

Classification and Logistic Regression

By Vipul Goyal

Logistic Regression

Linear Regression: Prediction

- $y \in \mathbb{R}$ is a continuous variable
- Stock market data, housing price, ..

Logistic regression: classification (despite the name)

- The label y is a discrete (typically small) variable

Spam Filtering



Nikki Wypych <nikki.wypych@accountingprincipals.com>
to Iuliia.getman1 ▾

Fri, Mar 5, 2:47 PM (1 day ago)



Why is this message in spam? It is similar to messages that were identified as spam in the past.

Report not spam



Happy Friday Iuliia-

I hope this email finds you well. I am hoping to network with you today and ask that you **only respond to this email if you are interested in the project.**

We are partnering with one of our clients **on a 2-4-week project** assisting with vaccine distribution in Pennsylvania. We will need to hire almost 50 associates across the state for Patient Administrator openings. The Patient Administrator interacts with individuals interested in registering to receive a COVID vaccination and ensures all of the relevant paperwork is completed in full. They will collect and enter patient data into the provided vaccination information system in an accurate and expeditious manner. They will also be responsible for maintaining and tracking electronic records and logs.

Compensation: We can pay \$15.00/hr

Location: We will try to match you up to the closest location to you. ***This job will be IN PERSON, not from home!***

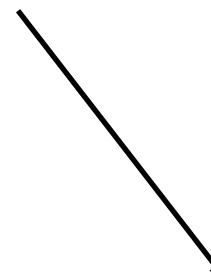
*****Multiple locations - we especially need people in Central and Northern PA!!***

The hours are Tues-Sat, 8am-5pm. Some site shifts could be 9-6 or 10-7 and will take candidates that can't work weekends and only Part-Time as well.

There will be no interview process for this - all YOU need to do is respond back to me letting me know you are interested in helping with this initiative, we will set up an interview, and I will give you a call to explain further details.

What a great way to give back to your local community and if you are not working, a great way to make some quick money!

If you are not interested in this and know someone that may be interested, feel free to pass along my contact information and they can reach out to me directly.



Spam vs.
Not Spam

Types of Logistic Regression

Types of Logistic regression

- Binary (Email: Spam/Not Spam, rain/no rain)
- Multi (Cats, Dogs, Sheep)
- Ordinal (Low, Medium, High)

Tumor Example

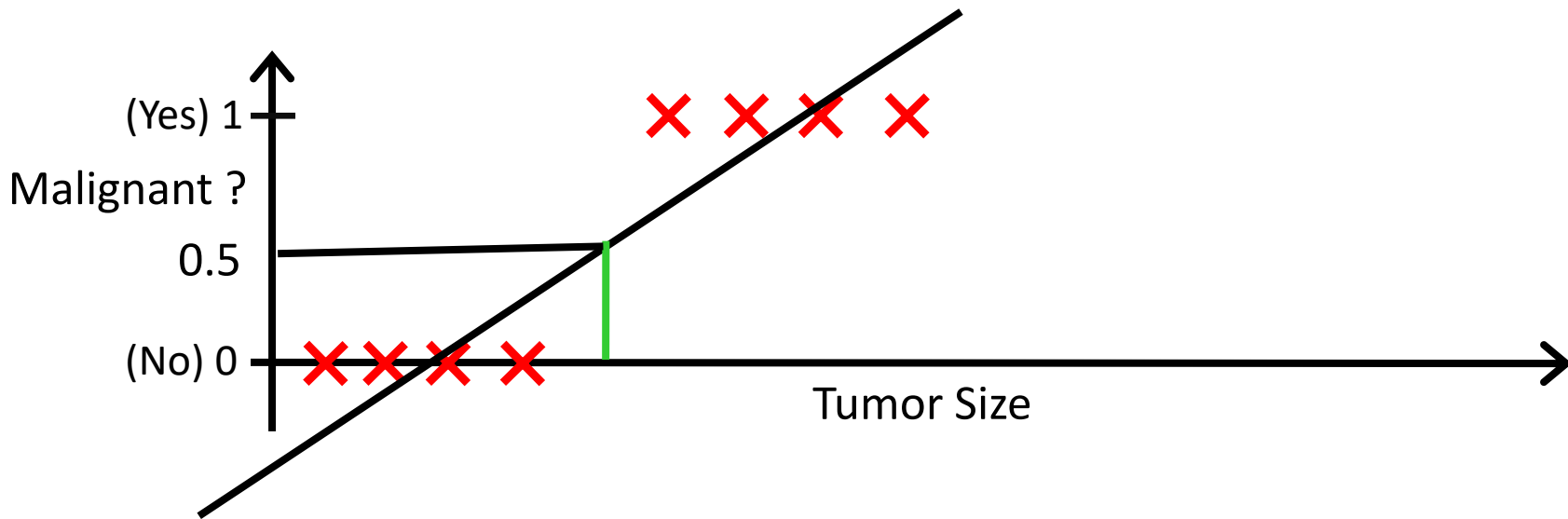
Given tumor size/features: Malignant / Benign ?

