# Introduction to Machine Learning

Machine Learning: Jordan Boyd-Graber
University of Maryland
STOCASTIC GRADIENT DESCENT FOR LOGISTIC REGRESSION

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

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

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

## Algorithm

1. Initialize a vector $B$ to be all zeros
2. For $t = 1, \ldots, T$
   - For each example $\vec{x}_i, y_i$ and feature $j$:
     - Compute $\pi_i \equiv \Pr(y_i = 1 \,|\, \vec{x}_i)$
     - Set $\beta[j] = \beta[j]' + \lambda(y_i - \pi_i)x_i$
3. Output the parameters $\beta_1, \ldots, \beta_d$.

## Example Documents

$$\beta[j] = \beta[j] + \lambda(y_i - \pi_i)x_i$$
$$\vec{\beta} = \langle \beta_{bias} = 0, \beta_A = 0, \beta_B = 0, \beta_C = 0, \beta_D = 0 \rangle$$

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

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

You first see the positive example. First, compute $\pi_1$

## Example Documents

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

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

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

You first see the positive example. First, compute $\pi_1$

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

## Example Documents

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

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

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

You first see the positive example. First, compute $\pi_1$

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

## Example Documents

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

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

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

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

## Example Documents

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

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

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

What's the update for $\beta_{bias}$?

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

## Example Documents

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

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

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

What's the update for $\beta_{bias}$?

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

## Example Documents

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

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

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

What's the update for $\beta_A$?

## Example Documents

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

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

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

What's the update for $\beta_A$?

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

## Example Documents

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

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

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

What's the update for $\beta_A$?

$\beta_A = \beta'_A + \lambda \cdot (y_1 - \pi_1) \cdot x_{1,A} = 0.0 + 1.0 \cdot (1.0 - 0.5) \cdot 4.0 \quad = 2.0$

## Example Documents

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

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

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

What's the update for $\beta_B$?

## Example Documents

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

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

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

What's the update for $\beta_B$?

$\beta_B = \beta'_B + \lambda \cdot (y_1 - \pi_1) \cdot x_{1,B} = 0.0 + 1.0 \cdot (1.0 - 0.5) \cdot 3.0$

## Example Documents

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

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

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

What's the update for $\beta_B$?

$\beta_B = \beta'_B + \lambda \cdot (y_1 - \pi_1) \cdot x_{1,B} = 0.0 + 1.0 \cdot (1.0 - 0.5) \cdot 3.0 \quad =1.5$

## Example Documents

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

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

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

What's the update for $\beta_C$?

## Example Documents

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

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

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

What's the update for $\beta_C$?

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

## Example Documents

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

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

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

What's the update for $\beta_C$?

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

## Example Documents

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

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

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

What's the update for $\beta_D$?

## Example Documents

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

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

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

What's the update for $\beta_D$?

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

## Example Documents

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

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

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

What's the update for $\beta_D$?

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

## Example Documents

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

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

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

Now you see the negative example. What's $\pi_2$?

## Example Documents

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

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

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

Now you see the negative example. What's $\pi_2$?

$$\pi_2 = \Pr(y_2 = 1 \,|\, \vec{x}_2) = \frac{\exp \beta^T x_i}{1 + \exp \beta^T x_i} = \frac{\exp\{.5 + 1.5 + 1.5 + 0\}}{\exp\{.5 + 1.5 + 1.5 + 0\} + 1} =$$

## Example Documents

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

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

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

Now you see the negative example. What's $\pi_2$?

$$\pi_2 = \Pr(y_2 = 1 \,|\, \vec{x}_2) = \frac{\exp \beta^T x_i}{1 + \exp \beta^T x_i} = \frac{\exp\{.5+1.5+1.5+0\}}{\exp\{.5+1.5+1.5+0\}+1} = 0.97$$

## Example Documents

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

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

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

Now you see the negative example. What's $\pi_2$?

$\pi_2 = 0.97$

What's the update for $\beta_{bias}$?

## Example Documents

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

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

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

What's the update for $\beta_{bias}$?

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

## Example Documents

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

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

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

What's the update for $\beta_{bias}$?

