Gradient Descent Optimization Techniques. An overview

David Panou

Pierre & Marie Curie University

david.panou@gmail.com

October 3, 2016

Overview

1. Introduction
2. Standard Gradient Descent Techniques
   - Stochastic Gradient Formalization
   - Batch Gradient Descent
   - Stochastic Gradient Descent
   - Mini-batch Gradient Descent
3. More in-depth Gradient Descent Procedures
   - Momentum
   - Nesterov Accelerated Gradient
4. Automated Learning Rate Selection
   - Introduction
   - AdaGrad
   - AdaDelta
   - RMSprop
   - Adam
5. Annex

Plan

1. Introduction
2. Standard Gradient Descent Techniques
   - Stochastic Gradient Formalization
   - Batch Gradient Descent
   - Stochastic Gradient Descent
   - Mini-batch Gradient Descent
3. More in-depth Gradient Descent Procedures
   - Momentum
   - Nesterov Accelerated Gradient
4. Automated Learning Rate Selection
   - Introduction
   - AdaGrad
   - AdaDelta
   - RMSprop
   - Adam
5. Annex

Introduction

This slide is a summary of [?] that introduces optimization techniques for Gradient Descent parameter optimization techniques.

We will go through the most commonly used and implemented Gradient Descent Strategies and give intuitive idea behind there formalization.

Plan

1. Introduction
2. Standard Gradient Descent Techniques
   - Stochastic Gradient Formalization
   - Batch Gradient Descent
   - Stochastic Gradient Descent
   - Mini-batch Gradient Descent
3. More in-depth Gradient Descent Procedures
   - Momentum
   - Nesterov Accelerated Gradient
4. Automated Learning Rate Selection
   - Introduction
   - AdaGrad
   - AdaDelta
   - RMSprop
   - Adam
5. Annex

Introduction

Before jumping into Gradient Descent strategy, we would like to remind to the reader the different existing Gradient Descent Formalization.

Those methods are dealing with the way the trained dataset is presented to the network As simple method, they doesn't adapt themselves to the problem at hand as each parameter (i.e learning rate, stopping) have to be set by hand

Gradient Descent Formalization

Definition

Gradient Descent procedures takes a learning problem and try to find a solution of set of linear inequalities

It does so by defining a criterion function $J(\theta)$, that is minimized if $\theta^T y_i > 0$, with θ being the algorithm parameters.

This very simple procedure reduces the learning problem to minimizing a scalar function.

The gradient descent pseudo algorithm is given at the end of those slide as a reminder.

Batch Gradient Descent

Definition

The most simple formalization of Gradient Descent is the Batch Gradient Descent. In Batch Gradient Descent, the learning algorithm takes all of the training dataset points at each training epoch.

Characteristics

Batch gradient descent is guaranteed to converge to the global minimum for convex functions and to a local minimum for non-convex ones.

However it can be intractable for datasets that doesn't fit in memory and can be very slow. Finally, for large datasets it performs redundant computations. Those problems are addressed by Stochastic Gradient Descent.

Stochastic Gradient Descent

Definition

Stochastic Gradient Descent (SGD) contrary to the Batch version, doesn't compute all of the training set in one time per epoch. This method is called stochastic because the training data can be considered a random variable that is chosen randomly in the training set.

Characteristics

Given the way SGD proceeds data it is therefore usually much faster than the batch version and can also be used to learn online. A weight update may reduce the error on the single pattern being presented, yet increase the error on the full training set. Because of this, the Objective function fluctuates heavily.

Mini-batch Gradient Descent

Definition

Mini-batch Gradient Descent takes the best of both worlds and proceeds an update for every mini-batch n of training examples.

Characteristics

This method is the most commonly used one.

1. Since its variance of a set of selected points is smaller than one of a single point, the convergence is more stable than the one of SGD.

Plan

1. Introduction
2. Standard Gradient Descent Techniques
   - Stochastic Gradient Formalization
   - Batch Gradient Descent
   - Stochastic Gradient Descent
   - Mini-batch Gradient Descent
3. More in-depth Gradient Descent Procedures
   - Momentum
   - Nesterov Accelerated Gradient
4. Automated Learning Rate Selection
   - Introduction
   - AdaGrad
   - AdaDelta
   - RMSprop
   - Adam
5. Annex

Introduction

After reviewing the standard ways to perform Gradient Descent, Ruder goes through several more in-depth Gradient Optimization Procedures that will be introduced in this section. It is not notify that only popular practicable methods are mentioned here.

