The Logistic Regressor

Alexander G. Ororbia II

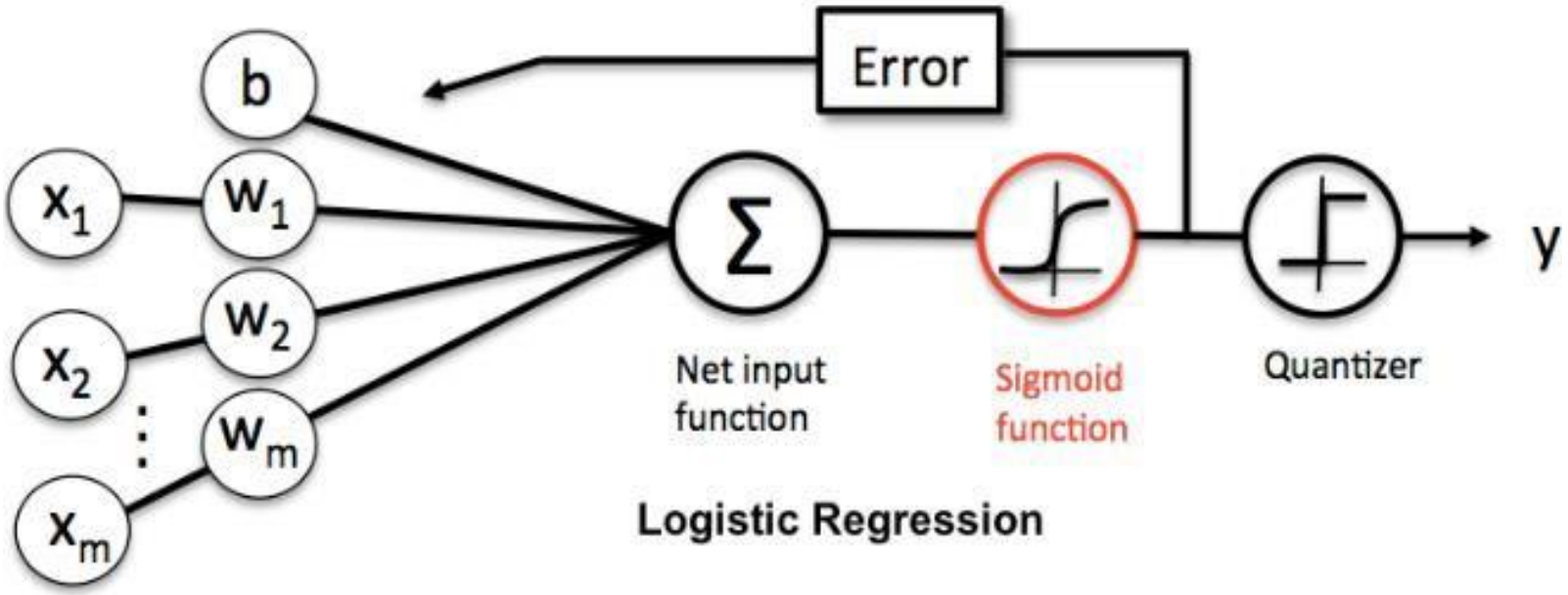
Introduction to Machine Learning

CSCI-635

10/4/2023

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

Logistic Regressor Architecture



(Representation!)

Where Does the Form Come From?

- Logistic regression hypothesis representation

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^\top x}} = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n)}}$$

- Consider learning f: $X \rightarrow Y$, where
  - X is a vector of real-valued features $[X_1, \cdots, X_n]^\top$
  - Y is Boolean
  - Assume all X_i are ***conditionally independent*** given Y
  - Model $P(X_i | Y = y_k)$ as Gaussian $N(\mu_{ik}, \sigma_i)$
  - Model $P(Y)$ as Bernoulli π

What is $P(Y | X_1, X_2, \cdots, X_n)$?

(Representation!)


WHITE BOARD TIME!

Logistic Regression

- Hypothesis representation
- **Cost function**
- Logistic regression with gradient descent
- 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**

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

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

Logistic Regression Cost Function

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

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

- 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 Cost Function

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

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

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?

