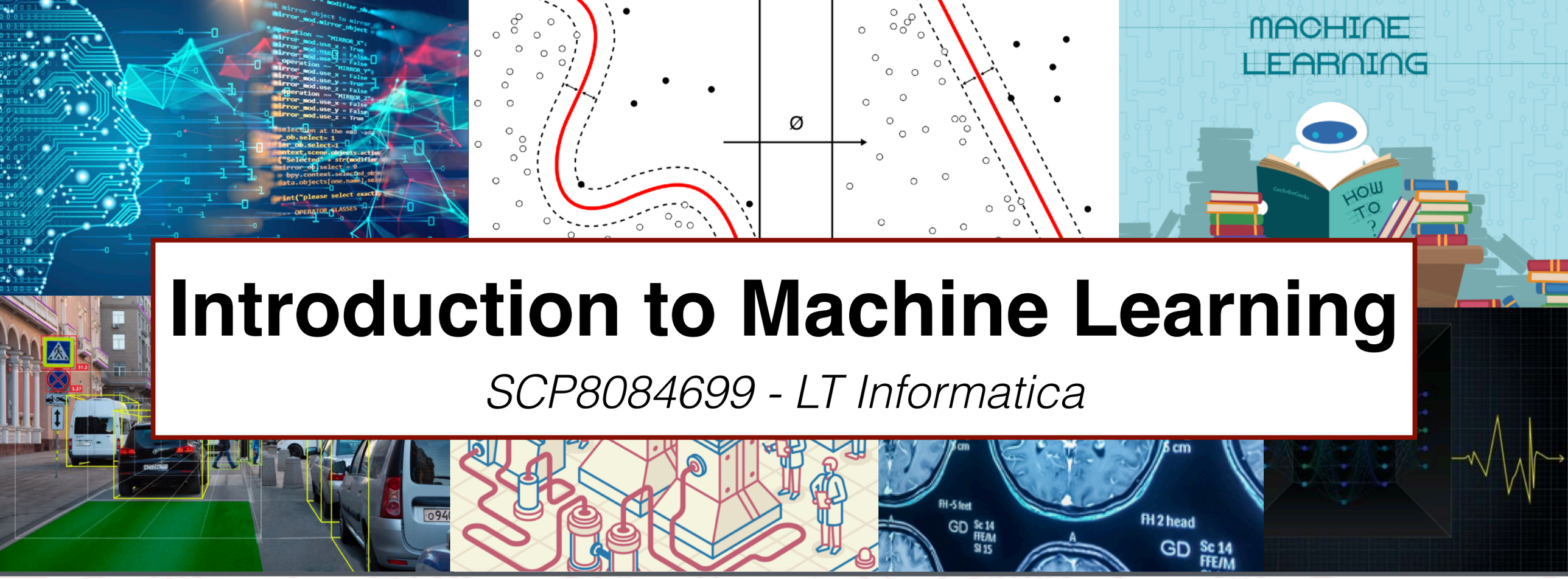


Introduction to Machine Learning

SCP8084699 - LT Informatica

Logistic Regression

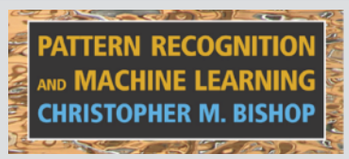
Prof. Lamberto Ballan

University of Padova, A.Y. 2022/23 March 23, 2023

DIPARTIMENTO MATEMATICA
Department of Mathematics “Tullio Levi-Civita”

Logistic Regression

- Applying linear regression to classification tasks usually is not a great idea
- A better approach is to use *logistic regression*
  - Note: although the term regression appears in its name, logistic regression is a classification algorithm
  - It has also a nice property: $0 \leq h_\theta(x) \leq 1$



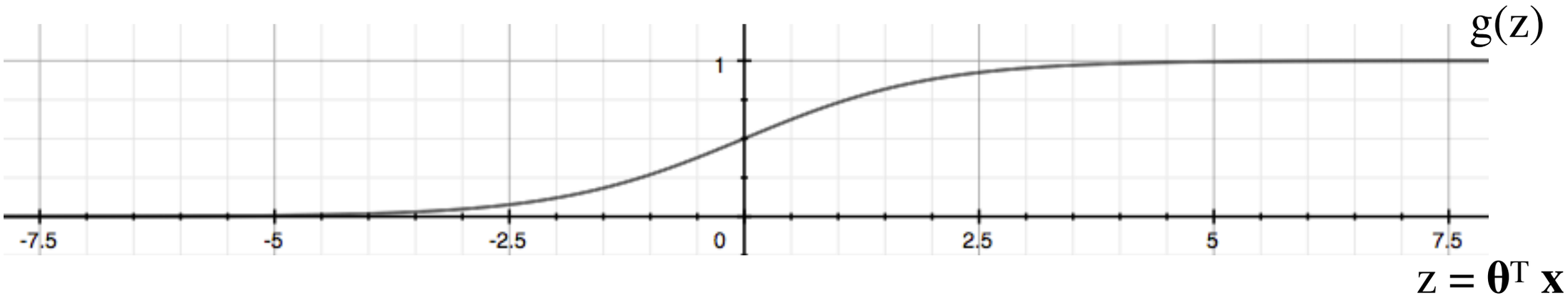
Pages 203-207

Logistic Regression

- Hypothesis representation:

$$h_\theta(\mathbf{x}) = g(\boldsymbol{\theta}^T \mathbf{x}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^T \mathbf{x}}}$$

where $g(z) = \frac{1}{1 + e^{-z}}$ (*Sigmoid* or *Logistic* function)



