CS 171: Intro to ML and DM

Christian Shelton

UC Riverside

Slide Set 7: Logistic Regression

Slides from CS 171

- From UC Riverside
  - CS 171: Introduction to Machine Learning and Data Mining
  - Professor Christian Shelton
- DO NOT REDISTRIBUTE
  - These slides contain copyrighted material (used with permission) from
    - Elements of Statistical Learning (Hastie, et al.)
    - Pattern Recognition and Machine Learning (Bishop)
    - An Introduction to Machine Learning (Kubat)
    - Machine Learning: A Probabilistic Perspective (Murphy)
  - For use only by enrolled students in the course

Problems with Perceptrons

- Does not work when data are not separable
- $w^\top x$ is treated the same as $(2w)^\top x$
  - Difficult if used for multi-class
  - Why the extra free parameter?

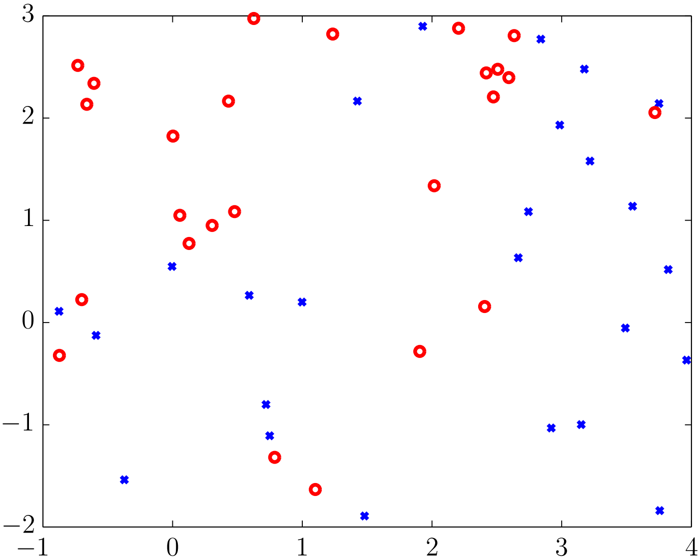
Solution?

Problems with Perceptrons

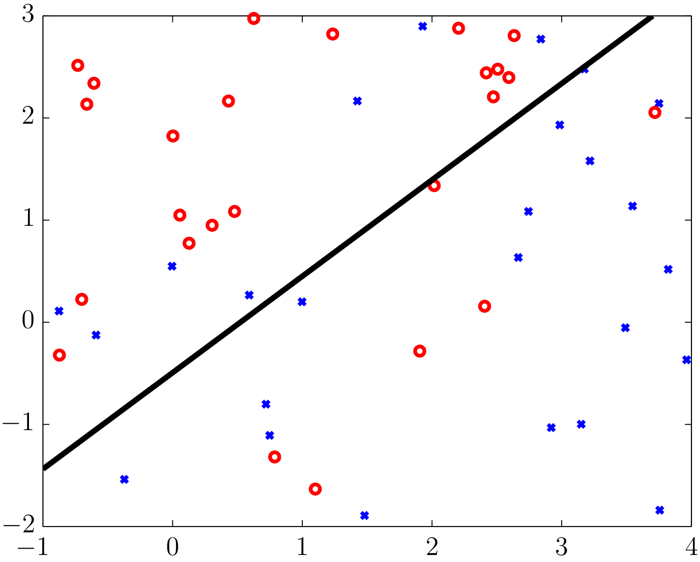
- Does not work when data are not separable
- $w^\top x$ is treated the same as $(2w)^\top x$
  - Difficult if used for multi-class
  - Why the extra free parameter?

Solution?
Make the output $w^\top x$ related to something.

