ECE 595: Machine Learning I

Lecture 9.1: Bayesian Decision - Review of High-Dimensional Gaussian

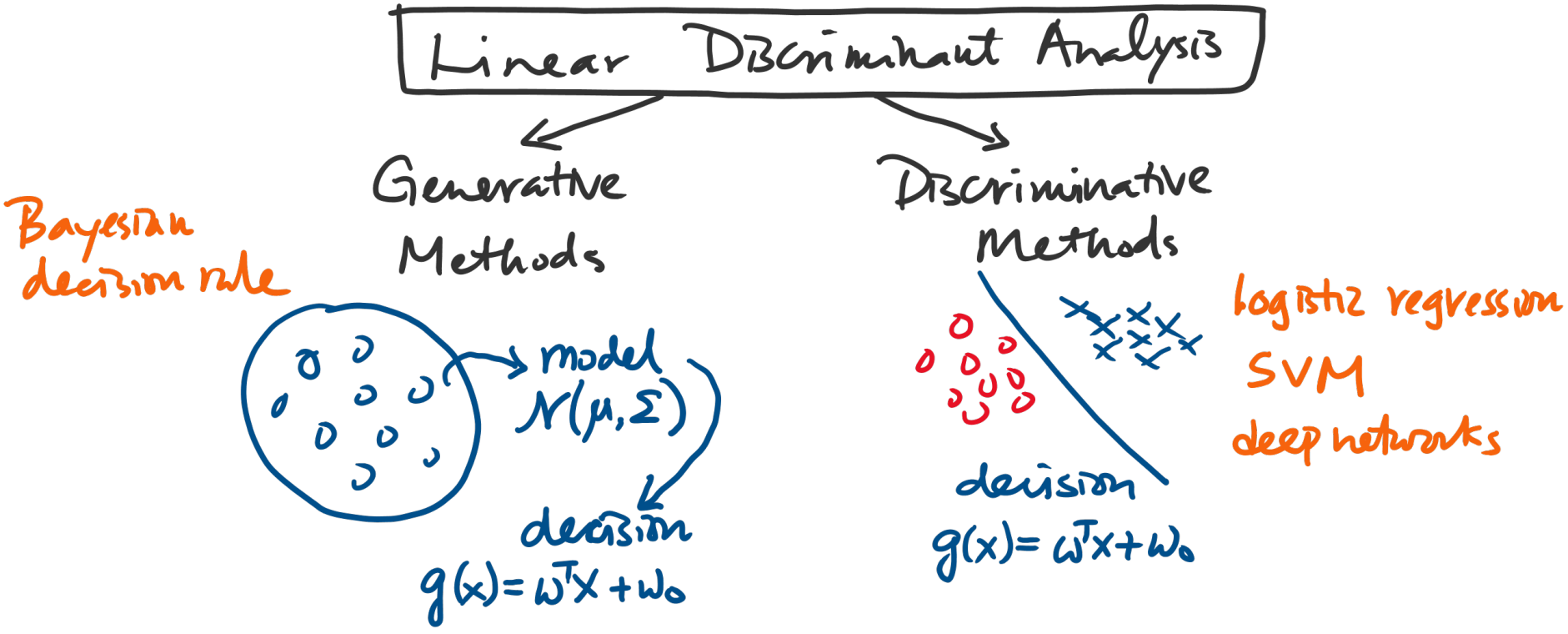
Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

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

Generative Approach

Goal: Construct a discriminant function $g(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0$ from the data.

- Suppose there are two classes C_1 and C_2.
- Each class is modeled as a Gaussian.
- We are going to utilize two concepts:
- **likelihood function**

$$p_{\boldsymbol{X}|Y}(\boldsymbol{x}|i) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

- **prior** distribution

$$p_Y(i) = \pi_i$$

Outline

Generative Approaches

- Lecture 9 Bayesian Decision Rules
- Lecture 10 Evaluating Performance
- Lecture 11 Bayesian Parameter Estimation
- Lecture 12 Bayesian Prior
- Lecture 13 Connecting Bayesian and Linear Regression

Today's Lecture

- Review of High-Dimensional Gaussian
  - Likelihood and prior
  - Gaussian PDF
- Basic Principle
  - Making the Bayesian decision
  - 1D Illustration
- The Three Cases
  - $\boldsymbol{\Sigma}_i = \sigma^2 \boldsymbol{I}$
  - $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$ (Next Lecture)
  - General $\boldsymbol{\Sigma}_i$ (Next Lecture)

High-dimensional Gaussian

An d-dimensional **Gaussian** has a PDF

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\},$$

where d denotes the dimensionality of the vector $\boldsymbol{x}$.

