ECE595 / STAT598: Machine Learning I

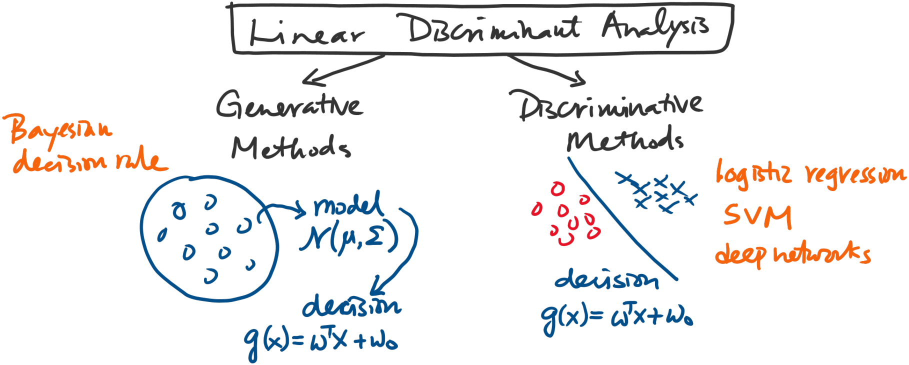
Lecture мп.1Y Logistic Regression 1 - From Linear to Logistic

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

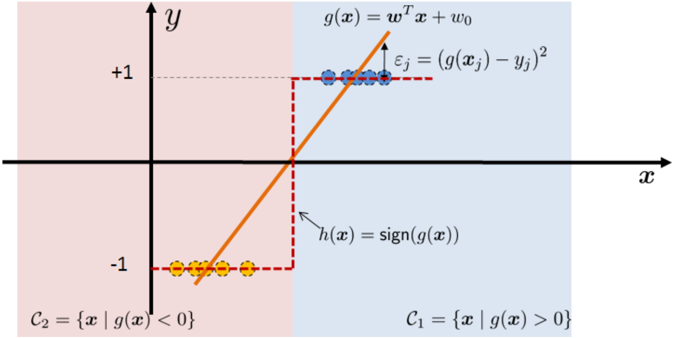
- Lecture 14 Logistic Regression 1
- Lecture 15 Logistic Regression 2

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

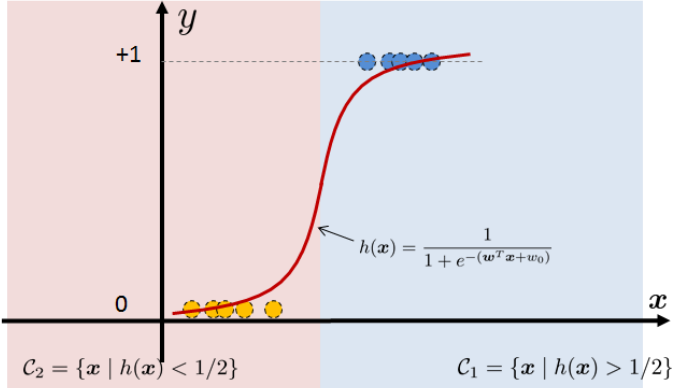
Geometry of Linear Regression

- 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

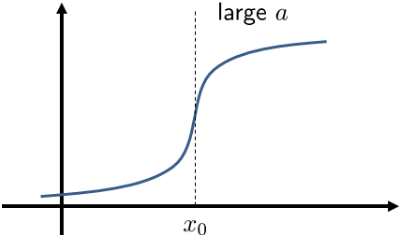
- The function

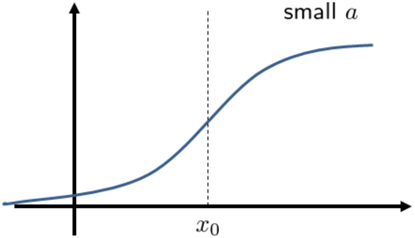
$$h(\boldsymbol{x}) = \frac{1}{1+e^{-g(\boldsymbol{x})}} = \frac{1}{1+e^{-(\boldsymbol{w}^T\boldsymbol{x}+w_0)}}$$

