REGRESSION

INTRODUCTION TO DATA SCIENCE

TIM KRASKA

teaching
datascience
.org

ANNOUNCEMENTS

- I forgot the chocolate at home ☹
- AWS Sign-Up
- User Study
- Viz Plagiarism (Marey’s Trains)
- Project mid-term grading and final project/report

CLICKER: WHAT IS GRADIENT DESCENT

A) A Regression Technique

B) An Optimization Technique

C) A Classification Technique

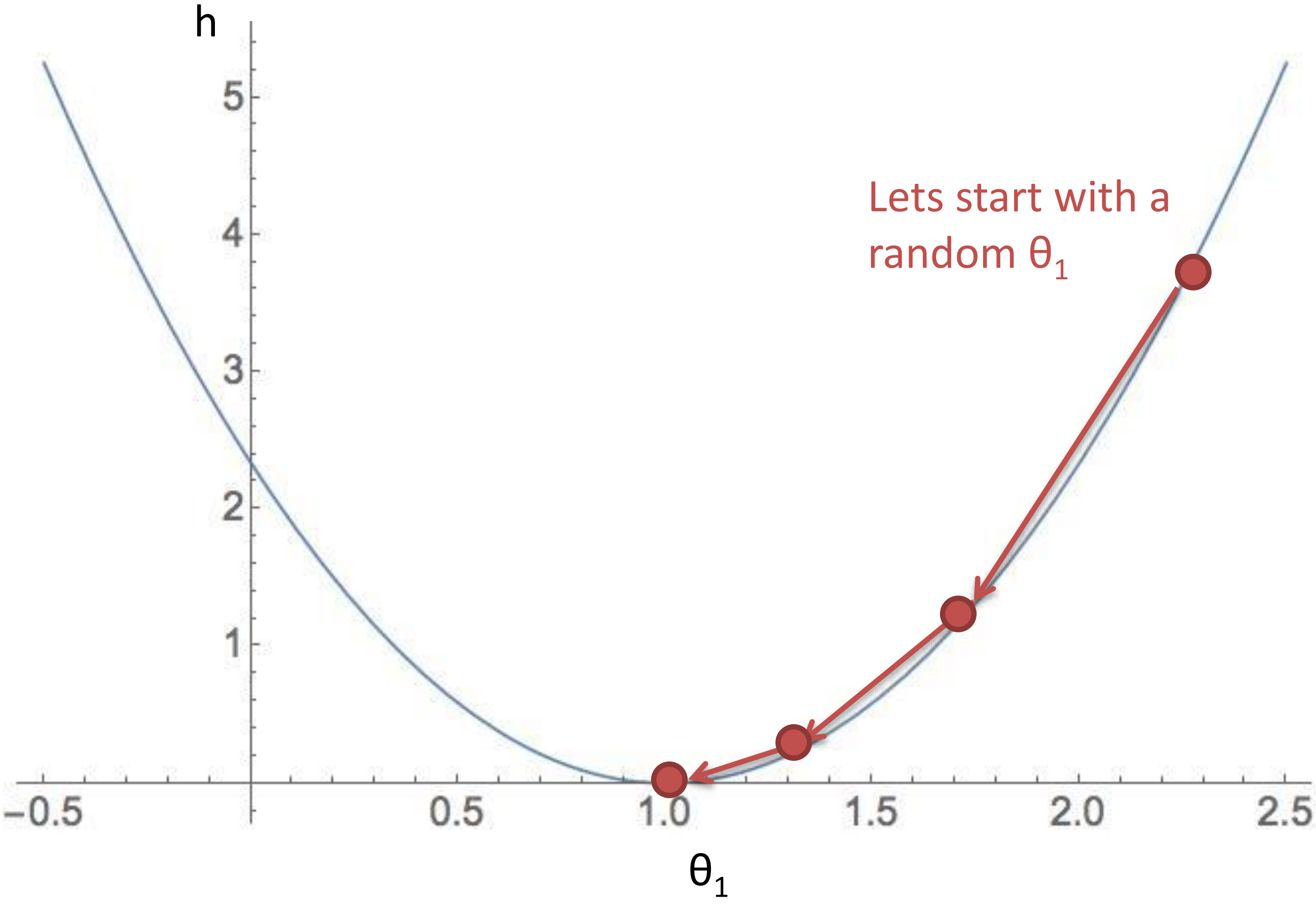
WHAT IS THE COST FUNCTION OF LINEAR REGRESSION

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

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

C) $$\theta_j := \theta_j + \alpha \left(y^{(i)} - h_{\theta}(x^{(i)})\right) x_j^{(i)}$$

How Can We Find The Minimum?

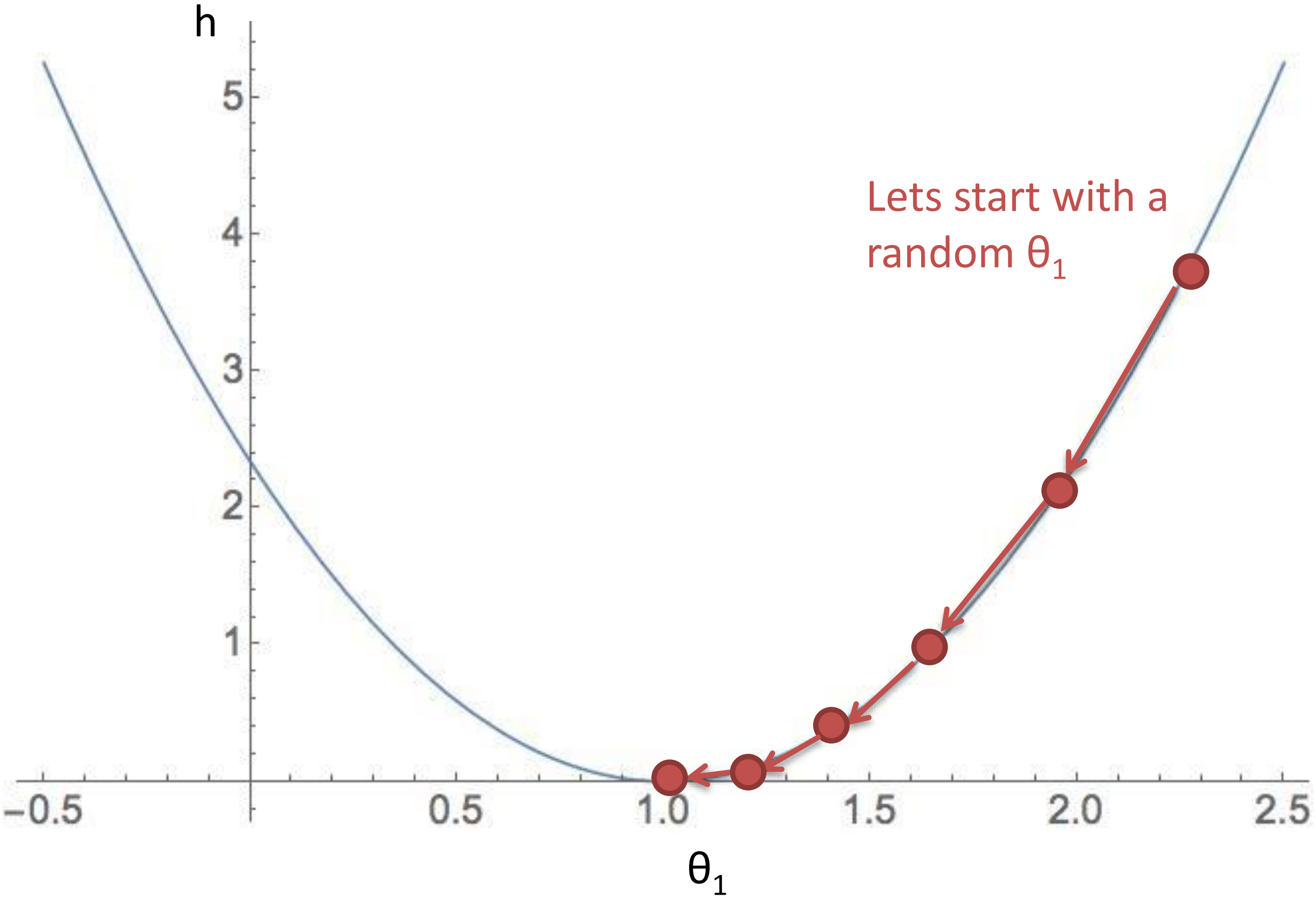


Lets start with a random θ_1

$$q_1 = q_1 - a \frac{\P}{\P q_1} J(q_1)$$

Step Size / Learning Rate

How Can We Find The Minimum?



