

Gradient Descent using Loss isn't our only way to calculate the predictor line(or hyperplane). We can use this as a general purpose way to predict some continuous output.

This isn't however, the only way to predict continuous output, there are a variety of ways to do this.

Normal Equation

The normal equation is another commonly used way of calculating the coefficients that minimize loss. The normal equation tries to solve the same problem as MSE LOSS, it aims to minimize the Sum of Residual Squares. This is a closed form solution however, and let's try to understand what it is doing.

$$\theta = (X^T X)^{-1} X^T y$$

What we're doing here in a very simple way is computing the values for the theta vector that will cause the partial derivatives to equal zero. Why would we want all of the partial derivatives to equal zero? Well that's because if we think about our loss function, when all of the partial derivatives are equal to zero, that means that we found the global minimum. Let's look at it from a 1D perspective.

If we have this closed form solution for finding the least squares, then why do we bother with Stochastic Gradient Descent?

The answer is because of time complexity.

The pros and cons list are as follow.

Gradient Descent

- Needs to find an optimal alpha.
- Needs many iterations.
- Works well even with a large dataset

Normal Equation

- No need to choose alpha
- No iteration
- Computationally Expensive