Machine Learning: Yoshinari Fujinuma

University of Colorado Boulder

LECTURE 10

Slides adapted from Chenhao Tan, Jordan Boyd-Graber, Chris Ketelsen

Logistics

- HW2 available on Github
- Small edits made on the slides from the last lecture
- Final project team formulation due on March 1st

Learning objectives

- Understand stochastic gradient descent

Outline

Stochastic Gradient Descent (SGD)

Problem with Gradient Descent

- $(\boldsymbol{x}, y) = \{(\boldsymbol{x}_0, y_0), ..., (\boldsymbol{x}_N, y_N)\}$: Training data with N examples. $\boldsymbol{x}_{ij}$ is the jth feature of ith example.
- $\beta = (\beta_0, ..., \beta_d)$: Parameters of logistic regression.
- η: Learning rate (step size)

Updating parameters by gradient descent is

$$\beta_j' \leftarrow \beta_j - \eta \frac{\partial \mathcal{L}}{\partial \beta_j} \tag{1}$$

where $\frac{\partial \mathcal{L}}{\partial \beta_j}$ for logistic regression is

$$\frac{\partial \mathcal{L}}{\partial \beta_j} = \sum_i^N - (y_i - \sigma_i)\boldsymbol{x}_{ij} \tag{2}$$

Probelm: $\sum_i^N$ indicates that you need to go through **all** trianing examples

Approximating the Gradient

- Training datasets are big these days (to fit into memory)
- What if we compute an update just from one observation?

Intuition of SGD: Analogy to Getting to Union Station

Pretend it's a pre-smartphone world and you want to get to Union Station

Stochastic Gradient Descent

- $(\boldsymbol{x}, y) = \{(\boldsymbol{x}_0, y_0), ..., (\boldsymbol{x}_N, y_N)\}$: Training data with N examples. x_{ij} is the jth feature of ith example.
- $\beta = (\beta_0, ..., \beta_n)$: Parameters of logistic regression.
- η: Learning rate (step size)

Updating the parameters by stochastic gradient descent

$$\beta'_j \leftarrow \beta_j - \eta \frac{\partial \mathcal{L}}{\partial \beta_j} \tag{3}$$

where $\frac{\partial \mathcal{L}}{\partial \beta_j}$ for logistic regression is

$$\frac{\partial \mathcal{L}}{\partial \beta_j} = -(y_i - \sigma_i)\boldsymbol{x}_{ij} \tag{4}$$

We now compute $\frac{\partial \mathcal{L}}{\partial \beta_j}$ without $\sum_i^N$ i.e., only from **one** trianing example

Gradient Descent vs. Stochastic Gradient Descent

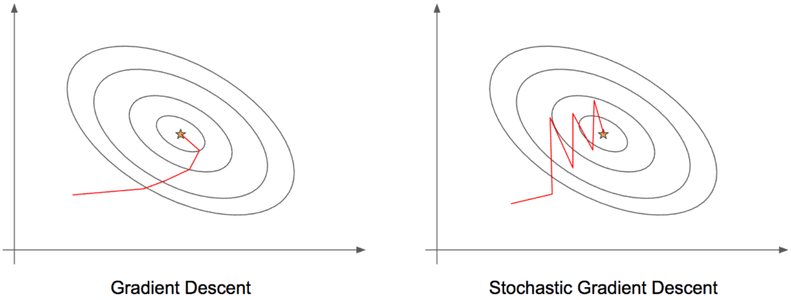


Image from `https://pythonmachinelearning.pro/complete-guide-to-deep-neural-networks-part-2/`

SGD Example

Note that $-\eta \frac{\partial \mathscr{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle \beta_0 = 0, \beta_1 = 0, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0 \rangle$$

$y_1 = 1$	$y_2 = 0$
$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$	$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

You first see the positive example. First, compute σ_1

SGD Example

Note that $-\eta \frac{\partial \mathscr{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$	$y_2 = 0$
$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$	$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