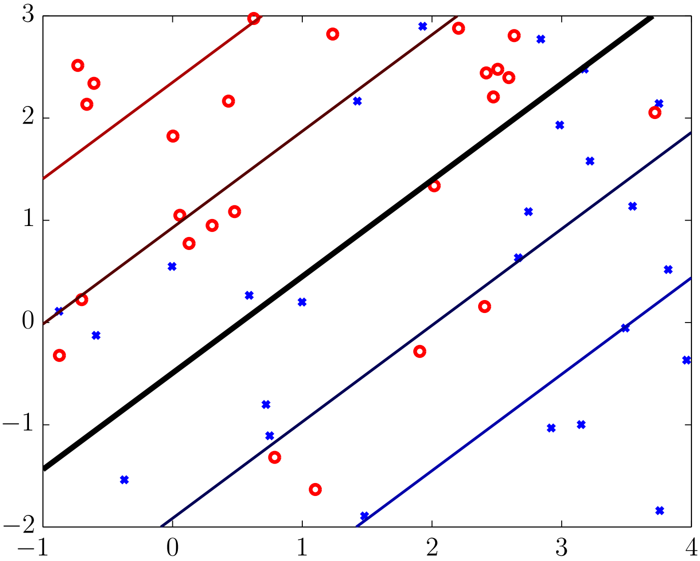
Calibrated Output



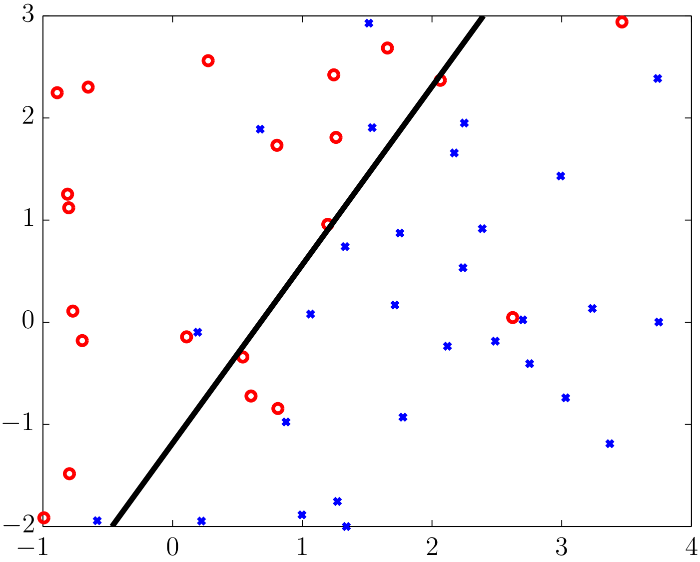
Calibrated Output



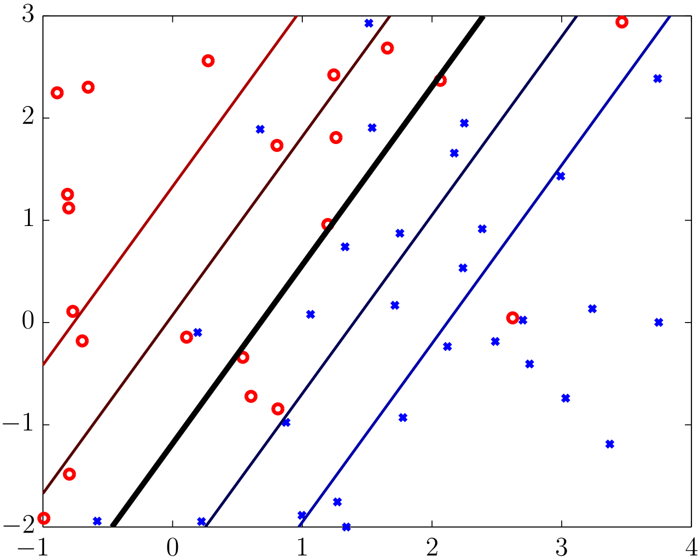
Calibrated Output



Calibrated Output



Calibrated Output

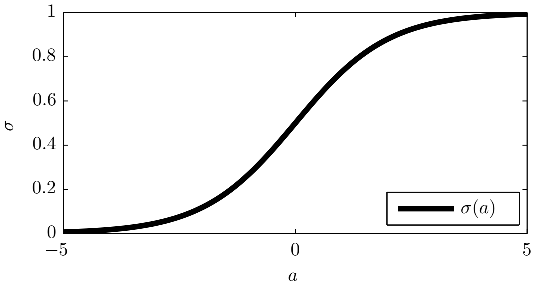


Sigmoid / Logistic

To make $w^\top x$ relate to the chance that the label is positive, we have to remap it from $(-\infty, \infty)$ to $(0, 1)$.

We use the "sigmoid" or "logistic" function:

$$\sigma(a) = \frac{1}{1 + e^{-a}}$$

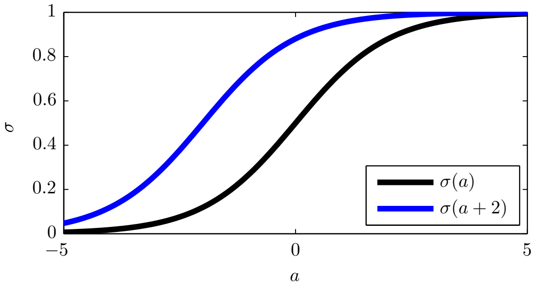


Sigmoid / Logistic

To make $w^\top x$ relate to the chance that the label is positive, we have to remap it from $(-\infty, \infty)$ to $(0, 1)$.

