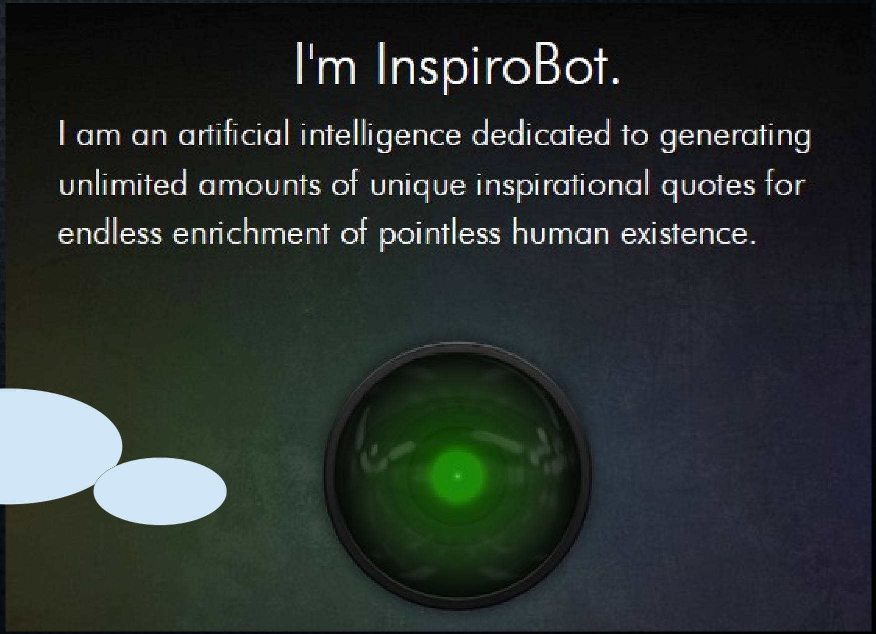
Daily Inspiration



Today

- Recap yesterday
- Logistic regression: using regression tools for classification
- Neural network basics

Yesterday

- Cost function: (differentiable) function that shows how wrong an estimate is for given parameters.
- Gradient descent: one common way to minimise the cost function automatically, i.e. to get optimal parameters
- Linear regression: very simple model that assumes that value to predict is linear combination of input features.
- Overfitting and underfitting, bias and variance: want our model to work wel for unseen data. Need just enough model freedom given the complexity of our problem. How:
  - Cross-validation to measure ability to generalise + get best hyperparameters
  - Use learning curves to diagnose bias vs. variance

Gradient descent in linear algebra

- Goal gradient descent: take a small step in every parameter such that you get closer to the minimum of the cost. Return new theta's.

$$\theta_{0new} = \theta_0 - \frac{\alpha}{m} \sum_{i=1}^{m} \left(\left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right) \cdot 1 \right)$$

$$\theta_{1new} = \theta_1 - \frac{\alpha}{m} \sum_{i=1}^{m} \left(\left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right) \cdot x_1^{(i)} \right)$$

$$\theta_{2new} = \theta_2 - \frac{\alpha}{m} \sum_{i=1}^{m} \left(\left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right) \cdot x_2^{(i)} \right)$$

Gradient descent in linear algebra

We have data, known values, and initial theta's:

$$X = \begin{bmatrix} 1 & feat_1 val_1 & feat_2 val_1 \\ 1 & feat_1 val_2 & feat_2 val_2 \end{bmatrix}; y = \begin{bmatrix} 10.23 \\ -4 \end{bmatrix}; params = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}$$

Get predicted values:

$$\begin{bmatrix} 1 & feat_1 val_1 & feat_2 val_1 \\ 1 & feat_1 val_2 & feat_2 val_2 \end{bmatrix} @ \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 9.23 \\ -2.5 \end{bmatrix}$$

2 by 3 times 3 by 1 gives 2 by 1 (rows by columns)

Get errors:

$$errs = \begin{bmatrix} 9.23 \\ -2.5 \end{bmatrix} - y = \begin{bmatrix} 9.23 \\ -2.5 \end{bmatrix} - \begin{bmatrix} 10.23 \\ -4 \end{bmatrix} = \begin{bmatrix} -1 \\ 1.5 \end{bmatrix}$$

Gradient descent in linear algebra

$$errs = \begin{bmatrix} -1 \\ 1.5 \end{bmatrix}$$

$$\theta_{0new} = \theta_0 - \frac{\alpha}{m} \sum_{i=1}^{m} \left(\boxed{\left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right)} \cdot 1 \right)$$

$$\theta_{1new} = \theta_1 - \frac{\alpha}{m} \sum_{i=1}^{m} \left(\boxed{\left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right)} \cdot x_1^{(i)} \right)$$

$$\theta_{2new} = \theta_2 - \frac{\alpha}{m} \sum_{i=1}^{m} \left(\boxed{\left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right)} \cdot x_2^{(i)} \right)$$

Gradient descent in linear algebra

Calculate, for each feature, sum of each error times that feature:

$$\begin{bmatrix}-1\\1.5\end{bmatrix}^T = \begin{bmatrix}-1 & 1.5\end{bmatrix}$$

$$\begin{bmatrix}-1 & 1.5\end{bmatrix} @ \begin{bmatrix}1 & feat_1val_1 & feat_2val_1\\1 & feat_1val_2 & feat_2val_2\end{bmatrix} =$$

$$\begin{bmatrix}-1\cdot 1 + 1.5\cdot 1 & -1\cdot feat_1val_1 + 1.5\cdot feat_1val_2 & -1\cdot feat_2val_1 + 1.5\cdot feat_2val_2\end{bmatrix}$$

$$\boxed{errs = \begin{bmatrix}-1\\1.5\end{bmatrix}}$$

Gradient descent in linear algebra

Calculate, for each feature, sum of each error times that feature:

$$\begin{bmatrix} -1 \\ 1.5 \end{bmatrix}^T = \begin{bmatrix} -1 & 1.5 \end{bmatrix}$$

$$\begin{bmatrix} -1 & 1.5 \end{bmatrix} @ \begin{bmatrix} 1 & feat_1val_1 & feat_2val_1 \\ 1 & feat_1val_2 & feat_2val_2 \end{bmatrix} =$$

$$errs = \begin{bmatrix} -1 \\ 1.5 \end{bmatrix}$$

$$\begin{bmatrix} \boxed{-1 \cdot 1 + 1.5 \cdot 1} & \boxed{-1 \cdot feat_1val_1 + 1.5 \cdot feat_1val_2} & \boxed{-1 \cdot feat_2val_1 + 1.5 \cdot feat_2val_2} \end{bmatrix}$$

$$\theta_{0new} = \theta_0 - \frac{\alpha}{m} \boxed{\sum_{i=1}^{m} \left(\left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right) \cdot 1 \right)}$$

$$\theta_{1new} = \theta_1 - \frac{\alpha}{m} \boxed{\sum_{i=1}^{m} \left(\left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right) \cdot x_1^{(i)} \right)}$$

$$\theta_{2new} = \theta_2 - \frac{\alpha}{m} \boxed{\sum_{i=1}^{m} \left(\left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right) \cdot x_2^{(i)} \right)}$$

Gradient descent in linear algebra

Now all that we need to do is multiply with α/m and subtract from our old theta's:

$$\alpha/m \cdot \begin{bmatrix} -1 \cdot 1 + 1.5 \cdot 1 & -1 \cdot feat_1 val_1 + 1.5 \cdot feat_1 val_2 & -1 \cdot feat_2 val_1 + 1.5 \cdot feat_2 val_2 \end{bmatrix}$$
$$= \begin{bmatrix} \frac{\alpha}{m}(-1 \cdot 1 + 1.5 \cdot 1) & \frac{\alpha}{m}(-1 \cdot feat_1 val_1 + 1.5 \cdot feat_1 val_2) & \frac{\alpha}{m}(-1 \cdot feat_2 val_1 + 1.5 \cdot feat_2 val_2) \end{bmatrix}$$

Transpose it:

$$\begin{bmatrix} \frac{\alpha}{m}(-1 \cdot 1 + 1.5 \cdot 1) & \frac{\alpha}{m}(-1 \cdot feat_1 val_1 + 1.5 \cdot feat_1 val_2) & \frac{\alpha}{m}(-1 \cdot feat_2 val_1 + 1.5 \cdot feat_2 val_2) \end{bmatrix}^T = \begin{bmatrix} \frac{\alpha}{m}(-1 \cdot 1 + 1.5 \cdot 1) \\ \frac{\alpha}{m}(-1 \cdot feat_1 val_1 + 1.5 \cdot feat_1 val_2) \\ \frac{\alpha}{m}(-1 \cdot feat_2 val_1 + 1.5 \cdot feat_2 val_2) \end{bmatrix}$$

So finally:

$$\begin{bmatrix} \theta_{0old} \\ \theta_{1old} \\ \theta_{2old} \end{bmatrix} - \begin{bmatrix} \frac{\alpha}{m}(-1 \cdot 1 + 1.5 \cdot 1) \\ \boxed{\frac{\alpha}{m}(-1 \cdot feat_1 val_1 + 1.5 \cdot feat_1 val_2)} \\ \frac{\alpha}{m}(-1 \cdot feat_2 val_1 + 1.5 \cdot feat_2 val_2) \end{bmatrix} = \begin{bmatrix} \theta_{0new} \\ \theta_{1new} \\ \theta_{2new} \end{bmatrix}$$

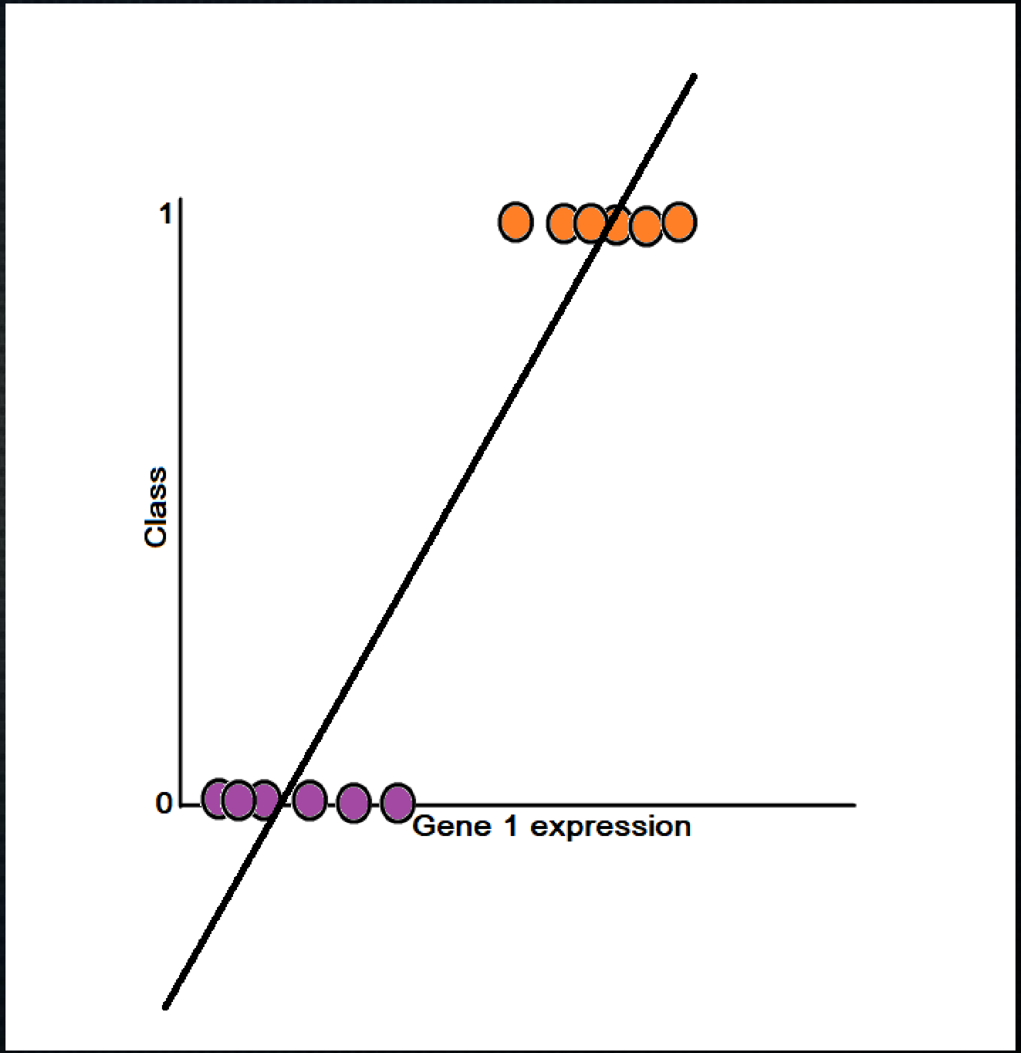
$$\theta_{1new} = \theta_1 - \boxed{\frac{\alpha}{m} \sum_{i=1}^{m} \left(\left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right) \cdot x_1^{(i)} \right)}$$

Logistic regression

- Use regression-like framework for classification

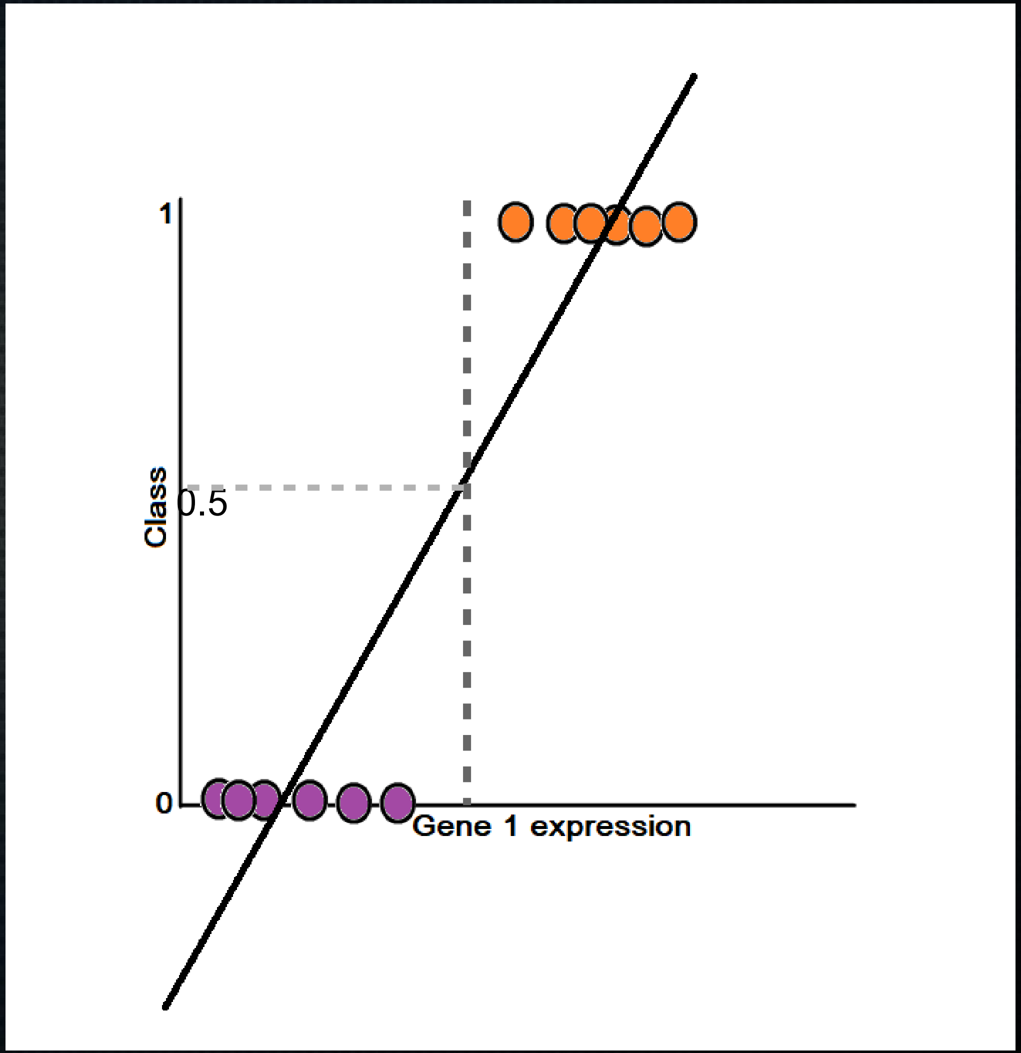
Logistic regression

- Naïve idea:
  Train a linear regression. If Class >= 0.5, predict class 1. Otherwise, class 0.



