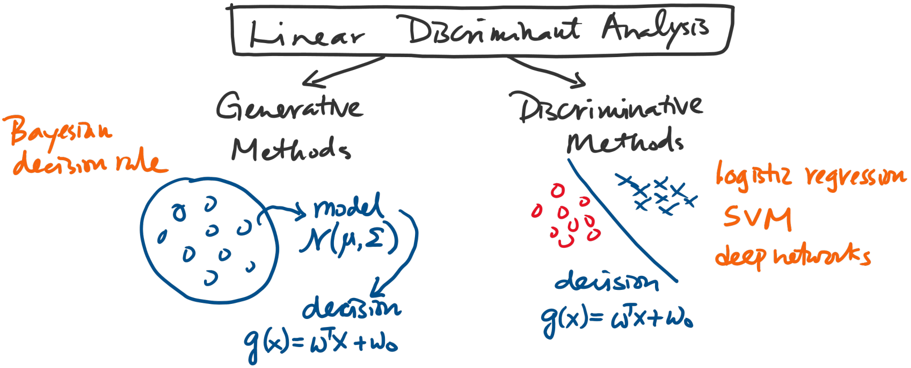
ECE595 / STAT598: Machine Learning I
Lecture 14 Logistic Regression

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

Overview



- In linear discriminant analysis (LDA), there are generally two types of approaches
- **Generative approach**: Estimate model, then define the classifier
- **Discriminative approach**: Directly define the classifier

Outline

Discriminative Approaches

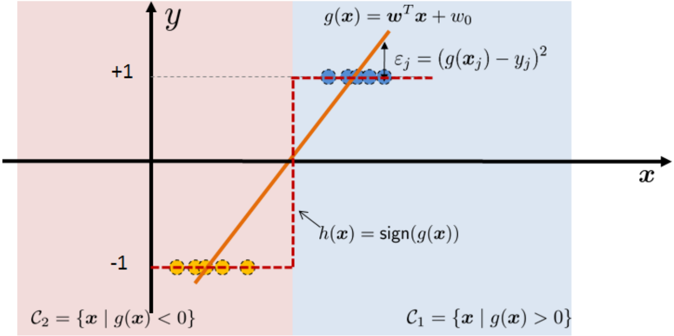
- Lecture 14 Logistic Regression 1
- Lecture 15 Logistic Regression 2

This lecture: Logistic Regression 1

- From Linear to Logistic
  - Motivation
  - Loss Function
  - Why not L2 Loss?
- Interpreting Logistic
  - Maximum Likelihood
  - Log-odd
- Convexity
  - Is logistic loss convex?
  - Computation

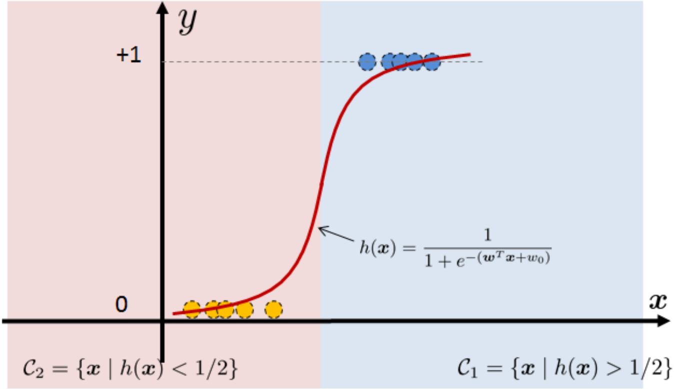
Geometry of Linear Regression

- The discriminant function $g(\boldsymbol{x})$ is linear
- The hypothesis function $h(\boldsymbol{x}) = \mathrm{sign}(g(\boldsymbol{x}))$ is a unit step



From Linear to Logistic Regression

- Can we replace $g(\boldsymbol{x})$ by $\text{sign}(g(\boldsymbol{x}))$?
- How about a soft-version of $\text{sign}(g(\boldsymbol{x}))$?
- This gives a logistic regression.



