Logistic Regression

Instructor: Jessica Wu -- Harvey Mudd College

The instructor gratefully acknowledges Andrew Ng (Stanford), Eric Eaton (UPenn), David Kauchak (Pomona), and the many others who made their course materials freely available online.

Logistic Regression Setup

Learning Goals

- Describe the logistic regression model
- Describe how to interpret a prediction under LogReg
- Describe the decision boundary for LogReg

Logistic Regression Setup

binary classification $y \in \{0, 1\}$

Instead of predicting class, give probability of instance being that class:

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = p(y = 1 \mid \boldsymbol{x}; \boldsymbol{\theta})$$

Why not just use linear regression with a threshold?

misnomer: logistic regression is a classification model!

Based on slide by Eric Eaton

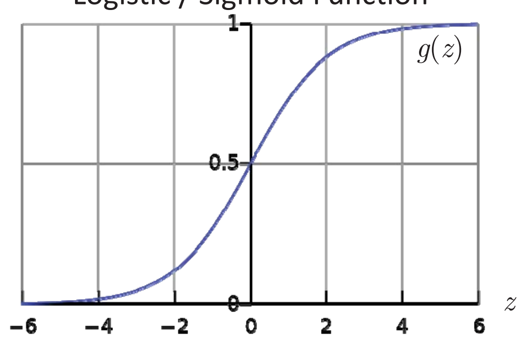
Logistic Regression Model

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = g(\boldsymbol{\theta}^T \boldsymbol{x})$$

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

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

Logistic / Sigmoid Function



as $z \to -\infty$, $g(z) \to 0$

$0 \leq g(z) \leq 1$

as $z \to \infty$, $g(z) \to 1$

for negative (y=0) instances, $\boldsymbol{\theta}^T \boldsymbol{x}$ should be large negative

for positive (y=1) instances, $\boldsymbol{\theta}^T \boldsymbol{x}$ should be large positive

Predict

- $y = 1$ if $h_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq 0.5$
- $y = 0$ if $h_{\boldsymbol{\theta}}(\boldsymbol{x}) < 0.5$

Based on slide by Eric Eaton

Interpretation of Hypothesis Output

Example: cancer diagnosis from tumor size

$$\boldsymbol{x} = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{tumor size} \end{bmatrix}$$

$y = 0$ benign tumor

$y = 1$ malignant tumor

You find $h_{\boldsymbol{\theta}}(\boldsymbol{x}) = 0.7$. What does this mean?

Based on slide by Eric Eaton
[example by Andrew Ng]

Another Interpretation

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

$$p(y = 0|\boldsymbol{x};\boldsymbol{\theta}) = 1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{e^{-\boldsymbol{\theta}^T\boldsymbol{x}}}{1+e^{-\boldsymbol{\theta}^T\boldsymbol{x}}}$$

$$\underbrace{\log \underbrace{\frac{p(y = 1|\boldsymbol{x};\boldsymbol{\theta})}{p(y = 0|\boldsymbol{x};\boldsymbol{\theta})}}_{\text{odds of 1}}}_{\text{log odds (logit) of 1}} = \log \frac{1}{e^{-\boldsymbol{\theta}^T\boldsymbol{x}}} = -\log e^{-\boldsymbol{\theta}^T\boldsymbol{x}} = \boldsymbol{\theta}^T\boldsymbol{x}$$

logistic regression assumes log odds is a linear function of x

Note: The **odds** in favor of an event is the quantity $p/(1-p)$, where p is the probability of the event.

e.g. If I toss a fair dice, what are the odds of a 6? $(1/6)\ /\ (5/6) = 1/5$

Based on slide by Eric Eaton
[originally by Xiaoli Fern]

Decision Boundary

What does the decision boundary of LogReg look like?

$$\log \frac{p(y=1|\boldsymbol{x};\boldsymbol{\theta})}{p(y=0|\boldsymbol{x};\boldsymbol{\theta})} = \log 1 = 0 = \boldsymbol{\theta}^T \boldsymbol{x}$$

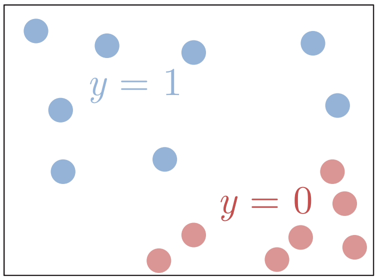


Figure by Eric Eaton

(This slide intentionally left blank.)

Solving Logistic Regression

Learning Goals

- Describe the optimization function $J(\boldsymbol{\theta})$ for LogReg (including the underlying probabilistic model)
- Describe how to optimize $\boldsymbol{\theta}$ using gradient descent

Cost Function

Can we use squared loss to find optimal $\boldsymbol{\theta}$?

