Do not distribute course material

You may not and may not allow others to reproduce or distribute lecture notes and course materials publicly whether or not a fee is charged.

Topic 3
Linear Classification & Logistic Regression

- http://cs229.stanford.edu/notes2020fall/notes2020fall/cs229-notes1.pdf
- https://eight2late.wordpress.com/2017/07/11/a-gentle-introduction-to-logistic-regression-and-lasso-regularisation-using-r/

PROF. LINDA SELLIE

Some slides/approaches used are from Prof. Rangan
Many approaches used are from CMU 18-661

Outline

- Motivating example: How can we classify? — How can we use a hyperplane for a classification problem?
- Estimating probabilities — Can we predict not only which class an example belongs to - but a confidence score of that classification
- Maximum likelihood — How can we find the most likely hyperplane? Could we write a function to describe how likely a hyperplane was to have generated the dataset?
  - Iterative approach - gradient ascent — Maximizing the function
- Thinking about different types of error — Some errors are more costly than other errors. Can we modify our predictions to decrease one type of error (and perhaps increase another type of error?
- Transformation of the features — Extending our algorithm to nonlinear decision boundaries
- Multiple classes — What if we have more than two classes?

Outline

- Motivating example: How can we classify?
  - How can we use a hyperplane for a classification problem?
- Which model
  - Can we predict not only which class an example belongs to - but a confidence score of that classification
- Finding an objective function
  - How can we find the most likely hyperplane? Could we write a function to describe how likely a hyperplane was to have generated the dataset?
- Optimizer
  - Ite - gradient ascent
    - Maximizing the function
- Thinking about different types of error
  - Some errors are more costly than other errors. Can we modify our predictions to decrease one type of error (and perhaps increase another type of error?
- Transformation of the features
  - Extending our algorithm to nonlinear decision boundaries
- Multiple classes
  - What if we have more than two classes?

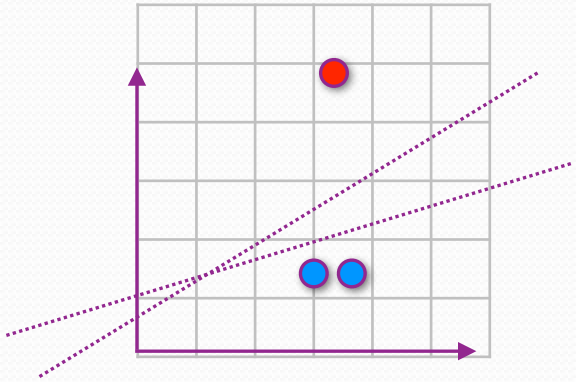
Given $D = \{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(N)}, y^{(N)})\}$ how can we find the "best" hyperplane, $\mathbf{w}$?

Optimize $\mathbf{w}$

We first need to decide what makes one hyperplane better than another (i.e. an objective function) Pair share

MLE!

Likelihood of seeing the data



- Our model: $p(y^{(i)} \mid \mathbf{x}^{(i)}; \mathbf{w}) = \begin{cases} \sigma(\mathbf{w}^T \mathbf{x}^{(i)}) & \text{for } y^{(i)} = 1 \\ 1 - \sigma(\mathbf{w}^T \mathbf{x}^{(i)}) & \text{for } y^{(i)} = 0 \end{cases}$

- Given the following data:

$\mathbf{x}^{(1)T}$ = [1, 3.2 4.7] $y^{(1)}$ = 0
$\mathbf{x}^{(2)T}$ = [1, 3.5 1.4] $y^{(2)}$ = 1
$\mathbf{x}^{(3)T}$ = [1, 3. 1.4] $y^{(3)}$ = 1

- How likely were we to see the data if the line was:

$$\mathbf{w} = \begin{bmatrix} 1/2 \\ 2/3 \\ -1 \end{bmatrix} \quad \left(1 - \frac{1}{1+e^{-(1/2+(2/3)3.2-4.7)}}\right) \left(\frac{1}{1+e^{-(1/2+(2/3)3.5-1.4)}}\right) \left(\frac{1}{1+e^{-(1/2+(2/3)3-1.4)}}\right) = 0.54$$

1-0.11 0.81 0.75

$$\mathbf{w} = \begin{bmatrix} 1 \\ 1/3 \\ -1 \end{bmatrix} \quad \left(1 - \frac{1}{1+e^{-(1+(1/3)3.2-4.7)}}\right) \left(\frac{1}{1+e^{-(1+(1/3)3.5-1.4)}}\right) \left(\frac{1}{1+e^{-(1+(1/3)3-1.4)}}\right) = 0.41$$

Pair share: Write the conditional likelihood function for these three examples

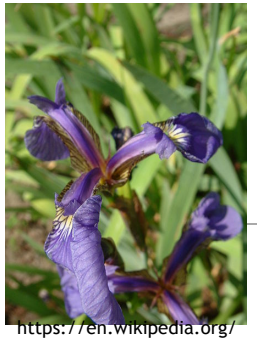
Our model:

$$p(y^{(i)} \mid \mathbf{x}^{(i)}; \mathbf{w}) = \begin{cases} \sigma(\mathbf{w}^T \mathbf{x}^{(i)}) & \text{for } y^{(i)} = 1 \\ 1 - \sigma(\mathbf{w}^T \mathbf{x}^{(i)}) & \text{for } y^{(i)} = 0 \end{cases}$$

$$p(y = 1 \mid \mathbf{x}; \mathbf{w}) = \sigma(\mathbf{w}^T \mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x})}}$$

$$p(y = 0 \mid \mathbf{x}; \mathbf{w}) = 1 - \sigma(\mathbf{w}^T \mathbf{x}) = 1 - \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}}}$$

$$D = \{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), (\mathbf{x}^{(3)}, y^{(3)})\}$$



https://en.wikipedia.org/wiki/Iris_flower_data_set#/media/

setosa

$\mathbf{x}^{(2)T}$ = [1, 3.5 1.4] $y^{(2)}$ = 1

$\mathbf{x}^{(3)T}$ = [1, 3. 1.4] $y^{(3)}$ = 1

https://commons.wikimedia.org/

versicolor

$\mathbf{x}^{(1)T}$ = [1, 3.2 4.7] $y^{(1)}$ = 0

The conditional likelihood function

versicolor

$\mathbf{x}^{(1)T}$ = [1 3.2 4.7] $y^{(1)} = 0$

setosa

$\mathbf{x}^{(2)T}$ = [1 3.5 1.4] $y^{(2)} = 1$

$\mathbf{x}^{(3)T}$ = [1 3. 1.4] $y^{(3)} = 1$

$$p(y = 1 \mid \mathbf{x}; \mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{x}}}$$

$$p(y = 0 \mid \mathbf{x}; \mathbf{w}) = 1 - \sigma(\mathbf{w}^T\mathbf{x}) = 1 - \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{x}}}$$

$$L(\mathbf{w}) = \left(1 - \frac{1}{1 + e^{-(\mathbf{w}^T\mathbf{x}^{(1)})}}\right)\left(\frac{1}{1 + e^{-(\mathbf{w}^T\mathbf{x}^{(2)})}}\right)\left(\frac{1}{1 + e^{-(\mathbf{w}^T\mathbf{x}^{(3)})}}\right)$$