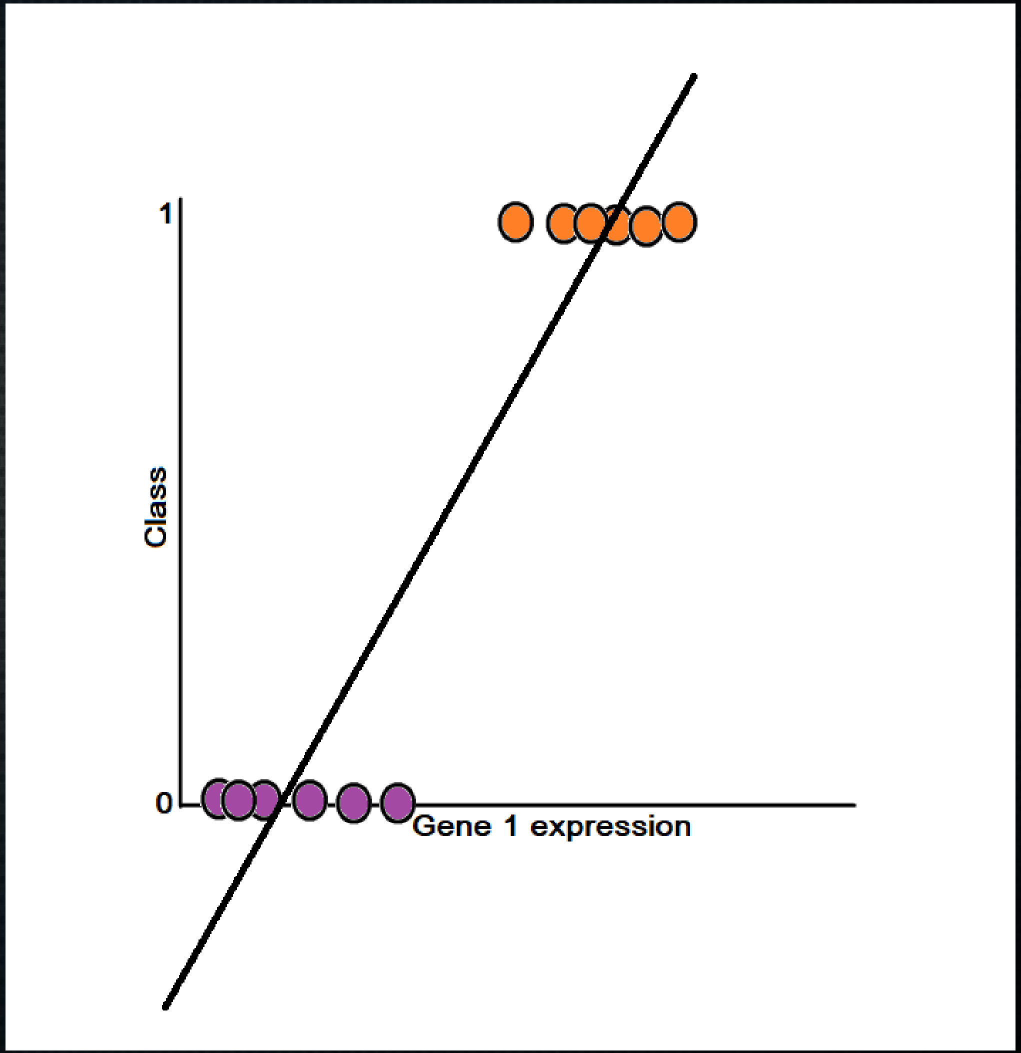
Logistic regression

- Naïve idea:
  Train a linear regression. If Class >= 0.5, predict class 1. Otherwise, class 0.



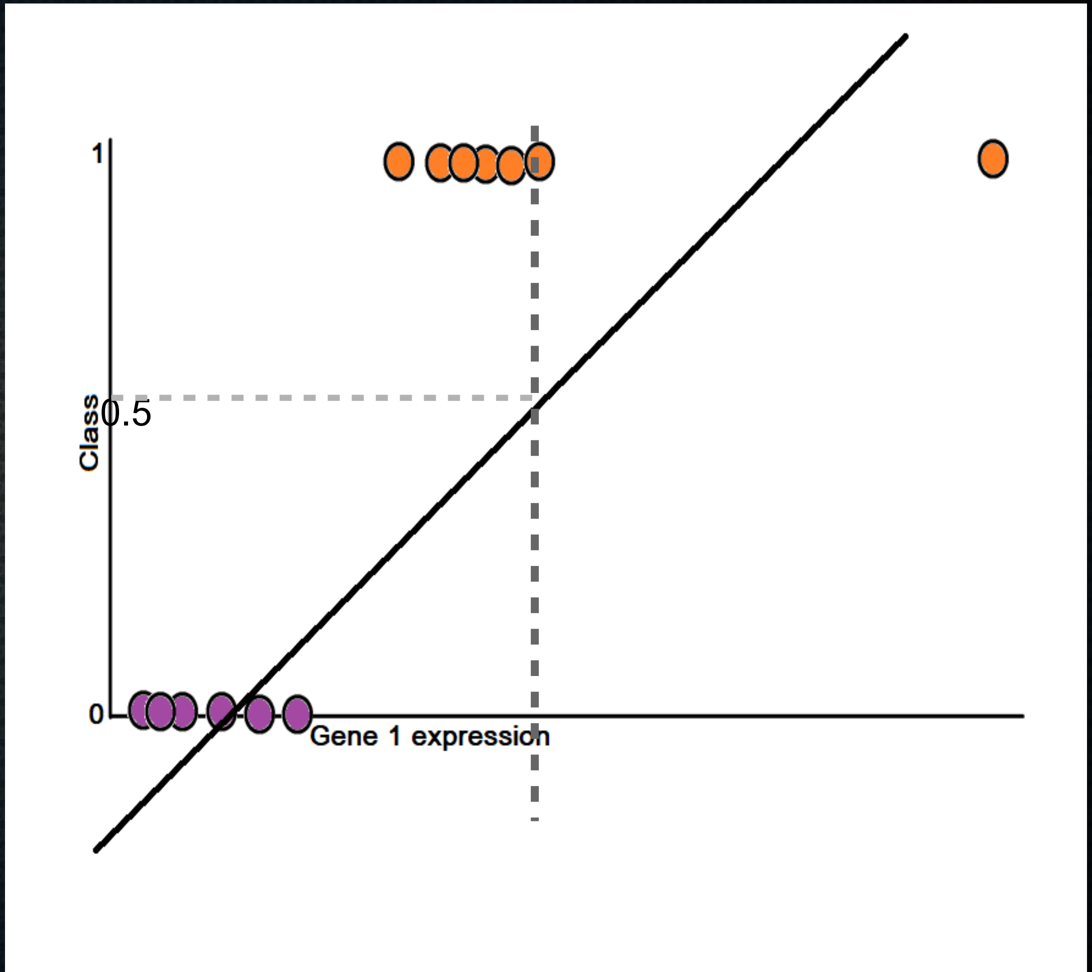
Logistic regression

- Naïve idea:
  Train a linear regression. If Class >= 0.5, predict class 1. Otherwise, class 0.
- Problems:
  -You can predict class > 1 and < 0, while that is not possible in reality.



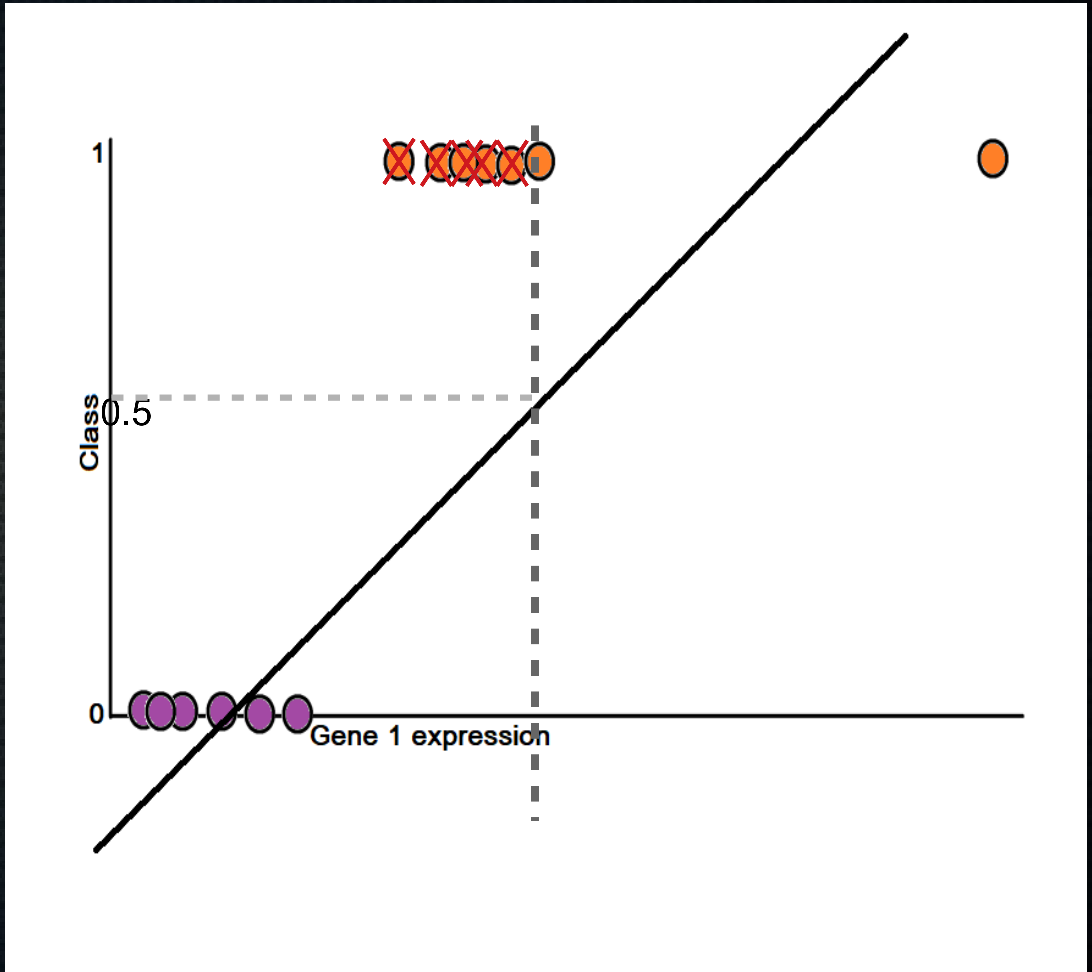
Logistic regression

- Naïve idea:
  Train a linear regression. If Class >= 0.5, predict class 1. Otherwise, class 0.
- Problems:
  -You can predict class > 1 and < 0, while that is not possible in reality.
  -This example seemed to work, but quickly breaks down →



Logistic regression

- Naïve idea:
  Train a linear regression. If Class >= 0.5, predict class 1. Otherwise, class 0.
- Problems:
  -You can predict class > 1 and < 0, while that is not possible in reality.
  -This example seemed to work, but quickly breaks down → get what is basically confirmation of hypothesis, but perform worse!



Logistic regression

- What we want:
  - Use the information that we only have two classes, 0 or 1.
  - Hypothesis function should output only numbers between 0 or 1.

Sigmoid or logistic function

- Before, our hypothesis function was of the form:

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

Sigmoid or logistic function

- Before, our hypothesis function was of the form:

$$h_\theta(x) = \theta^T \cdot x \longrightarrow \begin{bmatrix} 0.5 & 3 & -1.5 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 3 \\ 8 \end{bmatrix}$$

Sigmoid or logistic function

- Before, our hypothesis function was of the form:

$$h_\theta(x)=\theta^T\cdot x \longrightarrow \underbrace{\begin{bmatrix}0.5 & 3 & -1.5\end{bmatrix}}_{\text{Learned parameters (theta 0 – theta 2)}} \cdot \left.\begin{bmatrix}1\\3\\8\end{bmatrix}\right\}\text{Features for one sample (x0 = 1, intercept term, 2 data-derived features x1 and x2)}$$

Sigmoid or logistic function

- Before, our hypothesis function was of the form:

$$h_\theta(x) = \theta^T \cdot x \longrightarrow \begin{bmatrix} 0.5 & 3 & -1.5 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 3 \\ 8 \end{bmatrix} = 0.5 \cdot 1 + 3 \cdot 3 - 1.5 \cdot 8 = -2.5$$

Sigmoid or logistic function

- Before, our hypothesis function was of the form:

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

- Change that to the following:

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

Sigmoid or logistic function

- Before, our hypothesis function was of the form:

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

- Change that to the following:

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

Sigmoid or logistic function

- Before, our hypothesis function was of the form:

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

- Change that to the following:

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

- What does that look like?

Sigmoid or logistic function

- Before, our hypothesis function was of the form:

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

- Change that to the following:

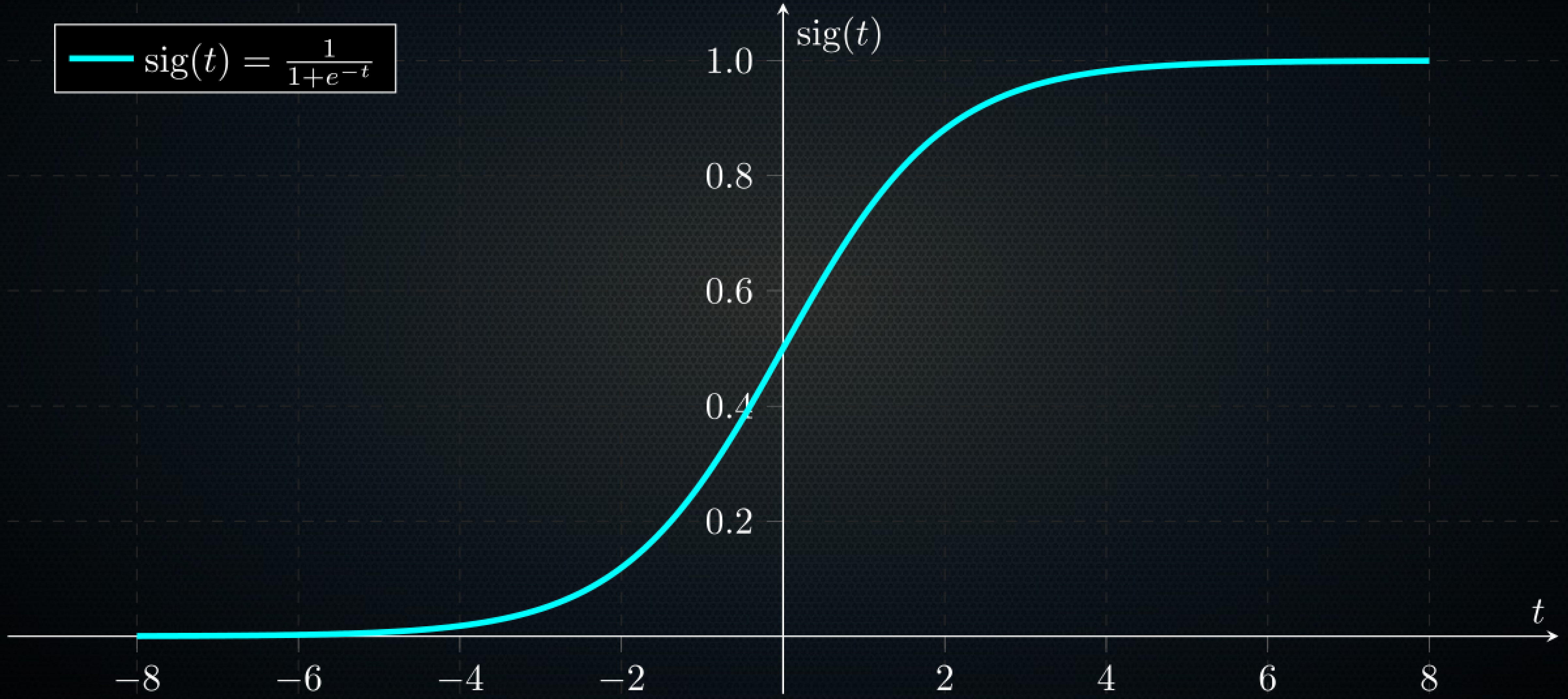
$$h_\theta(x)=g(\theta^T\cdot x) \longrightarrow g(z)=\frac{1}{1+e^{-z}}$$

- What does that look like?

