is absolutely no reason it would not fall on this hypothesis.

However, let’s really think about how we can express mathematically what is wrong with this model. Basically the weights associated with the features are too high. So why not add the weights themselves to the objective function? Right now the l1 norm of the feature weights is abs(3000) + abs(-1000) + abs(2) = 4002. That should be lower right? So let’s add that to the minimization. Now instead of minimizing:

(F11 \* W1 + F12 \*W2 + F13 \* W3 – R1) ^2 + (F21 \* W1 + F22 \* W2 + F23 \* W3 – R2) ^2 …

Minimize

(F11 \* W1 + F12 \*W2 + F13 \* W3 – R1) ^2 + (F21 \* W1 + F22 \* W2 + F23 \* W3 – R2) ^2 …

+ lambda \* (abs(W1) + abs(W2) + abs(W3))

This puts a penalty on the weights themselves! The above model would now have a loss of 4.002 if you set lambda to 0.01. So the hypothesis, 2 \* Feature3 becomes much more attractive because 1) it has no sum of squared error loss and 2) it’s sum of absolute values of weights is just 2 instead of 4002. So any reasonable machine learning training method would now fall on the simpler hypothesis of 2 \* F3.

In feature selection this is critical because the first model assigned weights of 3000 and -1000 to features 1 and 2. This makes them look HIGHLY important to the regression task. However, by visual inspection they clearly are not. By adding regularization we discourage the complex hypothesis and land on the simpler and more correct hypothesis of 2 \* F3. This also appropriately identifies the only feature that is actually important in the regression task, F3.