ECE 595: Machine Learning I

Lecture 9.3: Bayesian Decision - The Three Cases I

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

PURDUE
UNIVERSITY

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)

Three Cases of Gaussians

Discriminant function of Gaussian:

$$\begin{aligned} g_i(\boldsymbol{x}) &= \log p_{\boldsymbol{X}|Y}(\boldsymbol{x}|i) + \log \pi_i \\ &= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| + \log \pi_i. \end{aligned}$$

- $\boldsymbol{\Sigma}_i = \sigma^2 \boldsymbol{I}$
  - All Gaussians have the same covariance matrix
  - The covariance matrix is diagonal and same variance
- $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$
  - All Gaussians have the same covariance matrix
  - The covariance matrix can be anything
- arbitrary $\boldsymbol{\Sigma}_i$
  - Any positive semi-definite covariance matrix

Case 1: $\boldsymbol{\Sigma}_i = \sigma^2 \boldsymbol{I}$

Put $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$:

$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\log|\boldsymbol{\Sigma}| + \log\pi_i.$$

Let us do some simplification:

$$\begin{aligned}
g_i(\boldsymbol{x}) &= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) - \cancel{\frac{1}{2}\log|\boldsymbol{\Sigma}|} + \log\pi_i \\
&= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) + \log\pi_i \\
&= -\frac{1}{2\sigma^2}\|\boldsymbol{x}-\boldsymbol{\mu}_i\|^2 + \log\pi_i \\
&= -\frac{1}{2\sigma^2}\left(\|\boldsymbol{x}\|^2 - 2\boldsymbol{x}^T\boldsymbol{\mu}_i + \|\boldsymbol{\mu}_i\|^2\right) + \log\pi_i \\
&= -\frac{1}{2\sigma^2}\left(\cancel{\|\boldsymbol{x}\|^2} - 2\boldsymbol{x}^T\boldsymbol{\mu}_i + \|\boldsymbol{\mu}_i\|^2\right) + \log\pi_i \\
&= \left(\frac{\boldsymbol{\mu}_i}{\sigma^2}\right)^T\boldsymbol{x} - \left(\frac{\|\boldsymbol{\mu}_i\|^2}{2\sigma^2} - \log\pi_i\right).
\end{aligned}$$

Case 1: $\boldsymbol{\Sigma}_i = \sigma^2 \boldsymbol{I}$

$$g_i(\boldsymbol{x}) = \underbrace{\left(\frac{\boldsymbol{\mu}_i}{\sigma^2}\right)^T}_{\boldsymbol{w}_i} \boldsymbol{x} - \underbrace{\left(\frac{\|\boldsymbol{\mu}_i\|^2}{2\sigma^2} - \log \pi_i\right)}_{w_{i0}}$$

$$= \boldsymbol{w}_i^T \boldsymbol{x} + w_{i0}$$

So if the i-th and the j-th discriminant functions are

$$g_i(\boldsymbol{x}) = \boldsymbol{w}_i^T \boldsymbol{x} + w_{i0}$$

$$g_j(\boldsymbol{x}) = \boldsymbol{w}_j^T \boldsymbol{x} + w_{j0},$$

then,

$$g(\boldsymbol{x}) = g_i(\boldsymbol{x}) - g_j(\boldsymbol{x})$$

$$= \underbrace{\left(\frac{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j}{\sigma^2}\right)^T}_{\boldsymbol{w}_i - \boldsymbol{w}_j} \boldsymbol{x} + \underbrace{\left(-\frac{\|\boldsymbol{\mu}_i\|^2 - \|\boldsymbol{\mu}_j\|^2}{2\sigma^2} + \log \frac{\pi_i}{\pi_j}\right)}_{w_{i0} - w_{j0}}.$$

Case 1: $\boldsymbol{\Sigma}_i = \sigma^2 \boldsymbol{I}$

Theorem

If $\boldsymbol{\Sigma}_i = \sigma^2 \boldsymbol{I}$, then the separating hyperplane is given by

$$g(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0 = 0,$$

where

$$\boldsymbol{w} = \frac{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j}{\sigma^2}, \qquad \text{and} \qquad w_0 = -\frac{\|\boldsymbol{\mu}_i\|^2 - \|\boldsymbol{\mu}_j\|^2}{2\sigma^2} + \log \frac{\pi_i}{\pi_j}.$$

- You tell me the two Gaussians: $\boldsymbol{\mu}_i$, $\boldsymbol{\mu}_j$, π_i, π_j, σ
- I return you a separating hyperplane