You first see the positive example. First, compute σ_1

$\sigma_1 = \Pr(y_1 = 1 \,|\, \boldsymbol{x}_1) = \frac{1}{1 + \exp - \beta^T \boldsymbol{x}_1} =$

SGD Example

Note that $-\eta \frac{\partial \mathscr{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$	$y_2 = 0$
$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$	$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

You first see the positive example. First, compute σ_1

$\sigma_1 = \mathrm{Pr}(y_1 = 1 \,|\, \boldsymbol{x}_1) = \frac{1}{1+\exp -\beta^T \boldsymbol{x}_1} = \frac{1}{1+\exp 0} = 0.5$

SGD Example

Note that $-\eta \frac{\partial \mathscr{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$	$y_2 = 0$
$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$	$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

$\sigma_1 = 0.5$ What's the updated β_0'?

SGD Example

Note that $-\eta \frac{\partial \mathcal{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\begin{aligned} \beta_j' =& \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij} \\ \vec{\beta} =& \langle 0, 0, 0, 0, 0 \rangle \end{aligned}$$

$y_1 = 1$	$y_2 = 0$
$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$	$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_0'? $\beta_0' = \beta_0 + \eta \cdot (y_1 - \sigma_1) \cdot \boldsymbol{x}_{1,0} = 0.0 + 1.0 \cdot (1.0 - 0.5) \cdot 1.0$

SGD Example

Note that $-\eta \frac{\partial \mathcal{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$

$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$

$y_2 = 0$

$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_0'? $\beta_0' = \beta_0 + \eta \cdot (y_1 - \sigma_1) \cdot \boldsymbol{x}_{1,0} = 0.0 + 1.0 \cdot (1.0 - 0.5) \cdot 1.0$
=0.5

SGD Example

Note that $-\eta \frac{\partial \mathcal{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$

$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$

$y_2 = 0$

$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_1'?

SGD Example

Note that $-\eta \frac{\partial \mathcal{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$	$y_2 = 0$
$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$	$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_1'? $\beta_1' = \beta_1 + \eta \cdot (y_1 - \sigma_1) \cdot \boldsymbol{x}_{1,1} = 0.0 + 1.0 \cdot (1 - 0.5) \cdot 4$

SGD Example

Note that $-\eta\frac{\partial \mathcal{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$

$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$

$y_2 = 0$

$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_1'? $\beta_1' = \beta_1 + \eta \cdot (y_1 - \sigma_1) \cdot \boldsymbol{x}_{1,1} = 0.0 + 1.0 \cdot (1 - 0.5) \cdot 4$ =2.0

SGD Example

Note that $-\eta \frac{\partial \mathscr{L}}{\partial \beta_j} = \eta (y_i - \sigma_i) \boldsymbol{x}_{ij}$

$$\begin{aligned} \beta_j' =& \beta_j + \eta (y_i - \sigma_i) \boldsymbol{x}_{ij} \\ \vec{\beta} =& \langle 0, 0, 0, 0, 0 \rangle \end{aligned}$$

$y_1 =$1	$y_2 =$0
$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$	$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_2'?

SGD Example

Note that $-\eta \frac{\partial \mathcal{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$

$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$

$y_2 = 0$

$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_2'? $\beta_2' = \beta_2 + \eta \cdot (y_1 - \sigma_1) \cdot \boldsymbol{x}_{1,2} = 0.0 + 1.0 \cdot (1 - 0.5) \cdot 3$

SGD Example

Note that $-\eta \frac{\partial \mathscr{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$	$y_2 = 0$
$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$	$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_2'? $\beta_2' = \beta_2 + \eta \cdot (y_1 - \sigma_1) \cdot \boldsymbol{x}_{1,2} = 0.0 + 1.0 \cdot (1 - 0.5) \cdot 3$ =1.5

SGD Example

Note that $-\eta\frac{\partial\mathscr{L}}{\partial\beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$
$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$

$y_2 = 0$
$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_3'?