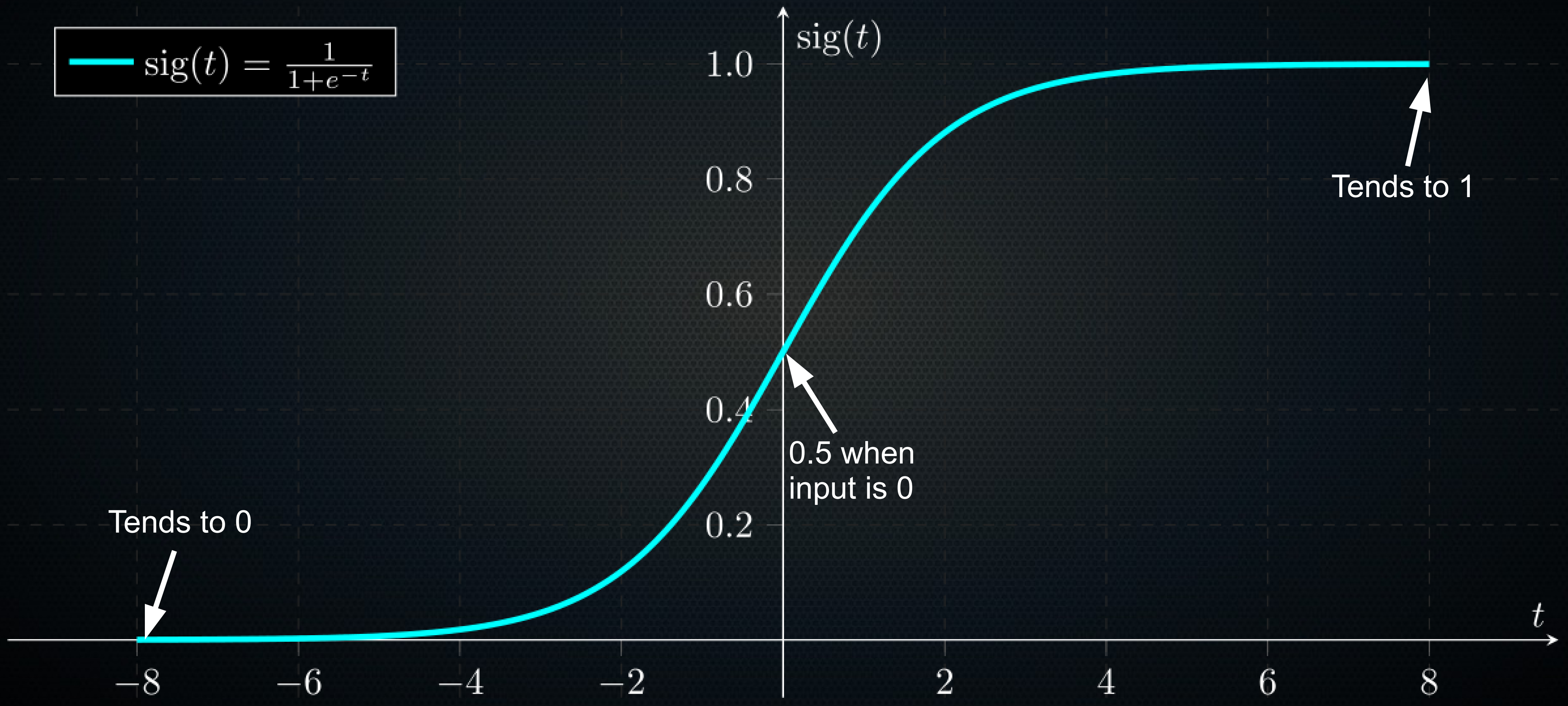
$$z\rightarrow\infty, e^{-z}\rightarrow 0$$

$$z\rightarrow-\infty, e^{-z}\rightarrow\infty$$

What does the sigmoid function look like?

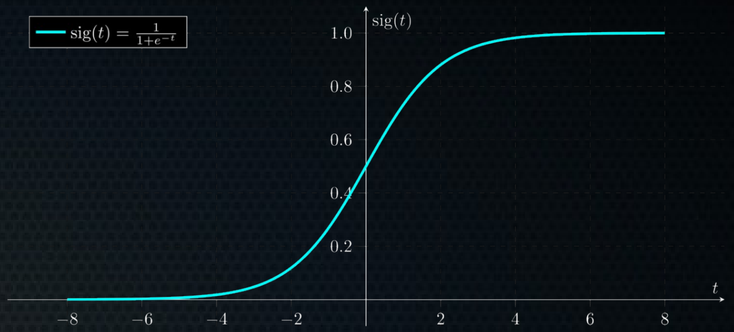


What does the sigmoid function look like?



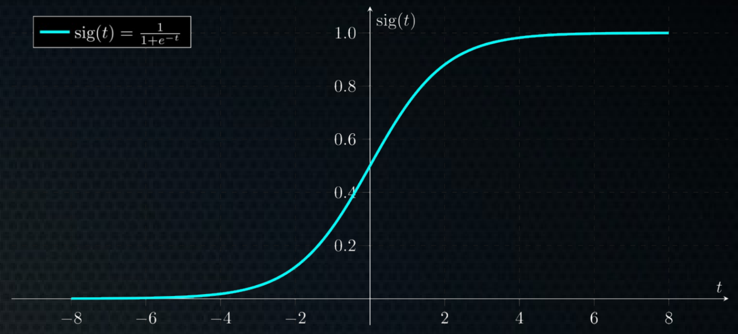
Sigmoid or logistic function

- How do we work with this? $h_\theta(x) = \frac{1}{1+e^{-(\theta^T \cdot x)}}$

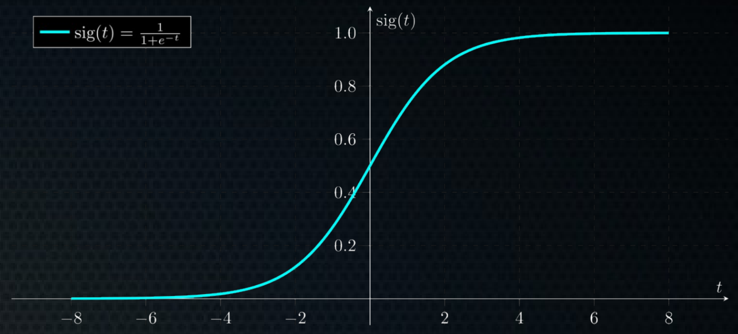


Sigmoid or logistic function

- How do we work with this? $h_\theta(x) = \dfrac{1}{1+e^{-(\theta^T \cdot x)}}$
  - Interpret outcome of $h_\theta(x)$ as probability that class = 1 given the features.



Sigmoid or logistic function



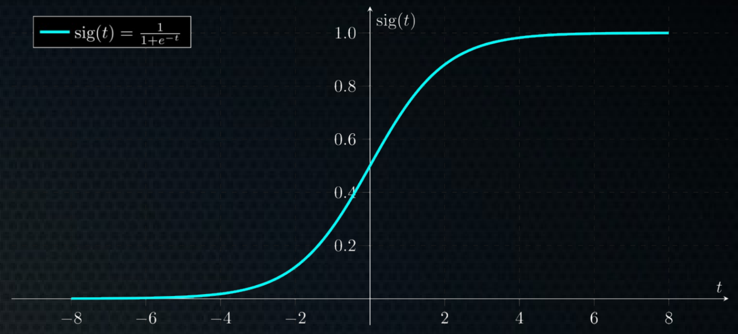
- How do we work with this? $h_\theta(x) = \frac{1}{1+e^{-(\theta^T \cdot x)}}$
  - Interpret outcome of $h_\theta(x)$ as probability that class = 1 given the features. Example:

$$x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{Tumor size} \\ \text{Neovascularisation level} \end{bmatrix}$$

$h_\theta(x) = 0.8$ ⟶ 80% chance of tumor being malignant

Sigmoid or logistic function

- How do we work with this? $h_\theta(x) = \dfrac{1}{1 + e^{-(\theta^T \cdot x)}}$



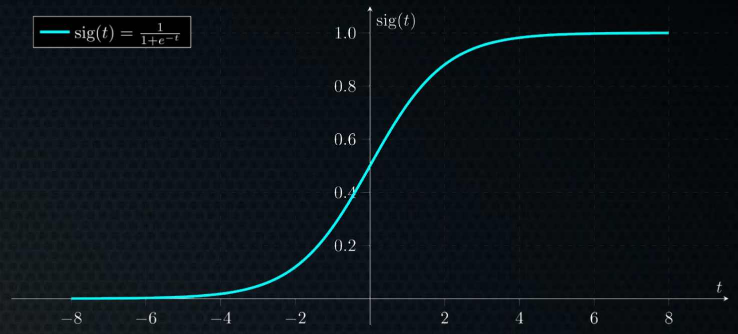
  - Interpret outcome of $h_\theta(x)$ as probability that class = 1 given the features. Example:

$$x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{Tumor size} \\ \text{Neovascularisation level} \end{bmatrix}$$

$h_\theta(x) = 0.8$ ⟶ 80% chance of tumor being malignant (class 1)

100% - 80% → 20 % chance of being benign (class 0)

Sigmoid or logistic function



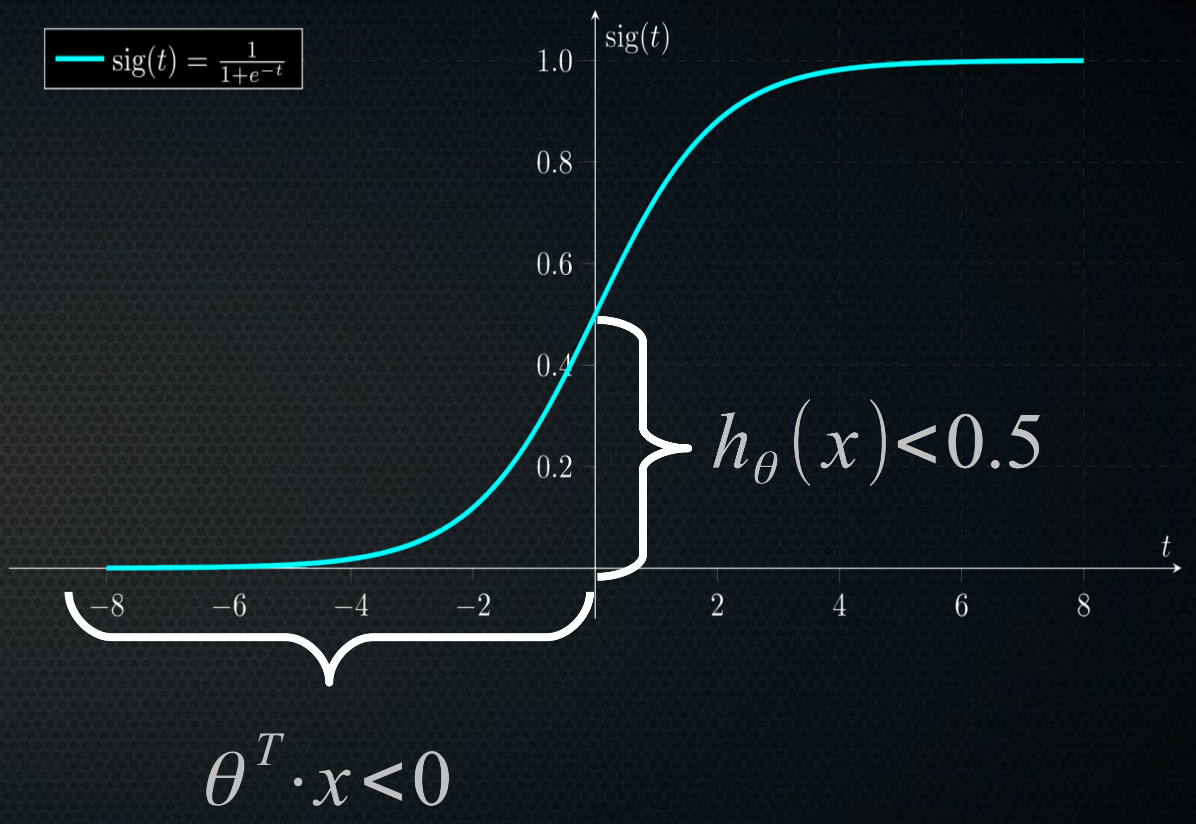
- How do we work with this? $h_\theta(x) = \frac{1}{1+e^{-(\theta^T \cdot x)}}$
  - Interpret outcome of $h_\theta(x)$ as probability that class = 1 given the features.
  - Formally:

$$h_\theta(x) = \frac{1}{1+e^{-(\theta^T \cdot x)}} = p(y=1|x;\theta)$$

$$p(y=0|x;\theta) = 1 - h_\theta(x)$$

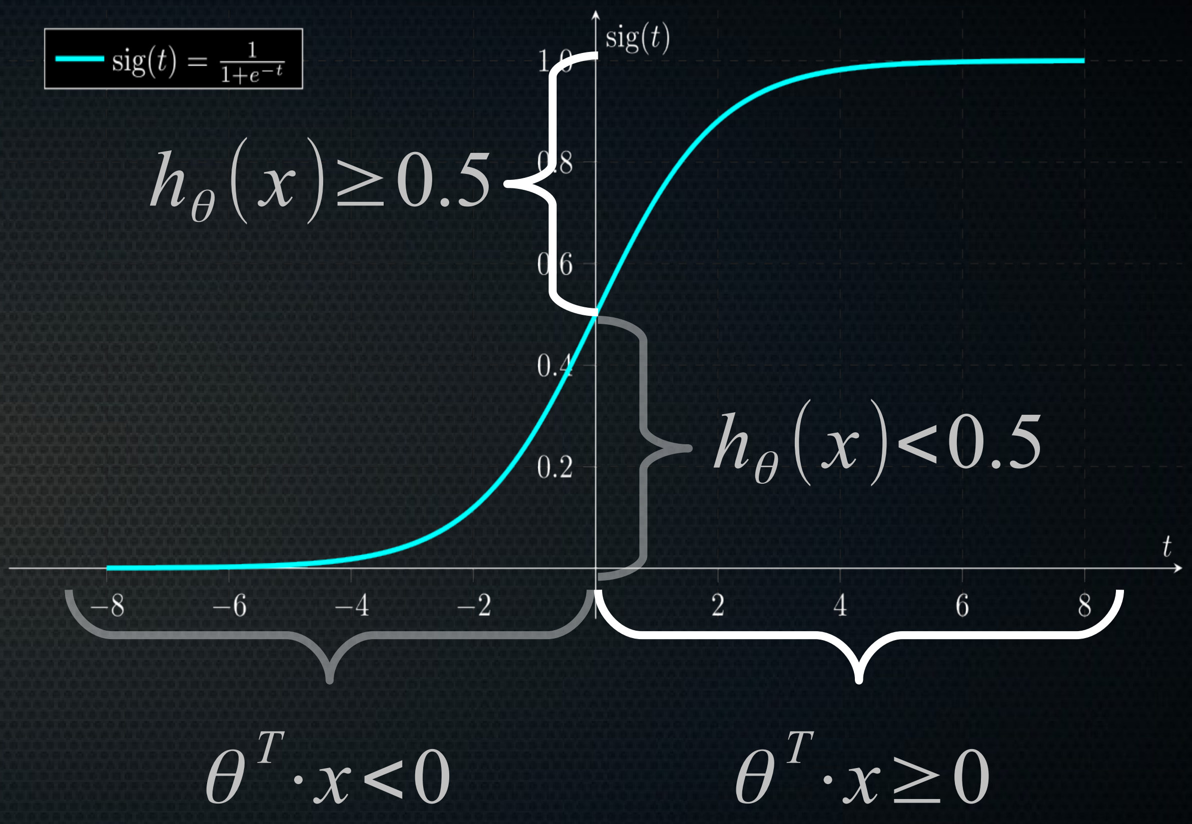
Decision boundary

- Threshold:



Decision boundary

- Threshold:

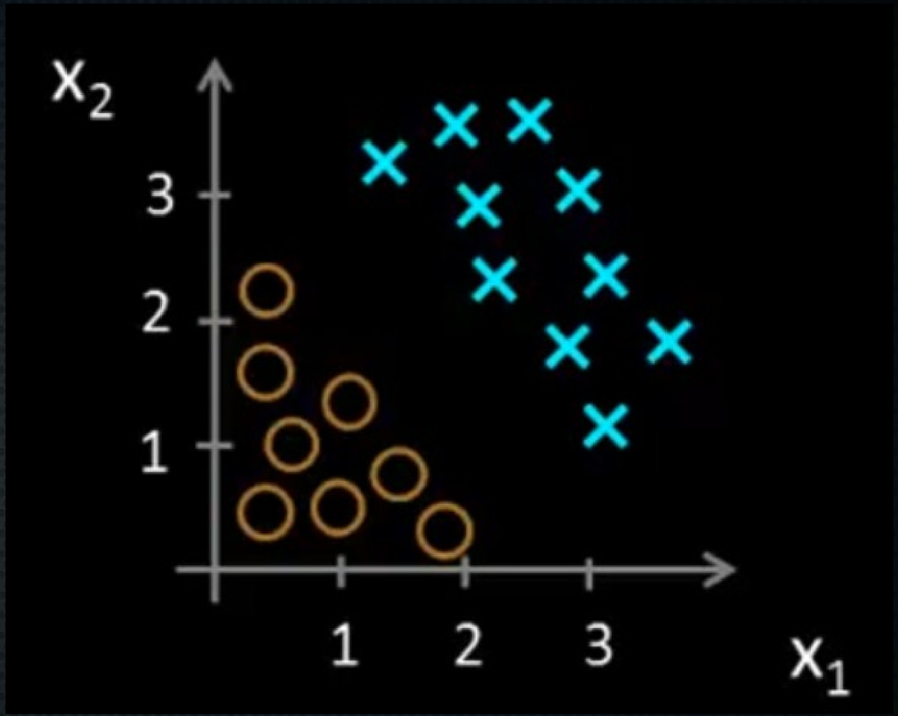


Decision boundary

- How does it look?

