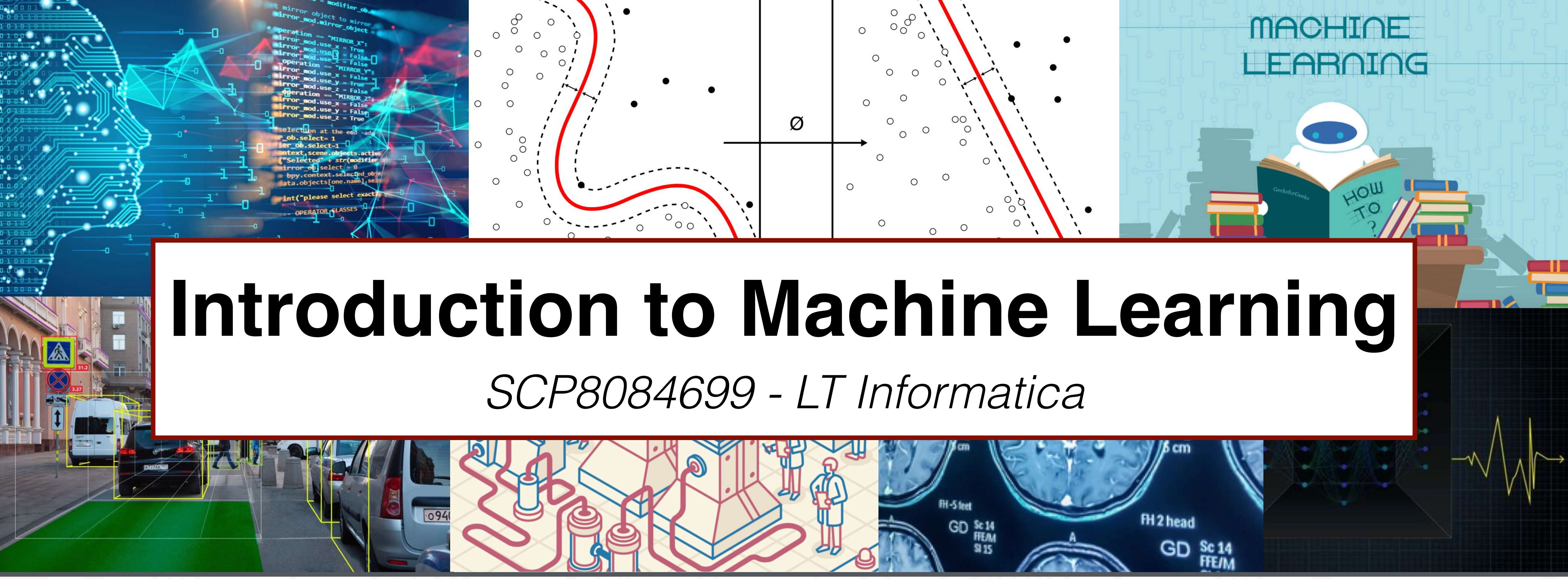


Introduction to Machine Learning

SCP8084699 - LT Informatica

Linear Classification, Logistic Regression

Prof. Lamberto Ballan

DIPARTIMENTO MATEMATICA
Department of Mathematics "Tullio Levi-Civita"

A bit more on Gradient Descent

- We have introduced *batch gradient descent*
  (i.e. each step of gradient descent uses all training examples)
  - There is another way to optimize across the training set…

- Stochastic Gradient Descent: update the parameters for each training case in turn, according to its own gradients

```
Randomly shuffle examples in the training set

for i=1 to m do{
```

$$\theta_0 \coloneqq \theta_0 - \eta \; (\theta_0 + \theta_1 x^{(i)} - y^{(i)})$$

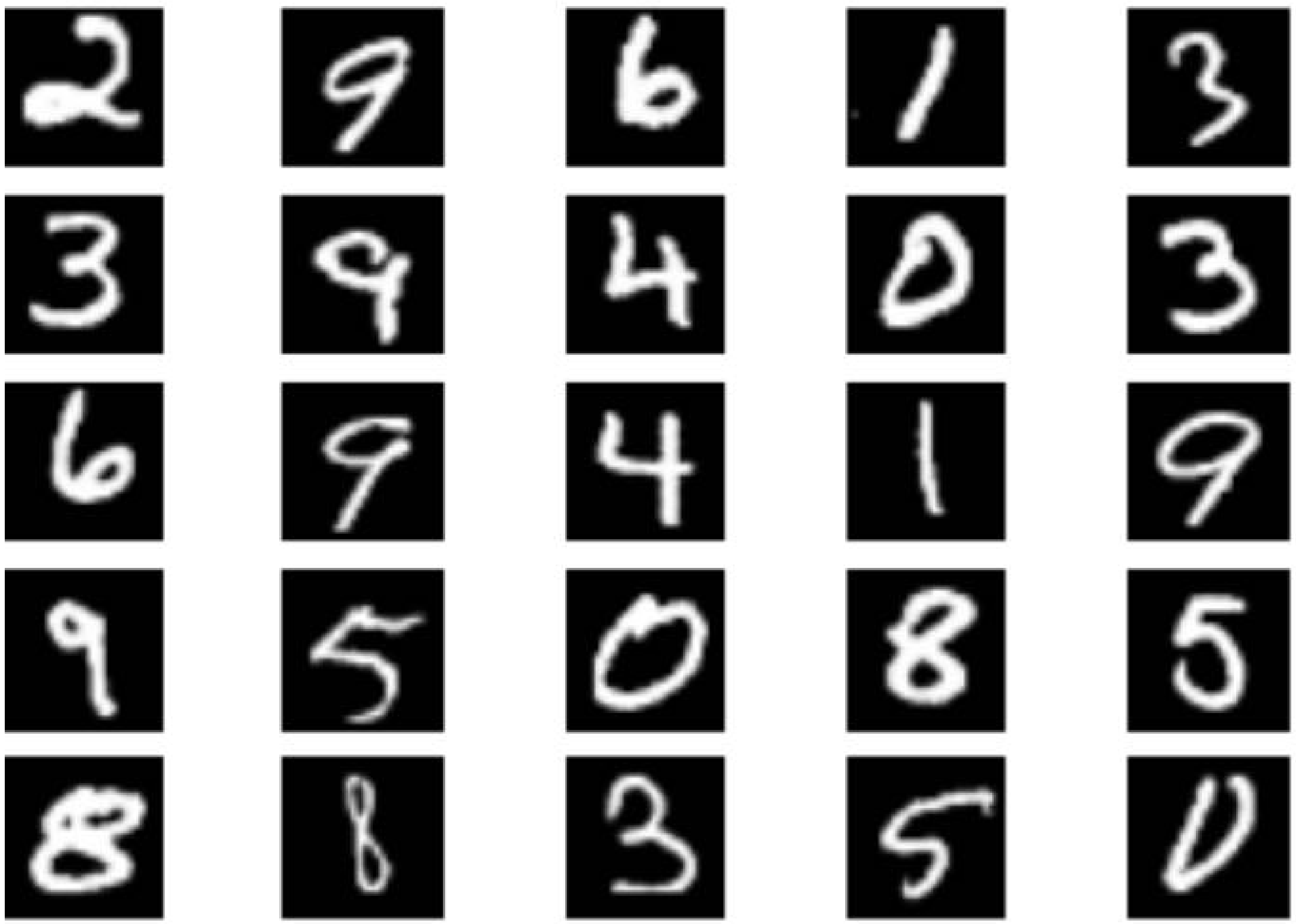
$$\theta_1 \coloneqq \theta_1 - \eta \; (\theta_0 + \theta_1 x^{(i)} - y^{(i)}) \, x^{(i)}$$

```
}
```

Underlying assumption: samples are independent and identically distributed (i.i.d.)

Learning is useful in many tasks

- **Classification:** determine to which discrete category a specific example belongs to



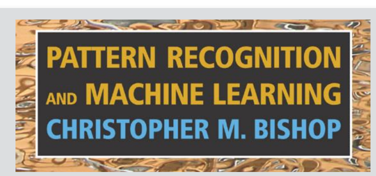
Example 1

What digit is this?

