University of Colorado **Boulder**

Department of Computer Science

CSCI 5622: Machine Learning

Chenhao Tan

Lecture 8: Stochastic Gradient Descent

Slides adapted from Chris Ketelsen, Jordan Boyd-Graber, and Noah Smith

Administrivia

- HW1 grading is done
  - Common issues

Learning Objectives

- Understand gradient descent
- Understand structural risk minimization
- Understand stochastic gradient descent

Outline

- Objective function
- Gradient descent
- Structural risk minimization
- Stochastic gradient descent

Outline

- Objective function
- Gradient descent
- Structural risk minimization
- Stochastic gradient descent

Reminder: Logistic Regression

$$P(Y = 0 \mid \boldsymbol{x}) = \frac{1}{1 + \exp\left[\beta_0 + \sum_j \beta_j \boldsymbol{x}_j\right]} \tag{1}$$

$$P(Y = 1 \mid \boldsymbol{x}) = \frac{\exp\left[\beta_0 + \sum_j \beta_j \boldsymbol{x}_j\right]}{1 + \exp\left[\beta_0 + \sum_j \beta_j \boldsymbol{x}_j\right]} \tag{2}$$

- Discriminative prediction: $P(y \mid \boldsymbol{x})$
- Classification uses: sentiment analysis, spam detection
- What we didn't talk about is how to learn β from data

Logistic Regression: Objective Function

One idea: find the parameter that maximize the likelihood of observing the training data.

Logistic Regression: Objective Function

One idea: find the parameter that maximize the likelihood of observing the training data.
Maximize likelihood

$$\text{Obj} \equiv \log P(Y \mid X, \beta) = \sum_i \log P(y^{(i)} \mid \boldsymbol{x}^{(i)}, \beta)$$

$$= \sum_i y^{(i)} \left(\beta_0 + \sum_j \beta_j \boldsymbol{x}_j^{(i)} \right) - \log \left[1 + \exp \left(\beta_0 + \sum_j \beta_j \boldsymbol{x}_j^{(i)} \right) \right]$$

Logistic Regression: Objective Function

Minimize negative log likelihood (loss)

$$\mathscr{L} \equiv -\log P(Y \mid X, \beta) = -\sum_i \log P(y^{(i)} \mid \boldsymbol{x}^{(i)}, \beta))$$

$$= \sum_i -y^{(i)} \left(\beta_0 + \sum_j \beta_i \boldsymbol{x}_j^{(i)} \right) + \log \left[1 + \exp \left(\beta_0 + \sum_j \beta_i \boldsymbol{x}_j^{(i)} \right) \right]$$

Logistic Regression: Objective Function

Minimize negative log likelihood (loss)

$$\mathscr{L} \equiv -\log P(Y \mid X, \beta) = -\sum_i \log P(y^{(i)} \mid \boldsymbol{x}^{(i)}, \beta))$$

$$= \sum_i -y^{(i)} \left(\beta_0 + \sum_j \beta_i \boldsymbol{x}_j^{(i)} \right) + \log \left[1 + \exp \left(\beta_0 + \sum_j \beta_i \boldsymbol{x}_j^{(i)} \right) \right]$$

Training data $\{(\boldsymbol{x}, y)\}$ are fixed. Objective function is a function of β . . . what values of β give a good value?

Logistic Regression: Objective Function

Minimize negative log likelihood (loss)

$$\mathscr{L} \equiv -\log P(Y \mid X, \beta) = -\sum_i \log P(y^{(i)} \mid \boldsymbol{x}^{(i)}, \beta))$$

$$= \sum_i -y^{(i)} \left(\beta_0 + \sum_j \beta_i \boldsymbol{x}_j^{(i)} \right) + \log \left[1 + \exp \left(\beta_0 + \sum_j \beta_i \boldsymbol{x}_j^{(i)} \right) \right]$$

Training data $\{(\boldsymbol{x}, y)\}$ are fixed. Objective function is a function of β . . . what values of β give a good value?

$$\beta^* = \arg\min_\beta \mathscr{L}(\beta)$$

Convexity

$\mathscr{L}(\beta)$ is convex for logistic regression.

Proof.

- Logistic loss $-yv + \log(1 + \exp(v))$ is convex.
- Composition with linear function maintains convexity.
- Sum of convex functions is convex.

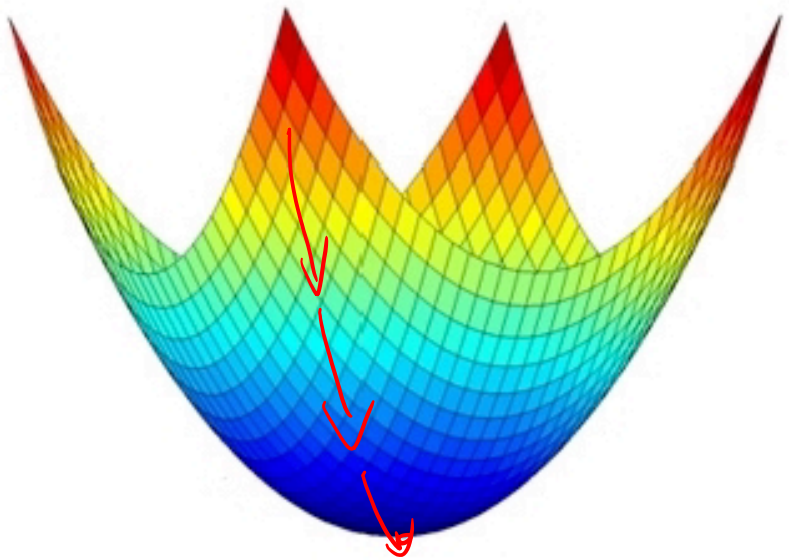
$\frac{\partial^2 f}{\partial v^2} \geq 0$