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

$$h_\theta(x)=g(\theta_0\cdot x_0,\theta_1\cdot x_1,\theta_2\cdot x_2)$$



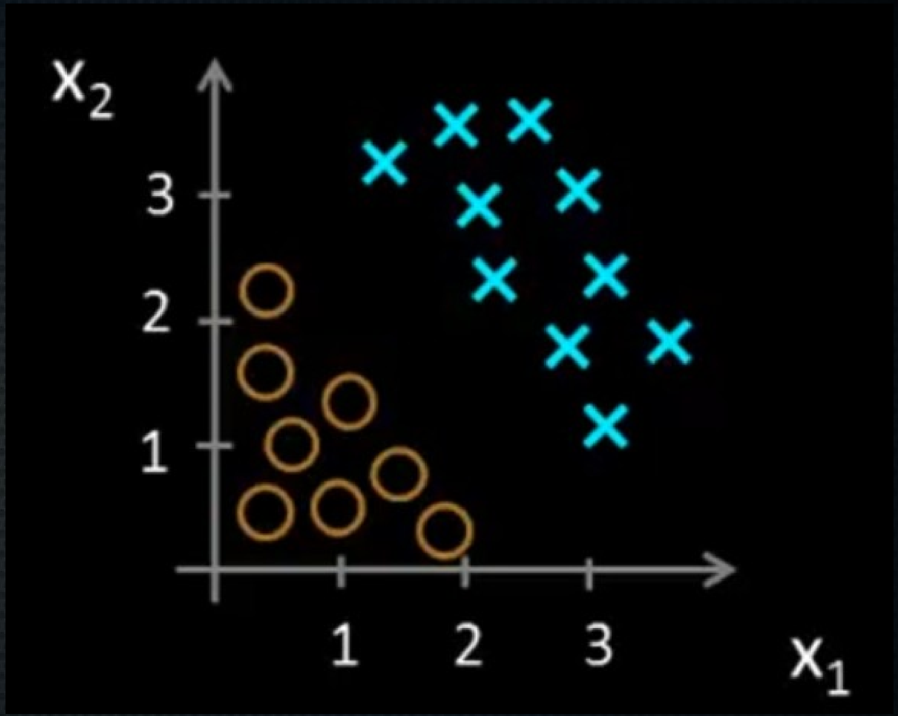
Decision boundary

- How does it look?

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

$$h_\theta(x)=g(\theta_0\cdot x_0,\theta_1\cdot x_1,\theta_2\cdot x_2)$$

$$\theta=\begin{bmatrix}-3\\1\\1\end{bmatrix}$$



Decision boundary

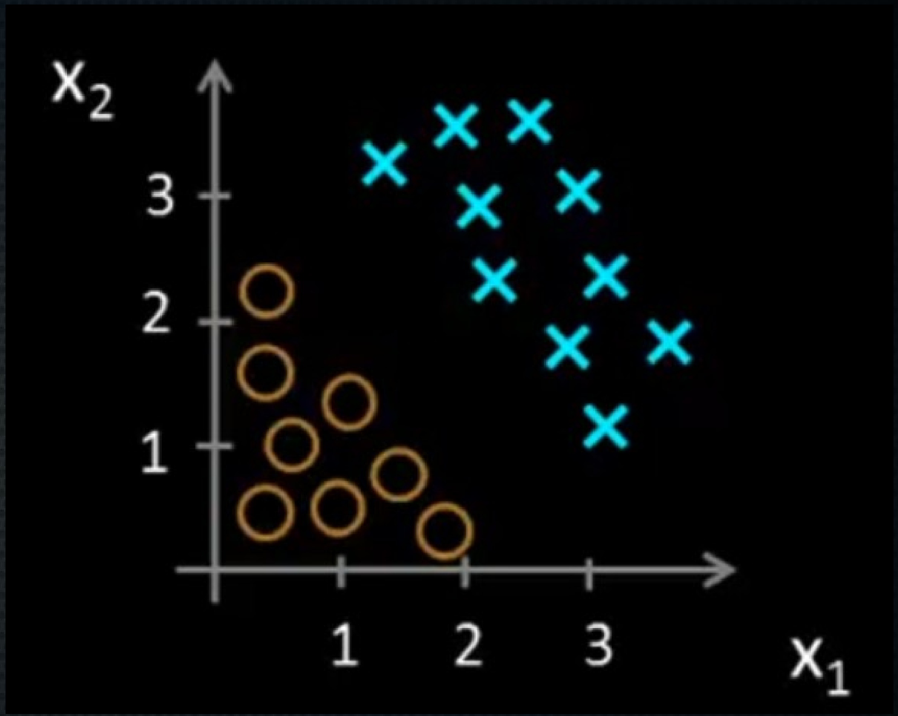
- How does it look?

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

$$h_\theta(x) = g(\theta_0 \cdot x_0, \theta_1 \cdot x_1, \theta_2 \cdot x_2)$$

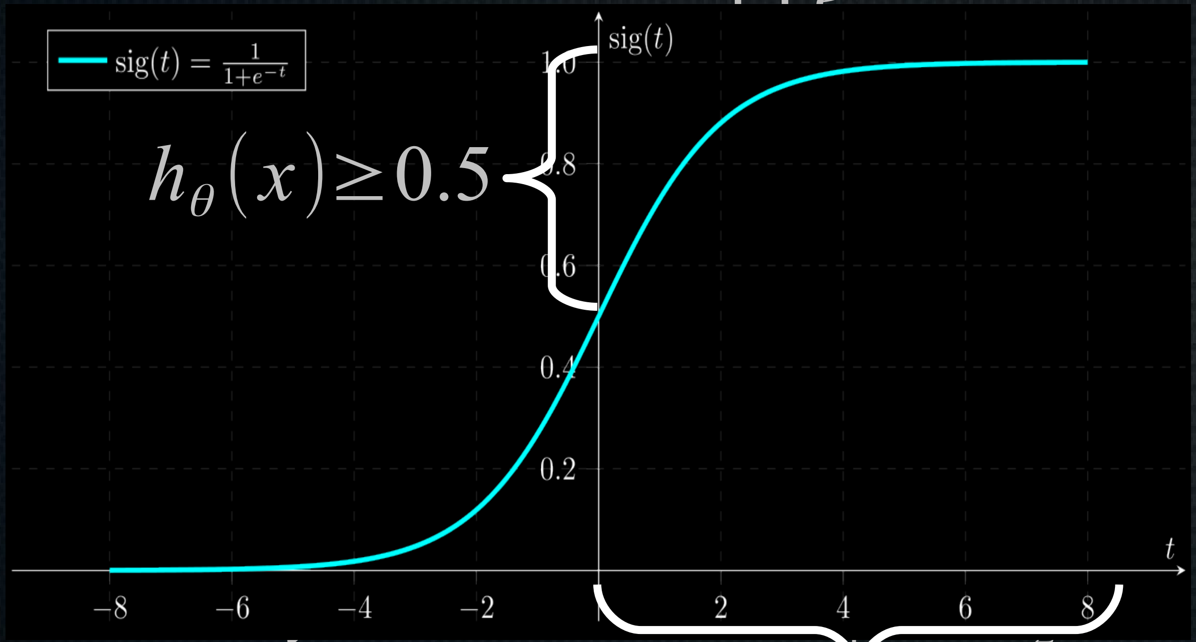
$$\underbrace{\theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}}_{}$$

$$y = 1 \text{ if} -3 \cdot 1 + 1 \cdot x_1 + 1 \cdot x_2 \geq 0$$



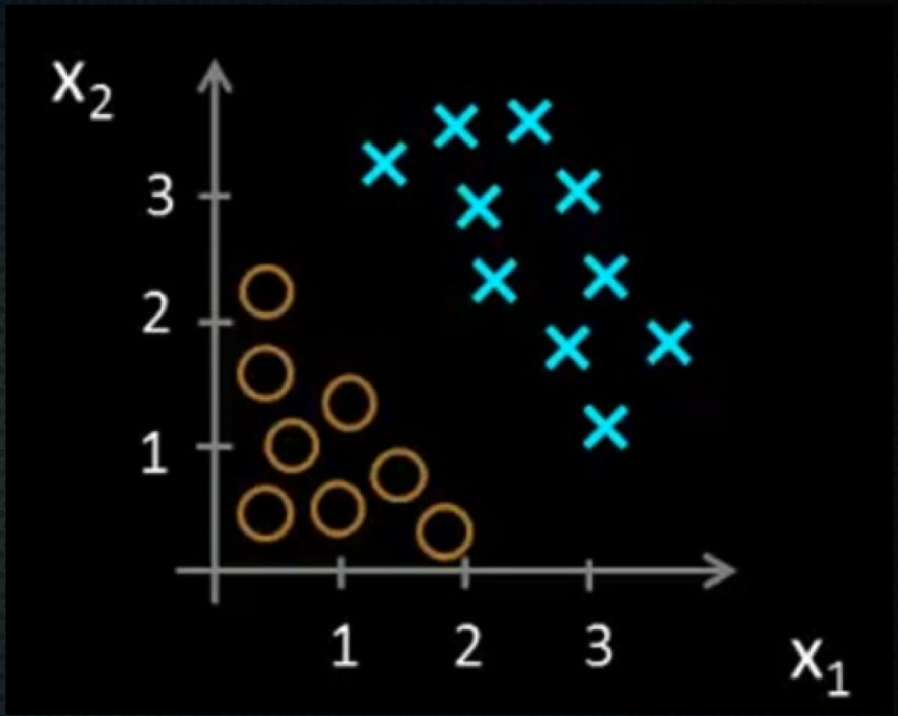
Decision boundary

- How does it look? $g(z)=\frac{1}{1+e^{-z}}$



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

$$\theta^T \cdot x \geq 0$$



Decision boundary

- How does it look?

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

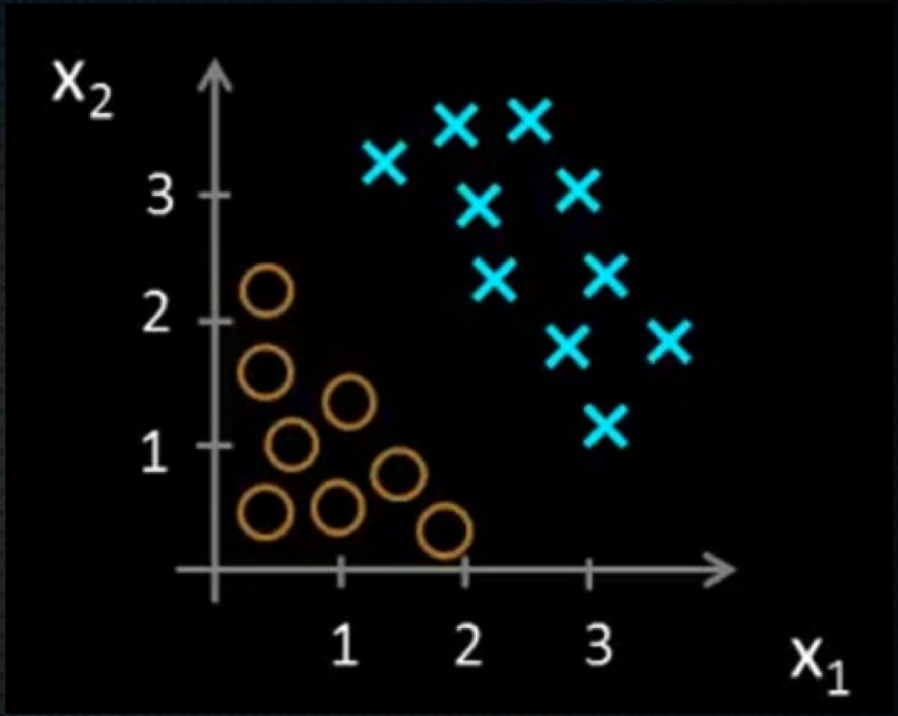
$$h_\theta(x)=g(\theta_0\cdot x_0,\theta_1\cdot x_1,\theta_2\cdot x_2)$$

$$\theta=\begin{bmatrix}-3\\1\\1\end{bmatrix}$$

$$y=1 \text{ if} -3\cdot 1+1\cdot x_1+1\cdot x_2\geq 0$$

$$-3\cdot 1+\cancel{1}\cdot x_1+\cancel{1}\cdot x_2\geq 0$$

$$x_1+x_2=3$$



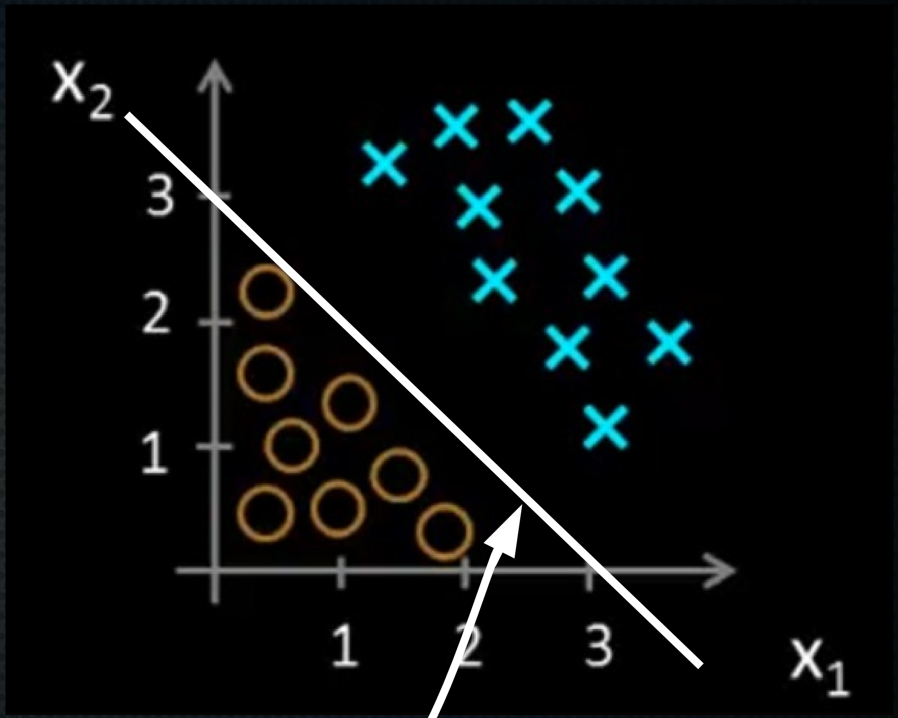
Decision boundary

- How does it look?

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

$$h_\theta(x)=g(\theta_0\cdot x_0,\theta_1\cdot x_1,\theta_2\cdot x_2)$$

$$\theta=\begin{bmatrix}-3\\1\\1\end{bmatrix}$$



$$y=1 \text{ if} -3\cdot 1+1\cdot x_1+1\cdot x_2\geq 0$$

$$-3\cdot 1+\cancel{1}\cdot x_1+\cancel{1}\cdot x_2\geq 0$$

$$x_1+x_2=3$$

Decision boundary

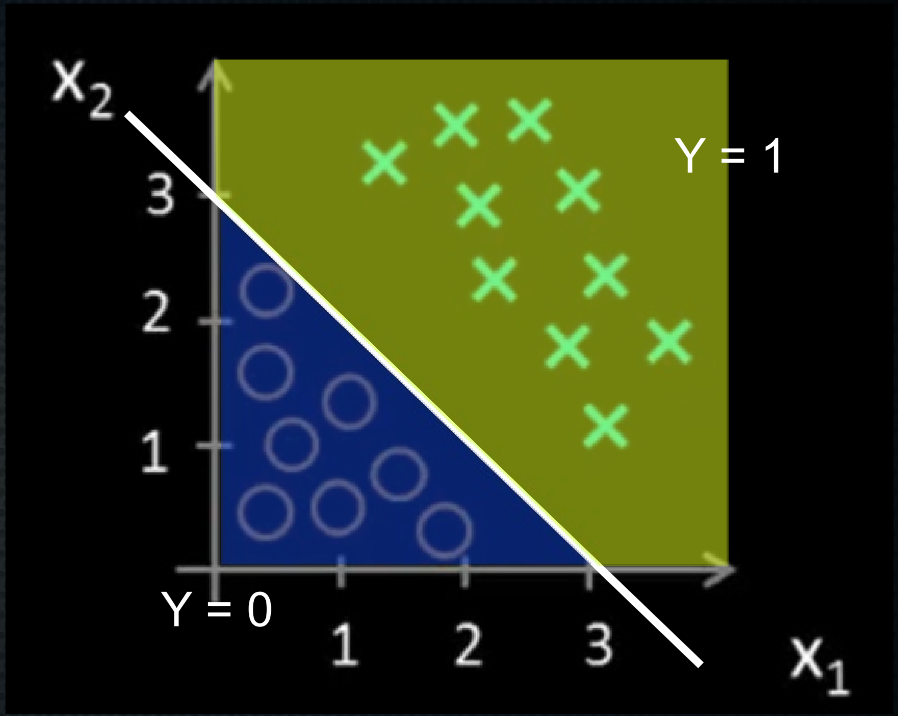
- How does it look?

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

$$h_\theta(x) = g(\theta_0 \cdot x_0, \theta_1 \cdot x_1, \theta_2 \cdot x_2)$$

$$\theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$$

$$y = 1 \text{ if} -3 \cdot 1 + 1 \cdot x_1 + 1 \cdot x_2 \geq 0$$



Decision boundary

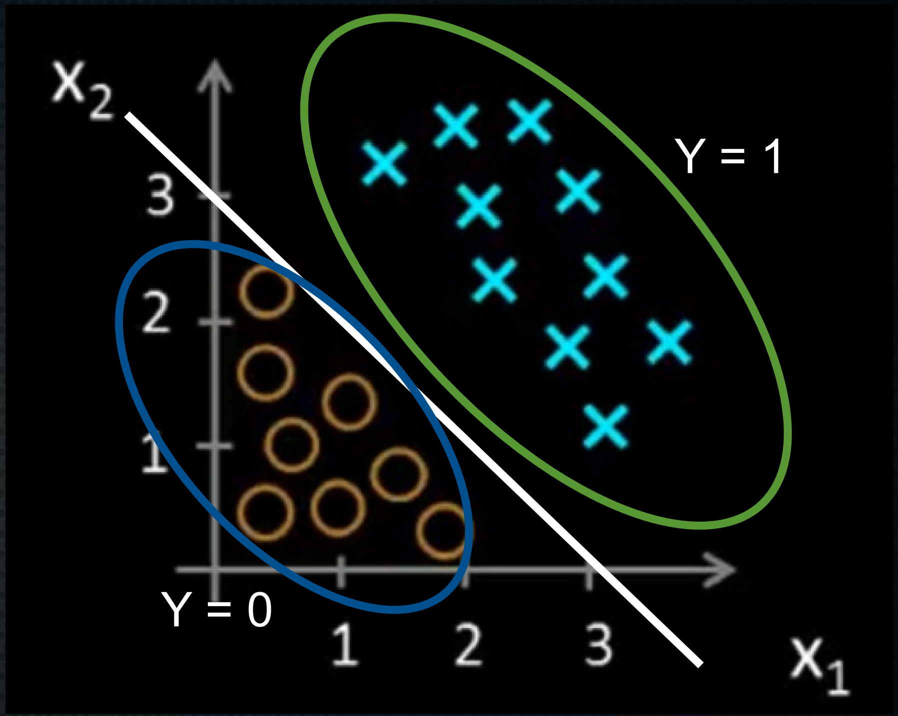
- How does it look?

