ECE595 / STAT598: Machine Learning I
Lecture 10 Minimum Probability of Error Rule

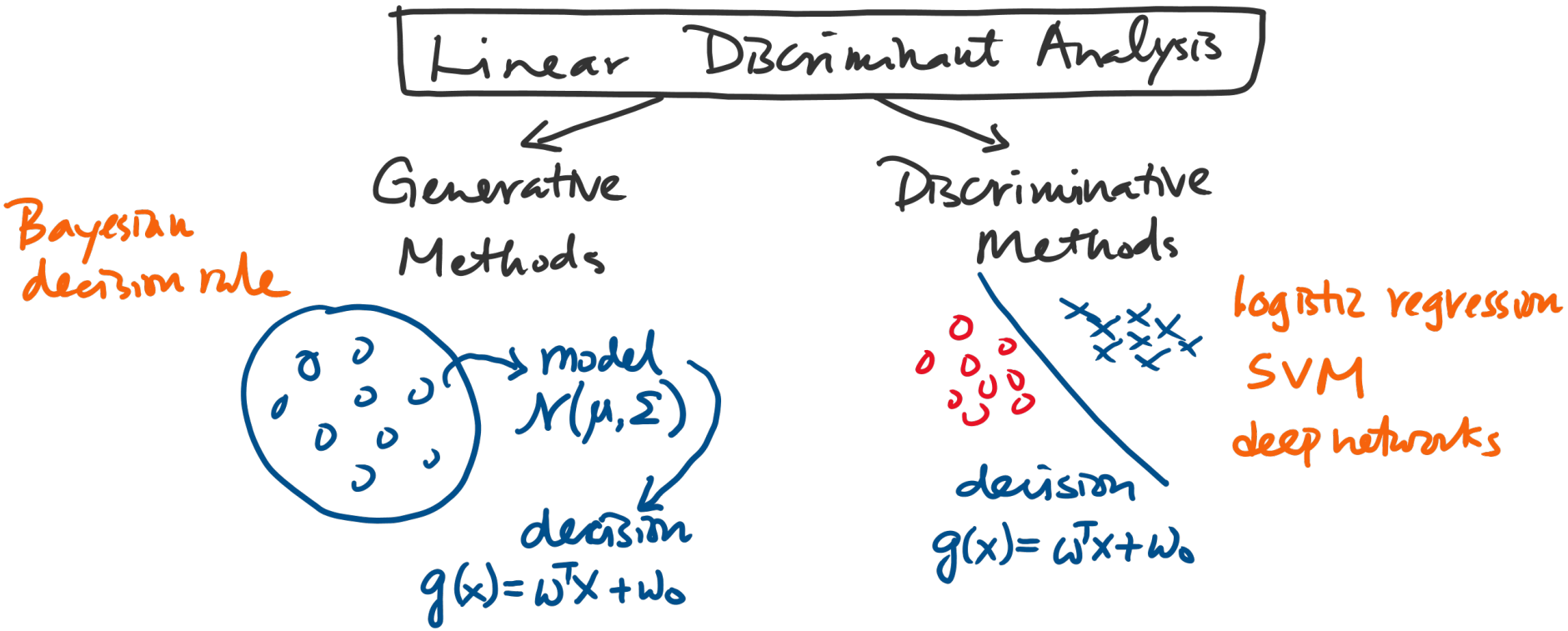
Spring 2020

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

Outline

Generative Approaches

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

Today's Lecture

- The Three Cases
  - $\boldsymbol{\Sigma}_i = \sigma^2 \boldsymbol{I}$ (Last Lecture)
  - $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$
  - General $\boldsymbol{\Sigma}_i$
- Evaluating Performance
  - Probability of Error
  - Bayesian Decision Rule is also Minimum Error Rule
  - ROC Curve

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

Last Lecture

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

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

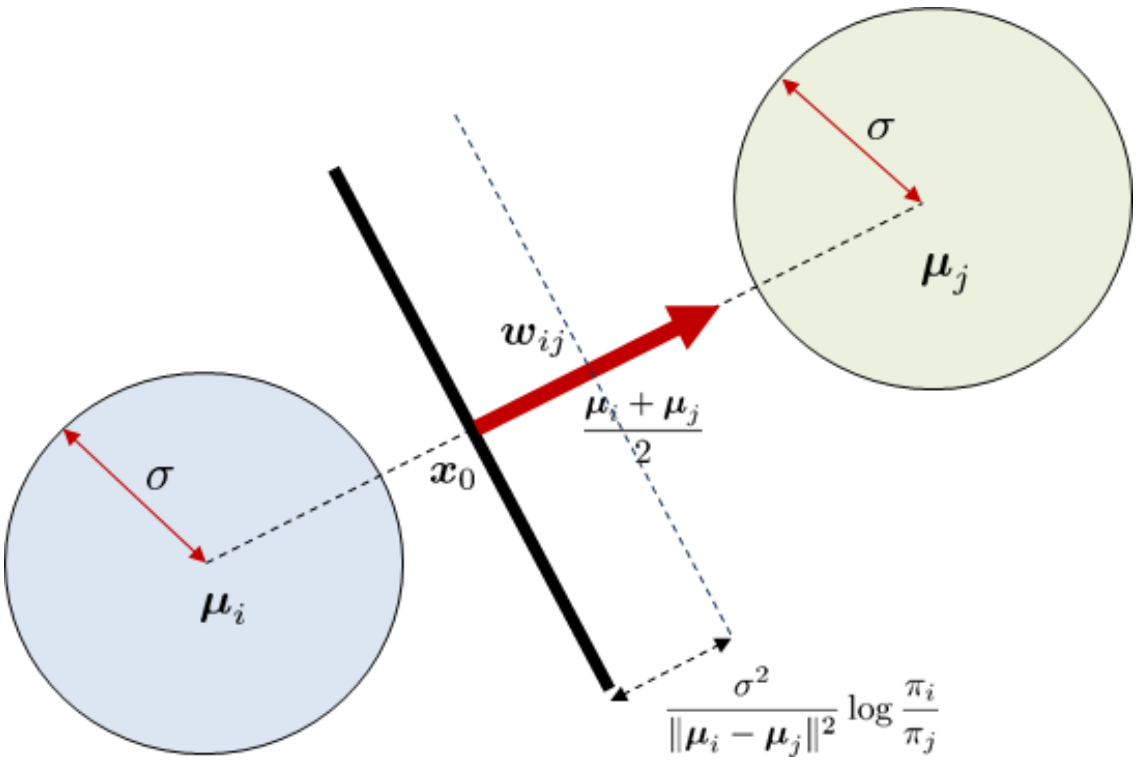
- You tell me the two Gaussians: $\boldsymbol{\mu}_i$, $\boldsymbol{\mu}_j$, π_i, π_j, σ
- I return you a separating hyperplane