□

Outline

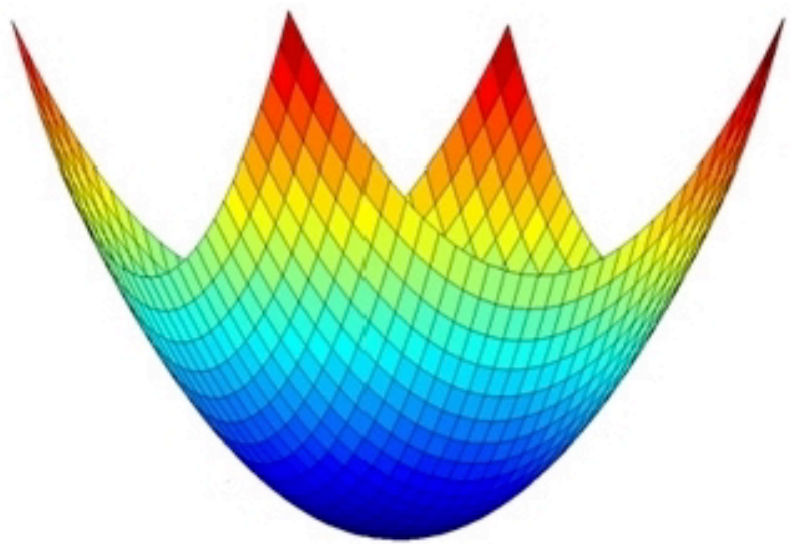
- Objective function
- Gradient descent
- Structural risk minimization
- Stochastic gradient descent

Convexity



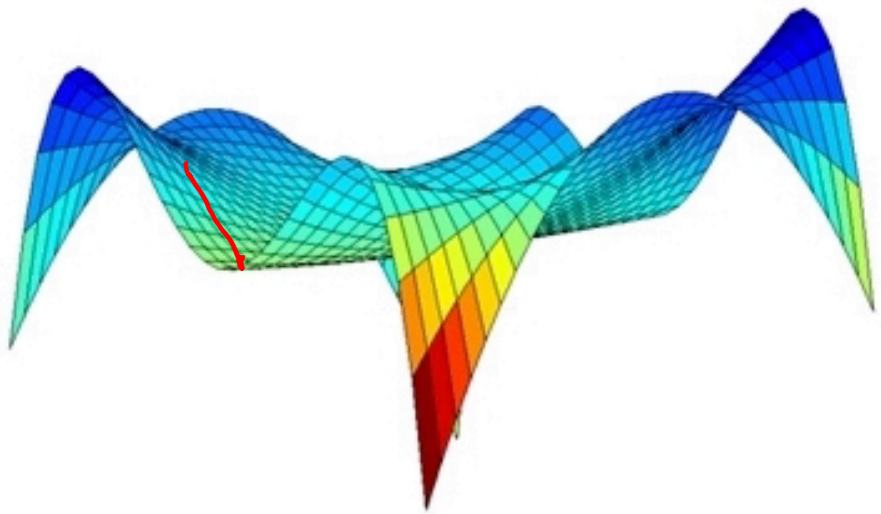
- Convex function
- Doesn't matter where you start, if you go down along the gradient

Convexity



- Convex function
- Doesn't matter where you start, if you go down along the gradient
- Gradient!

Convexity

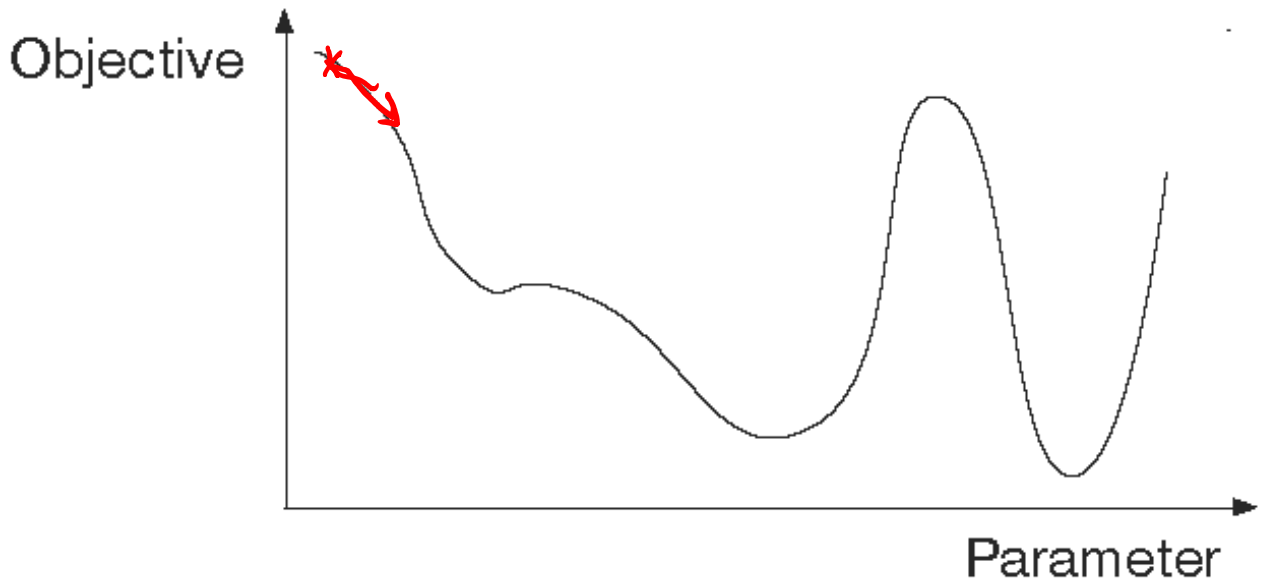


- It would have been much harder if this is not convex.

