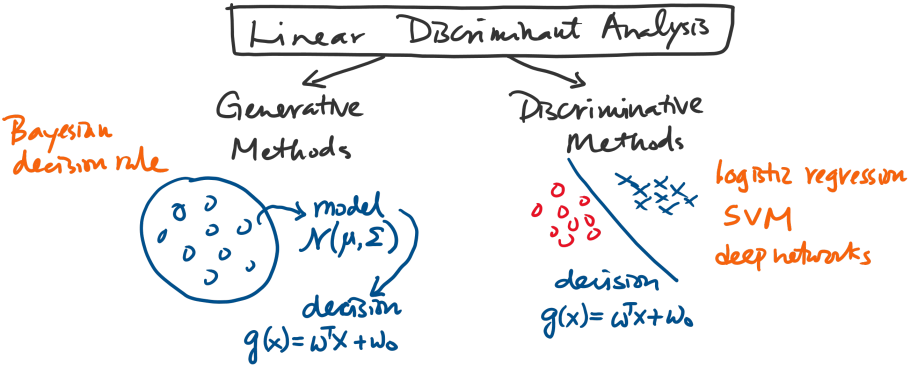
ECE 595: Machine Learning I
Lecture 09 Bayesian Decision

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

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

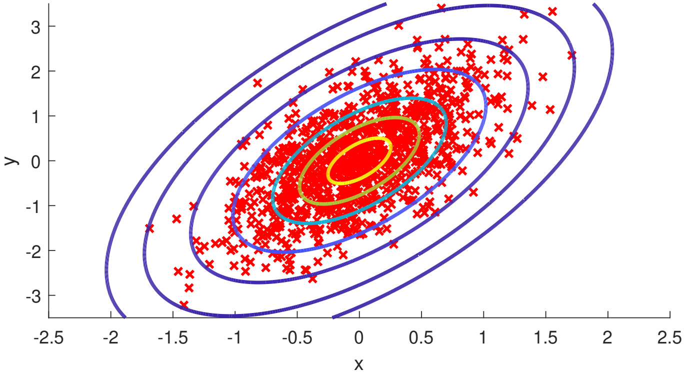
$$(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2}.$$

- 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:

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

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

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)

Interaction between Likelihood and Prior

- According to **Bayes Theorem**, we have that

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

- Posterior: **After** you have seen $\boldsymbol{x}$
- Likelihood: **Before** you see $\boldsymbol{x}$
- Prior: You subjective believe of class label

- You cannot just use $p_Y(i)$; Otherwise you are not using data
- You cannot just use $p_{\boldsymbol{X}|Y}(\boldsymbol{x}|i)$; Otherwise you cannot explain "Y given $\boldsymbol{X}$"

Making the Bayesian Decision

Which class is more likely?

$$
\begin{aligned}
i^* &= \underset{i}{\operatorname{argmax}} \; p_{Y|\boldsymbol{X}}(i|\boldsymbol{x}) \\
&= \underset{i}{\operatorname{argmax}} \; \frac{p_{\boldsymbol{X}|Y}(\boldsymbol{x}|i)p_Y(i)}{p_{\boldsymbol{X}}(\boldsymbol{x})} \\
&= \underset{i}{\operatorname{argmax}} \; \log p_{\boldsymbol{X}\,|\,Y}(\boldsymbol{x}|i) + \log \pi_i - \log p_{\boldsymbol{X}}(\boldsymbol{x}) \\
&= \underset{i}{\operatorname{argmax}} \; \log p_{\boldsymbol{X}\,|\,Y}(\boldsymbol{x}|i) + \log \pi_i - \cancel{\log p_{\boldsymbol{X}}(\boldsymbol{x})} \quad \text{remove}
\end{aligned}
$$

- Solution = the most likely class according to posterior
- This involves a likelihood which depends on the model you choose
- This involves a prior term which is subjective

Let us Plug-in Multi-dimensional Gaussian

Recall d-dimensional Gaussian.

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

Plug this into the discriminant function

$$\begin{aligned} i^* &= \underset{i}{\operatorname{argmax}} \; \log p_{\boldsymbol{X} \mid Y}(\boldsymbol{x} \mid i) + \log \pi_i \\ &= \underset{i}{\operatorname{argmax}} \; -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i) - \cancel{\frac{d}{2} \log(2\pi)} - \frac{1}{2} \log |\boldsymbol{\Sigma}_i| + \log \pi_i \\ &= \underset{i}{\operatorname{argmax}} \; \underbrace{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i)}_{\text{depend on } \boldsymbol{x}} \underbrace{-\frac{1}{2} \log |\boldsymbol{\Sigma}_i| + \log \pi_i}_{\text{does not depend on } \boldsymbol{x}}. \end{aligned}$$

