Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

# Classification: Logistic Regression from Data

Machine Learning: Jordan Boyd-Graber
University of Colorado Boulder
LECTURE 3

Slides adapted from William Cohen

## Content Questions

## Content Questions

## Content Questions

# Content Questions

## Content Questions

## Administrivia Questions

## Administrivia Questions

## Administrivia Questions

## Reminder: Logistic Regression

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

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

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

## Logistic Regression: Objective Function

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

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

## 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 $\beta_1, \dots, \beta_d$.

## Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

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

You first see the positive example. What's the update for $\beta_0$?

## Example Documents

$$\beta[j] = \beta[j] + \lambda(y-p)x_i$$
$$\vec{\beta} = \langle 0,0,0,0,0 \rangle$$

| y=1 | y=0 |
|---|---|
| A A A A B B B C | B C C C D D D D |

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

You first see the positive example. What's the update for $\beta_0$?
$\beta_0 = 0 + 1.0 * (1.0 - .5)1.0 = 0.5$

## Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

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

What's the update for $\beta_A$?

## Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

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

What's the update for $\beta_A$?

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

## Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

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

What's the update for $\beta_B$?

## Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

| y=1 | y=0 |
|---|---|
| A A A A B B B C | B C C C D D D D |

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

What's the update for $\beta_B$?
$\beta_B = 0 + 1.0 * (1.0 - .5)3.0 = 1.5$

## Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

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

What's the update for $\beta_C$?

## Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

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

What's the update for $\beta_C$?

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

## Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

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

What's the update for $\beta_D$?

## Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

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

What's the update for $\beta_D$?

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

$$\beta[j] = \beta[j] + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

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

Now you see the negative example. What's the update for $\beta_0$?

## Example Documents

$$\beta[j] = \beta[j] + \lambda(y-p)x_i$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

| y=1 | y=0 |
|---|---|
| A A A A B B B C | B C C C D D D D |

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

Now you see the negative example. What's the update for $\beta_0$?
What's the activation?

## Example Documents

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

$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

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

Now you see the negative example. What's the update for $\beta_0$?

$\sigma(.5 + 1.5 + 1.5 + 0) = 0.97$

## Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

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

Now you see the negative example. What's the update for $\beta_0$?
$\beta_0 = 0.5 + 1.0 * (0.0 - 0.97) = -0.47$

## Example Documents

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

$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

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

What's the update for $\beta_A$?

## Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

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

What's the update for $\beta_A$?

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

## Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

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

What's the update for $\beta_B$?

## Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

| y=1 | y=0 |
|---|---|
| A A A A B B B C | B C C C D D D D |

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

What's the update for $\beta_B$?
$\beta_B = 1.5 + 1.0 * (0.0 - 0.97)1.0 = 0.53$

## Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

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

What's the update for $\beta_C$?

## Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

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

What's the update for $\beta_C$?
$\beta_C = 0.5 + 1.0 * (0.0 - 0.97) 3.0 = -2.41$

## Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

| y=1 | y=0 |
|---|---|
| A A A A B B B C | B C C C D D D D |

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

What's the update for $\beta_D$?

## Example Documents

$$\beta[j] = \beta[j] + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

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

What's the update for $\beta_D$?

## Example Documents

$$\beta[j] = \beta[j] + \lambda(y-p)x_i$$
$$\vec{\beta} = \langle .5, 2, 1.5, 0.5, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

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

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

## 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

$$\beta_j = \beta_j - \lambda 2 \mu \beta_j = \beta_j \cdot (1 - 2\lambda\mu) \tag{5}$$

## How does the gradient change with regularization?

- You can do your normal update
- Then

$$\beta_j = \beta_j - \lambda 2\mu\beta_j = \beta_j \cdot (1 - 2\lambda\mu) \tag{5}$$

- 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

$$\beta_j = \beta_j - \lambda 2\mu\beta_j = \beta_j \cdot (1 - 2\lambda\mu) \tag{5}$$

- 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

$$\beta_j = \beta_j - \lambda 2 \mu \beta_j = \beta_j \cdot (1 - 2\lambda\mu) \tag{5}$$

- 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 $\beta_j$

## Example Documents (Regularized)

$$\beta[j] = (\beta[j] + \lambda(y - p)x_i) \cdot (1 - 2\lambda\mu)^m$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

| y=1 | y=0 |
|---|---|
| A A A A B B B C | B C C C D D D D |

Assume step size $\lambda = 1.0$ and $\mu = \frac{1}{4}$.

You first see the positive example. What's the update for $\beta_0$?

## Example Documents (Regularized)

$$\beta[j] = (\beta[j] + \lambda(y-p)x_i) \cdot (1-2\lambda\mu)^m$$
$$\vec{\beta} = \langle 0,0,0,0,0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

Assume step size $\lambda = 1.0$ and $\mu = \frac{1}{4}$.