Lets start with a random θ_1

Steps are automatically smaller the closer they get to the minimum

$$q_1 = q_1 - a \frac{¶}{¶q_1} J(q_1)$$

Step Size / Learning Rate

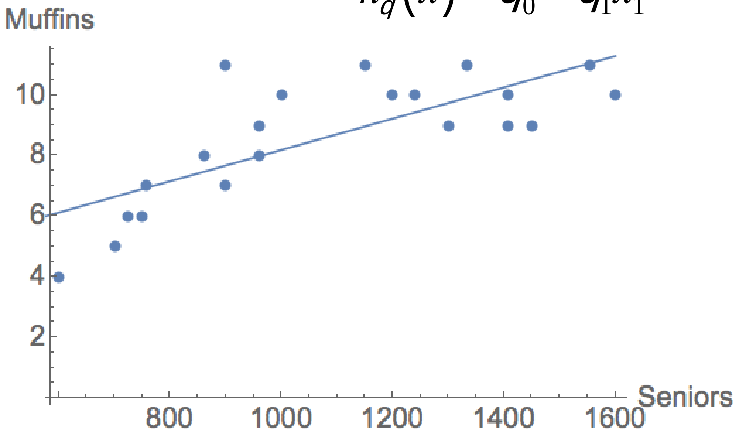
Clicker: Is Mini-Batch Guaranteed to Converge

A) Yes

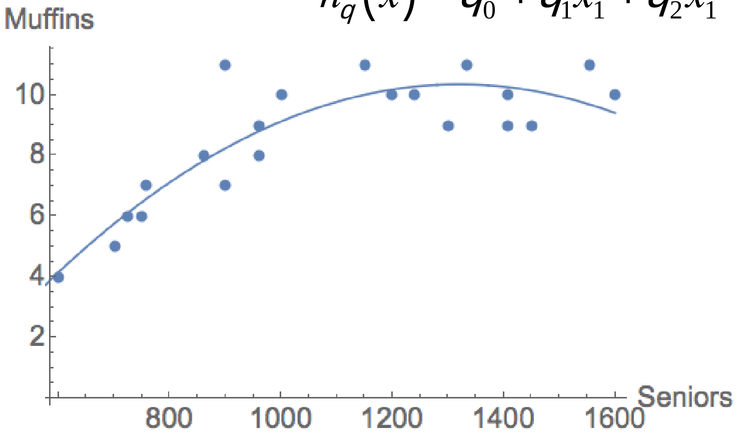
B) No

Polynomial Regression

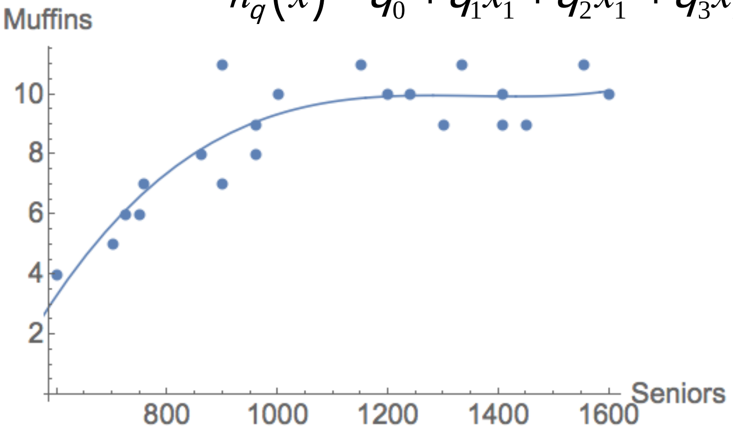
$h_q(x) = q_0 + q_1 x_1$



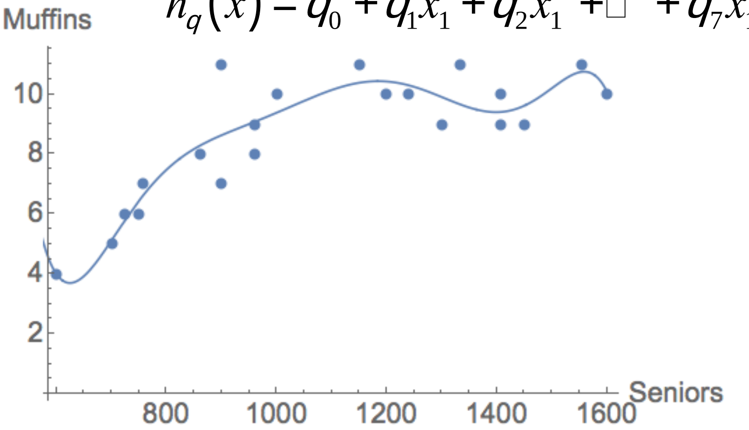
$h_q(x) = q_0 + q_1 x_1 + q_2 x_1^2$



$h_q(x) = q_0 + q_1 x_1 + q_2 x_1^2 + q_3 x_1^3$



$h_q(x) = q_0 + q_1 x_1 + q_2 x_1^2 + \square + q_7 x_1^7$

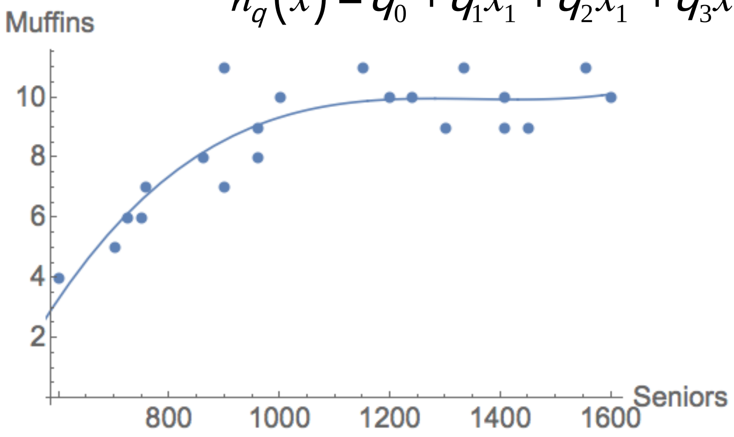


How To Prevent Overfitting

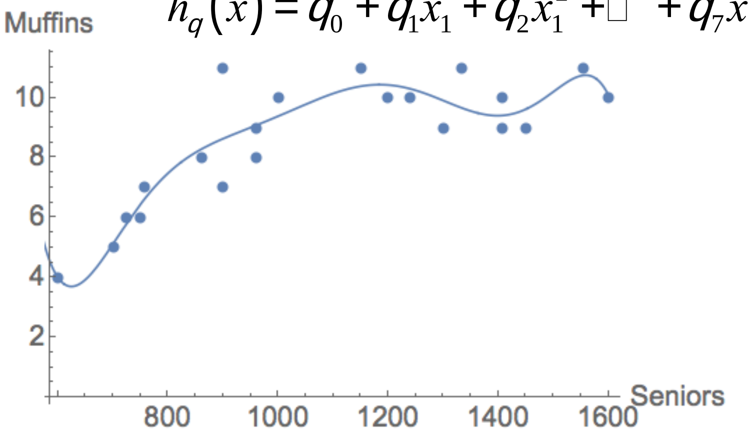
- Adjust features
  - Reduce number of polynomials
  - Reduce number of features
- Regularization
  - Keep all the features, but reduce their impact
  - Works well when we have a lot of features, each of which contributes a little

Regularization: Intuition

$$h_q(x) = q_0 + q_1 x_1 + q_2 x_1^2 + q_3 x_1^3$$



$$h_q(x) = q_0 + q_1 x_1 + q_2 x_1^2 + \square \ + q_7 x_1^7$$



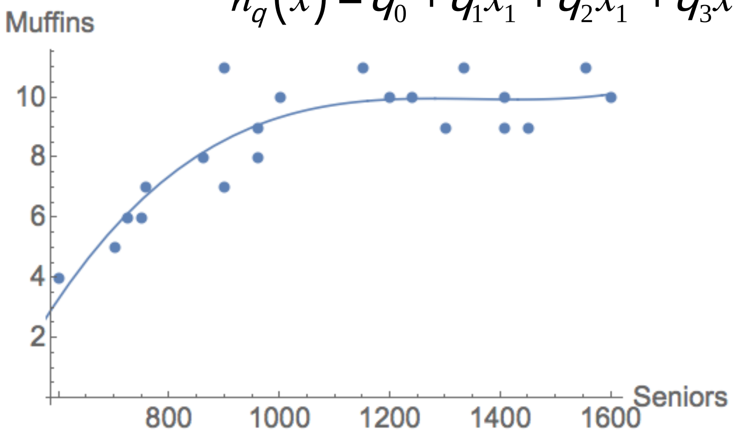
$$\min_q J(q) = \frac{1}{2m} \overset{m}{\underset{i=1}{\mathring{a}}} \left(h_q(x^{(i)}) - y^{(i)} \right)^2$$