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

- This is the best possible hyperplane according to posterior distribution

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

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



Case 2: $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$

Make all $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$.

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

In this case, we can show that

$$\begin{aligned} g_i(\boldsymbol{x}) &= -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i) - \cancel{\frac{1}{2}\log|\boldsymbol{\Sigma}|} + \log \pi_i \\ &= -\frac{1}{2}\left(\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{x} - 2\boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i\right) + \log \pi_i \\ &= -\frac{1}{2}\left(\cancel{\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{x}} - 2\boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i\right) + \log \pi_i \\ &= \underbrace{\boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1}\boldsymbol{x}}_{\boldsymbol{w}_i^T \boldsymbol{x}} \quad \underbrace{-\frac{1}{2}\boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \log \pi_i}_{w_{i0}} \end{aligned}$$

Case 2: $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$

The discriminant function is therefore:

$$\begin{aligned} g(\boldsymbol{x}) &= g_i(\boldsymbol{x}) - g_j(\boldsymbol{x}) \\ &= (\boldsymbol{w}_i - \boldsymbol{w}_j)^T \boldsymbol{x} + (w_{i0} - w_{j0}) \\ &= \underbrace{\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\right]^T \boldsymbol{x}}_{\boldsymbol{w}^T\boldsymbol{x}} \underbrace{-\frac{1}{2}\left(\boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i - \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j\right) + \log\frac{\pi_i}{\pi_j}}_{w_0}. \end{aligned}$$

If we want to write $g(\boldsymbol{x}) = \boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{x}_0)$, we can show that

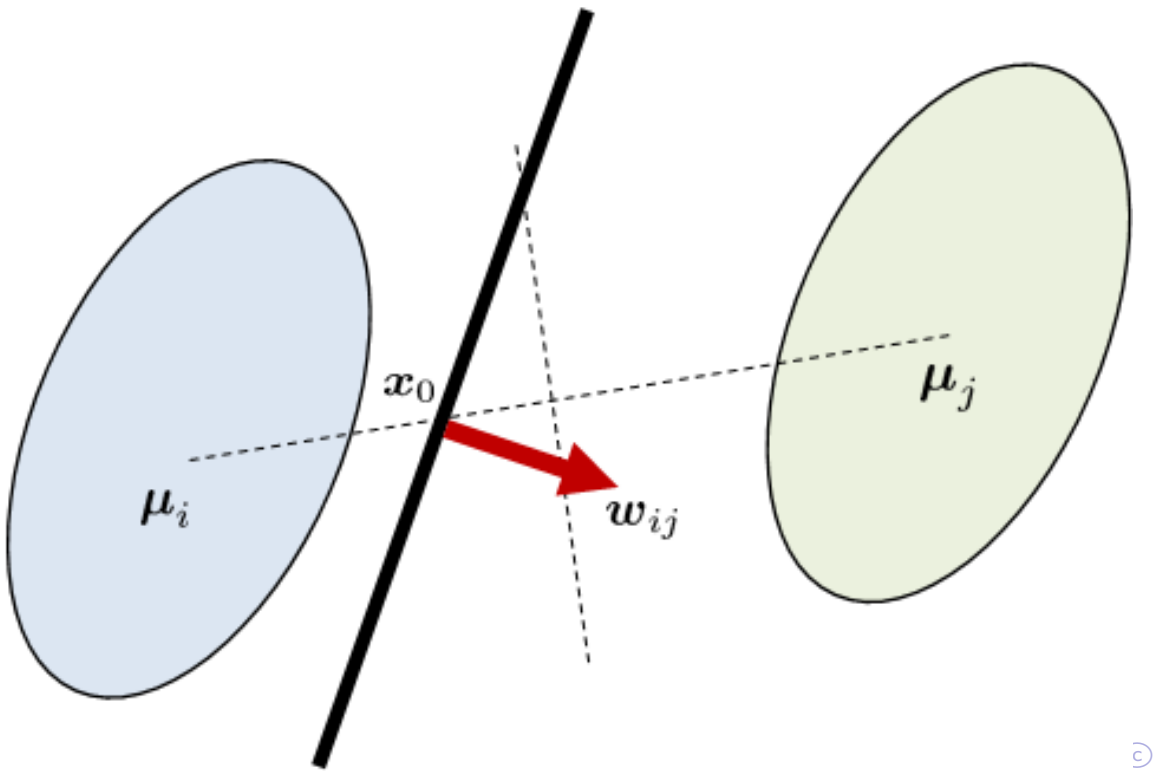
$$\begin{aligned} \boldsymbol{w} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j), \\ \boldsymbol{x}_0 &= \frac{\boldsymbol{\mu}_i + \boldsymbol{\mu}_j}{2} - \frac{\log\frac{\pi_i}{\pi_j}}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j). \end{aligned}$$

Case 2: $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$: Geometry

The discriminant function is $g(\boldsymbol{x}) = \boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{x}_0)$, with

$$\boldsymbol{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j),$$

$$\boldsymbol{x}_0 = \frac{\boldsymbol{\mu}_i + \boldsymbol{\mu}_j}{2} - \frac{\log \frac{\pi_i}{\pi_j}}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j).$$



Case 3: Arbitrary $\boldsymbol{\Sigma}_i$

This is the most general setting.

