Machine Learning and Pattern Recognition
A High Level Overview

Prof. Anderson Rocha
(Main bulk of slides kindly provided by **Prof. Sandra Avila**
and largely based on other materials as well (e.g., Andrew Ng's))
Institute of Computing (IC/Unicamp)

MC886/MO444

Today’s Agenda

- Logistic Regression
    - Classification
    - Hypothesis Representation
    - Decision Boundary
    - Cost Function
    - Simplified Cost Function and Gradient Descent
    - Multiclass Classification

Classification

Spam Filtering

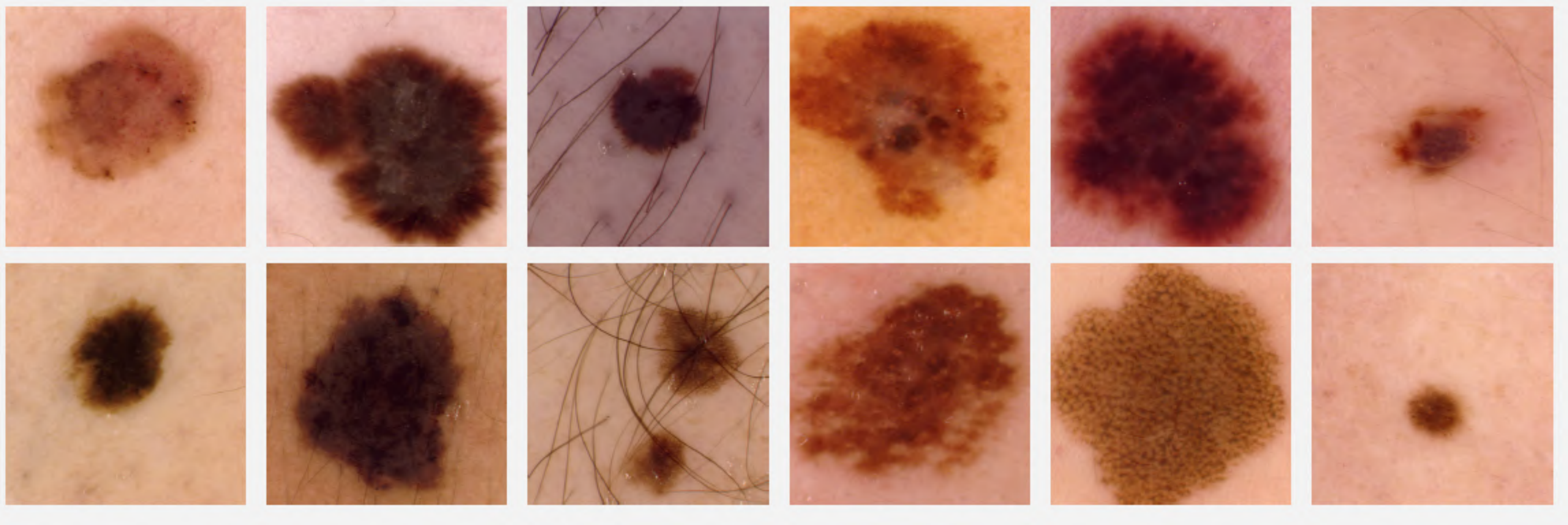
Bad Cures fast and effective! - Canadian *** Pharmacy #1 Internet Inline Drugstore Viagra Cheap Our price $1.99 …

Good Interested in your research on graphical models - Dear Prof., I have read some of your papers on probabilistic graphical models. Because I …

Sensitive Content Classification

Image Credit (left) : http://ehow-blog.com/how-to-avoid-internet-addiction/
Image Credit (right) : http://www.telegraph.co.uk/culture/tvandradio/9840832/Children-under-5-should-not-watch-TV-alone-Jackanory-creator-argues.html

Skin Cancer Classification



Melanomas (top row) and **benign** skin lesions (bottom row)

Image Credit: M. Fornaciali, S. Avila, M. Carvalho, E. Valle, Statistical learning approach for robust melanoma screening, SIBGRAPI, 2014

Classification

Email: **Spam** / **Not Spam**?

Content Video: **Sensitive** / **Non-sensitive**?

Skin Lesion: **Malignant** / **Benign**?

Classification

Email: **Spam** / **Not Spam**?

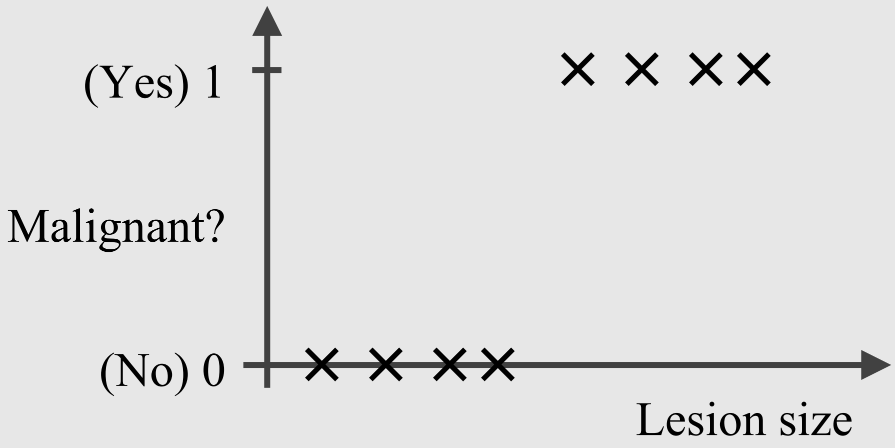
Content Video: **Sensitive** / **Non-sensitive**?

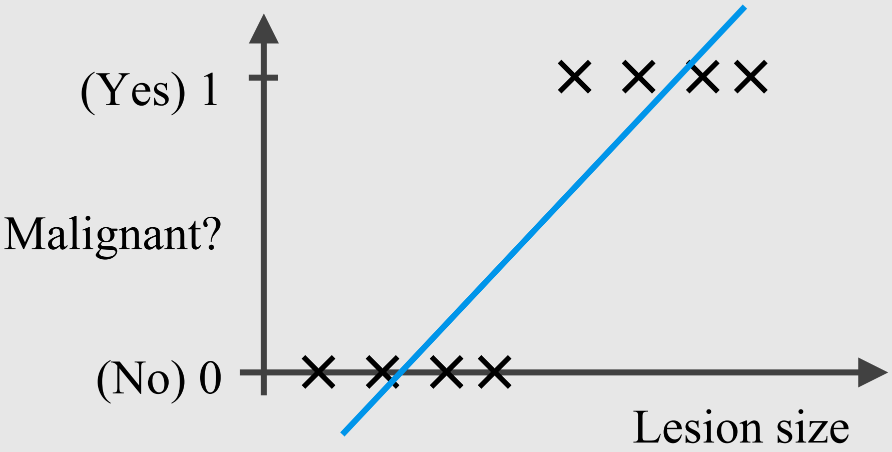
Skin Lesion: **Malignant** / **Benign**?

$y \in \{0,1\}$

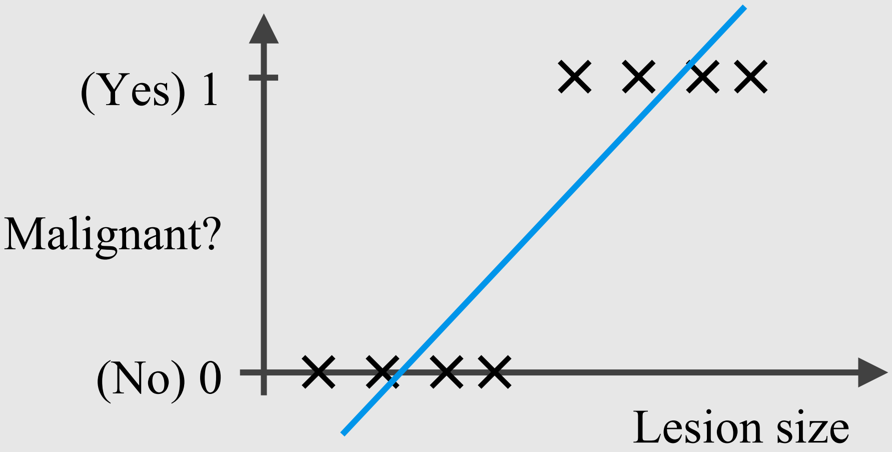
0: "Negative Class" (e.g., Benign skin lesion)

1: "Positive Class" (e.g., Malignant skin lesion)