Gradient Descent (non-convex)

Goal

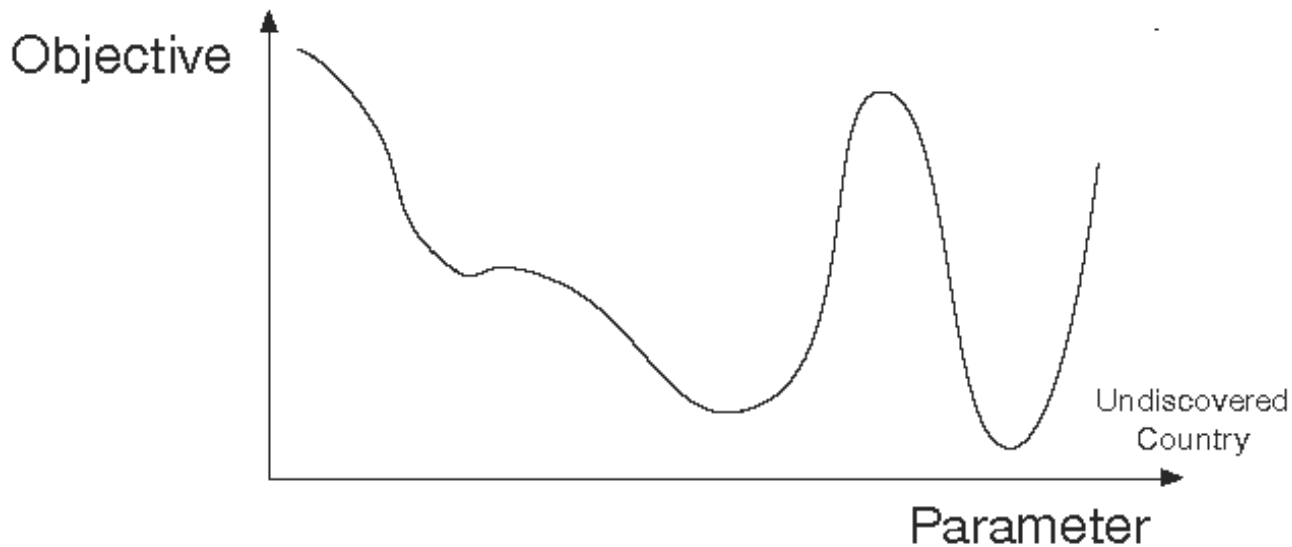
Optimize loss function with respect to variables β



Gradient Descent (non-convex)

Goal

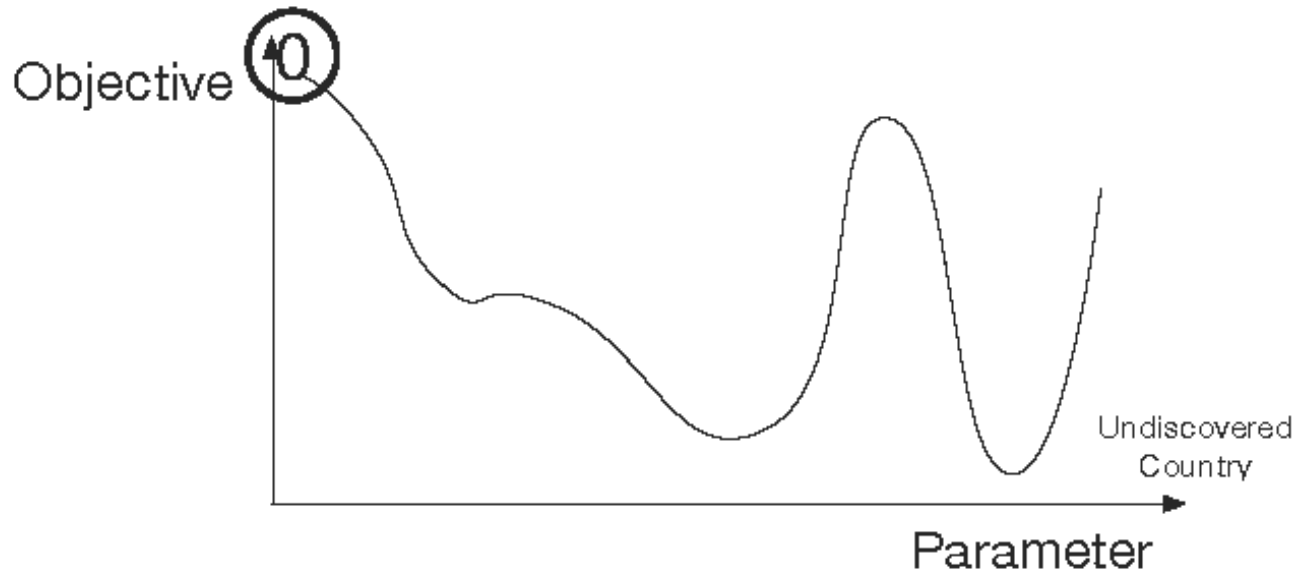
Optimize loss function with respect to variables β



Gradient Descent (non-convex)