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

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

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

First Idea: Just use Linear Regression?



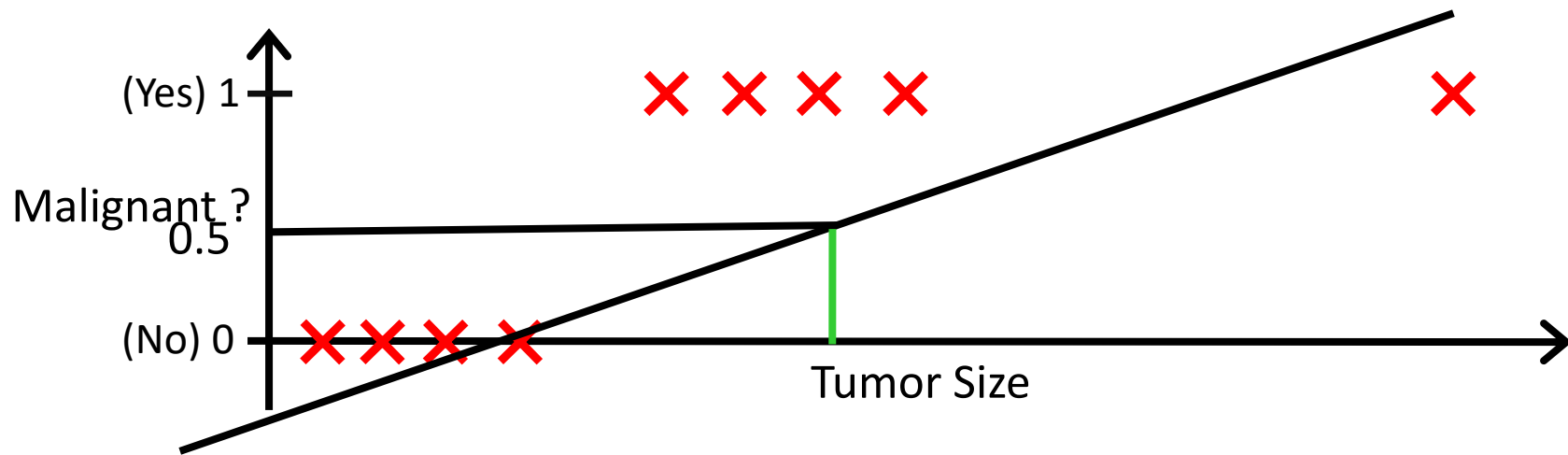
Threshold classifier output $h_{\theta}(x)$ at 0.5:

If $h_{\theta}(x) \geq 0.5$, predict “y = 1”

If $h_{\theta}(x) < 0.5$, predict “y = 0”

Anything to the right of the green line classified as malignant

The Problem



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 “ $y = 0$ ”

Green line doesn't give correct prediction

Classification Problem

Linear Regression: $h_{\theta}(x)$ can be > 1 or < 0

Classification: $y = 0$ or 1

- So ideally: $h_{\theta}(x)$ is either 0 or 1

But that is not always possible, neither desirable. Why?