(Yes) 1
Malignant?
(No) 0
Lesion size



$$h_\theta(x) = \theta^T x$$

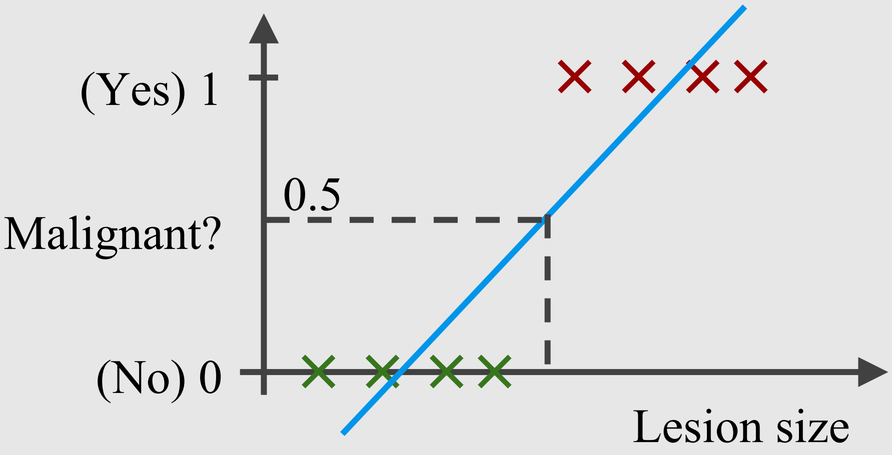


$$h_\theta(x) = \theta^T x$$

Threshold classifier output $h_\theta(x)$ at 0.5:

If $h_\theta(x) \geq 0.5$, predict "$y = 1$"

If $h_\theta(x) < 0.5$, predict "$y = 0$"



$$h_\theta(x) = \theta^T x$$

Threshold classifier output $h_\theta(x)$ at 0.5:

If $h_\theta(x) \geq 0.5$, predict "$y = 1$"

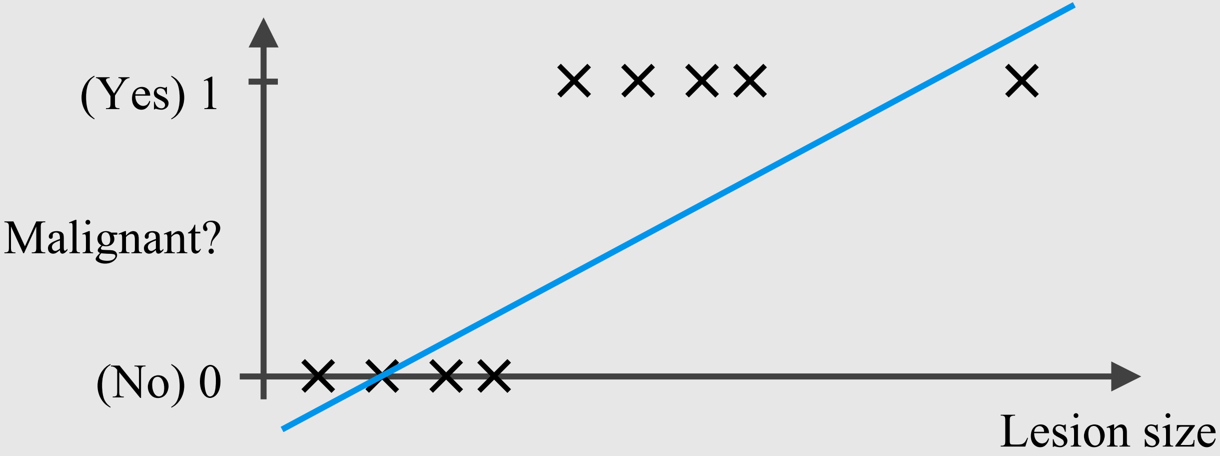
If $h_\theta(x) < 0.5$, predict "$y = 0$"

$$h_\theta(x) = \theta^T x$$

Threshold classifier output $h_\theta(x)$ at 0.5:

If $h_\theta(x) \geq 0.5$, predict "$y = 1$"

If $h_\theta(x) < 0.5$, predict "$y = 0$"

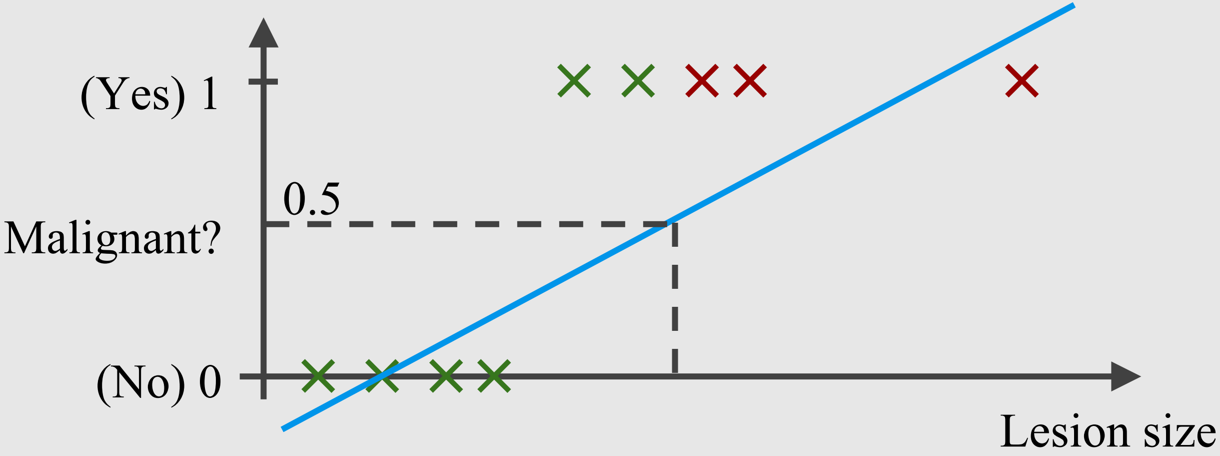


$$h_\theta(x) = \theta^T x$$

Threshold classifier output $h_\theta(x)$ at 0.5:

If $h_\theta(x) \geq 0.5$, predict "$y = 1$"

If $h_\theta(x) < 0.5$, predict "$y = 0$"



$$h_\theta(x) = \theta^T x$$

Threshold classifier output $h_\theta(x)$ at 0.5:

If $h_\theta(x) \geq 0.5$, predict "$y = 1$"

If $h_\theta(x) < 0.5$, predict "$y = 0$"

Classification: $y = 0$ or $y = 1$

$h_\theta(x)$ can be > 1 or < 0

Logistic Regression: $0 \leq h_\theta(x) \leq 1$

Hypothesis Representation

Logistic Regression Model

Want $0 \le h_\theta(x) \le 1$

Logistic Regression Model

Want $0 \leq h_\theta(x) \leq 1$

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

Logistic Regression Model

Want $0 \leq h_\theta(x) \leq 1$

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

Logistic Regression Model

Want $0 \leq h_\theta(x) \leq 1$

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

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

Logistic Regression Model

Want $0 \leq h_\theta(x) \leq 1$

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

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

Sigmoid Function

Logistic Function

Logistic Regression Model

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

Want $0 \le h_\theta(x) \le 1$

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

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

Sigmoid Function

Logistic Function

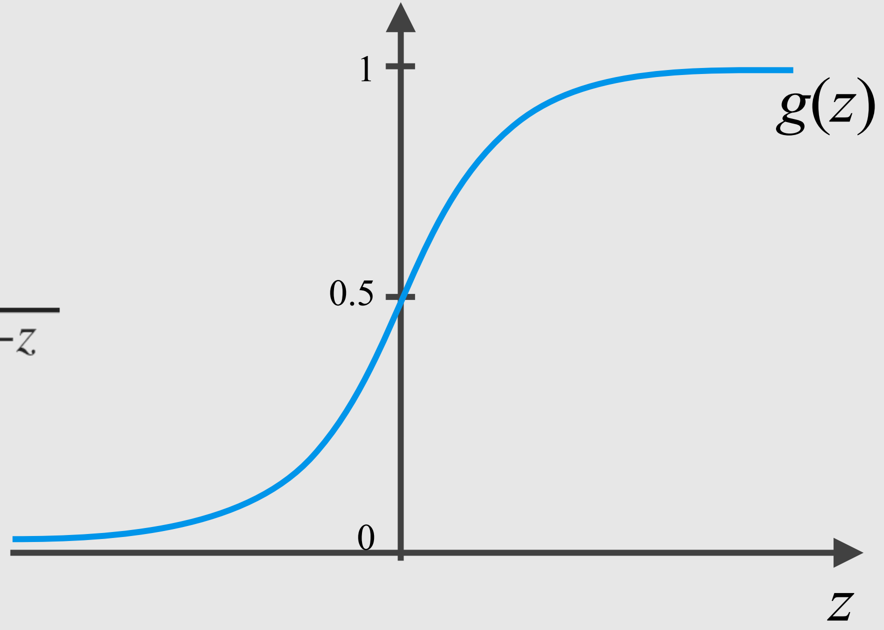
Logistic Regression Model

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

Want $0 \le h_\theta(x) \le 1$

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

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



Sigmoid Function

Logistic Function

Interpretation of Hypothesis Output

$h_\theta(x)$ = estimated probability that $y = 1$ on input x

Interpretation of Hypothesis Output

$h_\theta(x)$ = estimated probability that $y = 1$ on input x

Example: If $x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{tumorSize} \end{bmatrix}$