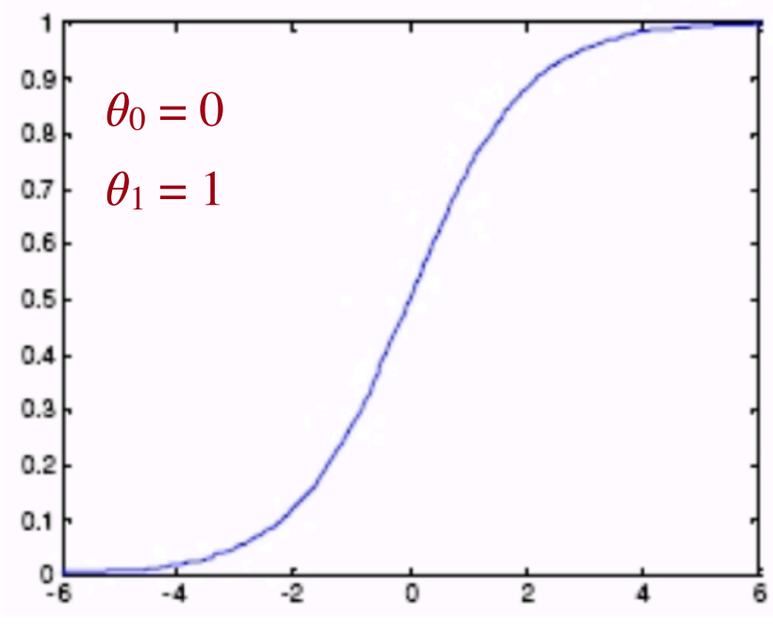
Logistic Regression

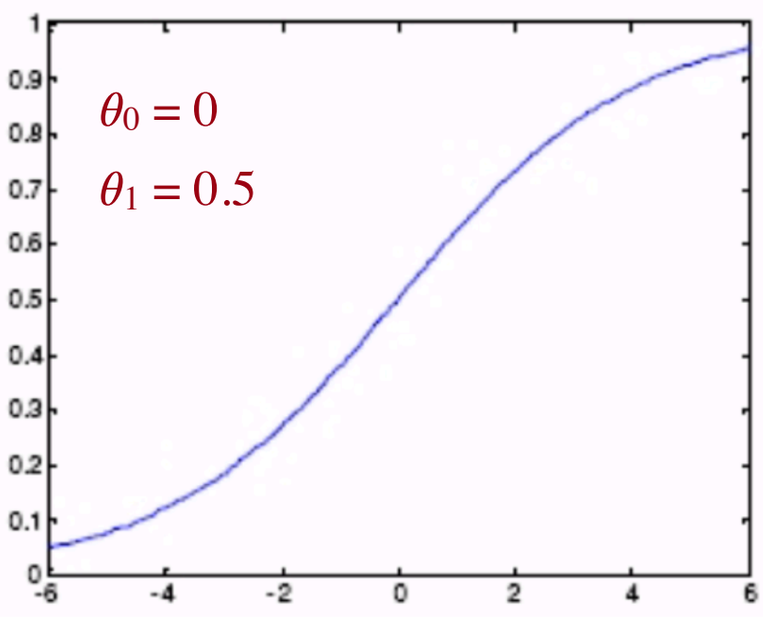
- A bit more about the shape of the logistic function:

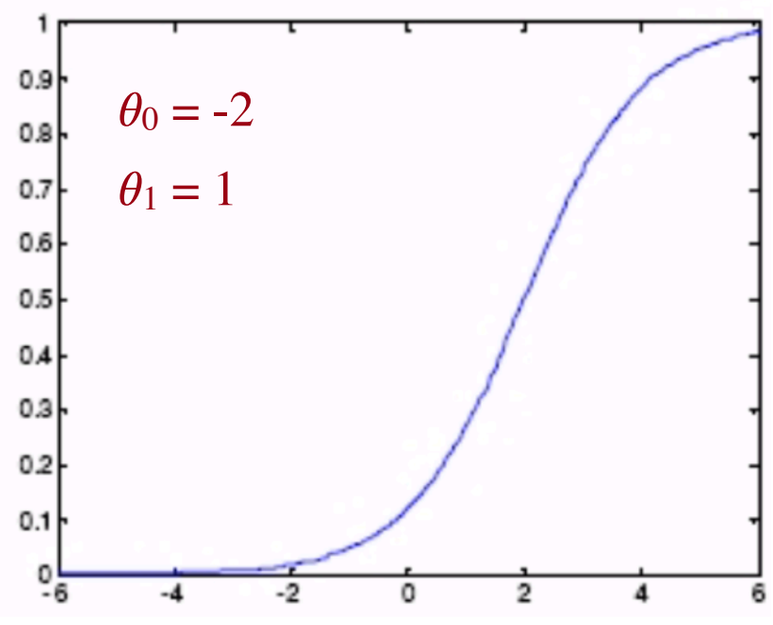
$$h_\theta(\mathbf{x}) = g(\boldsymbol{\theta}^T \mathbf{x}) \qquad \text{where} \qquad g(z) = \sigma(z) = \frac{1}{1 + e^{-z}}$$

(*Sigmoid* or *Logistic* function)

1D example: $h_\theta(\mathbf{x}) = \dfrac{1}{1 + e^{-(\theta_0 + \theta_1 x)}}$







Probabilistic Interpretation

- Interpretation of hypothesis output:
  - $h_\theta(\mathbf{x})$ = estimated probability that $\mathrm{y}=1$ on input $\mathbf{x}$
  - More formally: $h_\theta(\mathbf{x}) = \mathrm{P}\,(\mathrm{y}=1 \mid \mathrm{x};\, \theta)$

- An example:

$$\mathbf{x} = \begin{bmatrix} \mathrm{x}_0 \\ \mathrm{x}_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \mathrm{tumor\_size} \end{bmatrix} \qquad h_\theta(\mathbf{x}) = 0.7$$