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

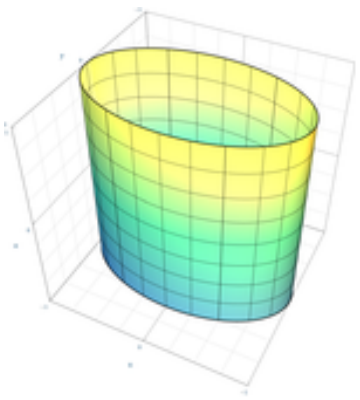
We can show that

$$\begin{aligned} g_i(\boldsymbol{x}) &= \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) + \frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \log\pi_i \\ &= \underbrace{\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{x}}_{\frac{1}{2}\boldsymbol{x}^T\boldsymbol{W}_i\boldsymbol{x}} \underbrace{-\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{x}}_{\boldsymbol{w}_i^T\boldsymbol{x}} \underbrace{+\frac{1}{2}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i + \frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \log\pi_i}_{w_{i0}}. \end{aligned}$$

Therefore, the discriminant function is

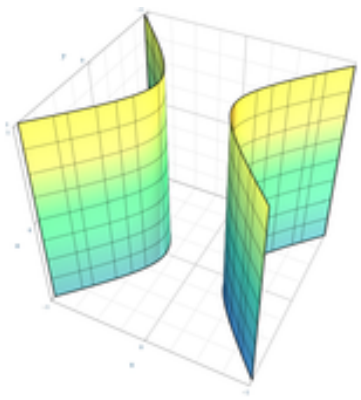
$$\begin{aligned} g(\boldsymbol{x}) &= g_i(\boldsymbol{x}) - g_j(\boldsymbol{x}) \\ &= \frac{1}{2}\boldsymbol{x}^T(\boldsymbol{W}_i - \boldsymbol{W}_j)\boldsymbol{x} + (\boldsymbol{w}_i - \boldsymbol{w}_j)^T\boldsymbol{x} + (w_{i0} - w_{j0}). \end{aligned}$$

Case 3: Arbitrary $\boldsymbol{\Sigma}_i$: Geometry



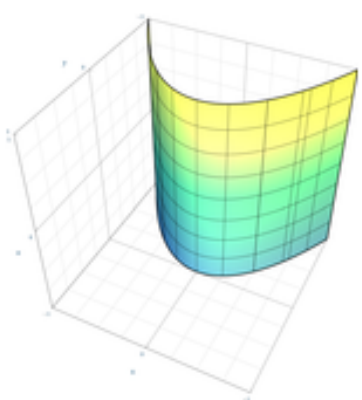
$$\boldsymbol{W} = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$\boldsymbol{w} = 0$



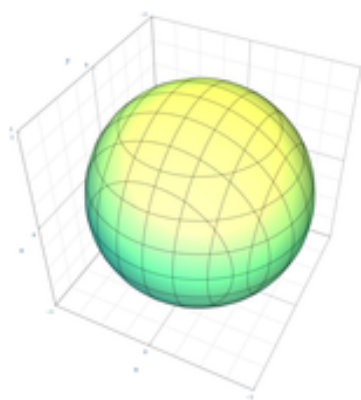
$$\boldsymbol{W} = \begin{bmatrix} a & 0 & 0 \\ 0 & -b & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$\boldsymbol{w} = 0$



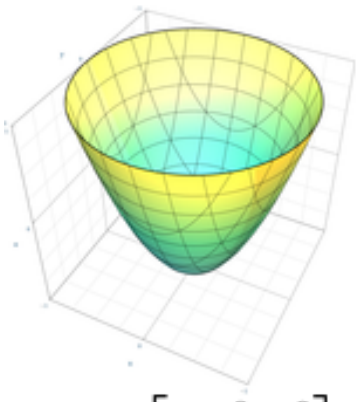
$$\boldsymbol{W} = \begin{bmatrix} a & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$\boldsymbol{w} \neq 0$



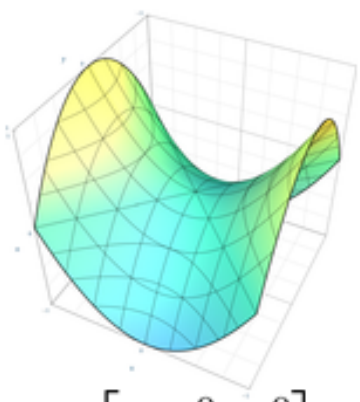
$$\boldsymbol{W} = \begin{bmatrix} a & 0 & 0 \\ 0 & a & 0 \\ 0 & 0 & a \end{bmatrix}$$

$\boldsymbol{w} = 0$



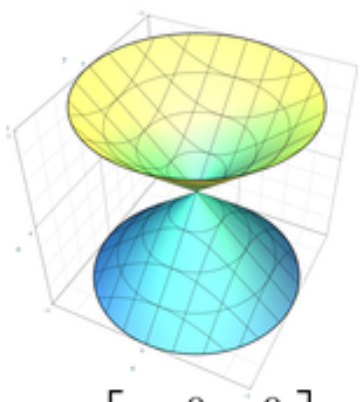
$$\boldsymbol{W} = \begin{bmatrix} a & 0 & 0 \\ 0 & a & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$\boldsymbol{w} \neq 0$



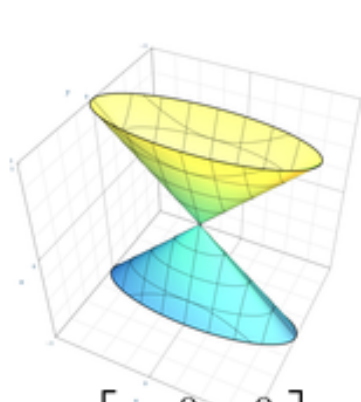
$$\boldsymbol{W} = \begin{bmatrix} a & 0 & 0 \\ 0 & -b & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$\boldsymbol{w} \neq 0$



$$\boldsymbol{W} = \begin{bmatrix} a & 0 & 0 \\ 0 & a & 0 \\ 0 & 0 & -b \end{bmatrix}$$