$h_\theta(x) = 0.7$

Tell patient that 70% chance of tumor being malignant

Interpretation of Hypothesis Output

$h_\theta(x)$ = estimated probability that $y = 1$ on input x

Example: If $x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{tumorSize} \end{bmatrix}$

$h_\theta(x) = 0.7$

Tell patient that 70% chance of tumor being malignant

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

Interpretation of Hypothesis Output

$h_\theta(x)$ = estimated probability that $y = 1$ on input x

Example: If $x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{tumorSize} \end{bmatrix}$

$h_\theta(x) = 0.7$

Tell patient that 70% chance of tumor being malignant

“probability that $y = 1$, given x, parameterized by θ”

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

Interpretation of Hypothesis Output

$h_\theta(x)$ = estimated probability that $y = 1$ on input x

Example: If $x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{tumorSize} \end{bmatrix}$

$h_\theta(x) = 0.7$

⬇

Tell patient that 70% chance of tumor being malignant

"probability that $y = 1$, given x, parameterized by θ"

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

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

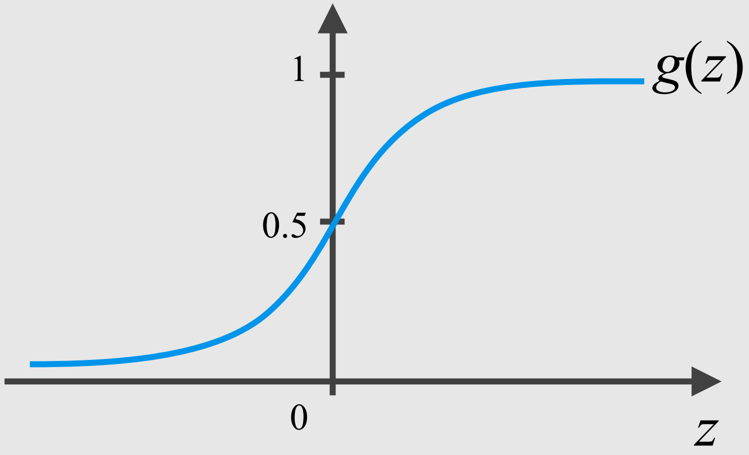
$$h_\theta(x) = P(y = 1 \mid x;\theta)$$

Decision Boundary

Logistic Regression

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

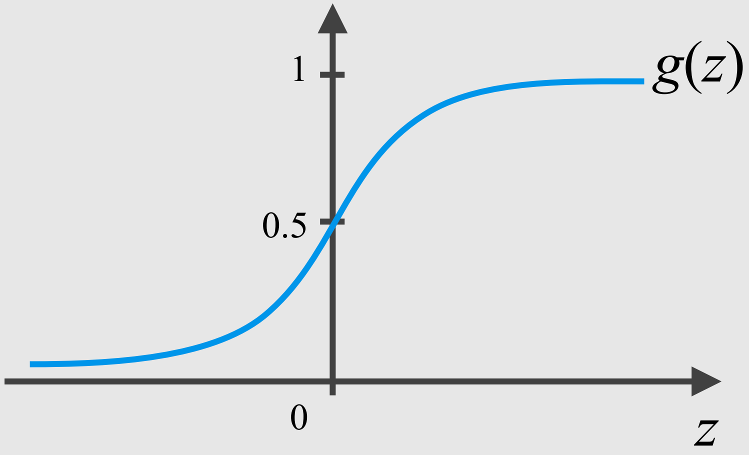
$$g(z) = \frac{1}{1+\mathrm{e}^{-z}}$$



Logistic Regression

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

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



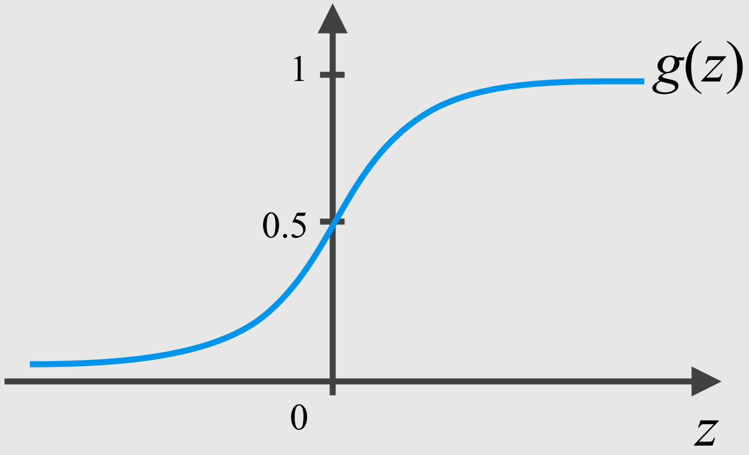
Suppose predict "$y = 1$" if $\quad h_\theta(x) \geq 0.5$

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

Logistic Regression

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

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



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

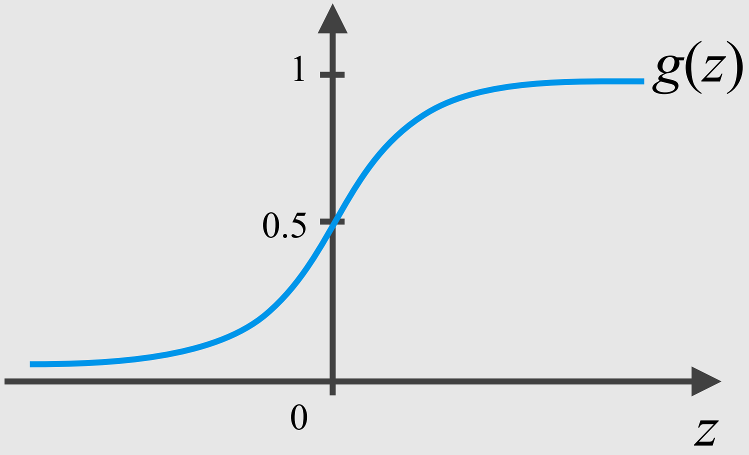
predict "$y = 0$" if $h_\theta(x) < 0.5$

$g(z) \geq 0.5$ when $z \geq 0$

Logistic Regression

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

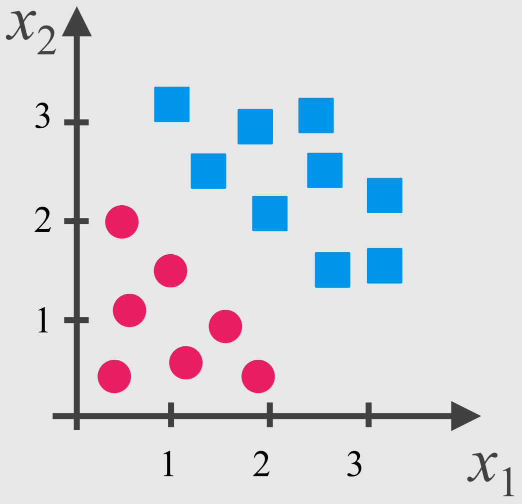
$$g(z) = \frac{1}{1+e^{-z}}$$



Suppose predict "$y = 1$" if $h_\theta(x) \geq 0.5$ $\quad$ $g(z) \geq 0.5$ when $z \geq 0$

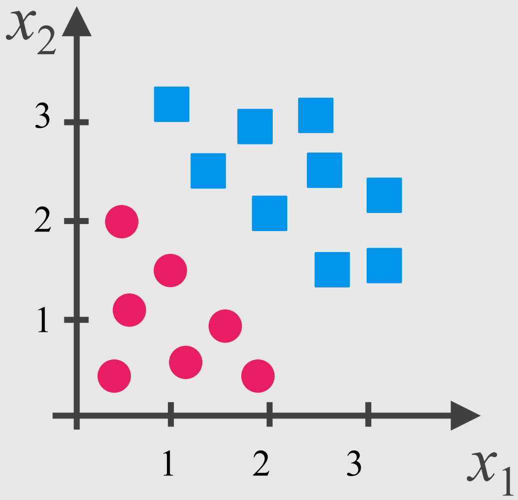
predict "$y = 0$" if $h_\theta(x) < 0.5$ $\quad$ $g(z) < 0.5$ when $z < 0$

Decision Boundary



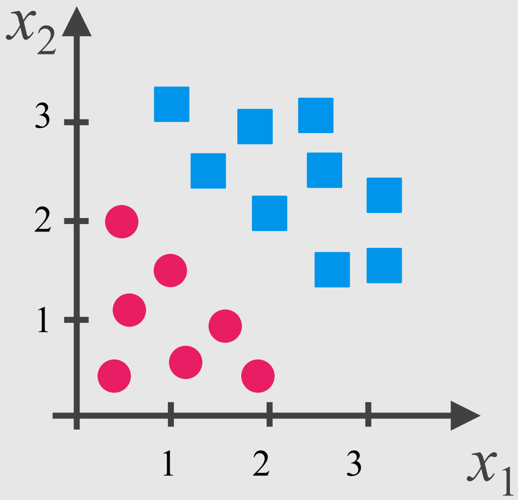
$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