Tell patient that 70% chance of tumor being malignant

Probabilistic Interpretation

- Interpretation of hypothesis output:
  - $h_\theta(\mathbf{x})$ = estimated probability that $\mathrm{y}=1$ on input $\mathbf{x}$
  - More formally: $h_\theta(\mathbf{x}) = \mathrm{P}\,(\mathrm{y}=1 \mid \mathrm{x};\, \theta)$

- If we have two classes, what about $\mathrm{P}\,(\mathrm{y}=0 \mid \mathrm{x};\, \theta)$?
  - *Marginalization* property: $\mathrm{P}\,(\mathrm{y}=1 \mid \mathrm{x};\, \theta) + \mathrm{P}\,(\mathrm{y}=0 \mid \mathrm{x};\, \theta) = 1$

    therefore $\mathrm{P}\,(\mathrm{y}=0 \mid \mathrm{x};\, \theta) = 1 - \mathrm{P}\,(\mathrm{y}=1 \mid \mathrm{x};\, \theta)$

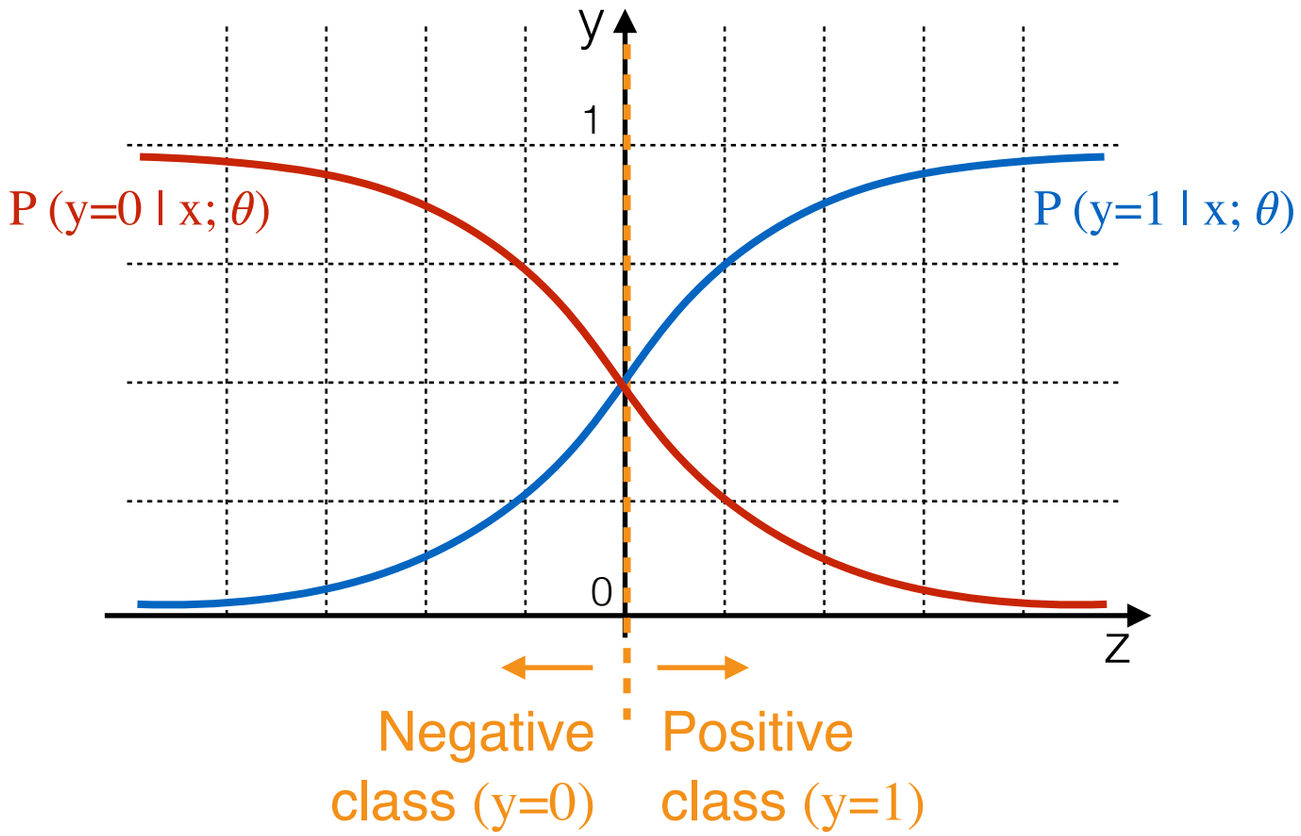
    i.e. $\mathrm{P}\,(\mathrm{y}=0 \mid \mathrm{x};\, \theta) = 1 - \dfrac{1}{1 + e^{-\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}}} = \dfrac{e^{-\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}}}{1 + e^{-\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}}}$

Decision Boundary

- What is the decision boundary for logistic regression?