Learning is useful in many tasks

- **Classification:** determine to which discrete category a specific example belongs to

- Other examples:
  - Email: spam *vs* not spam (*ham*)
  - Online transactions: fraudulent *vs* not fraudulent
  - Tumor: malignant *vs* benign



Pages 179-195

Classification *vs* Regression

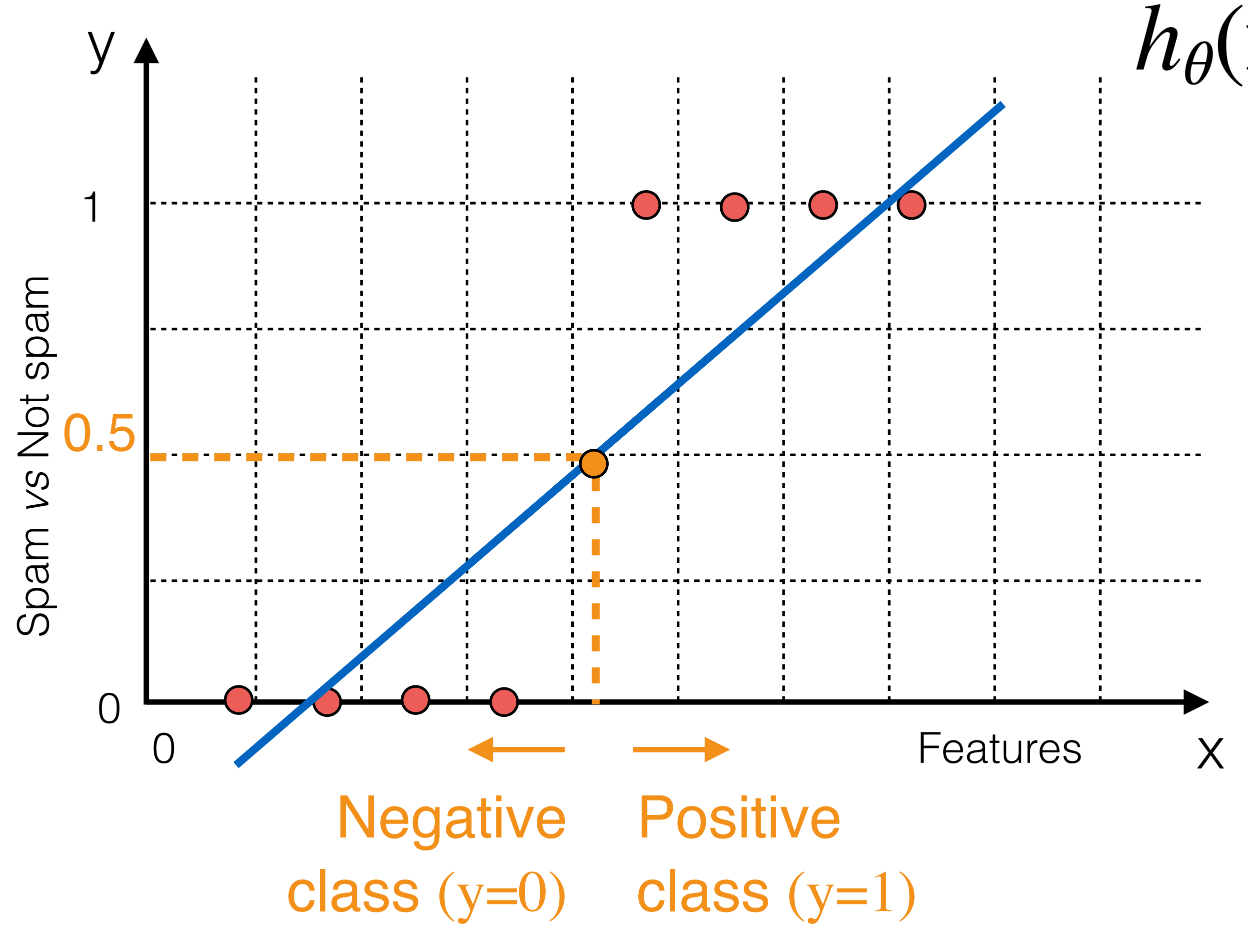
- Categorical outputs called labels (or classes)
  - e.g. yes/no, 1/2/3/…/9, cat/dog/person/…
  - Then we are interested in: $h \sim f: X \rightarrow Y$, where Y is categorical (while in regression typically $Y=\mathbb{R}$)

- Binary classification: two possible labels
- Multi-class classification: multiple possible labels

We will first look at binary problems and then discuss multi-class problems

Classification as Regression

- Can we do (binary) classification using what we have learned until now?



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

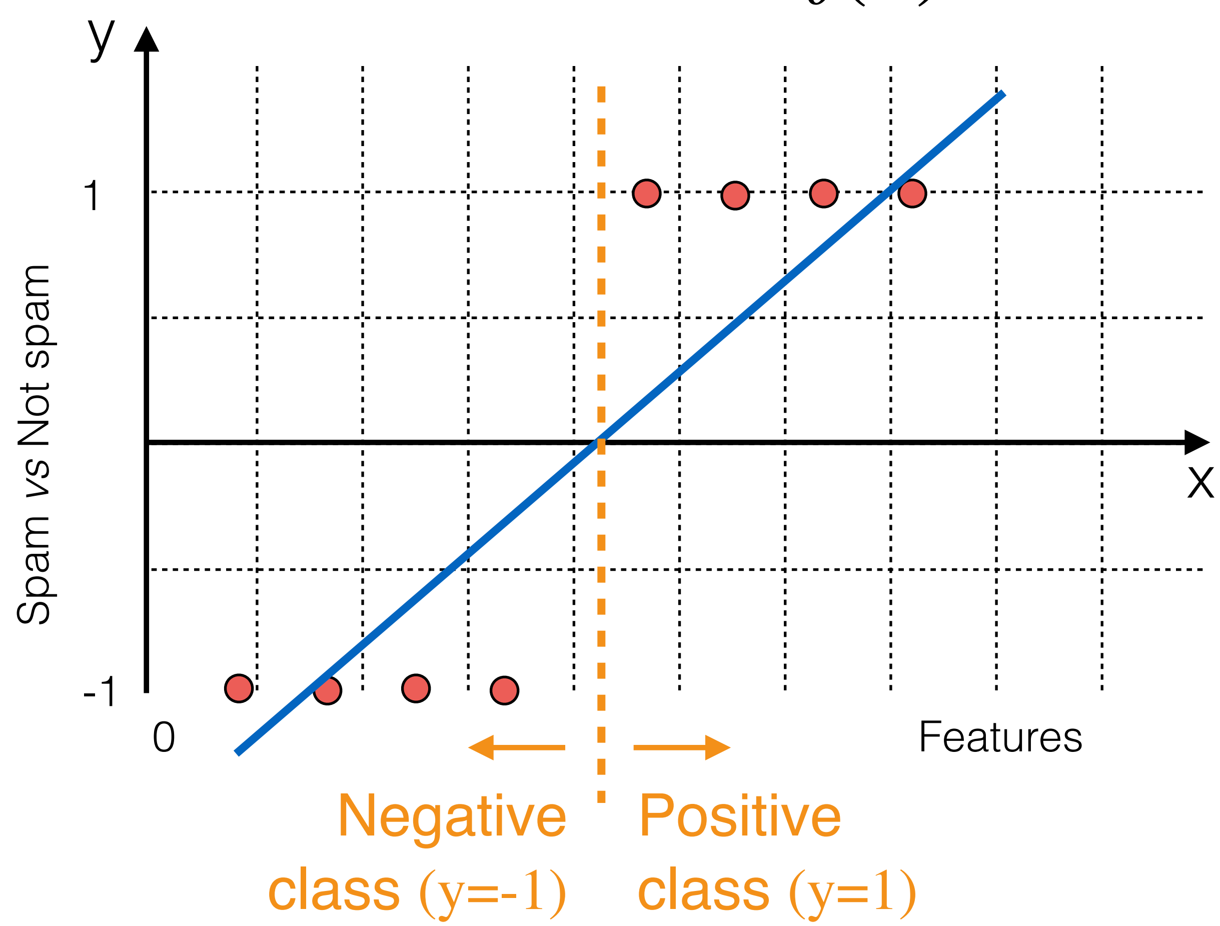
Threshold classifier:

- If $h_\theta(\mathbf{x}) \geq 0.5$, predict $y = 1$
- If $h_\theta(\mathbf{x}) < 0.5$, predict $y = 0$

Classification as Regression

- Let's use a slightly different notation

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



Threshold classifier:

- If $h_\theta(\mathbf{x}) \geq 0$, predict $y = 1$
- If $h_\theta(\mathbf{x}) < 0$, predict $y = -1$

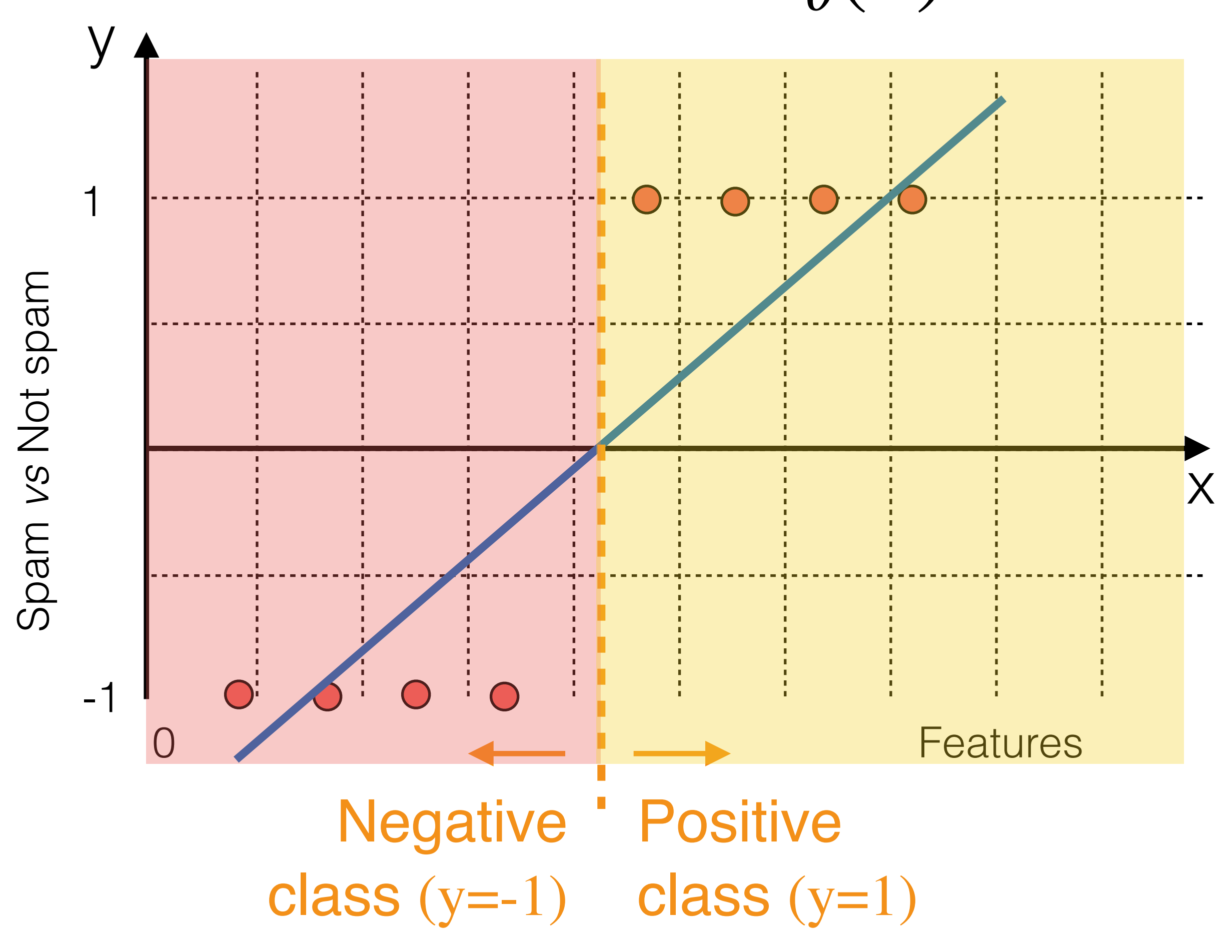
Decision rule *(mathematically)*:

- $y = \mathrm{sign}(h_\theta(\mathbf{x}))$

Linear Classification

- This specifies a *linear classifier*: it has a linear boundary (hyperplane) which separates the space

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



Decision rule:

- $\mathrm{y} = \mathrm{sign}(h_\theta(\mathbf{x}))$

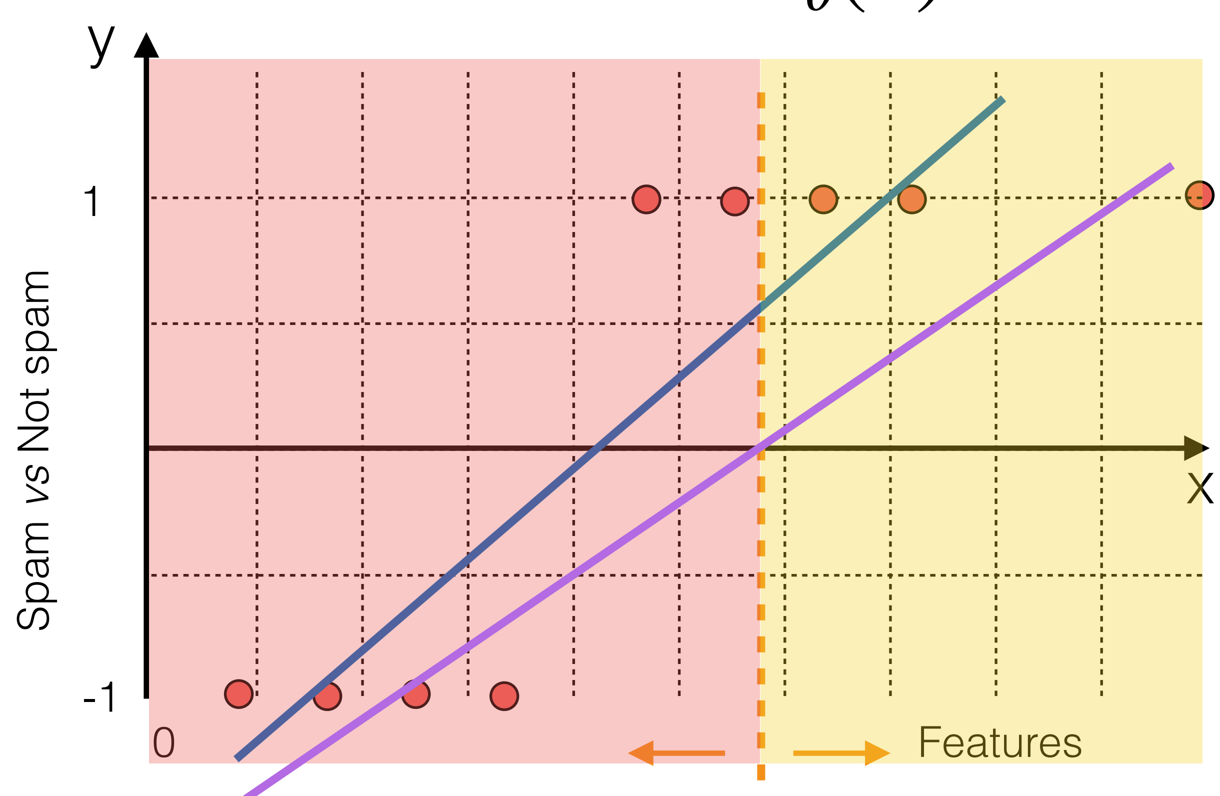
The linear boundary separates the space into two “half-spaces”

In 1D this is simply a threshold

Linear Classification

- Applying linear regression to classification tasks is not always a great idea…

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



Decision rule:

- $\mathrm{y} = \mathrm{sign}(h_\theta(\mathbf{x}))$

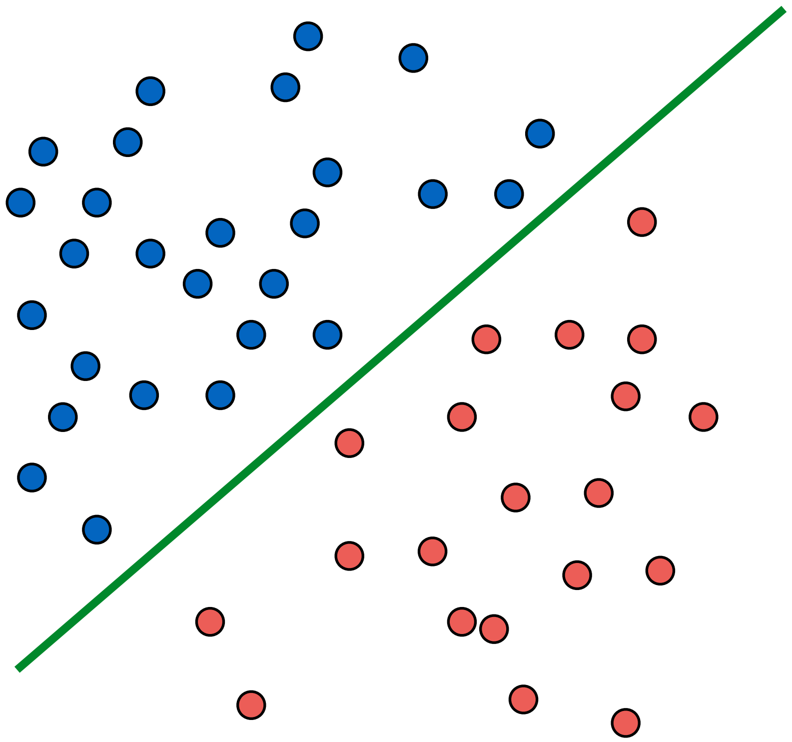
The linear boundary separates the space into two “half-spaces”

In 1D this is simply a threshold

Linear Classification

- This specifies a *linear classifier*: it has a linear boundary (hyperplane) which separates the space

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



Decision rule:

- $\mathrm{y} = \mathrm{sign}(h_\theta(\mathbf{x}))$

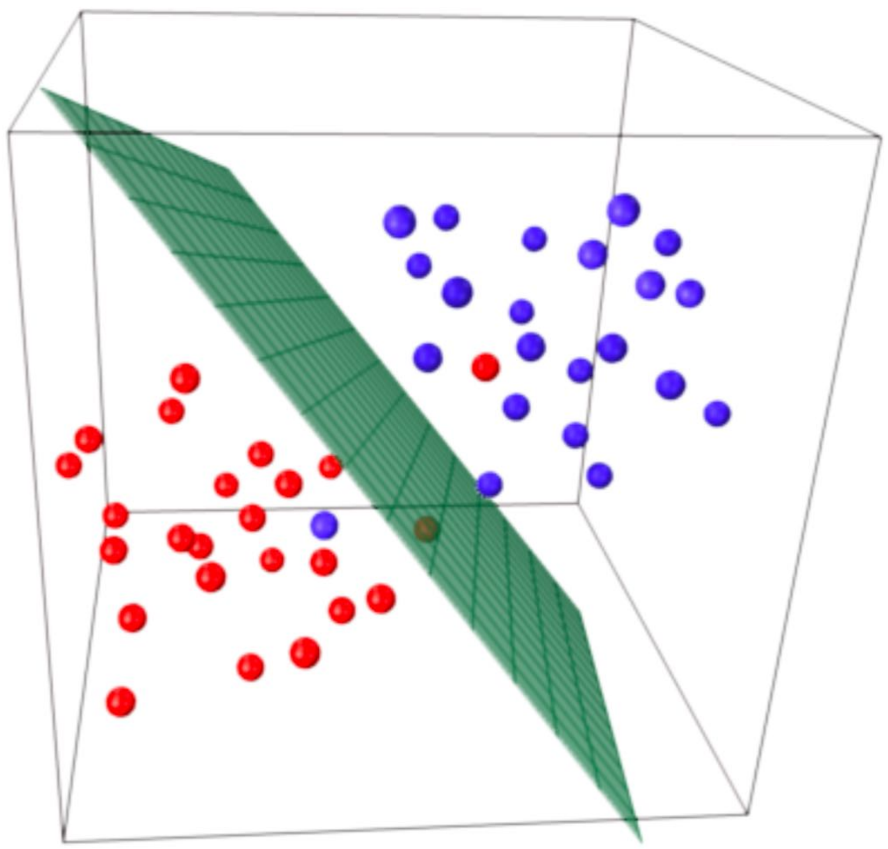
The linear boundary separates the space into two "half-spaces"

In 2D this is a line

Linear Classification

- This specifies a *linear classifier*: it has a linear boundary (hyperplane) which separates the space

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



Decision rule:

- $\mathrm{y} = \mathrm{sign}(h_\theta(\mathbf{x}))$

The linear boundary separates the space into two "half-spaces"

In 3D this is a plane

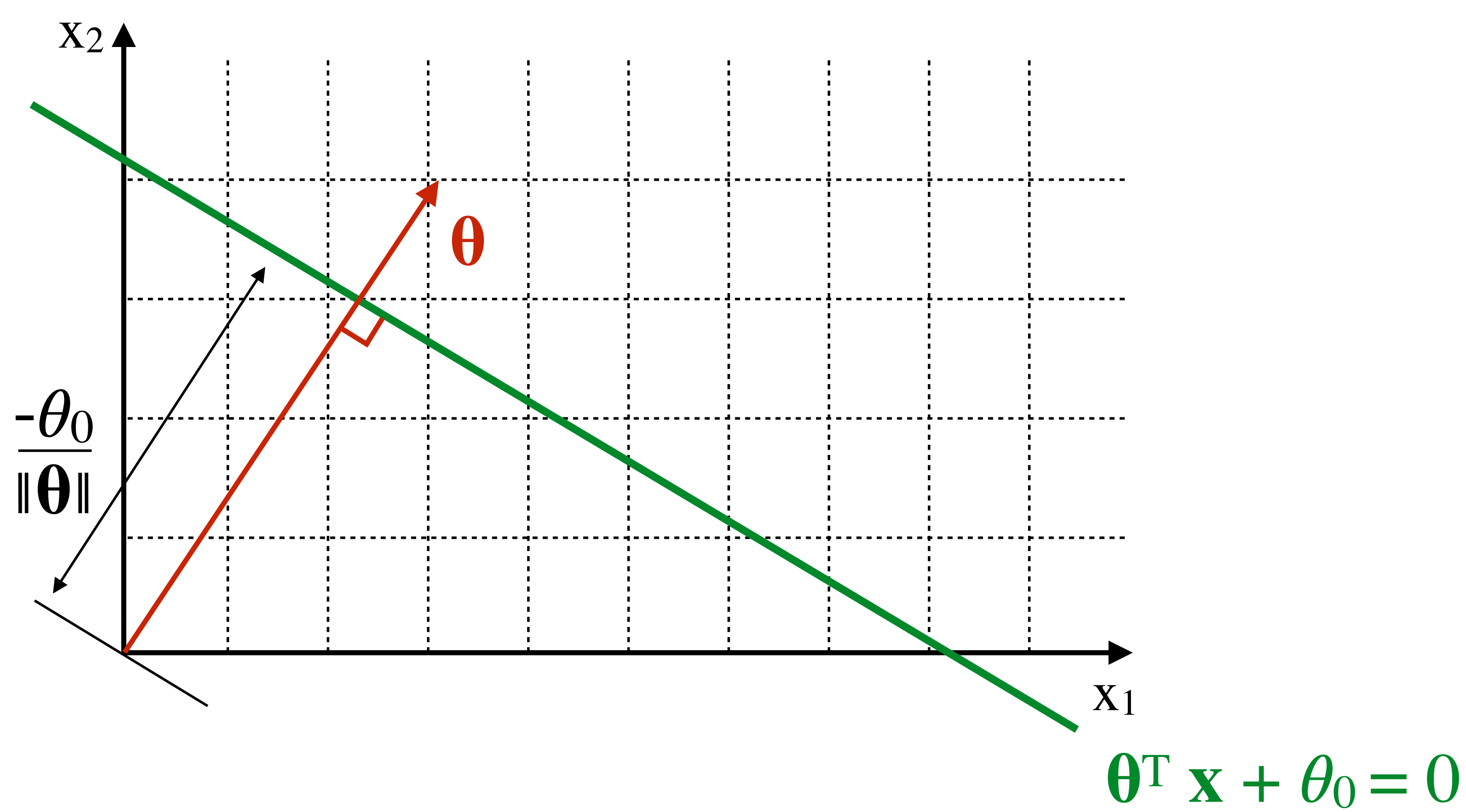
Image credit: S. Fidler

Geometric Interpretation

- What about higher-dimensional spaces?

