On Logistic Regression

Alexander G. Ororbia II

Introduction to Machine Learning

CSCI-335

3/18/2026

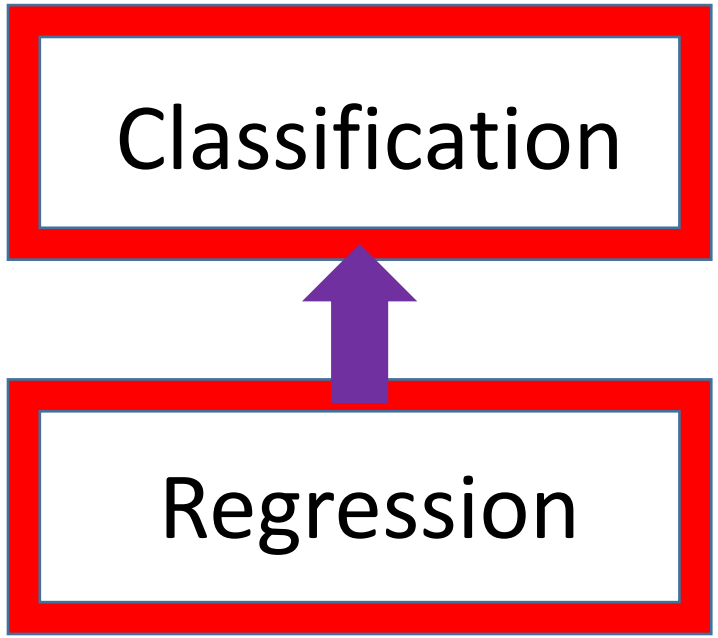
Special Thanks: Sargur N. Srihari, Andrew Ng, Carlos Guestrin

Machine Learning Algorithms

	Supervised Learning	**Unsupervised Learning**
Discrete	Classification K-NN classification	Clustering
Continuous	Regression Linear regression Polynomial regression K-NN regression	Dimensionality reduction Principal components analysis

Machine Learning Algorithms

	Supervised Learning	**Unsupervised Learning**
Discrete	Classification	Clustering
Continuous	Regression	Dimensionality reduction