Goal

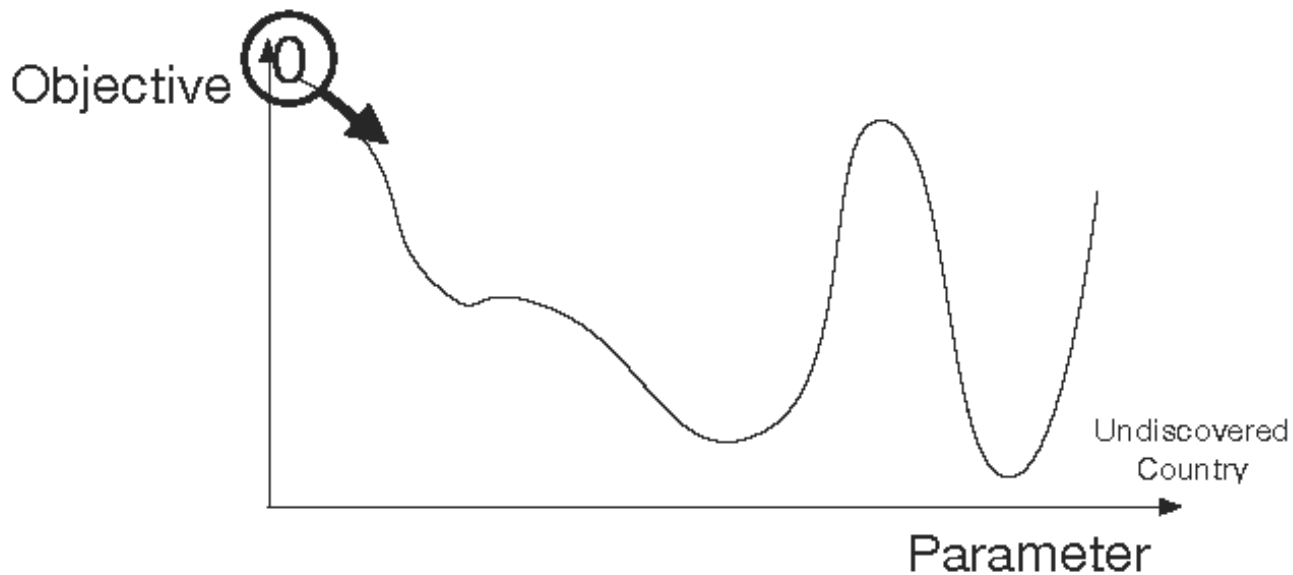
Optimize loss function with respect to variables β



Gradient Descent (non-convex)

Goal

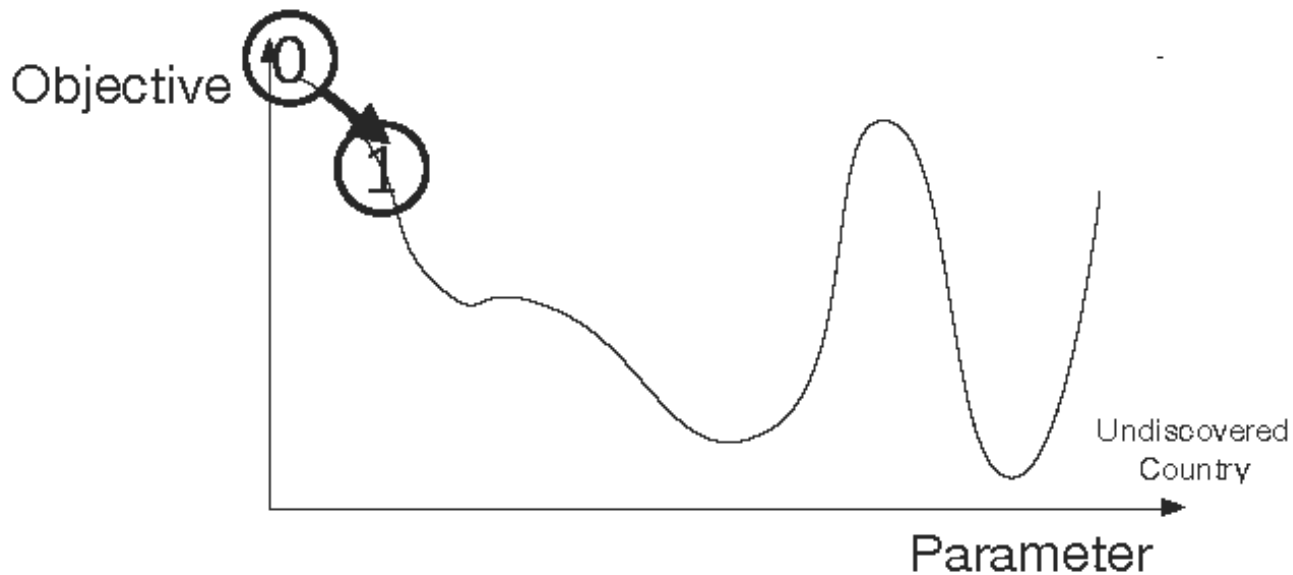
Optimize loss function with respect to variables β



Gradient Descent (non-convex)