Decision Boundary



$$\begin{array}{c} \quad\quad\;\; -3 \quad\;\; 1 \quad\quad\;\; 1 \\ \quad\quad\;\; \Uparrow \quad\;\; \Uparrow \quad\quad\;\; \Uparrow \\ h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2) \end{array}$$

Decision Boundary

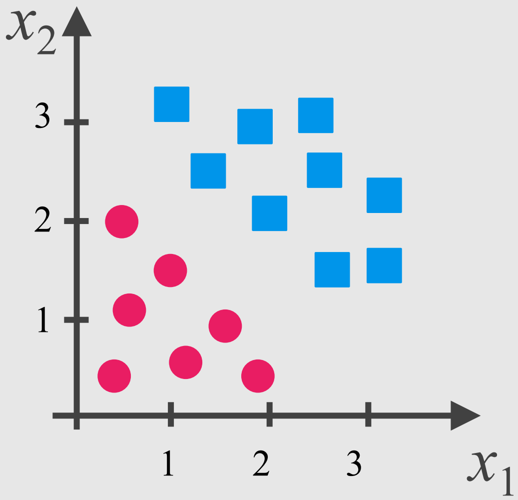


$$
\begin{array}{ccc} -3 & 1 & 1 \\ \Uparrow & \Uparrow & \Uparrow \end{array}
$$

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

Predict "$y = 1$" if $-3 + x_1 + x_2 \geq 0$

Decision Boundary



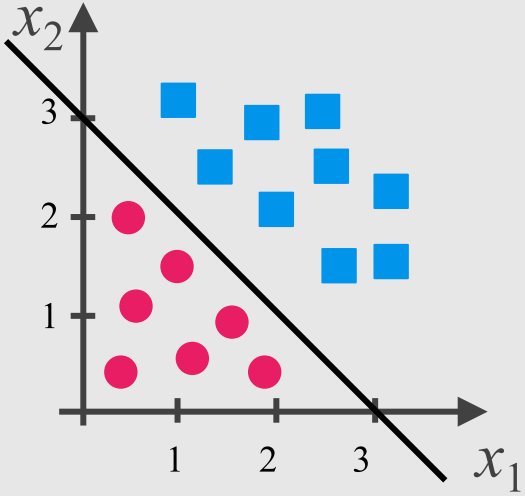
$$\begin{array}{ccc} -3 & 1 & 1 \\ \uparrow & \uparrow & \uparrow \end{array}$$

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

Predict "$y = 1$" if $-3 + x_1 + x_2 \geq 0$

$$x_1 + x_2 \geq 3$$

Decision Boundary

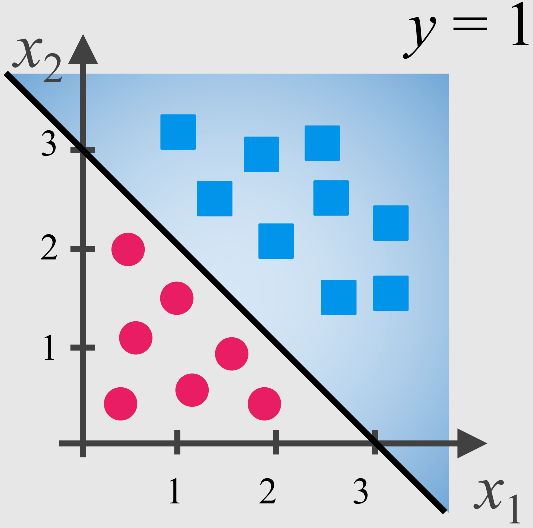


$$\begin{array}{c} \; -3 \;\;\;\;\;\;\; 1 \;\;\;\;\;\;\;\;\;\;\;\; 1 \\ \; \Uparrow \;\;\;\;\;\;\;\; \Uparrow \;\;\;\;\;\;\;\;\;\;\;\; \Uparrow \\ h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2) \end{array}$$

Predict "$y = 1$" if $-3 + x_1 + x_2 \geq 0$

$$x_1 + x_2 \geq 3$$

Decision Boundary



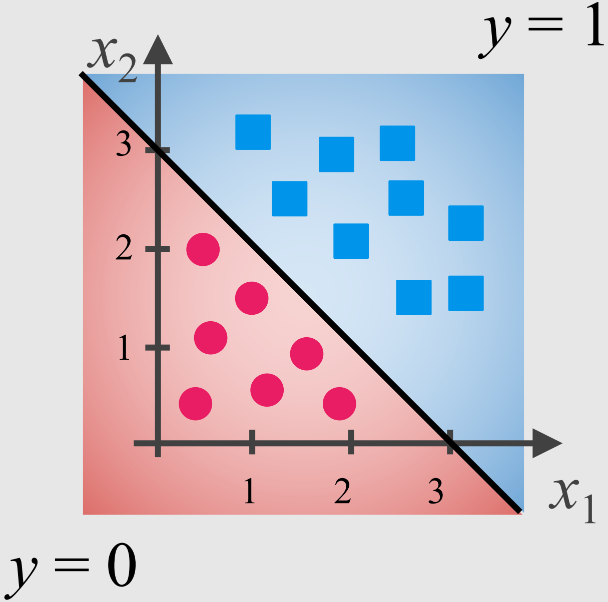
$$\begin{array}{c} \quad\quad\quad\quad -3 \quad\quad 1 \quad\quad\quad 1 \\ \quad\quad\quad\quad \Uparrow \quad\quad \Uparrow \quad\quad\quad \Uparrow \end{array}$$

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

Predict "$y = 1$" if $-3 + x_1 + x_2 \geq 0$

$$x_1 + x_2 \geq 3$$

Decision Boundary



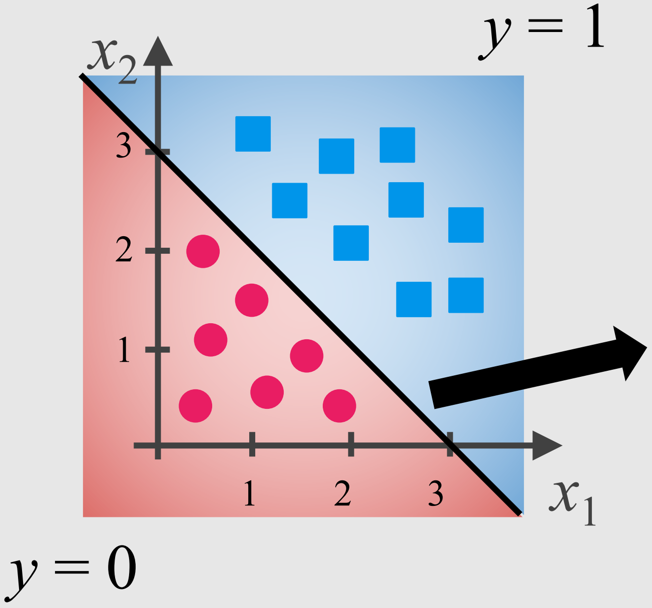
$$h_\theta(x) = g(\underset{\uparrow}{\overset{-3}{\theta_0}} + \underset{\uparrow}{\overset{1}{\theta_1}} x_1 + \underset{\uparrow}{\overset{1}{\theta_2}} x_2)$$