- We might also want some measure of confidence
- So ideally $h_{\theta}(x)$ represents the probability

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

Interpreting the Output of h

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

Example: If $x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ tumor\ size \end{bmatrix}$

$$h_{\theta}(x) = 0.7$$

Tell patient that 70% chance of tumor being malignant

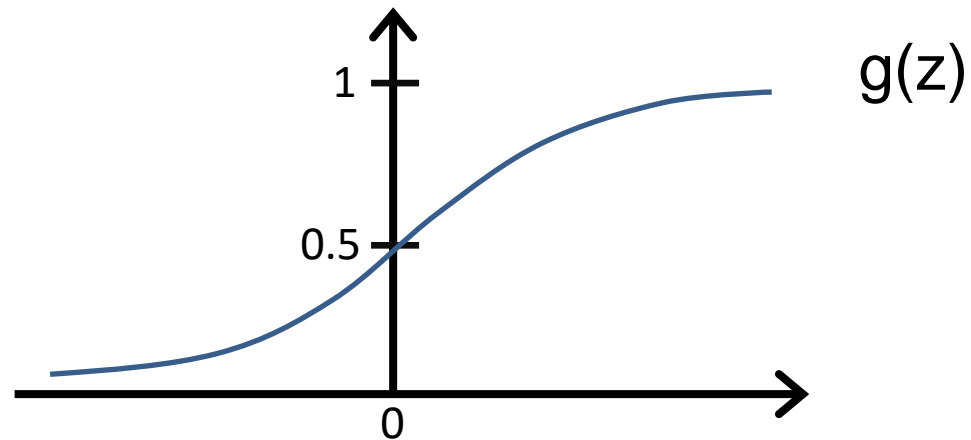
Defining the Hypothesis

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

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

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

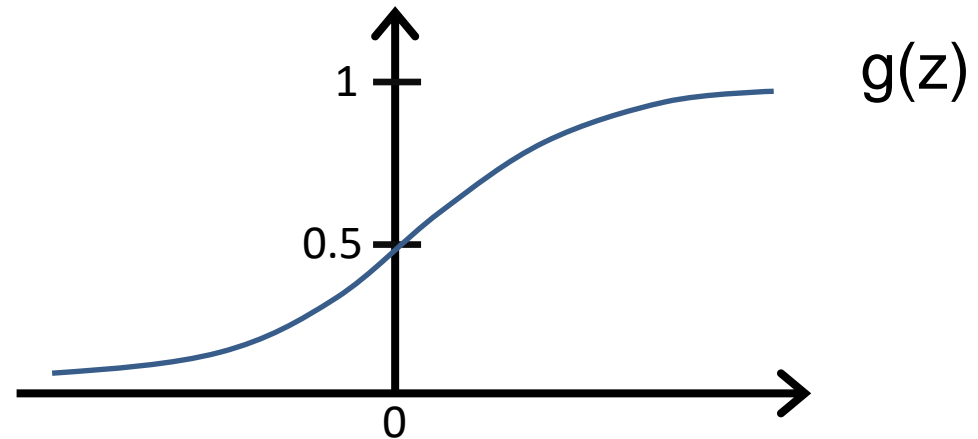
$$g(z) = \frac{1}{1 + e^{-z}}$$



Sigmoid function or Logistic function

Sigmoid Function

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

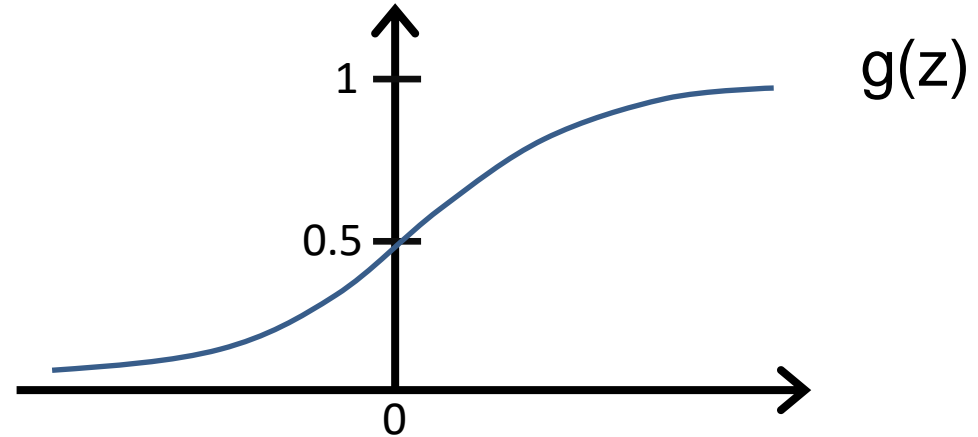


- If z is (large) $-ve$, e^{-z} is very large. Hence $g(z)$ is very close to 0
- If z is 0, e^{-z} is 1. Hence $g(z)$ is $\frac{1}{2}$