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

$$h_\theta(x)=g(\theta_0\cdot x_0,\theta_1\cdot x_1,\theta_2\cdot x_2)$$

$$\theta=\begin{bmatrix}-3\\1\\1\end{bmatrix}$$

$$y=1\text{ if}-3\cdot 1+1\cdot x_1+1\cdot x_2\geq 0$$



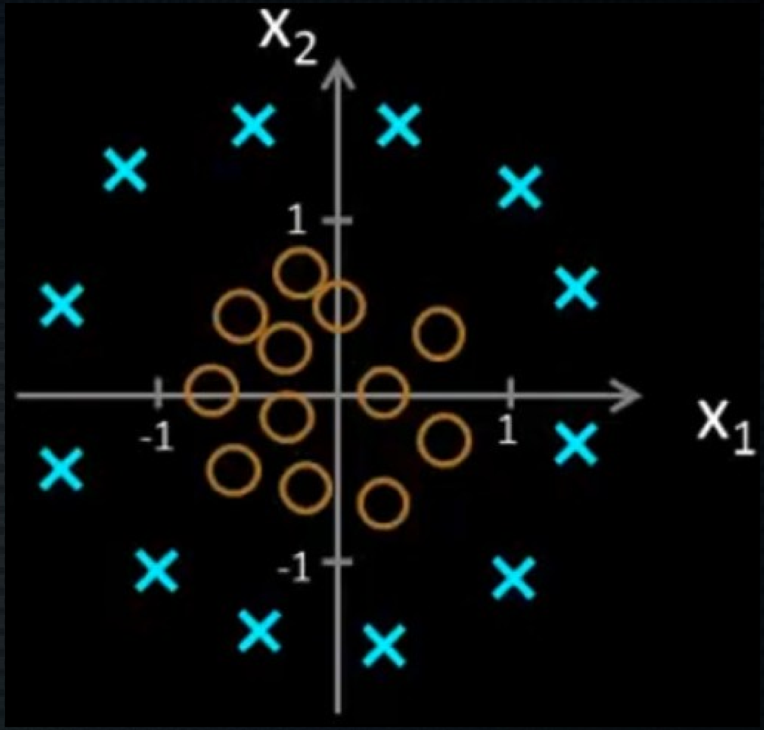
Decision boundary

- How does it look?

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

$$h_\theta(x)=g(\theta_0\cdot x_0,\theta_1\cdot x_1,\theta_2\cdot x_2)$$

$$\theta=\begin{bmatrix}-3\\1\\1\end{bmatrix}$$



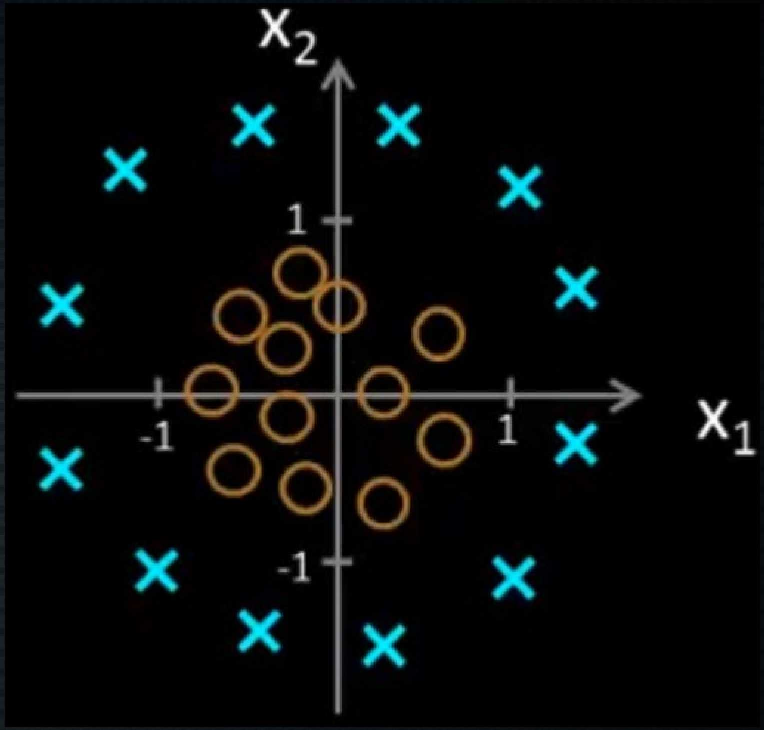
$$y=1\ \text{if} -3\cdot 1+1\cdot x_1+1\cdot x_2\geq 0$$

Non-linear decision boundary

- How does it look? $g(z) = \dfrac{1}{1+e^{-z}}$

$$h_\theta(x) = g(\theta_0 x_0, \theta_1 x_1, \theta_2 x_2, \boxed{\theta_3 x_1^2, \theta_4 x_2^2})$$

- Add two polynomial features



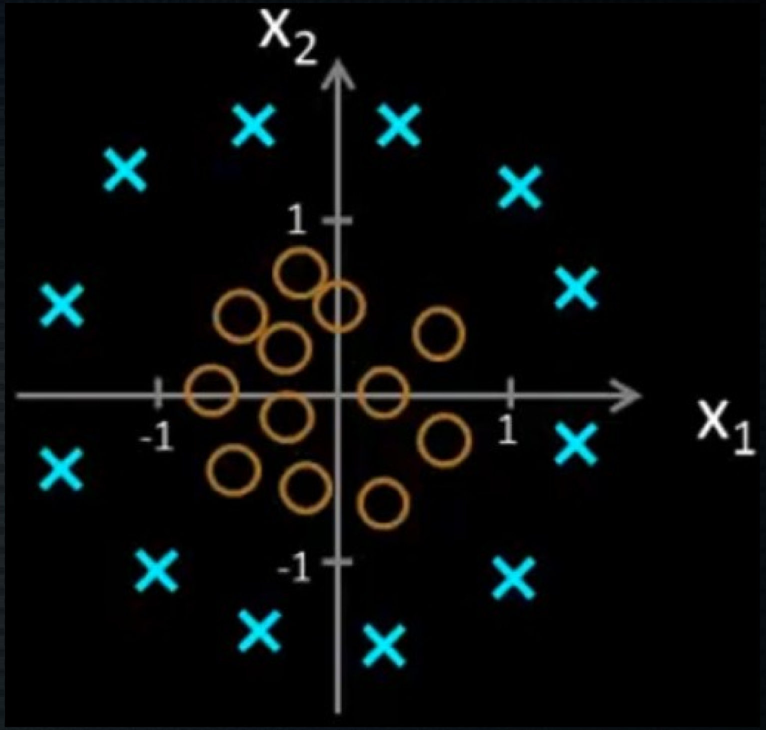
Non-linear decision boundary

- How does it look? $g(z) = \frac{1}{1+e^{-z}}$

$$h_\theta(x) = g(\theta_0 x_0, \theta_1 x_1, \theta_2 x_2, \boxed{\theta_3 x_1^2, \theta_4 x_2^2})$$

- Add two polynomial features

$$\theta = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$



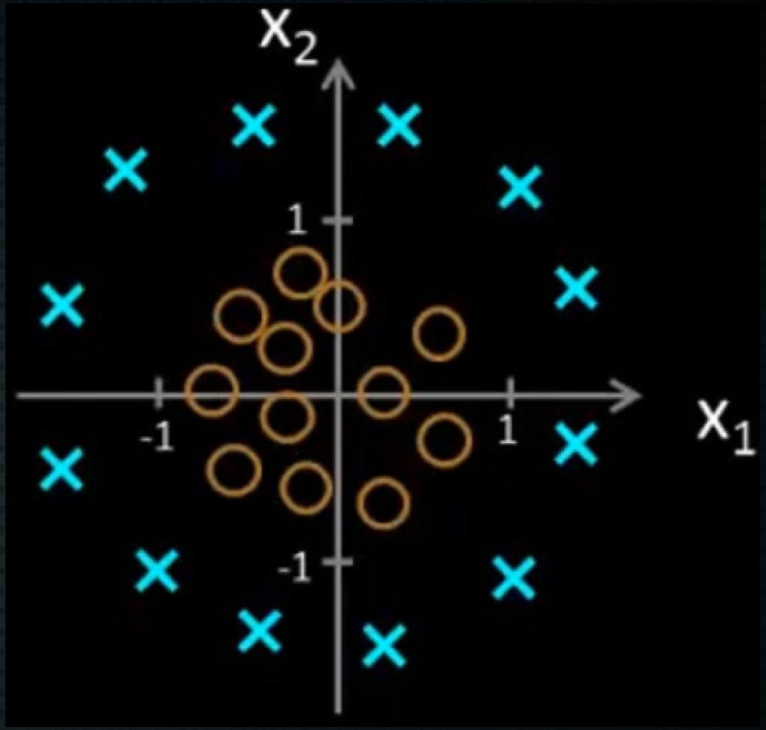
Non-linear decision boundary

- How does it look? $g(z)=\frac{1}{1+e^{-z}}$

$$h_\theta(x)=g(\theta_0 x_0,\theta_1 x_1,\theta_2 x_2,\boxed{\theta_3 x_1^2,\theta_4 x_2^2})$$

- Add two polynomial features

$$\theta=\begin{bmatrix}-1\\0\\0\\1\\1\end{bmatrix} \longrightarrow -1+x_1^2+x_2^2\geq 0$$



Non-linear decision boundary

- How does it look? $g(z)=\dfrac{1}{1+e^{-z}}$

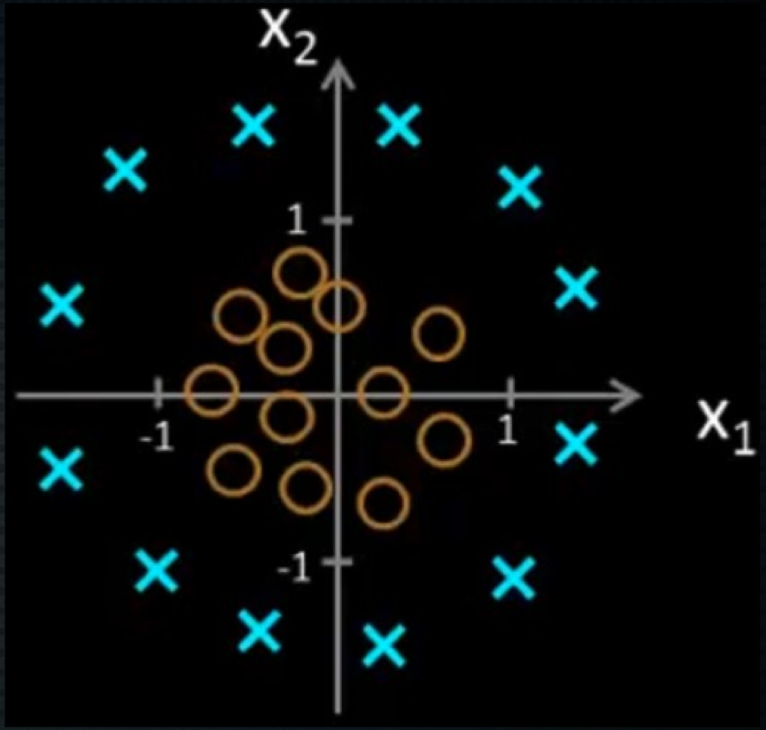
$$h_\theta(x)=g(\theta_0 x_0, \theta_1 x_1, \theta_2 x_2, \boxed{\theta_3 x_1^2, \theta_4 x_2^2)}$$

- Add two polynomial features

$$\theta=\begin{bmatrix}-1\\0\\0\\1\\1\end{bmatrix} \longrightarrow -1+x_1^2+x_2^2\geq 0$$

$$\downarrow$$

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



Non-linear decision boundary

- How does it look? $g(z) = \frac{1}{1+e^{-z}}$

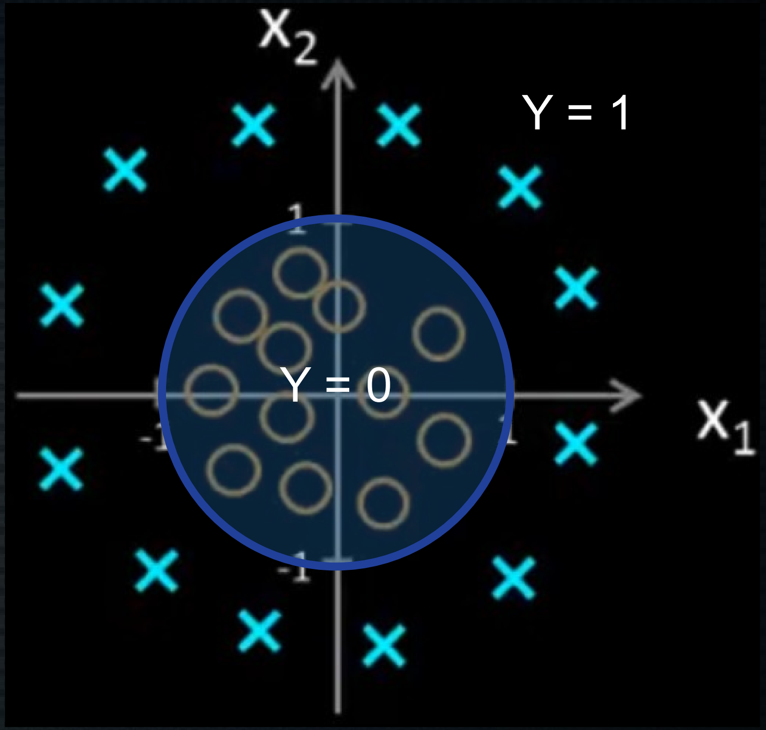
$$h_\theta(x) = g(\theta_0 x_0, \theta_1 x_1, \theta_2 x_2, \boxed{\theta_3 x_1^2, \theta_4 x_2^2})$$

- Add two polynomial features

$$\theta = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \longrightarrow -1 + x_1^2 + x_2^2 \geq 0$$

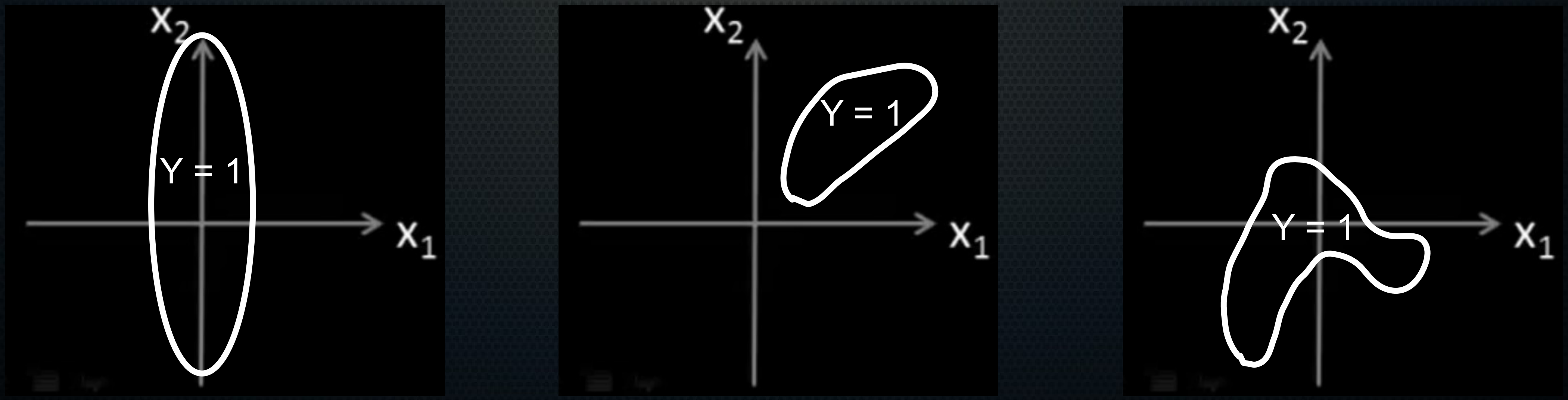
$$\downarrow$$

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



Non-linear decision boundary

- How does it look?
- If you add more and higher-order polynomial features, you can get complex boundaries:



So how do we get theta‘s?

- Need a cost function

So how do we get theta's?

- Need a cost function
- Before: $J(\theta)=\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right)-y^{(i)}\right)^2$

So how do we get theta's?

