Mosaic + Cassandra

Workshop 2

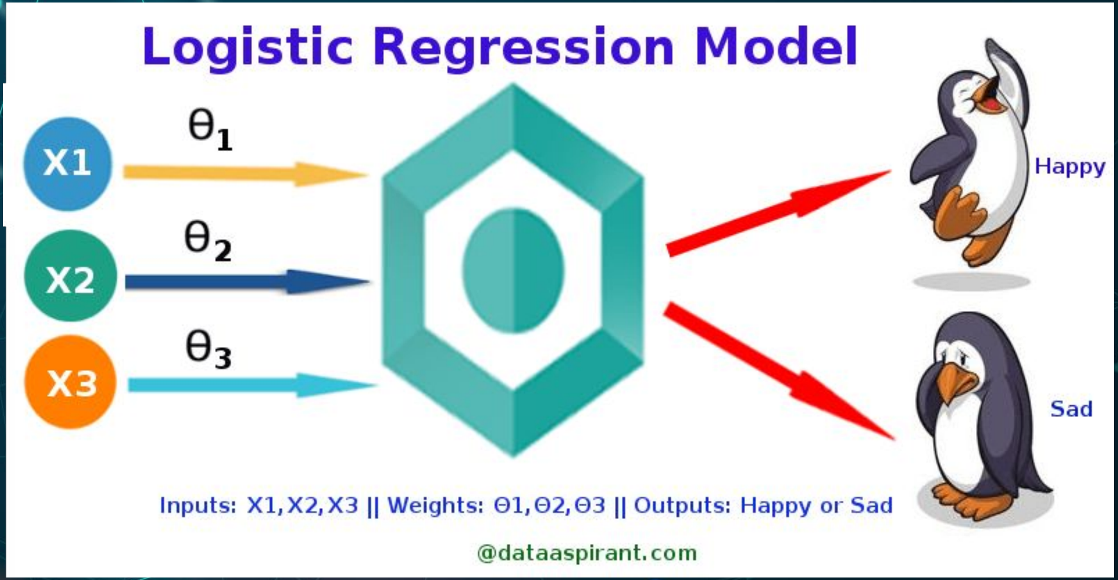
Logistic Regression

We Will basically divide the discussion into 3 major parts

- What is Logistic Regression?
- Why Logistic Regression?
- How is it Different?

What is Logistic Regression ?

Logistic Regression is basically the tweaked version of Linear Regression wherein at the output of Linear Regression a sigmoid function is applied to make a Binary Predictive Model.



Classification

Recap: Classification

- Email: Spam / Not Spam?
- Online Transactions: Fraudulent (Yes / No)?
- Tumor: Malignant / Benign ?

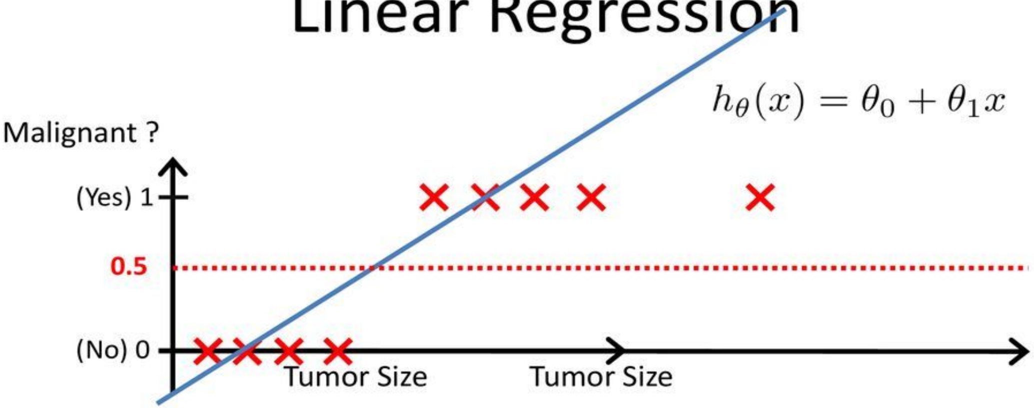
$y \in \{0, 1\}$

0: “Negative Class” (e.g., benign tumor)
1: “Positive Class” (e.g., malignant tumor)

Why ?