Predict "$y = 1$" if $-3 + x_1 + x_2 \geq 0$

$x_1 + x_2 \geq 3$

$y = 0,\ x_1 + x_2 < 3$

Decision Boundary



Decision Boundary

$$h_\theta(x) = g(\overset{-3}{\theta_0} + \overset{1}{\theta_1} x_1 + \overset{1}{\theta_2} x_2)$$

$$x_1 + x_2 = 3$$

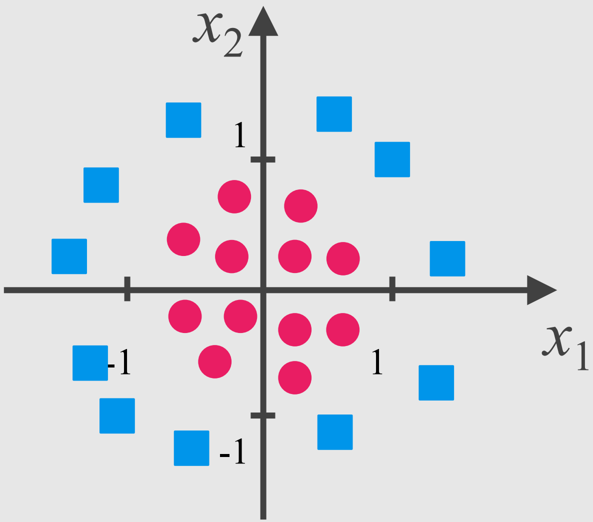
$$h_\theta(x) = 0.5$$

Predict "$y = 1$" if $-3 + x_1 + x_2 \geq 0$

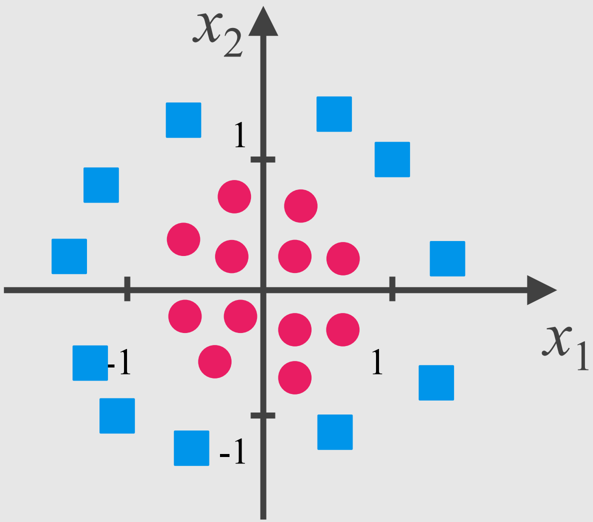
$$x_1 + x_2 \geq 3$$

$$y = 0,\ x_1 + x_2 < 3$$

Non-linear Decision Boundaries

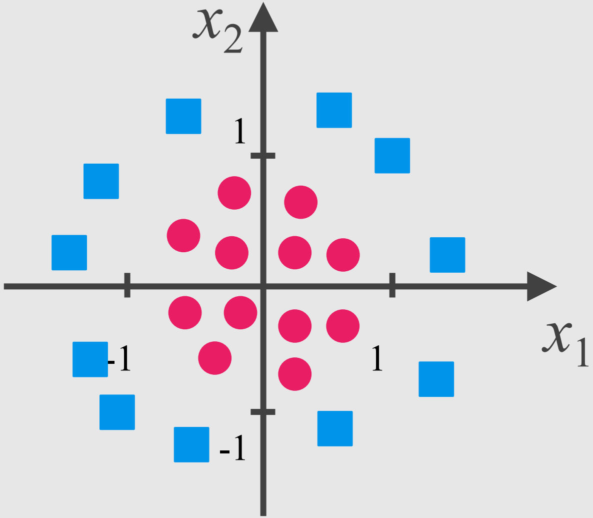


Non-linear Decision Boundaries



$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

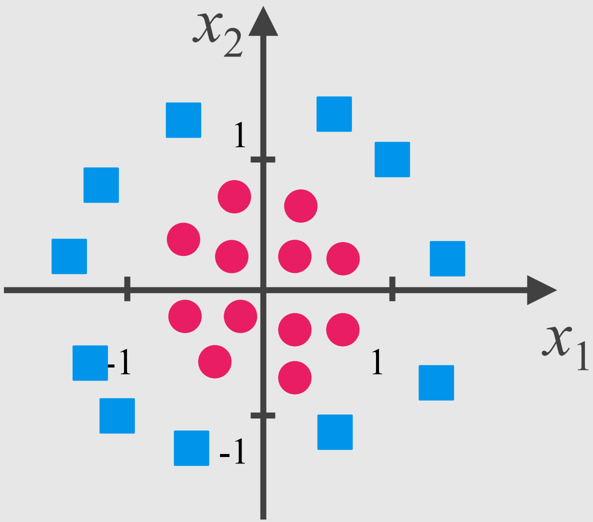
Non-linear Decision Boundaries



$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

$$\begin{array}{ccccc} \Downarrow & \Downarrow & \Downarrow & \Downarrow & \Downarrow \\ -1 & 0 & 0 & 1 & 1 \end{array}$$

Non-linear Decision Boundaries



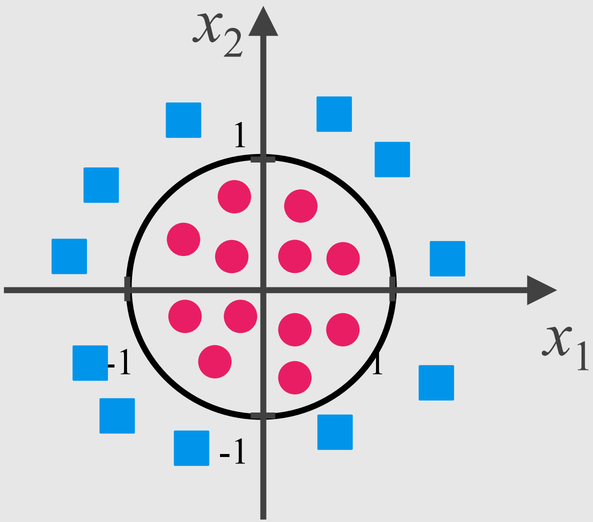
$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

θ_0	θ_1	θ_2	θ_3	θ_4
-1	0	0	1	1

Predict "$y = 1$" if $-1 + x_1^2 + x_2^2 \geq 0$

$$x_1^2 + x_2^2 \geq 1$$

Non-linear Decision Boundaries



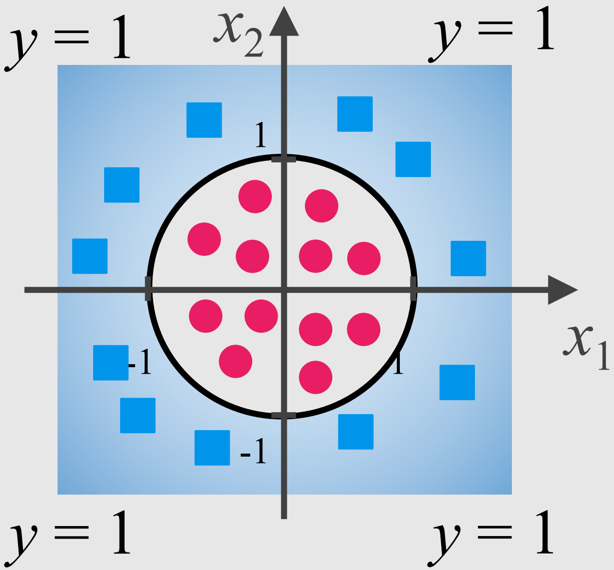
$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

$$\begin{array}{ccccc} \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ -1 & 0 & 0 & 1 & 1 \end{array}$$

Predict "$y = 1$" if $-1 + x_1^2 + x_2^2 \geq 0$

$$x_1^2 + x_2^2 \geq 1$$

Non-linear Decision Boundaries



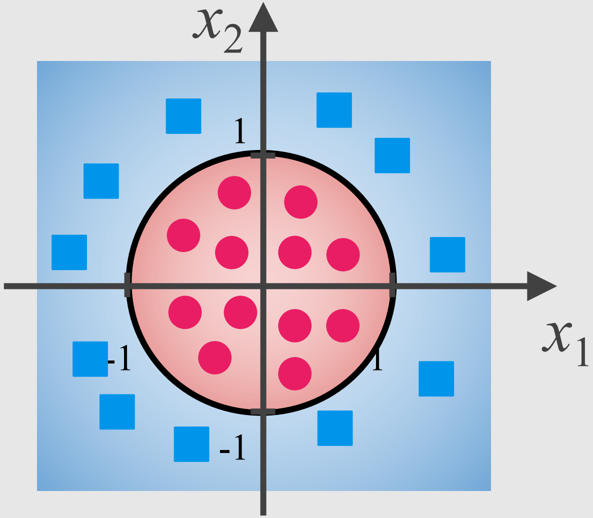
$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

$$\theta_0 = -1,\quad \theta_1 = 0,\quad \theta_2 = 0,\quad \theta_3 = 1,\quad \theta_4 = 1$$

Predict "$y = 1$" if $-1 + x_1^2 + x_2^2 \geq 0$

$$x_1^2 + x_2^2 \geq 1$$

Non-linear Decision Boundaries



$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

$$\begin{array}{ccccc} \theta_0 & \theta_1 & \theta_2 & \theta_3 & \theta_4 \\ \Downarrow & \Downarrow & \Downarrow & \Downarrow & \Downarrow \\ -1 & 0 & 0 & 1 & 1 \end{array}$$

Predict "$y = 1$" if $-1 + x_1^2 + x_2^2 \geq 0$

$$x_1^2 + x_2^2 \geq 1$$

Cost Function

Training set: $\{(x^{(1)},y^{(1)}),\ (x^{(2)},y^{(2)}),\ \ldots,\ (x^{(m)},y^{(m)})\}$

$$h_\theta(x) = \frac{1}{1+e^{-\theta^{\mathrm{T}}x}} \qquad x \in \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} \quad x_0 = 1,\ y \in \{0,1\}$$

How to choose parameters θ ?