$$+1000 q_4 + 1000 q_5 + 1000 q_6 + 1000 q_7$$

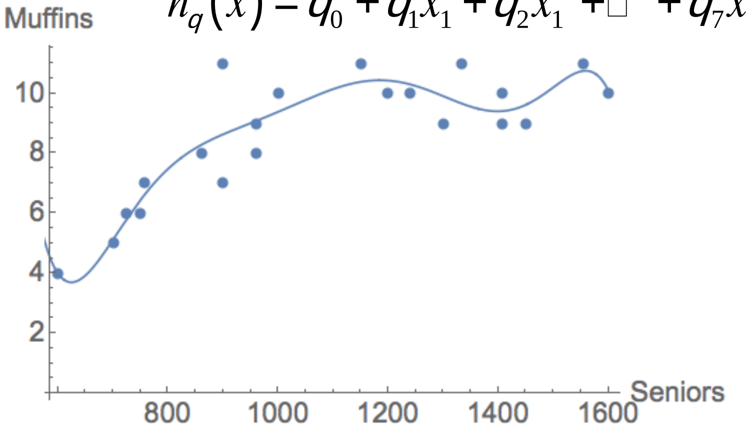
$$q_4, q_5, q_6, q_7 \to 0$$

Regularization: Intuition

$$h_q(x) = q_0 + q_1x_1 + q_2x_1^2 + q_3x_1^3$$



$$h_q(x) = q_0 + q_1x_1 + q_2x_1^2 + \square \ + q_7x_1^7$$



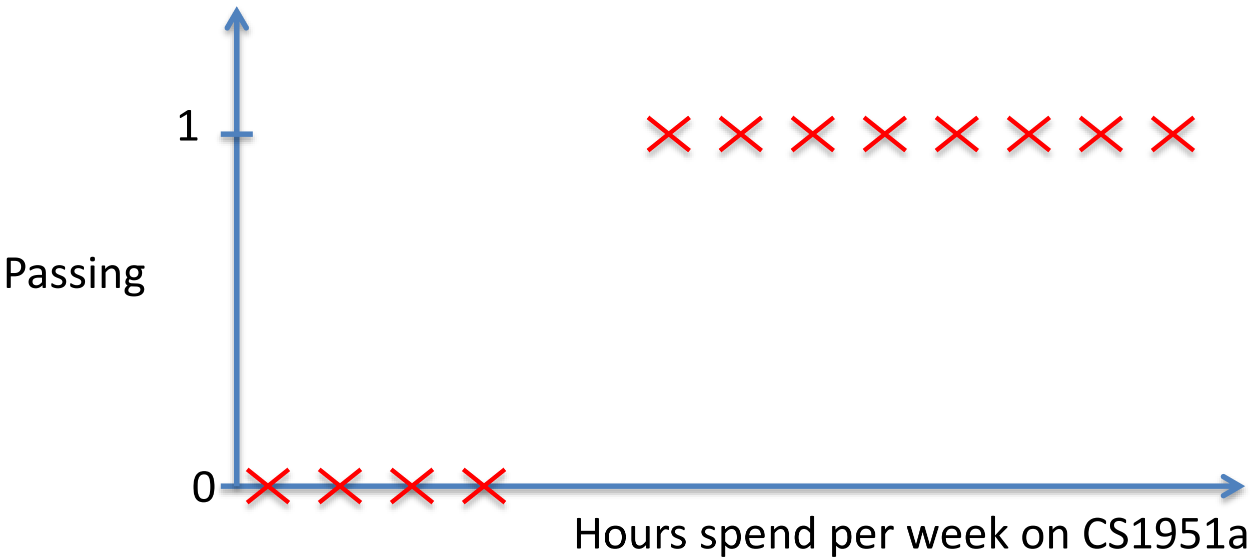
$$\min_q J(q) = \frac{1}{2m} \underset{i=1}{\overset{m}{\text{å}}} \left(h_q(x^{(i)}) - y^{(i)}\right)^2 + / \underbrace{\underset{j=1}{\overset{n}{\text{å}}} q_j^2}_{\text{Regularization term}}$$

Works ok for reducing the impact of polynomials (better to reduce nb of polynomials directly)
Works great for a bag of equally important features.

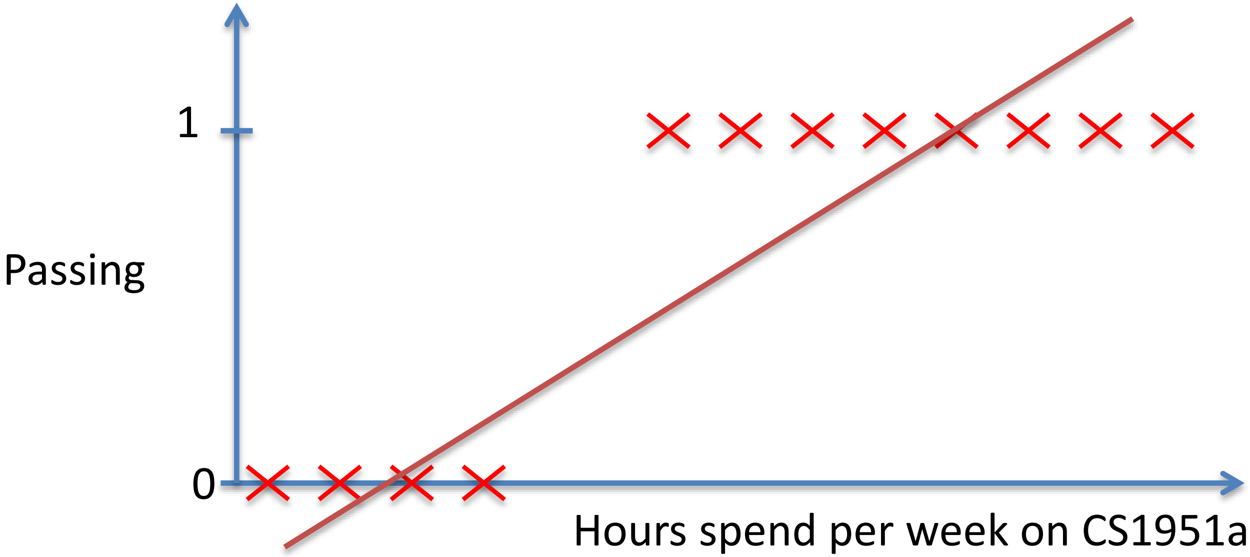
Logistic Regression

Adapted from notes and slides by Andrew Ng, Kurt Miller and Romain Thibaux

Can We do Classification with Linear Regression?



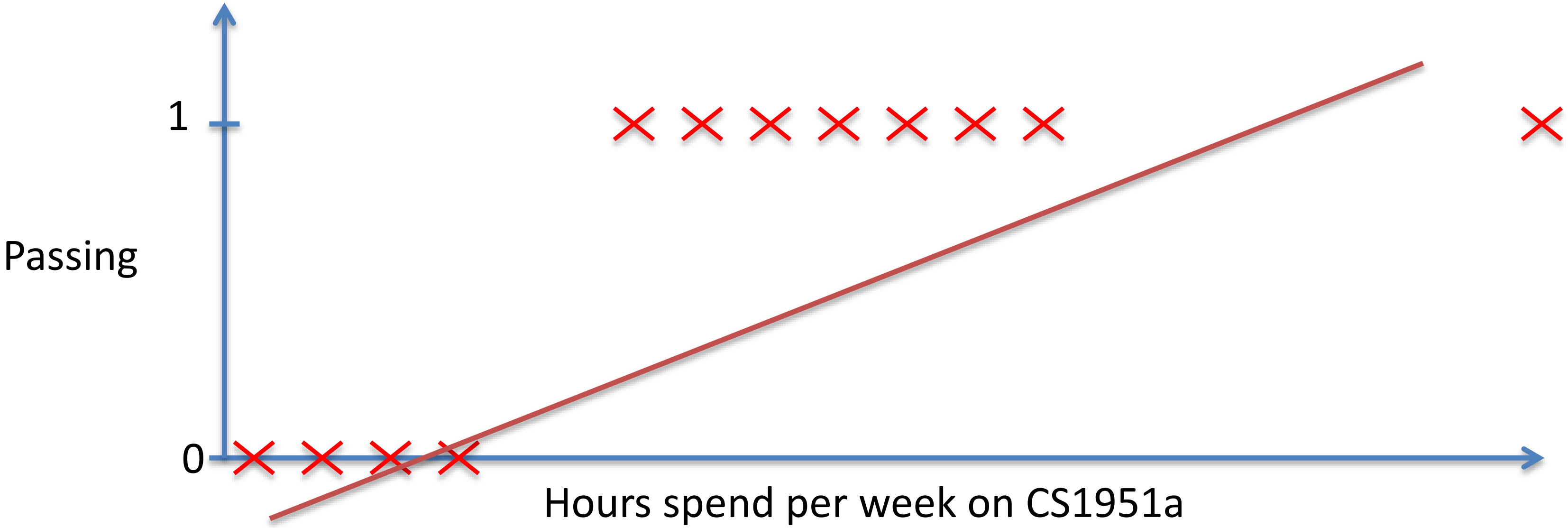
Can We do Classification with Linear Regression?



$$\hat{y} < 0.5 \rightarrow 0$$

$$\hat{y} \geq 0.5 \rightarrow 1$$

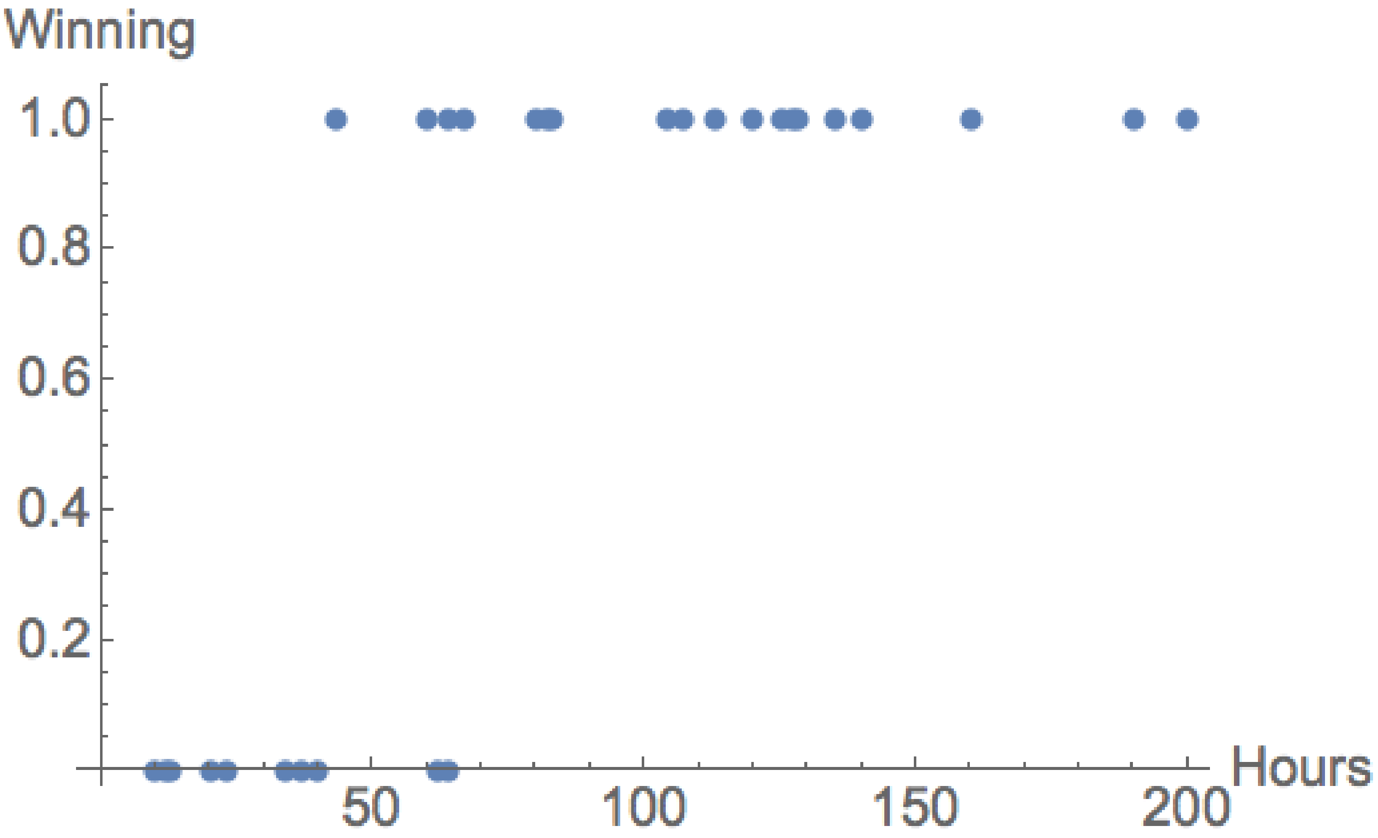
Can We do Classification with Linear Regression?