$$L(\mathbf{w}) = \left(1 - \frac{1}{1 + e^{-(w_0 + w_1 3.2 + w_2 4.7)}}\right)\left(\frac{1}{1 + e^{-(w_0 + w_1 3.5 + w_2 1.4)}}\right)\left(\frac{1}{1 + e^{-(w_0 + w_1 3 + w_2 1.4)}}\right)$$

$$L(\mathbf{w}) = \left(1 - p(y = 1 \mid \mathbf{x}^{(1)}; \mathbf{w})\right)\ p(y = 1 \mid \mathbf{x}^{(2)}; \mathbf{w}) p(y = 1 \mid \mathbf{x}^{(3)}; \mathbf{w}) = \prod_{i=1}^{N} p(y^{(i)} \ \begin{matrix}\text{correctly} \\ \text{predicted}\end{matrix} \mid \mathbf{x}^{(i)}; \mathbf{w})$$

$$L(\mathbf{w}) = \prod_{i:y^{(i)}=1} p(y^{(i)} = 1 \mid \mathbf{x}^{(i)}; \mathbf{w}) \prod_{i:y^{(i)}=0} \left(1 - p(y^{(i)} = 1 \mid \mathbf{x}^{(i)}; \mathbf{w})\right)$$

The conditional likelihood function

Conditional likelihood function (conditioned on **x**). Larger value means more likely

$$L(\mathbf{w}) = \prod_{i:y^{(i)}=1} p(y^{(i)}=1 \mid \mathbf{x}^{(i)};\mathbf{w}) \prod_{i:y^{(i)}=0} \left(1 - p(y^{(i)}=1 \mid \mathbf{x}^{(i)};\mathbf{w})\right)$$

Here we assume all the examples are independent

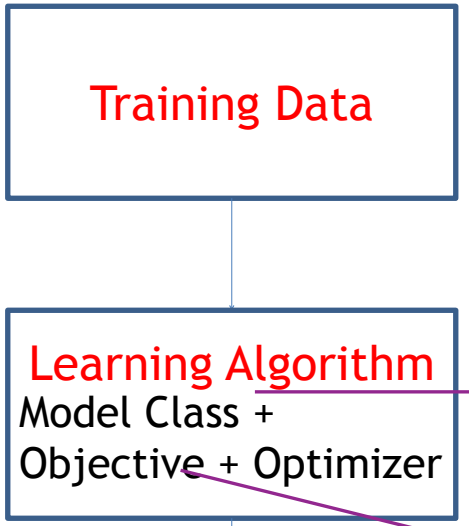
$$\prod_{i:y^{(i)}=1} p(y^{(i)}=1 \mid \mathbf{x}^{(i)};\mathbf{w})^{y^{(i)}} \left(1 - p(y^{(i)}=1 \mid \mathbf{x}^{(i)};\mathbf{w})\right)^{1-y^{(i)}} \prod_{i:y^{(i)}=0} \left(1 - p(y^{(i)}=1 \mid \mathbf{x}^{(i)};\mathbf{w})\right)^{1-y^{(i)}} p(y^{(i)}=1 \mid \mathbf{x}^{(i)};\mathbf{w})^{y^{(i)}}$$

$$L(\mathbf{w}) = \prod_{i=1}^{N} p(y^{(i)}=1 \mid \mathbf{x}^{(i)};\mathbf{w})^{y^{(i)}} (1 - p(y^{(i)}=1 \mid \mathbf{x}^{(i)};\mathbf{w}))^{1-y^{(i)}}$$

$$= \prod_{i=1}^{N} \sigma(\mathbf{w}^T\mathbf{x}^{(i)})^{y^{(i)}} \left(1 - \sigma(\mathbf{w}^T\mathbf{x}^{(i)})\right)^{1-y^{(i)}}$$

Define: $p(y^{(i)}=1 \mid \mathbf{x}^{(i)};\mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x}^{(i)}) = \dfrac{1}{1+e^{-\mathbf{w}^T\mathbf{x}^{(i)}}}$

Supervised Learning



$$p(y = 1 \mid \mathbf{x}; \mathbf{w}) = \sigma(\mathbf{w}^T \mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{x})}$$

$$L(\mathbf{w}) = \prod_{i=1}^{N} \sigma(\mathbf{w}^T \mathbf{x}^{(i)})^{y^{(i)}} \left(1 - \sigma(\mathbf{w}^T \mathbf{x}^{(i)})\right)^{1-y^{(i)}}$$

Input (say) patient data

x

Hypothesis or function

Output 0/1. 0 no cancer, 1 cancer

(Estimated probability of cancer)

Pair share: how do we find the $\mathbf{w}$ that maximizes this function

$$\text{Maximize } L(\mathbf{w}) = \prod_{i=1}^{N} \sigma(\mathbf{w}^T \mathbf{x}^{(i)})^{y^{(i)}} \left(1 - \sigma(\mathbf{w}^T \mathbf{x}^{(i)})\right)^{1-y^{(i)}}$$

The log-likelihood function

Define: $p(y^{(i)} = 1 \mid \mathbf{x}^{(i)}; \mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{x}}}$

- We wanted to maximize $L(\mathbf{w}) = \prod_{i=1}^{N} \sigma(\mathbf{w}^T\mathbf{x}^{(i)})^{y^{(i)}} \left(1 - \sigma(\mathbf{w}^T\mathbf{x}^{(i)})\right)^{1-y^{(i)}}$

- This is the same as maximizing $\ell(\mathbf{w}) = \ln(L(\mathbf{w}))$

$$= \ln\left[\prod_{i=1}^{N} \sigma(\mathbf{w}^T\mathbf{x}^{(i)})^{y^{(i)}} \left(1 - \sigma(\mathbf{w}^T\mathbf{x}^{(i)})\right)^{1-y^{(i)}}\right]$$

$$= \sum_{i=1}^{N} \ln\left[\sigma(\mathbf{w}^T\mathbf{x}^{(i)})^{y^{(i)}} \left(1 - \sigma(\mathbf{w}^T\mathbf{x}^{(i)})\right)^{1-y^{(i)}}\right]$$

$\log a^c b^d = c \log a + d \log b$

a^c b^d

$$= \sum_{i=1}^{N} \left[y^{(i)} \ln \sigma(\mathbf{w}^T\mathbf{x}^{(i)}) + (1 - y^{(i)}) \ln\left(1 - \sigma(\mathbf{w}^T\mathbf{x}^{(i)})\right)\right]$$

$c \log a$ $d \log b$

- How do we maximize the conditional likelihood?