- If z is (large) $+ve$, e^{-z} is very small. Hence $g(z)$ is very close to 1

How to Predict using h

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

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



- If $g(z) > \frac{1}{2}$, we predict $y=1$, else we predict $y=0$
- As observed: If z is 0, $g(z)$ is $\frac{1}{2}$

Have We Achieved Anything?

- If $g(z) > \frac{1}{2}$, we predict $y=1$, else we predict $y=0$
- As observed: If z is 0, $g(z)$ is $\frac{1}{2}$
- How about: If $z > 0$, we predict $y=1$, else we predict $y=0$?
- It will achieve the same result! And $z = \theta^T x$
- So why did we bother introducing the strange looking logistic function g ?
 - Answer: good observation! But we haven't talked about the cost yet

Cost Function in Logistic Regression

Summary of the Question

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

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

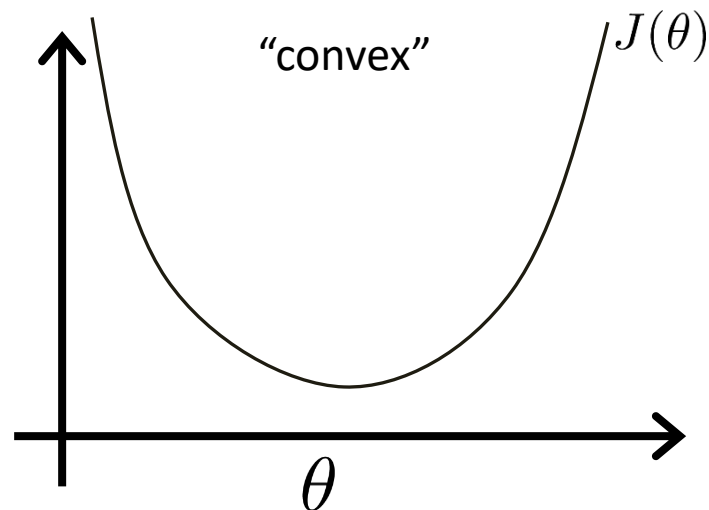
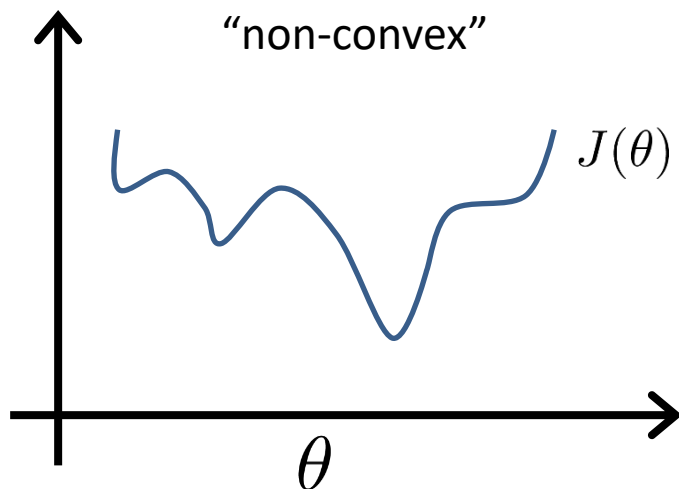
$$h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$

How to choose parameters θ ?