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

## Example Documents

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

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

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

What's the update for $\beta_A$?

## Example Documents

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

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

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

What's the update for $\beta_A$?

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

## Example Documents

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

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

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

What's the update for $\beta_A$?

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

## Example Documents

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

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

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

What's the update for $\beta_B$?

## Example Documents

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

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

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

What's the update for $\beta_B$?

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

## Example Documents

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

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

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

What's the update for $\beta_B$?

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

## Example Documents

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

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

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

What's the update for $\beta_C$?

## Example Documents

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

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

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

What's the update for $\beta_C$?

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

## Example Documents

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

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

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

What's the update for $\beta_C$?

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

## Example Documents

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

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

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

What's the update for $\beta_D$?

## Example Documents

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

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

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

What's the update for $\beta_D$?

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

## Example Documents

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

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

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

What's the update for $\beta_D$?

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

## How does the gradient change with regularization?

- First

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

## How does the gradient change with regularization?

- First

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

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

## How does the gradient change with regularization?

- First

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

- Doesn't depend on $X$ or $Y$. Just makes all your weights smaller
- Then do update as usual

## How does the gradient change with regularization?

- First

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

- Doesn't depend on $X$ or $Y$. Just makes all your weights smaller
- Then do update as usual
- But difficult to update every feature every time (if there are many features)

## How does the gradient change with regularization?

- First

$$\beta_j = \beta'_j - \lambda 2\mu\beta_j = \beta'_j \cdot (1 - 2\lambda\mu) \qquad (5)$$

- Doesn't depend on $X$ or $Y$. Just makes all your weights smaller
- Then do update as usual
- But difficult to update every feature every time (if there are many features)
- Following this up, we note that we can perform $m$ successive "regularization" updates by letting $\beta_j = \beta'_j \cdot (1 - 2\lambda\mu)^{m_j}$

### Basic Idea

Don't 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 $\beta_j$

## Revised Algorithm

1. Initialize a vector $\beta$ to be all zeros
2. Initialize a vector $A$ to be all zeros
3. For $t = 1, \dots, T$
   - For each example $\vec{x}_i, y_i$ and feature $j$:
     - Simulate regularization updates: $\beta[j] = \beta[j] \cdot (1 - 2\lambda\mu)^{k - A[j] - 1}$
     - Compute $\pi_i \equiv \Pr(y_i = 1 \,|\, \vec{x}_i)$
     - Set $\beta[j] = \beta[j] \cdot (1 - 2\lambda\mu) + \lambda(y_i - \pi_i)x_i$
     - Keep track of last update for feature $A[j] = k$
4. For each paramter, catch up on missing updates $\beta[j] = \beta[j] \cdot (1 - 2\lambda\mu)^{T - A[j]}$
5. Output the parameters $\beta_1, \dots, \beta_d$.

## Example Documents (Regularized)

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

| $y_1 = 1$ |
| --- |
| A A A A B B B C |

| $y_2 = 0$ |
| --- |
| B C C C D D D D |

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

## Example Documents (Regularized)

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

| $y_1 = 1$ |
| --- |
| A A A A B B B C |

| $y_2 = 0$ |
| --- |
| B C C C D D D D |

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

You first see the positive example. $\pi_1$ is still 0.5.

## Example Documents (Regularized)

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

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

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

You first see the positive example. $\pi_1$ is still 0.5.
What's the update for $\beta_{bias}$?

## Example Documents (Regularized)

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

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

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

What's the update for $\beta_{bias}$?

$\beta_{bias} = \beta'_{bias}(1 - 2 \cdot \lambda \cdot \mu)^{m_{bias}} + \lambda(y_1 - \pi_1)x_{1,bias} =$
$0.0\left(1 - 2 \cdot 1.0 \cdot \frac{1}{4}\right)^1 + 1.0 \cdot (1.0 - 0.5)1.0$

## Example Documents (Regularized)

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

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

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

What's the update for $\beta_{bias}$?