Equivalent objective function choices

$$\ell(\mathbf{w}) = \sum_{i=1}^{N} \left[y^{(i)} \ln \sigma(\mathbf{w}^T \mathbf{x}^{(i)}) + (1 - y^{(i)}) \ln \left(1 - \sigma(\mathbf{w}^T \mathbf{x}^{(i)}) \right) \right]$$

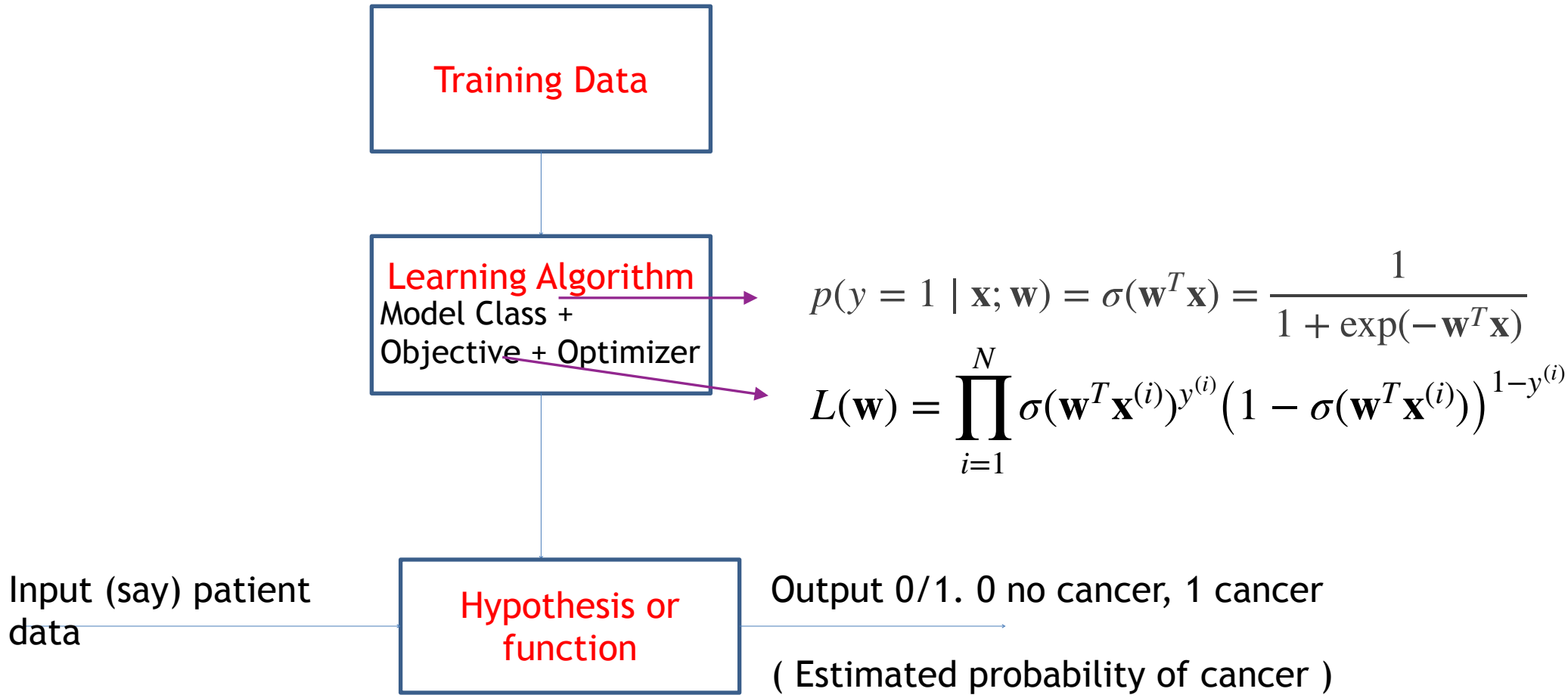
Finds same $\mathbf{w}$

$$\frac{1}{N}\ell(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} \left[y^{(i)} \ln \sigma(\mathbf{w}^T \mathbf{x}^{(i)}) + (1 - y^{(i)}) \ln \left(1 - \sigma(\mathbf{w}^T \mathbf{x}^{(i)}) \right) \right]$$

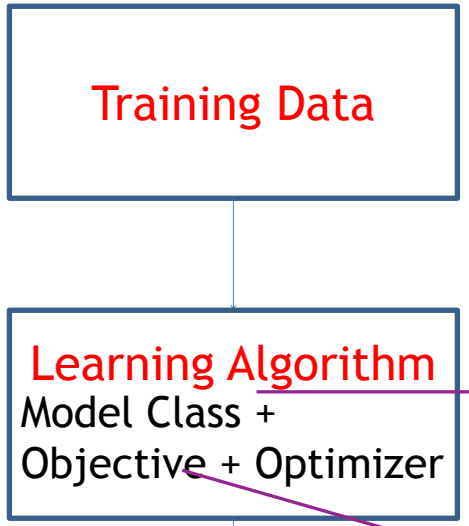
$$-\ell(\mathbf{w}) = \sum_{i=1}^{N} - \left[y^{(i)} \ln \sigma(\mathbf{w}^T \mathbf{x}^{(i)}) + (1 - y^{(i)}) \ln \left(1 - \sigma(\mathbf{w}^T \mathbf{x}^{(i)}) \right) \right]$$

Negative log likelihood is the Cross entropy error
(Find which parameters minimize the function)

Supervised Learning



Supervised Learning



$$p(y = 1 \mid \mathbf{x}; \mathbf{w}) = \sigma(\mathbf{w}^T \mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{x})}$$

$$\ell(\mathbf{w}) = \sum_{i=1}^{N} \left[y^{(i)} \ln \sigma(\mathbf{w}^T \mathbf{x}^{(i)}) + (1 - y^{(i)}) \ln \left(1 - \sigma(\mathbf{w}^T \mathbf{x}^{(i)})\right) \right]$$