Cost Function

Linear regression: $$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

Cost Function

$$\mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

Linear regression: $$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \boxed{\frac{1}{2}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2}$$

$$\mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)}) = \frac{1}{2}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

Cost Function

$$\mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

Linear regression: $J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \boxed{\frac{1}{2}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2}$

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

Cost Function

$$\mathrm{Cost}(h_\theta(x^{(i)}),\ y^{(i)})$$

~~Linear~~ regression:
Logistic

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\boxed{\frac{1}{2}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2}$$

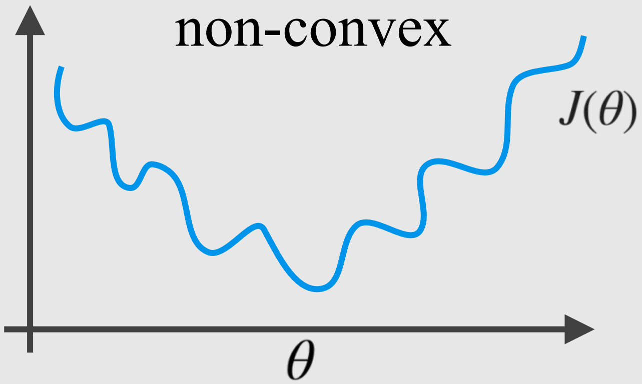
$$\mathrm{Cost}(h_\theta(x),\ y) = \frac{1}{2}(h_\theta(x) - y)^2 \qquad h_\theta(x) = \frac{1}{1+\mathrm{e}^{-\theta^{\mathrm{T}}x}}$$

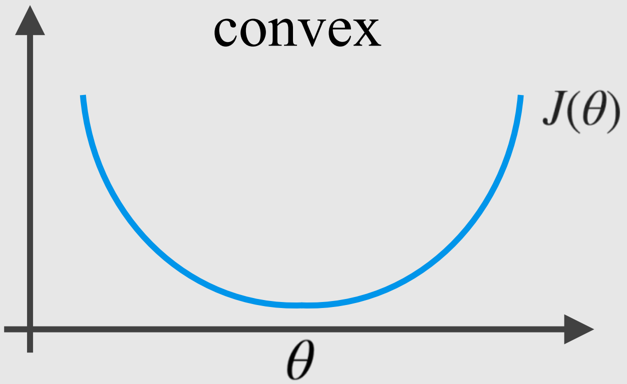
Cost Function

$$\mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

Logistic regression: $$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \boxed{\frac{1}{2}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2}$$

$$\mathrm{Cost}(h_\theta(x), y) = \frac{1}{2}\left(h_\theta(x) - y\right)^2 \qquad h_\theta(x) = \frac{1}{1+e^{-\theta^{\mathrm{T}}x}}$$





Derivative of Logistic Function

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

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

$$= \frac{0\cdot(1+e^{-z}) - 1\cdot(-e^{-z})}{(1+e^{-z})^2} \qquad \text{(quotient rule)}$$

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

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

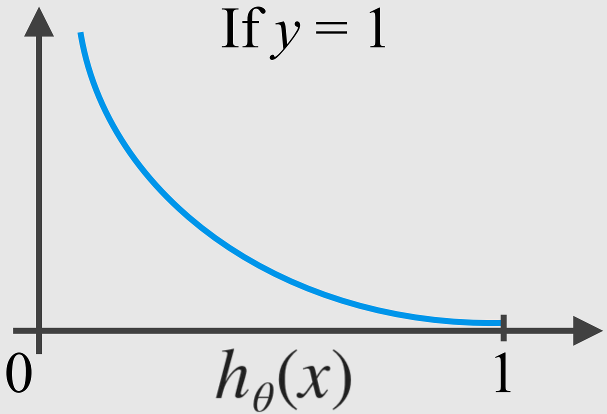
$$= g(z)(1-g(z))$$

Logistic Regression Cost Function

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

Logistic Regression Cost Function

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



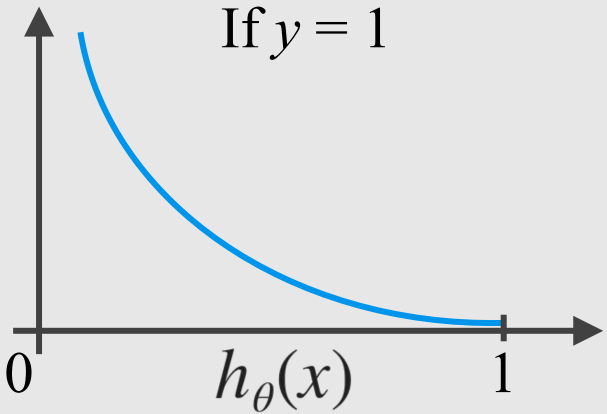
Cost $= 0$ if $y = 1, h_\theta(x) = 1$

But as $h_\theta(x) \to 0$

Cost $\to \infty$

Logistic Regression Cost Function

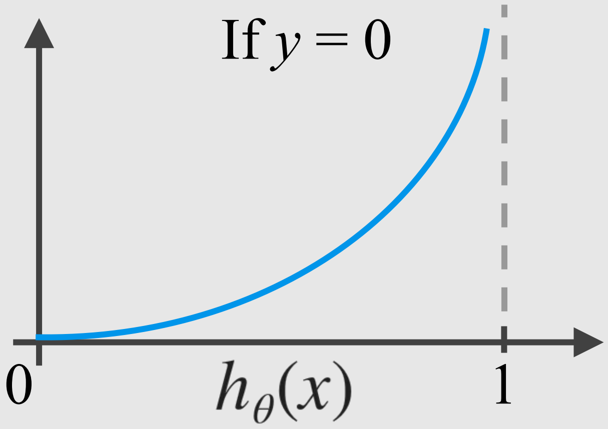
$$\mathrm{Cost}(h_\theta(x),\ y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$



Captures intuition that if $h_\theta(x) \to 0$, (predict $P(y = 1 \mid x;\theta) = 0$), but $y = 1$, we'll penalize learning algorithm by a very large cost.

Logistic Regression Cost Function

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$



Simplified Cost Function and Gradient Descent

Logistic Regression Cost Function

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}),\ y^{(i)})$$

$$\mathrm{Cost}(h_\theta(x),\ y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

Logistic Regression Cost Function

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}),\ y^{(i)})$$

$$\mathrm{Cost}(h_\theta(x),\ y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

$$\mathrm{Cost}(h_\theta(x),\ y) = -y\log(h_\theta(x)) - (1-y)\log(1 - h_\theta(x))$$

Logistic Regression Cost Function

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

$$\mathrm{Cost}(h_\theta(x), y) = -y\log(h_\theta(x)) - (1-y)\cancel{\log(1 - h_\theta(x))}$$

$$y = 1$$

Logistic Regression Cost Function

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

$$\mathrm{Cost}(h_\theta(x), y) = -y\cancel{\log(h_\theta(x))} - (1-y)\log(1 - h_\theta(x))$$

$$y = 0$$

Logistic Regression Cost Function

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}),\ y^{(i)})$$

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

Logistic Regression Cost Function

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\mathrm{Cost}(h_\theta(x^{(i)}),\ y^{(i)})$$

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

To fit parameters θ : $\min_{\theta} J(\theta)$

Logistic Regression Cost Function

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}),\ y^{(i)})$$

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

To fit parameters θ : $\min_\theta J(\theta)$

To make a new prediction given new x: Output $h_\theta(x) = \frac{1}{1 + \mathrm{e}^{-\theta^\mathrm{T} x}}$

Gradient Descent

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

Want $\min_\theta J(\theta)$:

repeat {

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial\theta_j} J(\theta)$$

} (simultaneously update θ_j for $j = 0, 1, \ldots, n$)

Gradient Descent

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

Want $\min_\theta J(\theta)$

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

repeat {

$$\theta_j := \theta_j - \alpha\boxed{\frac{\partial}{\partial\theta_j}J(\theta)}$$

} (simultaneously update θ_j for $j = 0, 1, \ldots, n$)

Gradient Descent

https://math.stackexchange.com/questions/477207/derivative-of-cost-function-for-logistic-regrssion

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

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

repeat {

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

} (simultaneously update θ_j for $j = 0, 1, \ldots, n$)

Gradient Descent

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

Want $\min_\theta J(\theta)$:

repeat {

$$\theta_j := \theta_j - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$$

} (simultaneously update θ_j for $j = 0, 1, \ldots, n$)

Gradient Descent

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

Want $\min_\theta J(\theta)$:

repeat {

$$\theta_j := \theta_j - \alpha \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$$

} (simultaneously update θ_j for $j = 0, 1, \ldots, n$)

Algorithm looks identical to linear regression!

Gradient Descent

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

Want $\min_\theta J(\theta)$

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

repeat {

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

} (simultaneously update θ_j for $j = 0, 1, \ldots, n$)

Algorithm looks identical to linear regression!