We use the "sigmoid" or "logistic" function:

$$\sigma(a) = \frac{1}{1 + e^{-a}}$$

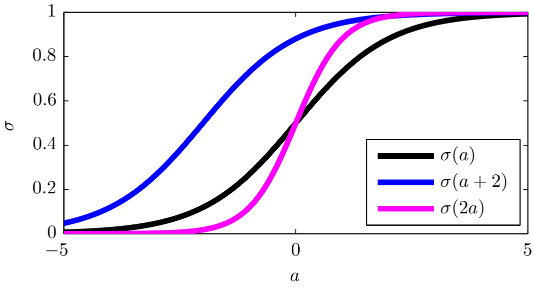


Sigmoid / Logistic

To make $w^\top x$ relate to the chance that the label is positive, we have to remap it from $(-\infty, \infty)$ to $(0, 1)$.

We use the "sigmoid" or "logistic" function:

$$\sigma(a) = \frac{1}{1 + e^{-a}}$$



Sigmoid Derivative

We will (later) need the derivative of $\sigma(a)$, so we'll do it now:

$$
\begin{aligned}
\sigma(a) &= \frac{1}{1+e^{-a}} \\
\sigma'(a) &= \frac{-1}{(1+e^{-a})^2}\left(-e^{-a}\right) \\
&= \frac{1}{1+e^{-a}}\frac{e^{-a}}{1+e^{-a}} \\
&= \frac{1}{1+e^{-a}}\left(1-\frac{1}{1+e^{-a}}\right) \\
&= \sigma(a)\left(1-\sigma(a)\right)
\end{aligned}
$$

(Binary) Logistic Regression

Let $f(x) = w^\top x$ be the output of our classifier.

(Binary) Logistic Regression

Let $f(x) = w^\top x$ be the output of our classifier.
We will let $\sigma(f(x))$ be the probability that the class is positive:

$$p(y = +1 \mid x) = \sigma(f(x)) = \frac{1}{1 + e^{-w^\top x}}$$

(Binary) Logistic Regression

Let $f(x) = w^\top x$ be the output of our classifier.
We will let $\sigma(f(x))$ be the probability that the class is positive:

$$p(y = +1 \mid x) = \sigma(f(x)) = \frac{1}{1 + e^{-w^\top x}}$$

Then the probability the class if negative is

$$p(y = -1 \mid x) = 1 - \sigma(f(x)) = \frac{e^{-w^\top x}}{1 + e^{-w^\top x}} = \frac{1}{1 + e^{w^\top x}} = \sigma(-f(x))$$

(Binary) Logistic Regression

Let $f(x) = w^\top x$ be the output of our classifier.
We will let $\sigma(f(x))$ be the probability that the class is positive:

$$p(y = +1 \mid x) = \sigma(f(x)) = \frac{1}{1 + e^{-w^\top x}}$$

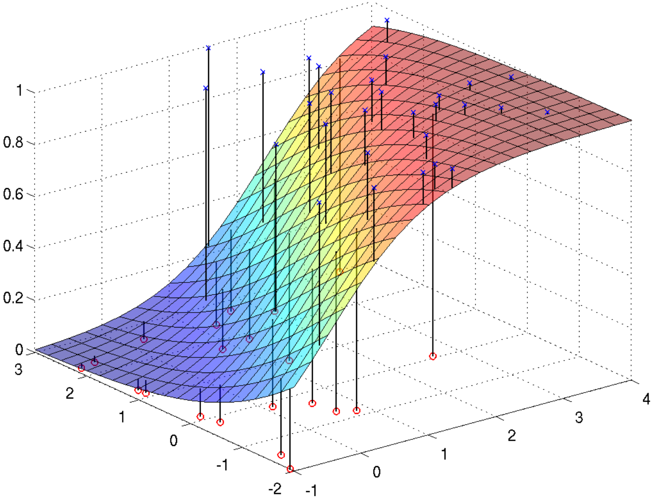
Then the probability the class if negative is

$$p(y = -1 \mid x) = 1 - \sigma(f(x)) = \frac{e^{-w^\top x}}{1 + e^{-w^\top x}} = \frac{1}{1 + e^{w^\top x}} = \sigma(-f(x))$$