Input (say) patient data

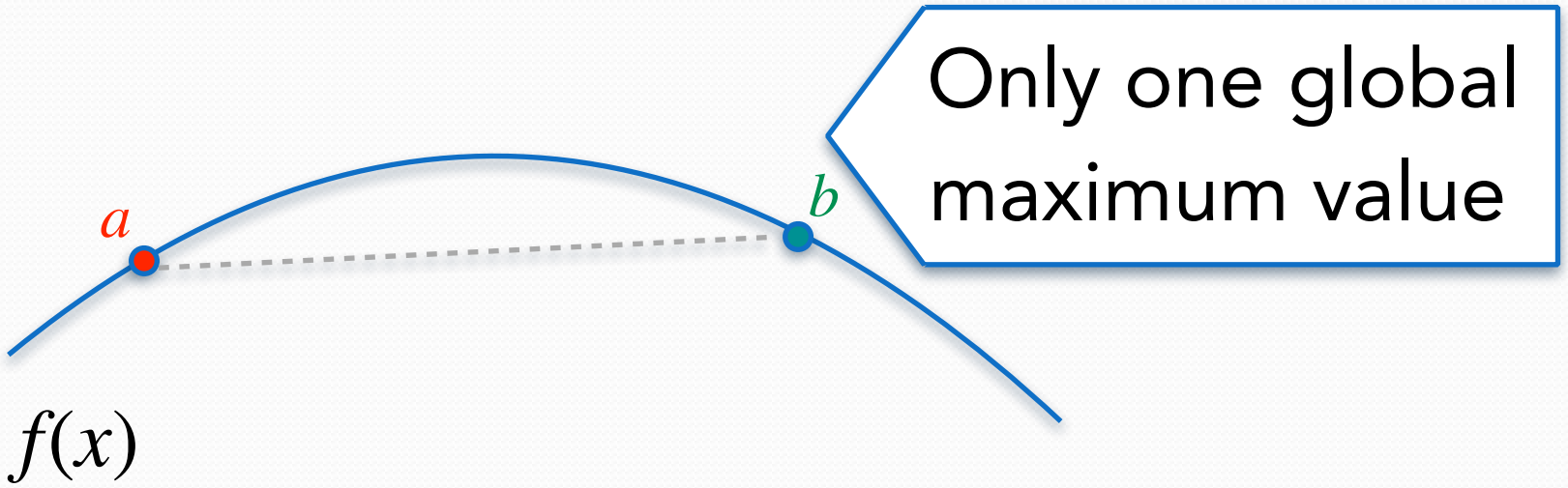
x

Hypothesis or function

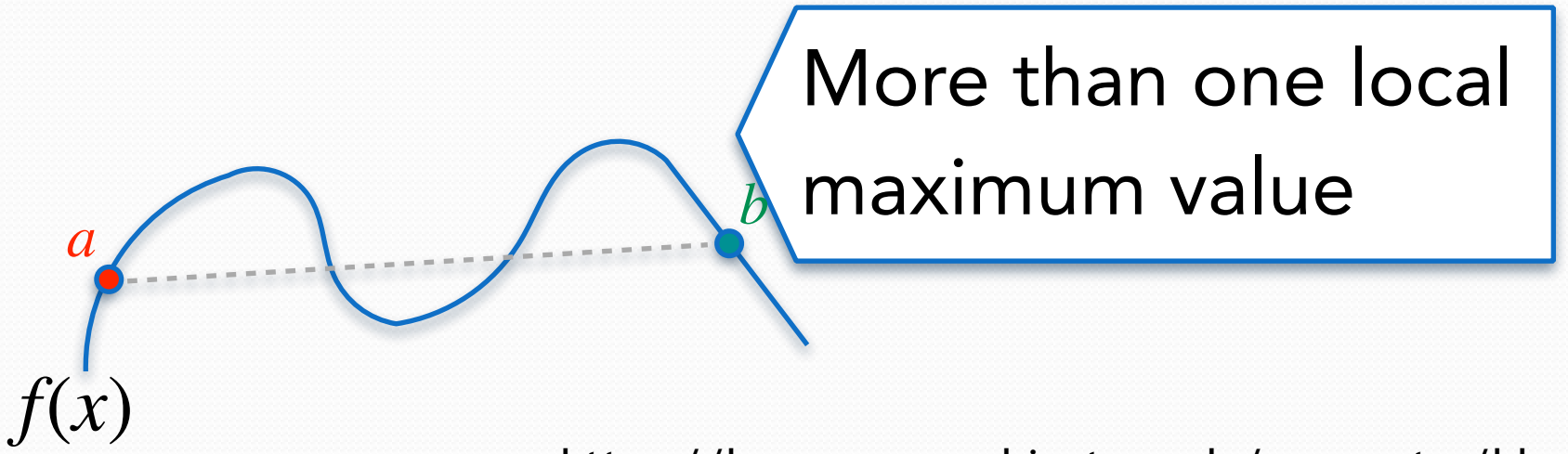
Output 0/1. 0 no cancer, 1 cancer

(Estimated probability of cancer)

Concave function



Not a concave function



https://homes.cs.washington.edu/~marcotcr/blog/concavity/

Finding $\mathbf{w}$

$$\ell(\mathbf{w}) = \sum_{i=1}^{N} \left[y^{(i)} \ln \sigma(\mathbf{w}^T \mathbf{x}^{(i)}) + (1 - y^{(i)}) \ln \left(1 - \sigma(\mathbf{w}^T \mathbf{x}^{(i)}) \right) \right]$$

- There is no closed form solution (i.e. we don't have a global optimization technique)
- We can move toward the optimal value using gradient *ascent*
- To find which way to move, we take the gradient of $\ell(\mathbf{w})$.

Outline

- Motivating example: How can we classify? — How can we use a hyperplane for a classification problem?
- Estimating probabilities — Can we predict not only which class an example belongs to - but a confidence score of that classification
- Maximum likelihood — How can we find the most likely hyperplane? Could we write a function to describe how likely a hyperplane was to have generated the dataset?
  - Iterative approach - gradient ascent — Maximizing the function
- Thinking about different types of error — Some errors are more costly than other errors. Can we modify our predictions to decrease one type of error (and perhaps increase another type of error?
- Transformation of the features — Extending our algorithm to nonlinear decision boundaries
- Multiple classes — What if we have more than two classes?

Maximize a function by repeatedly moving toward the maximum

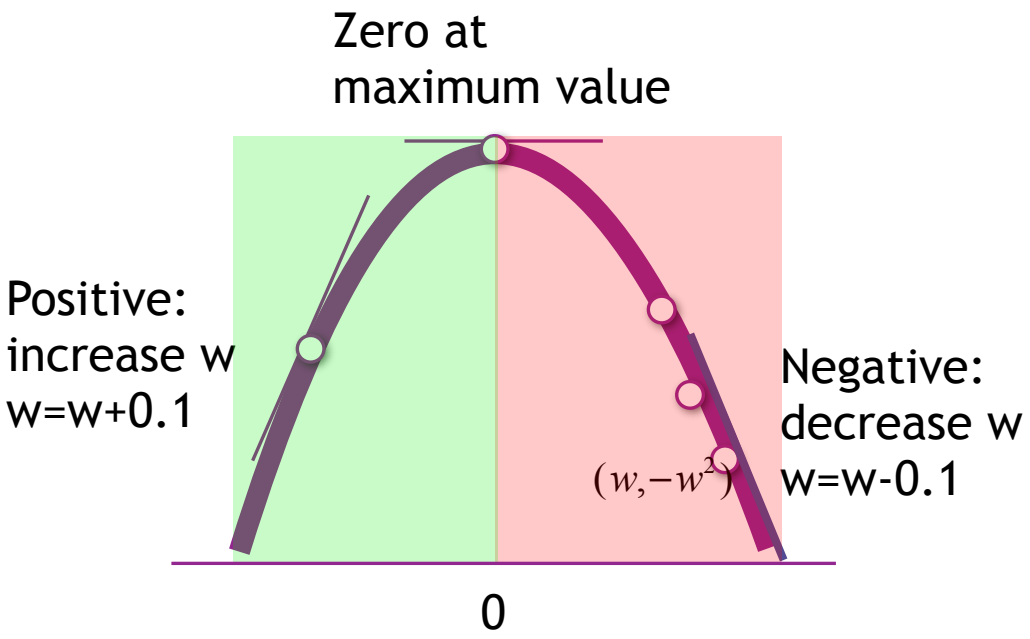
1. For i = 1 to num_iters:
if f'(w) > 0 then f is increasing,
move w a little to the ___

