

# Logistic Regression

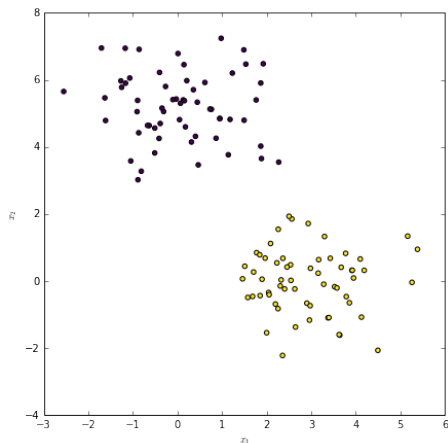
Advanced Topics in High-Performance Computing

Faisal Qureshi



# Logistic regression

Logistic regression is a binary classifier



In binary classification, the target variable  $y$  takes on values in  $\{0, 1\}$

## Bernoulli distribution

A Bernoulli random variable  $X$  takes values in  $\{0, 1\}$

$$\begin{aligned}\Pr(X|\theta) &= \begin{cases} \theta & \text{if } X = 1 \\ 1 - \theta & \text{otherwise} \end{cases} \\ &= \theta^X (1 - \theta)^{1-X}\end{aligned}$$

### Example usage

Bernoulli distribution  $\text{Ber}(X|\theta)$  can be used to model coin tosses.

## Binary classification

The goal of binary classification is to learn  $h_{\theta}(\mathbf{x})$ , which can be used to assign a label  $y \in \{0, 1\}$  to the input  $\mathbf{x}$ . Label  $y$  takes values in  $\{0, 1\}$ , so we can use Bernoulli distribution to specify its probability distribution. Specifically

$$\Pr(y = 1) = h_{\theta}(\mathbf{x})$$

$$\Pr(y = 0) = 1 - h_{\theta}(\mathbf{x})$$

Or more succinctly

$$\Pr(y) = h_{\theta}(\mathbf{x})^y (1 - h_{\theta}(\mathbf{x}))^{1-y}$$

## Likelihood for binary classification

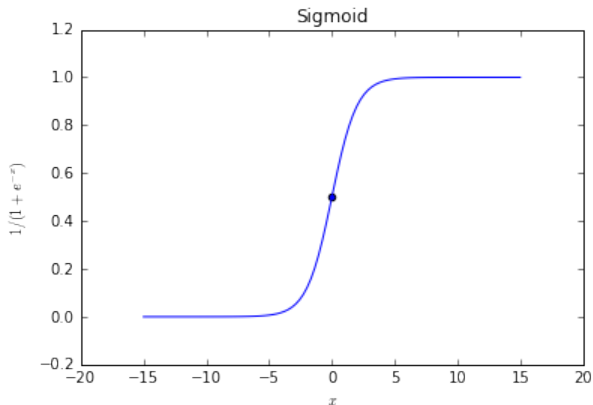
Under the assumption that data is i.i.d.

$$\Pr(y|\mathbf{X}, \theta) = \prod_{i=1}^N h_{\theta}(\mathbf{x}^{(i)})^{y^{(i)}} \left(1 - h_{\theta}(\mathbf{x}^{(i)})\right)^{1-y^{(i)}}$$

# Sigmoid function

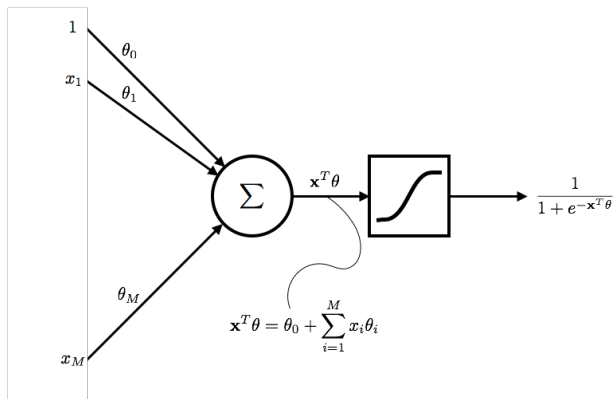
$\text{sigm}(x)$  refers to a *sigmoid* function, also known as the *logistic* or *logit* function.