  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



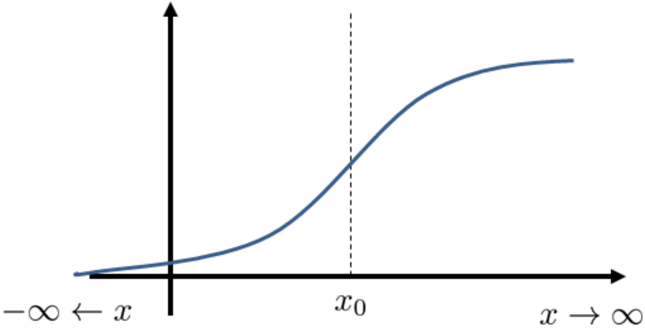


Sigmoid Function

- 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".



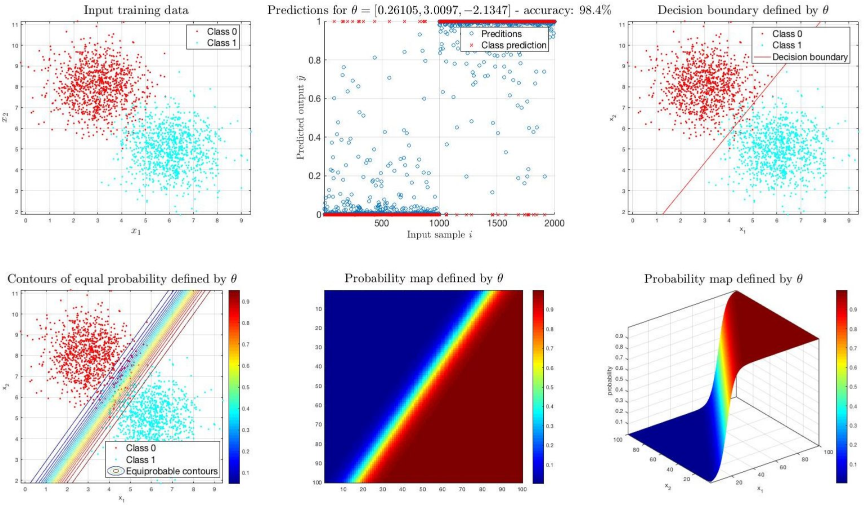
Sigmoid Function

- 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".

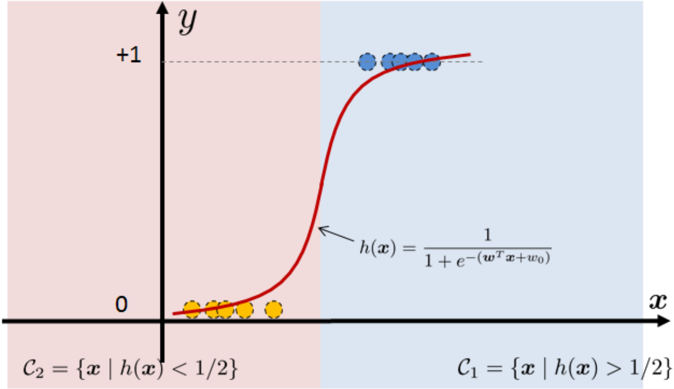
Sigmoid Function



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?

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

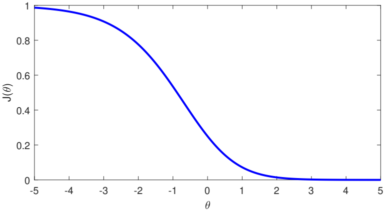
- Let's look at the 1D case:

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

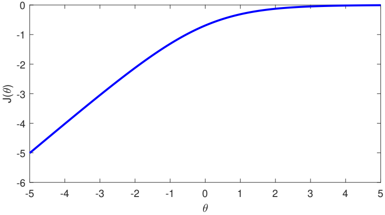
- This is NOT convex!
- How about the logistic loss?
- $$J(\theta) = y \log \left(\frac{1}{1 + e^{-\theta x}} \right) + (1 - y) \log \left(1 - \frac{1}{1 + e^{-\theta x}} \right)$$
- This is convex!

Why Not L2 Loss?

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



L2



Logistic

- 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