$\boldsymbol{\theta}^T \mathbf{x} = 0$ a line passing through the origin and orthogonal to $\boldsymbol{\theta}$

$\boldsymbol{\theta}^T \mathbf{x} + \theta_0 = 0$ shifts it by θ_0 ← *Note: this is usually referred as to the "bias term"*

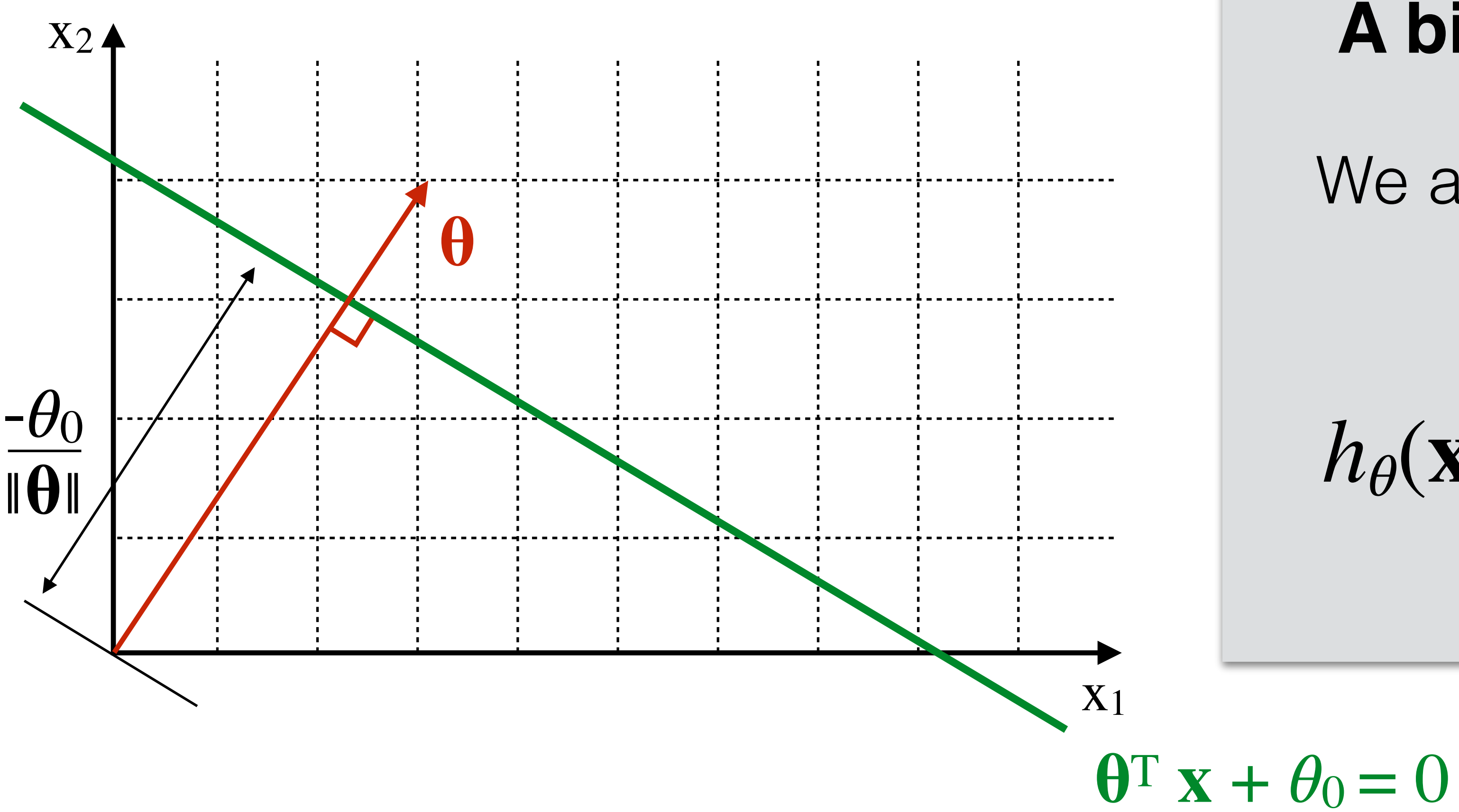


Geometric Interpretation

- What about higher-dimensional spaces?

$\boldsymbol{\theta}^{\mathrm{T}}\, \mathbf{x} = 0$ a line passing through the origin and orthogonal to $\boldsymbol{\theta}$

$\boldsymbol{\theta}^{\mathrm{T}}\, \mathbf{x} + \theta_0 = 0$ shifts it by θ_0 ← *Note: this is usually referred as to the "bias term"*



A bit more about the notation

We are using this trick/assumption:

$$h_\theta(\mathbf{x}) = \boldsymbol{\theta}^{\mathrm{T}}\, \mathbf{x} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix}$$

Learning Linear Classifiers

- Learning = estimating a “good” decision boundary
  - Find $\boldsymbol{\theta}$ (direction) and θ_0 (location) of the boundary
  - We need a criteria to select the parameters

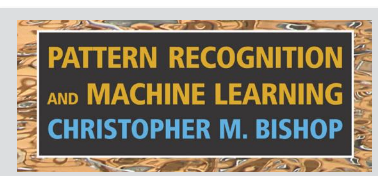
- Loss (cost) functions:
  - Zero/One: $J_{01}(\boldsymbol{\theta}) = \dfrac{1}{m} \sum_{i=1}^{m} \{0 \text{ if } h_\theta(x^{(i)})=y^{(i)}, 1 \text{ otherwise}\}$
  - Absolute: $J_{abs}(\boldsymbol{\theta}) = \dfrac{1}{m} \sum_{i=1}^{m} |h_\theta(x^{(i)}) - y^{(i)}|$
  - Squared: $J_{sqr}(\boldsymbol{\theta}) = \dfrac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$

Learning Linear Classifiers

- Learning = estimating a “good” decision boundary
  - Find $\boldsymbol{\theta}$ (direction) and θ_0 (location) of the boundary
  - We need a criteria to select the parameters
- Loss function: $J(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} cost(h_\theta(x^{(i)}), y^{(i)})$
  - Zero/One: $cost(h_\theta(x^{(i)}), y^{(i)}) = \{0 \text{ if } h_\theta(x^{(i)})=y^{(i)}, 1 \text{ otherwise}\}$
  - Absolute: $cost(h_\theta(x^{(i)}), y^{(i)}) = |h_\theta(x^{(i)}) - y^{(i)}|$
  - Squared: $cost(h_\theta(x^{(i)}), y^{(i)}) = \frac{1}{2}(h_\theta(x^{(i)}) - y^{(i)})^2$

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: $0 \leq h_\theta(\mathrm{x}) \leq 1$