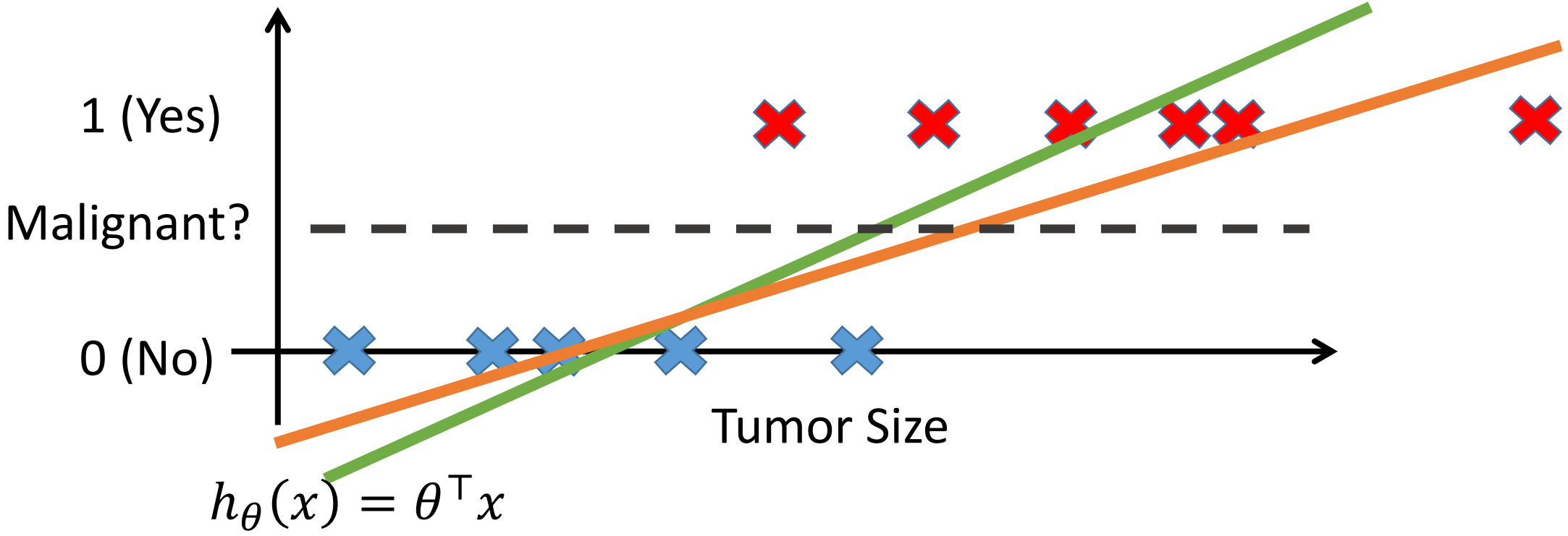


Logistic Regression

- Hypothesis *representation*
- Cost function (*evaluation*)
- Logistic regression with gradient descent (*optimization*)
- Regularization (again!)
- Multi-class classification

Logistic Regression

- **Hypothesis *representation***
- Cost function (*evaluation*)
- Logistic regression with gradient descent (*optimization*)
- Regularization (again!)
- Multi-class classification



- *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 "$y = 0$"

Why use Logistic Regression?

- There are many important research topics for which the dependent variable is "limited" in some way

- *Example*: whether or not a person smokes/drinks/skips class/takes advanced mathematics
  - For these, outcome is neither continuous nor distributed normally

- **Binary** logistic *regression* is a type of regression analysis where dependent variable is a ***dummy*** variable:
  coded 0 (*negative class*: did not smoke) or 1 (*positive class*: did smoke)