Special Case: 1D; Two classes

The decision rule is

$$i^* = \underset{i}{\operatorname{argmax}} \; \underbrace{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)}_{\text{depend on } \boldsymbol{x}} \underbrace{-\frac{1}{2}\log|\boldsymbol{\Sigma}_i| + \log \pi_i}_{\text{does not depend on } \boldsymbol{x}}.$$

Substitute $\boldsymbol{\Sigma}_i = \sigma^2$, and $\boldsymbol{\mu}_i = \mu_i$. Do two classes.

$$
\begin{aligned}
-\frac{(x-\mu_1)^2}{2\sigma^2} - \log\sigma + \log\pi_1 &\gtrless_{\mathcal{C}_2}^{\mathcal{C}_1} -\frac{(x-\mu_2)^2}{2\sigma^2} - \log\sigma + \log\pi_2 \\
-\frac{(x-\mu_1)^2}{2\sigma^2} - \cancel{\log\sigma} + \log\pi_1 &\gtrless_{\mathcal{C}_2}^{\mathcal{C}_1} -\frac{(x-\mu_2)^2}{2\sigma^2} - \cancel{\log\sigma} + \log\pi_2 \\
&\vdots \\
x &\gtrless_{\mathcal{C}_2}^{\mathcal{C}_1} \underbrace{\frac{\mu_1-\mu_2}{2} - \frac{\sigma^2}{\mu_1-\mu_2}\log\frac{\pi_1}{\pi_2}}_{\text{does not depend on } x}.
\end{aligned}
$$

Connecting to Linear Discriminant Function

Recall: A hypothesis function is

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

If there are only two classes, then we can define

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

where the i-th discriminant function is

$$g_i(\boldsymbol{x}) = \log p_{\boldsymbol{X}|Y}(\boldsymbol{x}|i) + \log \pi_i.$$

- Class i if $g(\boldsymbol{x}) > 0 \iff g_i(\boldsymbol{x}) > g_j(\boldsymbol{x})$
- Class j if $g(\boldsymbol{x}) < 0 \iff g_i(\boldsymbol{x}) < g_j(\boldsymbol{x})$
- Either if $g(\boldsymbol{x}) = 0 \iff g_i(\boldsymbol{x}) = g_j(\boldsymbol{x})$

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

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

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

$$
\begin{aligned}
g_i(\boldsymbol{x}) &= \boldsymbol{w}_i^T \boldsymbol{x} + w_{i0} \\
g_j(\boldsymbol{x}) &= \boldsymbol{w}_j^T \boldsymbol{x} + w_{j0},
\end{aligned}
$$

then,

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

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

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

- 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:

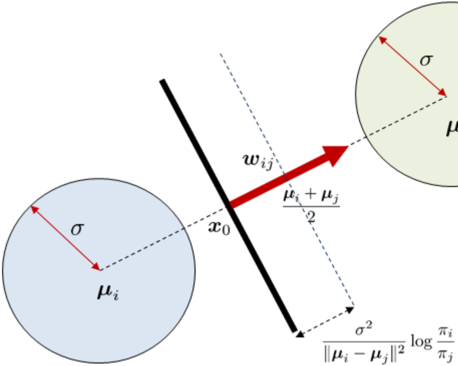
$$\begin{aligned}
g(\boldsymbol{x}) &= \left(\frac{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j}{\sigma^2}\right)^T \boldsymbol{x} - \left(\frac{\|\boldsymbol{\mu}_i\|^2}{2\sigma^2} - \frac{\|\boldsymbol{\mu}_j\|^2}{2\sigma^2}\right) + \log \frac{\pi_i}{\pi_j} \\
&= \left(\frac{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j}{\sigma^2}\right)^T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

Therefore, we have

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

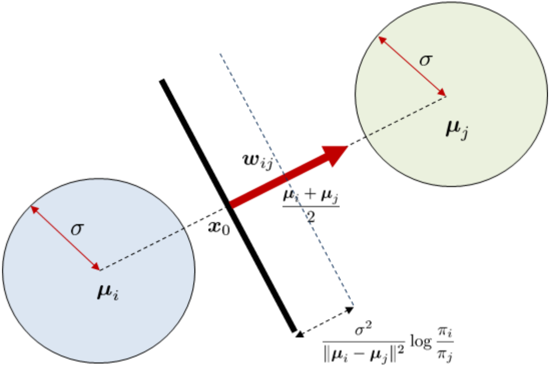
Case 1: $\boldsymbol{\Sigma}_i = \sigma^2 \boldsymbol{I}$: Geometry

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



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