Sigmoid Function

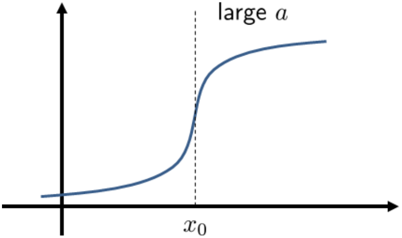
- The function

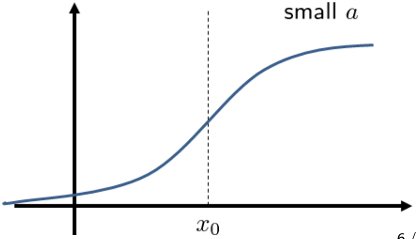
$$h(\boldsymbol{x}) = \frac{1}{1 + e^{-g(\boldsymbol{x})}} = \frac{1}{1 + e^{-(\boldsymbol{w}^T\boldsymbol{x} + w_0)}}$$

  is called a **sigmoid function**.
- Its 1D form is

$$h(x) = \frac{1}{1 + e^{-a(x - x_0)}}, \qquad \text{for some } a \text{ and } x_0,$$

- a controls the transient speed
- x_0 controls the cutoff location



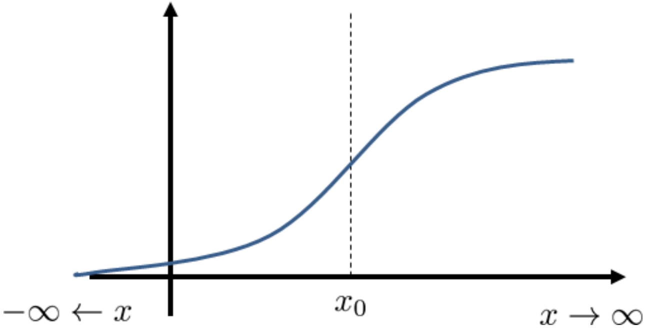


Sigmoid Function

- Note that

$$h(x) \to 1, \quad \text{as} \quad x \to \infty,$$
$$h(x) \to 0, \quad \text{as} \quad x \to -\infty,$$

- So $h(x)$ can be regarded as a "probability".



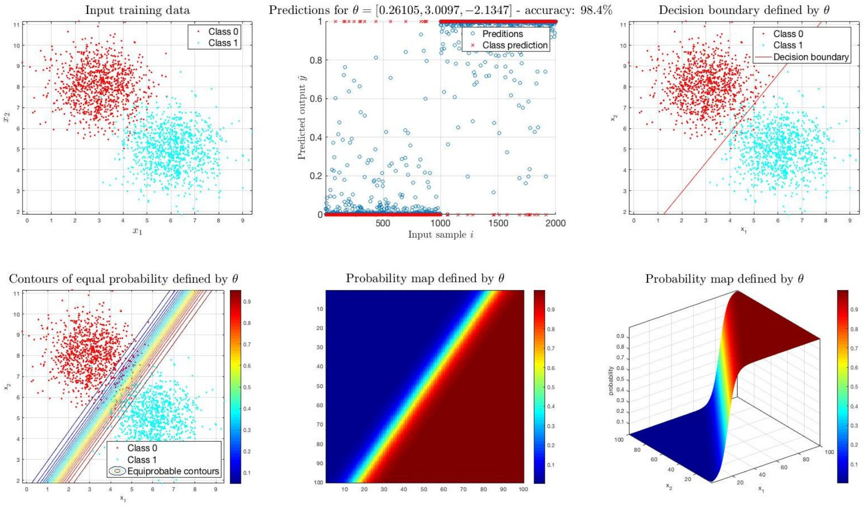
Sigmoid Function

- Derivative is

$$
\begin{aligned}
\frac{d}{dx}\left(\frac{1}{1+e^{-a(x-x_0)}}\right) &= -\left(1+e^{-a(x-x_0)}\right)^{-2}\left(e^{-a(x-x_0)}\right)(-a) \\
&= a\left(\frac{e^{-a(x-x_0)}}{1+e^{-a(x-x_0)}}\right)\left(\frac{1}{1+e^{-a(x-x_0)}}\right) \\
&= a\left(1-\frac{1}{1+e^{-a(x-x_0)}}\right)\left(\frac{1}{1+e^{-a(x-x_0)}}\right) \\
&= a[1-h(x)][h(x)].
\end{aligned}
$$

- Since $0 < h(x) < 0$, we have $0 < 1 - h(x) < 1$.
- Therefore, the derivative is always positive.
- So h is an increasing function.
- Hence h can be considered as a "CDF".