So, in general

$$p(y \mid x) = \sigma(yf(x))$$

(Binary) Logistic Regression



(Binary) Logistic Regression

Goal: pick w so that the ys are most likely, given the xs

$$\max_w \prod_{i=1}^{m} p(y_i \mid x_i)$$

(Binary) Logistic Regression

Goal: pick w so that the ys are most likely, given the xs

$$\max_w \prod_{i=1}^{m} p(y_i \mid x_i)$$

Same as

$$\max_w \sum_{i=1}^{m} \ln p(y_i \mid x_i)$$

(Binary) Logistic Regression

Goal: pick w so that the ys are most likely, given the xs

$$\max_w \prod_{i=1}^{m} p(y_i \mid x_i)$$

Same as

$$\max_w \sum_{i=1}^{m} \ln p(y_i \mid x_i)$$

So,

$$L = -\sum_{i=1}^{m} \ln p(y_i \mid x_i) = -\sum_{i=1}^{m} \ln \sigma(y_i f(x_i))$$

(Binary) Logistic Regression

How to minimize

$$L = -\sum_{i=1}^{m} \ln \sigma(y_i w^\top x_i)$$

(Binary) Logistic Regression

How to minimize

$$L = -\sum_{i=1}^{m} \ln \sigma(y_i w^\top x_i)$$

Gradient Descent... need derivative:

$$p_i = \sigma(y_i w^\top x_i)$$

$$L = -\sum_{i=1}^{m} \ln p_i$$

$$\begin{aligned}
\frac{\partial L}{\partial w_k} &= \sum_{i=1}^{m} \frac{-1}{p_i} \frac{\partial p_i}{\partial w_k} \\
&= \sum_{i=1}^{m} \frac{-1}{p_i} p_i(1-p_i) \frac{\partial y_i w^\top x_i}{\partial w_k} \\
&= \sum_{i=1}^{m} \frac{-1}{p_i} p_i(1-p_i) y_i x_{i,k} \\
-\nabla_w L &= \sum_{i=1}^{m} (1-p_i) y_i x_i
\end{aligned}$$

(Binary) Logistic Regression Algorithm

Recall: $L = \sum_{i=1}^{m} -\ln \sigma(y_i w^\top x_i)$

(Binary) Logistic Regression Algorithm

Recall: $L = \sum_{i=1}^{m} -\ln \sigma(y_i w^\top x_i)$

Gradient descent algorithm:

1. Let w be a random weight vector
2. While w is not at a local minimum of L
    1. Let $g \leftarrow 0$
    2. For $i = 1, \ldots, m$
        1. Let $p_i \leftarrow \sigma(y_i w^\top x_i)$
        2. Let $g_i \leftarrow -(1 - p_i) y_i x_i$
        3. Let $g \leftarrow g + g_i$
    3. Let $w \leftarrow w - \eta g$

(Binary) Logistic Regression Algorithm

Recall: $L = \sum_{i=1}^{m} -\ln \sigma(y_i w^\top x_i)$

Gradient descent algorithm:

1. Let w be a random weight vector
2. While w is not at a local minimum of L
    1. Let $g \leftarrow 0$
    2. For $i = 1, \ldots, m$
        1. Let $p_i \leftarrow \sigma(y_i w^\top x_i)$
        2. Let $g_i \leftarrow -(1-p_i) y_i x_i$
        3. Let $g \leftarrow g + g_i$
    3. Let $w \leftarrow w - \eta g$

Stochastic gradient descent algorithm:

1. Let w be a random weight vector
2. While w is not at a local minimum of L
    1. Let $g \leftarrow 0$
    2. For $i = 1, \ldots, m$
        1. Let $p_i \leftarrow \sigma(y_i w^\top x_i)$
        2. Let $g \leftarrow -(1-p_i) y_i x_i$
        3. Let $w \leftarrow w - \eta g$

Regularized (Binary) Logistic Regression

Just like linear regression, we can regularize the weights to smooth it:

$$L = -\sum_{i=1}^{m} \ln p_i + \lambda \sum_{j=1}^{n} w_j^2$$

$$\left[p_i = \sigma(y_i w^\top x_i)\right]$$

The algorithm is much the same. The gradient changes only slightly:

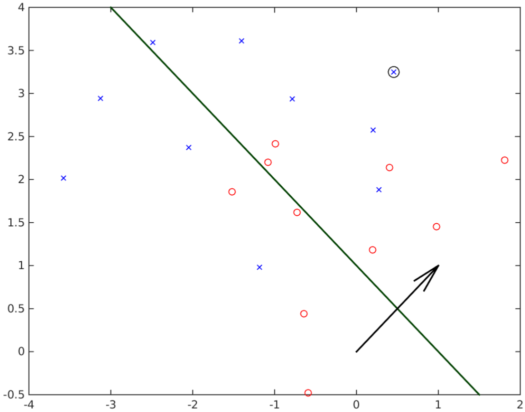
$$-\nabla_w L = -2\lambda w + \sum_{i=1}^{m} (1 - p_i) y_i x_i$$