$\boldsymbol{w} = 0$



$$\boldsymbol{W} = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & -c \end{bmatrix}$$

$\boldsymbol{w} = 0$

Case 3: Arbitrary $\boldsymbol{\Sigma}_i$: Computing

Quadratic classifier is not a big deal, on computer!

$$\begin{aligned} g(\boldsymbol{x}) &= g_i(\boldsymbol{x}) - g_j(\boldsymbol{x}) \\ &= \frac{1}{2}\boldsymbol{x}^T(\boldsymbol{W}_i - \boldsymbol{W}_j)\boldsymbol{x} + (\boldsymbol{w}_i - \boldsymbol{w}_j)^T\boldsymbol{x} + (w_{i0} - w_{j0}). \end{aligned}$$

Recall: A hypothesis function is

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

- All you need is to check whether $g(\boldsymbol{x}) > 0$ or $g(\boldsymbol{x}) < 0$.
- No need to obtain a closed form solution.

Outline

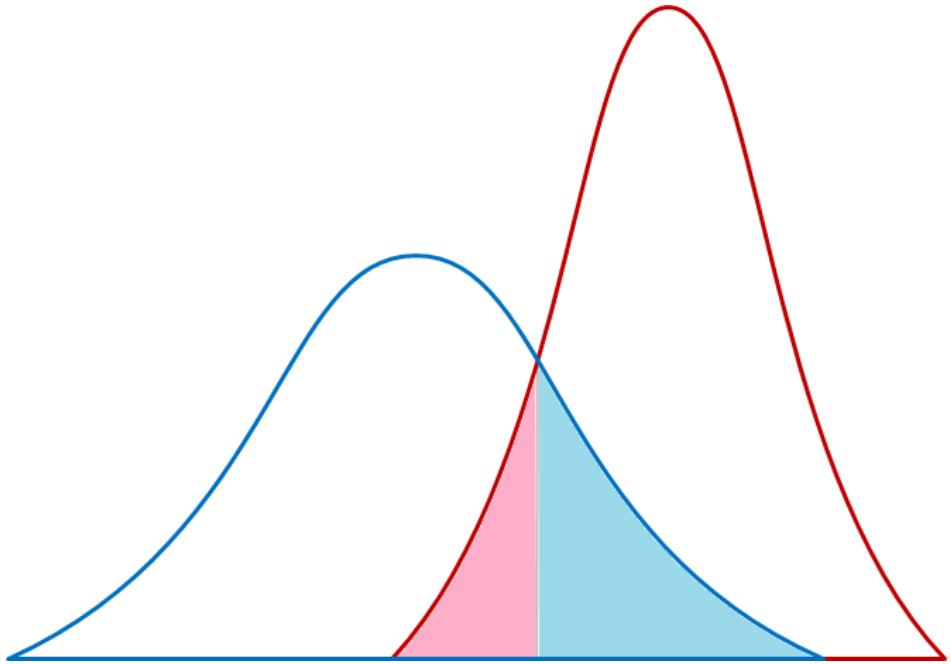
Generative Approaches

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

Today's Lecture

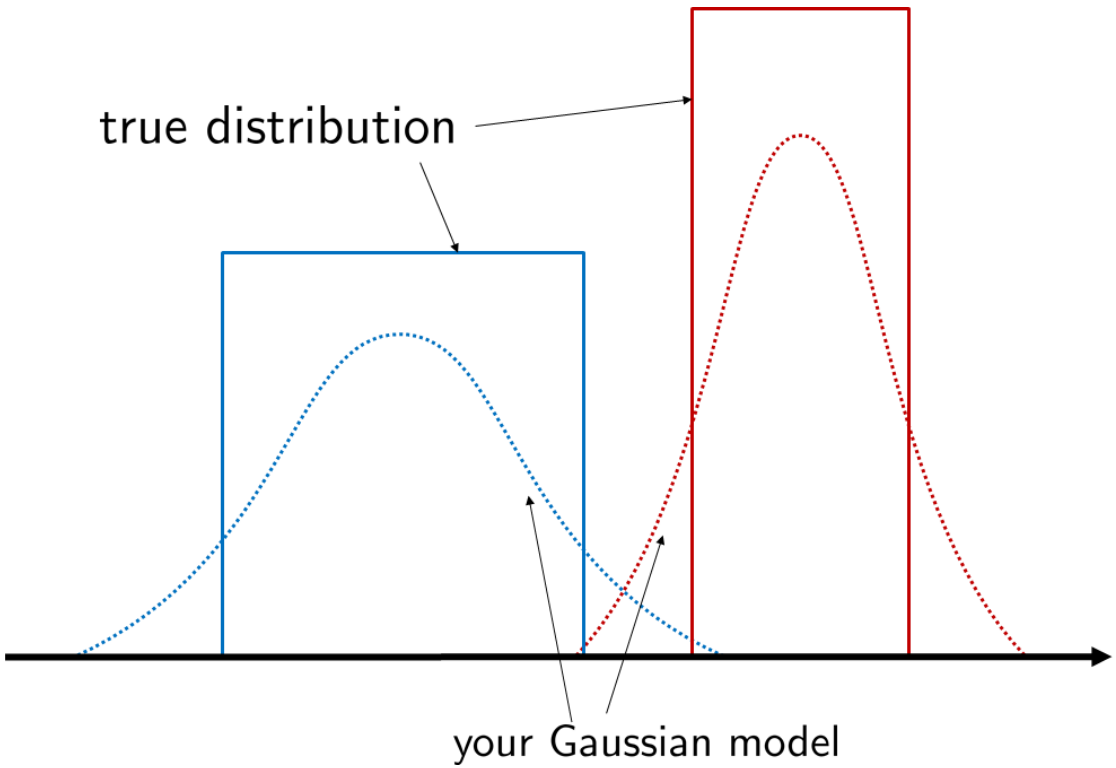
- The Three Cases
  - $\boldsymbol{\Sigma}_i = \sigma^2 \boldsymbol{I}$ (Last Lecture)
  - $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$
  - General $\boldsymbol{\Sigma}_i$
- Evaluating Performance
  - Probability of Error
  - Bayesian Decision Rule is also Minimum Error Rule
  - ROC Curve

How Well does Bayesian Decision Rule Perform?