$\beta_{bias} = \beta'_{bias}(1 - 2 \cdot \lambda \cdot \mu)^{m_{bias}} + \lambda(y_1 - \pi_1)x_{1,bias} =$

$0.0\left(1 - 2 \cdot 1.0 \cdot \frac{1}{4}\right)^1 + 1.0 \cdot (1.0 - 0.5)1.0 \quad =2$

## Example Documents (Regularized)

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

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

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

What's the update for $\beta_A$?

## Example Documents (Regularized)

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

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

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

What's the update for $\beta_A$? $\beta_A = \beta_A'(1 - 2 \cdot \lambda \cdot \mu)^{m_A} + \lambda(y_1 - \pi_1)x_{1,A} =$
$0.0\left(1 - 2 \cdot 1.0 \cdot \frac{1}{4}\right)^1 + 1.0 \cdot (1.0 - 0.5)4.0$

## Example Documents (Regularized)

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

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

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

What's the update for $\beta_A$? $\beta_A = \beta_A'(1 - 2 \cdot \lambda \cdot \mu)^{m_A} + \lambda(y_1 - \pi_1)x_{1,A} =$ $0.0\left(1 - 2 \cdot 1.0 \cdot \frac{1}{4}\right)^1 + 1.0 \cdot (1.0 - 0.5)4.0 \quad = 1.0$

## Example Documents (Regularized)

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

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

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

What's the update for $\beta_B$?

## Example Documents (Regularized)

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

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

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

What's the update for $\beta_B$? $\beta_B = \beta_B'(1 - 2 \cdot \lambda \cdot \mu)^{m_B} + \lambda(y_1 - \pi_1)x_{1,B} =$
$0.0\left(1 - 2 \cdot 1.0 \cdot \frac{1}{4}\right)^1 + 1.0 \cdot (1.0 - 0.5)3.0$

## Example Documents (Regularized)

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

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

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

What's the update for $\beta_B$? $\beta_B = \beta_B'(1 - 2 \cdot \lambda \cdot \mu)^{m_B} + \lambda(y_1 - \pi_1)x_{1,B} =$
$0.0\left(1 - 2 \cdot 1.0 \cdot \frac{1}{4}\right)^1 + 1.0 \cdot (1.0 - 0.5)3.0 \quad =1.5$

## Example Documents (Regularized)

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

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

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

What's the update for $\beta_C$?

## Example Documents (Regularized)

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

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

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

What's the update for $\beta_C$? $\beta_C = \beta_C'(1 - 2 \cdot \lambda \cdot \mu)^{m_C} + \lambda(y_1 - \pi_1)x_{1,C} = 0.0\left(1 - 2 \cdot 1.0 \cdot \frac{1}{4}\right)^1 + 1.0 \cdot (1.0 - 0.5)1.0$

## Example Documents (Regularized)

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

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

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

What's the update for $\beta_C$? $\beta_C = \beta'_C (1 - 2 \cdot \lambda \cdot \mu)^{m_C} + \lambda(y_1 - \pi_1)x_{1,C} =$
$0.0\left(1 - 2 \cdot 1.0 \cdot \frac{1}{4}\right)^1 + 1.0 \cdot (1.0 - 0.5)1.0 \quad =0.5$

## Example Documents (Regularized)

$$\beta[j] = \beta[j]' \cdot (1 - 2\lambda\mu)^{m_j} + \lambda(y - p)x_i$$

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

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

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

What's the update for $\beta_D$?

## Example Documents (Regularized)

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

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

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

What's the update for $\beta_D$?
We don't care: leave it for later.

## Example Documents (Regularized)

$$\beta[j] = \beta[j]' \cdot (1 - 2\lambda\mu)^{m_j} + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle .5, 2.0, 1.5, 0.5, 0 \rangle$$

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

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

## Example Documents (Regularized)

$$\beta[j] = \beta[j]' \cdot (1 - 2\lambda\mu)^{m_j} + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle .5, 2.0, 1.5, 0.5, 0 \rangle$$

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

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

Now you see the negative example. What's $\pi_2$?