(Binary) Logistic Regression

Optimization notes:

- (Unique) global minimum
  - Except if data are separable and $\lambda = 0$
- More advanced optimization possible and often used
  - Second-order methods (uses second derivatives)
  - Does not require picking step sizes
  - Based on Newton's method
  - In this case, it is called iteratively reweighted least squares (IRLS)

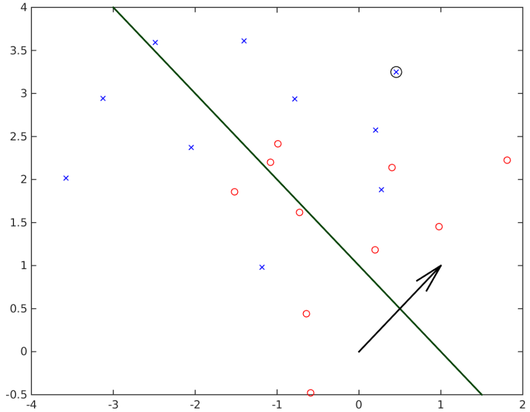
Logistic Regression SGD Example



$$\eta = 0.10$$
$$w = \begin{bmatrix} -1.00 & 1.00 & 1.00 \end{bmatrix}^\top$$

$$y_i = 1.00$$
$$x_i = \begin{bmatrix} 1.00 & 0.45 & 3.25 \end{bmatrix}^\top$$

Logistic Regression SGD Example

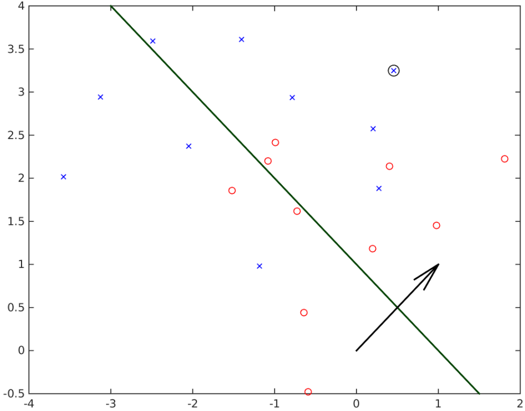


$$\eta = 0.10$$
$$w = \begin{bmatrix} -1.00 & 1.00 & 1.00 \end{bmatrix}^\top$$

$$y_i = 1.00$$
$$x_i = \begin{bmatrix} 1.00 & 0.45 & 3.25 \end{bmatrix}^\top$$

$$w^\top x_i = 2.70$$
$$\sigma(w^\top x_i) = 0.94$$
$$p_i = \sigma(y_i w^\top x_i) = 0.94$$

Logistic Regression SGD Example



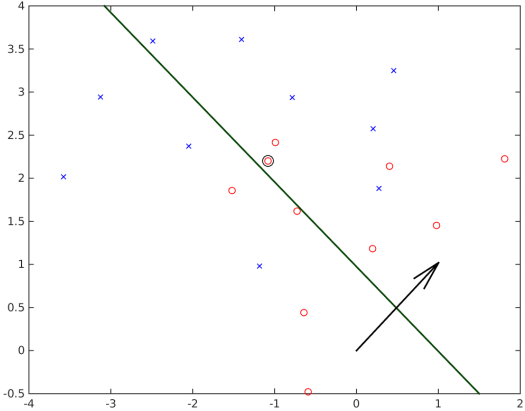
$$\eta = 0.10$$
$$w = \begin{bmatrix} -1.00 & 1.00 & 1.00 \end{bmatrix}^\top$$

$$y_i = 1.00$$
$$x_i = \begin{bmatrix} 1.00 & 0.45 & 3.25 \end{bmatrix}^\top$$

$$w^\top x_i = 2.70$$
$$\sigma(w^\top x_i) = 0.94$$
$$p_i = \sigma(y_i w^\top x_i) = 0.94$$

