What is the difference between L1 and L2 regularization? How does it solve the problem of overfitting? Which regularizer to use and when?

This question previously had details. They are now in a comment.

Justin Solomon, works at Doctor of Philosophy Degrees
Answered Nov 24, 2012 · Upvoted by Carl Shan, data scientist, 2014 Eric Schmidt Data Science Fellow and Naran Bayanbat, MSCS with focus in machine learning

There are many ways to understand the need for and approaches to regularization. I won't attempt to summarize the ideas here, but you should explore statistics or machine learning literature to get a high-level view. In particular, you can view regularization as a prior on the distribution from which your data is drawn (most famously Gaussian for least-squares), as a way to punish high values in regression coefficients, and so on. I prefer a more naive but somewhat more understandable (for me!) viewpoint.

Let's say you wish to solve the linear problem $Ax = b$Ax=b. Here, AA is a matrix and bb is a vector. We spend lots of time in linear algebra worrying about the *exactly-* and *over-determined* cases, in which AA is at least as tall as it is wide, but instead let's assume the system is *under-determined*, e.g. AA is wider than it is tall, in which case there generally exist infinitely many solutions to the problem at hand.

This case is troublesome, because there are **multiple** possible xx's you might want to recover. To choose one, we can solve the following optimization problem:

> MINIMIZE $\|x\|$‖x‖ WITH RESPECT TO $Ax = b$Ax=b

This is called the **least-norm solution**. In many ways, it says "In the absence of any other information, I might as well make xx small."

But there's one thing I've neglected in the notation above: The norm $\|x\|$‖x‖. It turns out, this makes all the difference!

In particular, consider the vectors $a = (0.5, 0.5)$a=(0.5,0.5) and $b = (-1, 0)$b=(−1,0). We can compute two possible norms:

- $\|a\|_1 = |0.5| + |0.5| = 1$‖a‖1=|0.5|+|0.5|=1 and $\|b\|_1 = |-1| + |0| = 1$‖b‖1=|−1|+|0|=1
- $\|a\|_2 = \sqrt{0.5^2 + 0.5^2} = 1/\sqrt{2} < 1$‖a‖2=0.52+0.52=1/2<1 and $\|b\|_2 = \sqrt{(-1)^2 + (0)^2} = 1$‖b‖2=(−1)2+(0)2=1

So, the two vectors are equivalent with respect to the L1 norm but different with respect to the L2 norm. This is because **squaring a number punishes large values more than it punishes small values**.

Thus, solving the minimization problem above with $\|x\|_2$‖x‖2 (so-called "Tikhonov regularization") *really* wants small values in all slots of xx, whereas solving the L1 version doesn't care if it

puts all the large values into a single slot of x.

Practically speaking, we can see L2 regularization spreads error throughout the vector x, whereas L1 is happy (in many cases) with a *sparse* x, meaning that some values in x are **exactly** zero while others may be relatively large. The former case is sufficient and indeed suitable for a variety of statistical problems, but the latter is gaining traction through the field of compressive sensing. From a non-rigorous standpoint, compressive sensing assumes not that observations come from Gaussian-distributed sources about ground truth but rather that sparse or simple solutions to equations are preferable or more likely (the "Occam's Razor" approach).

66.4k Views · 367 Upvotes