if f'(w) < 0 then f is decreasing,
move w a little to the ___

- right, right
- left, left
- right, left
- left, right
- none of these options

$f(w) = -w^2$

$df/dw = -2w$



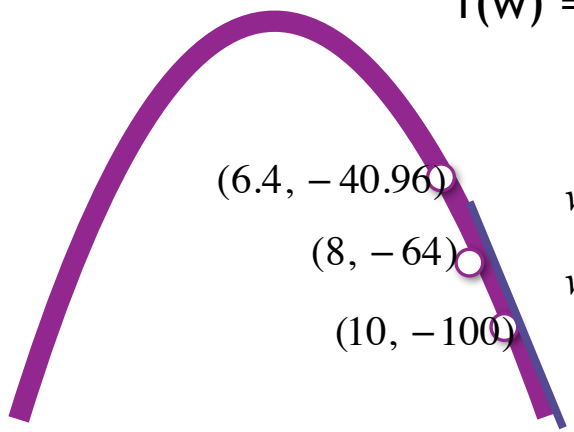
Maximize a function by repeatedly moving toward the maximum

Algorithm:
For i = 1 to num_iters:

if $f'(w) > 0$ then f is increasing,
move w a little to the right

if $f'(w) < 0$ then f is decreasing,
move w a little to the left

f(w) = -w² df/dw = -2w



$$w_{new} = w + \alpha \cdot (-2 \cdot w) \quad \alpha = 0.1$$

$$w_{new} = 10 + \alpha \cdot (-2 \cdot 10) = 10 - (0.1) \cdot 2 \cdot 10 = 8$$

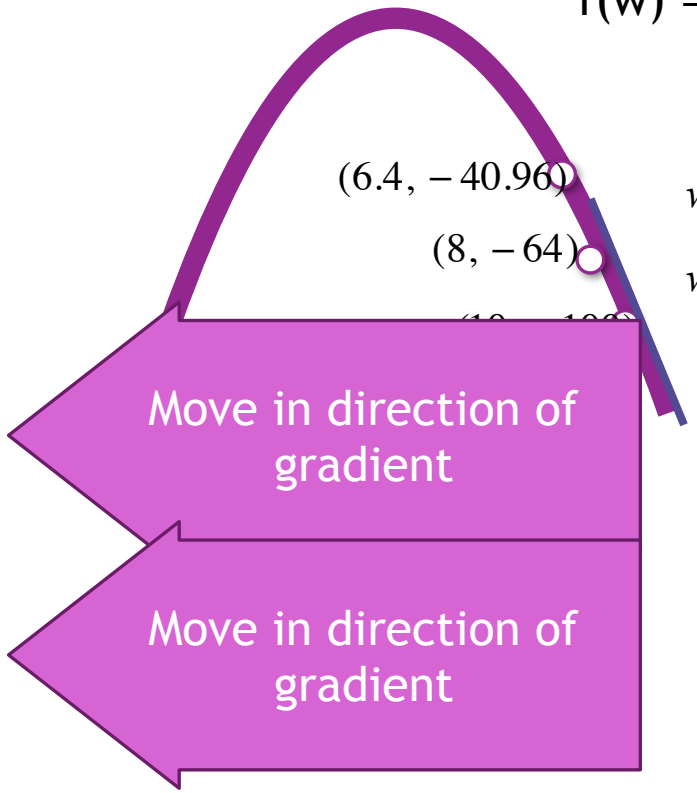
$$w_{new} = 8 - \alpha \cdot 2 \cdot 8 \quad = 8 - (0.1) \cdot 2 \cdot 8 = 6.4$$

Maximize a function by repeatedly moving toward the maximum

Algorithm:
For i = 1 to num_iters:
if $f'(w) > 0$ then f is increasing,
move w a little to the right

if $f'(w) < 0$ then f is decreasing,
move w a little to the left

f(w) = -w² df/dw = -2w



$$w_{new} = w + \alpha \cdot (-2 \cdot w) \quad \alpha = 0.1$$

$$w_{new} = 10 + \alpha \cdot (-2 \cdot 10) = 10 - (0.1) \cdot 2 \cdot 10 = 8$$

$$w_{new} = 8 - \alpha \cdot 2 \cdot 8 \quad = 8 - (0.1) \cdot 2 \cdot 8 = 6.4$$

The Gradient

- We wanted to maximize $\frac{1}{N}\ell(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}\left[y^{(i)}\ln\sigma(\mathbf{w}^T\mathbf{x}^{(i)}) + (1-y^{(i)})\ln\left(1-\sigma(\mathbf{w}^T\mathbf{x}^{(i)})\right)\right]$

- We would like to find where the derivative is zero (ie the most likely values for **w**.)
  We do this by taking steps towards the maximum value

$$w_j = w_j + \frac{\alpha}{N}\frac{\partial}{\partial w_j}\ell(\mathbf{w})$$

for k = 1 to num_iter

$$temp0 = w_0 + \frac{\alpha}{N}\frac{\partial\,\ell(\mathbf{w})}{\partial w_0}$$

$$temp1 = w_1 + \frac{\alpha}{N}\frac{\partial\,\ell(\mathbf{w})}{\partial w_1}$$

$$temp2 = w_2 + \frac{\alpha}{N}\frac{\partial\,\ell(\mathbf{w})}{\partial w_2}$$

$$w_0^{new} = temp0$$

$$w_1^{new} = temp1$$

$$w_2^{new} = temp2$$

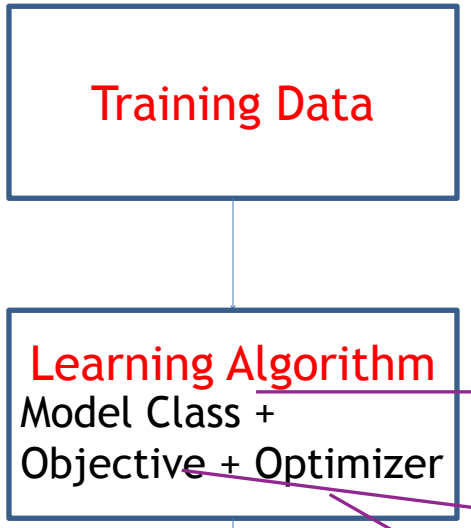
```
[1  3.2   4.7 ]
[1  3.2   4.5 ]
[1  3.1   4.9 ]
[1  2.3   4.  ]
[1  2.8   4.6 ]
[1  2.8   4.5 ]
[1  3.3   4.7 ]
[1  2.4   3.3 ]
[1  2.9   4.6 ]
[1  2.7   3.9 ]
```

Label 0 examples

```
[1  3.5   1.4 ]
[1  3.    1.4 ]
[1  3.2   1.3 ]
[1  3.1   1.5 ]
[1  3.6   1.4 ]
[1  3.9   1.7 ]
[1  3.4   1.4 ]
[1  3.4   1.5 ]
[1  2.9   1.4 ]
[1  3.1   1.5 ]
```

