Machine Learning: Chenhao Tan

University of Colorado Boulder

LECTURE 10

Slides adapted from Jordan Boyd-Graber, Chris Ketelsen

Logistics

- HW2 available on Github, due in 7 days

Learning objectives

- Understand stochastic gradient descent

Outline

Stochastic Gradient Descent

Outline

Stochastic Gradient Descent

Review of Wednesday's lecture

Objective function:

$$\begin{aligned}\mathscr{L} &= -\sum_i \log P(y^{(i)} \mid \boldsymbol{x}^{(i)}, \beta)) + \frac{1}{2}\lambda \sum_j \beta_j^2 \\ &= \sum_i -y^{(i)} \left(\beta_0 + \sum_j \beta_i \boldsymbol{x}_j^{(i)}\right) + \log \left[1 + \exp\left(\beta_0 + \sum_j \beta_i \boldsymbol{x}_j^{(i)}\right)\right] + \frac{1}{2}\lambda \sum_j \beta_j^2\end{aligned}$$

Gradient descent:

$$\beta_j^{l+1} = \beta_j^l - \eta \frac{\partial \mathscr{L}}{\partial \beta_j}$$

Gradient:

$$\frac{\partial \mathscr{L}}{\partial \beta_j} = \sum_i [-(y_i - \pi_i)x_j] + \lambda \beta_j$$

Approximating the Gradient

- Our datasets are big (to fit into memory)
- . . . or data are changing / streaming

Approximating the Gradient

- Our datasets are big (to fit into memory)
- ...or data are changing / streaming
- Hard to compute true gradient

$$\mathscr{L}(\beta) \equiv \mathbb{E}_{\boldsymbol{x}}\left[\nabla \mathscr{L}(\beta, \boldsymbol{x})\right] \tag{1}$$

- Average over all observations

Approximating the Gradient

- Our datasets are big (to fit into memory)
- ...or data are changing / streaming
- Hard to compute true gradient

$$\mathscr{L}(\beta) \equiv \mathbb{E}_{\boldsymbol{x}}\left[\nabla \mathscr{L}(\beta, \boldsymbol{x})\right] \tag{1}$$

- Average over all observations
- What if we compute an update just from one observation?

Getting to Union Station

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

Stochastic Gradient for Regularized Regression

$$\mathscr{L} = -\log p(y \,|\, \boldsymbol{x}; \beta) + \frac{1}{2}\lambda \sum_j \beta_j^2 \tag{2}$$

Stochastic Gradient for Regularized Regression

$$\mathscr{L} = -\log p(y \,|\, \boldsymbol{x}; \beta) + \frac{1}{2}\lambda \sum_j \beta_j^2 \tag{2}$$

Taking the derivative (with respect to example x_i)

$$\frac{\partial \mathscr{L}}{\partial \beta_j} = -(y_i - \pi_i)x_{ij} + \lambda \beta_j \tag{3}$$

Stochastic Gradient for Logistic Regression

Given a **single observation** x_i chosen at random from the dataset,

$$\beta_j \leftarrow \beta_j' - \eta \left(\lambda \beta_j' - x_{ij} \left[y_i - \pi_i \right] \right) \tag{4}$$

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle \beta_{bias} = 0, \beta_A = 0, \beta_B = 0, \beta_C = 0, \beta_D = 0 \rangle$$

$y_1 = 1$	$y_2 = 0$
A A A A B B B C	B C C C D D D D

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

You first see the positive example. First, compute π_1

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$

A A A A B B B C

$y_2 = 0$

B C C C D D D D

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

You first see the positive example. First, compute π_1

$\pi_1 = \Pr(y_1 = 1 \mid \boldsymbol{x}_1) = \frac{\exp \beta^T \boldsymbol{x}_i}{1 + \exp \beta^T \boldsymbol{x}_i} =$

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$

A A A A B B B C

$y_2 = 0$

B C C C D D D D

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

You first see the positive example. First, compute π_1

$$\pi_1 = \Pr(y_1 = 1 \mid \boldsymbol{x}_1) = \frac{\exp \beta^T \boldsymbol{x}_i}{1 + \exp \beta^T \boldsymbol{x}_i} = \frac{\exp 0}{\exp 0 + 1} = 0.5$$

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$

A A A A B B B C

$y_2 = 0$

B C C C D D D D

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

$\pi_1 = 0.5$ What's the update for β_{bias}?

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$

A A A A B B B C

$y_2 = 0$

B C C C D D D D

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

What's the update for β_{bias}?

$\beta_{bias} = \beta'_{bias} + \eta \cdot (y_1 - \pi_1) \cdot x_{1,bias} = 0.0 + 1.0 \cdot (1.0 - 0.5) \cdot 1.0$

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$

A A A A B B B C

$y_2 = 0$

B C C C D D D D

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

What's the update for β_{bias}?

$\beta_{bias} = \beta'_{bias} + \eta \cdot (y_1 - \pi_1) \cdot x_{1,bias} = 0.0 + 1.0 \cdot (1.0 - 0.5) \cdot 1.0$ =0.5

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$

A A A A B B B C

$y_2 = 0$

B C C C D D D D

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

What's the update for β_A?