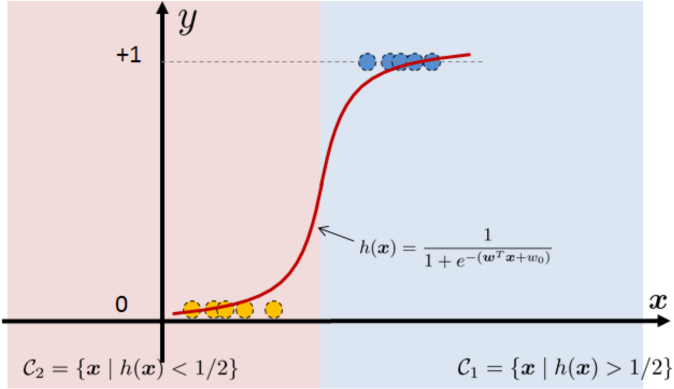
Sigmoid Function



http://georgepavlides.info/wp-content/uploads/2018/02/logistic-binary-e1517639495140.jpg

From Linear to Logistic Regression

- Can we replace $g(\boldsymbol{x})$ by $\text{sign}(g(\boldsymbol{x}))$?
- How about a soft-version of $\text{sign}(g(\boldsymbol{x}))$?
- This gives a logistic regression.



Loss Function for Linear Regression

- All discriminant algorithms have a **Training Loss Function**

$$J(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}(g(\boldsymbol{x}_n), y_n).$$

- In linear regression,

$$\begin{aligned} J(\boldsymbol{\theta}) &= \frac{1}{N}\sum_{n=1}^{N}(g(\boldsymbol{x}_n) - y_n)^2 \\ &= \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{w}^T\boldsymbol{x}_n + w_0 - y_n)^2 \\ &= \frac{1}{N}\left\| \begin{bmatrix} \boldsymbol{x}_1^T & 1 \\ \vdots & \vdots \\ \boldsymbol{x}_N^T & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{w} \\ w_0 \end{bmatrix} - \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \right\|^2 = \frac{1}{N}\|\boldsymbol{A\theta} - \boldsymbol{y}\|^2. \end{aligned}$$

Training Loss for Logistic Regression

$$
\begin{aligned}
J(\boldsymbol{\theta}) &= \sum_{n=1}^{N} \mathcal{L}(h_\theta(\boldsymbol{x}_n), y_n) \\
&= \sum_{n=1}^{N} -\Big\{ y_n \log h_{\boldsymbol{\theta}}(\boldsymbol{x}_n) + (1 - y_n) \log(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}_n)) \Big\}
\end{aligned}
$$

- This loss is also called the **cross-entropy loss**.
- Why do we want to choose this cost function?
- Consider two cases

$$
y_n \log h_{\boldsymbol{\theta}}(\boldsymbol{x}_n) = \begin{cases} 0, & \text{if} \quad y_n = 1, \quad \text{and} \quad h_{\boldsymbol{\theta}}(\boldsymbol{x}_n) = 1, \\ -\infty, & \text{if} \quad y_n = 1, \quad \text{and} \quad h_{\boldsymbol{\theta}}(\boldsymbol{x}_n) = 0, \end{cases}
$$

$$
(1 - y_n)(1 - \log h_{\boldsymbol{\theta}}(\boldsymbol{x}_n)) = \begin{cases} 0, & \text{if} \quad y_n = 0, \quad \text{and} \quad h_{\boldsymbol{\theta}}(\boldsymbol{x}_n) = 0, \\ -\infty, & \text{if} \quad y_n = 0, \quad \text{and} \quad h_{\boldsymbol{\theta}}(\boldsymbol{x}_n) = 1. \end{cases}
$$

- No solution if mismatch

Why Not L2 Loss?

- Why not use L2 loss?

$$J(\boldsymbol{\theta}) = \sum_{n=1}^{N} (h_{\boldsymbol{\theta}}(\boldsymbol{x}_n) - y_n)^2$$

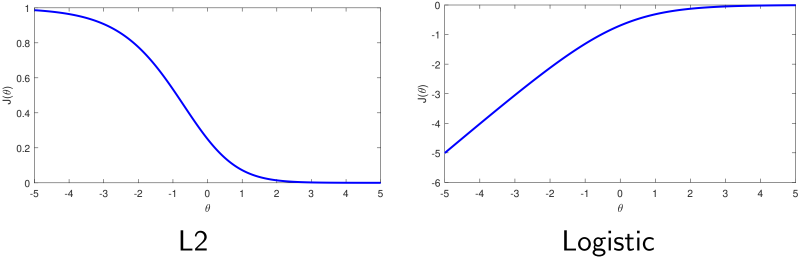
- Let's look at the 1D case:

$$J(\theta) = \left(\frac{1}{1 + e^{-\theta x}} - y \right)^2.$$

- This is NOT convex!
- How about the logistic loss?
- $$J(\theta) = y \log \left(\frac{1}{1 + e^{-\theta x}} \right) + (1 - y) \log \left(1 - \frac{1}{1 + e^{-\theta x}} \right)$$
- This is convex!