$$h_\theta(\mathbf{x}) = g(\boldsymbol{\theta}^T \mathbf{x})$$

where $g(z) = \dfrac{1}{1 + e^{-z}}$



$$h_\theta(x) = P\,(y=1 \mid x;\, \theta)$$

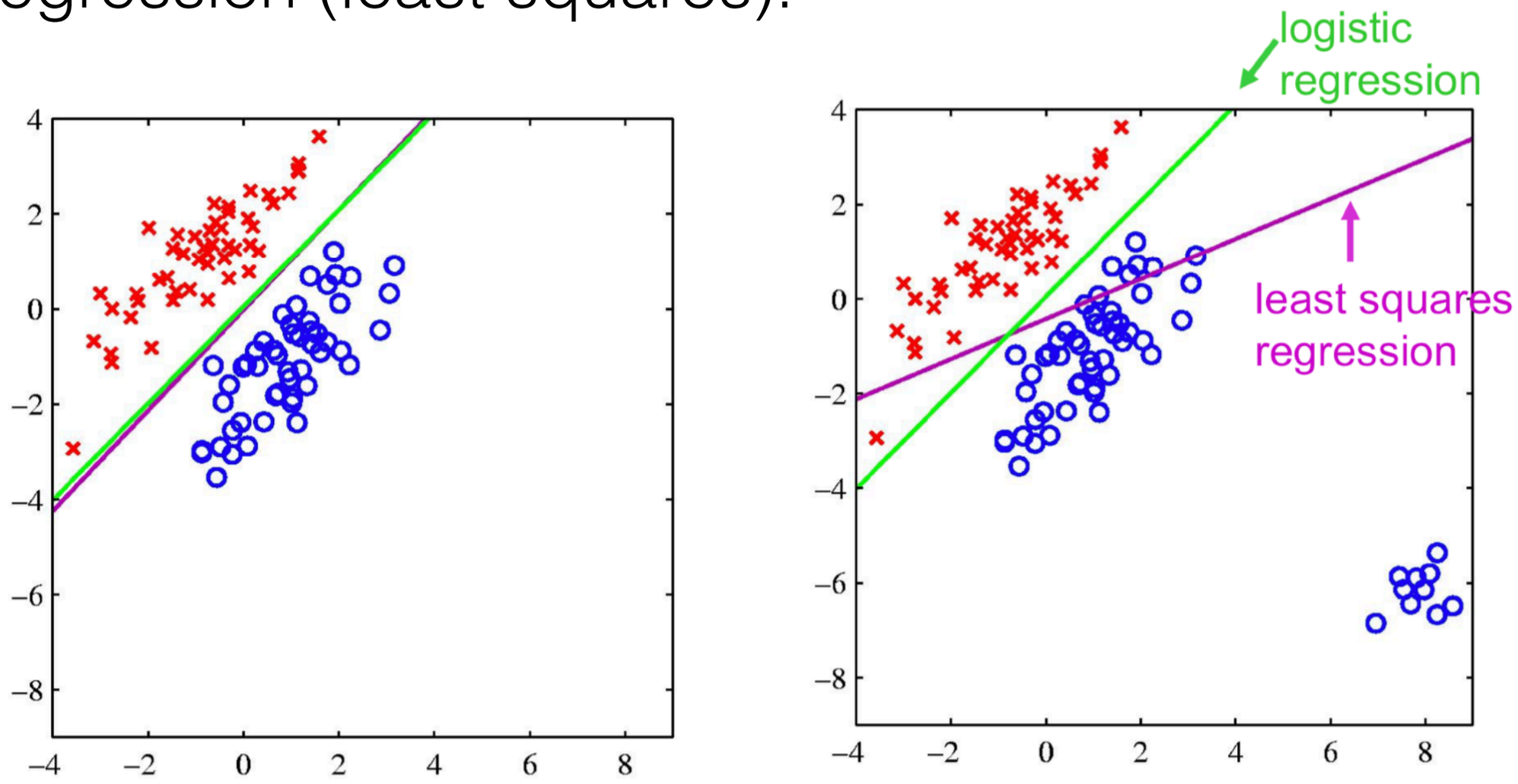
Suppose predict y=1 if $h_\theta(x) \geq 0.5$

predict y=0 if $h_\theta(x) < 0.5$

Logistic Regression has a linear decision boundary

Logistic *vs* Linear Regression

- A qualitative example of logistic regression vs linear regression (least squares):



If the right answer is 1 and the model says 1.5, it loses, so it changes the boundary to avoid being "too correct" (tilts away from outliers)

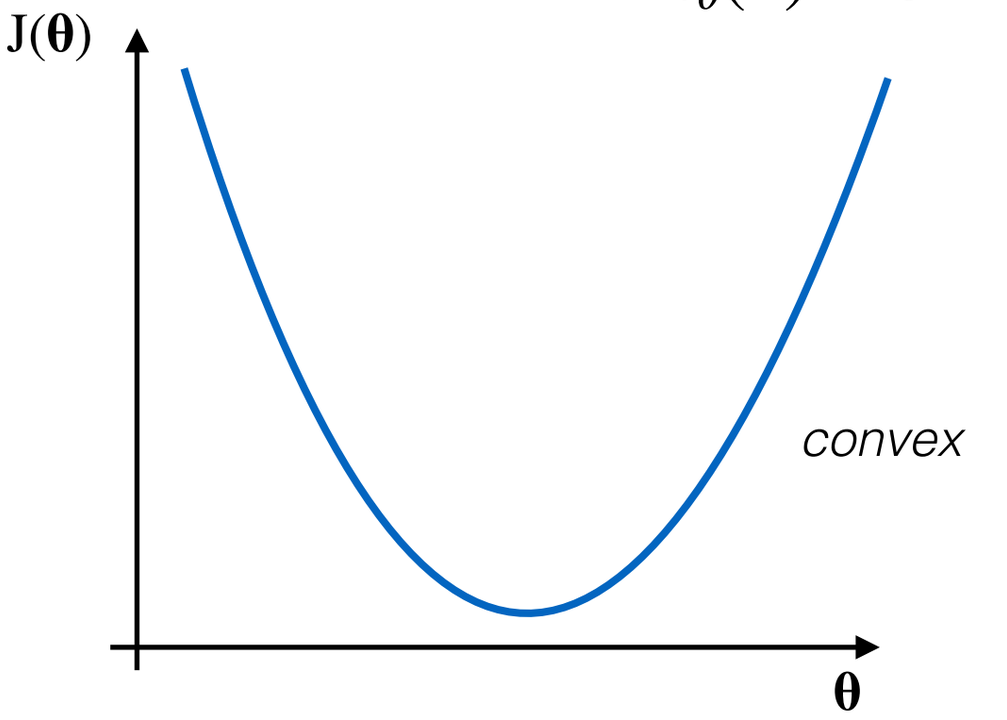
Logistic *vs* Linear Regression

- Loss function: $\mathrm{J}(\boldsymbol{\theta}) = \frac{1}{\mathrm{m}} \sum_{\mathrm{i}=1}^{\mathrm{m}} cost(h_\theta(\mathrm{x}^{(\mathrm{i})}), \mathrm{y}^{(\mathrm{i})})$