Therefore, the author do not mention technics such as Newton's second order algorihtm that provides a analytical solution for fixing η, but that is not always practicable because of requirement on the Hessian Matrix of J.

Momentum

Why Momentum

In order to avoid getting in stuck in regions in which the error surfaces is a plateaus - regions in which the slope $\frac{dJ(\theta)}{dw}$ is very small-, which happens very often, we can had momentum in order to specify that

Momentum

Definition

Adding momentum is done by modifying the learning rule in stochastic backpropagation.

$$\gamma_t = \gamma v_{t-1} + \eta \nabla_\theta J(\theta)$$

It is to note that the momentum term (η) has to be less than 1 for stability issues. Thus it is usually set to 0.9.

Momentum rarely changes the final answer of Gradient Descent, but helps getting to it faster.

Momentum Illustrations

Remarks

1. The momentum helps to have faster convergence.
2. With or without, the convergence is usually the same

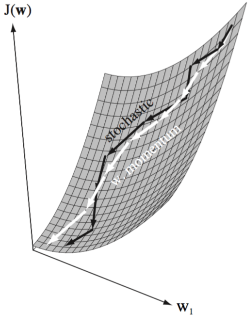

Illustration taken from [?]

Nesterov Accelerated Gradient

Description

Like momentum, Nesterov Accelerated Gradient is a first-order optimization method with better convergence rate guarantee.

As we have seen with Momentum, adding momentum to the parameter optimization procedure gives us a way to achieve faster convergences, but by momentum, we are adding velocity, we might as well give it a way to "slow down" when it reaches other ramp up.

Nesterov Accelerated Gradient Examples

Figure: Top is a classical momentum trajectory. Down is Nesterov Accelerated Gradient.

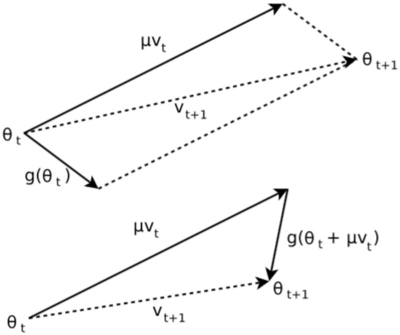

Illustration taken from [?]

Plan

1. Introduction
2. Standard Gradient Descent Techniques
   - Stochastic Gradient Formalization
   - Batch Gradient Descent
   - Stochastic Gradient Descent
   - Mini-batch Gradient Descent
3. More in-depth Gradient Descent Procedures
   - Momentum
   - Nesterov Accelerated Gradient
4. Automated Learning Rate Selection
   - Introduction
   - AdaGrad
   - AdaDelta
   - RMSprop
   - Adam
5. Annex

Introduction

Description

Up until now, all method had to be given an explicit learning rate.

We remind that this one can be calculated through second order method but that those methods are not always practicable - and even most of the time, unpracticable - thus they don't figure in this list.

AdaGrad

Description

Adagrad provides a way to only update parameters that are relevants for a given training example.

Doing so, Adagrad provides a way to achieve better robustness. But more : it also remove the needs to tune the learning rate manually

AdaDelta

Description

Adagrad that seeks to reduce its aggressive, monotonically decreasing learning rate. Instead of accumulating all past squared gradients, Adadelta restricts the window of accumulated past gradients to some fixed size w.

RMSprop

Description

RMSprop comes with the same intuition as AdaGrad. They in fact, have both been developed independently to solve the same issue.

It divides the learning rate by an exponentially decaying average of squared gradients.

Adam

Description

Adam stands for Adaptifve Moment Estimation
Like AdaGrad and RMSprop Adam also keeps an exponentially decaying average of past gradient

Plan

1. Introduction
2. Standard Gradient Descent Techniques
   - Stochastic Gradient Formalization
   - Batch Gradient Descent
   - Stochastic Gradient Descent
   - Mini-batch Gradient Descent
3. More in-depth Gradient Descent Procedures
   - Momentum
   - Nesterov Accelerated Gradient
4. Automated Learning Rate Selection
   - Introduction
   - AdaGrad
   - AdaDelta
   - RMSprop
   - Adam
5. **Annex**

Gradient Descent Formalization

Basic Gradient Descent Algorithm

```
Input: $a, \theta, \eta(.), criterion$
    $k \leftarrow 0$
    repeat
        $a \leftarrow a - \eta(k)\nabla J(\theta)$
    until $\eta(k)\nabla J(\theta) < criterion$
Output: $a$
```