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

Based on slide by Eric Eaton

Probabilistic Model

Given $\left\{\left(\boldsymbol{x}^{(i)}, y^{(i)}\right)\right\}_{i=1}^{n}$

$\boldsymbol{x}^{(i)} \in \mathbb{R}^d$

$y^{(i)} \in \{0, 1\}$

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

MLE and Cost Function

Intuition behind the Objective

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

rewrite objection function as

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{n} \text{cost} \left(h_{\boldsymbol{\theta}} \left(\boldsymbol{x}^{(i)} \right), y^{(i)} \right)$$

recall for linear regression

$$\text{cost} \left(h_{\boldsymbol{\theta}}(\boldsymbol{x}), y \right) = \left(h_{\boldsymbol{\theta}}(\boldsymbol{x}) - y \right)^2$$

cost of a single instance

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

aside: If $y \in \{-1, +1\}$ rather than $y \in \{0, +1\}$, then common to use

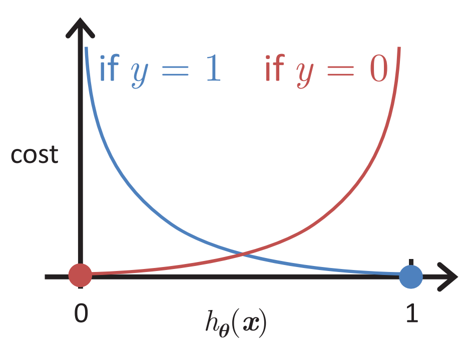
$$p(y | \boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{1 + e^{-y \boldsymbol{\theta}^T \boldsymbol{x}}} \qquad \text{cost} \left(h_{\boldsymbol{\theta}}(\boldsymbol{x}), y \right) = \log \left(1 + e^{-y h_{\boldsymbol{\theta}}(\boldsymbol{x})} \right)$$

logistic loss

Based on slide by Eric Eaton

Intuition behind the Objective

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



if $y = 1$

- $\text{cost} = 0$ if $h_\theta(\boldsymbol{x}) = 1$
- as $h_\theta(\boldsymbol{x}) \rightarrow 0$, $\text{cost} \rightarrow \infty$

if $y = 0$

- $\text{cost} = 0$ if $h_\theta(\boldsymbol{x}) = 0$
- as $h_\theta(\boldsymbol{x}) \rightarrow 1$, $\text{cost} \rightarrow \infty$

captures intuition that larger mistakes should get larger penalties

Based on slide by Eric Eaton
[example by Andrew Ng]

Gradient Descent

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\boldsymbol{\theta})$$

useful property of sigmoid $g(z) = \frac{1}{1+e^{-z}}$

Gradient Descent

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\boldsymbol{\theta})$$

For one training example $(\boldsymbol{x},\ y)$

Gradient Descent

Stochastic Gradient Descent

$$\theta_j \leftarrow \theta_j - \alpha \left(h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) - y^{(i)} \right) x_j^{(i)}$$

This looks <u>identical</u> to linear regression!
But the underlying model is different.

linear regression $h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^T \boldsymbol{x}$

logistic regression $h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \dfrac{1}{1 + e^{-\boldsymbol{\theta}^T \boldsymbol{x}}}$

Aside

Why do linear regression and logistic regression have the same update rule?

linear regression $y \mid \boldsymbol{x}; \boldsymbol{\theta} \sim N(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}, \sigma^2)$

logistic regression $y \mid \boldsymbol{x}; \boldsymbol{\theta} \sim \mathrm{Bernoulli}(g(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}))$

Both $p(y \mid \boldsymbol{x}; \boldsymbol{\theta})$ for belong to the **exponential family** of distributions.

Both regression models are **generalized linear models (GLMs)**.

See Andrew Ng's notes.

Summary

- Logistic regression is a linear classifier (of log odds ratio)
- Logistic regression uses a logistic loss function
- We can apply most linear regression tools
  - probabilistic interpretation
  - gradient descent
  - basis functions
  - regularization (in practice, you need to regularize since $l(\boldsymbol{\theta})$ tends to overfit)
- Homework
  - show that $J(\boldsymbol{\theta})$ is convex so GD gives <u>global</u> minimum

(This slide intentionally left blank.)

(This slide intentionally left blank.)