- Gaussian is a model.
- If the underlying true distribution is Gaussian, then there is always overlapping.
- No matter how you do, there is always classification error.

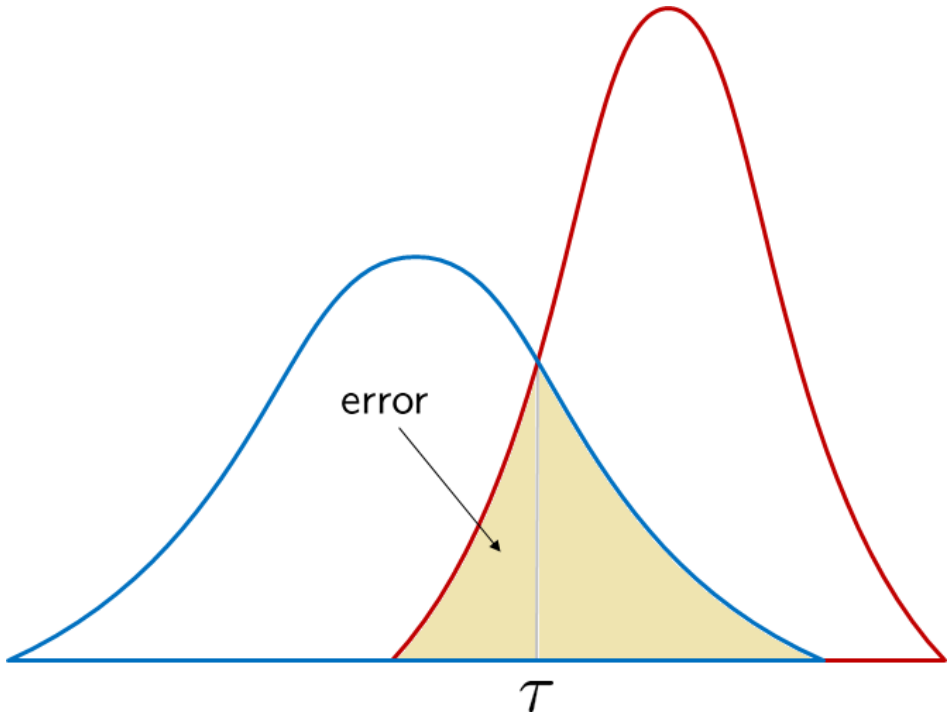
What if the distributions do not match



- If the true distribution is not overlapping, then the overlapping of the Gaussian is not a problem.
- In this lecture we assume the true distribution is indeed Gaussian.

Probability of Error

Suppose we have found a decision boundary, which is a threshold τ in 1D.



Probability of error is

$$P_e = \mathbb{P}[X < \tau \textbf{ and } Y = 1] + \mathbb{P}[X > \tau \textbf{ and } Y = 0]$$

Minimize the Probability of Error

Let us do some calculation:

$$\begin{aligned}
P_e &= \mathbb{P}[X < \tau \text{ and } Y = 1] + \mathbb{P}[X > \tau \text{ and } Y = 0] \\
&= \mathbb{P}[X < \tau \mid Y = 1]\mathbb{P}[Y = 1] + \mathbb{P}[X > \tau \mid Y = 0]\mathbb{P}[Y = 0] \\
&= \pi_1 \mathcal{N}(X < \tau \mid \mathcal{C}_1) + \pi_0 \mathcal{N}(X > \tau \mid \mathcal{C}_0) \\
&= \pi_1 \int_{-\infty}^{\tau} \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} dx + \pi_0 \int_{\tau}^{\infty} \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}} dx
\end{aligned}$$

Can we minimize the P_e by finding a good τ?

$$\begin{aligned}
\frac{d}{d\tau} P_e = \pi_1 \frac{d}{d\tau} \left(\int_{-\infty}^{\tau} \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} dx \right) \\
+ \pi_0 \frac{d}{d\tau} \left(\int_{\tau}^{\infty} \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}} dx \right)
\end{aligned}$$

Minimize the Classification Error

Fundamental Theorem of Calculus:

$$\frac{d}{d\tau}\int_{-\infty}^{\tau} f(x)dx = f(\tau)$$

Therefore, in our problem,

$$\frac{d}{d\tau}\left(\int_{-\infty}^{\tau} \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} dx\right) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(\tau-\mu_1)^2}{2\sigma_1^2}}$$

$$\frac{d}{d\tau}\left(\int_{\tau}^{\infty} \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}} dx\right) = -\frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(\tau-\mu_0)^2}{2\sigma_0^2}}$$

Special Case: $\sigma_1 = \sigma_0 = \sigma$

Let us assume that $\sigma_1 = \sigma_0 = \sigma$. Then,

$$\frac{d}{d\tau}P_e = \pi_1 \frac{d}{d\tau}\int_{-\infty}^{\tau} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} dx + \pi_0 \frac{d}{d\tau}\int_{\tau}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_0)^2}{2\sigma^2}} dx$$

$$= \pi_1 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\tau-\mu_1)^2}{2\sigma^2}} - \pi_0 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\tau-\mu_0)^2}{2\sigma^2}} = 0$$

