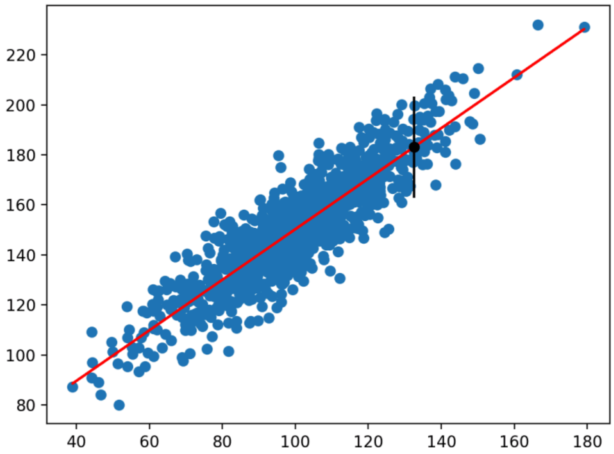
Author : Fahd Labba

Gradient Descent Clearly Explained

As a Data scientist I had many questions about gradient descent, how it works , why we need gradient descent in ML(Machine Learning).So I decided to write this article to clearly explain gradient descent without any technical requirement (you just need to be familiar with one of ML algorithms,in our case Simple Linear Regression).

1) Simple Linear Regression (Recap) :

Linear Regression is a supervised algorithm designed for regression tasks .To work with Linear Regression you first need a linear data point distribution like the figure above .



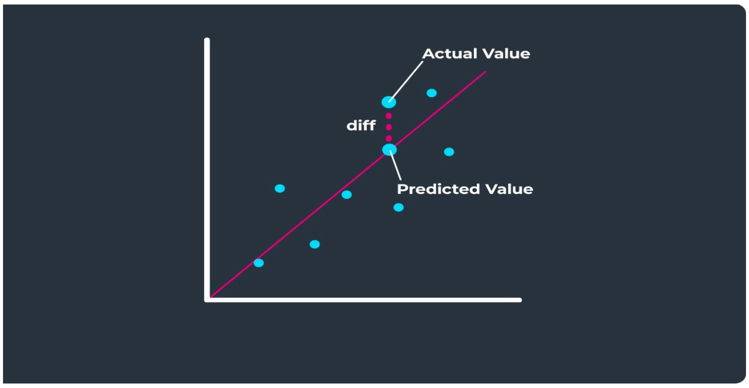
The main concept of Linear Regression is to fit a line (like the red line in the figure) to generalize well the distribution of the data point,so the equation for this line is y=f(x)=$\vec{w}\,\vec{x}$+b , b called bias and w weight vector,x called feature vector,for more simplicity we gonna work with one feature so line equation become y=f(x)=w*x+b.

2) Cost Function :

The main concept for cost function is to measure the error we make in our training phase (our main mission or let's say goal is to make this cost function as minimal as we can).Cost function noted as J(w,b).For regression tasks,the cost function J(w,b) is expressed as J(w,b)= $\sum_{i=0}^{m}(y_i$-f(x_i)$)^2$ which is also known as the Mean Squared Error (MSE).

f(x_i) also noted y_hat and its predicted values .

The figure below illustrates what this function is gonna calculate.



.

Author : Fahd Labba

In Generale Cost function :J(w,b) = $\sum_{i=0}^{m}$ L(w,b) with L(w,b) called loss function which is specific to your task.For example, in regression, we often use the Mean Squared Error (MSE).In classification,Binary Cross-Entropy is common.

3) Gradient Descent:

So if we recap what we've seen right now,first we fit a line to the data point and then measure our error value and by this error we know how our model does with the data point . so in order to get the best line (so the best perfermance) we need to get the w,b that have the min(J(w,b)).

And Here comes the magic of gradient descent which is gonna give you the best(w,b) so the best line that generalizes well our data point .

1) **How gradient descent finds w,b ?**

   The first step is to calculate the cost function J(w,b) and then we're gonna update w,b until our cost function converges to the global minimum.

2) **How does gradient descent update w,b ?**

   Here is the equation for w and b that gradient descent uses in order to update w ,b .

   $w=w-\alpha\frac{d}{dw}J(w,b)$

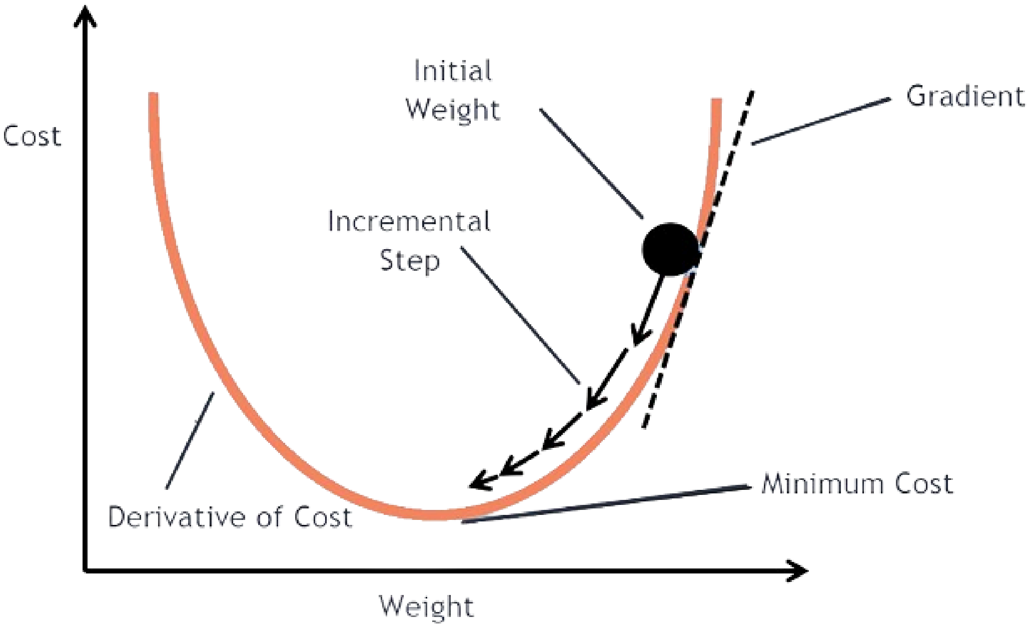
   $b=b-\alpha\frac{d}{db}J(w,b)$

   α **: Alpha or learning rate is the factor that decides how many steps we should take to update our parametre . is value is really sensitive and can cause many problems like(Vanishing or Exploring Gradient descent)**

   In case of vanishing GD alpha is a small number so we can't converge to our minimum and by this we don't get the best line .

   In other cases Exploring GD alpha is a big number so our cost function diverges,so we didn't get the best parametre w,b.

   NB : w,b should update simultaneous



3) **What’s next ?**
   after update our parameter we keep repeating the same process until we converge to our minimum or we achieve all the iteration number
   so the sum up :
   1) **Initialise w,b with random value (there are many techniques for this step and it can affect directly in your GD calculation).**
   2) **calculate cost function.**
   3) **update w,b(gradient descent).**
   4) **repeat 2 and 3 until we got the best w,b or achiever iteration number.**

Summary :

Gradient Descent plays a crucial part in your model performance(grantiring that you got the best parametre means grantiring the best model performance(most of the time)).And by Understanding GD very well means you’re able to understand all the ML algorithms.