You first see the positive example. What's the update for $\beta_0$?
$\beta_0 = (0 + 1.0 * (1.0 - .5)1.0) * \left(1 - \frac{2}{4}\right) = 0.25$

## Example Documents (Regularized)

$$\beta[j] = (\beta[j] + \lambda(y - p)x_i) \cdot (1 - 2\lambda\mu)^m$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

Assume step size $\lambda = 1.0$ and $\mu = \frac{1}{4}$.

What's the update for $\beta_A$?

## Example Documents (Regularized)

$$\beta[j] = (\beta[j] + \lambda(y-p)x_i) \cdot (1 - 2\lambda\mu)^m$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

Assume step size $\lambda = 1.0$ and $\mu = \frac{1}{4}$.

What's the update for $\beta_A$?

$\beta_A = (0 + 1.0 * (1.0 - .5)4.0) * \left(1 - \frac{2}{4}\right) = 1.0$

## Example Documents (Regularized)

$$\beta[j] = (\beta[j] + \lambda(y - p)x_i) \cdot (1 - 2\lambda\mu)^m$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

Assume step size $\lambda = 1.0$ and $\mu = \frac{1}{4}$.

What's the update for $\beta_B$?

## Example Documents (Regularized)

$$\beta[j] = (\beta[j] + \lambda(y-p)x_i) \cdot (1-2\lambda\mu)^m$$
$$\vec{\beta} = \langle 0,0,0,0,0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

Assume step size $\lambda = 1.0$ and $\mu = \frac{1}{4}$.

What's the update for $\beta_B$?

$\beta_B = (0 + 1.0 * (1.0 - .5)3.0) * \left(1 - \frac{2}{4}\right) = 0.75$

## Example Documents (Regularized)

$$\beta[j] = (\beta[j] + \lambda(y - p)x_i) \cdot (1 - 2\lambda\mu)^m$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

Assume step size $\lambda = 1.0$ and $\mu = \frac{1}{4}$.

What's the update for $\beta_C$?

## Example Documents (Regularized)

$$\beta[j] = (\beta[j] + \lambda(y - p)x_i) \cdot (1 - 2\lambda\mu)^m$$
$$\vec{\beta} = \langle 0, 0, 0, 0, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

Assume step size $\lambda = 1.0$ and $\mu = \frac{1}{4}$.

What's the update for $\beta_C$?

$\beta_C = (0 + 1.0 * (1.0 - .5)1.0) * \left(1 - \frac{2}{4}\right) = 0.25$

## Example Documents (Regularized)

$$\beta[j] = (\beta[j] + \lambda(y-p)x_i) \cdot (1-2\lambda\mu)^m$$
$$\vec{\beta} = \langle 0,0,0,0,0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

Assume step size $\lambda = 1.0$ and $\mu = \frac{1}{4}$.

What's the update for $\beta_D$?

## Example Documents (Regularized)

$$\beta[j] = (\beta[j] + \lambda(y-p)x_i) \cdot (1-2\lambda\mu)^m$$
$$\vec{\beta} = \langle 0,0,0,0,0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

Assume step size $\lambda = 1.0$ and $\mu = \frac{1}{4}$.

What's the update for $\beta_D$?

We don't care: leave it for later.

## Example Documents (Regularized)

$$\beta[j] = (\beta[j] + \lambda(y-p)x_i) \cdot (1-2\lambda\mu)^m$$
$$\vec{\beta} = \langle .25, 1, 0.75, 0.25, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

Assume step size $\lambda = 1.0$ and $\mu = \frac{1}{4}$.

Now you see the negative example. What's the update for $\beta_0$?

## Example Documents (Regularized)

$$\beta[j] = (\beta[j] + \lambda(y-p)x_i) \cdot (1 - 2\lambda\mu)^m$$
$$\vec{\beta} = \langle .25, 1, 0.75, 0.25, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

Assume step size $\lambda = 1.0$ and $\mu = \frac{1}{4}$.

Now you see the negative example. What's the update for $\beta_0$?

What's the activation?

## Example Documents (Regularized)

$$
\begin{aligned}
\beta[j] =& (\beta[j] + \lambda(y-p)x_i)\cdot(1-2\lambda\mu)^m \\
\vec{\beta} =& \langle .25, 1, 0.75, 0.25, 0 \rangle
\end{aligned}
$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

Assume step size $\lambda = 1.0$ and $\mu = \frac{1}{4}$.