Label 1 examples

Supervised Learning



$$p(y = 1 \mid \mathbf{x}; \mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x})}$$

$$\frac{1}{N}\ell(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}\left[y^{(i)} \ln \sigma(\mathbf{w}^T\mathbf{x}^{(i)}) + (1 - y^{(i)}) \ln\left(1 - \sigma(\mathbf{w}^T\mathbf{x}^{(i)})\right)\right]$$

Gradient ascent

Input (say) patient data

x

Hypothesis or function

Output 0/1. 0 no cancer, 1 cancer

(Estimated probability of cancer)

Its just math…

$$h(\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1+e^{-(\mathbf{w}^T\mathbf{x})}}$$

Our update rule:

- We wanted to maximize $\frac{1}{N}\ell(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}\left[y^{(i)}\ln\sigma(\mathbf{w}^T\mathbf{x}^{(i)}) + (1-y^{(i)})\ln\left(1-\sigma(\mathbf{w}^T\mathbf{x}^{(i)})\right)\right]$
- We would like to find where the gradient is zero (ie the most likely values for **w**.) We do this by taking steps towards the maximum value $w_j = w_j + \frac{\alpha}{N}\frac{\partial}{\partial w_j}\ell(\mathbf{w})$ for each j

The derivative of the sigmoid function:

$$\frac{d}{dz}\left(\frac{1}{1+e^{-z}}\right) = \frac{1}{\left(1+e^{-z}\right)^2}e^{-z}$$

$$= \frac{1}{\left(1+e^{-z}\right)^2}(e^{-z}+1-1) \quad = \frac{1+e^{-z}}{\left(1+e^{-z}\right)^2} - \frac{1}{\left(1+e^{-z}\right)^2}$$

$$= \frac{1}{\left(1+e^{-z}\right)} - \frac{1}{\left(1+e^{-z}\right)^2} \quad = \frac{1}{\left(1+e^{-z}\right)}\left(1 - \frac{1}{\left(1+e^{-z}\right)}\right)$$

You will not be tested on this material with the pink background

Thus:

$$\sigma(z) = \left(\frac{1}{1+e^{-z}}\right)$$

$$\frac{d\sigma(z)}{dz} = \sigma(z)(1-\sigma(z))$$

Useful to know before we derive $\frac{\partial}{\partial w_j}\ell(\mathbf{w})$

$$\frac{\partial\, \mathbf{w}^T\mathbf{x}}{\partial w_j} = \frac{\partial\,(w_0x_0 + w_1x_1 + \cdots + w_jx_j + \cdots w_dx_d)}{\partial w_j} = x_j$$

$$\frac{\partial\sigma(\mathbf{w}^T\mathbf{x})}{\partial w_j} = \sigma(\mathbf{w}^T\mathbf{x})(1-\sigma(\mathbf{w}^T\mathbf{x}))\frac{\partial \mathbf{w}^T\mathbf{x}}{w_j} = \sigma(\mathbf{w}^T\mathbf{x})\left(1-\sigma(\mathbf{w}^T\mathbf{x})\right)x_j$$

$$\boxed{\frac{\partial\sigma(\mathbf{w}^T\mathbf{x})}{\partial w_j} = \sigma(\mathbf{w}^T\mathbf{x})\left(1-\sigma(\mathbf{w}^T\mathbf{x})\right)x_j}$$

The Gradient

$$\sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1+e^{-(\mathbf{w}^T\mathbf{x})}} \qquad \frac{\partial \sigma(\mathbf{w}^T\mathbf{x})}{\partial w_j} = \sigma(\mathbf{w}^T\mathbf{x})(1-\sigma(\mathbf{w}^T\mathbf{x}))\frac{\partial \mathbf{w}^T\mathbf{x}}{w_j} \qquad \frac{\partial \mathbf{w}^T\mathbf{x}}{\partial w_j} = x_j$$

$$\frac{\partial}{\partial w_j}\ell(\mathbf{w}) = \frac{\partial}{w_j}\sum_{i=1}^{N}\left[y^{(i)}\ln\sigma(\mathbf{w}^T\mathbf{x}^{(i)}) + (1-y^{(i)})\ln\left(1-\sigma(\mathbf{w}^T\mathbf{x}^{(i)})\right)\right] = \sum_{i=1}^{N}\left[y^{(i)}\frac{\partial}{w_j}\ln\sigma(\mathbf{w}^T\mathbf{x}^{(i)}) + (1-y^{(i)})\frac{\partial}{w_j}\ln\left(1-\sigma(\mathbf{w}^T\mathbf{x}^{(i)})\right)\right]$$

$$= \sum_{i=1}^{N}\left[y^{(i)}\frac{1}{\sigma(\mathbf{w}^T\mathbf{x}^{(i)})}\frac{\partial\,\sigma(\mathbf{w}^T\mathbf{x}^{(i)})}{w_j} - (1-y^{(i)})\frac{1}{1-\sigma(\mathbf{w}^T\mathbf{x}^{(i)})}\frac{\partial\,\sigma(\mathbf{w}^T\mathbf{x}^{(i)})}{w_j}\right] = \sum_{i=1}^{N}\left[y^{(i)}\frac{1}{\sigma(\mathbf{w}^T\mathbf{x}^{(i)})} - (1-y^{(i)})\frac{1}{1-\sigma(\mathbf{w}^T\mathbf{x}^{(i)})}\right]\frac{\partial\,\sigma(\mathbf{w}^T\mathbf{x}^{(i)})}{w_j}$$

$$= \sum_{i=1}^{N}\left[y^{(i)}\frac{1}{\sigma(\mathbf{w}^T\mathbf{x}^{(i)})} - (1-y^{(i)})\frac{1}{1-\sigma(\mathbf{w}^T\mathbf{x}^{(i)})}\right]\sigma(\mathbf{w}^T\mathbf{x}^{(i)})\left(1-\sigma(\mathbf{w}^T\mathbf{x}^{(i)})\right)\frac{\partial\,\mathbf{w}^T\mathbf{x}^{(i)}}{w_j} = \sum_{i=1}^{N}\left(y^{(i)}\left(1-\sigma(\mathbf{w}^T\mathbf{x}^{(i)})\right) - (1-y^{(i)})\sigma(\mathbf{w}^T\mathbf{x}^{(i)})\right)x_j^{(i)}$$

$$= \sum_{i=1}^{N}\left(y^{(i)}\cdot 1 - y^{(i)}\cdot\sigma(\mathbf{w}^T\mathbf{x}^{(i)}) - 1\cdot\sigma(\mathbf{w}^T\mathbf{x}^{(i)}) + y^{(i)}\sigma(\mathbf{w}^T\mathbf{x}^{(i)})\right)x_j^{(i)}$$

$$= \sum_{i=1}^{N}\left(y^{(i)} - \sigma(\mathbf{w}^T\mathbf{x}^{(i)})\right)x_j^{(i)}$$

Error of i^{th} training example

You will not be tested on this material with the purple background

Interpretation:

- We wanted to maximize $\frac{1}{N}\ell(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}\left[y^{(i)} \ln \sigma(\mathbf{w}^T\mathbf{x}^{(i)}) + (1-y^{(i)})\ln\left(1-\sigma(\mathbf{w}^T\mathbf{x}^{(i)})\right)\right]$