where $cost(h_\theta(\mathrm{x}^{(\mathrm{i})}), \mathrm{y}^{(\mathrm{i})}) = \frac{1}{2} (h_\theta(\mathrm{x}^{(\mathrm{i})}) - \mathrm{y}^{(\mathrm{i})})^2$

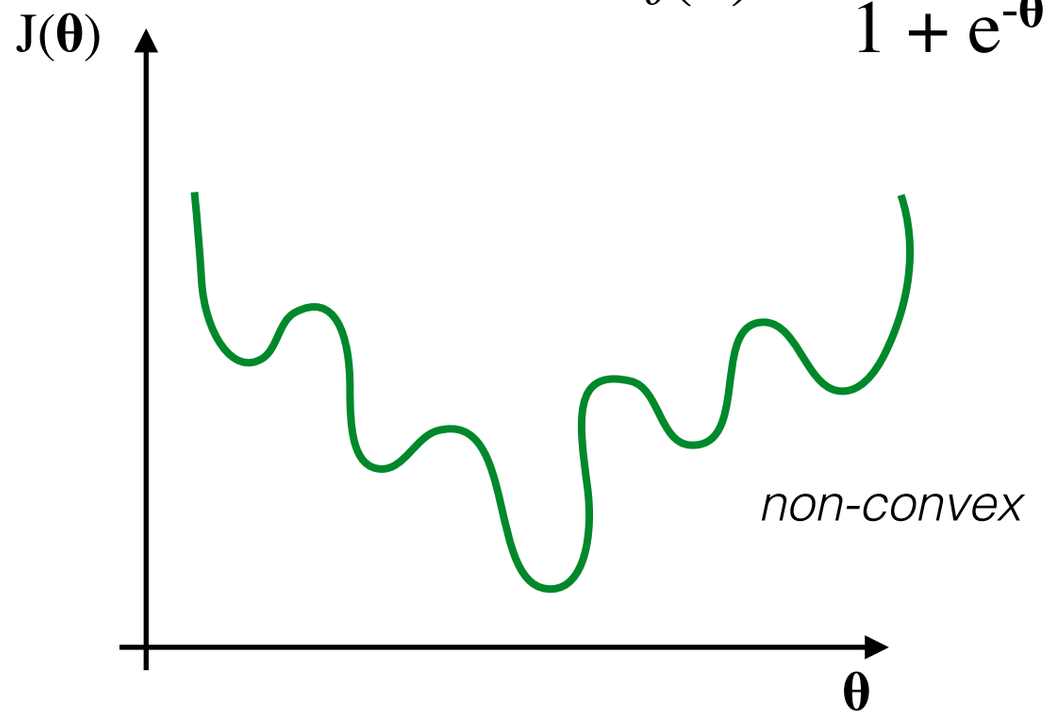
Linear Regression

$h_\theta(\mathbf{x}) = \boldsymbol{\theta}^{\mathrm{T}} \mathbf{x}$



Logistic Regression

$h_\theta(\mathbf{x}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^{\mathrm{T}} \mathbf{x}}}$