Why Not L2 Loss?

- Experiment: Set $x = 1$ and $y = 1$.
- Plot $J(\theta)$ as a function of θ.



- So the L2 loss is not convex, but the logistic loss is concave (negative is convex)
- If you do gradient descent on L2, you will be trapped at local minima

Outline

Discriminative Approaches

- Lecture 14 Logistic Regression 1
- Lecture 15 Logistic Regression 2

This lecture: Logistic Regression 1

- From Linear to Logistic
  - Motivation
  - Loss Function
  - Why not L2 Loss?
- Interpreting Logistic
  - Maximum Likelihood
  - Log-odd
- Convexity
  - Is logistic loss convex?
  - Computation

The Maximum-Likelihood Perspective

- We can show that

$$
\begin{aligned}
&\underset{\boldsymbol{\theta}}{\operatorname{argmin}} \quad J(\boldsymbol{\theta}) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \quad \sum_{n=1}^{N} -\Big\{ y_n \log h_{\boldsymbol{\theta}}(\boldsymbol{x}_n) + (1 - y_n) \log(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}_n)) \Big\} \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \quad -\log \left(\prod_{n=1}^{N} h_{\boldsymbol{\theta}}(\boldsymbol{x}_n)^{y_n} (1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}_n))^{1-y_n} \right) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \quad \prod_{n=1}^{N} \Big\{ h_{\boldsymbol{\theta}}(\boldsymbol{x}_n)^{y_n} (1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}_n))^{1-y_n} \Big\}.
\end{aligned}
$$

- This is maximum-likelihood for a Bernoulli random variable y_n
- The underlying probability is $h_{\boldsymbol{\theta}}(\boldsymbol{x}_n)$

Interpreting $h(\boldsymbol{x}_n)$

- Maximum-likelihood Bernoulli:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\text{argmax}} \prod_{n=1}^{N} \left\{ h_{\boldsymbol{\theta}}(\boldsymbol{x}_n)^{y_n} (1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}_n))^{1-y_n} \right\}.$$

- We can interpret $h_{\boldsymbol{\theta}}(\boldsymbol{x}_n)$ as a probability p. So:

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}_n) = p, \quad \text{and} \quad 1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}_n) = 1 - p.$$

- But p is a function of $\boldsymbol{x}_n$. So how about

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}_n) = p(\boldsymbol{x}_n), \quad \text{and} \quad 1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}_n) = 1 - p(\boldsymbol{x}_n).$$

- And this probability is "after" you see $\boldsymbol{x}_n$. So how about

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}_n) = p(1 \mid \boldsymbol{x}_n), \quad \text{and} \quad 1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}_n) = 1 - p(1 \mid \boldsymbol{x}_n) = p(0 \mid \boldsymbol{x}_n).$$

- So $h_{\boldsymbol{\theta}}(\boldsymbol{x}_n)$ is the **posterior** of observing $\boldsymbol{x}_n$.

Log-Odds

- Let us rewrite J as

$$J(\boldsymbol{\theta}) = \sum_{n=1}^{N} -\Big\{ y_n \log h_{\boldsymbol{\theta}}(\boldsymbol{x}_n) + (1 - y_n) \log(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}_n)) \Big\}$$

$$= \sum_{n=1}^{n} -\Big\{ y_n \log \left(\frac{h_{\boldsymbol{\theta}}(\boldsymbol{x}_n)}{1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}_n)} \right) + \log(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}_n)) \Big\}$$

- In statistics, the term $\log \left(\frac{h_{\boldsymbol{\theta}}(\boldsymbol{x}_n)}{1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}_n)} \right)$ is called the log-odd.
- If we put $h_{\boldsymbol{\theta}}(\boldsymbol{x}_n) = \frac{1}{1+e^{-\boldsymbol{\theta}^T x}}$, we can show that

$$\log \left(\frac{h_{\boldsymbol{\theta}}(\boldsymbol{x})}{1 - h_{\boldsymbol{\theta}}(\boldsymbol{x})} \right) = \log \left(\frac{\frac{1}{1+e^{-\boldsymbol{\theta}^T x}}}{\frac{e^{-\boldsymbol{\theta}^T x}}{1+e^{-\boldsymbol{\theta}^T x}}} \right) = \log \left(e^{\boldsymbol{\theta}^T \boldsymbol{x}} \right) = \boldsymbol{\theta}^T \boldsymbol{x}.$$

- Logistic regression is linear in the log-odd.