$$(1 - p_i) y_i x_i = \begin{bmatrix} 0.06 & 0.03 & 0.20 \end{bmatrix}^\top$$

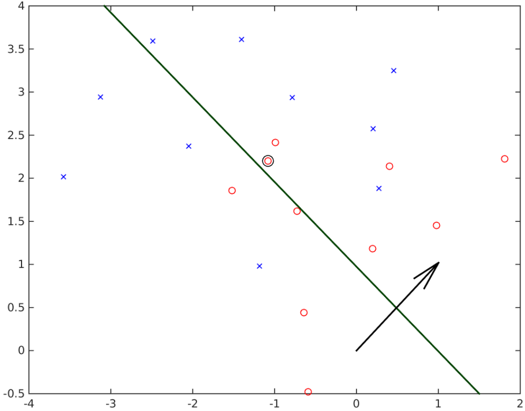
Logistic Regression SGD Example



$$\eta = 0.10$$
$$w = \begin{bmatrix} -0.99 & 1.00 & 1.02 \end{bmatrix}^\top$$

$$y_i = -1.00$$
$$x_i = \begin{bmatrix} 1.00 & -1.08 & 2.20 \end{bmatrix}^\top$$

Logistic Regression SGD Example

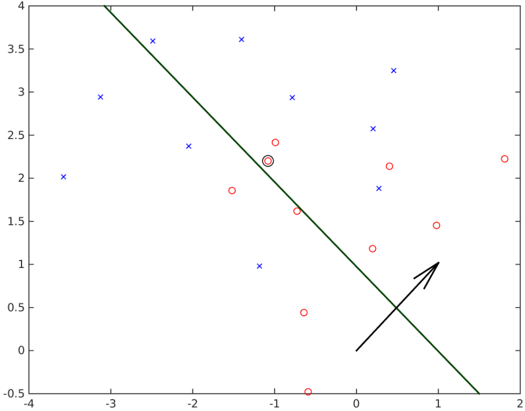


$$\eta = 0.10$$
$$w = \begin{bmatrix} -0.99 & 1.00 & 1.02 \end{bmatrix}^\top$$

$$y_i = -1.00$$
$$x_i = \begin{bmatrix} 1.00 & -1.08 & 2.20 \end{bmatrix}^\top$$

$$w^\top x_i = 0.17$$
$$\sigma(w^\top x_i) = 0.54$$
$$p_i = \sigma(y_i w^\top x_i) = 0.46$$

Logistic Regression SGD Example



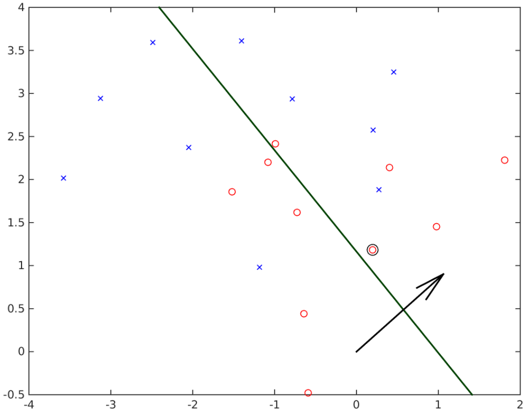
$$\eta = 0.10$$
$$w = \begin{bmatrix} -0.99 & 1.00 & 1.02 \end{bmatrix}^\top$$

$$y_i = -1.00$$
$$x_i = \begin{bmatrix} 1.00 & -1.08 & 2.20 \end{bmatrix}^\top$$

$$w^\top x_i = 0.17$$
$$\sigma(w^\top x_i) = 0.54$$
$$p_i = \sigma(y_i w^\top x_i) = 0.46$$