Categorization: $y = 1$ or $y = 0$

$h_\theta(x) = \theta^\top x$ (from linear regression)
can be > 1 or < 0

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

Logistic regression is actually for **classification**

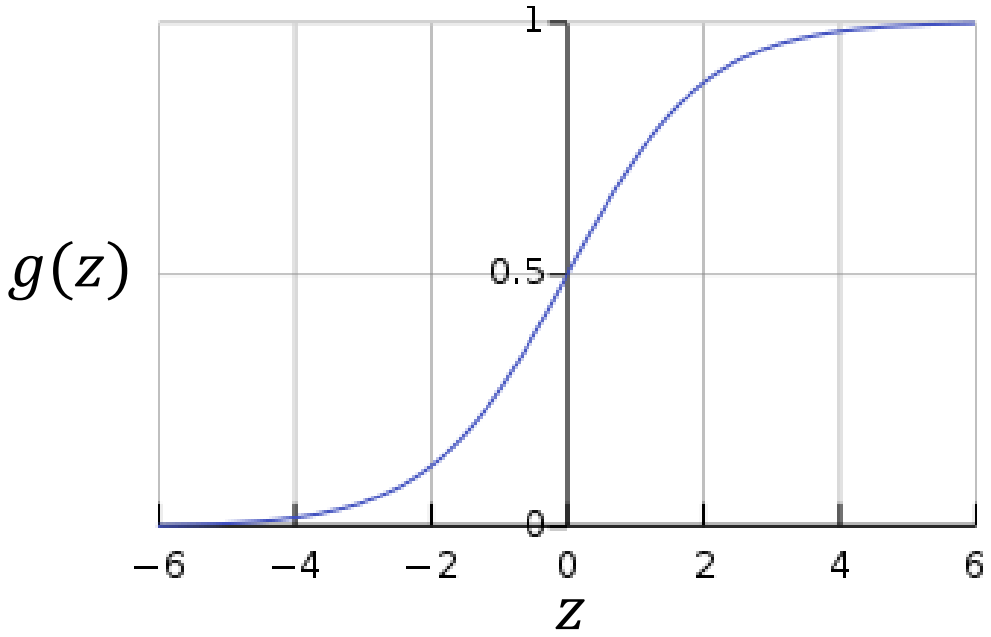
Hypothesis Representation

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

- $h_\theta(x) = g(\theta^\top x)$,

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

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



- Sigmoid function
- Logistic (link) function

Interpretation of Hypothesis Output

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

- *Example*: If $x = \begin{bmatrix} x_0 \\ \mathrm{x}_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \mathrm{tumorSize} \end{bmatrix}$
- $h_\theta(x) = 0.7$

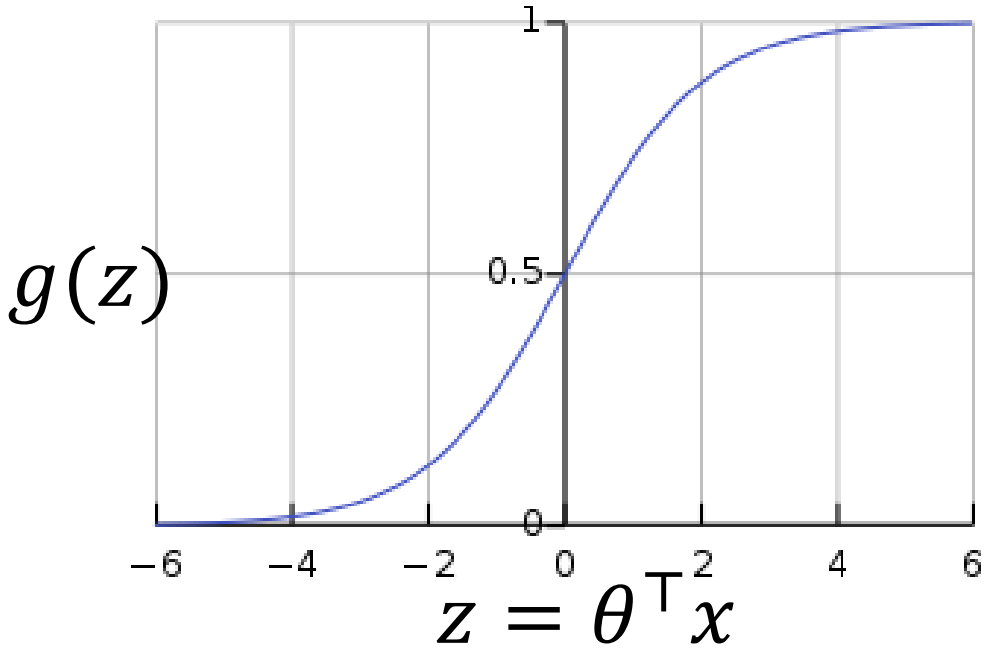
- Tell patient that 70% chance of tumor being malignant