Outline

Discriminative Approaches

- Lecture 14 Logistic Regression 1
- Lecture 15 Logistic Regression 2

This lecture: Logistic Regression 1

- From Linear to Logistic
  - Motivation
  - Loss Function
  - Why not L2 Loss?
- Interpreting Logistic
  - Maximum Likelihood
  - Log-odd
- Convexity
  - Is logistic loss convex?
  - Computation

Convexity of Logistic Training Loss

Recall that

$$J(\boldsymbol{\theta}) = \sum_{n=1}^{n} -\left\{ y_n \log\left(\frac{h_{\boldsymbol{\theta}}(\boldsymbol{x}_n)}{1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}_n)}\right) + \log(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}_n)) \right\}$$

- The first term is linear, so it is convex.
- The second term: Gradient:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}}[-\log(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}))] &= -\nabla_{\boldsymbol{\theta}}\left[\log\left(1 - \frac{1}{1 + e^{-\boldsymbol{\theta}^T\boldsymbol{x}}}\right)\right] \\
&= -\nabla_{\boldsymbol{\theta}}\left[\log \frac{e^{-\boldsymbol{\theta}^T\boldsymbol{x}}}{1 + e^{-\boldsymbol{\theta}^T\boldsymbol{x}}}\right] = -\nabla_{\boldsymbol{\theta}}\left[\log e^{-\boldsymbol{\theta}^T\boldsymbol{x}} - \log(1 + e^{-\boldsymbol{\theta}^T\boldsymbol{x}})\right] \\
&= -\nabla_{\boldsymbol{\theta}}\left[-\boldsymbol{\theta}^T\boldsymbol{x} - \log(1 + e^{-\boldsymbol{\theta}^T\boldsymbol{x}})\right] = \boldsymbol{x} + \nabla_{\boldsymbol{\theta}}\left[\log\left(1 + e^{-\boldsymbol{\theta}^T\boldsymbol{x}}\right)\right] \\
&= \boldsymbol{x} + \left(\frac{-e^{-\boldsymbol{\theta}^T\boldsymbol{x}}}{1 + e^{-\boldsymbol{\theta}^T\boldsymbol{x}}}\right)\boldsymbol{x} = h_{\boldsymbol{\theta}}(\boldsymbol{x})\boldsymbol{x}.
\end{aligned}$$

Convexity of Logistic Training Loss

- Gradient of second term is

$$\nabla_{\boldsymbol{\theta}}[-\log(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}))] = h_{\boldsymbol{\theta}}(\boldsymbol{x})\boldsymbol{x}.$$

- Hessian is:

$$\begin{aligned}
\nabla^2_{\boldsymbol{\theta}}[-\log(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}))] &= \nabla_{\boldsymbol{\theta}}\left[h_{\boldsymbol{\theta}}(\boldsymbol{x})\boldsymbol{x}\right] \\
&= \nabla_{\boldsymbol{\theta}}\left[\left(\frac{1}{1+e^{-\boldsymbol{\theta}^T\boldsymbol{x}}}\right)\boldsymbol{x}\right] \\
&= \left(\frac{1}{(1+e^{-\boldsymbol{\theta}^T\boldsymbol{x}})^2}\right)\left(-e^{-\boldsymbol{\theta}^T\boldsymbol{x}}\right)\boldsymbol{x}\boldsymbol{x}^T \\
&= \left(\frac{1}{1+e^{-\boldsymbol{\theta}^T\boldsymbol{x}}}\right)\left(1-\frac{1}{1+e^{-\boldsymbol{\theta}^T\boldsymbol{x}}}\right)\boldsymbol{x}\boldsymbol{x}^T \\
&= h_{\boldsymbol{\theta}}(\boldsymbol{x})[1 - h_{\boldsymbol{\theta}}(\boldsymbol{x})]\boldsymbol{x}\boldsymbol{x}^T.
\end{aligned}$$

Convexity of Logistic Training Loss

- For any $\boldsymbol{v} \in \mathbb{R}^d$, we have that

$$\begin{aligned} \boldsymbol{v}^T \nabla^2_{\boldsymbol{\theta}}[-\log(1-h_{\boldsymbol{\theta}}(\boldsymbol{x}))]\boldsymbol{v} &= \boldsymbol{v}^T \left[h_{\boldsymbol{\theta}}(\boldsymbol{x})[1-h_{\boldsymbol{\theta}}(\boldsymbol{x})]\boldsymbol{x}\boldsymbol{x}^T \right] \boldsymbol{v} \\ &= (h_{\boldsymbol{\theta}}(\boldsymbol{x})[1-h_{\boldsymbol{\theta}}(\boldsymbol{x})]) \, \|\boldsymbol{v}^T\boldsymbol{x}\|^2 \geq 0. \end{aligned}$$

