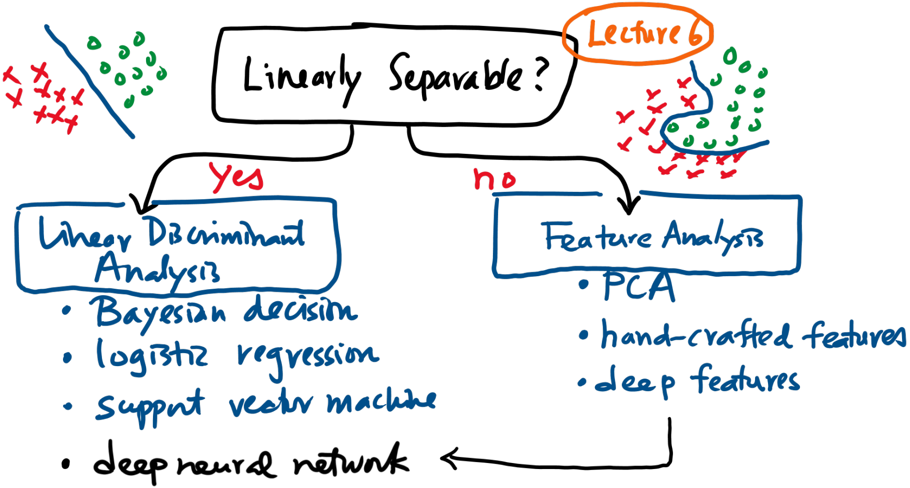
ECE595 / STAT598: Machine Learning I
Lecture 6.1: Linear Separability - Notations

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

Overview


Linearly Separable?
Lecture 6
yes
no
Linear Discriminant Analysis
• Bayesian decision
• logistic regression
• support vector machine
• deep neural network
Feature Analysis
• PCA
• hand-crafted features
• deep features

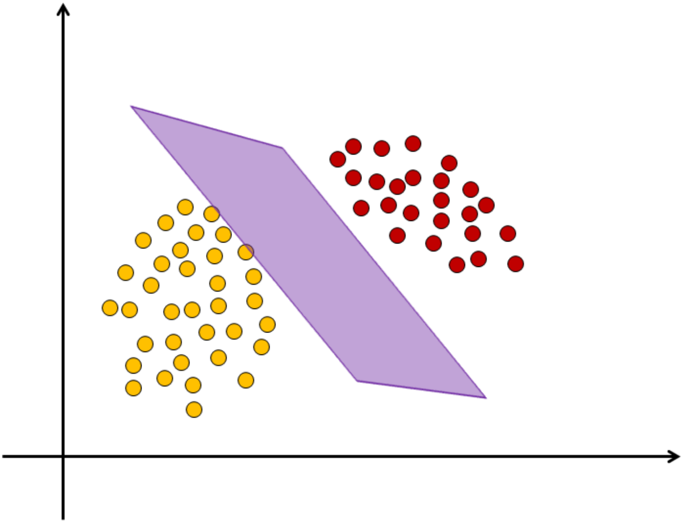
Outline

Goal: Understand the geometry of linear separability.

- Notations
  - Input Space, Output Space, Hypothesis
  - Discriminant Function
- Geometry of Discriminant Function
  - Separating Hyperplane
  - Normal Vector
  - Distance from Point to Plane
- Linear Separability
  - Which set is linearly separable?
  - Separating Hyperplane Theorem
  - What if theorem fails?

Supervised Classification

The goal of supervised classification is to construct a **decision boundary** such that the two classes can be (maximally) **separated**.



Terminology

- **Input vectors**: $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$.
  - E.g., images, speech, EEG signal, rating, etc
- **Input space**: $\mathcal{X}$. Every $\mathbf{x}_n \in \mathcal{X}$.
- **Labels** $y_1, y_2, \ldots, y_N$.
- **Label space**: $\mathcal{Y}$. Every $y_n \in \mathcal{Y}$.
  - If labels are binary, e.g., $y_n = \pm 1$, then

$$\mathcal{Y} = \{+1, -1\}.$$

  - Labels are arbitrary. $\{+1, -1\}$ and $\{0, 1\}$ has no difference.
- **Target function** $f : \mathcal{X} \to \mathcal{Y}$. Unknown.
  - Relationship:

$$y_n = f(\mathbf{x}_n).$$

- **Hypothesis** $h : \mathcal{X} \to \mathcal{Y}$. Ideally, want

$$h(\mathbf{x}) \approx f(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X}.$$

Binary Case

If we restrict ourselves to binary classifier, then

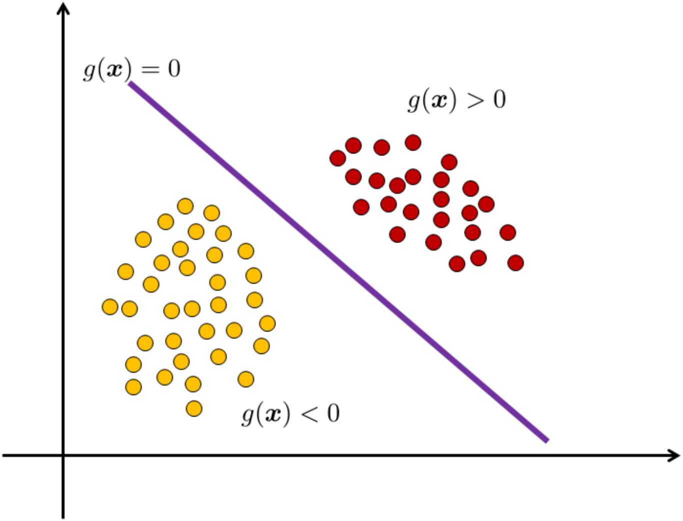
$$h(\boldsymbol{x}) = \begin{cases} 1, & \text{if} \quad g(\boldsymbol{x}) > 0 \\ 0, & \text{if} \quad g(\boldsymbol{x}) < 0 \\ \text{either}, & \text{if} \quad g(\boldsymbol{x}) = 0 \end{cases}$$

- $g : \mathcal{X} \to \mathbb{R}$ is called a **discriminant function**.
- $g(\boldsymbol{x}) > 0$: $\boldsymbol{x}$ lives on the positive side of g.
- $g(\boldsymbol{x}) < 0$: $\boldsymbol{x}$ lives on the negative side of g.
- $g(\boldsymbol{x}) = 0$: The decision boundary.
- You can also claim

$$h(\boldsymbol{x}) = \begin{cases} +1, & \text{if} \quad g(\boldsymbol{x}) > 0 \\ -1, & \text{if} \quad g(\boldsymbol{x}) < 0 \\ \text{either}, & \text{if} \quad g(\boldsymbol{x}) = 0 \end{cases}$$

No difference as far as decision is concerned.

Binary Case



Linear Discriminant Function

A linear discriminant function takes the form

$$g(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0.$$

- $\boldsymbol{w} \in \mathbb{R}^d$: linear coefficients
- $w_0 \in \mathbb{R}$: bias / offset
- Define the overall **parameter**

$$\boldsymbol{\theta} = \{\boldsymbol{w}, w_0\} \in \mathbb{R}^{d+1}.$$

- Example:
  - If $d = 2$, then

$$g(\boldsymbol{x}) = w_2 x_2 + w_1 x_1 + w_0.$$

  - $g(\boldsymbol{x}) = 0$ means

$$x_2 = \underbrace{-\frac{w_1}{w_2}}_{\text{slope}} x_1 + \underbrace{-\frac{w_0}{w_2}}_{\text{y-intercept}}.$$

Linear Discriminant Function