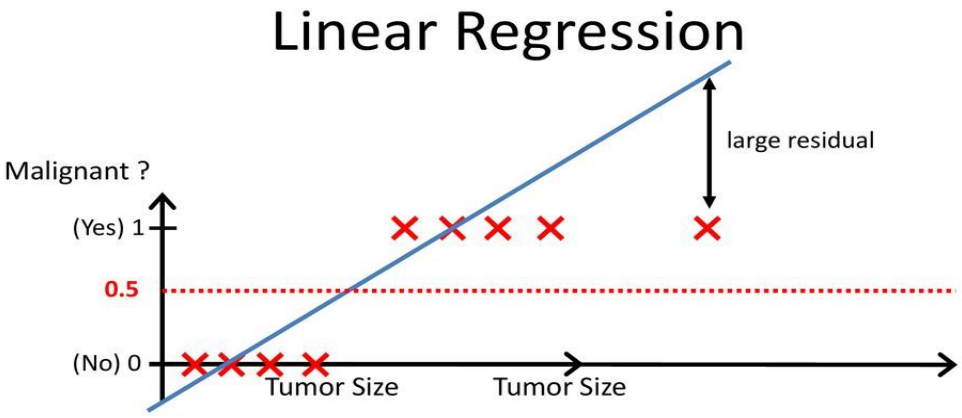
Linear Regression

$$h_\theta(x) = \theta_0 + \theta_1 x$$



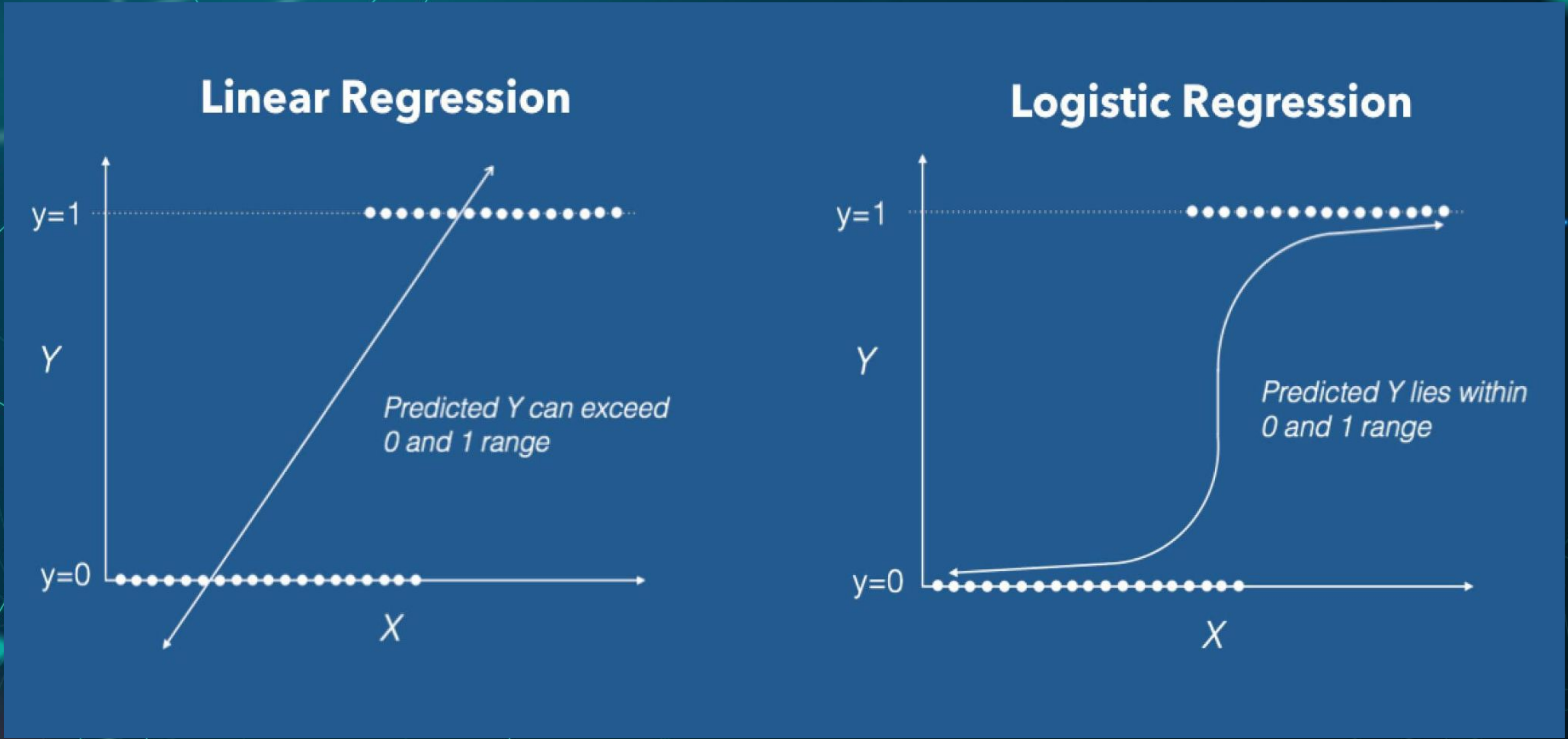
- Threshold classifier output $h_\theta(x)$ at 0.5:
  - If $h_\theta(x) \geq 0.5$, predict "y = 1"
  - If $h_\theta(x) < 0.5$, predict "y = 0"