Logistic Regression

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

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

Log-linear model



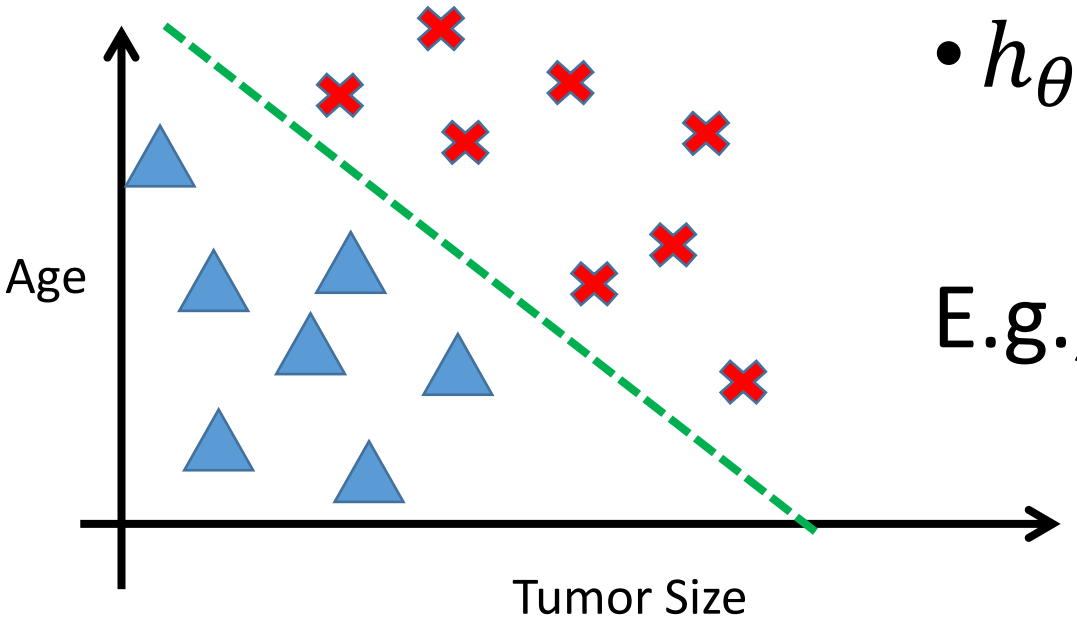
Suppose we predict "y = 1" if $h_\theta(x) \geq 0.5$