$$(1 - p_i) y_i x_i = \begin{bmatrix} -0.54 & 0.58 & -1.19 \end{bmatrix}^\top$$

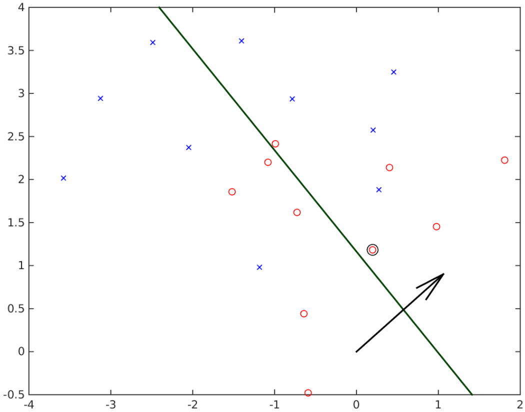
Logistic Regression SGD Example



$$\eta = 0.10$$
$$w = \begin{bmatrix} -1.05 & 1.06 & 0.90 \end{bmatrix}^\top$$

$$y_i = -1.00$$
$$x_i = \begin{bmatrix} 1.00 & 0.20 & 1.18 \end{bmatrix}^\top$$

Logistic Regression SGD Example

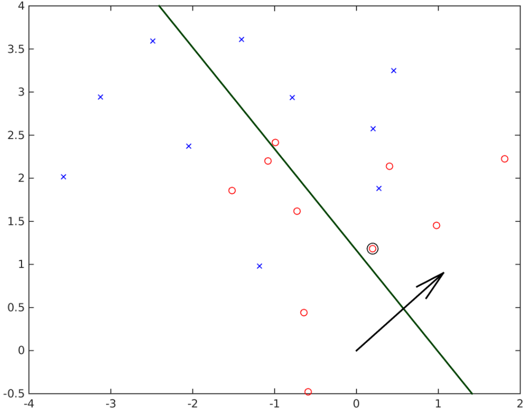


$$\eta = 0.10$$
$$w = \begin{bmatrix}-1.05 & 1.06 & 0.90\end{bmatrix}^\top$$

$$y_i = -1.00$$
$$x_i = \begin{bmatrix}1.00 & 0.20 & 1.18\end{bmatrix}^\top$$

$$w^\top x_i = 0.23$$
$$\sigma(w^\top x_i) = 0.56$$
$$p_i = \sigma(y_i w^\top x_i) = 0.44$$

Logistic Regression SGD Example



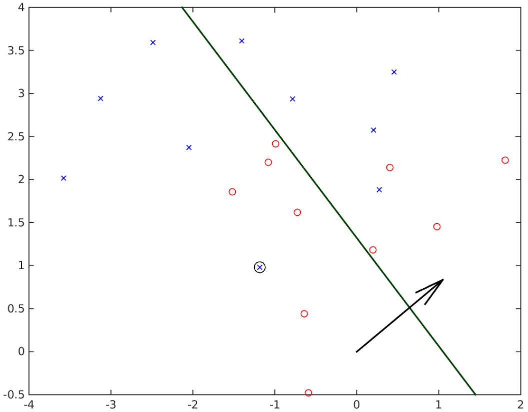
$$\eta = 0.10$$
$$w = \begin{bmatrix} -1.05 & 1.06 & 0.90 \end{bmatrix}^\top$$

$$y_i = -1.00$$
$$x_i = \begin{bmatrix} 1.00 & 0.20 & 1.18 \end{bmatrix}^\top$$

$$w^\top x_i = 0.23$$
$$\sigma(w^\top x_i) = 0.56$$
$$p_i = \sigma(y_i w^\top x_i) = 0.44$$

$$(1 - p_i) y_i x_i = \begin{bmatrix} -0.56 & -0.11 & -0.66 \end{bmatrix}^\top$$

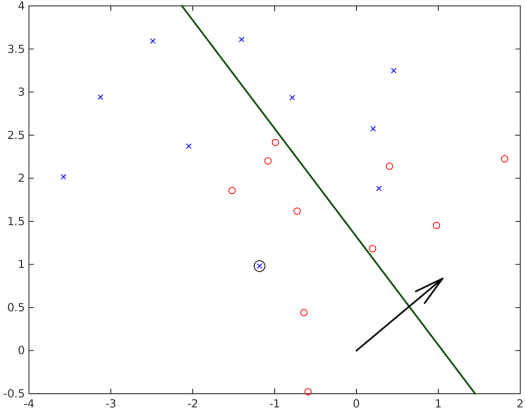
Logistic Regression SGD Example



$$\eta = 0.10$$
$$w = \begin{bmatrix} -1.10 & 1.05 & 0.84 \end{bmatrix}^\top$$

$$y_i = 1.00$$
$$x_i = \begin{bmatrix} 1.00 & -1.18 & 0.98 \end{bmatrix}^\top$$

Logistic Regression SGD Example

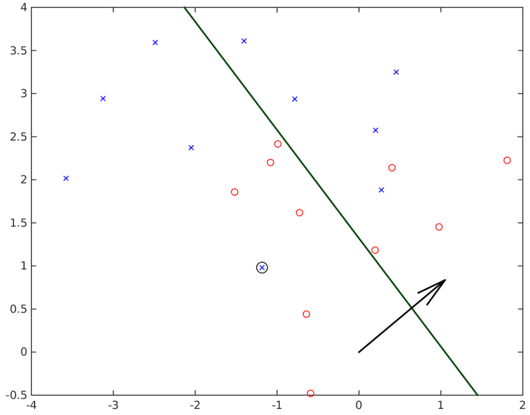


$$\eta = 0.10$$
$$w = \begin{bmatrix} -1.10 & 1.05 & 0.84 \end{bmatrix}^\top$$

$$y_i = 1.00$$
$$x_i = \begin{bmatrix} 1.00 & -1.18 & 0.98 \end{bmatrix}^\top$$

$$w^\top x_i = -1.53$$
$$\sigma(w^\top x_i) = 0.18$$
$$p_i = \sigma(y_i w^\top x_i) = 0.18$$