Problems with Linear Regression



- **Problem**: if y=1, it is more unhappy about h(x)=10 than h(x)=0
- The loss function hates when you make correct predictions with high confidence!

Large Residual Hence abrupt change in the Regression function.



Predicted Y can exceed 0 and 1 range

Logistic Regression Model

Logistic Regression Model

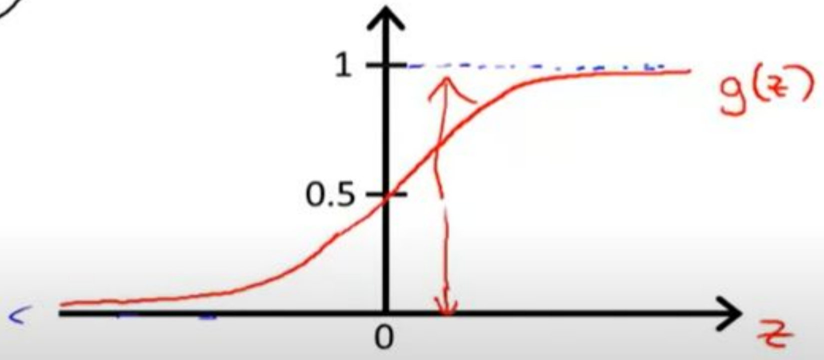
Want $\boxed{0 \leq h_\theta(x) \leq 1}$

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

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

$\theta^T x$

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

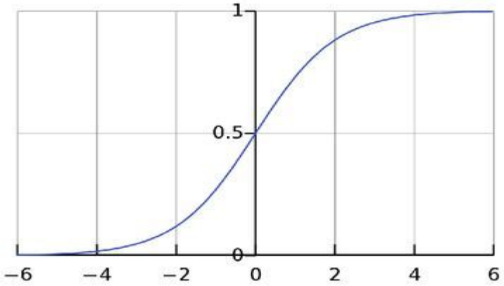


Parameters θ.

→ Sigmoid function
→ Logistic function

Decision Boundary

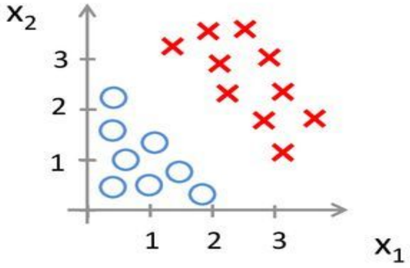
$$h_\theta(x) = g(\theta^T x)$$
$$g(z) = \frac{1}{1+e^{-z}}$$



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

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

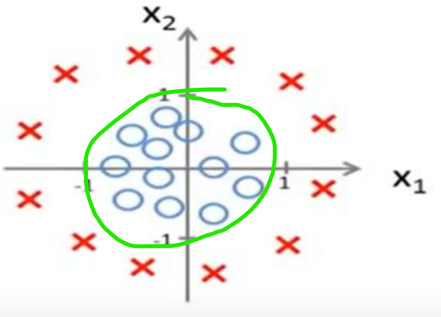
Examples



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

Example:

$\theta^T = [-3, 1, 1]$ ➡ Predict "$y = 1$" if $-3 + x_1 + x_2 \geq 0$



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

Exercise: how does the decision boundary look like?

$$\theta^T = [-1, 0, 0, 1, 1]$$

How?