Multiclass Classification: One-vs-all

Classification

Email tagging: Work, Friends, Family

Skin Lesion: Melanoma, Carcinoma, Nevus, Keratosis

Video: Pornography, Violence, Gore scenes, Child abuse

Classification

Email tagging: Work, Friends, Family

$y = 1 \quad y = 2 \quad y = 3$

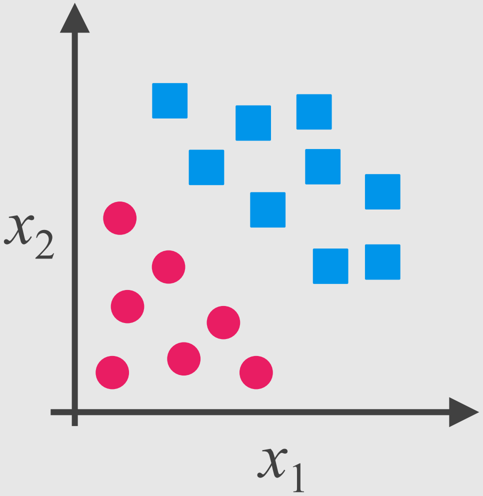
Skin Lesion: Melanoma, Carcinoma, Nevus, Keratosis

$y = 1 \quad y = 2 \quad y = 3 \quad y = 4$

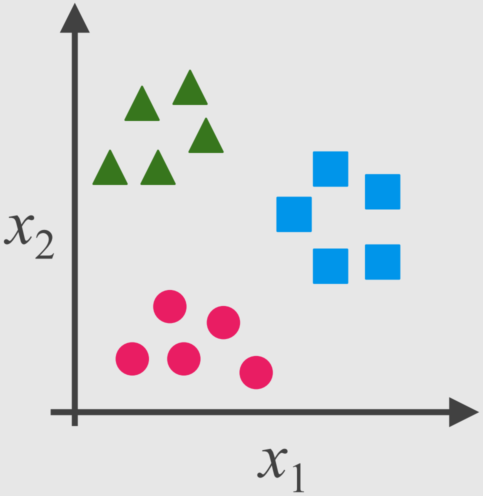
Video: Pornography, Violence, Gore scenes, Child abuse

Binary Classification

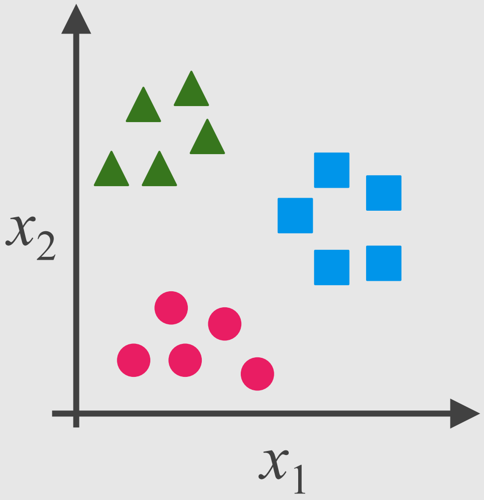
Classification



Multi-class



One-vs-All (One-vs-Rest)

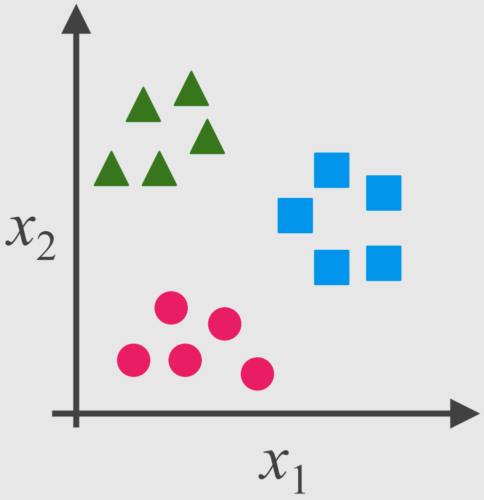


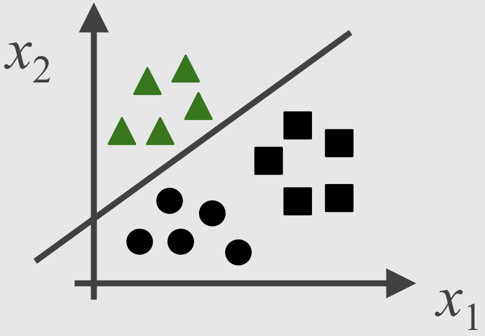
Class 1:

Class 2:

Class 3:

One-vs-All (One-vs-Rest)



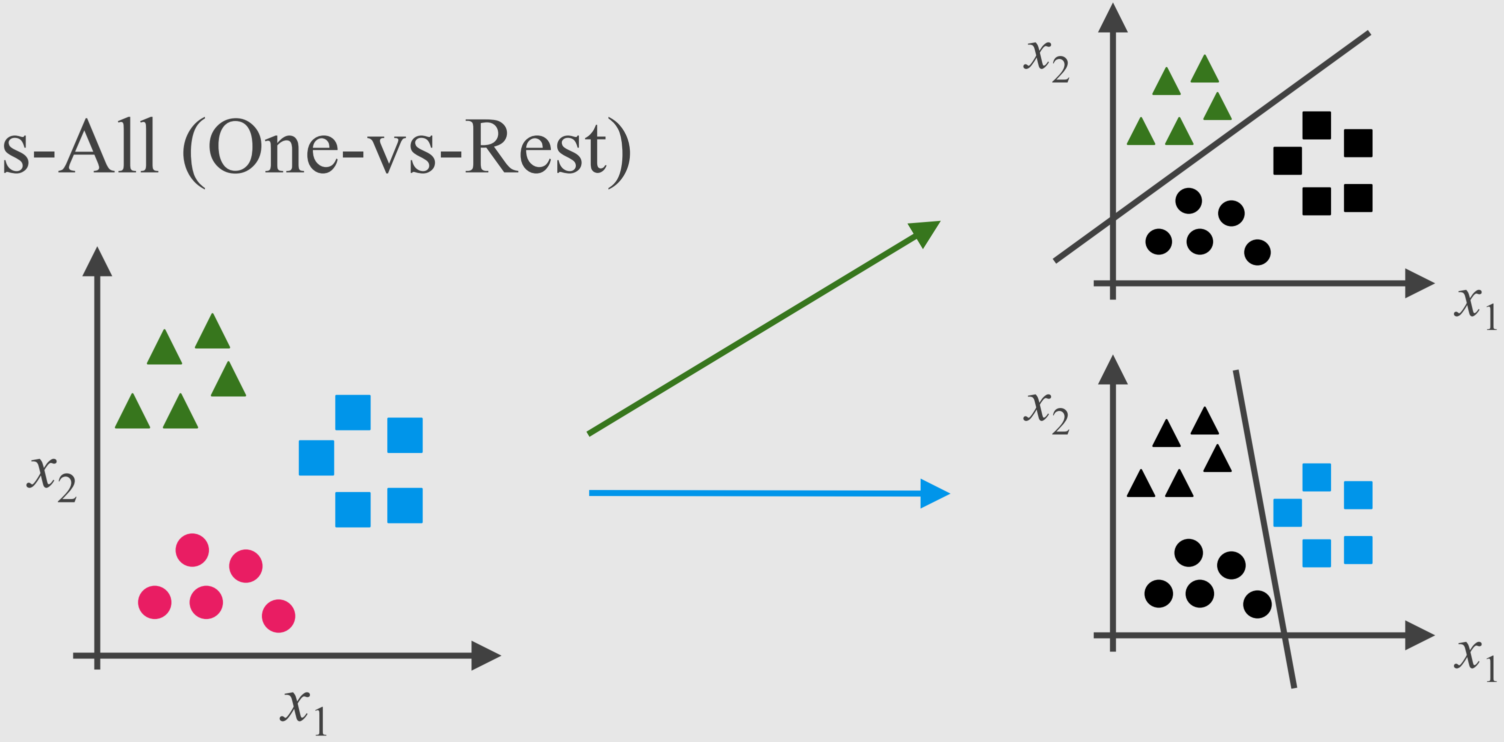


Class 1: ▲

Class 2: ■

Class 3: ●

One-vs-All (One-vs-Rest)