Goal

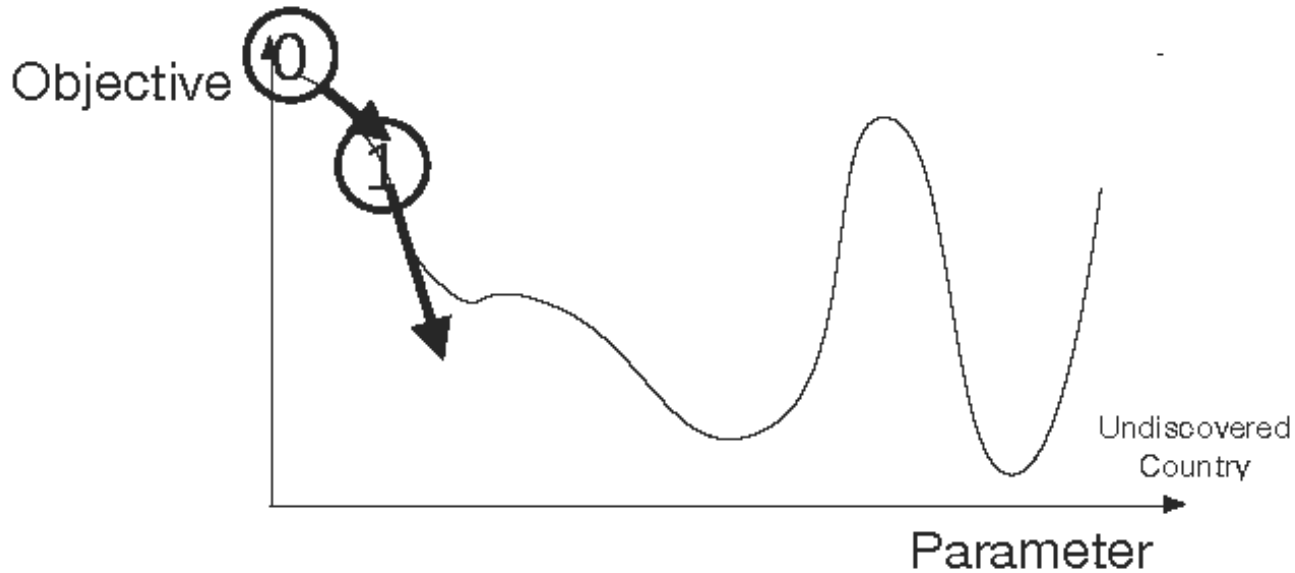
Optimize loss function with respect to variables β



Gradient Descent (non-convex)

Goal

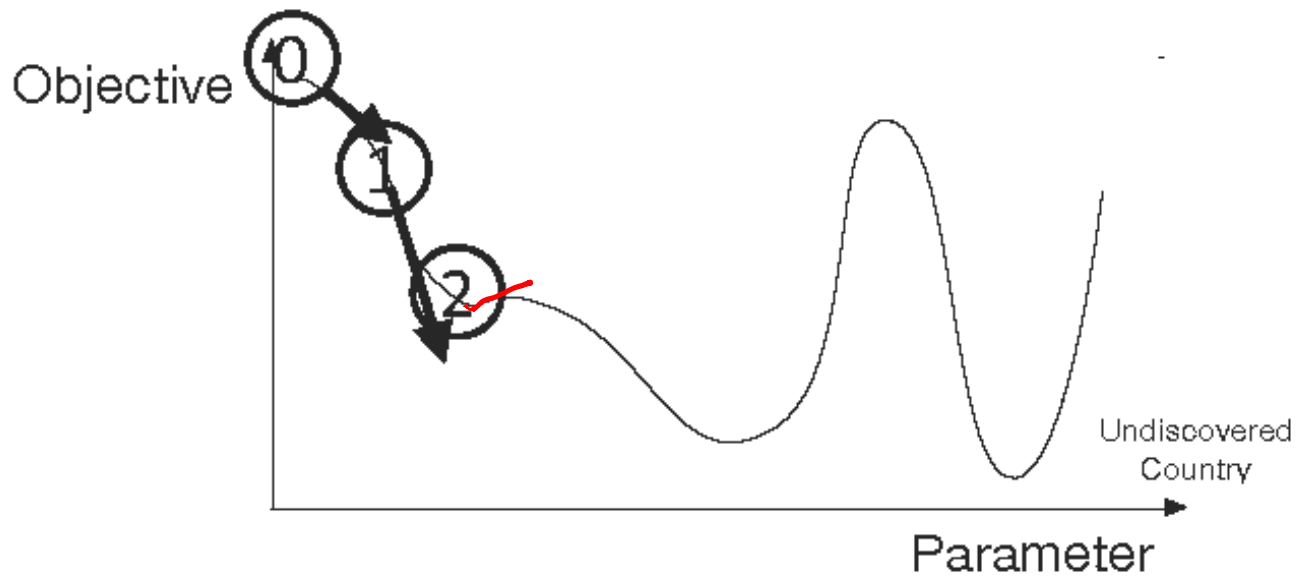
Optimize loss function with respect to variables β



Gradient Descent (non-convex)