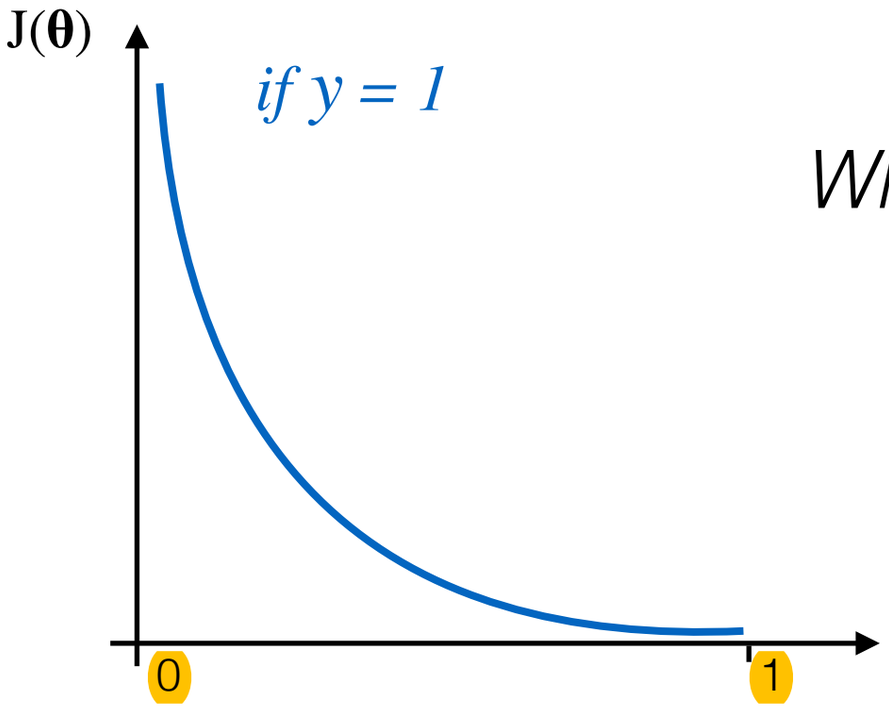


Logistic Regression Loss Function

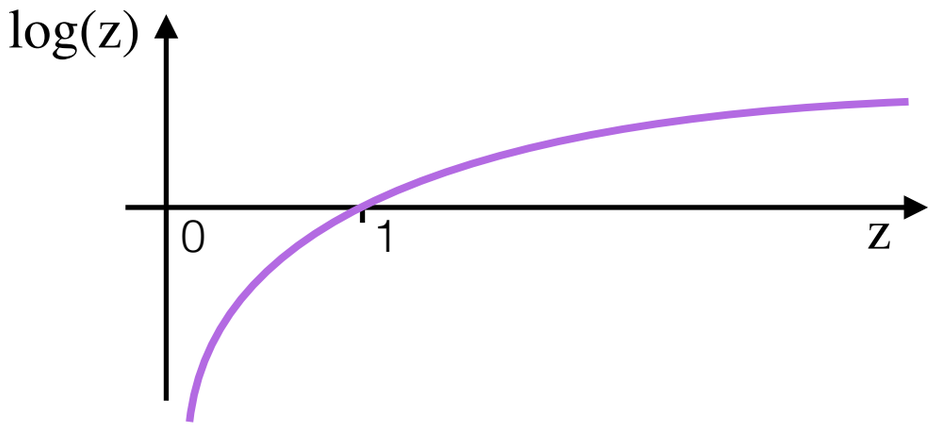
- Loss function: $J(\boldsymbol{\theta}) = \frac{1}{m} \sum_{i=1}^{m} cost(h_\theta(x^{(i)}), y^{(i)})$

where $cost(h_\theta(x^{(i)}), y^{(i)}) = \begin{cases} -\log(h_\theta(x^{(i)})) & \text{if } y^{(i)} = 1 \\ -\log(1 - h_\theta(x^{(i)})) & \text{if } y^{(i)} = 0 \end{cases}$

- Intuition:



Why is that?



$cost = 0$ if $y^{(i)} = 1$ and $h_\theta(x^{(i)}) = 1$

$cost \to \infty$ if $h_\theta(x^{(i)}) \to 0$ (and $y^{(i)} = 1$)

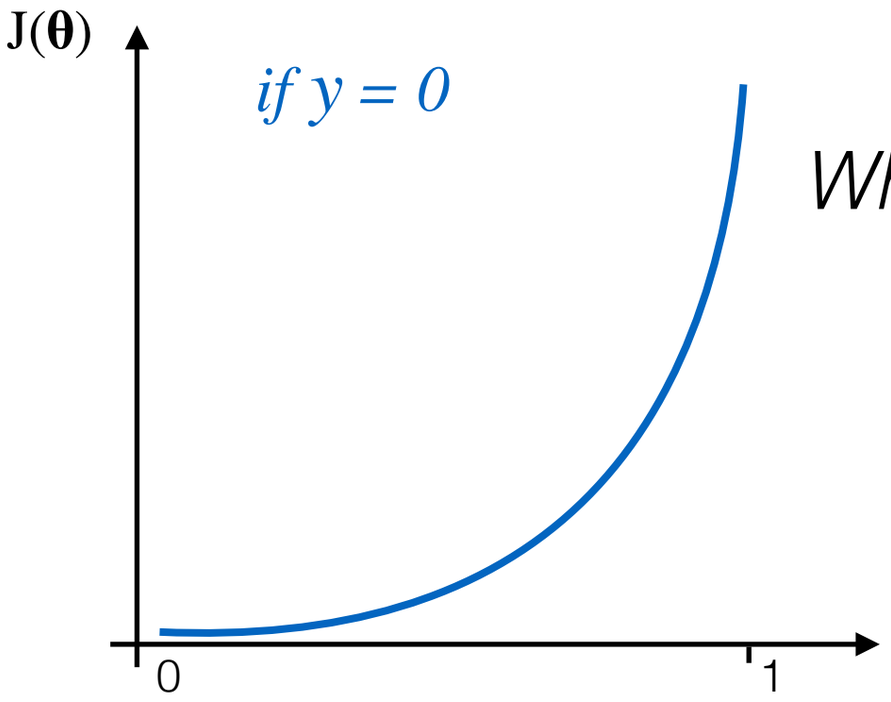
(i.e. predict $P(y=1 \mid x; \theta) = 0$ but $y=1$)

Logistic Regression Loss Function

- Loss function: $\mathrm{J}(\boldsymbol{\theta}) = \frac{1}{\mathrm{m}} \sum_{\mathrm{i}=1}^{\mathrm{m}} cost(h_\theta(\mathrm{x}^{(\mathrm{i})}), \mathrm{y}^{(\mathrm{i})})$

$$\textit{where} \quad cost(h_\theta(\mathrm{x}^{(\mathrm{i})}), \mathrm{y}^{(\mathrm{i})}) = \begin{cases} -\log(h_\theta(\mathrm{x}^{(\mathrm{i})})) & \text{if } \mathrm{y}^{(\mathrm{i})} = 1 \\ -\log(1 - h_\theta(\mathrm{x}^{(\mathrm{i})})) & \text{if } \mathrm{y}^{(\mathrm{i})} = 0 \end{cases}$$

- Intuition:



Why is that?