## Example Documents (Regularized)

$$\beta[j] = \beta[j]' \cdot (1 - 2\lambda\mu)^{m_j} + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle .5, 2.0, 1.5, 0.5, 0 \rangle$$

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

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

Now you see the negative example. What's $\pi_2$?

$$\pi_2 = \Pr(y_2 = 1 \mid \vec{x}_2) = \frac{\exp \beta^T x_i}{1 + \exp \beta^T x_i} = \frac{\exp\{.5+1.5+1.5+0\}}{\exp\{.5+1.5+1.5+0\}+1} =$$

## Example Documents (Regularized)

$$\beta[j] = \beta[j]' \cdot (1 - 2\lambda\mu)^{m_j} + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle .5, 2.0, 1.5, 0.5, 0 \rangle$$

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

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

Now you see the negative example. What's $\pi_2$?

$$\pi_2 = \Pr(y_2 = 1 \,|\, \vec{x}_2) = \frac{\exp \beta^T x_i}{1 + \exp \beta^T x_i} = \frac{\exp\{.5 + 1.5 + 1.5 + 0\}}{\exp\{.5 + 1.5 + 1.5 + 0\} + 1} = 0.97$$

## Example Documents (Regularized)

$$\beta[j] = \beta[j]' \cdot (1 - 2\lambda\mu)^{m_j} + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle .5, 2.0, 1.5, 0.5, 0 \rangle$$

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

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

$\pi_2 = 0.97$

Careful: You'd need to regularize $\beta_D$ if it weren't already zero (multiply it by $(1 - 2\lambda\mu)^{m_j}$)

What's the update for $\beta_{bias}$?

## Example Documents (Regularized)

$$\beta[j] = \beta[j]' \cdot (1 - 2\lambda\mu)^{m_j} + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle .5, 2.0, 1.5, 0.5, 0 \rangle$$

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

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

What's the update for $\beta_{bias}$?

$\beta_{bias} = \beta'_{bias}(1 - 2 \cdot \lambda \cdot \mu)^{m_{bias}} + \lambda(y_2 - \pi_2)x_{2,bias} =$

$0.5\left(1 - 2 \cdot 1.0 \cdot \frac{1}{4}\right)^1 + 1.0 \cdot (0.0 - 0.97)1.0$

## Example Documents (Regularized)

$$\beta[j] = \beta[j]' \cdot (1 - 2\lambda\mu)^{m_j} + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle .5, 2.0, 1.5, 0.5, 0 \rangle$$

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

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

What's the update for $\beta_{bias}$?

$\beta_{bias} = \beta'_{bias}(1 - 2 \cdot \lambda \cdot \mu)^{m_{bias}} + \lambda(y_2 - \pi_2)x_{2,bias} =$

$0.5\left(1 - 2 \cdot 1.0 \cdot \frac{1}{4}\right)^1 + 1.0 \cdot (0.0 - 0.97)1.0$ =-0.72

## Example Documents (Regularized)

$$\beta[j] = \beta[j]' \cdot (1 - 2\lambda\mu)^{m_j} + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle .5, 2.0, 1.5, 0.5, 0 \rangle$$

**$y_1 = 1$**

A A A A B B B C

**$y_2 = 0$**

B C C C D D D D

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

What's the update for $\beta_A$?

## Example Documents (Regularized)

$$\beta[j] = \beta[j]' \cdot (1 - 2\lambda\mu)^{m_j} + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle .5, 2.0, 1.5, 0.5, 0 \rangle$$

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

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

What's the update for $\beta_A$?

We don't care: leave it for later.

## Example Documents (Regularized)

$$\beta[j] = \beta[j]' \cdot (1 - 2\lambda\mu)^{m_j} + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle .5, 2.0, 1.5, 0.5, 0 \rangle$$

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

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

What's the update for $\beta_B$?

