Stochastic Gradient Descent

Mohammad Azodi, Shahid Beheshti University 27/07/2020

Optimization for data science
Teacher: Dr. Bijan Ahmadi

Stochastic Gradient Descent

A custom implementation of Stochastic Gradient Descent for Linear Regression that optimizes the weights **W(i)** of each component and the bias **b** term.

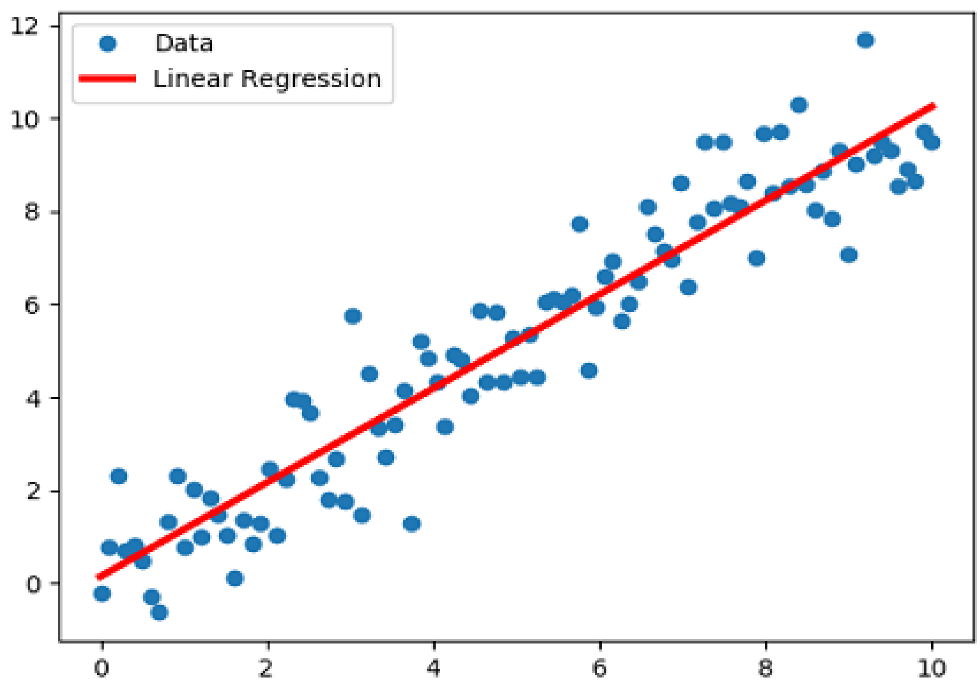
Gradient Descent:

Gradient Descent is an iterative optimization algorithm that can be used to converge to an optimal value easily with the use of modern computaional power.

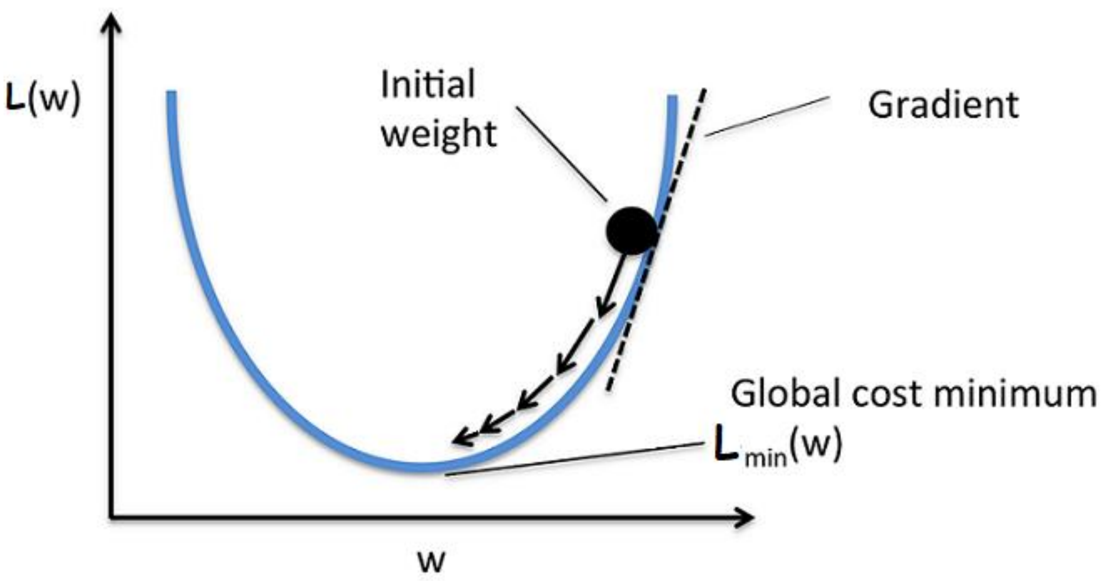
- The update equation for every iteration is, **x(i) = x(i-1) - r * [df/dx]_x(i-1)** ; **i : 1 -> n**.
- It mainly depends on the **learning rate** (or) **step size** denoted by **"r"**.
- The **learning rate** tells how fast to converge to the optimal value. So giving a right value of **"r"** matters.
- If the right value of **"r"** ain't given then the updation might jump over optimal value(min) and we'll not be converging at the right solution. Hence need to check with different values of **"r"**.

Let us consider a simple linear regression,

- **Objective :** Find the line/plane that best fits the data as,



- The dataset is **$D = \langle x_i, y_i \rangle$; $x_i \in R^d$; $y_i \in R$**. Here **d** is for # of dimensions.
- We can find the line/plane that fits the real values data of form **$y_i = W.T*x_i + B$** for given **x_i**. Here **B** is for bias.
- Then we've **Mean Squared Error(MSE)** = **$\mathrm{sum}([y_i - (W.T*x_i + B)]^2) / n$** ; **$i : 1 \rightarrow n$**.
- The optimal weight vector will be **$W^* = \mathrm{argmin}(W)\ \mathrm{sum}([y_i - (W.T*x_i + B)]^2) / n$** i.e the one which gives minimum sum of squared errors.



We can write the optimization problem **W* = argmin(W) sum([y_i - (W.T*x_i + B)]^2) / n** as,

L(W) = [*Σ*(y_i - (W.T*x_i + B))^2] / n ; **i : 1 -> n**.

Then the vector differentiation or grad of **L(w)** is ***∇*_w L = *Σ* { 2*(y_i - (W.T*x_i + B))(-x_i) } / n**