$$\text{sigm}(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}$$



# Artificial Neuron

Perceptron with Sigmoid activation function



# Logistic regression

For logistic regression, we set  $h_{\theta}(\mathbf{x}) = \text{sigm}(\mathbf{x}^T \theta)$ . So

$$\Pr(y|\mathbf{X}, \theta) = \prod_{i=1}^N \left[ \frac{1}{1 + e^{-\mathbf{x}^{(i)T} \theta}} \right]^{y^{(i)}} \left[ 1 - \frac{1}{1 + e^{-\mathbf{x}^{(i)T} \theta}} \right]^{1-y^{(i)}}$$

where

$$\mathbf{x}^T \theta = \theta_0 + \sum_{i=1}^M \theta_i \mathbf{x}_i$$



# MLE for logistic regression

## Likelihood

$$L(\theta) = \Pr(y|\mathbf{X}, \theta)$$

## Negative log-likelihood

$$\begin{aligned} l(\theta) &= -\log L(\theta) \\ &= -\sum_{i=1}^N y^{(i)} \log h_{\theta}(\mathbf{x}^{(i)}) + (1 - y^{(i)}) \log(1 - h_{\theta}(\mathbf{x}^{(i)})) \end{aligned}$$

## Goal

Our goal is to find parameters  $\theta$  that maximize the likelihood (or **minimize the negative log-likelihood**).

$$\theta^* = \arg \min_{\theta} l(\theta)$$

*We prefer to work in the log domain for mathematical convenience. There are also numerical advantages of working in the log domain.*

## Derivative of a sigmoid

$$\begin{aligned}\frac{d}{dx} \text{sigm}(x) &= \frac{d}{dx} \frac{1}{1 + e^{-x}} \\ &= \frac{-(-1)e^{-x}}{(1 + e^{-x})^2} \\ &= \left( \frac{e^{-x}}{1 + e^{-x}} \right) \left( \frac{1}{1 + e^{-x}} \right) \\ &= \left( \frac{1 - 1 + e^{-x}}{1 + e^{-x}} \right) \left( \frac{1}{1 + e^{-x}} \right) \\ &= \left( 1 - \frac{1}{1 + e^{-x}} \right) \left( \frac{1}{1 + e^{-x}} \right) \\ &= (1 - \text{sigm}(x)) \text{sigm}(x)\end{aligned}$$

### Gradient

$$\frac{d}{d\theta} \text{sigm}(\mathbf{x}^T \theta) = (1 - \text{sigm}(x)) \text{sigm}(x) \mathbf{x}$$

# MLE for logistic regression

Gradient of  $l(\theta)$  for  $i$ th example

$$\nabla l(\theta) = -\mathbf{x}^{(i)}(y^{(i)} - h_{\theta}(\mathbf{x}^{(i)}))$$

Stochastic gradient descent rule:

$$\begin{aligned}\theta^{(k+1)} &= \theta^{(k)} - \eta \nabla l(\theta) \\ &= \theta^{(k)} + \eta \mathbf{x}^{(i)}(y^{(i)} - h_{\theta}(\mathbf{x}^{(i)}))\end{aligned}$$

# Logistic regression for binary classification

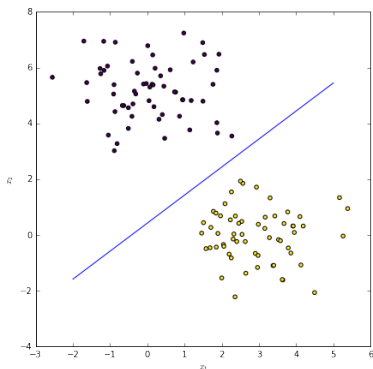
Given a point  $\mathbf{x}^{(*)}$ , classify using the following rule

$$y^{(*)} = \begin{cases} 1 & \text{if } \Pr(y|\mathbf{x}^{(*)}, \theta) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

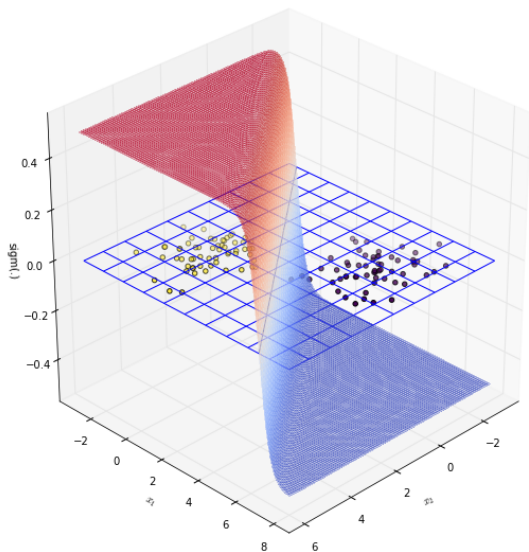
The decision boundary

is  $\mathbf{x}^T \theta = 0$ .

Recall that this is where the sigmoid function is 0.5.



# Logistic regression for binary classification



# Summary

- ▶ We looked at logistic regression, a binary classifier.
- ▶ Bernoulli distribution