## Example Documents (Regularized)

$$\beta[j] = \beta[j]' \cdot (1 - 2\lambda\mu)^{m_j} + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle .5, 2.0, 1.5, 0.5, 0 \rangle$$

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

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

What's the update for $\beta_B$? $\beta_B = \beta_B'(1 - 2 \cdot \lambda \cdot \mu)^{m_B} + \lambda(y_2 - \pi_2)x_{2,B} =$
$1.5\left(1 - 2 \cdot 1.0 \cdot \frac{1}{4}\right)^1 + 1.0 \cdot (0.0 - 0.97)1.0$

## Example Documents (Regularized)

$$\beta[j] = \beta[j]' \cdot (1 - 2\lambda\mu)^{m_j} + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle .5, 2.0, 1.5, 0.5, 0 \rangle$$

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

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

What's the update for $\beta_B$? $\beta_B = \beta_B'(1 - 2 \cdot \lambda \cdot \mu)^{m_B} + \lambda(y_2 - \pi_2)x_{2,B} =$
$1.5\left(1 - 2 \cdot 1.0 \cdot \frac{1}{4}\right)^1 + 1.0 \cdot (0.0 - 0.97)1.0$ =-0.22

## Example Documents (Regularized)

$$
\begin{aligned}
\beta[j] &= \beta[j]' \cdot (1 - 2\lambda\mu)^{m_j} + \lambda(y - p)x_i \\
\vec{\beta} &= \langle .5, 2.0, 1.5, 0.5, 0 \rangle
\end{aligned}
$$

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

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

What's the update for $\beta_C$?

## Example Documents (Regularized)

$$\beta[j] = \beta[j]' \cdot (1 - 2\lambda\mu)^{m_j} + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle .5, 2.0, 1.5, 0.5, 0 \rangle$$

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

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

What's the update for $\beta_C$? $\beta_C = \beta_C'(1 - 2 \cdot \lambda \cdot \mu)^{m_C} + \lambda(y_2 - \pi_2)x_{2,C} = 0.5\left(1 - 2 \cdot 1.0 \cdot \frac{1}{4}\right)^1 + 1.0 \cdot (0.0 - 0.97)3.0$

## Example Documents (Regularized)

$$\beta[j] = \beta[j]' \cdot (1 - 2\lambda\mu)^{m_j} + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle .5, 2.0, 1.5, 0.5, 0 \rangle$$

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

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

What's the update for $\beta_C$? $\beta_C = \beta'_C(1 - 2 \cdot \lambda \cdot \mu)^{m_C} + \lambda(y_2 - \pi_2)x_{2,C} =$
$0.5\left(1 - 2 \cdot 1.0 \cdot \frac{1}{4}\right)^1 + 1.0 \cdot (0.0 - 0.97)3.0$ =-2.7

## Example Documents (Regularized)

$$\beta[j] = \beta[j]' \cdot (1 - 2\lambda\mu)^{m_j} + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle .5, 2.0, 1.5, 0.5, 0 \rangle$$

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

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

What's the update for $\beta_D$?

## Example Documents (Regularized)

$$\beta[j] = \beta[j]' \cdot (1 - 2\lambda\mu)^{m_j} + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle .5, 2.0, 1.5, 0.5, 0 \rangle$$

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

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

What's the update for $\beta_D$? $\beta_D = \beta_D'(1 - 2 \cdot \lambda \cdot \mu)^{m_D} + \lambda(y_2 - \pi_2)x_{2,D} = 0.0\left(1 - 2 \cdot 1.0 \cdot \frac{1}{4}\right)^2 + 1.0 \cdot (0.0 - 0.97)4.0$

## Example Documents (Regularized)

$$\beta[j] = \beta[j]' \cdot (1 - 2\lambda\mu)^{m_j} + \lambda(y - p)x_i$$
$$\vec{\beta} = \langle .5, 2.0, 1.5, 0.5, 0 \rangle$$

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

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

What's the update for $\beta_D$? $\beta_D = \beta_D'(1 - 2 \cdot \lambda \cdot \mu)^{m_D} + \lambda(y_2 - \pi_2)x_{2,D} =$
$0.0\left(1 - 2 \cdot 1.0 \cdot \frac{1}{4}\right)^2 + 1.0 \cdot (0.0 - 0.97)4.0$ =-3.9

## If this were final iteration ...

- Need to remember that $\beta_A$ is still waiting for regularization

$$\beta_A^{\text{final}} = \beta_A \left(1 - 21.0\frac{1}{4}\right)^1 = 1.0 \tag{6}$$

## Next time . . .

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