$$\hat{y} < 0.5 \rightarrow 0$$

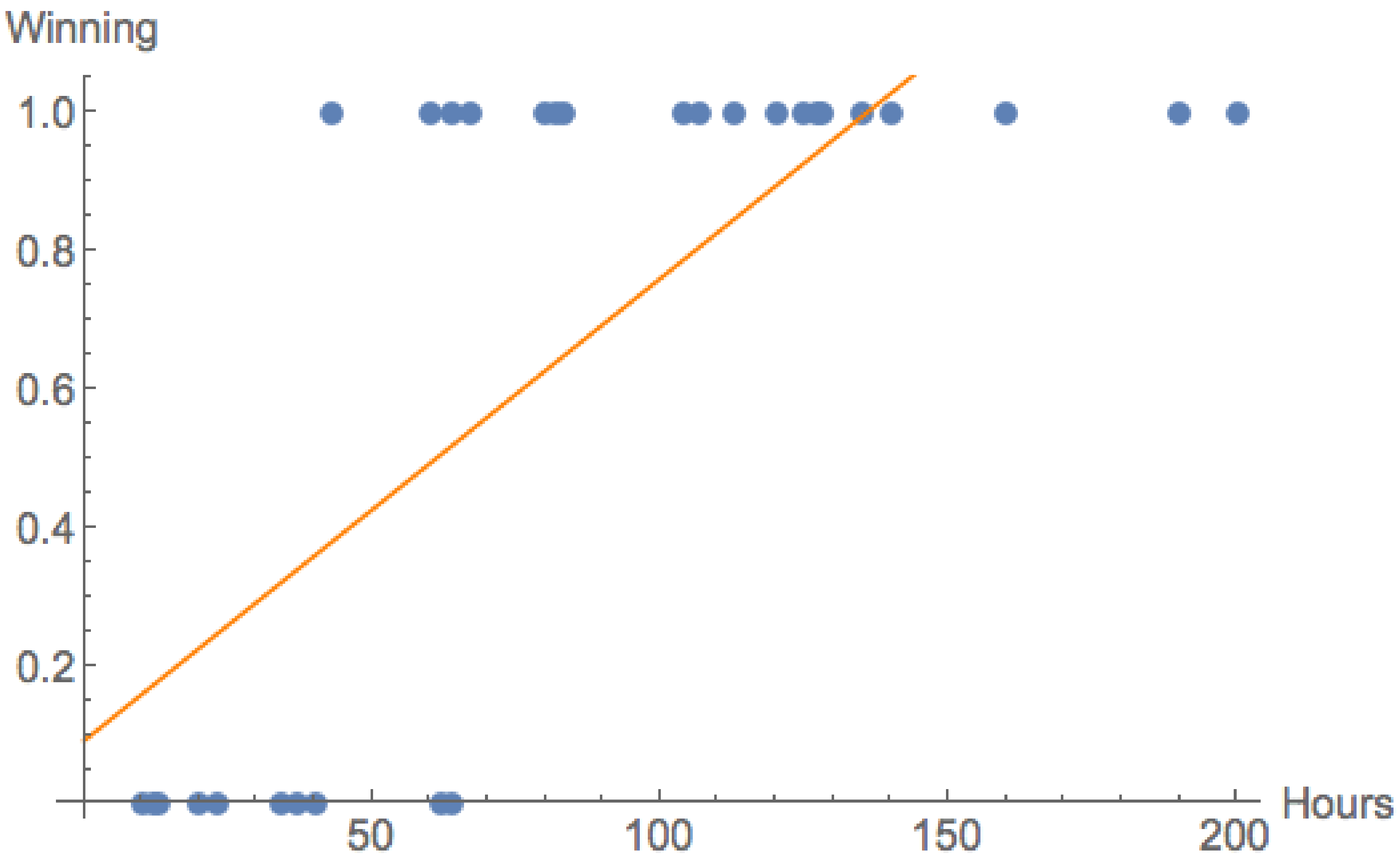
$$\hat{y} \geq 0.5 \rightarrow 1$$

On the Request of my PhD Students I changed my Logistic Regression example



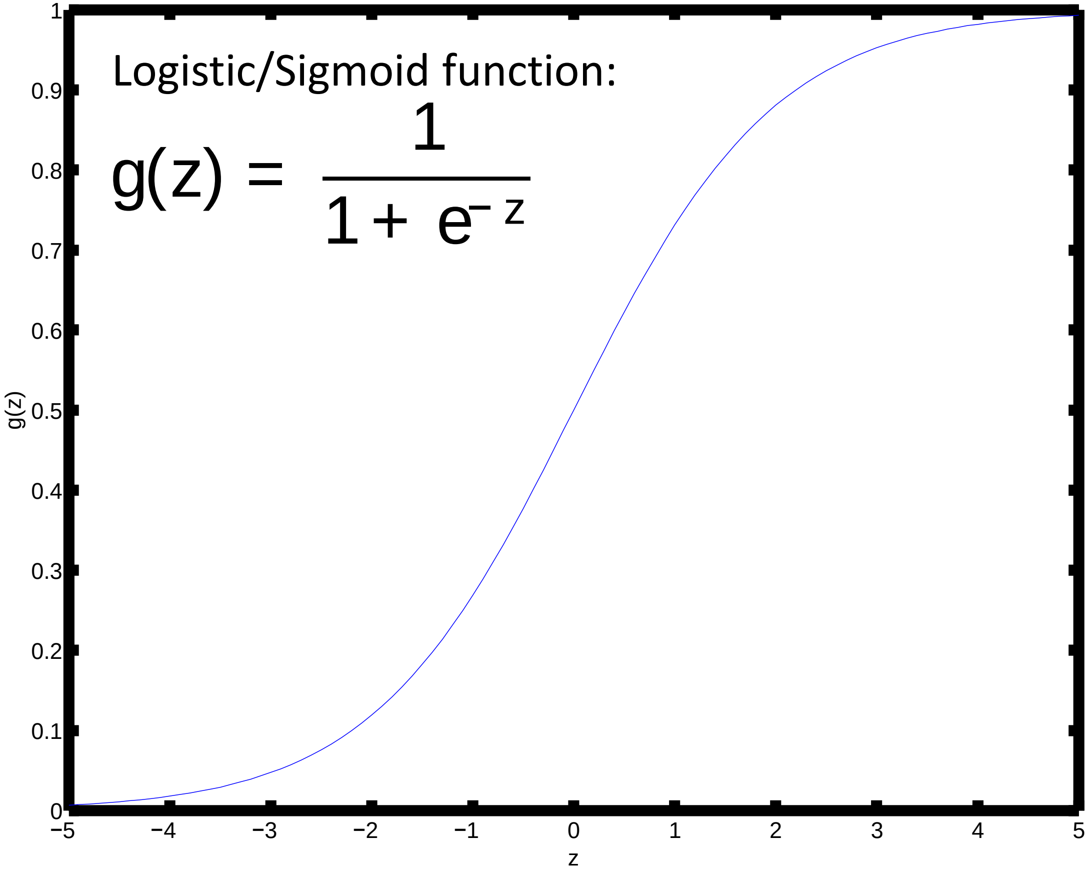
Can We do Classification with Linear Regression?

$$h_q(x) = q^T x = q_0 + q_1 hours = .09 + .007x$$



How do we adjust $h_{\boldsymbol{\theta}}(x)$ for $y \in \{0,1\}$?

