# ECE595 / STAT598: Machine Learning I
## Lecture 14.2: Logistic Regression 1 - Interpreting Logistic

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
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

# 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

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

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

- 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 $h_{\boldsymbol{\theta}}(\boldsymbol{x}_n) = \frac{1}{1+e^{-\theta^T x}}$, we can show that

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

- Logistic regression is linear in the log-odd.