Defining a Cost Function

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

Use same J but with new $h_{\theta}(x)$?



Using a Different Cost Function

Linear Regression:

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

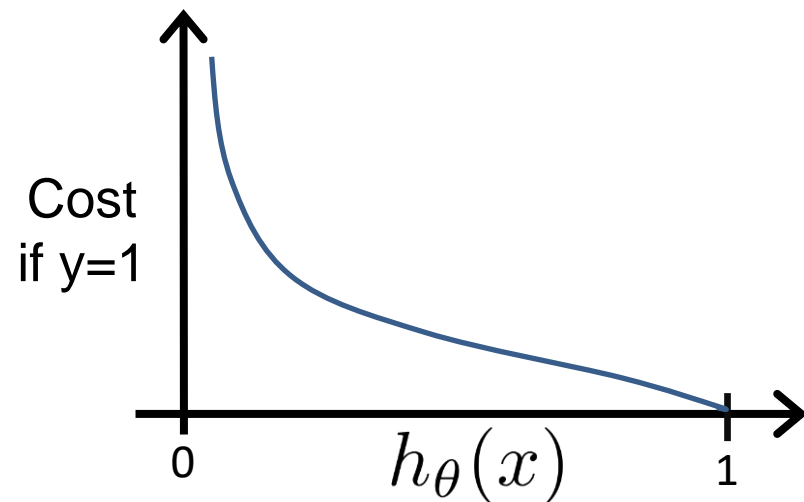
- Instead of Mean Squared Error, we use a cost function called Cross-Entropy, also known as Log Loss.
- Cross-entropy loss can be divided into two separate cost functions: one for $y=0$ and another for $y=1$

Logistic Regression Cost Function

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

if $h_{\theta}(x) = 1$ and $y=1$, cost = $-\log(1) = 0$

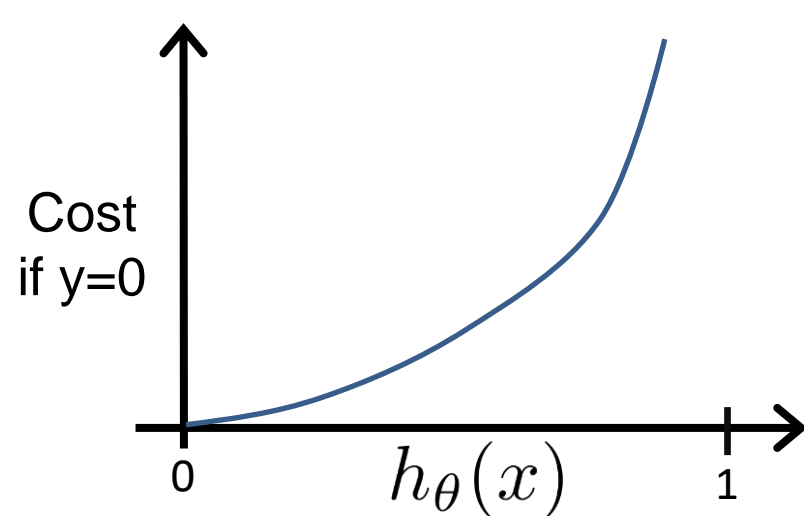
if $h_{\theta}(x) \rightarrow 0$ and $y=1$, cost $\rightarrow \infty$



So in case of a "blatantly wrong prediction", cost is very very high!

Logistic Regression Cost Function

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



if $h_{\theta}(x) = 0$ and $y=0$, cost = $-\log(1) = 0$

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

So in case of a "blatantly wrong prediction", cost is very very high!

