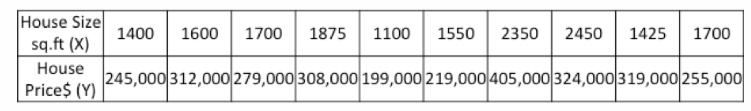
Let me explain to you using an example.

Most of the data science algorithms are optimization problems and one of the most used algorithms to do the same is the Gradient Descent Algorithm.

Now, for a starter, the name itself Gradient Descent Algorithm may sound intimidating, well, hopefully after going though this post,that might change.

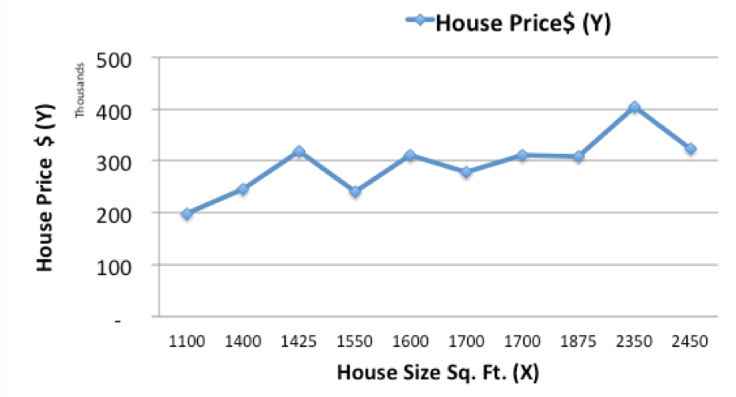
Lets take the example of predicting the price of a new price from housing data:

Now, given **historical housing data,**the task is to create a model that predicts the price of a new house given the house size.

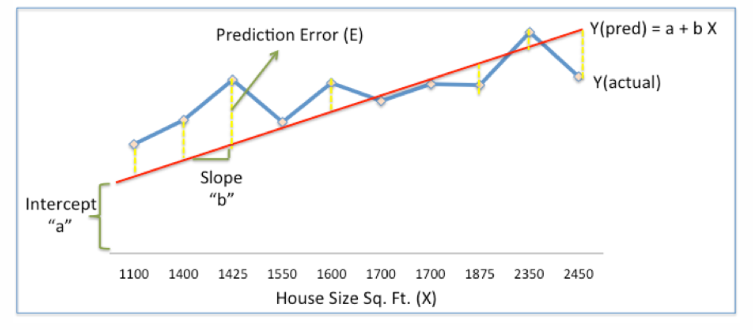


The task – for a new house, given its size (X), what will its price (Y) be?

Lets start off by plotting the historical housing data:



Now, we will use a simple linear model, where we fit a line on the historical data, to predict the price of a new house (Ypred) given its size (X)



In the above chart, the red line gives the predicted house price (Ypred) given house size (X).

Ypred = a+bX

The blue line gives the actual house prices from historical data (Yactual)

The difference between Yactual and Ypred (given by the yellow dashed lines) is the prediction error (E)

So, we need to find a line with optimal values of a,b (called weights) that best fits the historical data by reducing the prediction error and improving prediction accuracy.

So, our objective is to find optimal **a, b** that minimizes the error between actual and predicted values of house price (1/2 is for mathematical convenience since it helps in calculating gradients in calculus)

**Sum of Squared Errors (SSE) = ½ Sum (Actual House Price – Predicted House Price)2** = **½ Sum(Y – Ypred)2**

(Please note that there are other measures of Error. SSE is just one of them.)

This is where Gradient Descent comes into the picture. Gradient descent is an optimization algorithm that finds the optimal weights (a,b) that reduces  prediction error.

Lets now go step by step to understand the **Gradient Descent algorithm:**

**Step 1: Initialize the weights(a & b) with random values and calculate Error (SSE)**

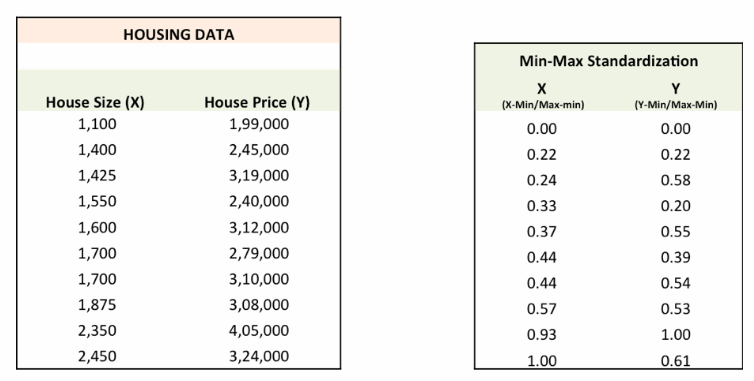
**Step 2: Calculate the gradient i.e. change in SSE when the weights (a & b) are changed by a very small value from their original randomly initialized value. This helps us move the values of a & b in the direction in which SSE is minimized.**