Goal

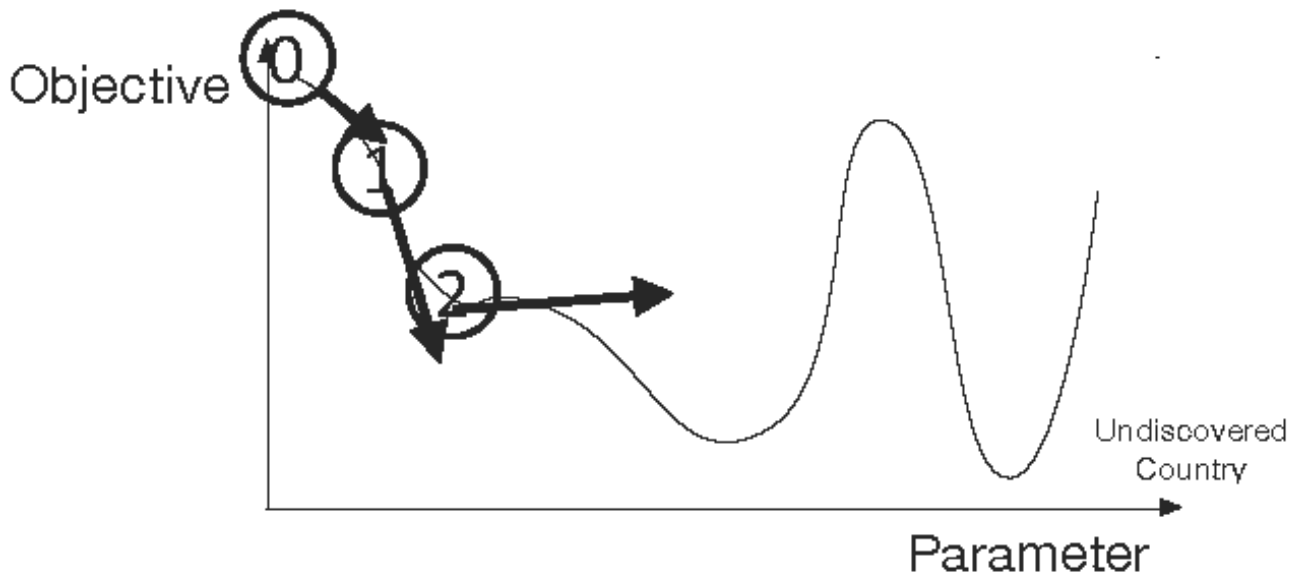
Optimize loss function with respect to variables β



Gradient Descent (non-convex)

Goal

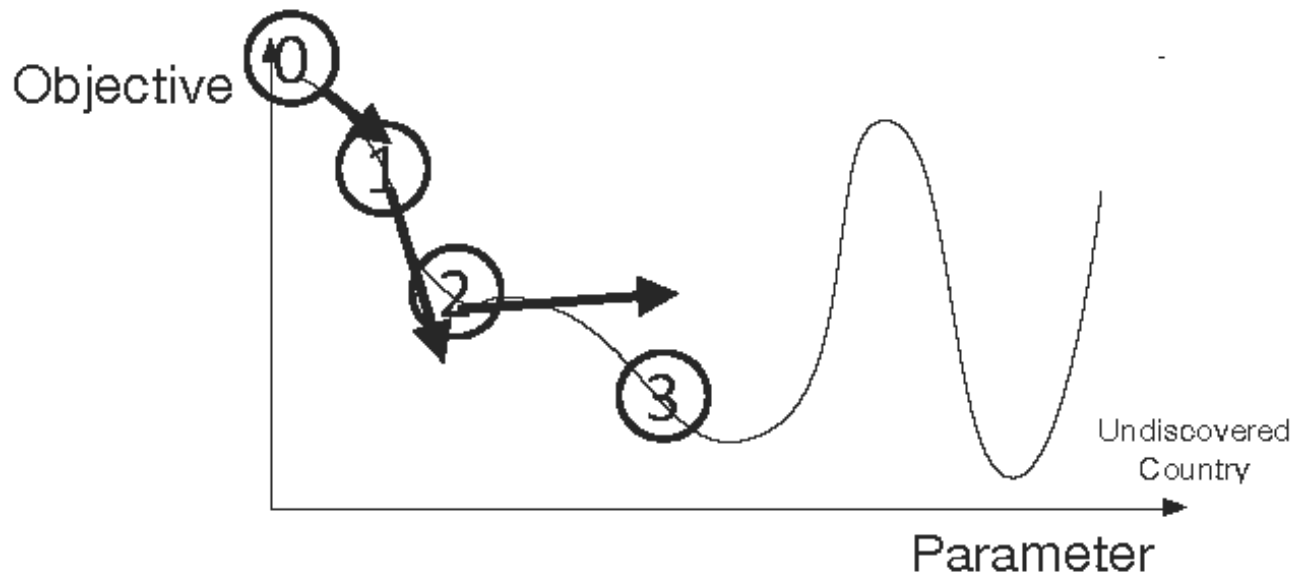
Optimize loss function with respect to variables β



Gradient Descent (non-convex)

Goal

Optimize loss function with respect to variables β



Gradient Descent (non-convex)

Goal

Optimize loss function with respect to variables β

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

Gradient Descent (non-convex)

Goal

Optimize loss function with respect to variables β

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

Luckily, (vanilla) logistic regression is convex

Gradient for Logistic Regression

To ease notation, let's define

$$\pi_i = \frac{\exp \beta^T x_i}{1 + \exp \beta^T x_i} \tag{3}$$

Our objective function is

$$\mathscr{L} = -\sum_i \log p(y_i \,|\, x_i) = \sum_i \mathscr{L}_i = \sum_i \begin{cases} -\log \pi_i & \text{if } y_i = 1 \\ -\log(1 - \pi_i) & \text{if } y_i = 0 \end{cases} \tag{4}$$

$$\frac{\partial \pi_i}{\partial \beta} = \frac{\partial \left(1 - \frac{1}{1+\exp(\beta^T x_i)}\right)}{\partial \beta} = -\cdot\left(-\frac{1}{(1+\exp(\beta^T x))^2}\right)\cdot\frac{\partial(1+\exp(\beta^T x))}{\partial \beta}$$

$$= \frac{1}{(1+\exp(\beta^T x))^2}\cdot \exp(\beta^T x)\cdot x$$

$$= \pi_i(1-\pi_i)\cdot x$$

$$\frac{\partial \mathscr{L}}{\partial \pi_i} = \begin{cases} -\frac{1}{\pi_i} & \text{if } y_i = 1 \\ \frac{1}{1-\pi_i} & y_i = 0 \end{cases}$$

