Gradient Descent

Moe Antar

January 4, 2017

General Idea

Gradient Descent is an optimization method that allows us to minimize the parameters of a function. It works by taking the function (and in essence the data that it's drawing from) and treating it as a convex plane (think of a valley that we want to climb to the bottom of but can't actually see).

We start at some point on this plane (sometimes 0 , sometimes a random point) and then measure our slope at that point (which gives us an idea of what direction the steepest descent to our goal is (in our analogy of climbing down the valley, this is like feeling around us with our foot and seeing which step feels steepest - and thus more likely to take us downward to our destination).

The size of each step we take is our alpha value. The "feeling of steepness and direction we take" is found by taking the partial derivative of the function we are trying to minimize with respect to the parameter we are trying to minimize. In linear regression this would be our beta*x plus intercept. In logistic regression this would be our 1 over (one plus "e" raised to the linear equation).

Learning Rate

As mentioned earlier the learning rate is basically our step-size (i.e how far we move down the plane after each calculation). This is important because if our steps are too big, we may over-shoot the optimal point when we reach it. Likewise if our steps are too small, it would take us a very long time to finally reach our goal.

Normalization

It is said in data-science etiquette that we should normalize data before applying gradient descent. This is mainly because as we discussed in the climber analogy above, gradient-descent relies on the plane that we're working with having some kind of curvature (like the valley in our example). This is important because if we didn't have a curved plane with as few "dips" as possible, there would be more than one dip and we might travel to a different "bottom" of this proverbial valley depending on our starting point.

Convergence

We say that the function has converged when taking additional "steps" no longer provides us with a significant decrease in the error of our output. Think of this as we're pretty sure that we're at the bottom of this valley and each time we feel with our feet to find a steeper point under us we either don't find one at all or find one too insignificant to be worth the effort. At this point we say that we've achieved our goal.