$$z = \theta^\top x \geq 0$$

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

$$z = \theta^\top x < 0$$

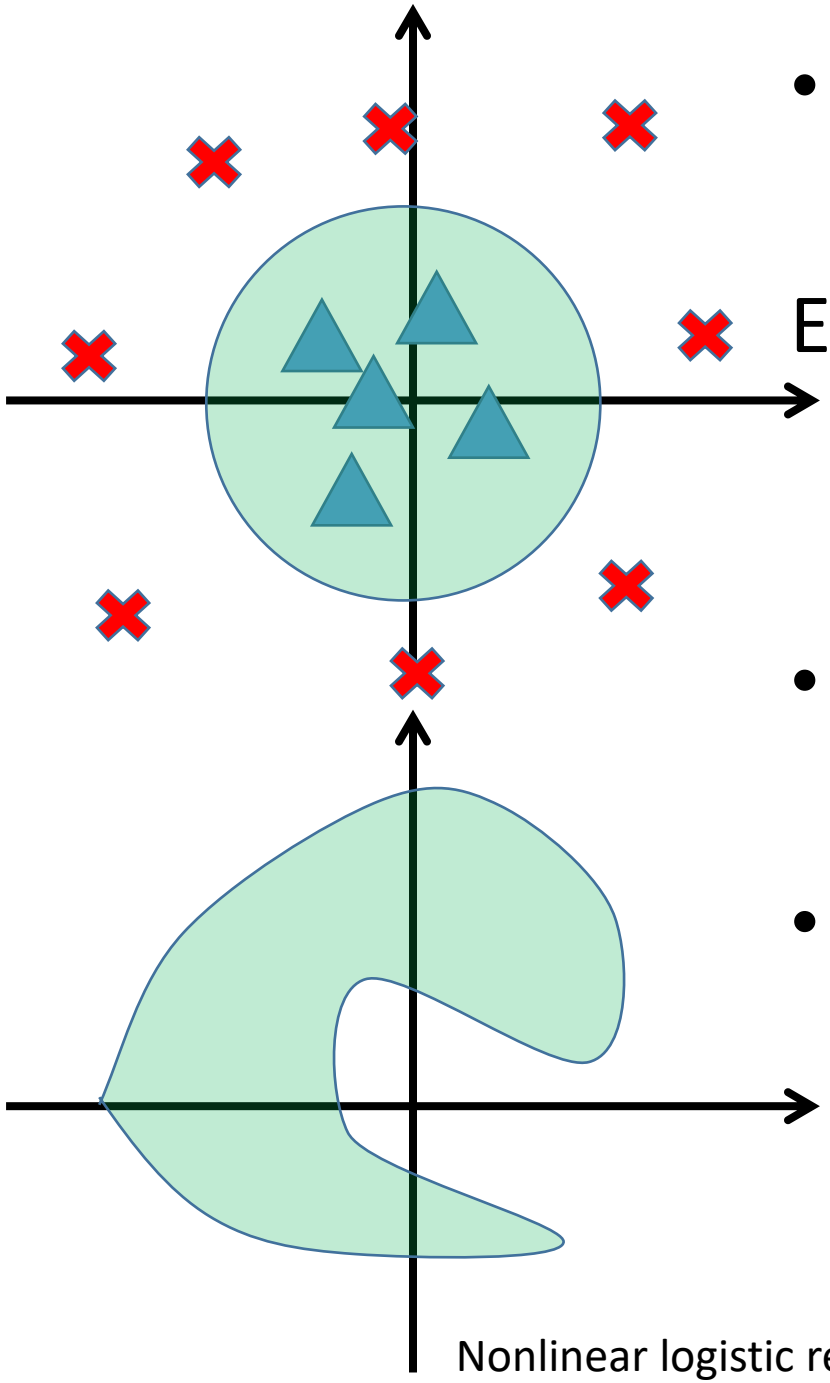
Decision Boundary



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

E.g., $\theta_0 = -3,\ \theta_1 = 1,\ \theta_2 = 1$

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



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

E.g., $\theta_0 = -1, \theta_1 = 0, \theta_2 = 0, \theta_3 = 1, \theta_4 = 1$

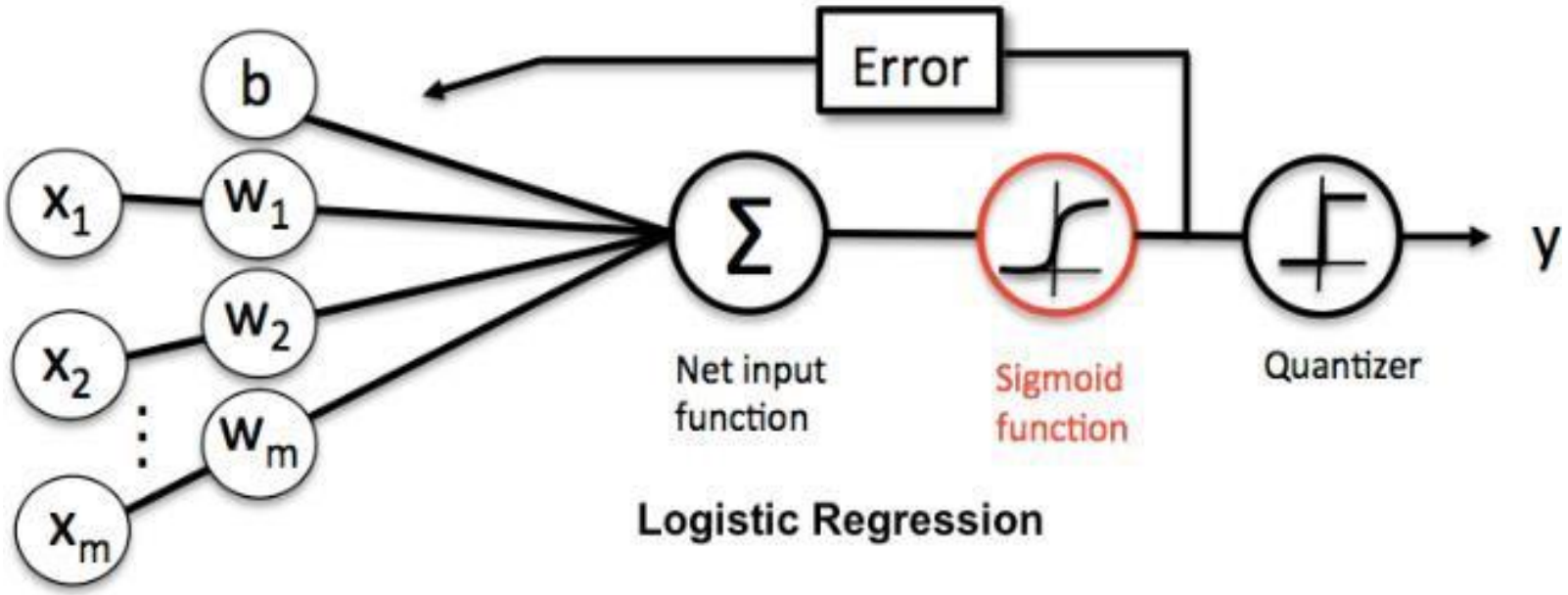
- Predict "$y = 1$" if $-1 + x_1^2 + x_2^2 \geq 0$

