CS 4980/6980: Introduction to Data Science © Spring 2018

Lecture 7: Gradient Descent

INSTRUCTOR: DANIEL L. PIMENTEL-ALARCÓN

Scribed by: John Le

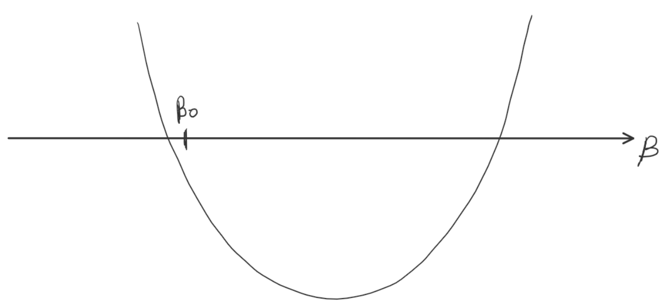
This is preliminary work and has not been reviewed by instructor. If you have comments about typos, errors, notation inconsistencies, etc., please email the scribers.

7.1 Introduction

Gradient descent is an optimization algorithm to find the local minimum through the steps of initializing a point and calculating the next point until you've reached a local minimum. When going positive to find the local maximum, its called Gradient Ascent.

7.2 Using Gradient Descent

Suppose that we have this for $f(\beta)$



Goal: Minimum $\nabla\ f(\beta)$

7.3 Steps

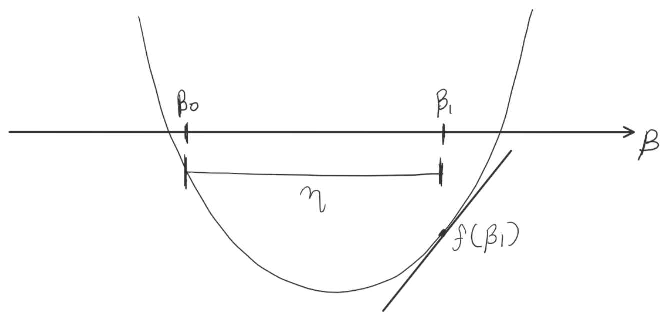
Step 1: Initialize with β and evaluate $f(\beta_0)$
.

Step 2: Compute the gradient $\nabla f(\beta_0)$
.

Step 3: Evaluate $\nabla f(\beta_0)$
.

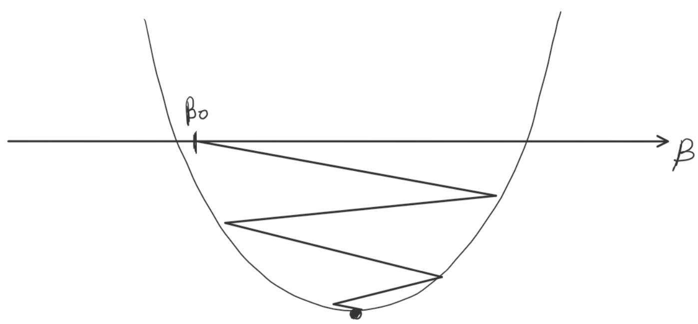
Step 4: Move in the direction of the gradient using a scalar step size η("eta"), where $\eta \epsilon \mathbb{R}$, to obtain

$$\beta_1 = \beta_0 + \eta \nabla f(\beta_0)$$



Step 5: For any t $\geq$ 1, repeat... until convergences.

$$\beta = \beta_{t-1} + \eta_t \nabla f(\beta_{t-1})$$



7.4 Now to find the log of gradient decent.

$$\ell(\beta) = \sum_{i=1}^{N} \ell og(\frac{1}{1+e^{\beta^T X_i}})^{y_i} + \ell(1 - \frac{1}{1+e^{\beta^T X_i}})^{1-y_i}$$

Where...

$$\nabla \ell(\beta) = \begin{bmatrix} \frac{\partial \ell(\beta)}{\partial \beta_1} \\ \frac{\partial \ell(\beta)}{\partial \beta_2} \\ \vdots \\ \frac{\partial \ell(\beta)}{\partial \beta_D} \end{bmatrix}$$

$$\ell(\beta_1\beta_2...\beta_D) = \sum_{i=1}^{N} \ell og(\frac{1}{1+e^{-\sum_{d=1}^{D}\beta_d X_{id}}})^{y_i} + \ell(1 - \frac{1}{1+e^{-\sum_{d=1}^{D}\beta_d X_{id}}})^{1-y_i}$$

$$= \sum_{i=1}^{N} y_i \ell og(\frac{1}{1+e^{-\beta^T X_i}}) + (1-y_i)(1 - \frac{1}{1+e^{-\beta^T X_i}})$$

$$= \sum_{i=1}^{N} -y_i \ell og(1+e^{-\beta^T X_i}) + (1-y_i)\Big[\ell og(e^{\beta^T x_i}) - \ell og(1+e^{\beta^T x_i})\Big]$$

$$= \sum_{i=1}^{N} \cancel{-y_i} \ell og(1+e^{-\beta^T X_i}) + (1-y_i)\ell og(e^{-\beta^T x_i}) - (1 \cancel{- yi})\ell og(1+e^{-\beta^T x_i})$$

$$= \sum_{i=1}^{N} (1-y_i)(-\beta^T x_i) - \ell og(1+e^{-\beta^T x_i})$$

Find the Derivative

$$\frac{\partial \ell(\beta)}{\partial \beta_1} = \frac{\partial}{\partial \beta_1}\Big[\sum_{i=1}^{N}(y_i - 1)\sum_{d=1}^{D}\beta_d X_{id} - \ell og(1+e^{-\sum_{d=1}^{D}\beta_d X_{id}})\Big]$$

$$= \sum_{i=1}^{N}(y-1)\underbrace{\frac{\partial}{\partial \beta_1}\sum_{d=1}^{D}\beta_1 X_{id}}_{x_{i1}\text{(Since we are only looking at }\beta_1)} - \sum_{i=1}^{N}\frac{\partial}{\partial \beta_1}\ell og(1+e^{-\sum_{d=1}^{D}\beta_d X_{id}})$$

$$= \sum_{i=1}^{N} (y_i - 1)x_{i1} - \sum_{i=1}^{N} \frac{1}{1+e^{-\beta^T X_i}} \underbrace{\frac{\partial}{\partial \beta_1}(1 + e^{-\sum_{d=1}^{D} \beta^T x_{id}})}$$

$$-e^{-\beta^T X_i} \frac{\partial}{\partial \beta_1}(-\sum_{d=1}^{D} \beta_2 X_i d)$$

$$= \sum_{i=1}^{N} (y_i - 1 + \frac{e^{-\beta^T X_i}}{1+e^{-\beta^T X_i}})X_i 1$$

$$= \sum_{i=1}^{N} (y_i - \frac{-1 - \cancel{e^{-\beta^T X_i}}}{1+e^{-\beta^T X_i}} + \frac{\cancel{e^{-\beta^T X_i}}}{1+e^{-\beta^T X_i}})X_{i1}$$

$$\boxed{\nabla \ell(\beta) = \sum_{i=1}^{N} (y_i - \frac{1}{1+e^{-\beta^T X_i}})X_{i1}}$$