- Therefore the Hessian is positive semi-definite.
- So $-\log(1-h_{\boldsymbol{\theta}}(\boldsymbol{x})$ is convex in $\boldsymbol{\theta}$.
- Conclusion: The training loss function

$$J(\boldsymbol{\theta}) = \sum_{n=1}^{n} -\left\{ y_n \log\left(\frac{h_{\boldsymbol{\theta}}(\boldsymbol{x}_n)}{1-h_{\boldsymbol{\theta}}(\boldsymbol{x}_n)}\right) + \log(1-h_{\boldsymbol{\theta}}(\boldsymbol{x}_n)) \right\}$$

  is **convex** in $\boldsymbol{\theta}$.
- So we can use convex optimization algorithms to find $\boldsymbol{\theta}$.

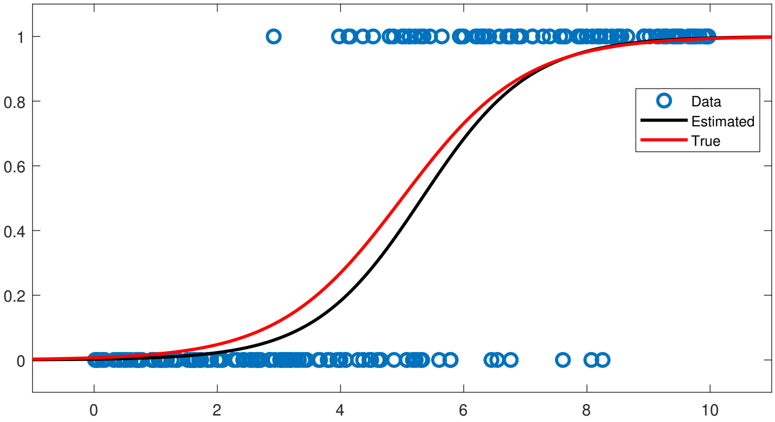
Convex Optimization for Logistic Regression

- We can use CVX to solve the logistic regression problem
- But it requires some re-organization of the equations

$$
\begin{aligned}
J(\boldsymbol{\theta}) &= \sum_{n=1}^{N} -\left\{ y_n \boldsymbol{\theta}^T \boldsymbol{x}_n + \log(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}_n)) \right\} \\
&= \sum_{n=1}^{N} -\left\{ y_n \boldsymbol{\theta}^T \boldsymbol{x}_n + \log\left(1 - \frac{e^{\boldsymbol{\theta}^T \boldsymbol{x}_n}}{1 + e^{\boldsymbol{\theta}^T \boldsymbol{x}_n}}\right) \right\} \\
&= \sum_{n=1}^{N} -\left\{ y_n \boldsymbol{\theta}^T \boldsymbol{x}_n - \log\left(1 + e^{\boldsymbol{\theta}^T \boldsymbol{x}_n}\right) \right\} \\
&= -\left\{ \left(\sum_{n=1}^{N} y_n \boldsymbol{x}_n \right)^T \boldsymbol{\theta} - \sum_{n=1}^{N} \log\left(1 + e^{\boldsymbol{\theta}^T \boldsymbol{x}_n}\right) \right\}.
\end{aligned}
$$

- The last term is a sum of `log-sum-exp`: $\log(e^0 + e^{\theta^T x})$.

Convex Optimization for Logistic Regression



Reading List

Logistic Regression (Machine Learning Perspective)

- Chris Bishop's *Pattern Recognition*, Chapter 4.3
- Hastie-Tibshirani-Friedman's *Elements of Statistical Learning*, Chapter 4.4
- Stanford CS 229 Discriminant Algorithms http://cs229.stanford.edu/notes/cs229-notes1.pdf
- CMU Lecture https://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12.pdf
- Stanford Language Processing https://web.stanford.edu/~jurafsky/slp3/ (Lecture 5)

Logistic Regression (Statistics Perspective)

- Duke Lecture https://www2.stat.duke.edu/courses/Spring13/sta102.001/Lec/Lec20.pdf
- Princeton Lecture https://data.princeton.edu/wws509/notes/c3.pdf