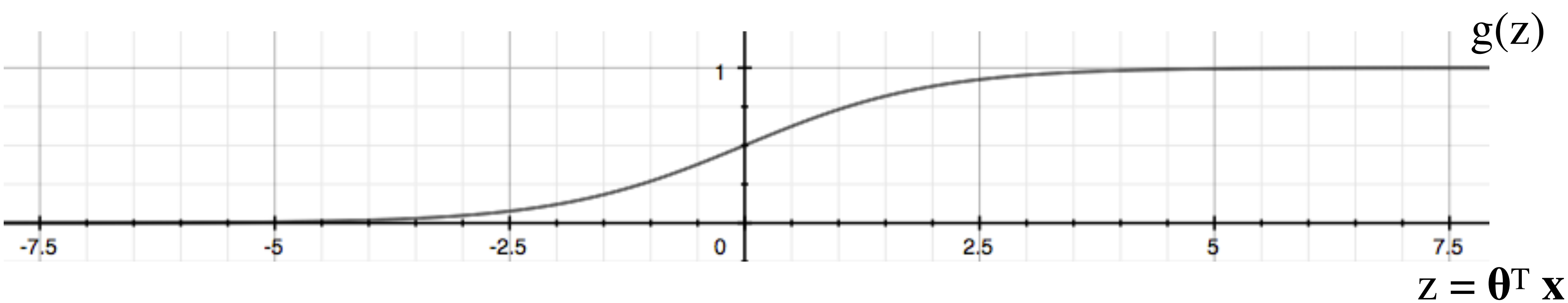
Pages 203-207

Logistic Regression

- Hypothesis representation:

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

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



Adapted from slide by A. Ng

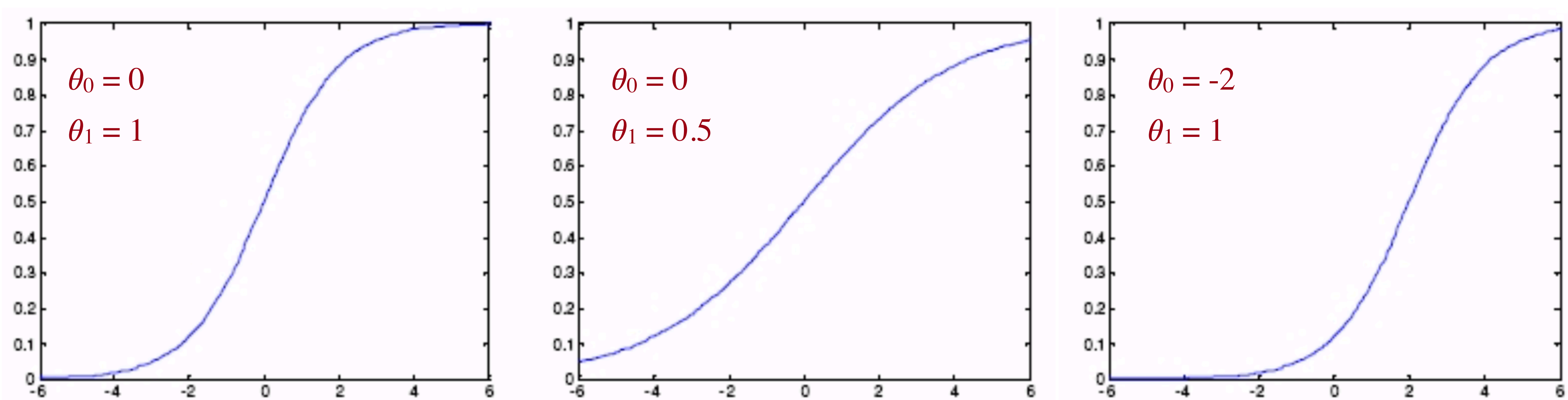
Logistic Regression

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

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

(*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 $y=1$ on input x
  - More formally: $h_\theta(\mathbf{x}) = P\,(y=1 \mid x;\, \theta)$

- An example:

$$\mathbf{x} = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{tumor\_size} \end{bmatrix} \qquad h_\theta(\mathbf{x}) = 0.7$$

Tell patient that 70% chance of tumor being malignant

Probabilistic Interpretation

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

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

    therefore $P(y=0 \mid x; \theta) = 1 - P(y=1 \mid x; \theta)$

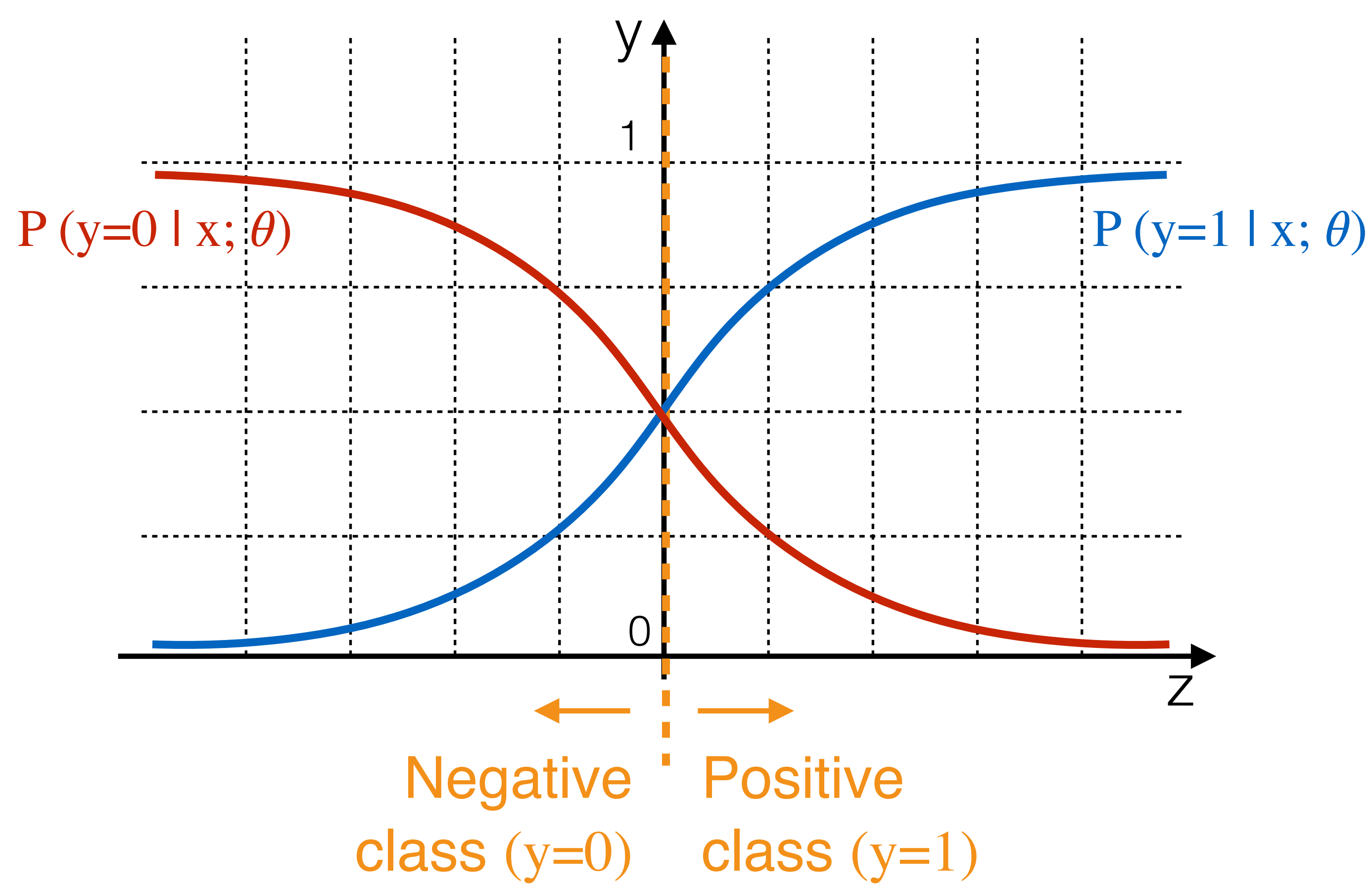
    i.e. $P(y=0 \mid x; \theta) = 1 - \dfrac{1}{1 + e^{-\boldsymbol{\theta}^T\mathbf{x}}} = \dfrac{e^{-\boldsymbol{\theta}^T\mathbf{x}}}{1 + e^{-\boldsymbol{\theta}^T\mathbf{x}}}$

Decision Boundary

- What is the decision boundary for logistic regression?