So finally, we can equate the two sides, and get

$$\pi_1 \cancel{\frac{1}{\sqrt{2\pi\sigma^2}}} e^{-\frac{(\tau-\mu_1)^2}{2\sigma^2}} = \pi_0 \cancel{\frac{1}{\sqrt{2\pi\sigma^2}}} e^{-\frac{(\tau-\mu_0)^2}{2\sigma^2}}$$

$$\log\left\{\pi_1 e^{-\frac{(\tau-\mu_1)^2}{2\sigma^2}}\right\} = \log\left\{\pi_0 e^{-\frac{(\tau-\mu_0)^2}{2\sigma^2}}\right\}$$

$$\tau = \frac{\mu_1 - \mu_0}{2} - \frac{\sigma^2}{\mu_1 - \mu_0}\log\frac{\pi_1}{\pi_0}.$$

Optimality of Bayesian Decision Rule

Therefore,

$$\tau = \frac{\mu_1 - \mu_0}{2} - \frac{\sigma^2}{\mu_1 - \mu_0} \log \frac{\pi_1}{\pi_0},$$

and this means that

$$h(x) = \begin{cases} 1, & \text{if } x > \frac{\mu_1-\mu_0}{2} - \frac{\sigma^2}{\mu_1-\mu_0} \log \frac{\pi_1}{\pi_0}, \\ 0, & \text{if } x < \frac{\mu_1-\mu_0}{2} - \frac{\sigma^2}{\mu_1-\mu_0} \log \frac{\pi_1}{\pi_0}, \\ \text{either}, & \text{if } x = \frac{\mu_1-\mu_0}{2} - \frac{\sigma^2}{\mu_1-\mu_0} \log \frac{\pi_1}{\pi_0}. \end{cases}$$

This is the **exact same result** as the Bayesian decision rule.

- Bayesian decision is optimal in posterior
- Bayesian decision is optimal in probability of error
- Optimal $\neq$ error free
- You still have error if you have infinite training data

Two Levels of Optimality

Error due to **Finite Training Samples**

- You need to estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$
- More samples, better estimate
- Controlled by Hoeffding inequality

Error due to **Linear Non-Separability**

- Error even if you have infinite training data
- Fundamental limit of the model
- Gaussians are not linearly separable
- Minimize error probability $\neq$ zero error

Risk and Probability of Error

Is Your "Optimal" Really Optimal?

$$P_e = \mathbb{P}[X < \tau \mid Y = 1]\pi_1 + \mathbb{P}[X > \tau \mid Y = 0]\pi_0.$$

- $Y = 1$: disease present. $Y = 0$: normal.
- $X < \tau$: report no disease. $X > \tau$: report disease.
- $\mathbb{P}[X < \tau \mid Y = 1]$: There is a disease, but you cannot find it.
- $\mathbb{P}[X > \tau \mid Y = 0]$: There is no disease, but you say there is.
- Which one is more serious?
- We need to define **Risk**

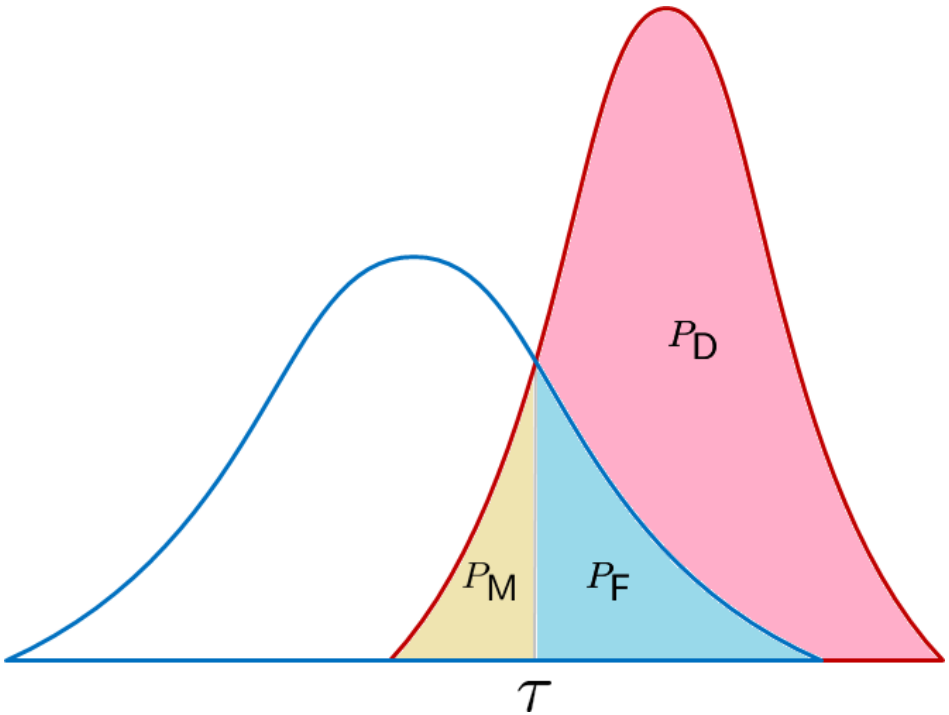
$$\underbrace{R}_{\text{risk}} = C_{01}\underbrace{\mathbb{P}[X < \tau \mid Y = 1]}_{\text{miss}}\pi_1 + C_{10}\underbrace{\mathbb{P}[X > \tau \mid Y = 0]}_{\text{false alarm}}\pi_0.$$

- C_{01}: cost of miss. C_{10}: cost of false alarm.