$cost = 0$ if $\mathrm{y}^{(\mathrm{i})} = 0$ and $h_\theta(\mathrm{x}^{(\mathrm{i})}) = 0$

$cost \rightarrow \infty$ if $h_\theta(\mathrm{x}^{(\mathrm{i})}) \rightarrow 1$ (and $\mathrm{y}^{(\mathrm{i})} = 0$)

(i.e. predict P (y=0 | x; θ) = 1 but y=0)

Logistic Regression Loss Function

- Loss function: $\mathrm{J}(\boldsymbol{\theta}) = \frac{1}{\mathrm{m}} \sum_{\mathrm{i}=1}^{\mathrm{m}} cost(h_\theta(\mathrm{x}^{(\mathrm{i})}), \mathrm{y}^{(\mathrm{i})})$

$$\textit{where} \quad cost(h_\theta(\mathrm{x}^{(\mathrm{i})}), \mathrm{y}^{(\mathrm{i})}) = \begin{cases} -\log(h_\theta(\mathrm{x}^{(\mathrm{i})})) & \text{if } \mathrm{y}^{(\mathrm{i})} = 1 \\ -\log(1 - h_\theta(\mathrm{x}^{(\mathrm{i})})) & \text{if } \mathrm{y}^{(\mathrm{i})} = 0 \end{cases}$$

  - Note: by definition $\mathrm{y}=1$ or $\mathrm{y}=0$ (binary classifier)

  - "Simplified notation":

$$cost(h_\theta(\mathrm{x}^{(\mathrm{i})}), \mathrm{y}^{(\mathrm{i})}) = -\mathrm{y}^{(\mathrm{i})} \cdot \log(h_\theta(\mathrm{x}^{(\mathrm{i})})) - (1 - \mathrm{y}^{(\mathrm{i})}) \cdot \log(1 - h_\theta(\mathrm{x}^{(\mathrm{i})}))$$

$$\Rightarrow \quad \mathrm{J}(\boldsymbol{\theta}) = -\frac{1}{\mathrm{m}} \sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{y}^{(\mathrm{i})} \cdot \log(h_\theta(\mathrm{x}^{(\mathrm{i})})) + (1 - \mathrm{y}^{(\mathrm{i})}) \cdot \log(1 - h_\theta(\mathrm{x}^{(\mathrm{i})}))$$

This is a convex function!

Parameter Learning

- We can learn our parameters with gradient descent

$$J(\boldsymbol{\theta}) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \cdot \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \cdot \log(1 - h_\theta(x^{(i)}))$$

$$\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

Note: this is usually referred as to "cross-entropy loss" or "log-loss"

```
repeat until convergence{
```

$$\theta_j := \theta_j - \eta \frac{\partial}{\partial \theta_j} J(\boldsymbol{\theta}) = \theta_j - \frac{\eta}{m}\sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})\, x_j^{(i)}$$

(simultaneously update all θ_j)

```
}
```

Logistic Regression - Update Rule

- The (gradient descent) update rule is exactly the same for both linear and logistic regression
  - That's great…. but how is it possible?
  - Let's take a look at the derivative of cost function for logistic regression

Logistic Regression - Update Rule

- We need to figure out what is the derivative $\frac{\partial}{\partial \theta_j} \mathbf{J}(\boldsymbol{\theta})$

Cost function $$\mathbf{J}(\boldsymbol{\theta}) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \cdot \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \cdot \log(1 - h_\theta(x^{(i)}))\right]$$

where $h_\theta(\mathbf{x}) = g(\boldsymbol{\theta}^T \mathbf{x})$ and $g(z) = \sigma(z) = \frac{1}{1 + e^{-z}}$

- Let's start by computing the derivative of $\sigma(z)$

$$\frac{d\,\sigma(z)}{dz} = \frac{d}{dz}\frac{f(z) = 1}{g(z) = 1 + e^{-z}}$$

Quotient rule

$$\frac{d}{dx}\frac{f(x)}{g(x)} = \frac{f'g - f\,g'}{g^2}$$

Logistic Regression - Update Rule

- We need to figure out what is the derivative $\frac{\partial}{\partial \theta_j} \mathbf{J}(\boldsymbol{\theta})$

Cost function $$\mathbf{J}(\boldsymbol{\theta}) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \cdot \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \cdot \log(1 - h_\theta(x^{(i)}))\right]$$

where $h_\theta(\mathbf{x}) = g(\boldsymbol{\theta}^T \mathbf{x})$ and $g(z) = \sigma(z) = \frac{1}{1 + e^{-z}}$

- Let's start by computing the derivative of $\sigma(z)$

$$\frac{d\,\sigma(z)}{dz} = \frac{0 \cdot (1 + e^{-z}) - (1) \cdot (e^{-z} \cdot (-1))}{(1 + e^{-z})^2} = \frac{(e^{-z})}{(1 + e^{-z})^2} = \frac{1-1+(e^{-z})}{(1 + e^{-z})^2} =$$

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

Logistic Regression - Update Rule

- We need to figure out what is the derivative $\frac{\partial}{\partial \theta_j} \mathbf{J}(\boldsymbol{\theta})$

Cost function $$\mathbf{J}(\boldsymbol{\theta}) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \cdot \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \cdot \log(1 - h_\theta(x^{(i)}))\right]$$

where $h_\theta(\mathbf{x}) = g(\boldsymbol{\theta}^T \mathbf{x})$ and $g(z) = \sigma(z) = \frac{1}{1 + e^{-z}}$

- Writing now in terms of partial derivatives:

$$\frac{\partial}{\partial \theta_j} \mathbf{J}(\boldsymbol{\theta}) =$$

$$f(x) = \log(x)$$
$$g(x) = h_\theta(x)$$

$$\frac{d}{dx} \log(x) = \frac{1}{x}$$

Chain rule

$$\frac{d}{dx} f(\, g(x)\,) = f'(g(x))\, g'(x)$$

Logistic Regression - Update Rule

- We need to figure out what is the derivative $\frac{\partial}{\partial \theta_j} \mathbf{J}(\boldsymbol{\theta})$