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

- This is the best possible hyperplane according to posterior distribution

Case 1: $\boldsymbol{\Sigma}_i = \sigma^2 \boldsymbol{I}$: Geometry

Can we write $g(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0$ in terms of

$$g(\boldsymbol{x}) = \boldsymbol{w}^T (\boldsymbol{x} - \boldsymbol{x}_0).$$

Not too difficult:

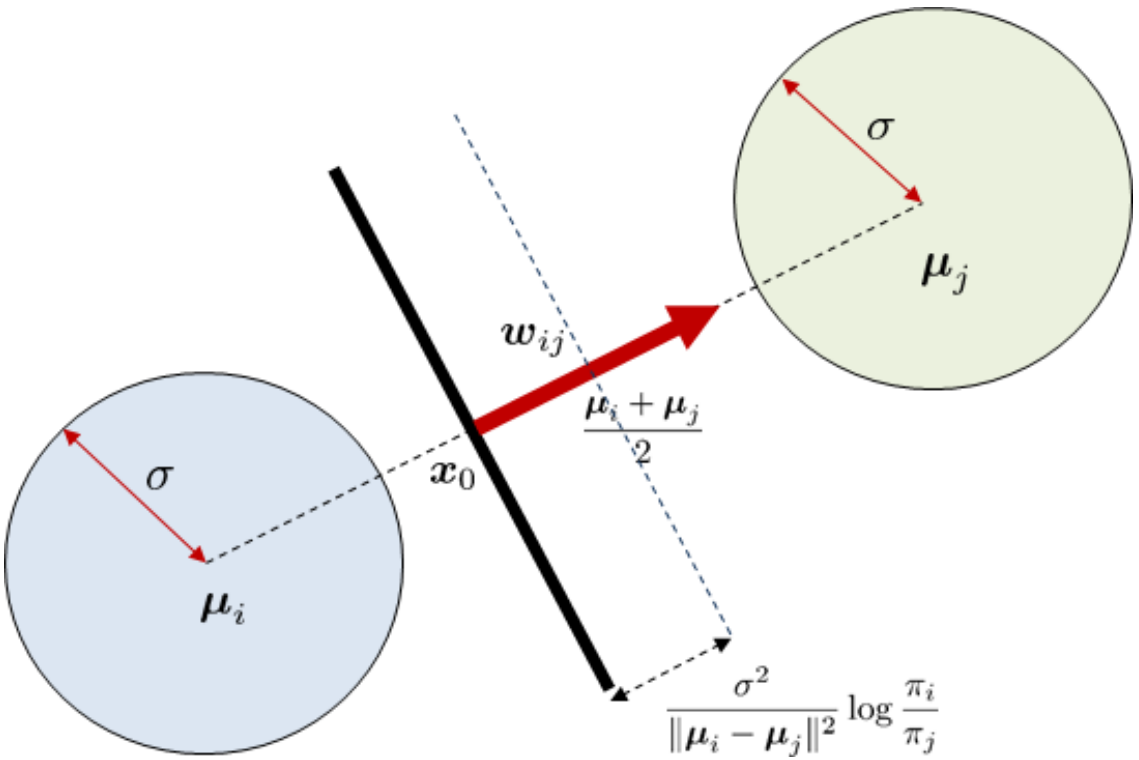
$$\begin{aligned} g(\boldsymbol{x}) &= \left(\frac{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j}{\sigma^2} \right)^T \boldsymbol{x} - \left(\frac{\|\boldsymbol{\mu}_i\|^2}{2\sigma^2} - \frac{\|\boldsymbol{\mu}_j\|^2}{2\sigma^2} \right) + \log \frac{\pi_i}{\pi_j} \\ &= \left(\frac{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j}{\sigma^2} \right)^T \Bigg[\boldsymbol{x} - \underbrace{\frac{\boldsymbol{\mu}_i + \boldsymbol{\mu}_j}{2} + \sigma^2 \left(\log \frac{\pi_i}{\pi_j} \right) \frac{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2}}_{\boldsymbol{x}_0} \Bigg] \end{aligned}$$

Therefore, we have

$$\boldsymbol{w} = \frac{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j}{\sigma^2}, \qquad \text{and} \qquad \boldsymbol{x}_0 = \frac{\boldsymbol{\mu}_i + \boldsymbol{\mu}_j}{2} - \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2} \left(\log \frac{\pi_i}{\pi_j} \right) (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j),$$