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$	$y_2 = 0$
A A A A B B B C	B C C C D D D D

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

What's the update for β_A? $\beta_A = \beta'_A + \eta \cdot (y_1 - \pi_1) \cdot x_{1,A} = 0.0 + 1.0 \cdot (1.0 - 0.5) \cdot 4.0$

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$	$y_2 = 0$
A A A A B B B C	B C C C D D D D

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

What's the update for β_A? $\beta_A = \beta'_A + \eta \cdot (y_1 - \pi_1) \cdot x_{1,A} = 0.0 + 1.0 \cdot (1.0 - 0.5) \cdot 4.0$
$=2.0$

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$

$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$

A A A A B B B C

$y_2 = 0$

B C C C D D D D

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

What's the update for β_B?

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$

A A A A B B B C

$y_2 = 0$

B C C C D D D D

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

What's the update for β_B? $\beta_B = \beta'_B + \eta \cdot (y_1 - \pi_1) \cdot x_{1,B} = 0.0 + 1.0 \cdot (1.0 - 0.5) \cdot 3.0$

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$

A A A A B B B C

$y_2 = 0$

B C C C D D D D

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

What's the update for β_B? $\beta_B = \beta'_B + \eta \cdot (y_1 - \pi_1) \cdot x_{1,B} = 0.0 + 1.0 \cdot (1.0 - 0.5) \cdot 3.0$
=1.5

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$

A A A A B B B C

$y_2 = 0$

B C C C D D D D

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

What's the update for β_C?

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$

A A A A B B B C

$y_2 = 0$

B C C C D D D D

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

What's the update for β_C?

$\beta_C = \beta'_C + \eta \cdot (y_1 - \pi_1) \cdot x_{1,C} = 0.0 + 1.0 \cdot (1.0 - 0.5) \cdot 1.0$

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$

A A A A B B B C

$y_2 = 0$

B C C C D D D D

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

What's the update for β_C?

$\beta_C = \beta'_C + \eta \cdot (y_1 - \pi_1) \cdot x_{1,C} = 0.0 + 1.0 \cdot (1.0 - 0.5) \cdot 1.0$ =0.5

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$

$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$

A A A A B B B C

$y_2 = 0$

B C C C D D D D

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

What's the update for β_D?

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$

A A A A B B B C

$y_2 = 0$

B C C C D D D D

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

What's the update for β_D?

$\beta_D = \beta'_D + \eta \cdot (y_1 - \pi_1) \cdot x_{1,D} = 0.0 + 1.0 \cdot (1.0 - 0.5) \cdot 0.0$

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

$y_1 = 1$

A A A A B B B C

$y_2 = 0$

B C C C D D D D

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

What's the update for β_D?

$\beta_D = \beta'_D + \eta \cdot (y_1 - \pi_1) \cdot x_{1,D} = 0.0 + 1.0 \cdot (1.0 - 0.5) \cdot 0.0$ =0.0

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$

A A A A B B B C

$y_2 = 0$

B C C C D D D D

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

Now you see the negative example. What's π_2?

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$

A A A A B B B C

$y_2 = 0$

B C C C D D D D

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

Now you see the negative example. What's π_2?

$\pi_2 = \Pr(y_2 = 1 \mid \vec{x_2}) = \frac{\exp \beta^T x_i}{1+\exp \beta^T x_i} = \frac{\exp\{.5+1.5+1.5+0\}}{\exp\{.5+1.5+1.5+0\}+1} =$

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 =$1

A A A A B B B C

$y_2 =$0

B C C C D D D D

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

Now you see the negative example. What's π_2?

$$\pi_2 = \Pr(y_2 = 1 \mid \vec{x_2}) = \frac{\exp \beta^T x_i}{1+\exp \beta^T x_i} = \frac{\exp\{.5+1.5+1.5+0\}}{\exp\{.5+1.5+1.5+0\}+1} = 0.97$$

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$	$y_2 = 0$
A A A A B B B C	B C C C D D D D

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

Now you see the negative example. What's π_2?
$\pi_2 = 0.97$
What's the update for β_{bias}?

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$	$y_2 = 0$
A A A A B B B C	B C C C D D D D

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

What's the update for β_{bias}?

$\beta_{bias} = \beta'_{bias} + \eta \cdot (y_2 - \pi_2) \cdot x_{2,bias} = 0.5 + 1.0 \cdot (0.0 - 0.97) \cdot 1.0$

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$	$y_2 = 0$
A A A A B B B C	B C C C D D D D

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

What's the update for β_{bias}?

$\beta_{bias} = \beta'_{bias} + \eta \cdot (y_2 - \pi_2) \cdot x_{2,bias} = 0.5 + 1.0 \cdot (0.0 - 0.97) \cdot 1.0$ =-0.47

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$

A A A A B B B C

$y_2 = 0$

B C C C D D D D

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

What's the update for β_A?

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$	$y_2 = 0$
A A A A B B B C	B C C C D D D D

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

What's the update for β_A?