SGD Example

Note that $-\eta \frac{\partial \mathscr{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$	$y_2 = 0$
$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$	$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_3'? $\quad \beta_3' = \beta_3 + \eta \cdot (y_1 - \sigma_1) \cdot \boldsymbol{x}_{1,3} = 0.0 + 1.0 \cdot (1 - 0.5) \cdot 1$

SGD Example

Note that $-\eta \frac{\partial \mathscr{L}}{\partial \beta_j} = \eta (y_i - \sigma_i) \boldsymbol{x}_{ij}$

$$\begin{aligned} \beta_j' =& \beta_j + \eta (y_i - \sigma_i) \boldsymbol{x}_{ij} \\ \vec{\beta} =& \langle 0, 0, 0, 0, 0 \rangle \end{aligned}$$

$y_1 =$1	$y_2 =$0
$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$	$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_3'? $\beta_3' = \beta_3 + \eta \cdot (y_1 - \sigma_1) \cdot \boldsymbol{x}_{1,3} = 0.0 + 1.0 \cdot (1 - 0.5) \cdot 1$ =0.5

SGD Example

Note that $-\eta\frac{\partial\mathcal{L}}{\partial\beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0\rangle$$

$y_1 = 1$	$y_2 = 0$
$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$	$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_4'?

SGD Example

Note that $-\eta \frac{\partial \mathcal{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$	$y_2 = 0$
$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$	$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_4'? $\beta_4' = \beta_4 + \eta \cdot (y_1 - \sigma_1) \cdot \boldsymbol{x}_{1,4} = 0.0 + 1.0 \cdot (1.0 - 0.5) \cdot 0$

SGD Example

Note that $-\eta\frac{\partial \mathscr{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$	$y_2 = 0$
$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$	$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_4'? $\beta_4' = \beta_4 + \eta \cdot (y_1 - \sigma_1) \cdot \boldsymbol{x}_{1,4} = 0.0 + 1.0 \cdot (1.0 - 0.5) \cdot 0$
=0.0

SGD Example

Note that $-\eta \frac{\partial \mathcal{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0.5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$
$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$

$y_2 = 0$
$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

Now you see the negative example. What's σ_2?

SGD Example

Note that $-\eta \frac{\partial \mathscr{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0.5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$	$y_2 = 0$
$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$	$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

Now you see the negative example. What's σ_2?

$\sigma_2 = \Pr(y_2 = 0 \,|\, \vec{x_2}) = \frac{\exp -\beta^T x_i}{1 + \exp -\beta^T x_i} = \frac{\exp\{-(.5+1.5+1.5+0)\}}{1+\exp\{-(.5+1.5+1.5+0)\}} =$

SGD Example

Note that $-\eta \frac{\partial \mathscr{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0.5, 2, 1.5, 0.5, 0\rangle$$

$y_1 = 1$	$y_2 = 0$
$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$	$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

Now you see the negative example. What's σ_2?

$$\sigma_2 = \Pr(y_2 = 0 \,|\, \vec{x_2}) = \frac{\exp -\beta^T x_i}{1+\exp -\beta^T x_i} = \frac{\exp\{-(.5+1.5+1.5+0)\}}{1+\exp\{-(.5+1.5+1.5+0)\}} = 0.97$$

SGD Example

Note that $-\eta \frac{\partial \mathscr{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0.5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$

$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$

$y_2 = 0$

$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

Now you see the negative example. What's σ_2?

$\sigma_2 = 0.97$

What's the updated β_0'?