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

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



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

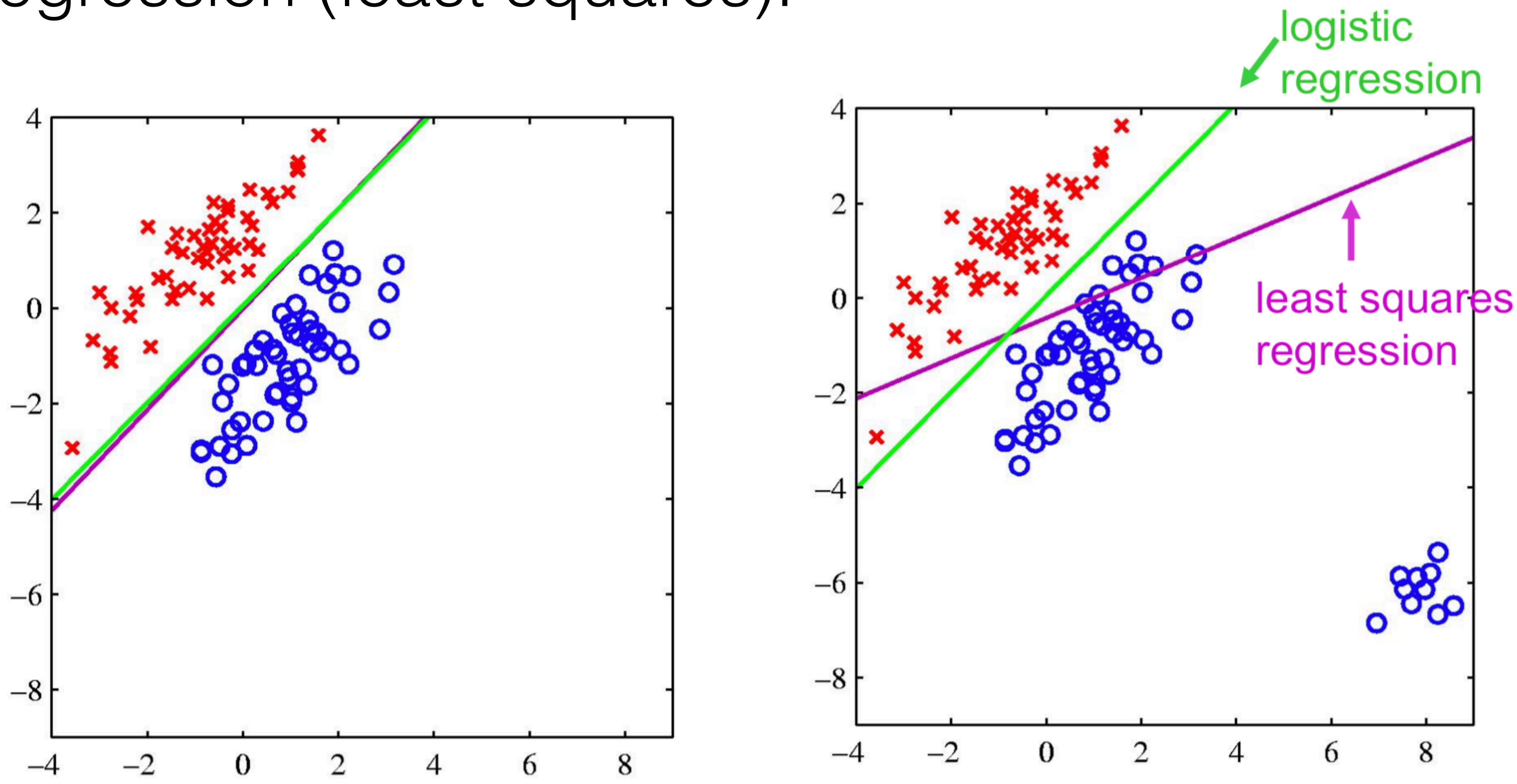
Suppose predict $y=1$ if $h_\theta(x) \geq 0.5$

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)