- Need a cost function
- Before: $$J(\theta)=\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right)-y^{(i)}\right)^2$$

$$\updownarrow$$

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

So how do we get theta's?

- Need a cost function
- Before: $$J(\theta)=\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right)-y^{(i)}\right)^2$$

$$\updownarrow$$

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

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

So how do we get theta‘s?

- Need a cost function
- Before: $$J(\theta)=\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)})-y^{(i)}\right)^2$$

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

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

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

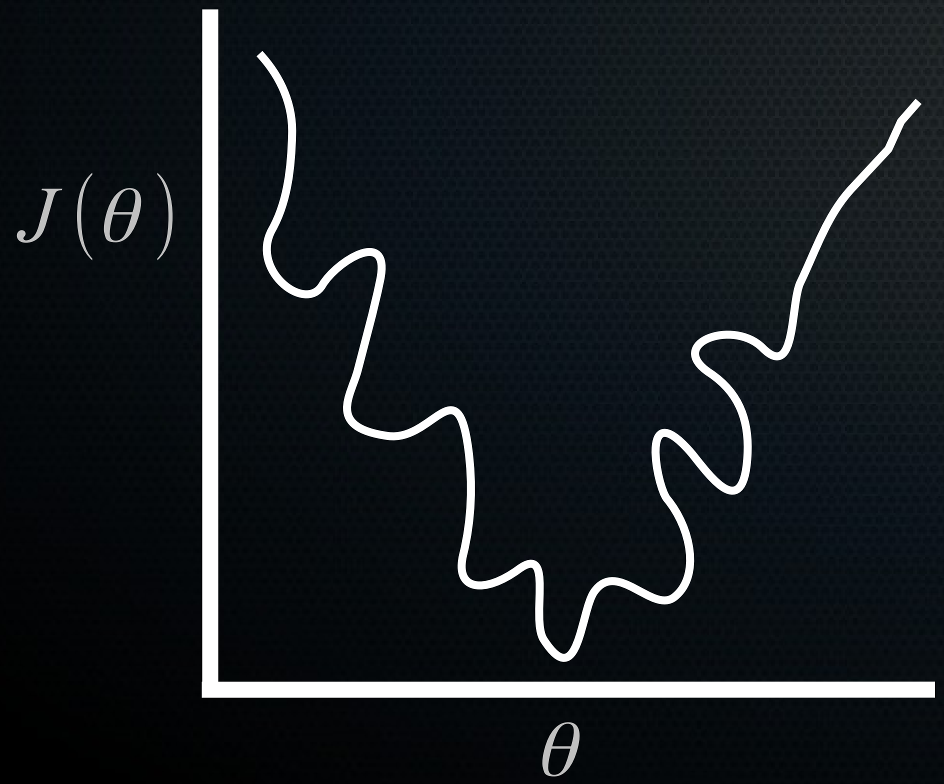
So how do we get theta‘s?

- Need a cost function $J(\theta)=\frac{1}{m}\sum_{i=1}^{m}\mathrm{Cost}(x^i)$ $\mathrm{Cost}(x)=\frac{1}{2}(h_\theta(x)-y)^2$
- Why not MSE? → not convex

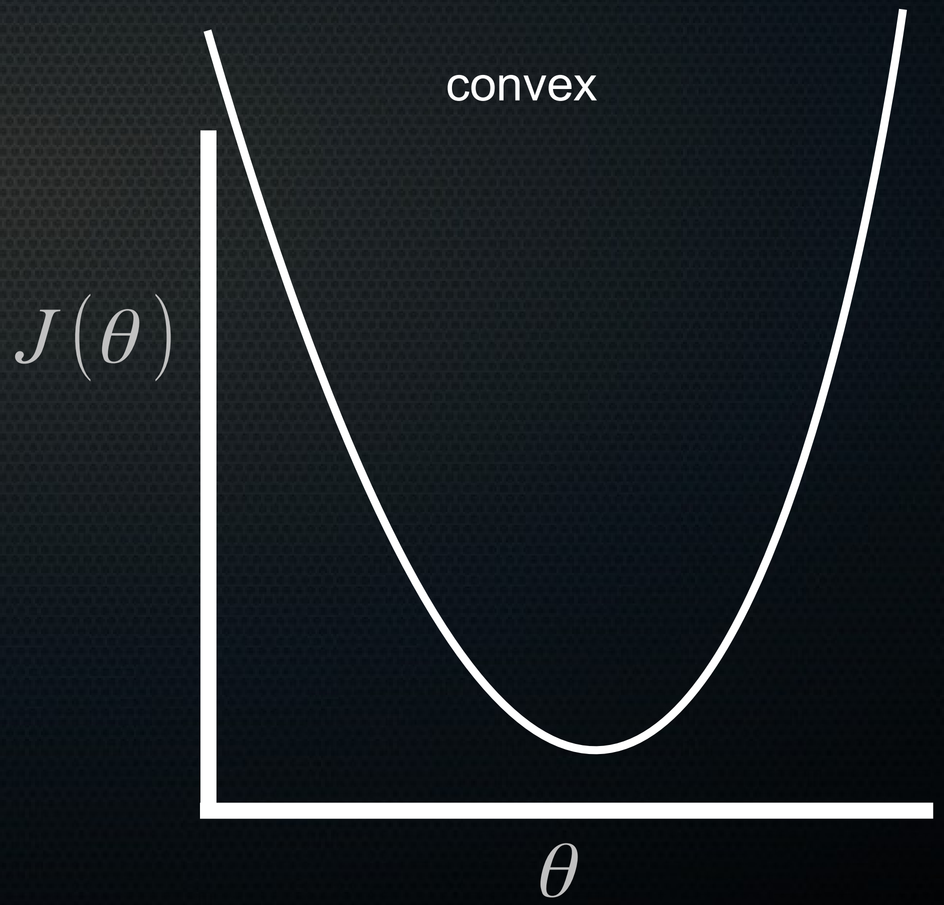
So how do we get theta's?

- Need a cost function $J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \mathrm{Cost}(x^i)$ $\quad \mathrm{Cost}(x) = \frac{1}{2}(h_\theta(x) - y)^2$
- Why not MSE? → not convex

non-convex

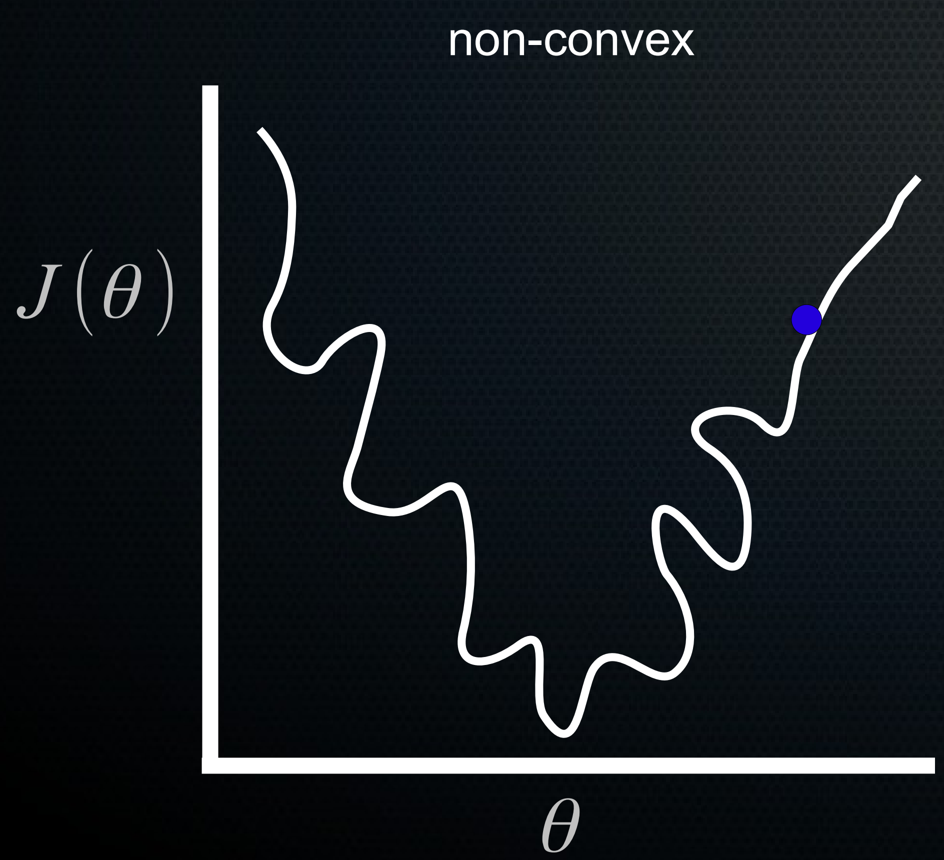


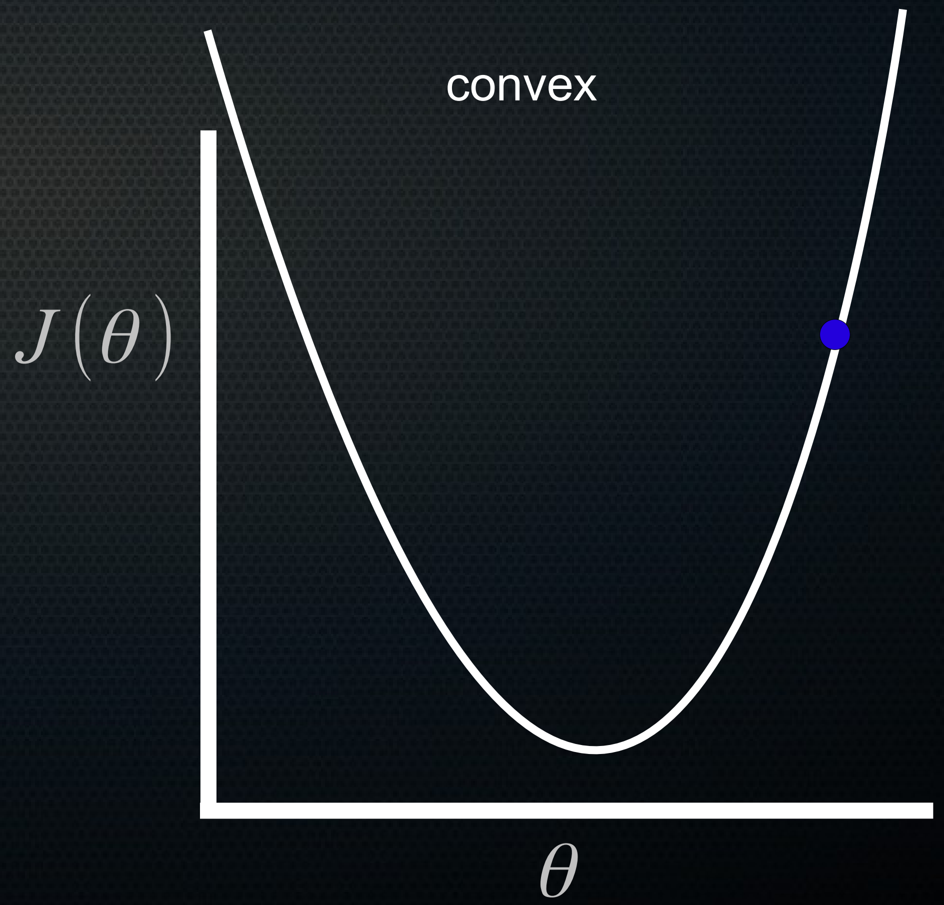
convex



So how do we get theta‘s?

- Need a cost function $J(\theta)=\frac{1}{m}\sum_{i=1}^{m}\mathrm{Cost}(x^i)$ $\quad \mathrm{Cost}(x)=\frac{1}{2}(h_\theta(x)-y)^2$
- Why not MSE? → not convex

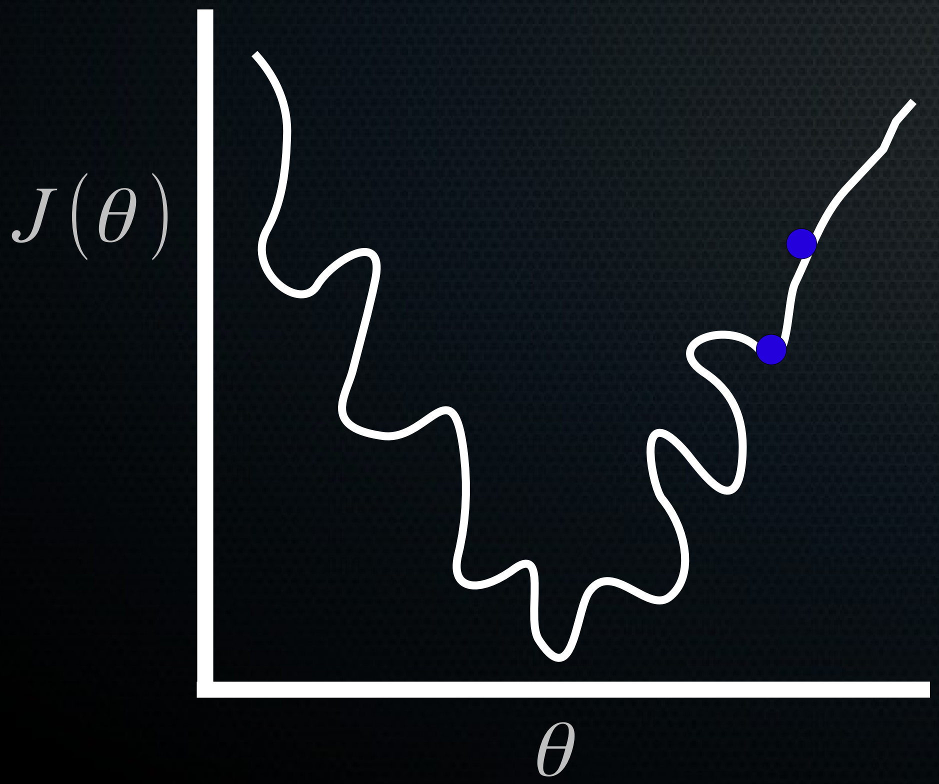




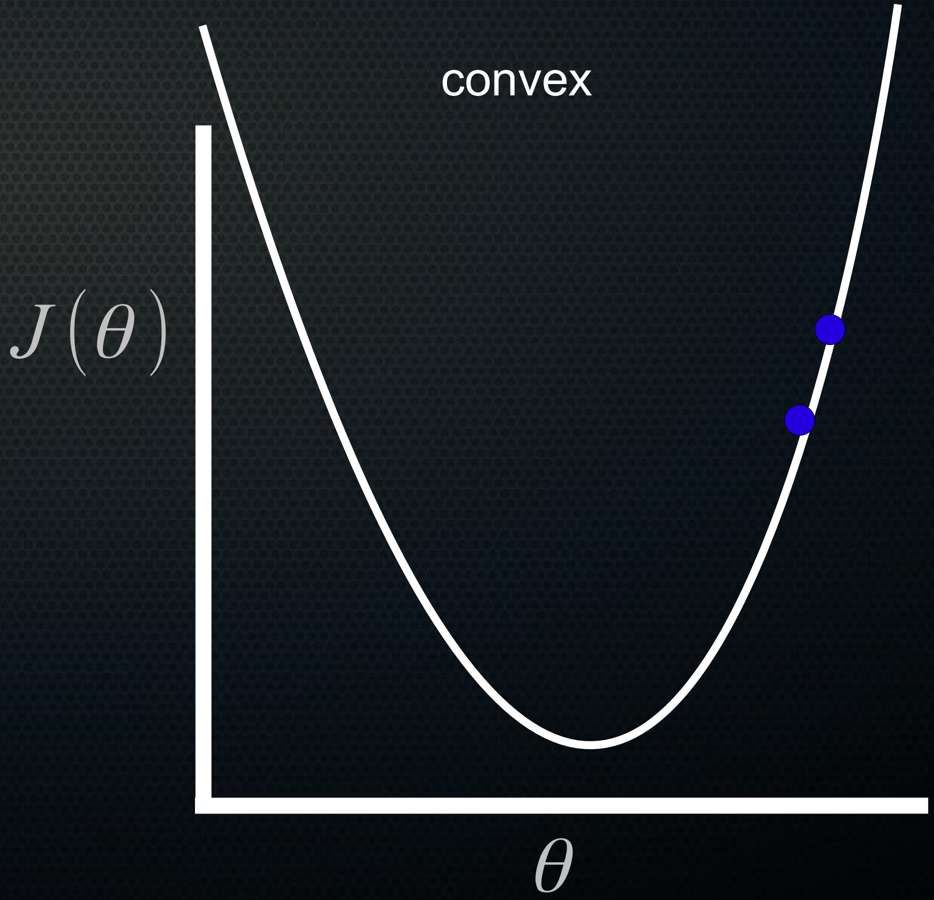
So how do we get theta's?