SGD Example

Note that $-\eta \frac{\partial \mathcal{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0.5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$

$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$

$y_2 = 0$

$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_0'? $\beta_0' = \beta_0 + \eta \cdot (y_2 - \sigma_2) \cdot \boldsymbol{x}_{2,0} = 0.5 + 1.0 \cdot (0 - 0.97) \cdot 1$

SGD Example

Note that $-\eta \frac{\partial \mathcal{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0.5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$

$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$

$y_2 = 0$

$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_0'? $\beta_0' = \beta_0 + \eta \cdot (y_2 - \sigma_2) \cdot \boldsymbol{x}_{2,0} = 0.5 + 1.0 \cdot (0 - 0.97) \cdot 1$
=-0.47

SGD Example

Note that $-\eta \frac{\partial \mathcal{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0.5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$

$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$

$y_2 = 0$

$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_1'?

SGD Example

Note that $-\eta \frac{\partial \mathcal{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0.5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$

$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$

$y_2 = 0$

$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_1'? $\beta_1' = \beta_1 + \eta \cdot (y_2 - \sigma_2) \cdot \boldsymbol{x}_{2,1} = 2.0 + 1.0 \cdot (0 - 0.97) \cdot 0$

SGD Example

Note that $-\eta\frac{\partial \mathcal{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0.5, 2, 1.5, 0.5, 0\rangle$$

$y_1 = 1$	$y_2 = 0$
$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$	$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_1'? $\beta_1' = \beta_1 + \eta \cdot (y_2 - \sigma_2) \cdot \boldsymbol{x}_{2,1} = 2.0 + 1.0 \cdot (0 - 0.97) \cdot 0$
=2.0

SGD Example

Note that $-\eta \frac{\partial \mathcal{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0.5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$

$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$

$y_2 = 0$

$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_2'?

SGD Example

Note that $-\eta\frac{\partial \mathcal{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0.5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$
$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$

$y_2 = 0$
$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_2'? $\beta_2' = \beta_2 + \eta \cdot (y_2 - \sigma_2) \cdot \boldsymbol{x}_{2,2} = 1.5 + 1.0 \cdot (0 - 0.97) \cdot 1$

SGD Example

Note that $-\eta \frac{\partial \mathcal{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0.5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$

$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$

$y_2 = 0$

$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_2'? $\beta_2' = \beta_2 + \eta \cdot (y_2 - \sigma_2) \cdot \boldsymbol{x}_{2,2} = 1.5 + 1.0 \cdot (0 - 0.97) \cdot 1$
=0.53

SGD Example

Note that $-\eta \frac{\partial \mathcal{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\begin{aligned} \beta_j' =& \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij} \\ \vec{\beta} =& \langle 0.5, 2, 1.5, 0.5, 0 \rangle \end{aligned}$$

$y_1 =$ 1

$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$

$y_2 =$ 0

$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_3'?

SGD Example

Note that $-\eta \frac{\partial \mathcal{L}}{\partial \beta_j} = \eta (y_i - \sigma_i) \boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta (y_i - \sigma_i) \boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0.5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$

$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$

$y_2 = 0$

$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_3'? $\beta_3' = \beta_3 + \eta \cdot (y_2 - \sigma_2) \cdot \boldsymbol{x}_{2,3} = 0.5 + 1.0 \cdot (0 - 0.97) \cdot 3$

SGD Example

Note that $-\eta \frac{\partial \mathscr{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0.5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$

$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$

$y_2 = 0$

$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_3'? $\beta_3' = \beta_3 + \eta \cdot (y_2 - \sigma_2) \cdot \boldsymbol{x}_{2,3} = 0.5 + 1.0 \cdot (0 - 0.97) \cdot 3$
=-2.41

SGD Example

Note that $-\eta \frac{\partial \mathcal{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0.5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$	$y_2 = 0$
$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$	$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_4'?

