



Classification: Logistic Regression from Data

Machine Learning: Jordan Boyd-Graber
University of Colorado Boulder

LECTURE 3

Slides adapted from Emily Fox

Content Questions

Content Questions

Content Questions

Content Questions

Content Questions

Administrivia Questions

Administrivia Questions

Administrivia Questions

Reminder: Logistic Regression

$$P(Y = 0|X) = \frac{1}{1 + \exp [\beta_0 + \sum_i \beta_i X_i]} \quad (1)$$

$$P(Y = 1|X) = \frac{\exp [\beta_0 + \sum_i \beta_i X_i]}{1 + \exp [\beta_0 + \sum_i \beta_i X_i]} \quad (2)$$

- Discriminative prediction: $p(y|x)$
- Classification uses: ad placement, spam detection
- What we didn't talk about is how to learn β from data

Logistic Regression: Objective Function

$$\mathcal{L} \equiv \ln p(Y|X, \beta) = \sum_j \ln p(y^{(j)} | x^{(j)}, \beta) \quad (3)$$

$$= \sum_j y^{(j)} \left(\beta_0 + \sum_i \beta_i x_i^{(j)} \right) - \ln \left[1 + \exp \left(\beta_0 + \sum_i \beta_i x_i^{(j)} \right) \right] \quad (4)$$

Algorithm

- ① Initialize a vector B to be all zeros
- ② For $t = 1, \dots, T$
 - For each example \vec{x}_i, y_i and feature j :
 - Compute $p \equiv \Pr(y = 1 \mid \vec{x}_i)$
 - Set $\beta[j] = \beta[j] + \lambda(y - p)x_i$
- ③ Output the parameters β_1, \dots, β_d .

Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i \quad (5)$$

y=1

A A A A B B B C

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

y=0

B C C C D D D D

You first see the positive example. What's the update for β_0 ?

Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i \quad (5)$$

y=1

A A A A B B B C

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

y=0

B C C C D D D D

You first see the positive example. What's the update for β_0 ?

$$\beta_0 = 0 + 1.0 * (1.0 - .5)1.0 = 0.5$$

Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i \quad (5)$$

y=1

A A A A B B B C

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

y=0

B C C C D D D D

What's the update for β_A ?

Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i \quad (5)$$

y=1

A A A A B B B C

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

y=0

B C C C D D D D

What's the update for β_A ?

$$\beta_A = 0 + 1.0 * (1.0 - .5)4.0 = 2.0$$

Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i \quad (5)$$

y=1

A A A A B B B C

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

y=0

B C C C D D D D

What's the update for β_B ?

Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i \quad (5)$$

y=1

A A A A B B B C

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

y=0

B C C C D D D D

What's the update for β_B ?

$$\beta_B = 0 + 1.0 * (1.0 - .5)3.0 = 1.5$$

Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i \quad (5)$$

y=1

A A A A B B B C

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

y=0

B C C C D D D D

What's the update for β_C ?

Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i \quad (5)$$

y=1

A A A A B B B C

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

y=0

B C C C D D D D

What's the update for β_C ?

$$\beta_C = 0 + 1.0 * (1.0 - .5)1.0 = 0.5$$

Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i \quad (5)$$

y=1

A A A A B B B C

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

y=0

B C C C D D D D

What's the update for β_D ?

Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i \quad (5)$$

y=1

A A A A B B B C

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

y=0

B C C C D D D D

What's the update for β_D ?

$$\beta_D = 0 + 1.0 * (1.0 - .5)0.0 = 0.0$$

Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i \quad (5)$$

y=1

A A A A B B B C

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

y=0

B C C C D D D D

Now you see the negative example. What's the update for β_0 ?

Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i \quad (5)$$

y=1

A A A A B B B C

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

y=0

B C C C D D D D

Now you see the negative example. What's the update for β_0 ?
What's the activation?

Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i \quad (5)$$

y=1

A A A A B B B C

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

y=0

B C C C D D D D

Now you see the negative example. What's the update for β_0 ?

$$0.5 + 1.5 + 1.5 + 0 = 0.97$$

Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i \quad (5)$$

y=1

A A A A B B B C

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

y=0

B C C C D D D D

Now you see the negative example. What's the update for β_0 ?

$$\beta_0 = 0.5 + 1.0 * (0.0 - 0.97) = -0.47$$

Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i \quad (5)$$

y=1

A A A A B B B C

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

y=0

B C C C D D D D

What's the update for β_A ?

Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i \quad (5)$$

y=1

A A A A B B B C

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

y=0

B C C C D D D D

What's the update for β_A ?

$$\beta_A = 2.0 + 1.0 * (0.0 - 0.97)0.0 = 2.0$$

Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i \quad (5)$$

y=1

A A A A B B B C

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

y=0

B C C C D D D D

What's the update for β_B ?

Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i \quad (5)$$

y=1

A A A A B B B C

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

y=0

B C C C D D D D

What's the update for β_B ?

$$\beta_B = 1.5 + 1.0 * (0.0 - 0.97)1.0 = 0.53$$

Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i \quad (5)$$

y=1

A A A A B B B C

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

y=0

B C C C D D D D

What's the update for β_C ?

Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i \quad (5)$$

y=1

A A A A B B B C

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

y=0

B C C C D D D D

What's the update for β_C ?

$$\beta_C = 0.5 + 1.0 * (0.0 - 0.97)3.0 = -2.41$$

Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i \quad (5)$$

y=1

A A A A B B B C

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

y=0

B C C C D D D D

What's the update for β_D ?

Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i \quad (5)$$

y=1

A A A A B B B C

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

y=0

B C C C D D D D

What's the update for β_D ?

Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i \quad (5)$$

y=1

A A A A B B B C

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

y=0

B C C C D D D D

$$\beta_D = 0.0 + 1.0 * (0.0 - 0.97)4.0 = -3.88$$

How does the gradient change with regularization?

- You can do your normal update

How does the gradient change with regularization?

- You can do your normal update
- Then

$$\beta_j = \beta_j - \lambda 2\mu \beta_j = \beta_j \cdot (1 - 2\lambda\mu) \quad (6)$$

How does the gradient change with regularization?

- You can do your normal update
- Then

$$\beta_j = \beta_j - \lambda 2\mu \beta_j = \beta_j \cdot (1 - 2\lambda\mu) \quad (6)$$

- Doesn't depend on X or Y . Just makes all your weights smaller

How does the gradient change with regularization?

- You can do your normal update
- Then

$$\beta_j = \beta_j - \lambda 2\mu \beta_j = \beta_j \cdot (1 - 2\lambda\mu) \quad (6)$$

- Doesn't depend on X or Y . Just makes all your weights smaller
- But difficult to update every feature every time

How does the gradient change with regularization?

- You can do your normal update
- Then

$$\beta_j = \beta_j - \lambda 2\mu \beta_j = \beta_j \cdot (1 - 2\lambda\mu) \quad (6)$$

- Doesn't depend on X or Y . Just makes all your weights smaller
- But difficult to update every feature every time
- Following this up, we note that we can perform m successive "regularization" updates by letting $B_j = B_j \cdot (1 - 2\lambda\mu)^m$. The basic idea of the new algorithm is to not perform regularization updates for zero-valued x_j 's, but instead to simply keep track of how many such updates would need to be performed to update β_j