Gradient Descent for Logistic Regression

Combined Cost Function

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

Note: $y = 0$ or 1 always

$$J(\theta) = \frac{1}{m} \left[\sum_{i=1}^m \text{Cost}(h_{\theta}(x^{(i)}), y^{(i)}) \right]$$

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

This is because if $y^{(i)} = 0$, first term is 0 and only the 2nd term remains. Similarly, if $y^{(i)} = 1$, second term is 0 and only the first one remains

Summary of the Problem

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

Find parameters $\theta : \min_{\theta} J(\theta)$

To make a prediction given new x

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

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 all θ_j)

Computing Derivative Term

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

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

Repeat {

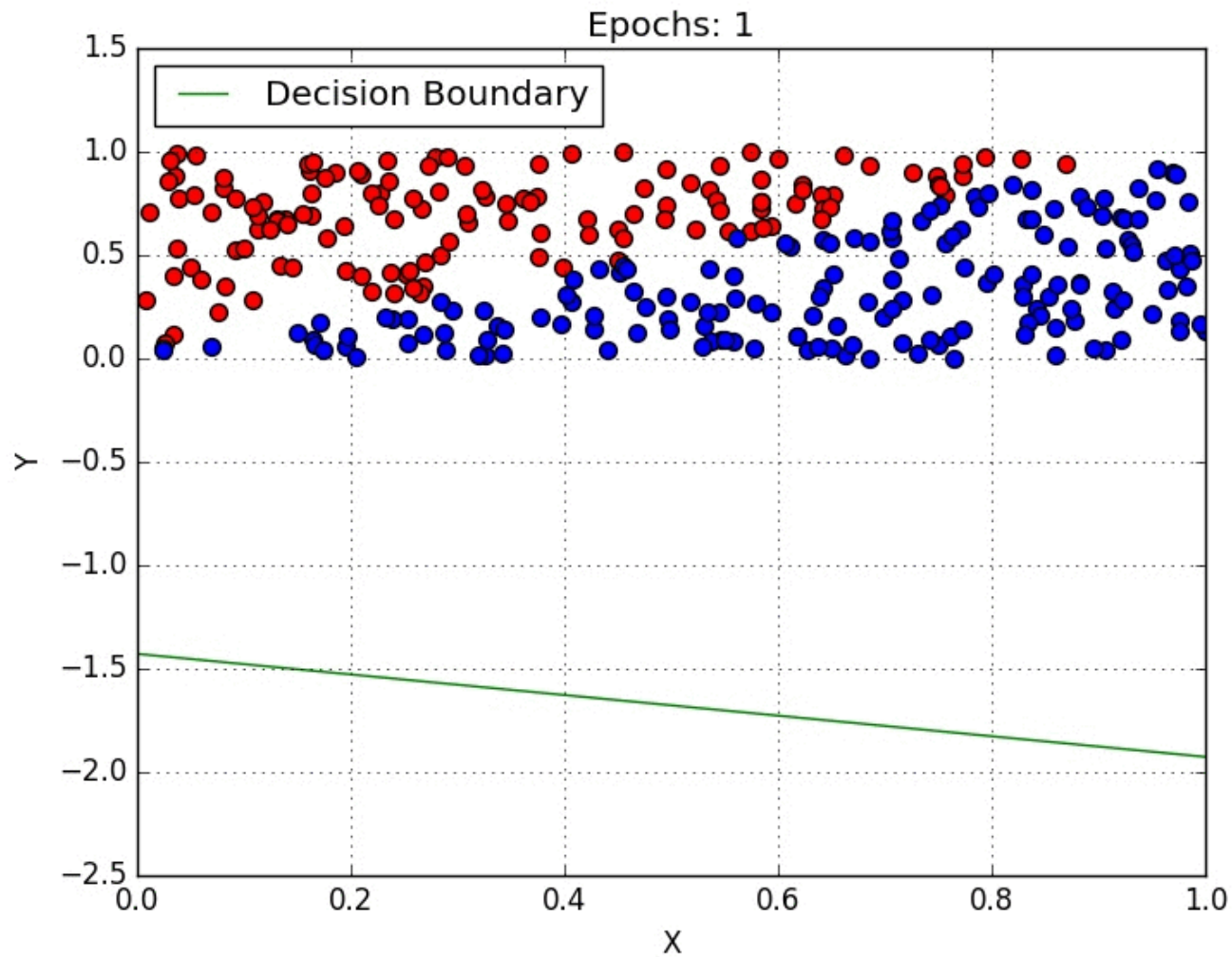
$$\theta_j := \theta_j - \alpha \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

}

(simultaneously update all θ_j)

Algorithm looks identical to linear regression! But the hypothesis h is different.