Cost Function

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

m examples $\quad x \in \begin{bmatrix} x_0 \\ x_1 \\ \cdots \\ x_n \end{bmatrix} \quad x_0 = 1, y \in \{0, 1\}$

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

How to choose parameters θ ?

The Problem With Linear Reg. Cost Function

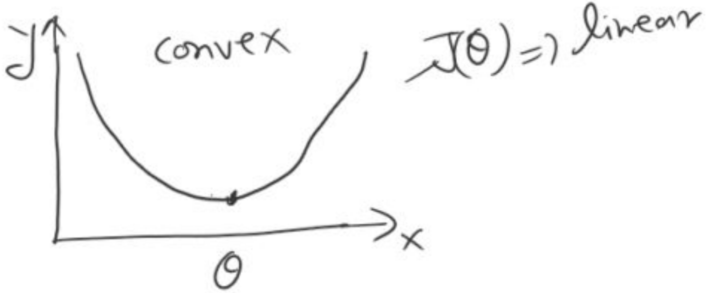
Linear Regression cost function

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

So for Logistic regression we can write

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

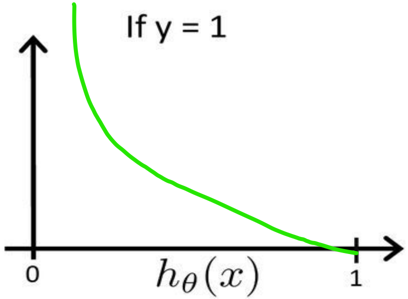
$$cost\left(h_\theta x^i ; y\right) = \frac{1}{2}\left(h_\theta(x) - y\right)^2$$



Logistic Regression Cost Function

Cost Function

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



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

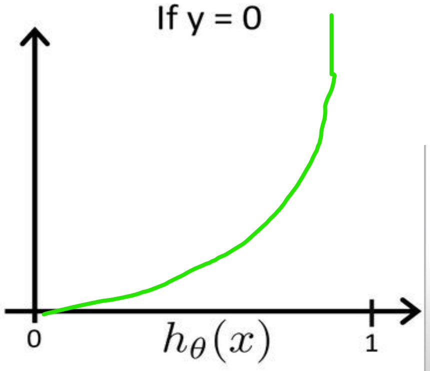
But as $h_\theta(x) \to 0$

$Cost \to \infty$

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

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 $y = 0, h_\theta(x) = 0$
But as $h_\theta(x) \rightarrow 1$
$Cost \rightarrow \infty$

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

Simplified Cost Function

Logistic regression cost function

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

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

Note: $y = \boxed{0}$ or $\boxed{1}$ always

$$\rightarrow \mathrm{Cost}(h_\theta(x), y) = \underbrace{-y \log(h_\theta(x))}_{=0} - \underbrace{(1-y)}_{=0} \underbrace{\log(1 - h_\theta(x))}_{=1}$$

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

If $y=0$: $\mathrm{Cost}(h_\theta(x), y) = -\log(1 - h_\theta(x))$

Moving Towards Gradient Descent

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

$$\begin{aligned} J(\theta) &= \frac{1}{m} \sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)}) \\ &= -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))] \end{aligned}$$

To fit parameters θ:

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

Gradient Descent

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 all θ_j)

}

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

Gradient Descent

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

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix}$$

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

Repeat {

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

(simultaneously update all θ_j)

}

Algorithm looks identical to linear regression!