$\beta_A = \beta'_A + \eta \cdot (y_2 - \pi_2) \cdot x_{2,A} = 2.0 + 1.0 \cdot (0.0 - 0.97) \cdot 0.0$

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$

A A A A B B B C

$y_2 = 0$

B C C C D D D D

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

What's the update for β_A?

$\beta_A = \beta'_A + \eta \cdot (y_2 - \pi_2) \cdot x_{2,A} = 2.0 + 1.0 \cdot (0.0 - 0.97) \cdot 0.0$ =2.0

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$

A A A A B B B C

$y_2 = 0$

B C C C D D D D

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

What's the update for β_B?

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$

A A A A B B B C

$y_2 = 0$

B C C C D D D D

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

What's the update for β_B?

$$\beta_B = \beta'_B + \eta \cdot (y_2 - \pi_2) \cdot x_{2,B} = 1.5 + 1.0 \cdot (0.0 - 0.97) \cdot 1.0$$

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$	$y_2 = 0$
A A A A B B B C	B C C C D D D D

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

What's the update for β_B?

$\beta_B = \beta'_B + \eta \cdot (y_2 - \pi_2) \cdot x_{2,B} = 1.5 + 1.0 \cdot (0.0 - 0.97) \cdot 1.0$ =0.53

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$

A A A A B B B C

$y_2 = 0$

B C C C D D D D

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

What's the update for β_C?

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$

A A A A B B B C

$y_2 = 0$

B C C C D D D D

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

What's the update for β_C?

$\beta_C = \beta'_C + \eta \cdot (y_2 - \pi_2) \cdot x_{2,C} = 0.5 + 1.0 \cdot (0.0 - 0.97) \cdot 3.0$

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$	$y_2 = 0$
A A A A B B B C	B C C C D D D D

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

What's the update for β_C?

$\beta_C = \beta'_C + \eta \cdot (y_2 - \pi_2) \cdot x_{2,C} = 0.5 + 1.0 \cdot (0.0 - 0.97) \cdot 3.0$ =-2.41

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$

A A A A B B B C

$y_2 = 0$

B C C C D D D D

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

What's the update for β_D?

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$

A A A A B B B C

$y_2 = 0$

B C C C D D D D

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

What's the update for β_D?

$\beta_D = \beta'_D + \eta \cdot (y_2 - \pi_2) \cdot x_{2,D} = 0.0 + 1.0 \cdot (0.0 - 0.97) \cdot 4.0$

Example Documents

$$\beta_j = \beta_j + \eta(y_i - \pi_i)x_{ij}$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

$y_1 = 1$

A A A A B B B C

$y_2 = 0$

B C C C D D D D

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

What's the update for β_D?

$\beta_D = \beta'_D + \eta \cdot (y_2 - \pi_2) \cdot x_{2,D} = 0.0 + 1.0 \cdot (0.0 - 0.97) \cdot 4.0$ =-3.88

Algorithm

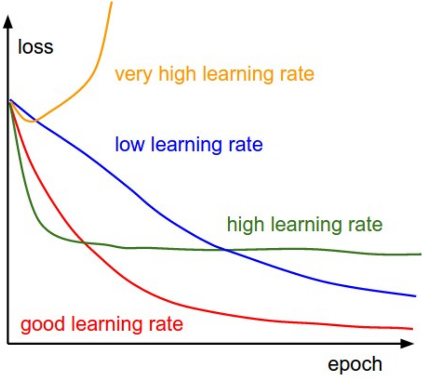
1. Initialize a vector β to be all zeros
2. For $t = 1, \dots, T$
   - For each example $\boldsymbol{x}_i, y_i$ and feature j:
     - Compute $\pi_i \equiv \text{Pr}(y_i = 1 \mid \boldsymbol{x}_i)$
     - Set $\beta_j = \beta_j' - \eta(\lambda\beta_j' - (y_i - \pi_i)\boldsymbol{x}_i)$
3. Output the parameters $\beta_1, \dots, \beta_d$.

Algorithm

1. Initialize a vector β to be all zeros
2. For $t = 1, \dots, T$
   - For each example $\boldsymbol{x}_i, y_i$ and feature j:
     - Compute $\pi_i \equiv \Pr(y_i = 1 \mid \boldsymbol{x}_i)$
     - Set $\beta_j = \beta_j' - \eta(\lambda \beta_j' - (y_i - \pi_i)\boldsymbol{x}_i)$
3. Output the parameters $\beta_1, \dots, \beta_d$.

How to decide η?

Choosing learning rate



http://cs231n.github.io/neural-networks-3/

Learning rate decay

- Decay after each epoch (e.g., $\frac{\eta_0}{t^2}$, $\eta_0 e^{-kt}$)
- Decay after each example (e.g., $\frac{\eta_0}{1+kn}$)

Decay schedule can be seen as a hyperparameter too.

Learning rate decay

- Decay after each epoch (e.g., $\frac{\eta_0}{t^2}$, $\eta_0 e^{-kt}$)
- Decay after each example (e.g., $\frac{\eta_0}{1+kn}$)

Decay schedule can be seen as a hyperparameter too.

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