Logistic Regression SGD Example



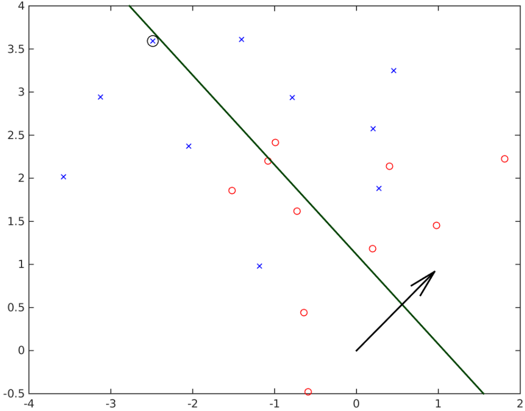
$$\eta = 0.10$$
$$w = \begin{bmatrix} -1.10 & 1.05 & 0.84 \end{bmatrix}^\top$$

$$y_i = 1.00$$
$$x_i = \begin{bmatrix} 1.00 & -1.18 & 0.98 \end{bmatrix}^\top$$

$$w^\top x_i = -1.53$$
$$\sigma(w^\top x_i) = 0.18$$
$$p_i = \sigma(y_i w^\top x_i) = 0.18$$

$$(1 - p_i) y_i x_i = \begin{bmatrix} 0.82 & -0.97 & 0.80 \end{bmatrix}^\top$$

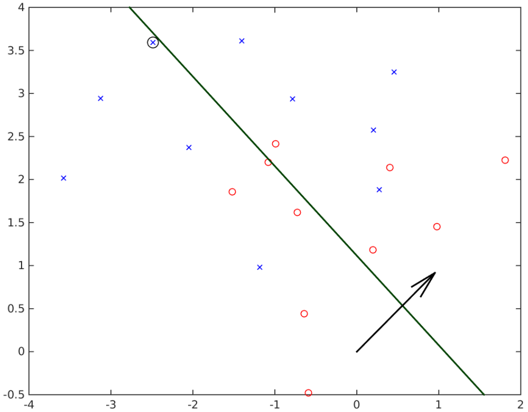
Logistic Regression SGD Example



$$\eta = 0.10$$
$$w = \begin{bmatrix} -1.02 & 0.95 & 0.92 \end{bmatrix}^\top$$

$$y_i = 1.00$$
$$x_i = \begin{bmatrix} 1.00 & -2.49 & 3.59 \end{bmatrix}^\top$$

Logistic Regression SGD Example

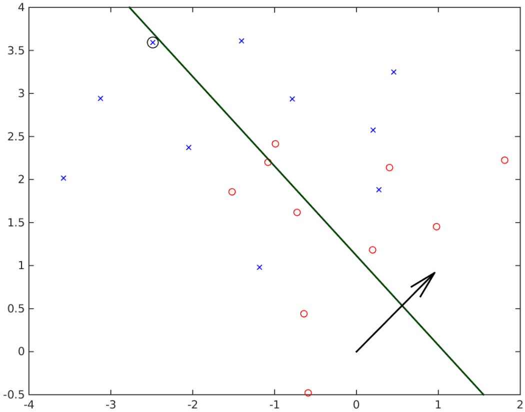


$$\eta = 0.10$$
$$w = \begin{bmatrix} -1.02 & 0.95 & 0.92 \end{bmatrix}^\top$$

$$y_i = 1.00$$
$$x_i = \begin{bmatrix} 1.00 & -2.49 & 3.59 \end{bmatrix}^\top$$

$$w^\top x_i = -0.11$$
$$\sigma(w^\top x_i) = 0.47$$
$$p_i = \sigma(y_i w^\top x_i) = 0.47$$

Logistic Regression SGD Example



$$\eta = 0.10$$
$$w = \begin{bmatrix} -1.02 & 0.95 & 0.92 \end{bmatrix}^\top$$

$$y_i = 1.00$$
$$x_i = \begin{bmatrix} 1.00 & -2.49 & 3.59 \end{bmatrix}^\top$$

$$w^\top x_i = -0.11$$
$$\sigma(w^\top x_i) = 0.47$$
$$p_i = \sigma(y_i w^\top x_i) = 0.47$$

$$(1 - p_i) y_i x_i = \begin{bmatrix} 0.53 & -1.31 & 1.89 \end{bmatrix}^\top$$