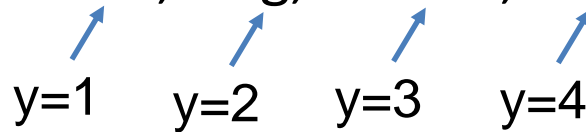
Sample Execution



Logistic Regression: Going Beyond 0/1

Multiclass Classification

Image recognition: Cat, Dog, Mouse, Fox


y=1 y=2 y=3 y=4

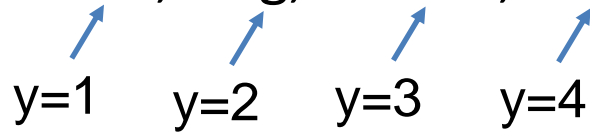
Email foldering/tagging: Family, Friends, Work, Hobby

Weather: Sunny, Cloudy, Rain, Snow

Question: logistic regression only gives us a prediction for $y=0$ or $y=1$. What about other values of y ?

The Idea

Image recognition: Cat, Dog, Mouse, Fox


y=1 y=2 y=3 y=4

Break the prediction of y into a series of smaller question:

Q1: Is the given image a cat?

Q2: Is the given image a dog?

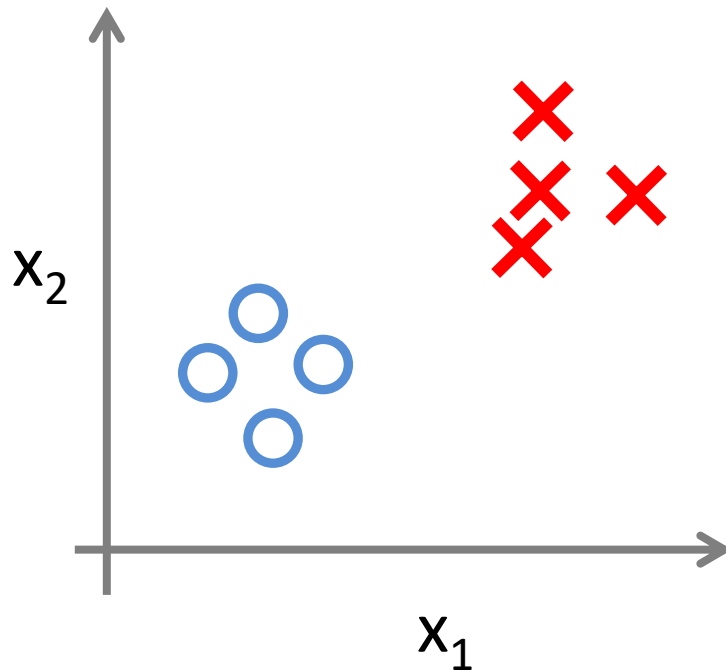
Q3: Is the given image a mouse?

Q4: Is the given image a fox?