- $h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_1^2 x_2 + \theta_5 x_1^2 x_2^2 + \theta_6 x_1^3 x_2 + \cdots)$

A multiplicative feature basis function!

Nonlinear logistic regression

Logistic Regressor Architecture



(Representation!)

Logistic Regression

- Hypothesis *representation*
- **Cost function (*evaluation*)**
- Logistic regression with gradient descent (*optimization*)
- Regularization
- Multi-class classification

Training set with m examples

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

$$x \in \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} \qquad x_0 = 1, y \in \{0, 1\}$$

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

How to choose parameters θ?

Cost Function for Linear Regression

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

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

Cost Function for Linear Regression

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

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

Where does this cost come from?
Remember the Gaussian distribution?

Cost Function for Logistic Regression

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

- $\text{Cost}(h_\theta(x), y) = -y \log\left(\text{h}_\theta(\text{x})\right) - (1 - \text{y}) \log\left(1 - h_\theta(x)\right)$

But, where does it come from?

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

Logistic Regression

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

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

Learning: fit parameter θ

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

Prediction: given new x

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

Where Does the Cost Come From?

- Training set with m examples

$$\left\{\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \cdots, \left(x^{(m)}, y^{(m)}\right)\right\}$$

- Maximum likelihood estimate for parameter θ

$$\theta_{\text{MLE}} = \underset{\theta}{\text{argmax}}\, P_\theta\left(\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \cdots, \left(x^{(m)}, y^{(m)}\right)\right)$$

$$= \underset{\theta}{\text{argmax}} \prod_{i=1}^{m} P_\theta\left(\left(x^{(i)}, y^{(i)}\right)\right)$$

- Maximum <u>conditional</u> likelihood estimate for parameter θ

Remember the Bernoulli Distribution?

$$P(\mathrm{x} = 1) = \phi$$

$$P(\mathrm{x} = 0) = 1 - \phi$$

$$P(\mathrm{x} = x) = \phi^{x}(1 - \phi)^{1-x}$$

$$\mathbb{E}_{\mathrm{x}}[\mathrm{x}] = \phi$$

$$\mathrm{Var}_{\mathrm{x}}(\mathrm{x}) = \phi(1 - \phi)$$

Remember the Bernoulli Distribution?

$$P(\mathrm{x} = 1) = \phi$$

$$P(\mathrm{x} = 0) = 1 - \phi$$

Bernoulli probability mass function (likelihood)

$$P(\mathrm{x} = x) = \phi^{x}(1 - \phi)^{1-x}$$

$$\mathbb{E}_{\mathrm{x}}[\mathrm{x}] = \phi$$

$$\mathrm{Var}_{\mathrm{x}}(\mathrm{x}) = \phi(1 - \phi)$$

- **Goal**: choose θ to maximize conditional likelihood of training data
  - $P_\theta(Y = 1|X = x) = h_\theta(x) = \frac{1}{1+e^{-\theta^\top x}}$
  - $P_\theta(Y = 0|X = x) = 1 - h_\theta(x) = \frac{e^{-\theta^\top x}}{1+e^{-\theta^\top x}}$