Adapted from slide by S. Fidler

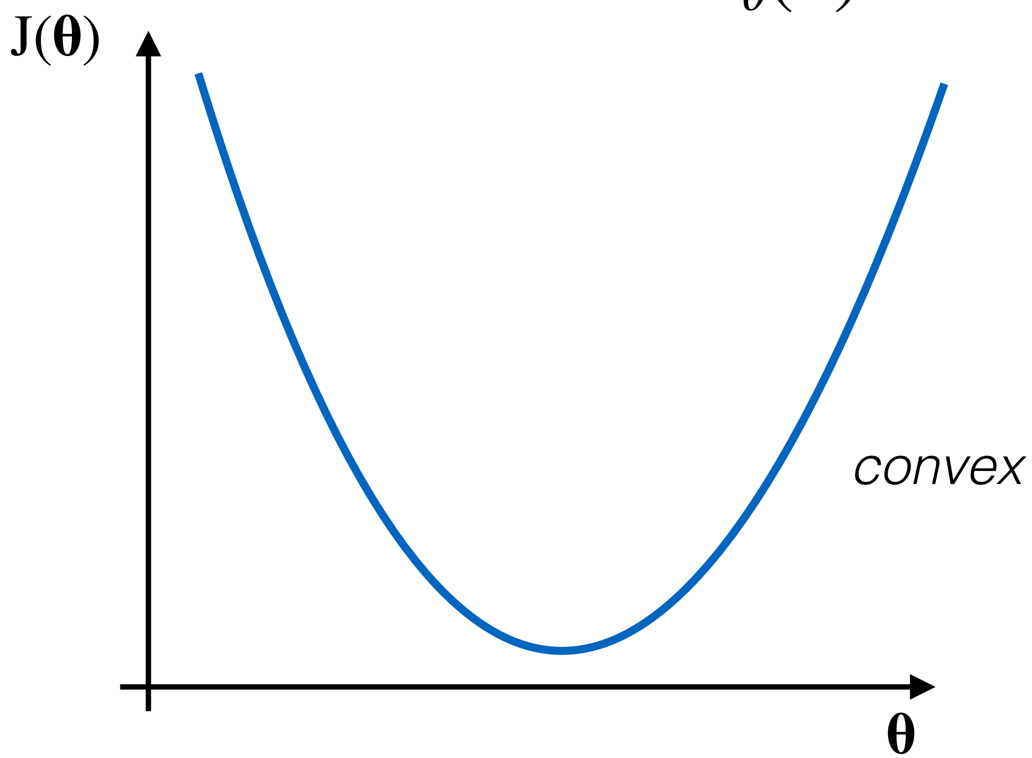
Logistic *vs* Linear Regression

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

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

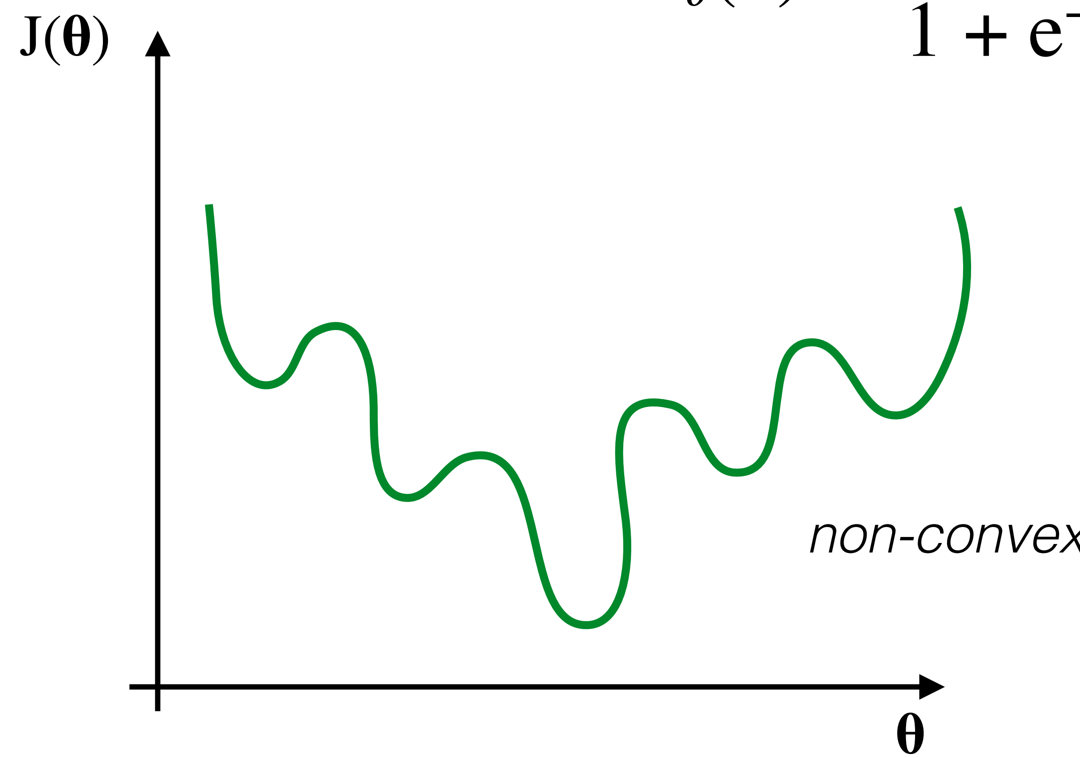
Linear Regression

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



Logistic Regression

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

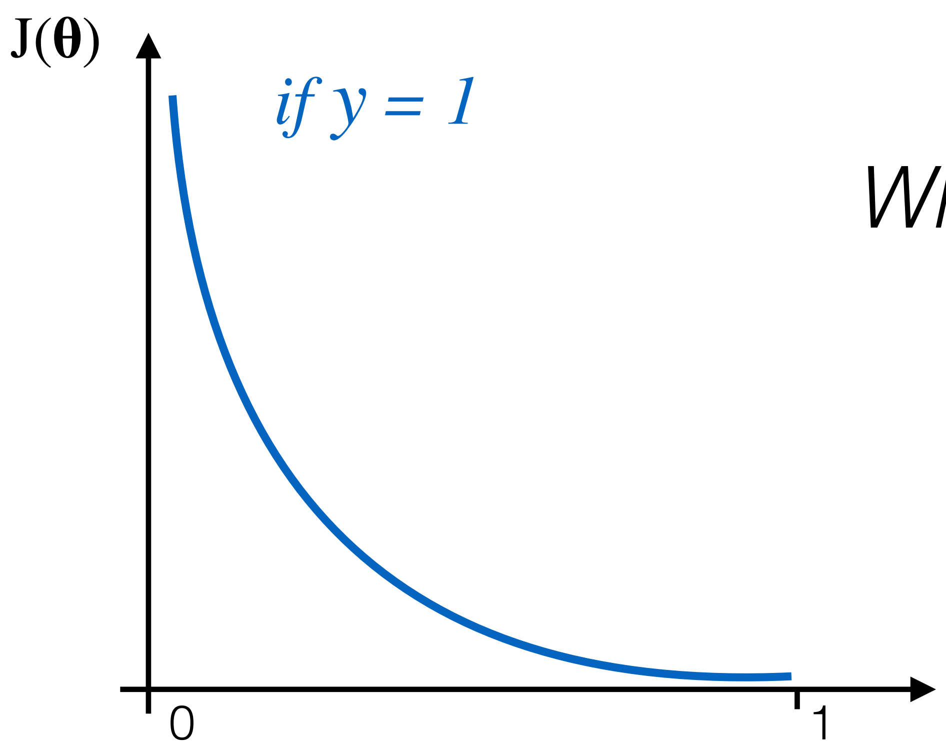


Logistic Regression Loss Function

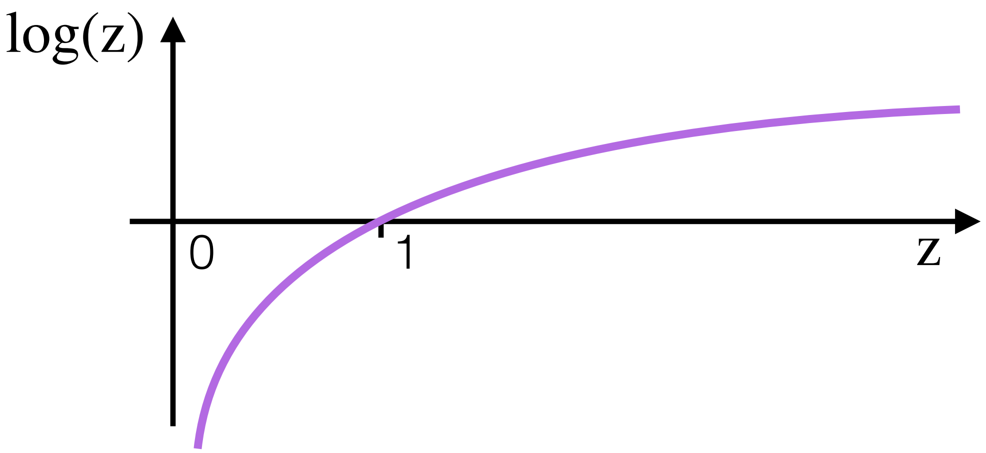
- Loss function: $J(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} cost(h_\theta(x^{(i)}), y^{(i)})$

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

- Intuition:



Why is that?



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

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

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

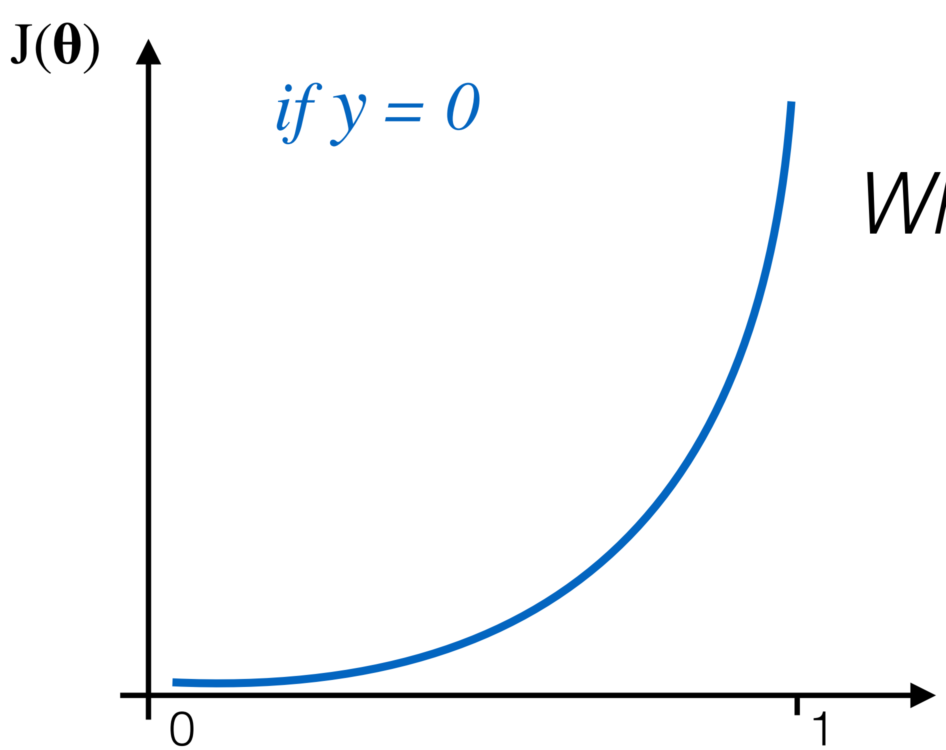
Adapted from slide by A. Ng

Logistic Regression Loss Function

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

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

- Intuition:



Why is that?

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

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

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

Logistic Regression Loss Function

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

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

  - Note: by definition $y=1$ or $y=0$ (binary classifier)
  - "Simplified notation":

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

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

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

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

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

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

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

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

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:

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

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

$f(x) = \log(x)$

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

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

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

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

- Writing now in terms of partial derivatives:

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

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

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:

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

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

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

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

Contact

- **Office:** Torre Archimede 6CD, room 622
- **Office hours** (ricevimento)**:** Friday 11:00-13:00

✉ lamberto.ballan@unipd.it

http://www.lambertoballan.net

http://vimp.math.unipd.it

@ twitter.com/lambertoballan