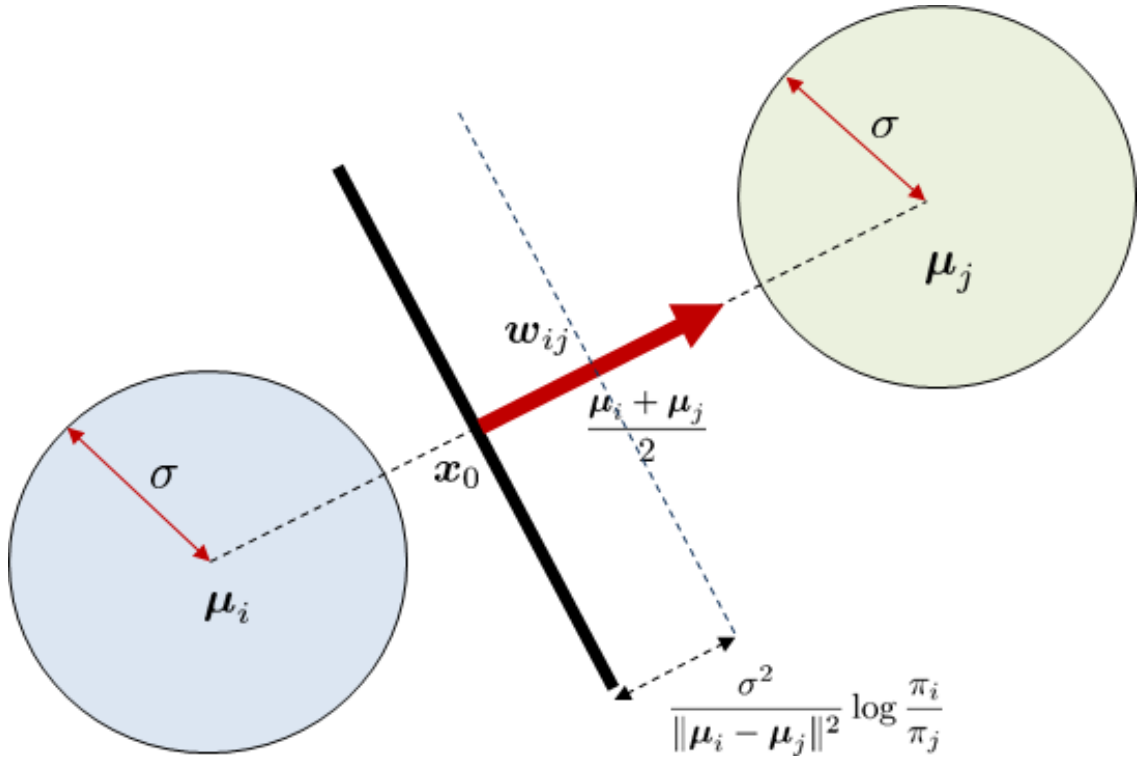
Case 1: $\boldsymbol{\Sigma}_i = \sigma^2 \boldsymbol{I}$: Geometry

$$\boldsymbol{w} = \frac{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j}{\sigma^2}, \qquad \text{and} \qquad \boldsymbol{x}_0 = \frac{\boldsymbol{\mu}_i + \boldsymbol{\mu}_j}{2} - \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2} \left(\log \frac{\pi_i}{\pi_j}\right) (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j),$$



Interpreting Results

Here are the geometric interpretations:



- Normal vector is $\boldsymbol{w} = \frac{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j}{\sigma^2}$. It points from one center to another.
- Midpoint is $\boldsymbol{x}_0 = \frac{\boldsymbol{\mu}_i + \boldsymbol{\mu}_j}{2}$
- The prior creates an offset. Offset direction is also $\boldsymbol{\mu}_i - \boldsymbol{\mu}_j$. If $\pi_i = \pi_j = 1/2$, then $\log(\pi_1/\pi_j) = 0$.

Reading List

High Dimensional Gaussian

- Bishop, Pattern Recognition and Machine Learning, Chapter 2.3
- Stanford CS 229 Tutorial on Gaussian
  http://cs229.stanford.edu/section/gaussians.pdf

Bayesian Decision Rule

- Bishop, Pattern Recognition and Machine Learning, Chapter 4.1
- Duda, Hart and Stork's Pattern Classification, Chapter 2.1, 2.2, 2.6
- Stanford CS 229 Generative Algorithms
  http://cs229.stanford.edu/notes/cs229-notes2.pdf
- UCSD ECE 271A, Lecture 4 and 5
  http://www.svcl.ucsd.edu/courses/ece271A/ece271A.htm