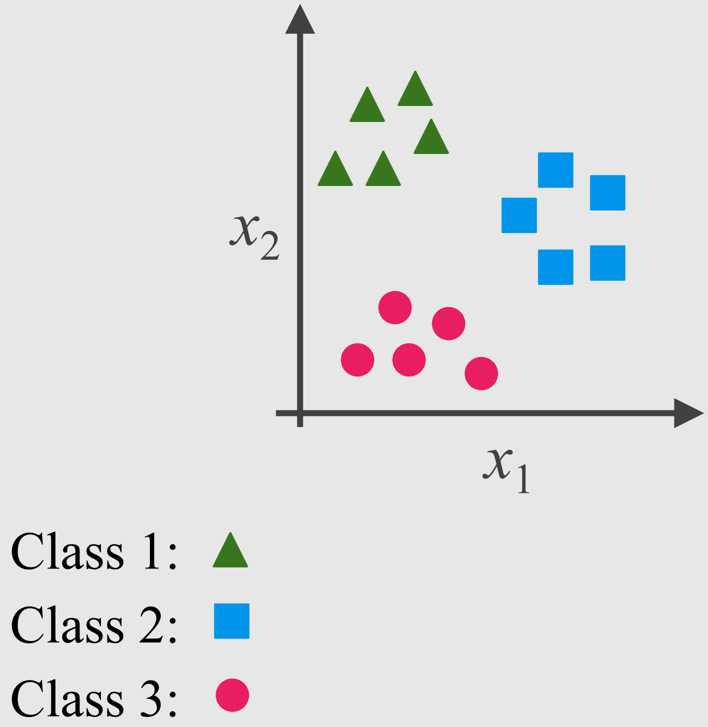


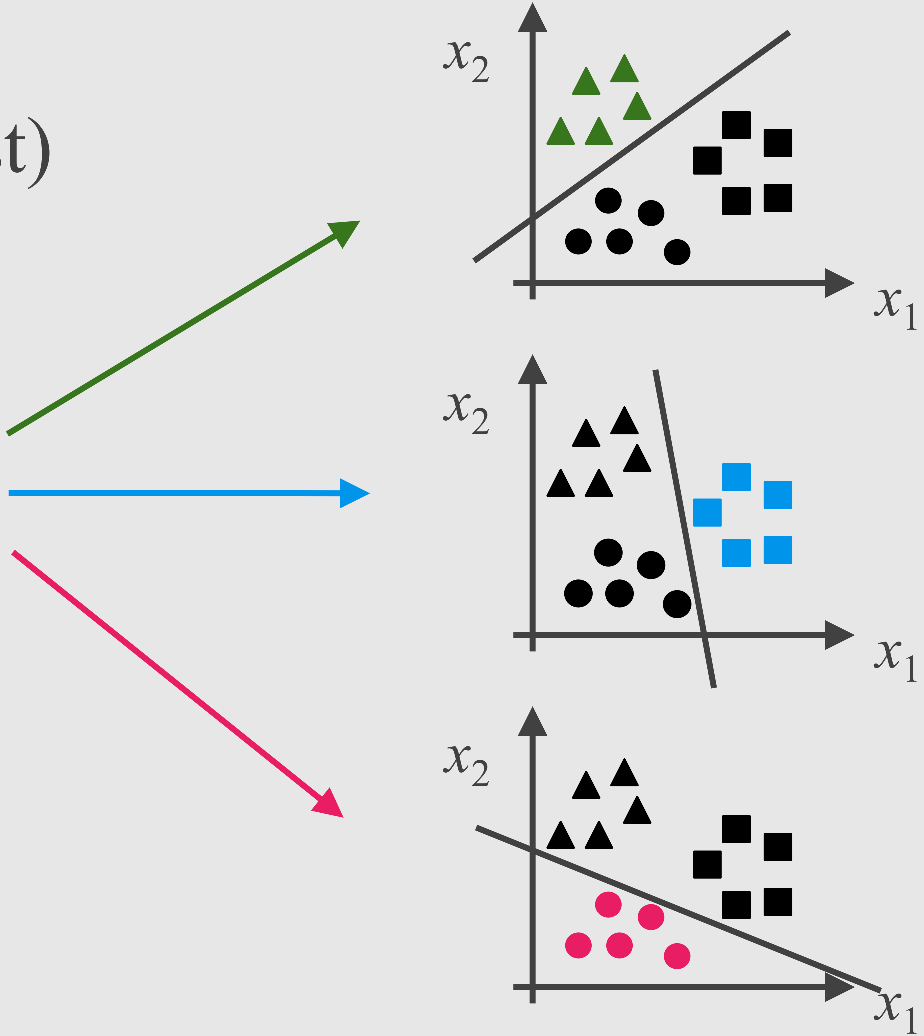
Class 1: ▲

Class 2: ■

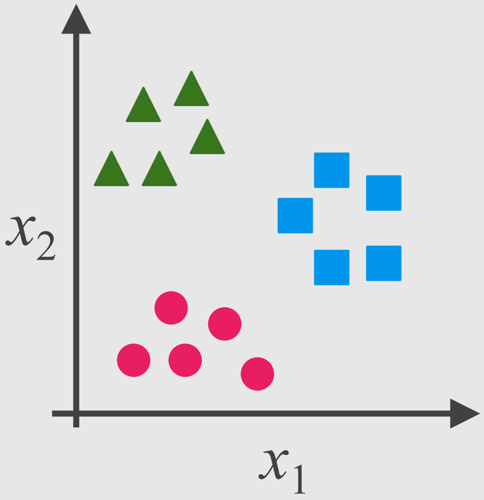
Class 3: ●

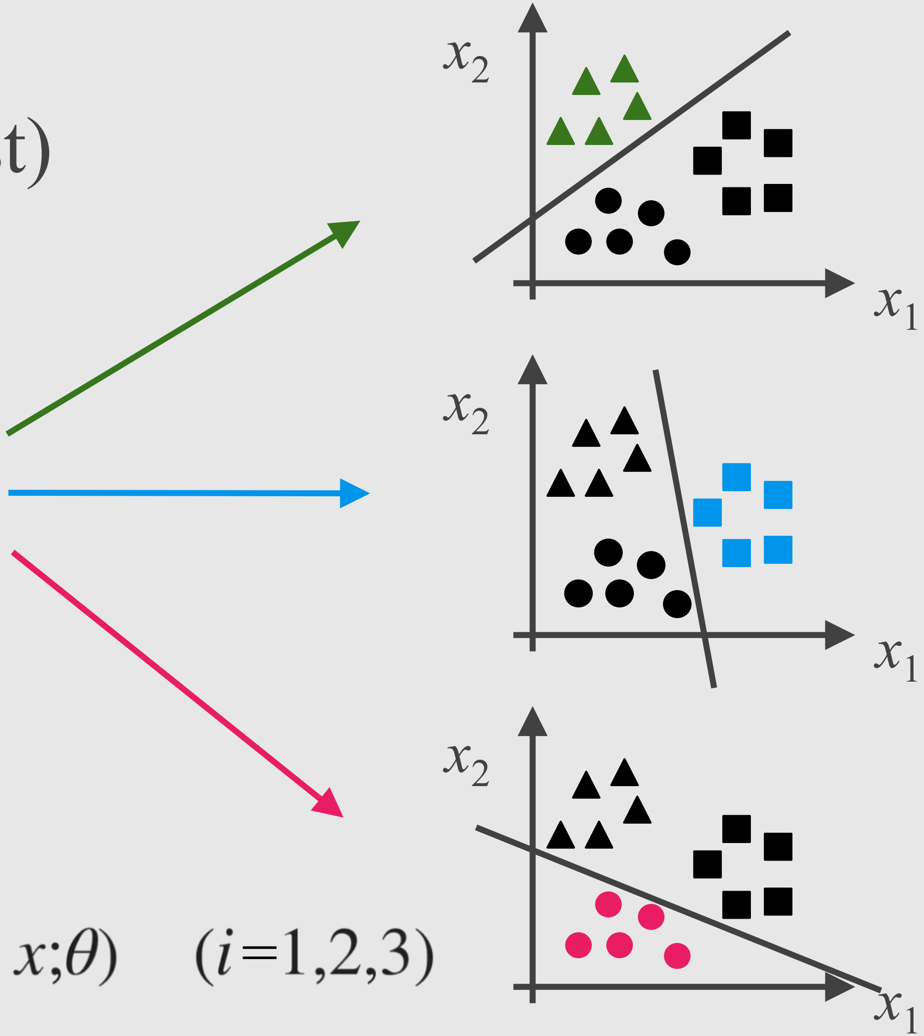
One-vs-All (One-vs-Rest)





One-vs-All (One-vs-Rest)





Class 1:

Class 2:

Class 3:

$$h_\theta^{(i)}(x) = P(y = i \mid x;\theta) \quad (i=1,2,3)$$

One-vs-All (One-vs-Rest)

Train a logistic regression classifier $h_\theta^{(i)}(x)$ for each class i to predict the probability that $y = i$.

One a new input x, to make a prediction, pick the class i that maximizes

$$\max_i \; h_\theta^{(i)}(x)$$

References

Machine Learning Books

- Hands-On Machine Learning with Scikit-Learn and TensorFlow, Chap. 4
- Pattern Recognition and Machine Learning, Chap. 4

Machine Learning Courses

- https://www.coursera.org/learn/machine-learning, Week 3