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



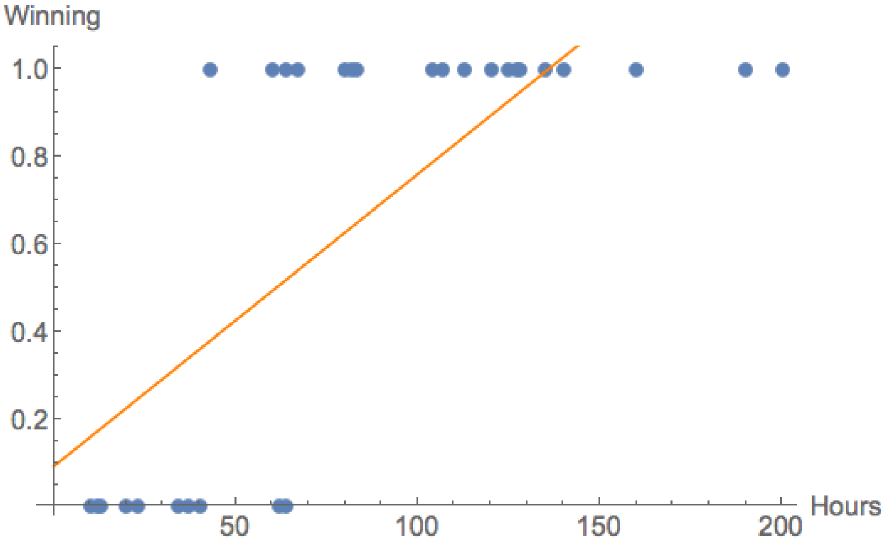
Logistic/Sigmoid function:

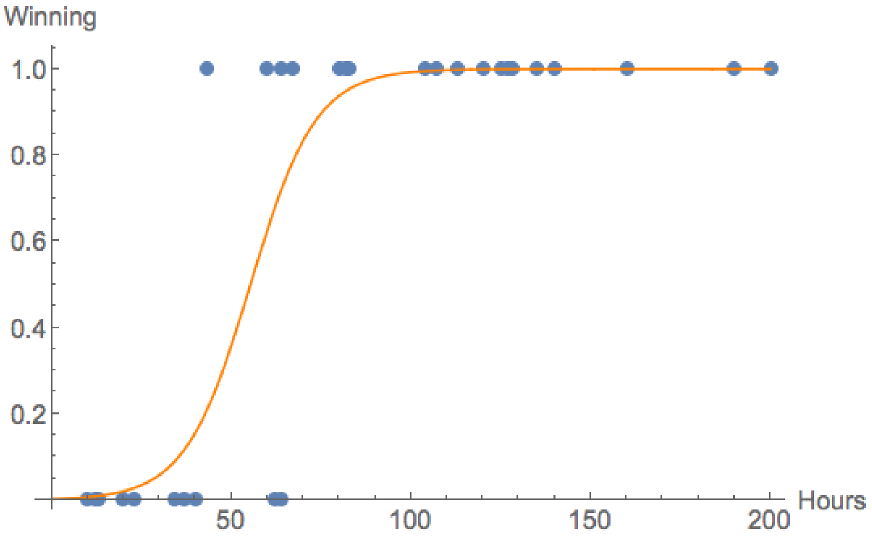
$$g(z) = \frac{1}{1 + e^{-z}}$$

Logistic Regression

$$h_q(x) = q^T x$$
$$= q_0 + q_1 hours$$
$$= .09 + .007x$$

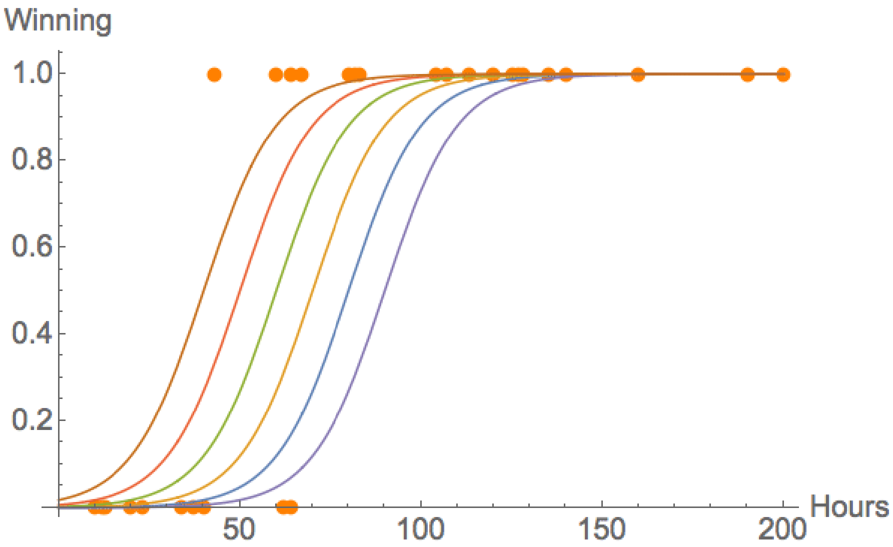
$$h_q(x) = g(q^T x) = g(q_0 + q_1 hours)$$
$$= \frac{1}{1+e^{-q^T x}} = \frac{1}{1+e^{-q_0 - q_1 hours}}$$
$$= \frac{1}{1+e^{-6-0.11 hours}}$$



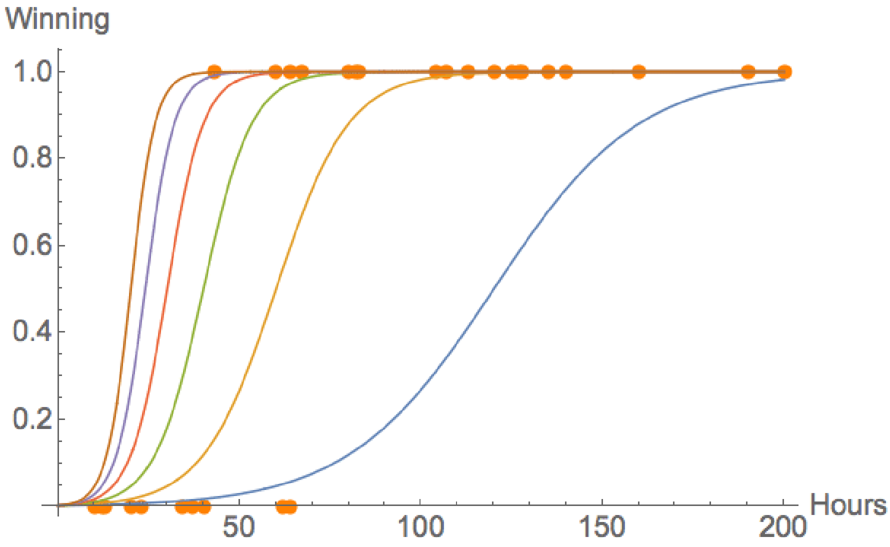


Fitting the SIGMOID Function

$$h_q(x) = g(q^T x) = g(q_0 + q_1 hours) = \frac{1}{1 + e^{-q_0 - q_1 hours}}$$



$\Theta_0 = \{4,5,6,7,8,9\}$
$\Theta_1 = 0.1$



$\Theta_0 = 6$
$\Theta_1 = \{0.3,0.25,0.2,0.15,0.1,0.5\}$

Cost Function

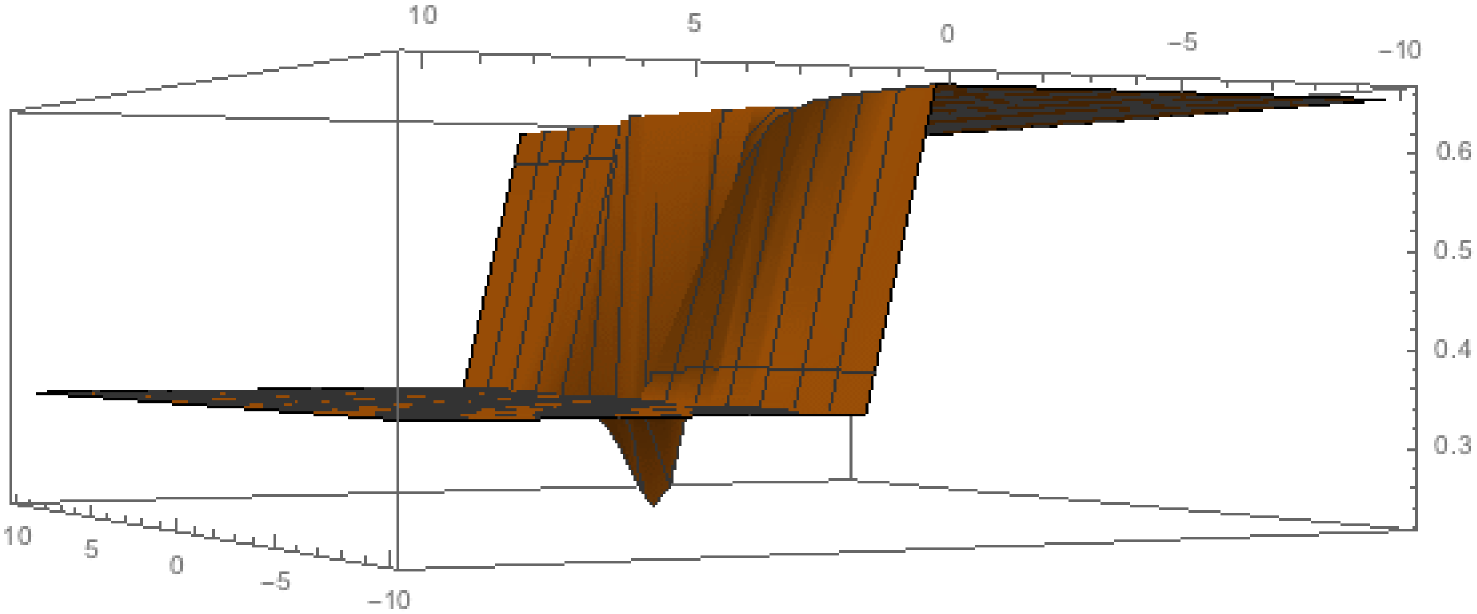
$$h_q = g\left(q^T x\right) = \frac{1}{1 + e^{- q^T x}}$$

From

Linear Regression: $\min\limits_{q} J(q) = \frac{1}{m} \overset{m}{\underset{i=1}{\text{å}}} \text{cost}\left(h_q(x^{(i)}), y^{(i)}\right)$

$$\text{cost}\left(h_q(x), y\right) = \underbrace{\frac{1}{2}\left(y - h_q(x)\right)^2}_{\text{No longer convex}}$$

$$J(q) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left(y - \frac{1}{1+e^{-q_0 - q_1 x}}\right)^2$$



Cost Function

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

From

Linear Regression: $\min_\theta J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{cost}\left(h_\theta(x^{(i)}), y^{(i)}\right)^2$