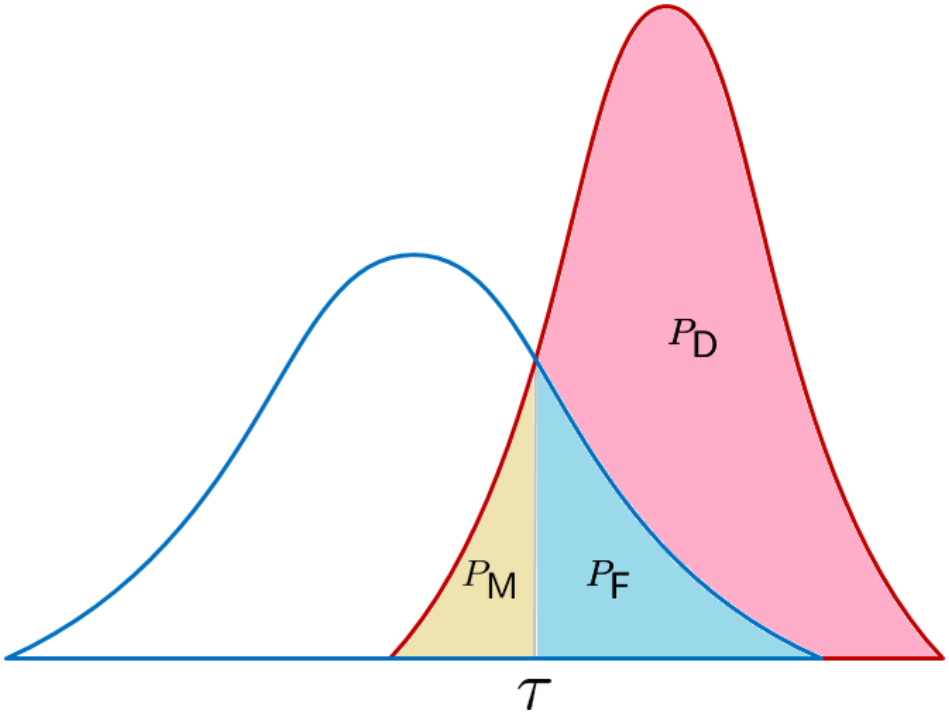
False Alarm and Detection

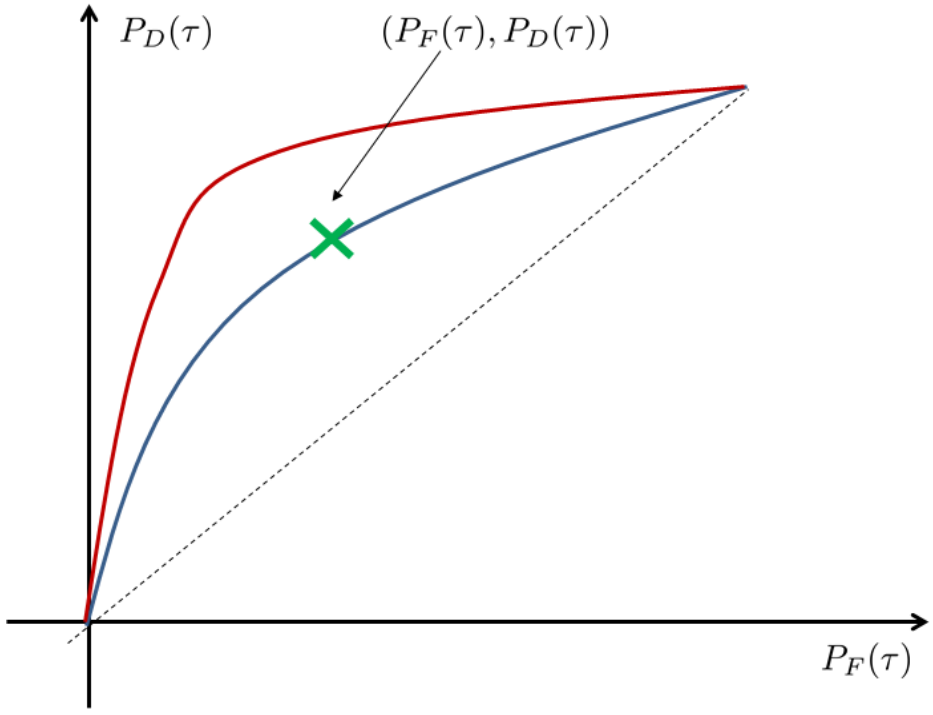
$$P_e = C_{01}\underbrace{\mathbb{P}[X < \tau \mid Y = 1]}_{P_M}\pi_1 + C_{10}\underbrace{\mathbb{P}[X > \tau \mid Y = 0]}_{P_F}\pi_0.$$

- P_M: Probability of Miss
- $P_D = 1 - P_M$: Probability of Detection
- P_F: Probability of False Alarm



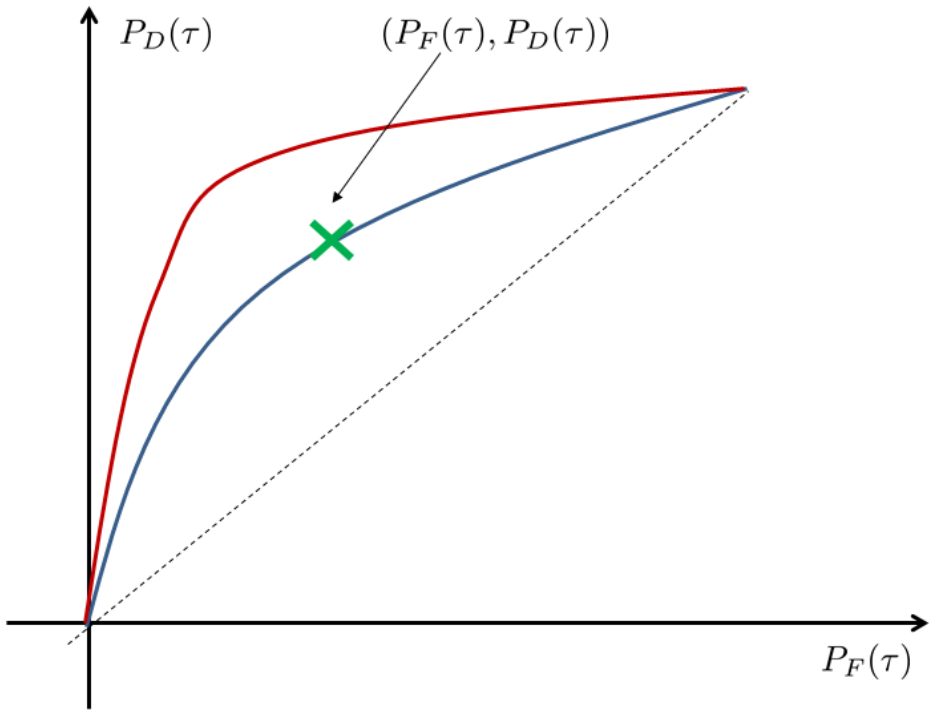
Receive Operating Characteristic (ROC) Curve





- They are all functions of τ.
- If you change τ, they will change correspondingly.
- Can plot $P_F(\tau)$ and $P_D(\tau)$ as τ changes.
- This called a Receive Operating Characteristic (ROC) curve.

