ECE595 / STAT598: Machine Learning I

Lecture 14.3: Logistic Regression 1 - Convexity

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

PURDUE
UNIVERSITY

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:

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

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

$$= -\nabla_{\boldsymbol{\theta}}\left[-\boldsymbol{\theta}^T\boldsymbol{x} - \log(1 + e^{-\boldsymbol{\theta}^T\boldsymbol{x}})\right] = \boldsymbol{x} + \nabla_{\boldsymbol{\theta}}\left[\log\left(1 + e^{-\boldsymbol{\theta}^T\boldsymbol{x}}\right)\right]$$

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

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

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

- 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 $\boldsymbol{\theta}$.

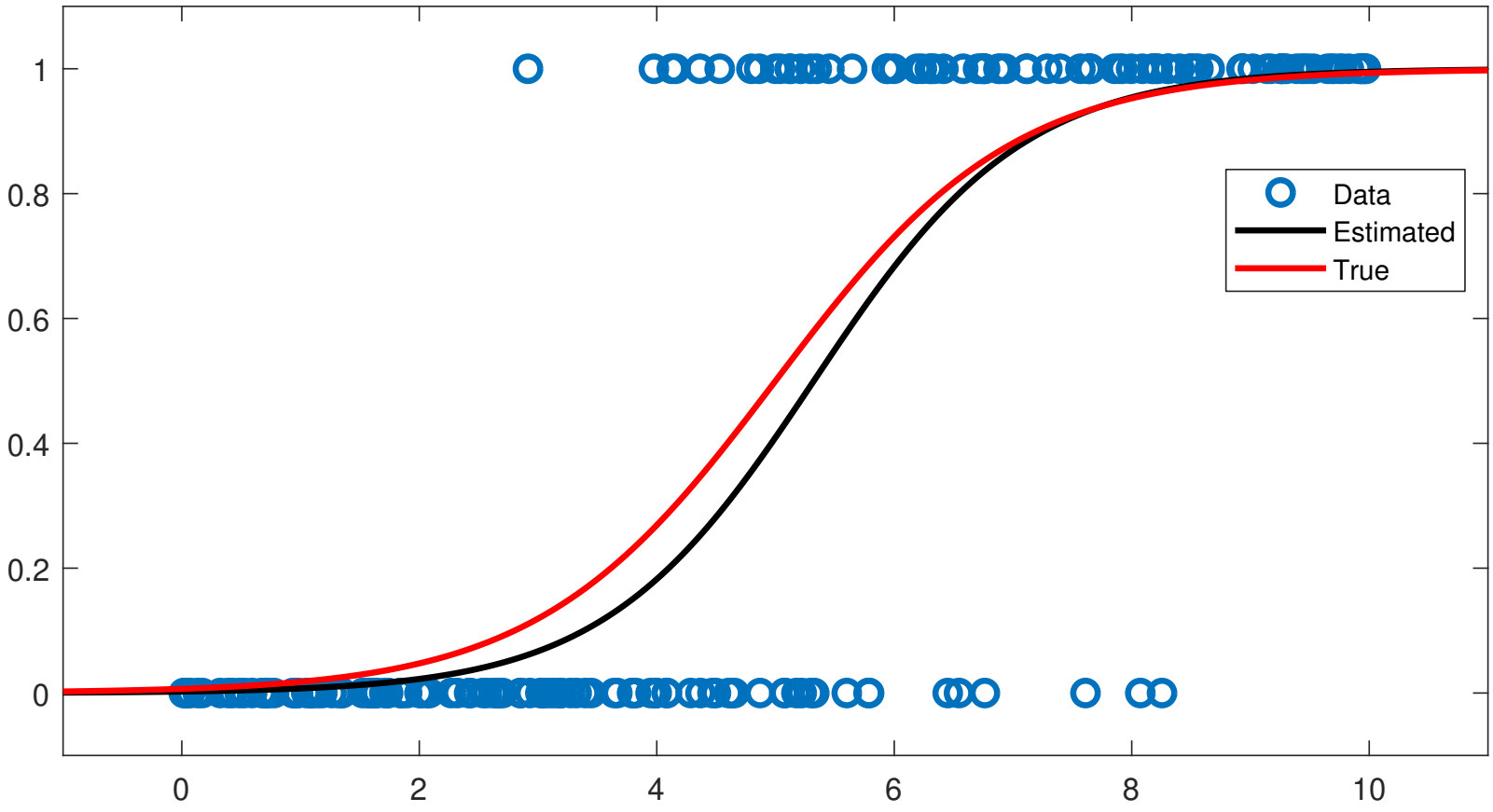
Convex Optimization for Logistic Regression

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

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

- 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