- The **mean vector** $\boldsymbol{\mu}$ is

$$\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}] = \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_d] \end{bmatrix}$$

- The **covariance matrix** $\boldsymbol{\Sigma}$ is

$$\boldsymbol{\Sigma} = \mathbb{E}[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^T] = \begin{bmatrix} \mathrm{Var}[X_1] & \mathrm{Cov}(X_1, X_2) & \dots & \mathrm{Cov}(X_1, X_d) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}[X_2] & \dots & \mathrm{Cov}(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_d, X_1) & \mathrm{Cov}(X_N, X_2) & \dots & \mathrm{Var}[X_d] \end{bmatrix}$$

- $\boldsymbol{\Sigma}$ is always positive semi-definite. (Why?)

Special Case: Diagonal Covariance

- Suppose that X_i and X_j are independent for all $i \neq j$.
- This implies $\mathrm{Cov}(X_i, X_j) = 0$
- Simplify $\boldsymbol{\Sigma}$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_d^2 \end{bmatrix},$$

- Then, the exponential is

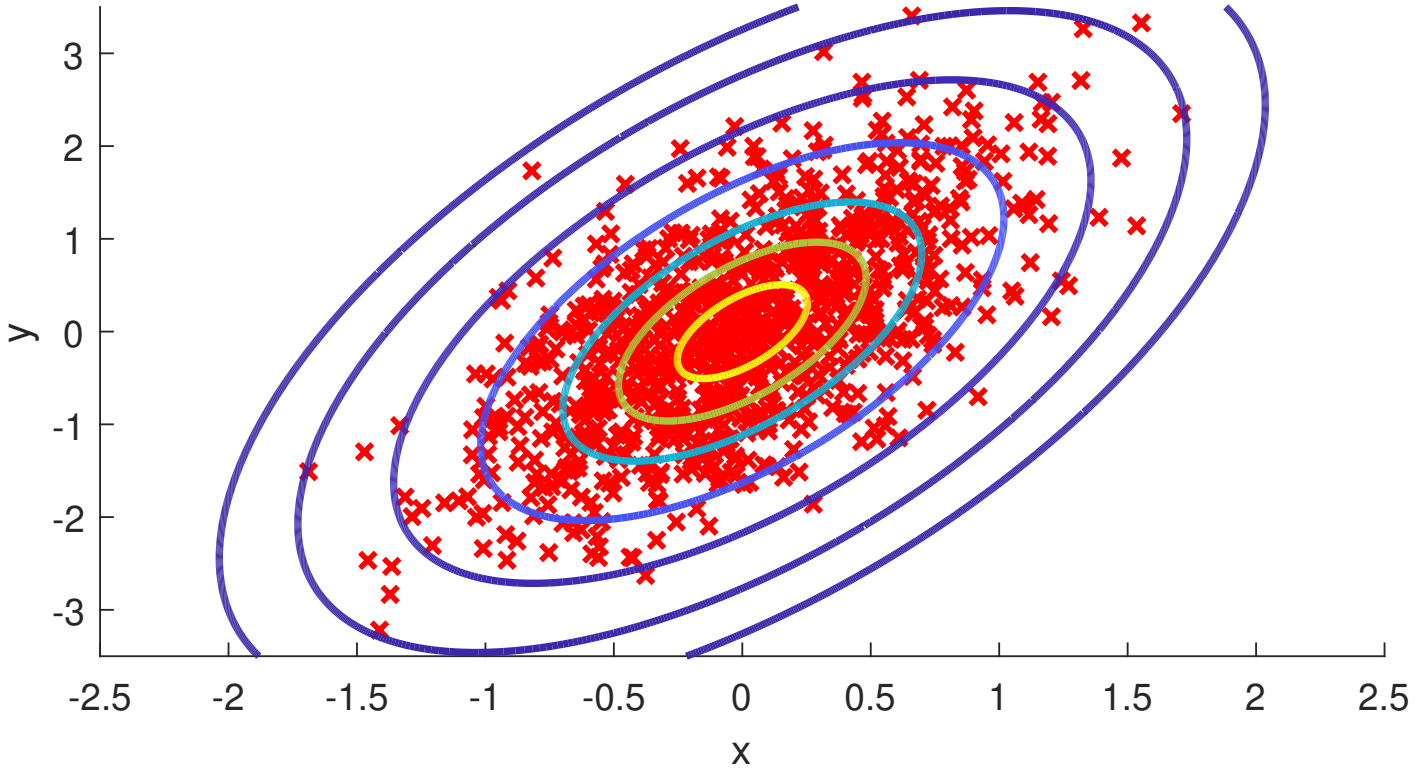
$$(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2}.$$