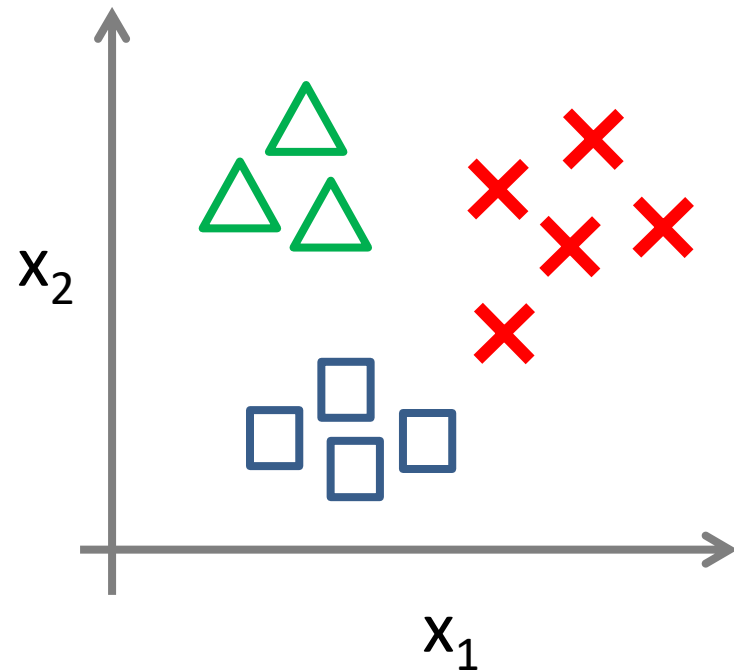
Each of these is a yes/no (or 0/1) question. Given all these answers, output a prediction for y from 1 to 4.

Pictorial Representation

Binary classification:

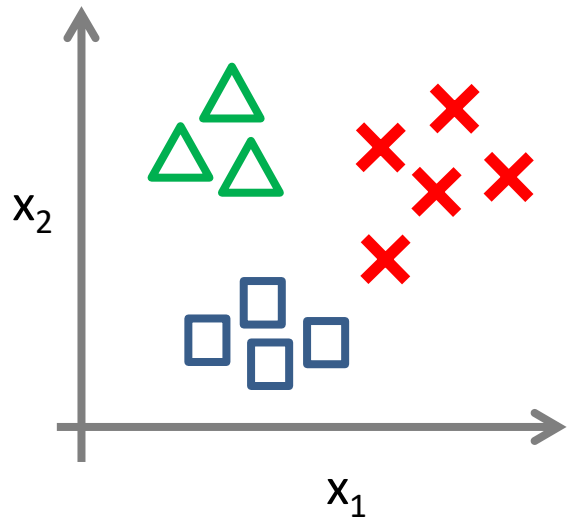





Multi-class classification:

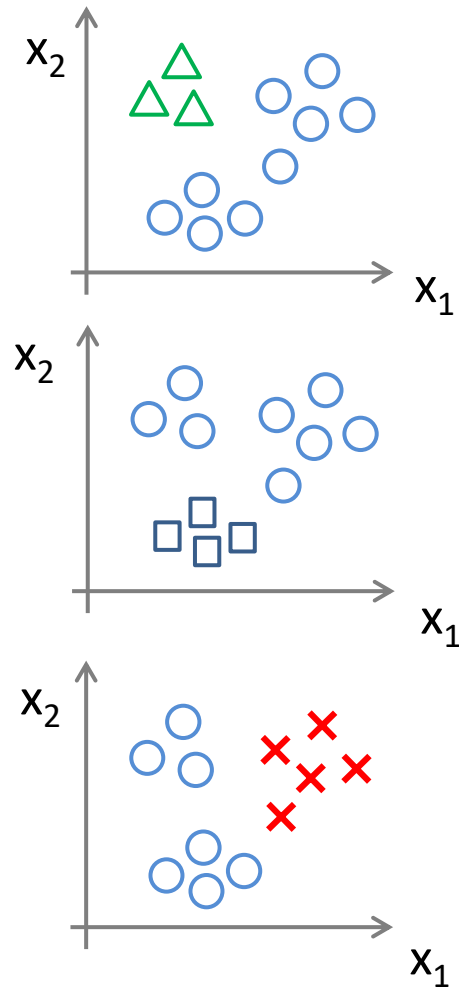


x_1 and x_2 are features, value of y given by the shape (circle, cross,...)

Pictorial Representation



Class 1: 
Class 2: 
Class 3: 



Algorithm for Multiclass Classification

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 i where the probability $y = i$ is highest

Questions?