- **Training data** $\mathrm{D} = \left\{\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \cdots, \left(x^{(m)}, y^{(m)}\right)\right\}$
- **Data likelihood** $= \prod_{i=1}^{m} P_\theta\left(\left(x^{(i)}, y^{(i)}\right)\right)$
- **Data conditional likelihood** $= \prod_{i=1}^{m} P_\theta\left(y^{(i)}|x^{(i)}\right)$

$$\theta_{\mathrm{MCLE}} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{m} P_\theta\left(y^{(i)}|x^{(i)}\right)$$

Expressing Conditional Log-Likelihood

$$L(\theta) = \log \prod_{i=1}^{m} P_\theta\left(y^{(i)}|x^{(i)}\right) = \sum_{i=1}^{m} \log P_\theta\left(y^{(i)}|x^{(i)}\right)$$

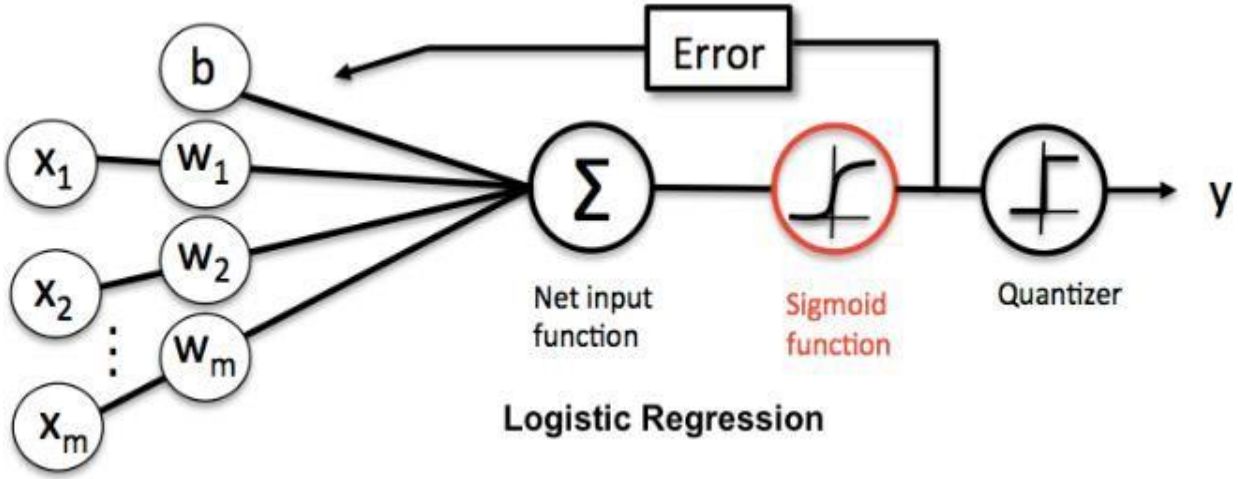
The logarithm is your *friend*!

$$= \sum_{i=1}^{m} y^{(i)} \log P_\theta\left(y^{(i)} = 1|x^{(i)}\right) + \left(1 - y^{(i)}\right) \log P_\theta\left(y^{(i)} = 0|x^{(i)}\right)$$

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

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

Logistic Regression



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

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

Bernoulli log likelihood!

Learning: fit parameter θ

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

Prediction: given new x

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

(Evaluation!)

Gradient Descent

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

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

Good news: Convex function!