**Step 3: Adjust the weights with the gradients to reach the optimal values where SSE is minimized**

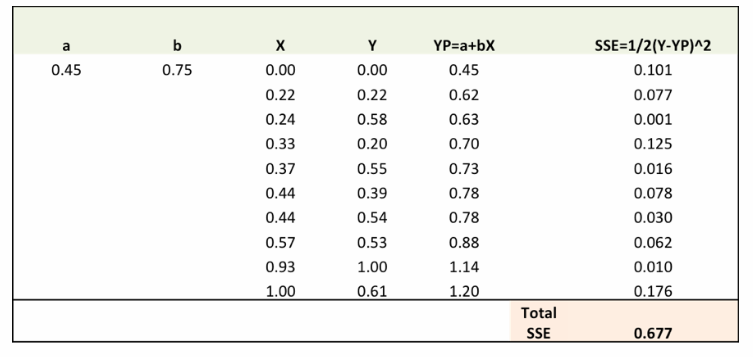
**Step 4: Use the new weights for prediction and to calculate the new SSE**

**Step 5: Repeat steps 2 and 3 till further adjustments to weights doesn’t significantly reduce the Error**

We will now go through each of the steps in detail (I worked out the steps in excel, which I have pasted below). But before that, we have to standardize the data as it makes the optimization process faster.



**Step 1:**To fit a line Ypred = a + b X, start off with random values of a and b and calculate prediction error (SSE)



**Step 2:**Calculate the error gradient w.r.t the weights

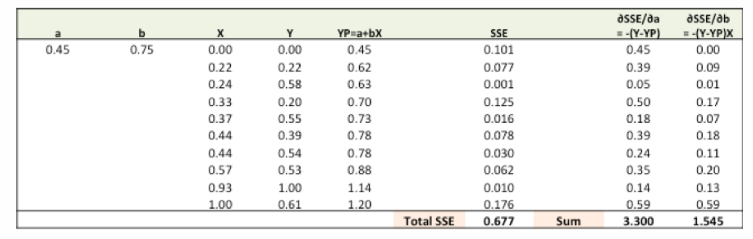
∂SSE/∂a = – (Y-YP)

∂SSE/∂b = – (Y-YP)X

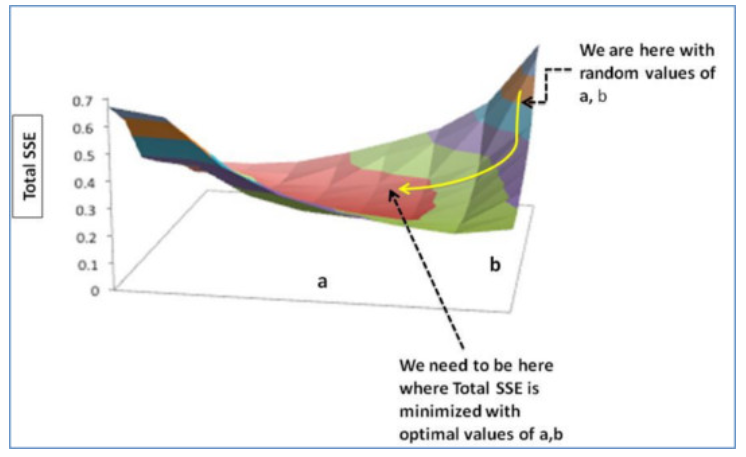
Here, SSE=½ (Y-YP)2 = ½(Y-(a+bX))2

You need to know a bit of calculus, but that’s about it!!

∂SSE/∂a and ∂SSE/∂b are the **gradients** and they give the direction of the movement of a,b w.r.t to SSE.



**Step 3:**Adjust the weights with the gradients to reach the optimal values where SSE is minimized



We need to update the random values of a,b so that we move in the direction of optimal a, b.

Update rules:

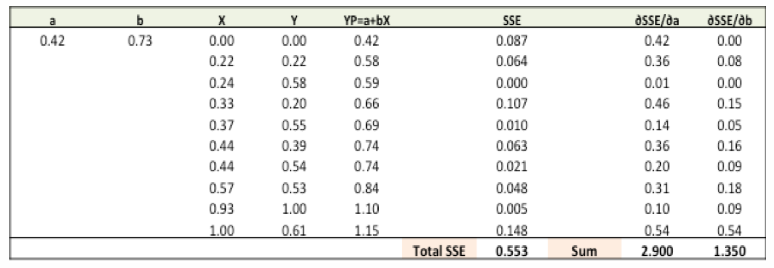
* a – ∂SSE/∂a
* b – ∂SSE/∂b

So, update rules:

1. New a = a – r \* **∂SSE/∂a =**0.45-0.01\*3.300 = 0.42
2. New b = b – r \* **∂SSE/∂b=**0.75-0.01\*1.545 = 0.73

here, r is the learning rate = 0.01, which is the pace of adjustment to the weights.

**Step 4:**Use new a and b for prediction and to calculate new Total SSE



You can see with the new prediction, the total SSE has gone down (0.677 to 0.553). That means prediction accuracy has improved.

**Step 5:**Repeat step 3 and 4 till the time further adjustments to a, b doesn’t significantly reduces the error. At that time, we have arrived at the optimal a,b with the highest prediction accuracy.

This is the Gradient Descent Algorithm. This optimization algorithm and its variants form the core of many machine learning algorithms like Neural Networks and even Deep Learning.