Cost function $$\mathbf{J}(\boldsymbol{\theta}) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \cdot \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \cdot \log(1 - h_\theta(x^{(i)}))\right]$$

where $h_\theta(\mathbf{x}) = g(\boldsymbol{\theta}^T \mathbf{x})$ and $g(z) = \sigma(z) = \frac{1}{1 + e^{-z}}$

- Writing now in terms of partial derivatives:

$$\frac{\partial}{\partial \theta_j} \mathbf{J}(\boldsymbol{\theta}) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \cdot \frac{1}{h_\theta(x^{(i)})} \cdot \frac{\partial}{\partial \theta_j} h_\theta(x^{(i)}) + \right.$$

$$\left. + (1 - y^{(i)}) \cdot \frac{1}{(1 - h_\theta(x^{(i)}))} \cdot \frac{\partial}{\partial \theta_j} (1 - h_\theta(x^{(i)}))\right]$$

Logistic Regression - Update Rule

- Writing now in terms of partial derivatives: $\frac{\partial}{\partial \theta_j} J(\boldsymbol{\theta}) =$

$$= -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \cdot \frac{1}{h_\theta(x^{(i)})} \cdot \frac{\partial}{\partial \theta_j} h_\theta(x^{(i)}) + (1 - y^{(i)}) \cdot \frac{1}{(1 - h_\theta(x^{(i)}))} \cdot \frac{\partial}{\partial \theta_j} (1 - h_\theta(x^{(i)}))\right] =$$

plugging in our previous results (and using the derivative pattern of sigmoids)

$$= -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \cdot \frac{1}{h_\theta(x^{(i)})} \cdot \sigma(z) \cdot (1 - \sigma(z)) \cdot \frac{\partial}{\partial \theta_j} (\boldsymbol{\theta}^T x) + (1 - y^{(i)}) \cdot \frac{1}{(1 - h_\theta(x^{(i)}))} \cdot \right.$$

$$\left. \cdot (-\sigma(z)) \cdot (1 - \sigma(z)) \cdot \frac{\partial}{\partial \theta_j} (\boldsymbol{\theta}^T x)\right] = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \cdot \frac{1}{h_\theta(x^{(i)})} \cdot h_\theta(x^{(i)}) \cdot (1 - h_\theta(x^{(i)})) \cdot x_j^{(i)} + \right.$$

$$\left. + (1 - y^{(i)}) \cdot \frac{1}{(1 - h_\theta(x^{(i)}))} \cdot (-h_\theta(x^{(i)})) \cdot (1 - h_\theta(x^{(i)})) \cdot x_j^{(i)}\right]$$

Logistic Regression - Update Rule

- Simplifying the terms by multiplication:

$$\frac{\partial}{\partial \theta_j} J(\boldsymbol{\theta}) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \cdot \frac{1}{h_\theta(x^{(i)})} \cdot h_\theta(x^{(i)}) \cdot (1 - h_\theta(x^{(i)})) \cdot x_j^{(i)} + \right.$$

$$\left. + (1 - y^{(i)}) \cdot \frac{1}{(1 - h_\theta(x^{(i)}))} \cdot (-h_\theta(x^{(i)})) \cdot (1 - h_\theta(x^{(i)})) \cdot x_j^{(i)} \right] =$$

$$= -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \cdot (1 - h_\theta(x^{(i)})) \cdot x_j^{(i)} - (1 - y^{(i)}) \cdot h_\theta(x^{(i)}) \cdot x_j^{(i)}\right] =$$

$$= -\frac{1}{m}\left[\sum_{i=1}^{m} \left(y^{(i)} - \cancel{y^{(i)} \cdot h_\theta(x^{(i)})} - h_\theta(x^{(i)}) + \cancel{y^{(i)} \cdot h_\theta(x^{(i)})} \right) \cdot x_j^{(i)}\right] =$$

$$\frac{\partial}{\partial \theta_j} J(\boldsymbol{\theta}) = -\frac{1}{m}\left[\sum_{i=1}^{m} \left(y^{(i)} - h_\theta(x^{(i)}) \right) \cdot x_j^{(i)}\right]$$

Regularized Logistic Regression

- We can regularize logistic regression in a similar way that we regularize linear regression

- Recall that our cost function was:

$$J(\boldsymbol{\theta}) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \cdot \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \cdot \log(1 - h_\theta(x^{(i)}))\right]$$

- We can regularize this equation by adding a term:

$$J(\boldsymbol{\theta}) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \cdot \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \cdot \log(1 - h_\theta(x^{(i)}))\right] + \frac{\lambda}{2m}\sum_{j=1}^{n} \theta_j^2$$

Regularized Logistic Regression

- We can learn our parameters with gradient descent

$$\min_{\boldsymbol{\theta}} -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \cdot \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \cdot \log(1 - h_\theta(x^{(i)}))\right] + \frac{\lambda}{2m}\sum_{j=1}^{n} \theta_j^2$$

repeat until convergence{ (simultaneously update all θ_j)

$$\theta_0 := \theta_0 - \eta \frac{1}{m}\sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})\, x_0^{(i)}$$

$$\theta_j := \theta_j - \eta \left[\frac{1}{m}\sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})\, x_j^{(i)} + \frac{\lambda}{m}\theta_j\right]$$

$(j = 1, \ldots, n)$

}

This is exactly the same algorithm we use for linear regression!

Contact

- **Office:** Torre Archimede, room 6CD3
- **Office hours** (ricevimento)**:** Friday 9:00-11:00

✉ lamberto.ballan@unipd.it

🏠 http://www.lambertoballan.net

🏠 http://vimp.math.unipd.it

@ twitter.com/lambertoballan