$$\frac{\partial \mathscr{L}}{\partial \beta} = \frac{\partial \mathscr{L}}{\partial \pi_i}\frac{\partial \pi_i}{\partial \beta} = \begin{cases} -(1-\pi_i)x & \text{if } y_i = 1 \\ \pi_i \cdot x & \text{if } y_i = 0 \end{cases}$$

$$= -(y_i - \pi_i)x$$

Taking the Derivative

Apply chain rule:

$$\frac{\partial \mathscr{L}}{\partial \beta_j} = \sum_i \frac{\partial \mathscr{L}_i(\vec{\beta})}{\partial \beta_j} = \sum_i \begin{cases} -\frac{1}{\pi_i}\frac{\partial \pi_i}{\partial \beta_j} & \text{if } y_i = 1 \\ -\frac{1}{1-\pi_i}\left(-\frac{\partial \pi_i}{\partial \beta_j}\right) & \text{if } y_i = 0 \end{cases} \tag{5}$$

If we plug in the derivative,

$$\frac{\partial \pi_i}{\partial \beta_j} = \pi_i(1-\pi_i)x_j, \tag{6}$$

we can merge these two cases

$$\frac{\partial \mathscr{L}_i}{\partial \beta_j} = -(y_i - \pi_i)x_j. \tag{7}$$

Gradient for Logistic Regression

Gradient

$$\nabla_\beta \mathscr{L}(\vec{\beta}) = \left[\frac{\partial \mathscr{L}(\vec{\beta})}{\partial \beta_0}, \ldots, \frac{\partial \mathscr{L}(\vec{\beta})}{\partial \beta_n}\right] \tag{8}$$

Update

$$\Delta\beta \equiv \eta \nabla_\beta \mathscr{L}(\vec{\beta}) \tag{9}$$

$$\beta_i' \leftarrow \beta_i - \eta \frac{\partial \mathscr{L}(\vec{\beta})}{\partial \beta_i} \tag{10}$$

Gradient for Logistic Regression

Gradient

$$\nabla_\beta \mathscr{L}(\vec{\beta}) = \left[\frac{\partial \mathscr{L}(\vec{\beta})}{\partial \beta_0}, \ldots, \frac{\partial \mathscr{L}(\vec{\beta})}{\partial \beta_n}\right] \tag{8}$$

Update

$$\Delta\beta \equiv \eta \nabla_\beta \mathscr{L}(\vec{\beta}) \tag{9}$$

$$\beta_i' \leftarrow \beta_i - \eta \frac{\partial \mathscr{L}(\vec{\beta})}{\partial \beta_i} \tag{10}$$

η: step size, must be greater than zero

Gradient for Logistic Regression

Gradient

$$\nabla_\beta \mathscr{L}(\vec{\beta}) = \left[\frac{\partial \mathscr{L}(\vec{\beta})}{\partial \beta_0}, \ldots, \frac{\partial \mathscr{L}(\vec{\beta})}{\partial \beta_n}\right] \tag{8}$$

Update

$$\Delta\beta \equiv \eta \nabla_\beta \mathscr{L}(\vec{\beta}) \tag{9}$$

$$\beta_i' \leftarrow \beta_i - \eta \frac{\partial \mathscr{L}(\vec{\beta})}{\partial \beta_i} \tag{10}$$

Overfitting

- It is not ideal to maximize the likelihood of training data

Overfitting

- It is not ideal to maximize the likelihood of training data
  - When to stop?
  - Simple models (avoid β to get too big)

Overfitting

- It is not ideal to maximize the likelihood of training data
  - When to stop?
  - Simple models (avoid β to get too big)
    Regularization

$\Sigma \beta^2$

Outline

- Objective function
- Gradient descent
- Structural risk minimization
- Stochastic gradient descent

Regularized Conditional Log Likelihood

Unregularized

$$\beta^* = \arg\min_{\beta} -\ln \left[p(y^{(j)} \mid x^{(j)}, \beta) \right] \tag{11}$$

Regularized

$$\beta^* = \arg\min_{\beta} -\ln \left[p(y^{(j)} \mid x^{(j)}, \beta) \right] + \frac{1}{2}\lambda \sum_i \beta_i^2 \tag{12}$$

Regularized Conditional Log Likelihood

Unregularized

$$\beta^* = \arg\min_{\beta} -\ln \left[p(y^{(j)} \mid x^{(j)}, \beta) \right] \tag{11}$$

Regularized

$$\beta^* = \arg\min_{\beta} -\ln \left[p(y^{(j)} \mid x^{(j)}, \beta) \right] + \frac{1}{2}\lambda \sum_i \beta_i^2 \tag{12}$$

λ is the "regularization" parameter (a hyperparameter) that trades off between likelihood and having small parameters

Alternative view of regularization

Can also get to regularization by putting prior beliefs on parameters

$$p(\beta \mid \mathcal{D}) \propto p(\mathcal{D} \mid \beta) p(\beta)$$

Then MAP estimate for β is $\hat{\beta}$ which maximizes posterior

Alternative view of regularization

Can also get to regularization by putting prior beliefs on parameters

$$p(\beta \mid \mathcal{D}) \propto p(\mathcal{D} \mid \beta) p(\beta)$$

Then MAP estimate for β is $\hat{\beta}$ which maximizes posterior

Ridge: Assume Gaussian prior $p(\beta_j) = \mathcal{N}(\beta_j \mid 0, \tau^2)$, we will obtain the same regularized objective function
You can learn more about this view in “Bayesian statistics”