Multiclass Classification

Email foldering/tagging: Work, Friends, Family, Hobby

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

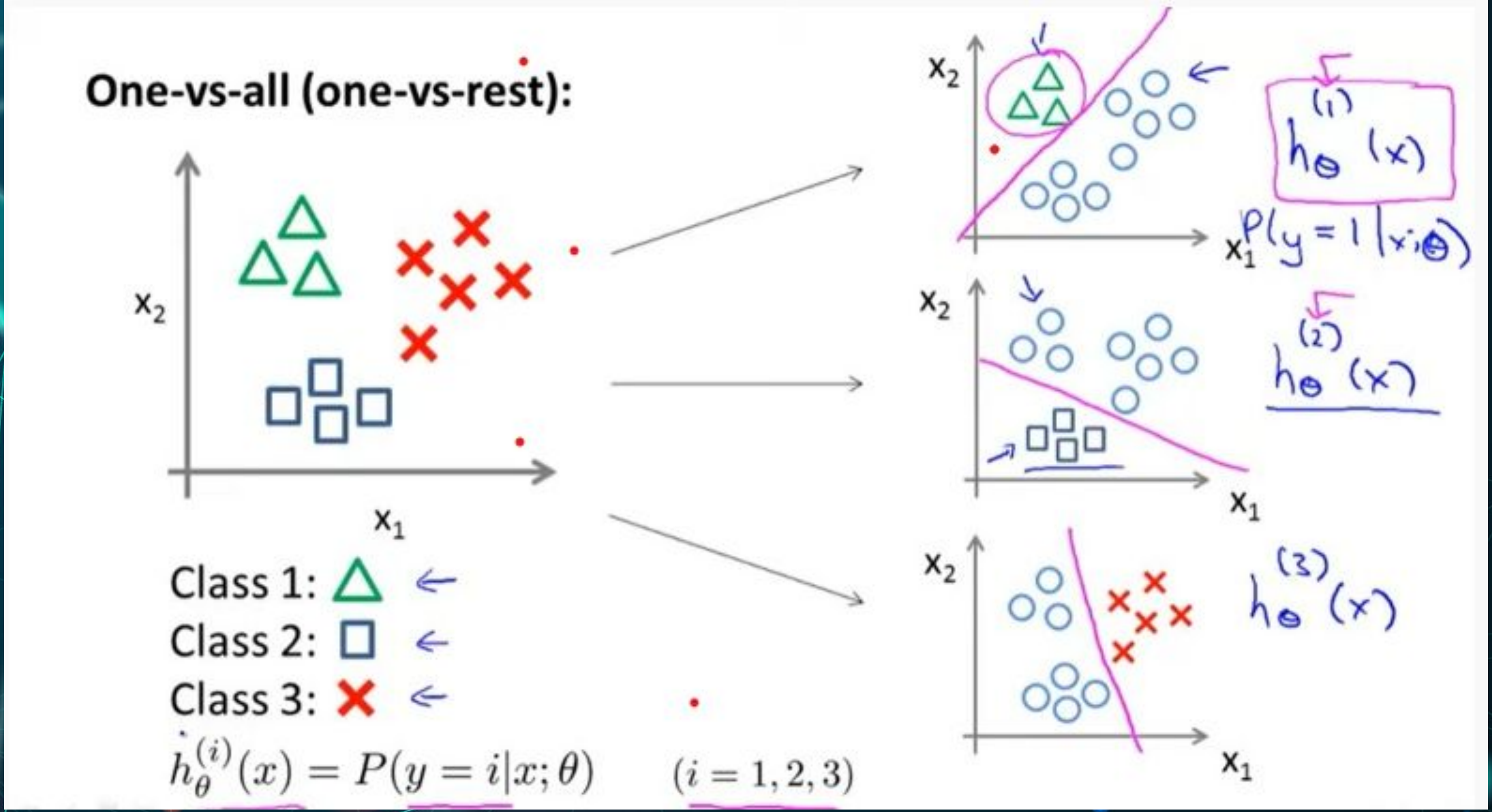
Medical diagrams: Not ill, Cold, Flu

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

Weather: Sunny, Cloudy, Rain, Snow

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

One Vs All



One-vs-all (one-vs-rest):
x_2
x_1
Class 1: △
Class 2: □
Class 3: ×
$h_\theta^{(i)}(x) = P(y = i|x;\theta)$
$(i = 1, 2, 3)$
$h_\theta^{(1)}(x)$
$P(y=1|x;\theta)$
$h_\theta^{(2)}(x)$
$h_\theta^{(3)}(x)$

Final Approach

One-vs-all

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

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

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

Assignment Discussion