- We would like to find where the partial derivative is zero (ie the most likely values for **w**). We do this by taking steps towards the maximum value

$$w_j = w_j + \frac{\alpha}{N}\frac{\partial}{\partial w_j}\ell(\mathbf{w}) = w_j + \frac{\alpha}{N}\sum_{i=1}^{N}\left(y^{(i)} - \sigma(\mathbf{w}^T\mathbf{x}^{(i)})\right)x_j^{(i)}$$

- How does an example affect the update?
  - if y=1 and $\sigma(\mathbf{w}^T\mathbf{x})\approx 1$, then $(y-\sigma(\mathbf{w}^T\mathbf{x}))\approx 0$, almost no change!
  - if y=1 and $\sigma(\mathbf{w}^T\mathbf{x})\approx 0$, then $(y-\sigma(\mathbf{w}^T\mathbf{x}))\approx 1$, approx α/N times the j^{th} feature
  - if y=0 and $\sigma(\mathbf{w}^T\mathbf{x})\approx 0$, then $(y-\sigma(\mathbf{w}^T\mathbf{x}))\approx 0$, almost no change!
  - if y=0 and $\sigma(\mathbf{w}^T\mathbf{x})\approx 1$, then $(y-\sigma(\mathbf{w}^T\mathbf{x}))\approx -1$, approx $-\alpha/N$ times the j^{th} feature

$$\frac{\partial}{\partial w_j}\ell(\mathbf{w}) = \sum_{i=1}^{N}\left(y^{(i)} - \sigma(\mathbf{w}^T\mathbf{x}^{(i)})\right)x_j^{(i)}$$

Gradient Ascent Algorithm

$$\frac{1}{N}\ell(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}\left[y^{(i)}\ln\sigma(\mathbf{w}^T\mathbf{x}^{(i)}) + (1-y^{(i)})\ln\left(1-\sigma(\mathbf{w}^T\mathbf{x}^{(i)})\right)\right]$$

for k = 1 to num_iter

$$temp0 = w_0 + \alpha\frac{\partial\ell(\mathbf{w})/N}{\partial w_0} = w_0 + \frac{\alpha}{N}\sum_{i=1}^{N}(y^{(i)} - \sigma(\mathbf{w}^T\mathbf{x}^{(i)}))x_0^{(i)}$$

$$temp1 = w_1 + \alpha\frac{\partial\ell(\mathbf{w})/N}{\partial w_1} = w_1 + \frac{\alpha}{N}\sum_{i=1}^{N}(y^{(i)} - \sigma(\mathbf{w}^T\mathbf{x}^{(i)}))x_1^{(i)}$$

$$\vdots$$

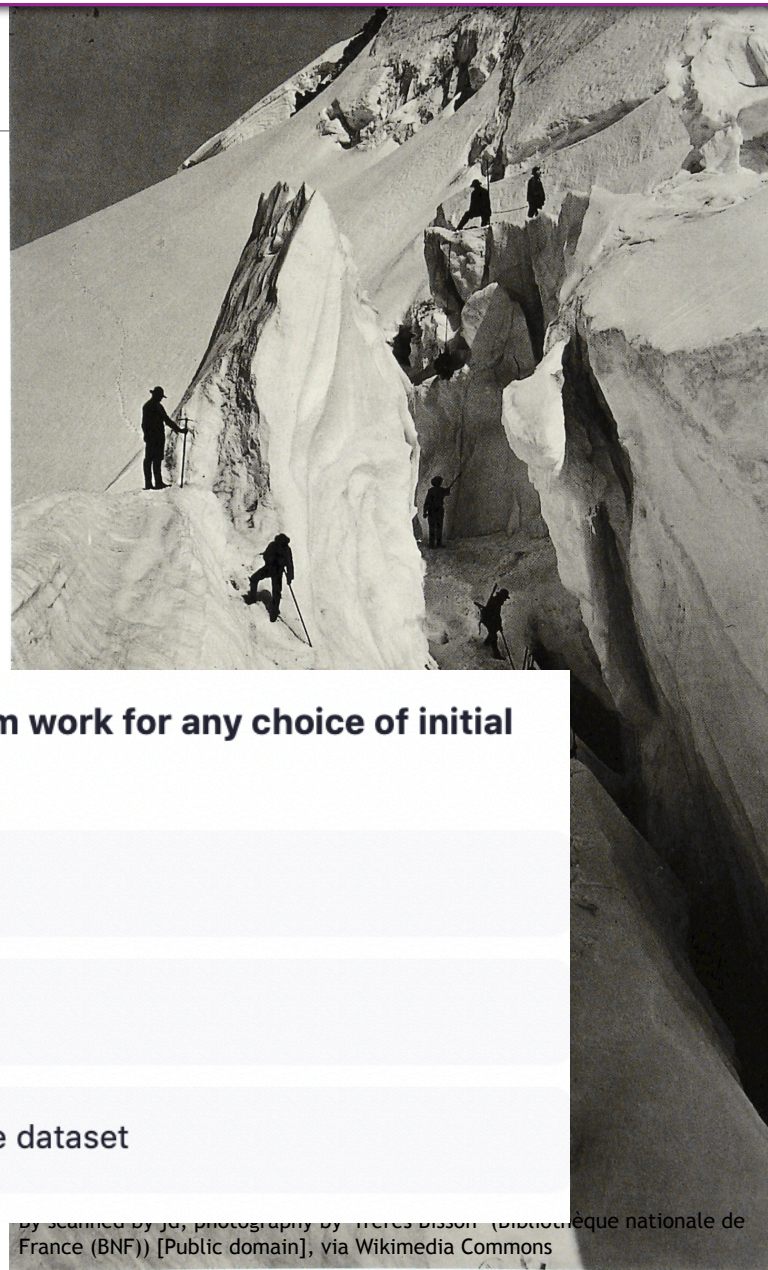
$$tempd = w_d + \alpha\frac{\partial\ell(\mathbf{w})/N}{\partial w_d} = w_d + \frac{\alpha}{N}\sum_{i=1}^{N}(y^{(i)} - \sigma(\mathbf{w}^T\mathbf{x}^{(i)}))x_d^{(i)}$$

$x^{(i)}_0=1$

$$w_0 = temp0$$
$$w_1 = temp1$$
$$\vdots$$
$$w_d = tempd$$

Simultaneous update

Running time:
$O(Nd \cdot$ # iter$)$



1. Does this algorithm work for any choice of initial values for w?

- Yes
- No
- It depends on the dataset

Gradient Ascent Step

$$X = \begin{bmatrix} 1 & 3 & 4 \\ 1 & 3 & 5 \\ 1 & 4 & 1 \\ 1 & 3 & 1.5 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \quad \mathbf{w} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow X\mathbf{w} = \begin{bmatrix} 1 & 3 & 4 \\ 1 & 3 & 5 \\ 1 & 4 & 1 \\ 1 & 3 & 1.5 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \hat{\mathbf{y}} = \begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{bmatrix}$$

for k = 1 to num_iter

$$temp0 = w_0 + \frac{\alpha}{N}\frac{\partial \ell(\mathbf{w})}{\partial w_0} = w_0 + \frac{\alpha}{N}\sum_{i=1}^{N}(y^{(i)} - \sigma(\mathbf{w}^T\mathbf{x}^{(i)}))x_0^{(i)} = w_0 + \frac{0.2}{4}\left[\left(0 - \sigma\left(\mathbf{w}^T\mathbf{x}^{(1)}\right)\right) + \left(0 - \sigma\left(\mathbf{w}^T\mathbf{x}^{(2)}\right)\right) + \left(1 - \sigma\left(\mathbf{w}^T\mathbf{x}^{(3)}\right)\right) + \left(1 - \sigma\left(\mathbf{w}^T\mathbf{x}^{(4)}\right)\right)\right]$$