Risk minimization

$$\min_{\beta} \sum_{i} \ell(y^{(i)}, h_{\beta}(\boldsymbol{x}^{(i)})) + \lambda R(\beta)$$

Risk minimization

$$\min_{\beta} \sum_{i} \ell(y^{(i)}, h_{\beta}(\boldsymbol{x}^{(i)})) + \lambda R(\beta)$$

Loss functions (ℓ)

Describe how well the model fits the training data

- $-y\hat{y} + log(1 + exp(\hat{y}))$

Regularization (R)

Control the complexity of the model

- $||\beta||^2 = \sum_j \beta_j^2$

Risk minimization

$$\min_{\beta} \sum_{i} \ell(y^{(i)}, h_{\beta}(\boldsymbol{x}^{(i)})) + \lambda R(\beta)$$

Loss functions (ℓ)

Describe how well the model fits the training data

- $-y\hat{y} + log(1 + exp(\hat{y}))$
- $(y - \hat{y})^2$
- $\max\{0, 1 - y\hat{y}\}$

Regularization (R)

Control the complexity of the model

- $||\beta||^2 = \sum_j \beta_j^2$
- $||\beta||_p = \left(\sum_j |\beta_j|^p\right)^{\frac{1}{p}}$
  - ℓ_1–regularization: $\sum_j |\beta_j|$

Outline

- Objective function
- Gradient descent
- Structual risk minimization
- Stochastic gradient descent

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

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

$$\vec{\beta} = \langle \beta_{bias} = 0, \beta_A = 0, \beta_B = 0, \beta_C = 0, \beta_D = 0 \rangle$$

$\frac{\exp(\beta^T x)}{1+\exp(\beta^T x)}$

$y_1 = 1$

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 $\pi_1 = 0.5$

$\beta_{bias} = 0.5$

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 $\eta = 1.0$.)

You first see the positive example. First, compute π_1

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

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

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

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

A:4
B:3
C:1

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

$y_2 = 0$

B C C C D D D D

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

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

$y_1 = 1$

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

Example Documents

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

$y_1 = 1$

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

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

$y_1 = 1$

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?

$\pi_2 = 0.97$

What's the update for β_{bias}?

Example Documents

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

$y_1 = 1$

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}?

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

Example Documents

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

$y_1 = 1$

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}?

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

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

$y_1 = 1$

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?

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

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

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

$y_1 = 1$

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?

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

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

$y_1 = 1$

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?

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

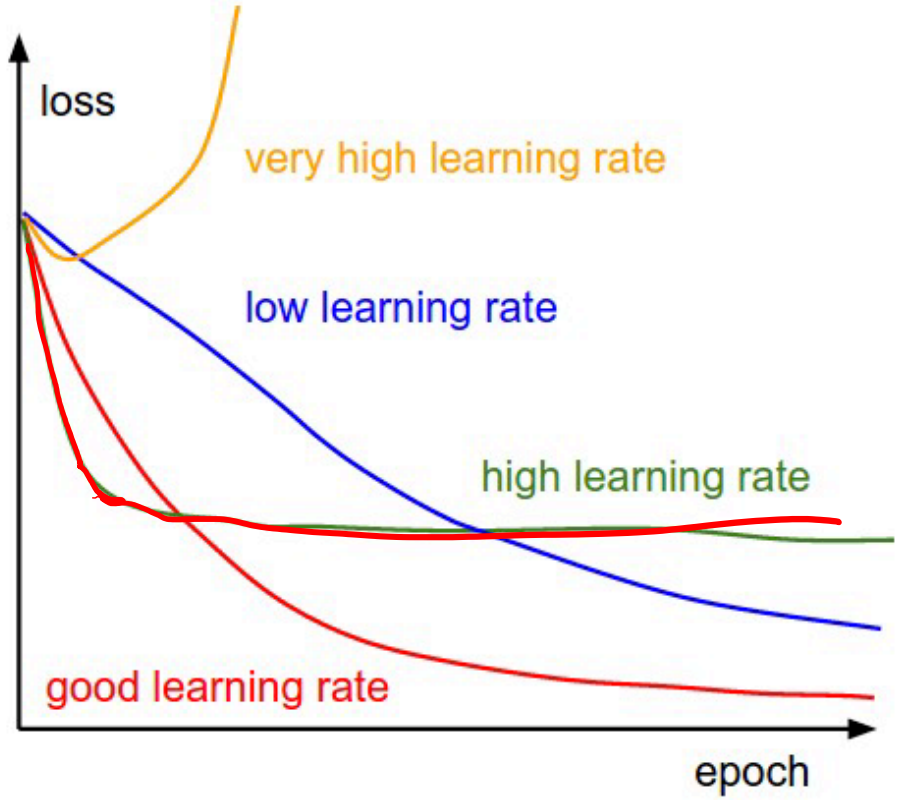
1. Initialize a vector β to be all zeros
2. For $t = 1, \ldots, T$
   - For each example $\boldsymbol{x}_i, y_i$ and feature j:
     - Compute $\pi_i \equiv \Pr(y_i = 1 \mid \boldsymbol{x}_i)$
     - Set $\beta_j = \beta'_j - \eta(\lambda\beta'_j - (y_i - \pi_i)\boldsymbol{x}_i)$
3. Output the parameters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

Algorithm

1. Initialize a vector β to be all zeros
2. For $t = 1, \ldots, T$
   - 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 $\beta_1, \ldots, \beta_d$.

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/`

Adam

Recap

- Follow the gradient to fit the logistic regression model
- Most machine learning methods fall into the framework of (loss + regularization)
- Stochastic gradient descent allows for approximating the gradient from one instance