- And hence, the PDF is

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{ -\frac{(x_i - \mu_i)^2}{2\sigma_i^2} \right\}.$$

Visualization

- Generate 1000 random samples from a 2D Gaussian
- $\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, and $\boldsymbol{\Sigma} = \begin{bmatrix} 0.25 & 0.3 \\ 0.3 & 1 \end{bmatrix}$



Conditional Gaussian

- Data $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$.
- Class $Y \in \{1, 2, \ldots, K\}$.
- **Likelihood**:

$$p_{\boldsymbol{X}|Y}(\boldsymbol{x}|k) = \text{Probability of getting } \boldsymbol{X} \text{ given } Y$$

- **Prior**:

$$p_Y(k) = \text{Probability of getting } Y$$

- **Posterior**:

$$p_{Y|\boldsymbol{X}}(k|\boldsymbol{x}) = \text{Probability of getting } Y \text{ given } \boldsymbol{X}$$

- Related by

$$p_{Y|\boldsymbol{X}}(k|\boldsymbol{x}) = \frac{p_{\boldsymbol{X}|Y}(\boldsymbol{x}|k)p_Y(k)}{p_{\boldsymbol{X}}(\boldsymbol{x})} = \frac{p_{\boldsymbol{X}|Y}(\boldsymbol{x}|k)p_Y(k)}{\sum_k p_{\boldsymbol{X}|Y}(\boldsymbol{x}|k)p_Y(k)}$$

Example

- Two Gaussian $\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$.
- **Prior** probability of getting a class is

$$p_Y(1) = \pi_1 \quad \text{and} \quad p_Y(2) = \pi_2.$$

- The **likelihood** term is

$$\begin{aligned} p_{\boldsymbol{X}|Y}(\boldsymbol{x}|k) &= \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ &= \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_k|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\} \end{aligned}$$

- The **posterior** is

$$\begin{aligned} p_{Y|\boldsymbol{X}}(k|\boldsymbol{x}) &= \frac{p_{\boldsymbol{X}|Y}(\boldsymbol{x}|k) p_Y(k)}{p_{\boldsymbol{X}}(\boldsymbol{x})} \\ &= \frac{\frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_k|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\} \cdot \pi_k}{\sum_{k=1}^{K} \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_k|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\} \cdot \pi_k} \end{aligned}$$

Negative Log-Likelihood

Negative Log-Likelihood for Gaussian:

$$
\begin{aligned}
& -\log p_{\boldsymbol{X}|Y}(\boldsymbol{x}|k) \\
&= -\log\left(\frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}_k|}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\}\right) \\
&= \underbrace{\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)}_{\text{contains } \boldsymbol{x}} \quad \underbrace{-\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}_k|}_{\text{no } \boldsymbol{x}}.
\end{aligned}
$$

- $(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \geq 0$, always.
- $\sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$ is called **Mahalanobis distance**.