Interpreting the ROC



- Higher ROC is better: Same level of false alarm, higher detection rate
- One classifier has one ROC. One ROC curve for all τ.
- Ways to improve ROC:
  - Train with more sample so that you have less training error
  - Find a better model than Gaussian
  - Find a better decision rule than linear classifier

Reading List

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

Probability of Error

- Duda, Hart and Stork's *Pattern Classification*, Chapter 2.7, 3.1.
- Poor, *Intro to Signal Estimation and Detection*, Chapter 2.

ROC Curve

- ECE645 Note. https://engineering.purdue.edu/ChanGroup/ECE645Notes/StudentLecture02.pdf

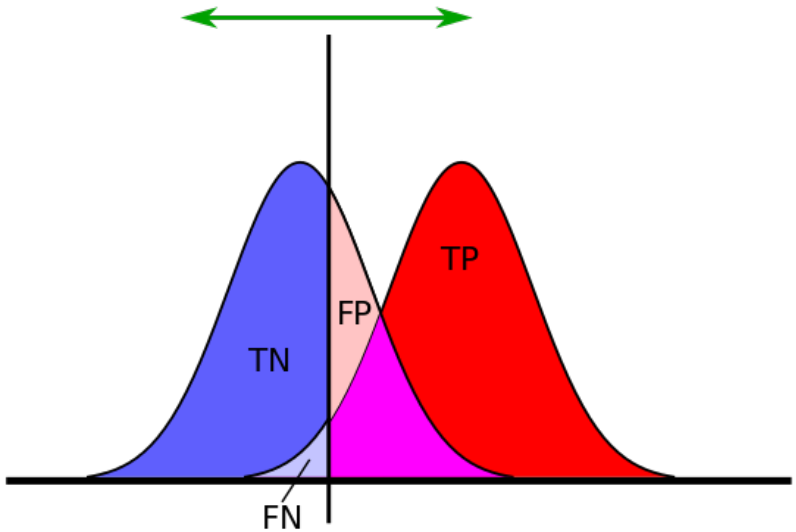
Appendix

Q&A 1 Is ROC curve always concave?

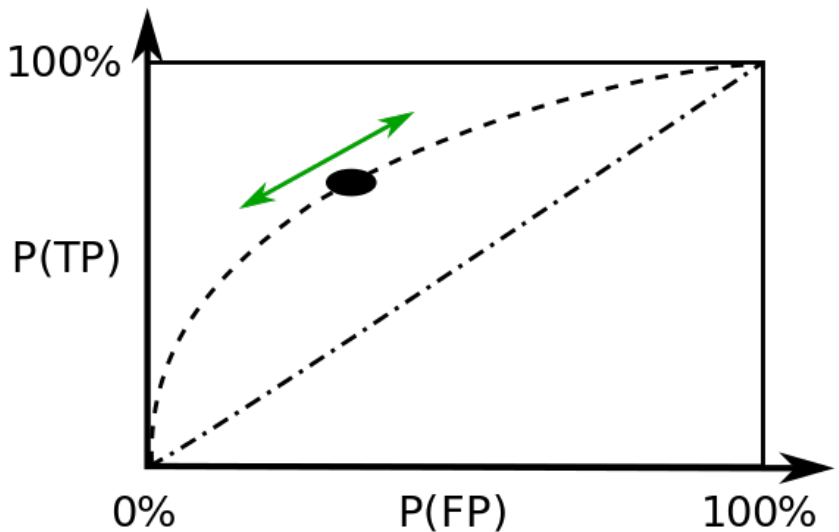
- No.
- The ROC curve is concave when the decision rule comes from Neyman-Pearson. See V. Poor's Intro to Signal Estimation and Detection Chapter 2.
- You can always create a decision rule that performs terribly on one part and very well on the other part.
- Most algorithms will generate ROC curves that shows stair-case effects.
- It is important to report the entire ROC curve.

Q&A 2 Type 1 and Type 2 Error

You probably have heard these two terms before. Below is an image I found from Wikipedia



TP	FP
FN	TN



Q&A 2 Type 1 and Type 2 Error

- A typical naming issue between statistics and engineering
- False Positive = Type 1 Error = False Alarm
- False Negative = Type 2 Error = Miss
- Both are **probabilities**
- In the previous figure, we see that a decision rule is the location of the black line.
- If you move the black line left and right, you will get a point on the ROC curve.

Q&A 3 What is the difference between ROC and Precision-Recall Curve (PRC)?

- ROC: Receiver Operating Curve.
  - Detection: No. TP / (No. TP + No. FN)
  - False Alarm: No. FP / (No. FP + No. TN)
  - ROC Curve always goes up.
- PRC: Precision Recall Curve.
  - Precision: No. TP / (No. TP + No. FP)
  - Recall: No. TP / (No. TP + No. FN)
  - PRC Curve always goes down.
- There is generally no definitive answer to which of ROC or PRC is better. However, according to the following paper, ROC is better for datasets with balanced classes whereas PRC is better for skewed classes. https://www.biostat.wisc.edu/~page/rocpr.pdf