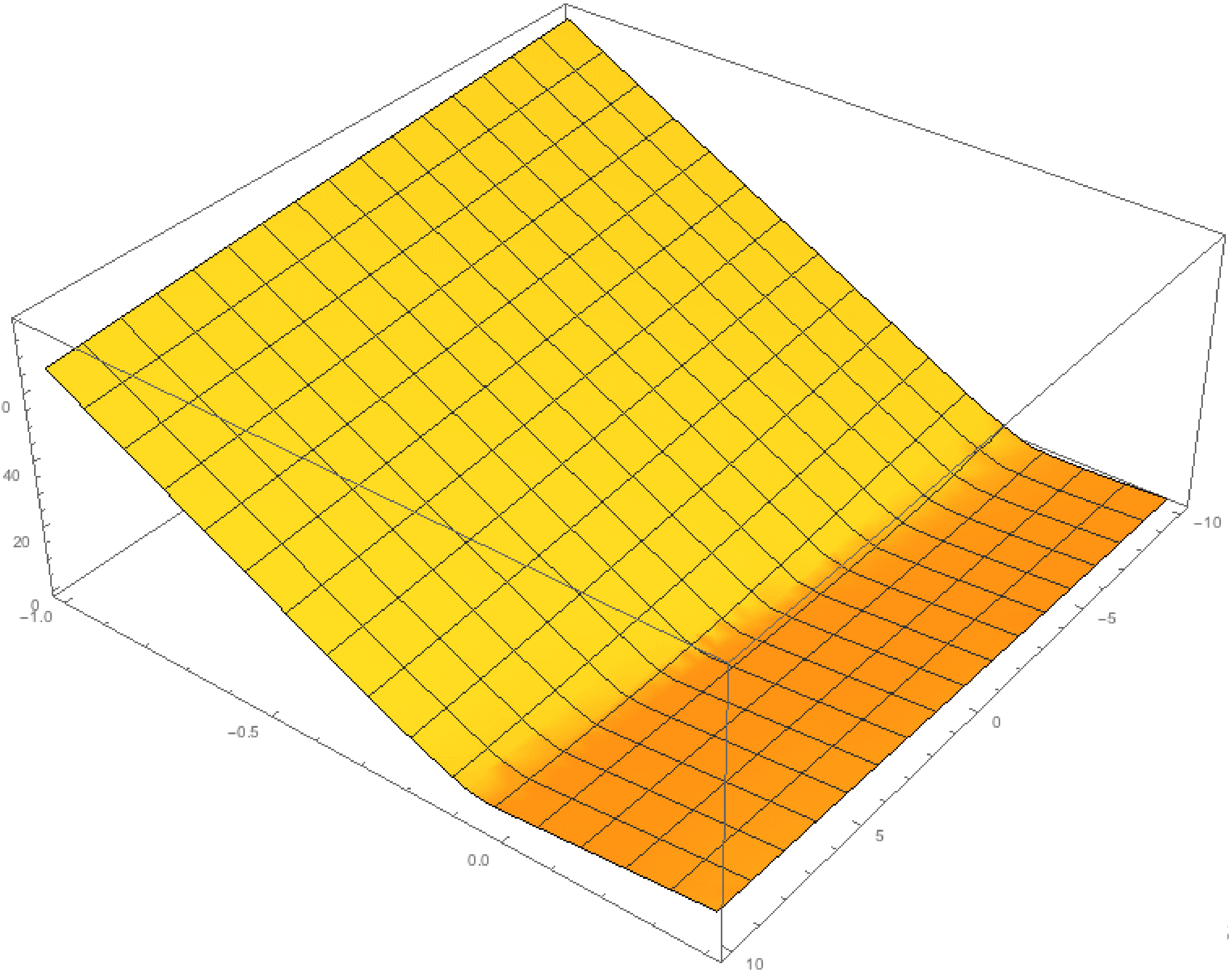
$$\mathrm{cost}\left(h_\theta(x), y\right) = \underbrace{\frac{1}{2}\left(h_\theta(x) - y\right)^2}_{\text{No longer convex}}$$

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

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

Why this cost function? → Can be shown it is equivalent to MLE estimations

$$\min_{q} J(q) = \frac{1}{m}\sum_{i=1}^{m} - y\log\left(\frac{1}{1+e^{-q_0-q_1x}}\right) - (1- y)\log\left(1- \frac{1}{1+e^{-q_0-q_1x}}\right)$$



Interpretation of Hypothesis

$$h_q = \text{estimated probability that } y = 1 \text{ on input } x$$

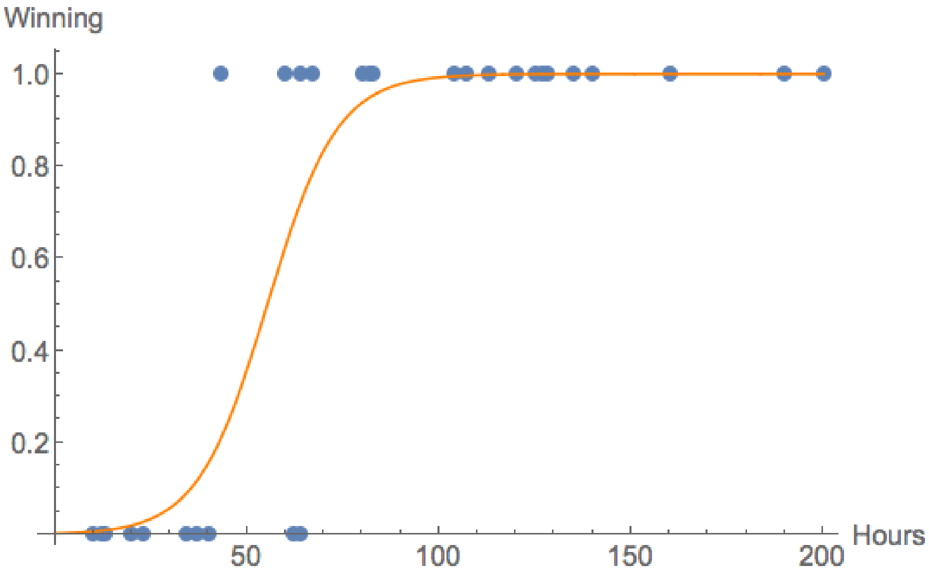
Example:

Erfan trained for 50h

$$x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 50h \end{bmatrix}$$

$$h_q(x) = \frac{1}{1 + e^{-q_0 x_0 - q_1 * x_1}}$$

$$= \frac{1}{1 + e^{-6 - 0.11*50}} = 0.378$$

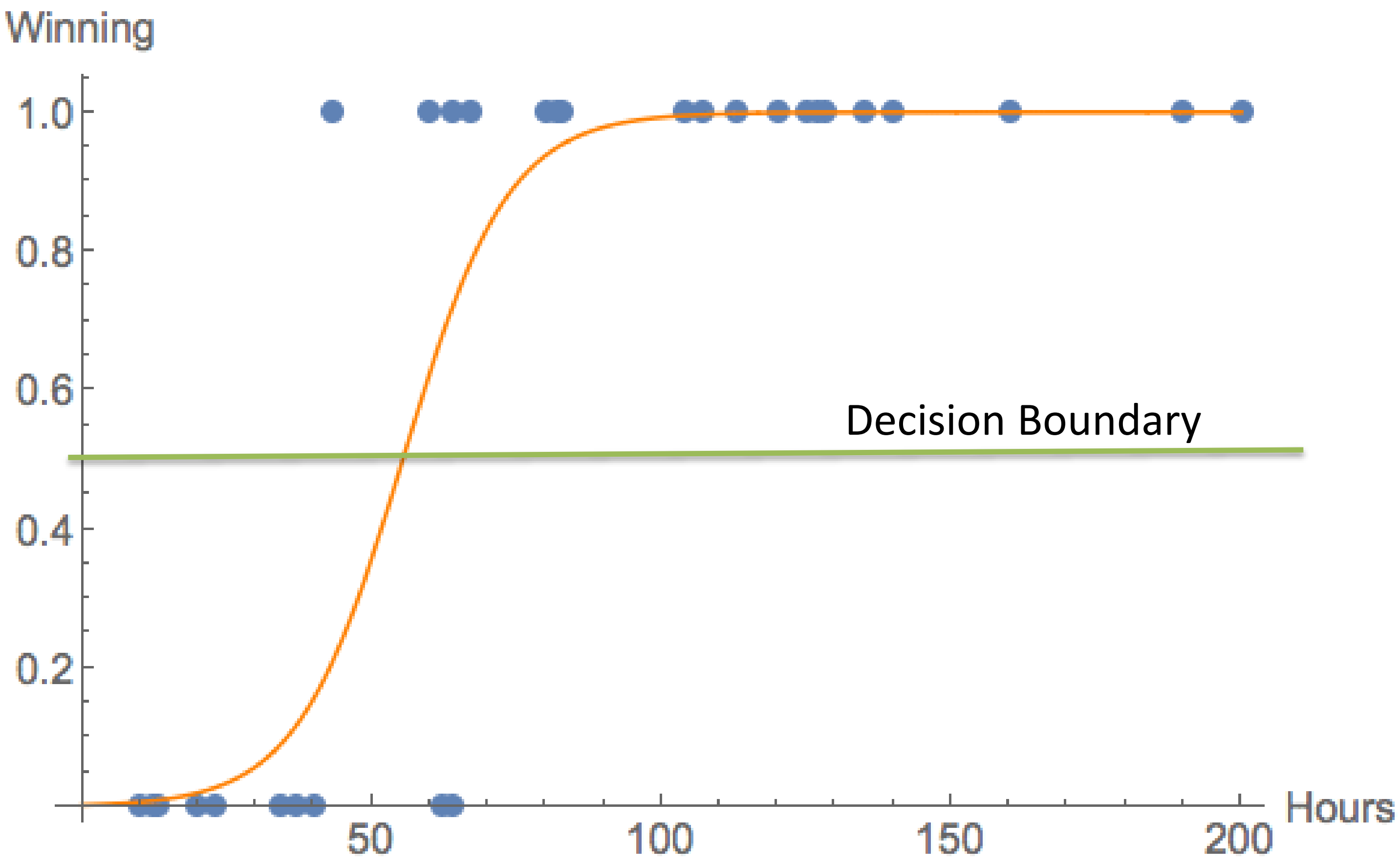


After 50 hours of training,
the PhD Student has a 37.8% chance of winning

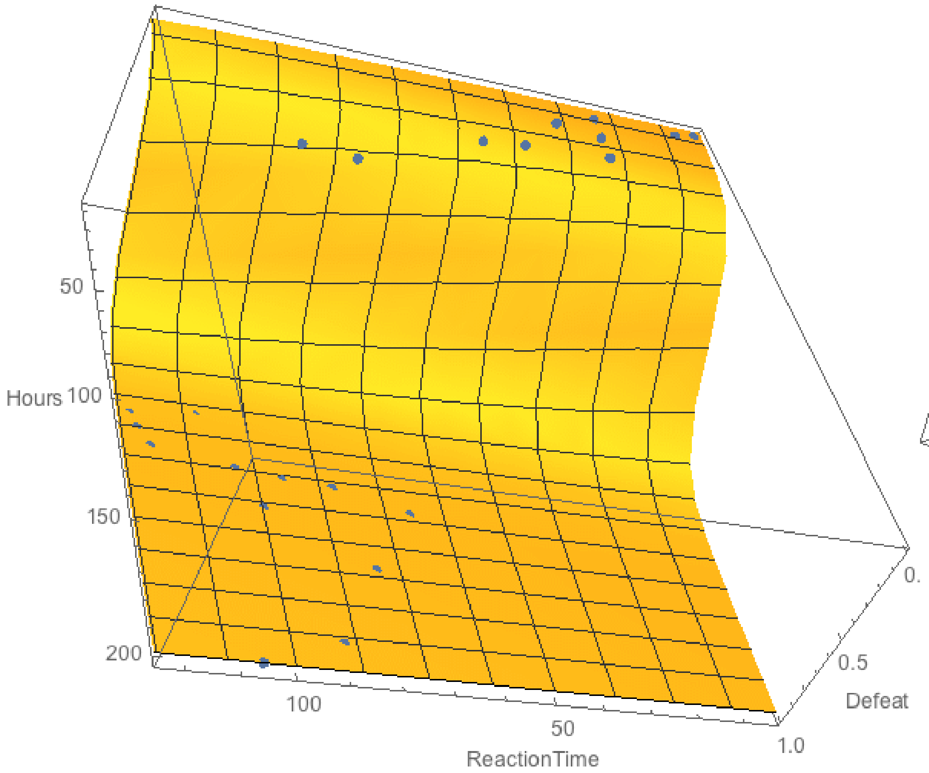
More formally:

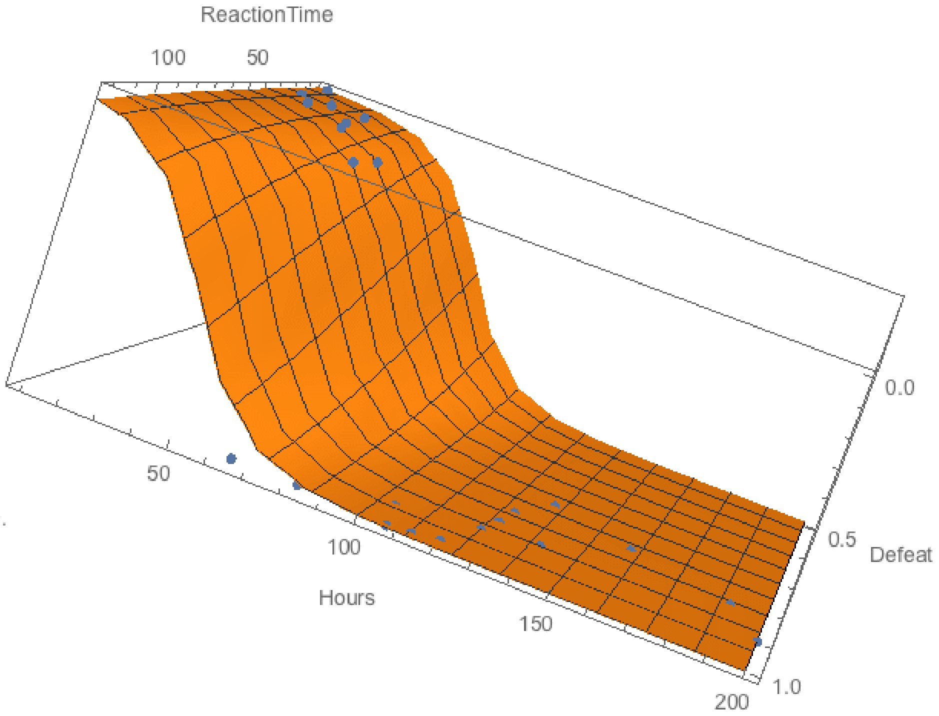
$$h_q(x) = p\left(y = 1 \middle| x, q\right)$$

When Should We Classify a New Data Point as 1 or 0?



Fitting in Higher Dimensions

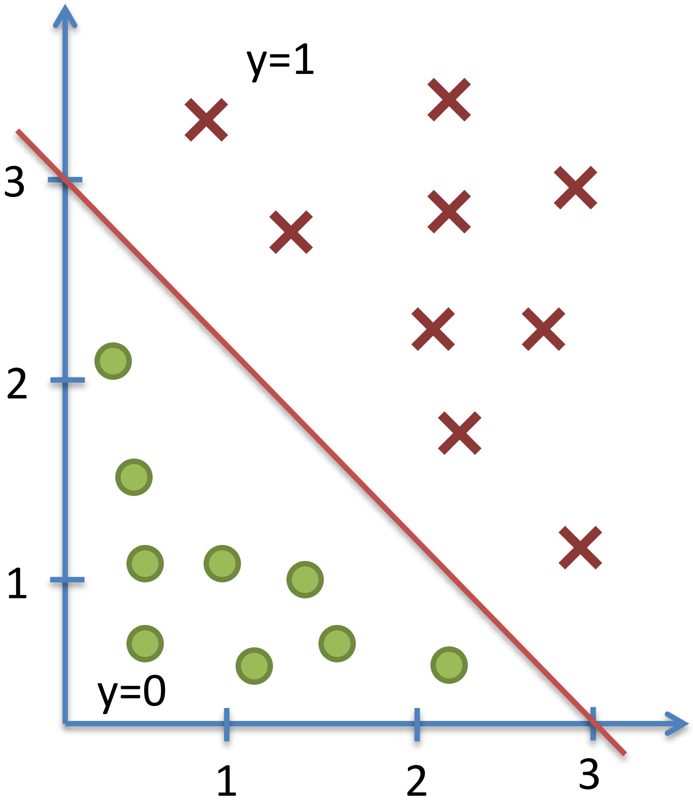




Decision Boundary with 2 Features

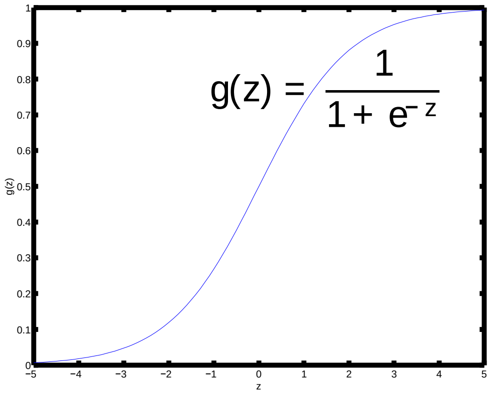
$$h_q(x) = g\left(q_0 + q_1 x + q_2 x\right) = \frac{1}{1 + e^{-\,(q_0 + q_1 x_1 + q_2 x_2)}}$$

$$= \frac{1}{1 + e^{-\,(-3 + 1x_1 + 1x_2)}}$$



When is h above 0.5

$$h_\theta(x) \geq 0.5 \;\rightarrow (-3 + \; x_1 + x_2) \geq 0$$
$$h_\theta(x) < 0.5 \;\rightarrow (-3 + \; x_1 + x_2) < 0$$



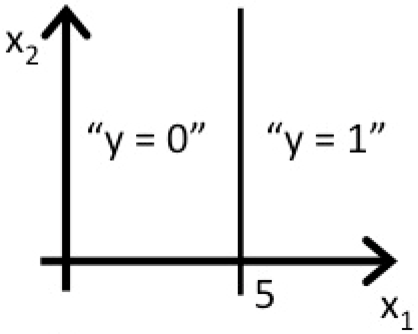
z	e^{-z}	$\frac{1}{1+e^{-z}}$
-1.00	2.72	0.27
0.00	1.00	0.50
1.00	0.37	0.73

Clicker

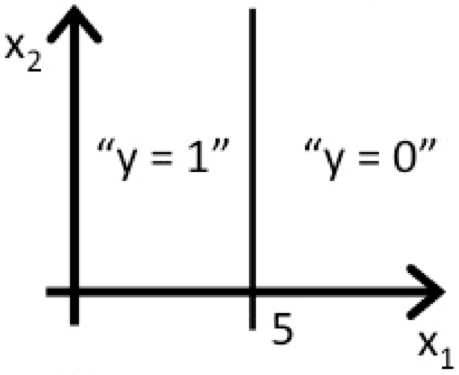
Assume you have two features x_1 and x_2, and $q_0 = 5$, $q_1 = - 1$, $q_2 = 0$.

So $h_q(x) = g(5 - x_1)$. Which of these shows the decision boundary of $h_0(x)$?