- Need a cost function $J(\theta)=\frac{1}{m}\sum_{i=1}^{m}\text{Cost}(x^i)$ $\text{Cost}(x)=\frac{1}{2}(h_\theta(x)-y)^2$
- Why not MSE? → not convex

non-convex

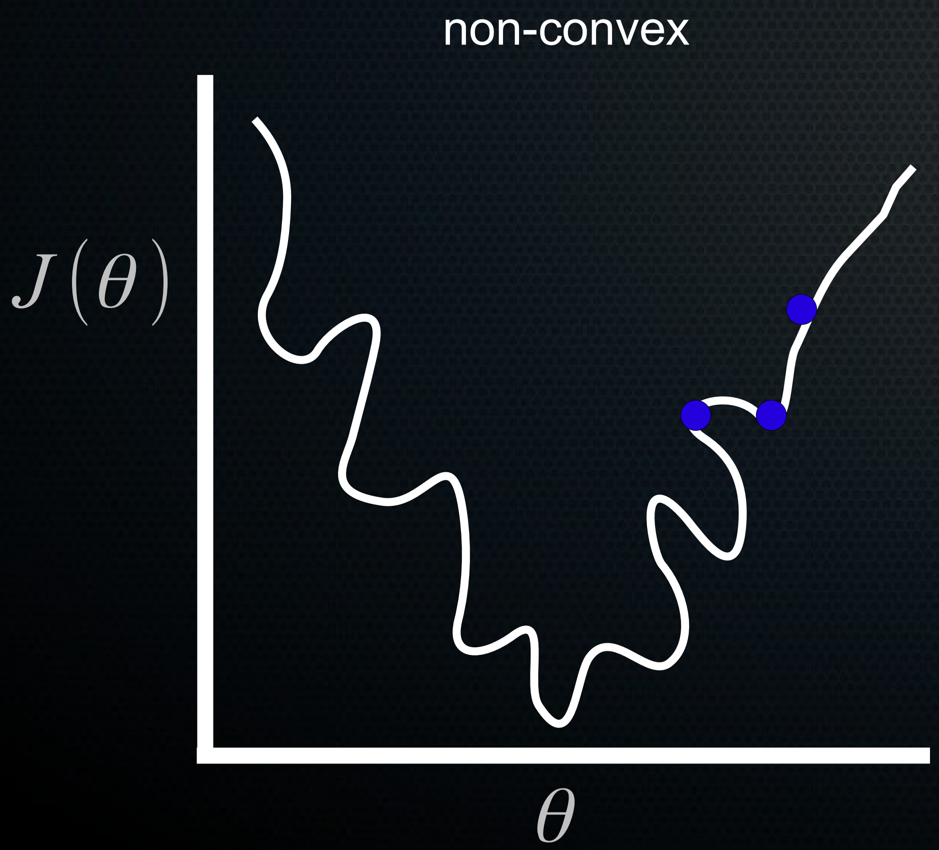


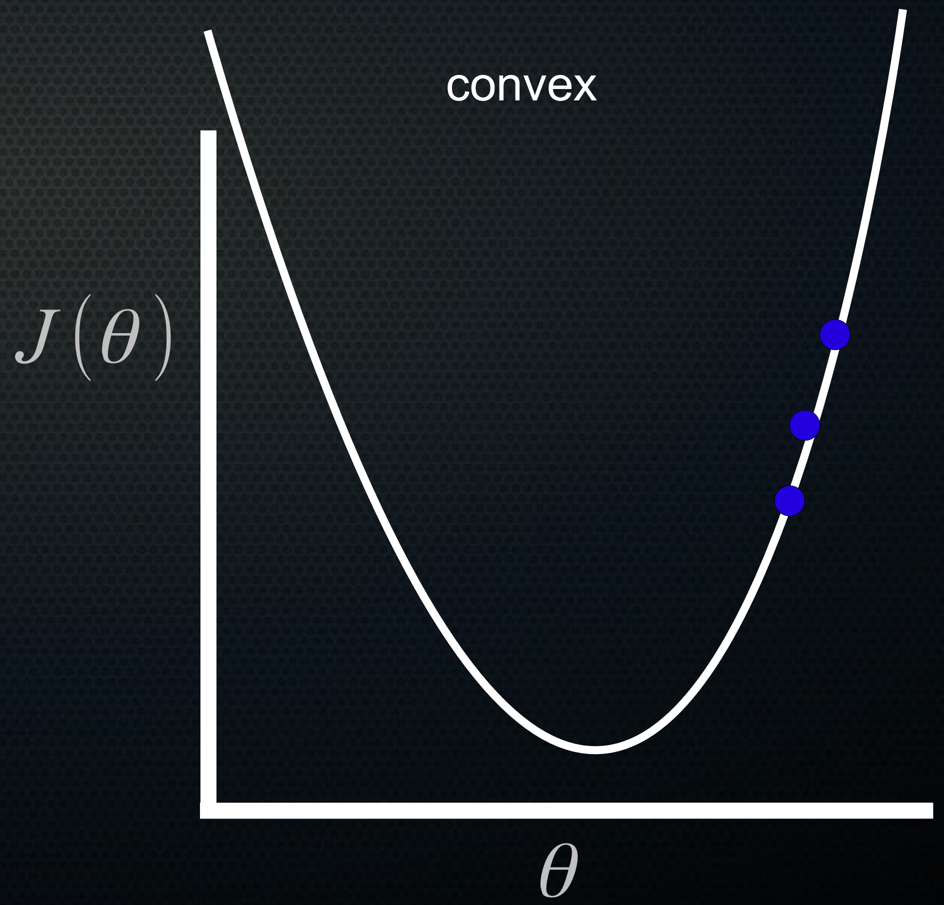
convex



So how do we get theta‘s?

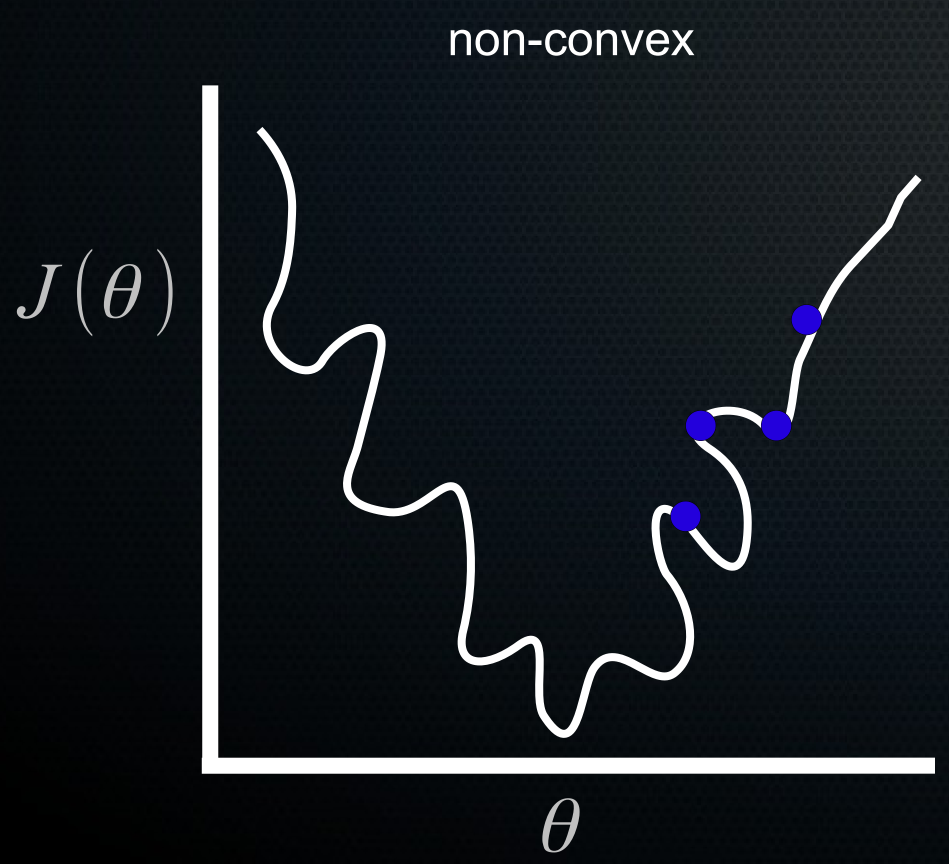
- Need a cost function $J(\theta)=\frac{1}{m}\sum_{i=1}^{m}\mathrm{Cost}(x^i)$ $\quad \mathrm{Cost}(x)=\frac{1}{2}(h_\theta(x)-y)^2$
- Why not MSE? → not convex

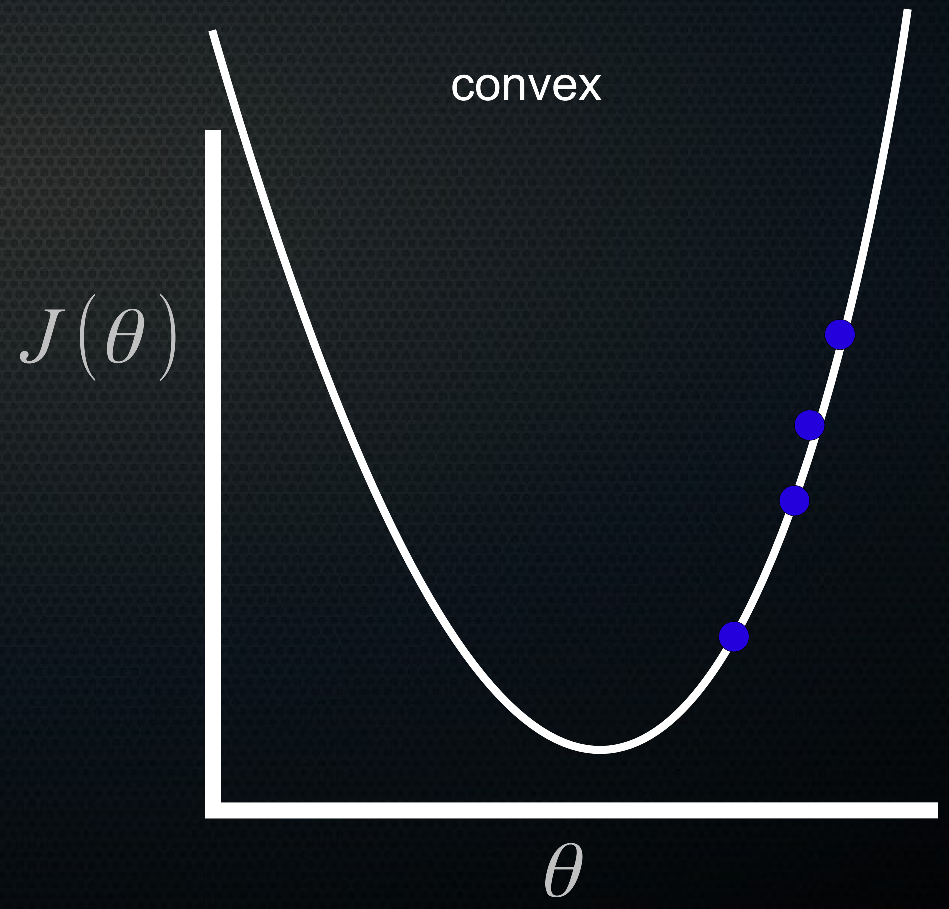




So how do we get theta's?

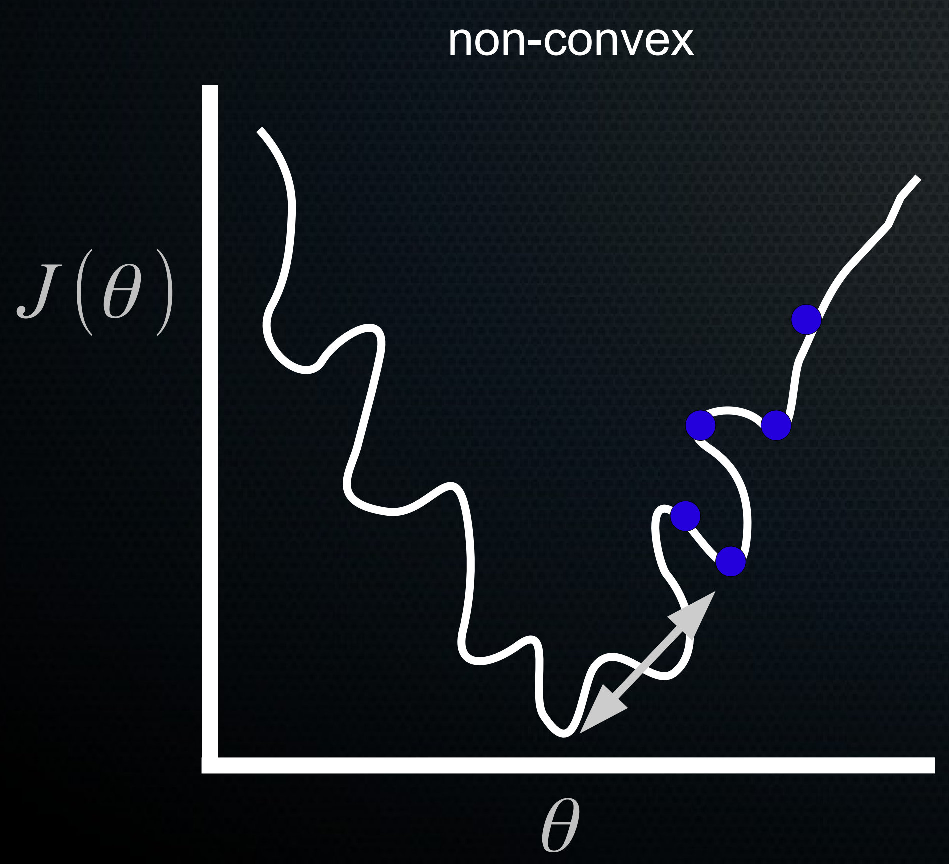
- Need a cost function $J(\theta)=\frac{1}{m}\sum_{i=1}^{m}\text{Cost}(x^i)$ $\text{Cost}(x)=\frac{1}{2}(h_\theta(x)-y)^2$
- Why not MSE? → not convex

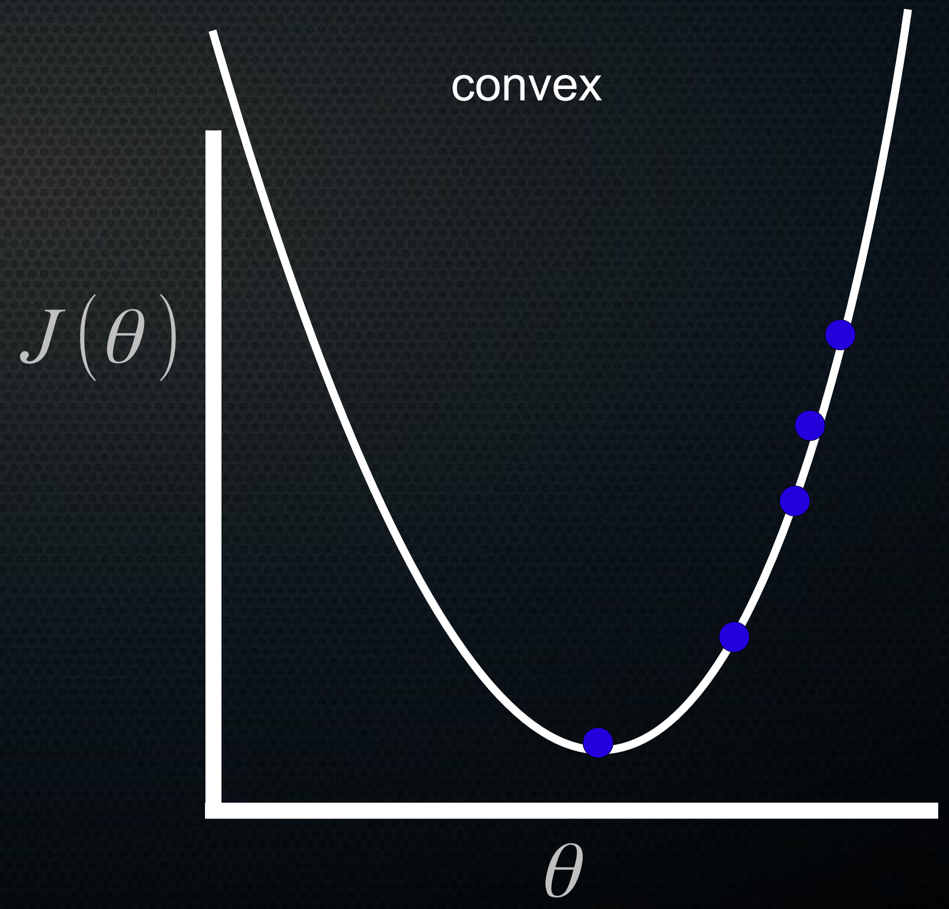




So how do we get theta's?

- Need a cost function $J(\theta)=\frac{1}{m}\sum_{i=1}^{m}\mathrm{Cost}(x^i)$ $\qquad \mathrm{Cost}(x)=\frac{1}{2}(h_\theta(x)-y)^2$
- Why not MSE? → not convex





So how do we get theta's?

- Need a cost function $J(\theta)=\frac{1}{m}\sum_{i=1}^{m}\mathrm{Cost}(x^i)$ $\cancel{\mathrm{Cost}(x)=\frac{1}{2}(h_\theta(x)-y)^2}$
- What then?

So how do we get theta's?