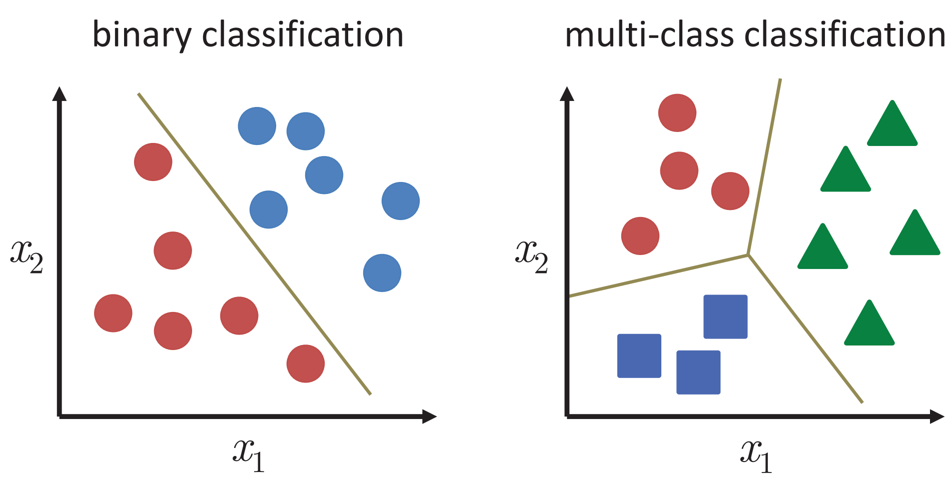
(extra slides)

Multi-Class Logistic Regression

Learning Goals

- Describe how to extend logistic regression to multiple classes

Multi-Class Classification



disease diagnosis: healthy / cold / flu / pneumonia

object classification: desk / chair / monitor / bookcase

Based on slide by Eric Eaton

Multi-Class Logistic Regression

For 2 classes

$$\underbrace{h_{\boldsymbol{\theta}}(\boldsymbol{x})}_{p(y=1|\boldsymbol{x};\boldsymbol{\theta})} = \frac{1}{1+\exp\left(-\boldsymbol{\theta}^T\boldsymbol{x}\right)} = \frac{\exp\left(\boldsymbol{\theta}^T\boldsymbol{x}\right)}{\underbrace{\exp\left(\boldsymbol{\theta}^T\boldsymbol{x}\right)}_{\text{weight assigned to } y=1} + \underbrace{1}_{\text{weight assigned to } y=0}}$$

logistic function

$y \mid \boldsymbol{x};\, \boldsymbol{\theta} \sim \mathrm{Bernoulli}(\phi)$ where $\phi = h_{\boldsymbol{\theta}}(\boldsymbol{x})$

For k classes

$$p(y=c|\boldsymbol{x};\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\ldots\boldsymbol{\theta}_k) = \frac{\exp\left(\boldsymbol{\theta}_c^T\boldsymbol{x}\right)}{\sum_{j=1}^{k} \underbrace{\exp\left(\boldsymbol{\theta}_j^T\boldsymbol{x}\right)}_{\text{weight assigned to class } y=j}}$$

softmax function

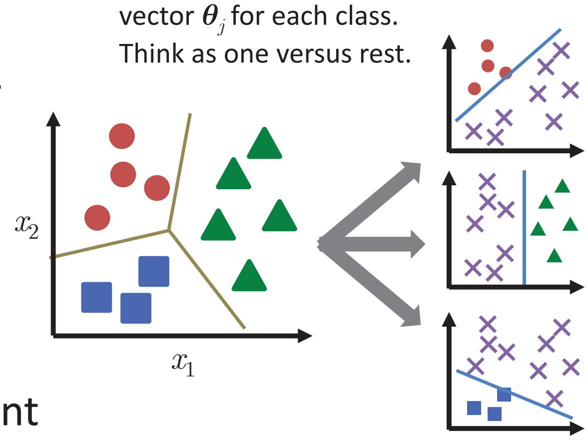
$y \mid \boldsymbol{x};\, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k \sim \mathrm{Multinomial}(\phi_1, \ldots, \phi_k)$
where $\phi_j = h_{\boldsymbol{\theta}_j}(\boldsymbol{x})$

Implementing Multi-Class Logistic Regression

Model for class c

$$h_{\boldsymbol{\theta}_c}(\boldsymbol{x}) = \frac{\exp\left(\boldsymbol{\theta}_c^T\boldsymbol{x}\right)}{\sum_{j=1}^{k}\exp\left(\boldsymbol{\theta}_j^T\boldsymbol{x}\right)}$$

Maintain separate weight vector $\boldsymbol{\theta}_j$ for each class.
Think as one versus rest.



Train using gradient descent

- simultaneously update all parameters for all models
- same update step, just with above hypothesis

Predict most probable class $\arg\max_c h_{\boldsymbol{\theta}_c}(\boldsymbol{x})$

Based on slide by Eric Eaton