Now you see the negative example. What's the update for $\beta_0$?
$\sigma(.25 + 0.75 + 0.75 + 0) = 0.85$

## Example Documents (Regularized)

$$\beta[j] = (\beta[j] + \lambda(y - p)x_i) \cdot (1 - 2\lambda\mu)^m$$
$$\vec{\beta} = \langle .25, 1, 0.75, 0.25, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

Assume step size $\lambda = 1.0$ and $\mu = \frac{1}{4}$.

Now you see the negative example. What's the update for $\beta_0$?
$\beta_0 = (0.5 + 1.0 * (0.0 - 0.85)) * \left(1 - \frac{2}{4}\right) = -0.30$

## Example Documents (Regularized)

$$\beta[j] = (\beta[j] + \lambda(y - p)x_i) \cdot (1 - 2\lambda\mu)^m$$
$$\vec{\beta} = \langle .25, 1, 0.75, 0.25, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

Assume step size $\lambda = 1.0$ and $\mu = \frac{1}{4}$.

What's the update for $\beta_A$?

## Example Documents (Regularized)

$$\beta[j] = (\beta[j] + \lambda(y-p)x_i) \cdot (1 - 2\lambda\mu)^m$$
$$\vec{\beta} = \langle .25, 1, 0.75, 0.25, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

Assume step size $\lambda = 1.0$ and $\mu = \frac{1}{4}$.

What's the update for $\beta_A$?

We don't care: leave it for later.

## Example Documents (Regularized)

$$\beta[j] = (\beta[j] + \lambda(y - p)x_i) \cdot (1 - 2\lambda\mu)^m$$
$$\vec{\beta} = \langle .25, 1, 0.75, 0.25, 0 \rangle$$

| y=1 | y=0 |
|---|---|
| A A A A B B B C | B C C C D D D D |

Assume step size $\lambda = 1.0$ and $\mu = \frac{1}{4}$.

What's the update for $\beta_B$?

## Example Documents (Regularized)

$$\beta[j] = (\beta[j] + \lambda(y-p)x_i) \cdot (1-2\lambda\mu)^m$$
$$\vec{\beta} = \langle .25, 1, 0.75, 0.25, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

Assume step size $\lambda = 1.0$ and $\mu = \frac{1}{4}$.

What's the update for $\beta_B$?

$\beta_B = (0.75 + 1.0 * (0.0 - 0.85)1.0) * \left(1 - \frac{2}{4}\right) = -0.05$

## Example Documents (Regularized)

$$\beta[j] = (\beta[j] + \lambda(y - p)x_i) \cdot (1 - 2\lambda\mu)^m$$
$$\vec{\beta} = \langle .25, 1, 0.75, 0.25, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

Assume step size $\lambda = 1.0$ and $\mu = \frac{1}{4}$.

What's the update for $\beta_C$?

## Example Documents (Regularized)

$$\beta[j] = (\beta[j] + \lambda(y - p)x_i) \cdot (1 - 2\lambda\mu)^m$$
$$\vec{\beta} = \langle .25, 1, 0.75, 0.25, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

Assume step size $\lambda = 1.0$ and $\mu = \frac{1}{4}$.

What's the update for $\beta_C$?

$\beta_C = (0.5 + 1.0 * (0.0 - 0.85)3.0) * \left(1 - \frac{2}{4}\right) = -1.15$

## Example Documents (Regularized)

$$\beta[j] = (\beta[j] + \lambda(y - p)x_i) \cdot (1 - 2\lambda\mu)^m$$
$$\vec{\beta} = \langle .25, 1, 0.75, 0.25, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

Assume step size $\lambda = 1.0$ and $\mu = \frac{1}{4}$.

What's the update for $\beta_D$?

## Example Documents (Regularized)

$$\beta[j] = (\beta[j] + \lambda(y-p)x_i) \cdot (1-2\lambda\mu)^m$$
$$\vec{\beta} = \langle .25, 1, 0.75, 0.25, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

Assume step size $\lambda = 1.0$ and $\mu = \frac{1}{4}$.

What's the update for $\beta_D$?

## Example Documents (Regularized)

$$\beta[j] = (\beta[j] + \lambda(y - p)x_i) \cdot (1 - 2\lambda\mu)^m$$
$$\vec{\beta} = \langle .25, 1, 0.75, 0.25, 0 \rangle$$

**y=1**

A A A A B B B C

**y=0**

B C C C D D D D

Assume step size $\lambda = 1.0$ and $\mu = \frac{1}{4}$.

$$\beta_D = (0.0 + 1.0 * (0.0 - 0.85)4.0) * \left(1 - \frac{2}{4}\right)^2 = -0.85$$

## Next time . . .

- Multinomial logistic regression (more than one option)
- Crafting effective features
- Preparation for third homework