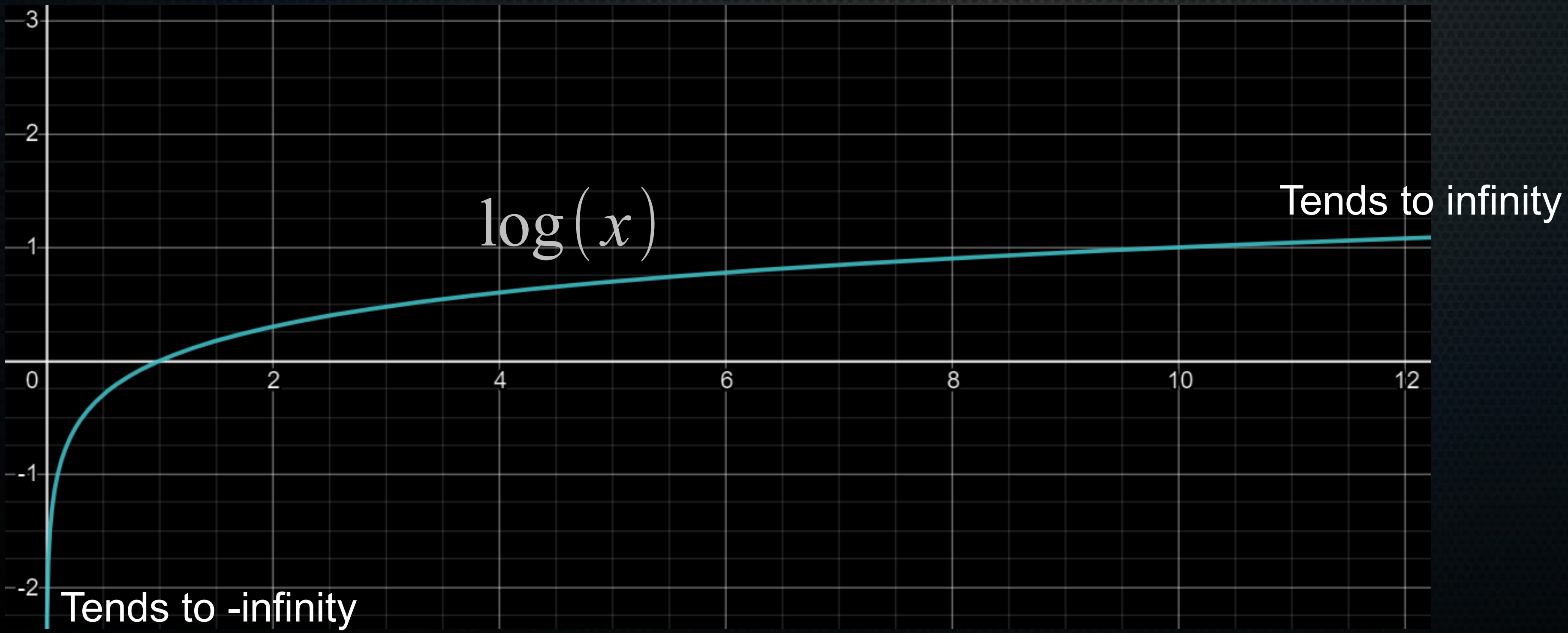
- Need a cost function $J(\theta)=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\text{Cost}(x^i)$ $\cancel{\text{Cost}(x)=\frac{1}{2}(h_\theta(x)-y)^2}$
- What then?

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

What does this function look like?

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

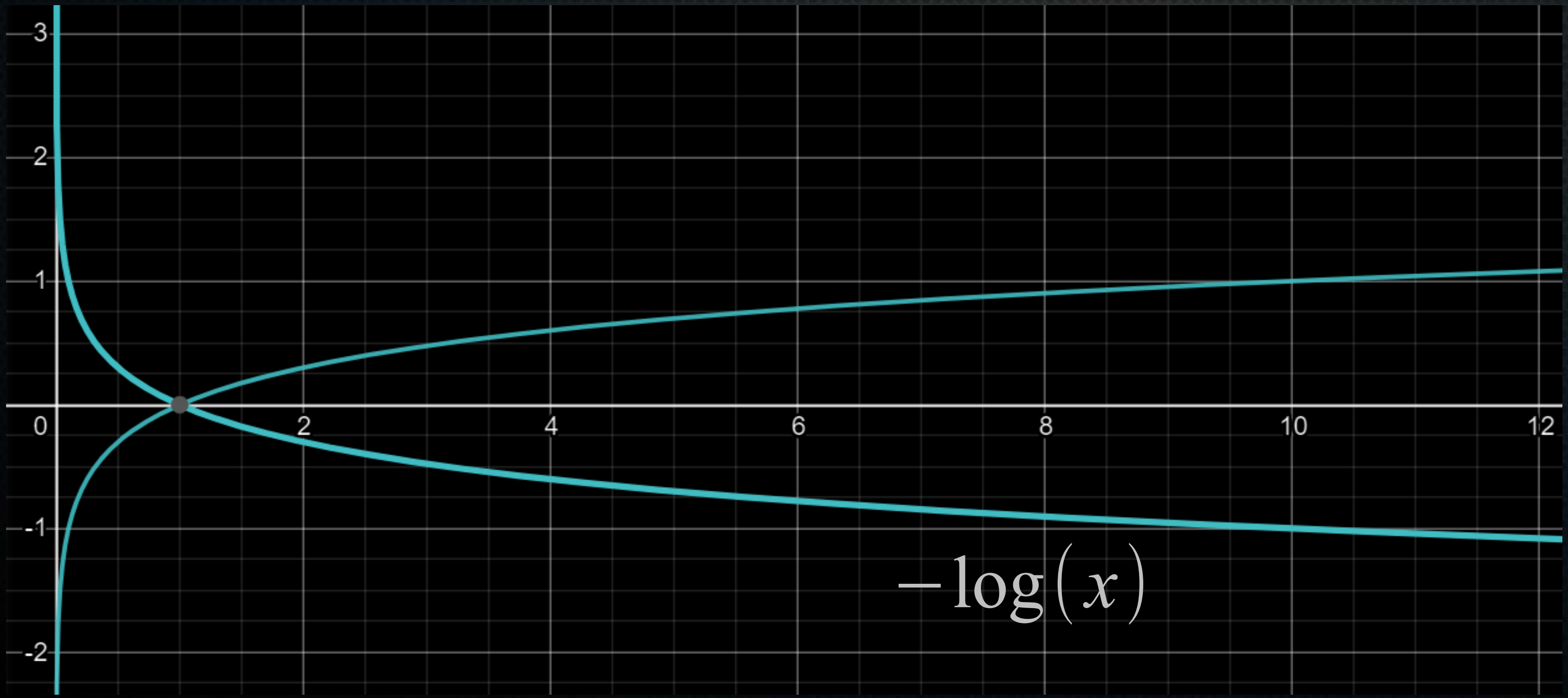
$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \text{Cost}(x^i)$$



What does this function look like?

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

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \text{Cost}(x^i)$$

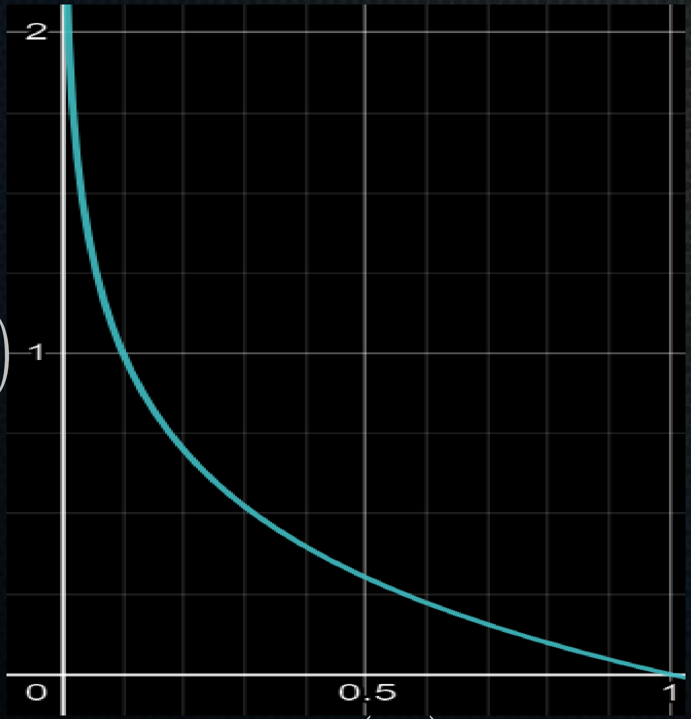


What does this function look like?

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

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

$-\log(h_\theta(x))$



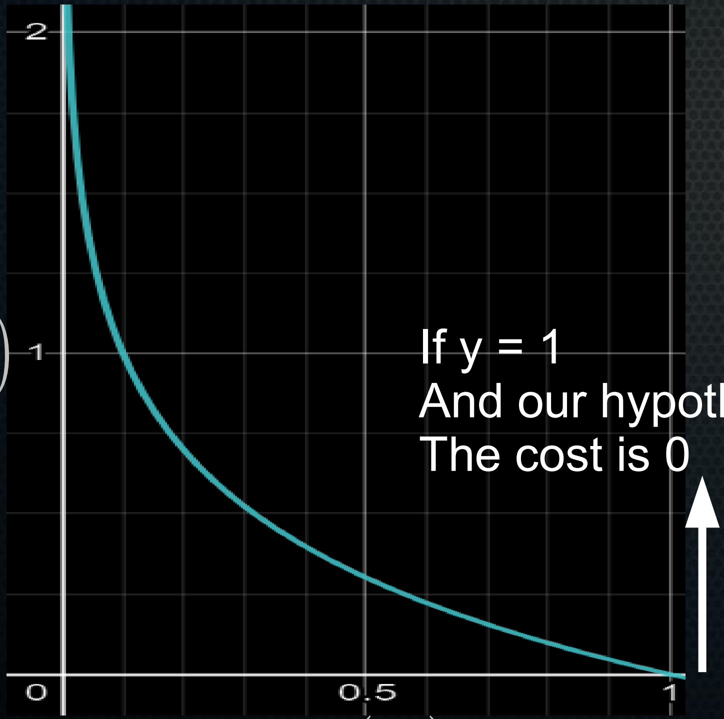
$h_\theta(x)$

What does this function look like?

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

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \text{Cost}(x^i)$$

$-\log(h_\theta(x))$



If y = 1
And our hypothesis predicts 1
The cost is 0

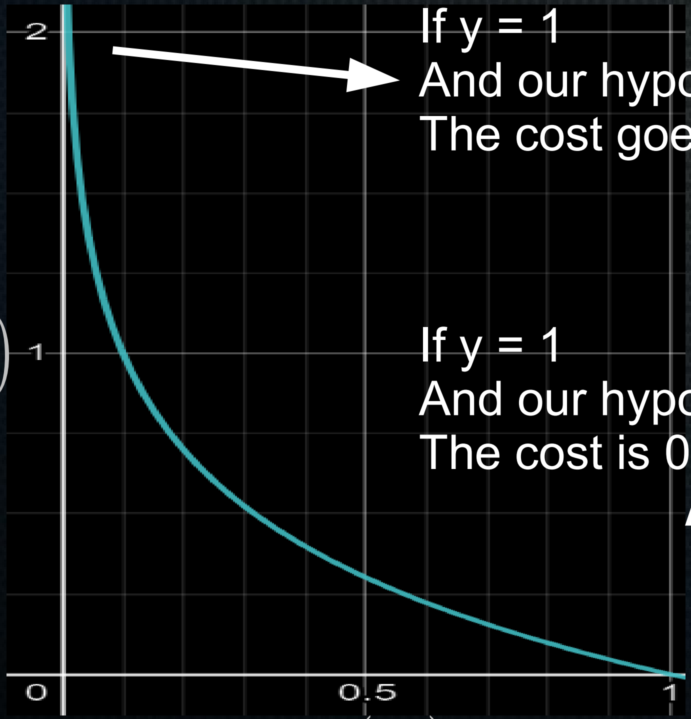
$h_\theta(x)$

What does this function look like?

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

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

$-\log(h_\theta(x))$



$h_\theta(x)$

If y = 1
And our hypothesis predicts 0
The cost goes to ∞

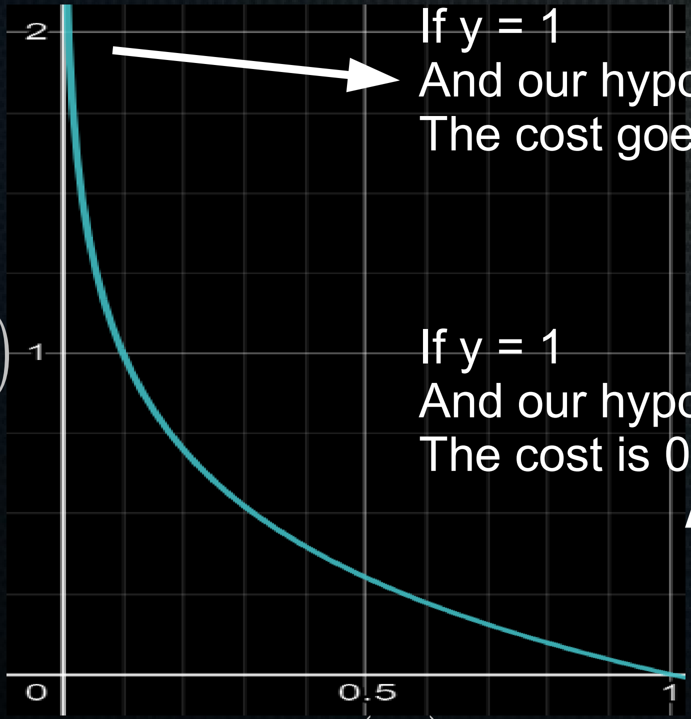
If y = 1
And our hypothesis predicts 1
The cost is 0

What does this function look like?

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

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \text{Cost}(x^i)$$

$-\log(h_\theta(x))$



$h_\theta(x)$

If y = 1
And our hypothesis predicts 0
The cost goes to ∞

If we had used MSE

$$(0-1)^2 = 1$$

Penalised far less!

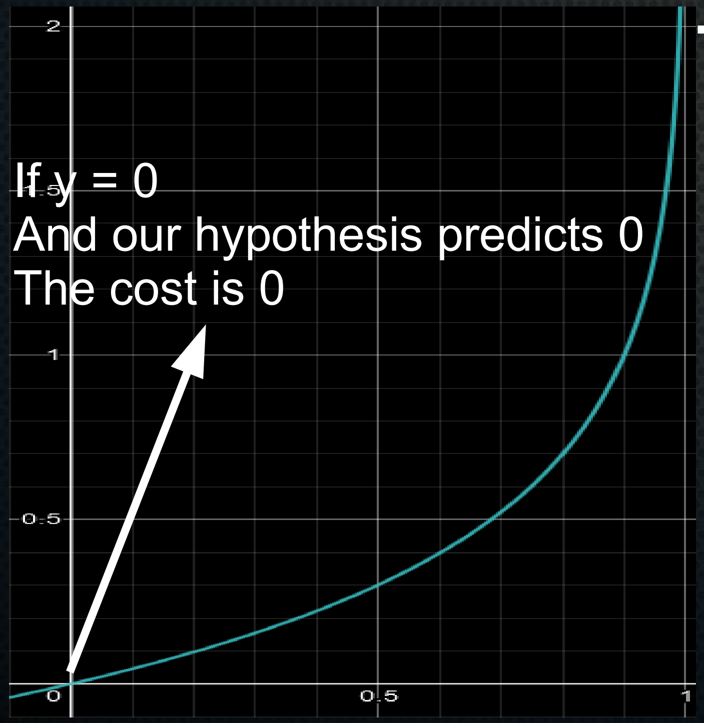
If y = 1
And our hypothesis predicts 1
The cost is 0

What does this function look like?

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

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

$-\log(1-h_\theta(x))$



$h_\theta(x)$

If y = 0
And our hypothesis predicts 1
The cost goes to ∞

Simplified notation

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

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \text{Cost}(x^i)$$

$$\text{Cost}(x) = -y \cdot \log(h_\theta(x)) - (1-y) \cdot \log(1-h_\theta(x))$$

Simplified notation

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

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \text{Cost}(x^i)$$

$$\text{Cost}(x) = -y \cdot \log(h_\theta(x)) - (1-y) \cdot \log(1-h_\theta(x))$$

$y = 0$

$$-0 \cdot \log(h_\theta(x)) - (1-0) \cdot \log(1-h_\theta(x))$$

Simplified notation

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

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \text{Cost}(x^i)$$

$$\text{Cost}(x) = -y \cdot \log(h_\theta(x)) - (1-y)\cdot\log(1-h_\theta(x))$$

$y = 0$

$$-0\cdot\log(h_\theta(x)) - (1-0)\cdot\log(1-h_\theta(x))$$

$$\boxed{-\log(1-h_\theta(x))}$$

Simplified notation

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

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

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

$y = 1$

$$-1 \cdot \log(h_\theta(x)) - (1-1) \cdot \log(1-h_\theta(x))$$

$$\boxed{-\log(h_\theta(x))}$$

Putting it all together

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

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

Putting it all together

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

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \boxed{-}\, y^{(i)} \cdot \log(h_\theta(x^{(i)})) \boxed{-} (1-y^{(i)}) \cdot \log(1-h_\theta(x^{(i)}))$$

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

Optimising the cost function

- Same form as for linear regression (only hypothesis function differs!)

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

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

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

Summary

- By using the sigmoid function as a transformation of normal regression and interpreting the output as a chance of being 0 or 1 we can do classification.
- Only the form of our hypothesis function is different
- Need a different cost function: should be smooth, and give logical values for large errors.

Break for practical