Bad news: No analytical solution

Repeat {

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

}

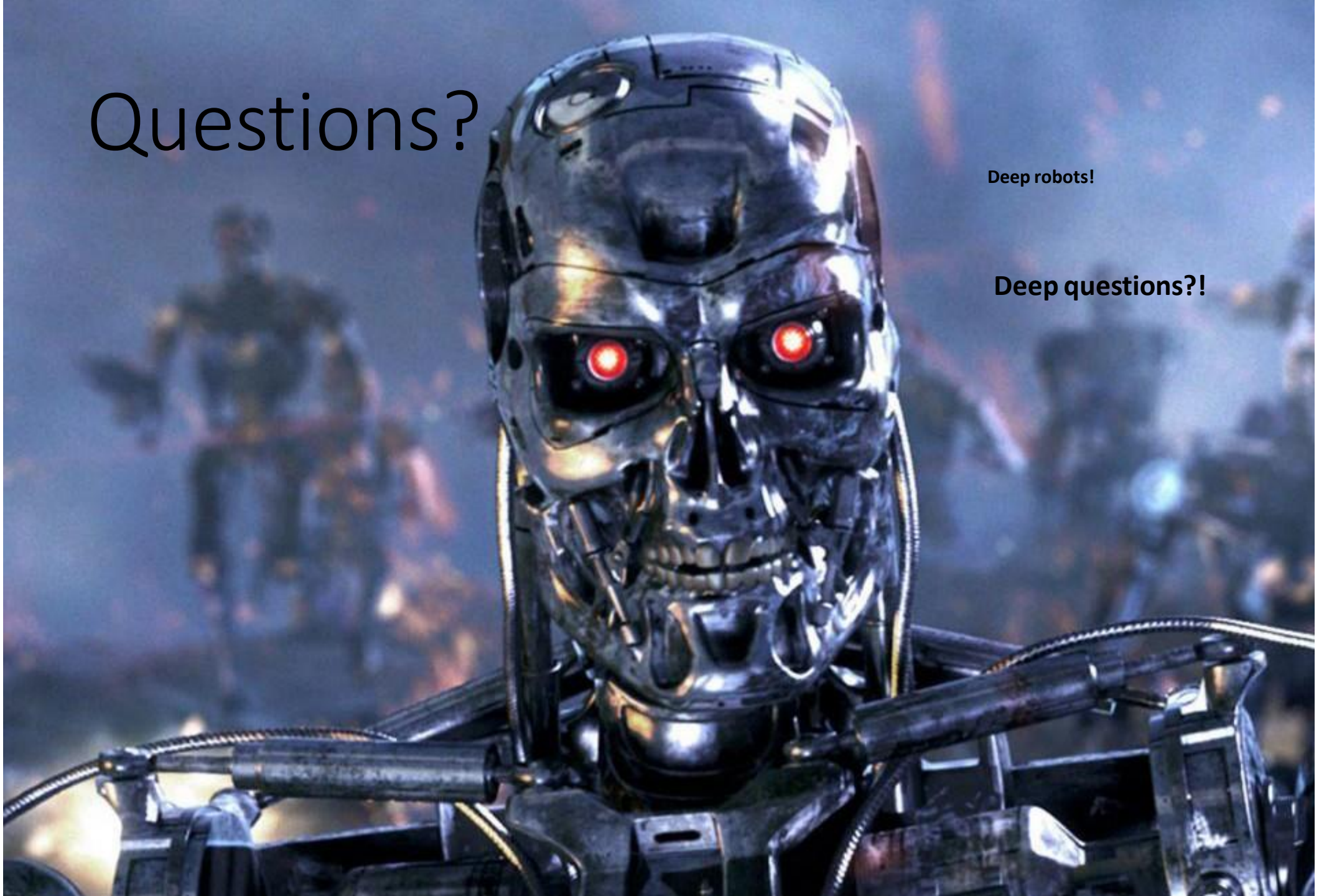
What do we still need?

Derive gradient of $J(\theta_j)$ with respect to each θ_j

Next time!

(Optimization!)

Questions?



Deep robots!

Deep questions?!