-0.5 -0.5 0.5 0.5

$$temp1 = w_1 + \frac{\alpha}{N}\frac{\partial \ell(\mathbf{w})}{\partial w_1} = w_1 + \frac{\alpha}{N}\sum_{i=1}^{N}(y^{(i)} - \sigma(\mathbf{w}^T\mathbf{x}^{(i)}))x_1^{(i)} = w_1 + \frac{0.2}{4}\left[\left(0 - \sigma\left(\mathbf{w}^T\mathbf{x}^{(1)}\right)\right)x_1^{(1)} + \left(0 - \sigma\left(\mathbf{w}^T\mathbf{x}^{(2)}\right)\right)x_1^{(2)} + \left(1 - \sigma\left(\mathbf{w}^T\mathbf{x}^{(3)}\right)x_1^{(3)}\right) + \left(1 - \sigma\left(\mathbf{w}^T\mathbf{x}^{(4)}\right)x_1^{(4)}\right)\right]$$

(-0.5)3 (-0.5)3 (0.5)4 (0.5)3

$$temp2 = w_2 + \frac{\alpha}{N}\frac{\partial \ell(\mathbf{w})}{\partial w_2} = w_2 + \frac{\alpha}{N}\sum_{i=1}^{N}(y^{(i)} - \sigma(\mathbf{w}^T\mathbf{x}^{(i)}))x_2^{(i)} = w_2 + \frac{0.2}{4}\left[\left(0 - \sigma\left(\mathbf{w}^T\mathbf{x}^{(1)}\right)\right)x_2^{(1)} + \left(0 - \sigma\left(\mathbf{w}^T\mathbf{x}^{(2)}\right)\right)x_2^{(2)} + \left(1 - \sigma\left(\mathbf{w}^T\mathbf{x}^{(3)}\right)x_2^{(3)}\right) + \left(1 - \sigma\left(\mathbf{w}^T\mathbf{x}^{(4)}\right)x_2^{(4)}\right)\right]$$

(-0.5)4 (-0.5)5 (0.5)1 (0.5)(1.5)

$w_0^{new} = temp0$ 0

$w_1^{new} = temp1$ 0.025

$w_2^{new} = temp2$ -0.1625

Gradient Ascent Example

$$X = \begin{bmatrix} 1 & 3 & 4 \\ 1 & 3 & 5 \\ 1 & 4 & 1 \\ 1 & 3 & 1.5 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \quad \mathbf{w} = \begin{bmatrix} 0 \\ 0.025 \\ -0.1625 \end{bmatrix} \Rightarrow X\mathbf{w} = \begin{bmatrix} 1 & 3 & 4 \\ 1 & 3 & 5 \\ 1 & 4 & 1 \\ 1 & 3 & 1.5 \end{bmatrix} \begin{bmatrix} 0 \\ 0.025 \\ -0.1625 \end{bmatrix} = \begin{bmatrix} -0.575 \\ -0.7375 \\ -0.0625 \\ -0.16875 \end{bmatrix} \Rightarrow \hat{\mathbf{y}} = \begin{bmatrix} 0.3601 \\ 0.3236 \\ 0.4844 \\ 0.4579 \end{bmatrix}$$

for k = 1 to num_iter

$$temp0 = w_0 + \frac{\alpha}{N}\frac{\partial \ell(\mathbf{w})}{\partial w_0} = w_0 + \frac{\alpha}{N}\sum_{i=1}^{N}(y^{(i)} - \sigma(\mathbf{w}^T\mathbf{x}^{(i)}))x_0^{(i)} = w_0 + \frac{0.2}{4}\left[\left(0 - \sigma\left(\mathbf{w}^T\mathbf{x}^{(1)}\right)\right) + \left(0 - \sigma\left(\mathbf{w}^T\mathbf{x}^{(2)}\right)\right) + \left(1 - \sigma\left(\mathbf{w}^T\mathbf{x}^{(3)}\right)\right) + \left(1 - \sigma\left(\mathbf{w}^T\mathbf{x}^{(4)}\right)\right)\right]$$

(0-0.3601) (0-0.3236) (1-0.4844) (1-0.4579)

$$temp1 = w_1 + \frac{\alpha}{N}\frac{\partial \ell(\mathbf{w})}{\partial w_1} = w_1 + \frac{\alpha}{N}\sum_{i=1}^{N}(y^{(i)} - \sigma(\mathbf{w}^T\mathbf{x}^{(i)}))x_1^{(i)} w_1 + \frac{0.2}{4}\left[\left(0 - \sigma\left(\mathbf{w}^T\mathbf{x}^{(1)}\right)\right)x_1^{(1)} + \left(0 - \sigma\left(\mathbf{w}^T\mathbf{x}^{(2)}\right)\right)x_1^{(2)} + \left(1 - \sigma\left(\mathbf{w}^T\mathbf{x}^{(3)}\right)x_1^{(3)}\right) + \left(1 - \sigma\left(\mathbf{w}^T\mathbf{x}^{(4)}\right)x_1^{(4)}\right)\right]$$

(0-0.3601)*3 (0-0.3236)*3 (1-0.4844)*4 (1-0.4579)*3

$$temp2 = w_2 + \frac{\alpha}{N}\frac{\partial \ell(\mathbf{w})}{\partial w_2} = w_2 + \frac{\alpha}{N}\sum_{i=1}^{N}(y^{(i)} - \sigma(\mathbf{w}^T\mathbf{x}^{(i)}))x_2^{(i)} = w_2 + \frac{0.2}{4}\left[\left(0 - \sigma\left(\mathbf{w}^T\mathbf{x}^{(1)}\right)\right)x_2^{(1)} + \left(0 - \sigma\left(\mathbf{w}^T\mathbf{x}^{(2)}\right)\right)x_2^{(2)} + \left(1 - \sigma\left(\mathbf{w}^T\mathbf{x}^{(3)}\right)x_2^{(3)}\right) + \left(1 - \sigma\left(\mathbf{w}^T\mathbf{x}^{(4)}\right)x_2^{(4)}\right)\right]$$

(0-0.3601)*4 (0-0.3236)*5 (1-0.4844)*1 (1-0.4579)*(1.5)

$w_0^{new} = temp0$ 0.019

$w_1^{new} = temp1$ 0.107

$w_2^{new} = temp2$ -0.249