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

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

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

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

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

- **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

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

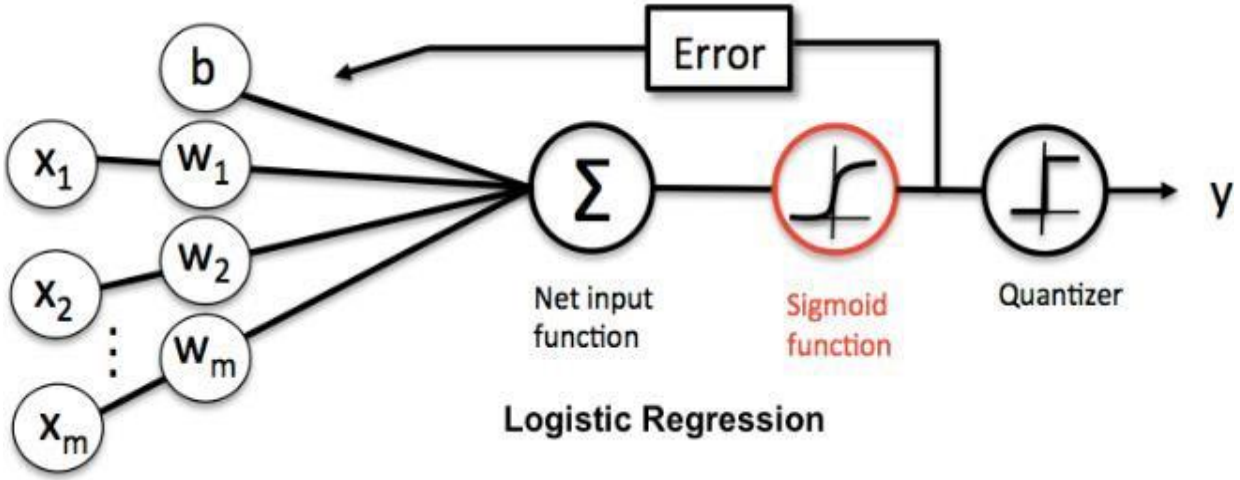
The logarithm is your *friend*!

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

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

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

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 {

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

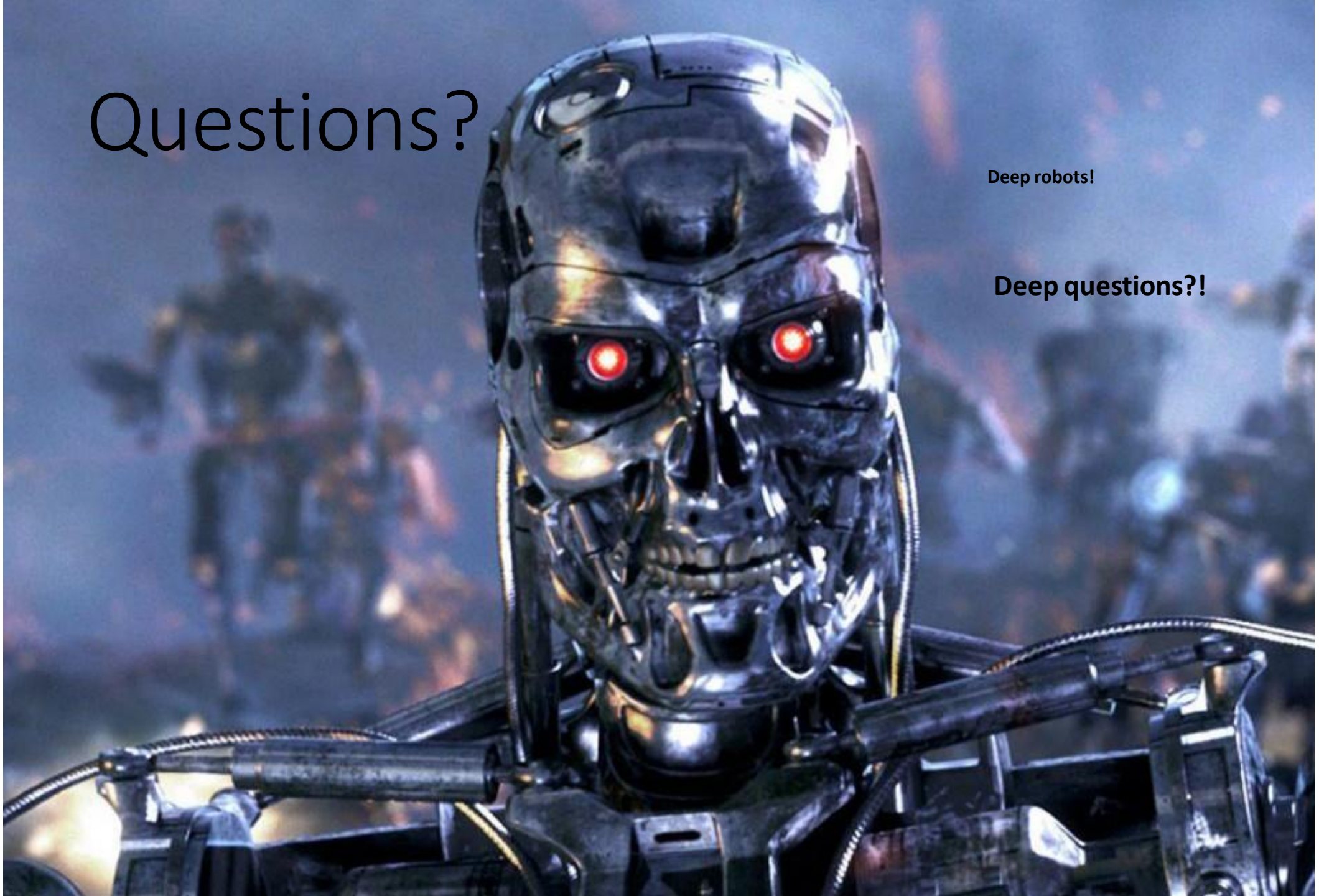
}

What do we still need?

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

(Optimization!)

Questions?



Deep robots!

Deep questions?!