SGD Example

Note that $-\eta \frac{\partial \mathscr{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\begin{aligned} \beta_j' =& \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij} \\ \vec{\beta} =& \langle 0.5, 2, 1.5, 0.5, 0 \rangle \end{aligned}$$

$y_1 =$1
$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$

$y_2 =$0
$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_4'? $\beta_4' = \beta_4 + \eta \cdot (y_2 - \sigma_2) \cdot \boldsymbol{x}_{2,4} = 0 + 1.0 \cdot (0 - 0.97) \cdot 4$

SGD Example

Note that $-\eta\frac{\partial\mathcal{L}}{\partial\beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$$
$$\vec{\beta} = \langle 0.5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$

$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$

$y_2 = 0$

$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

What's the updated β_4'? $\beta_4' = \beta_4 + \eta \cdot (y_2 - \sigma_2) \cdot \boldsymbol{x}_{2,4} = 0 + 1.0 \cdot (0 - 0.97) \cdot 4$
=-3.88

SGD Example

Note that $-\eta \frac{\partial \mathcal{L}}{\partial \beta_j} = \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$

$$\begin{aligned} \beta_j' &= \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij} \\ \vec{\beta} &= \langle -0.47, 2, 0.53, -2.41, -3.88 \rangle \end{aligned}$$

$y_1 = 1$	$y_2 = 0$
$\boldsymbol{x}_1 = (1, 4, 3, 1, 0)$	$\boldsymbol{x}_2 = (1, 0, 1, 3, 4)$

(Assume step size $\eta = 1.0$.)

Overview of Optimizing β using SGD

1. Initialize a vector β to be all zeros
2. For $t = 1, \ldots, T$ (i.e. number of epochs)
   - For each example $\boldsymbol{x}_i, y_i$ and each feature j:
     - Compute $\sigma_i = \text{Pr}(y_i \,|\, \boldsymbol{x}_i)$
     - Set $\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$
3. Output the parameters $\beta_0, \ldots, \beta_d$.

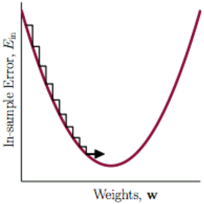
Overview of Optimizing β using SGD

1. Initialize a vector β to be all zeros
2. For $t = 1, \ldots, T$ (i.e. number of epochs)
   - For each example $\boldsymbol{x}_i, y_i$ and each feature j:
     - Compute $\sigma_i = \text{Pr}(y_i \,|\, \boldsymbol{x}_i)$
     - Set $\beta_j' = \beta_j + \eta(y_i - \sigma_i)\boldsymbol{x}_{ij}$
3. Output the parameters $\beta_0, \ldots, \beta_d$.

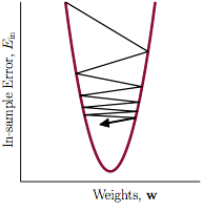
How to decide η?

Choosing learning rate

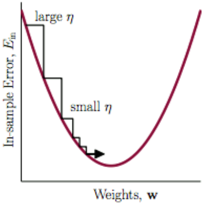
η **too small**



η **too large**



variable η_t **– just right**



Learning rate decay

Decay schedule can be seen as a hyperparameter too.

- Decay after each epoch (e.g., $\frac{\eta_0}{t^2}$)

Advanced stochastic gradient descent:

`http://ruder.io/optimizing-gradient-descent/`

(Bonus) Mini-Batch Stochastic Gradient Descent

- $(\boldsymbol{x}, y) = \{(\boldsymbol{x}_0, y_0), ..., (\boldsymbol{x}_N, y_N)\}$: Training data with N examples. x_{ij} is the jth feature of ith example.
- $\beta = (\beta_0, ..., \beta_n)$: Parameters of logistic regression.
- η: Learning rate (step size)

Updating the parameters by stochastic gradient descent

$$\beta'_j \leftarrow \beta_j - \eta \frac{\partial \mathcal{L}}{\partial \beta_j} \tag{5}$$

where $\frac{\partial \mathcal{L}}{\partial \beta_j}$ for logistic regression is

$$\frac{\partial \mathcal{L}}{\partial \beta_j} = \sum_{i}^{M} -(y_i - \sigma_i)\boldsymbol{x}_{ij} \tag{6}$$

We now compute $\frac{\partial \mathcal{L}}{\partial \beta_j}$ from M trianing examples (less noisy)