McGill University
Department of Computer Science

Gradient Descent on Linear Regression

A Quick Summary of Stochastic, Batch and Mini-batch Gradient Descent

Hair Albeiro Parra Barrera
260738619

March 5, 2020

1 Gradient Descent for Linear Regression

Suppose we have a hypothesis $h : \mathbb{R}^n \to \mathbb{R}$, $h_\theta(\mathbf{x}) = \hat{y}$ with paramters $\theta \in \mathbb{R}^n$. Recall the **Mean-Squared Loss** (MSE) metric, applied to linear regression:

$$MSE(y, h_\theta(\mathbf{x})) = \frac{1}{2n}||\mathbf{y} - h_\theta(\mathbf{x})||_2^2 = \frac{1}{2n}\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2 = \frac{1}{2n}\sum_{i=1}^{n}(y^{(i)} - \mathbf{w}^T x^{(i)})^2$$

In order to minimize the MSE, we take partials w.r.t. each parameter, and have the general update:

$$\mathbf{w}^{k+1} = \mathbf{w}^k + \alpha\frac{\partial}{\partial \mathbf{w}} MSE(y, \hat{y})$$

As the MSE is a **convex function**, it is guaranteed to have a global optimum, so given an appropriate choice of α, also called the **learning rate**, the algorithm will converge. In what follows, the algorithm stops whenever $||\mathbf{w}^k - \mathbf{w}^{k-1}|| < \epsilon$ or $|MSE_k - MSE_{k-1}| < \epsilon$,for some small ϵ.

Batch Gradient Descent: For $k = 0, 1, \dots$

1. For $k = 0, 1, \dots$

$$\mathbf{w}^{k+1} = \mathbf{w}^k + \alpha\frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - \mathbf{w}^T x^{(i)})\mathbf{x}^{(i)}$$

Mini-batch Gradient Descent:

1. For $k = 0, 1, \dots$
   (a) Split data D into T subsets D_t of sizes $n_0, \dots, n_{T-1}$, s.t. $\sum_t n_t = 1$.
   (b) For each subset D_t:

$$\mathbf{w} := \mathbf{w} + \alpha\frac{1}{n_t}\sum_{i:x^{(i)} \in D_t}^{n_t}(y^{(i)} - \mathbf{w}^T x^{(i)})\mathbf{x}^{(i)}$$

Stochastic Gradient Descent:

1. For $k = 0, 1, \dots$
   (a) For $i = 1, \dots, n$:

$$\mathbf{w} := \mathbf{w} + \alpha(y^{(i)} - \mathbf{w}^T x^{(i)})\mathbf{x}^{(i)}$$