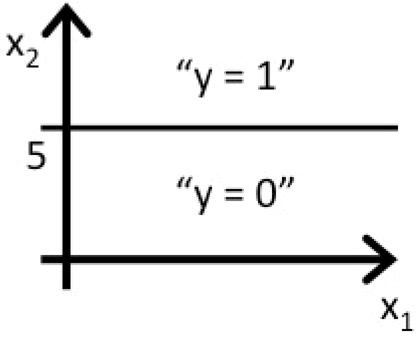
(a)



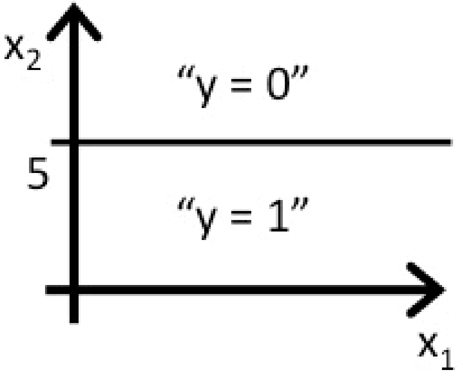
(b)



(c)

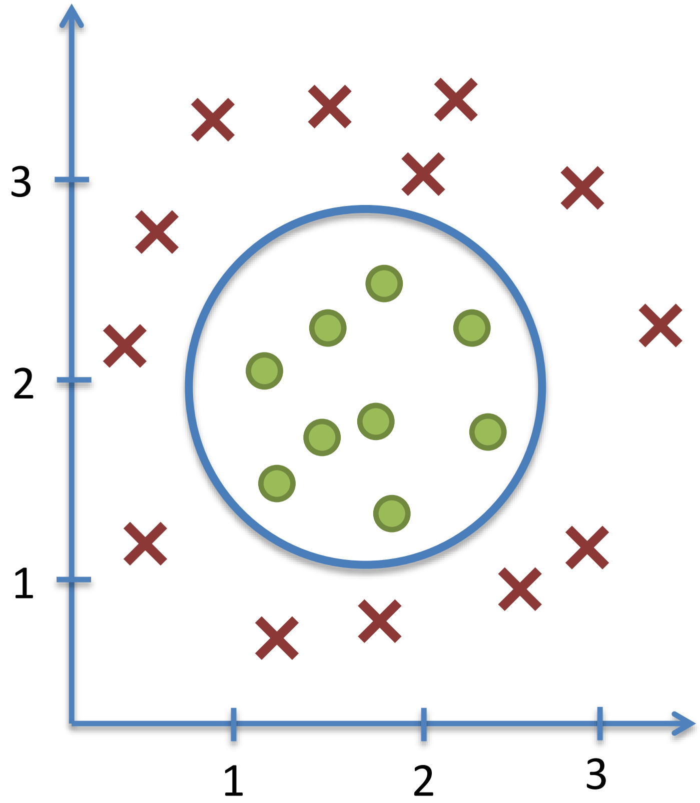


(d)

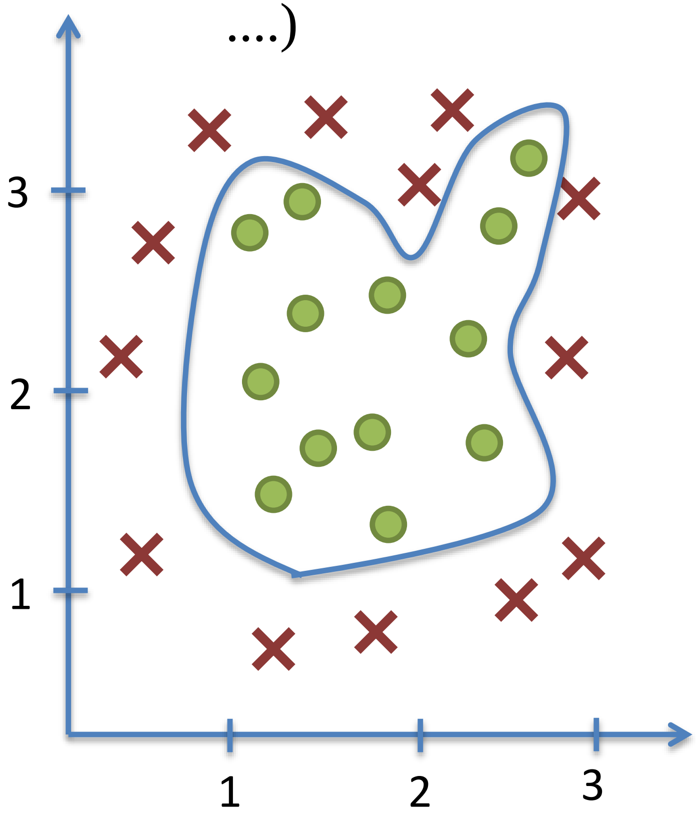


And what now?

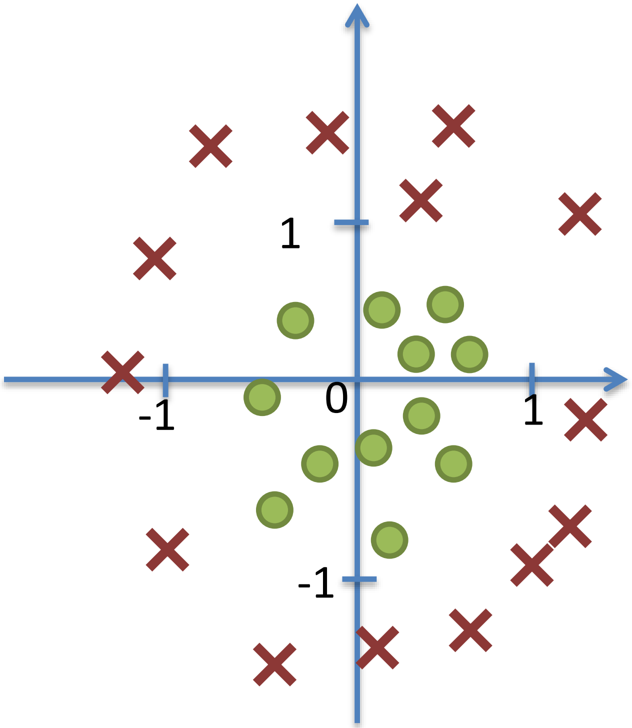
$$h_q(x) = g(q_0 + q_1 x_1 + q_2 x_2 + q_3 x_1^2 + q_4 x_2^2)$$



$$h_q(x) = g(q_0 + q_1 x_1 + q_2 x_2 + q_3 x_1^2 + q_4 x_2^2 + q_5 x_1^3 + q_6 x_2^3 + q_7 x_1^1 x_2^1 + q_8 x_1^2 x_2^2 \ldots.)$$



Clicker Question



Our hypothesis;

$$h_q(x) = g(q_0 + q_1 x_1 + q_2 x_2 + q_3 x_1^2 + q_4 x_2^2)$$

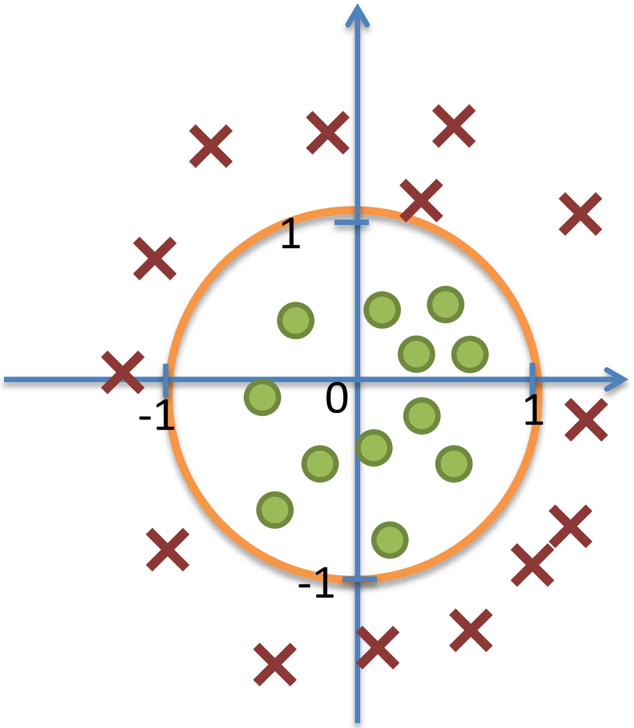
Which θ would predict all red (y=1) and Green dots (y=0) correctly

a) $q = \begin{pmatrix} -1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$

b) $q = \begin{pmatrix} -1 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}$

c) $q = \begin{pmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$

Clicker Question



Our hypothesis;

$$h_q(x) = g(q_0 + q_1 x_1 + q_2 x_2 + q_3 x_1^2 + q_4 x_2^2)$$

Which θ would predict all red (y=1) and Green dots (y=0) correctly

$$\begin{pmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

Predict “y=1” if

$$- 1 + x_1^2 + x_2^2 > 0$$

$$x_1^2 + x_2^2 > 1$$

Stochastic Gradient Descent

```
Loop {
    for i = 1 to m, {
        $\theta_j := \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}$        (for every $j$).
    }
}
```

Linear Regression

$$h(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T x$$

Logistic Regression

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

Interpretation of Hypothesis

$$h_q = \text{estimated probability that } y = 1 \text{ on input } x$$

Example:

$$x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ Hours \end{bmatrix}$$

$$h_q = 0.7$$

Student has a 70% chance of passing

More formally:

$$h_q(x) = p\left(y = 1 | x, q\right)$$

Summary

- (Linear) Regression → Regression technique
- Logistic Regression → Classification technique
- Batch/Mini-Batch/Stochastic – Gradient Descent → Optimization technique
- Important tuning parameters
  - Learning rate → speed and convergence
  - Polynomials → degrees of freedom
  - Regularization