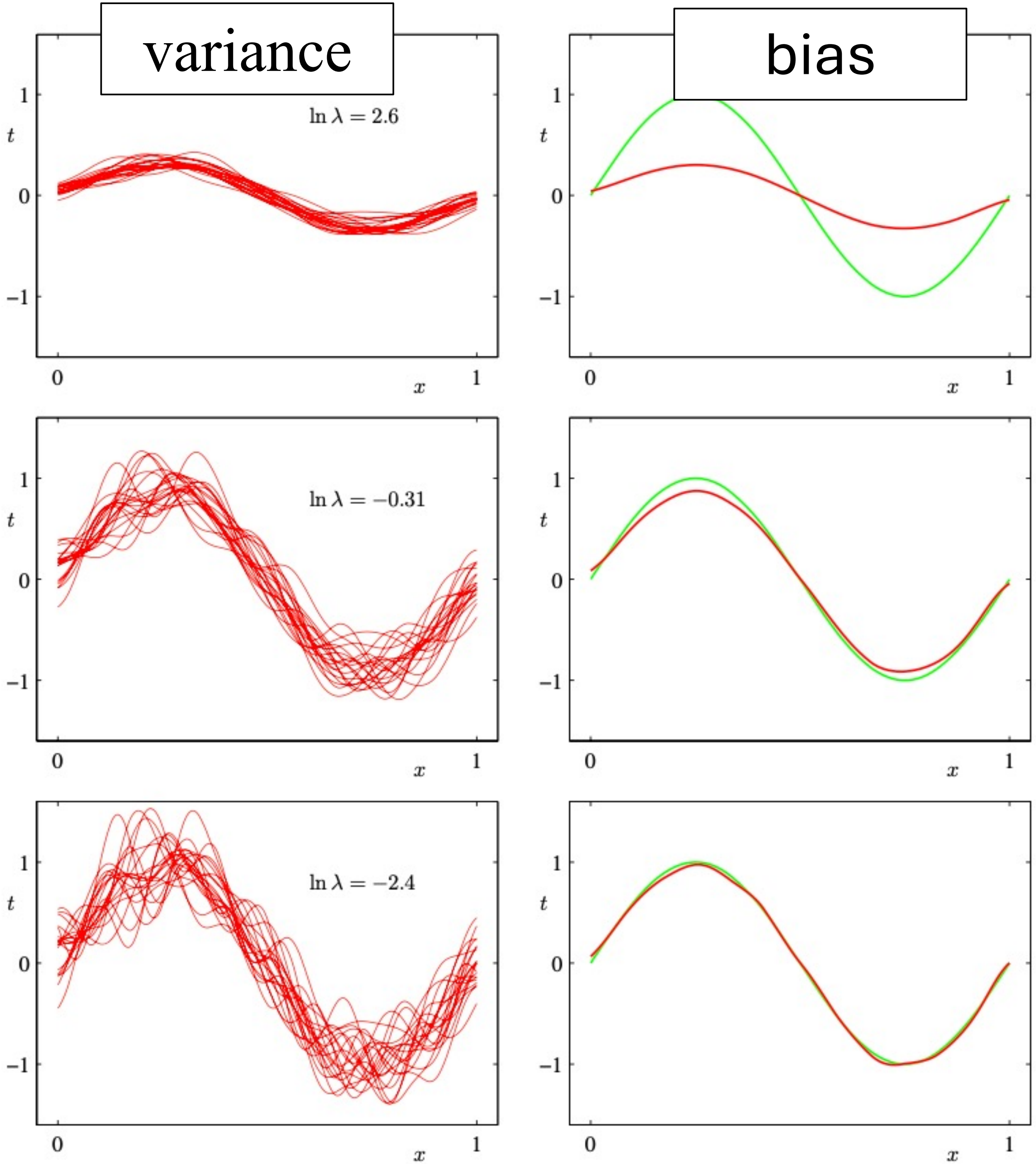
CS 461: Machine Learning Principles

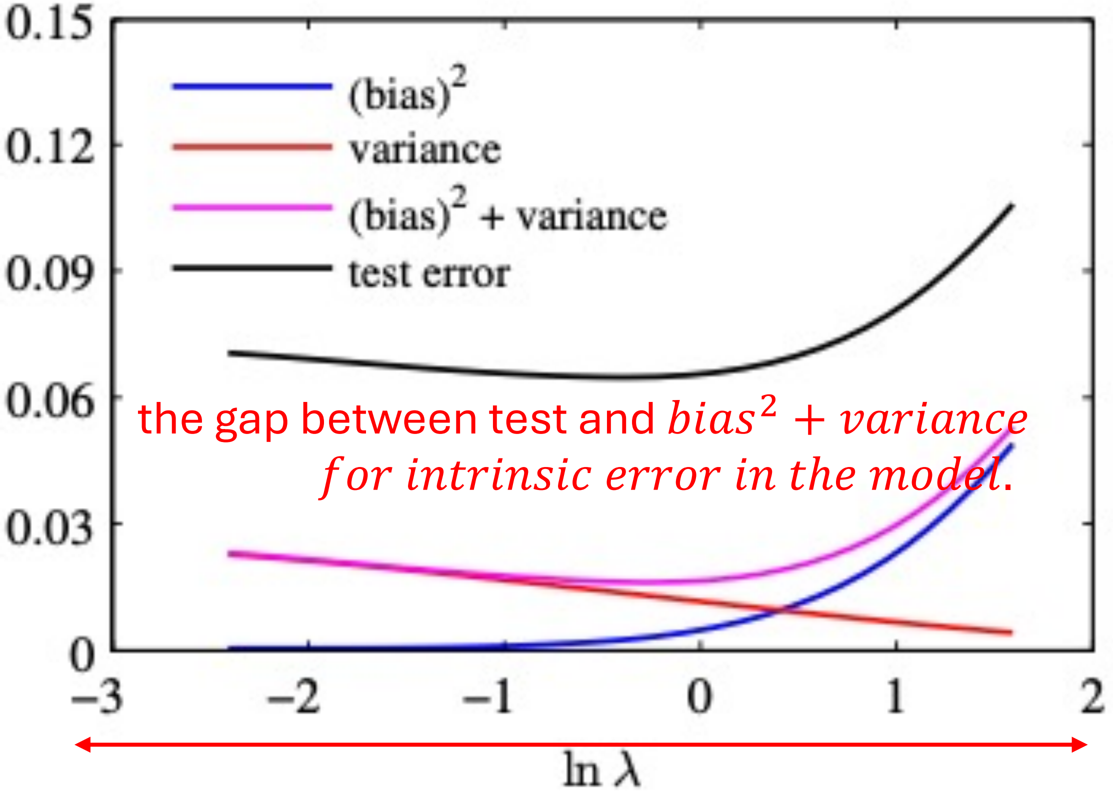
Class 8: Sept. 30

Binary Classification: **L**inear **D**iscriminant **A**nalysis (**LDA**)

Instructor: Diana Kim



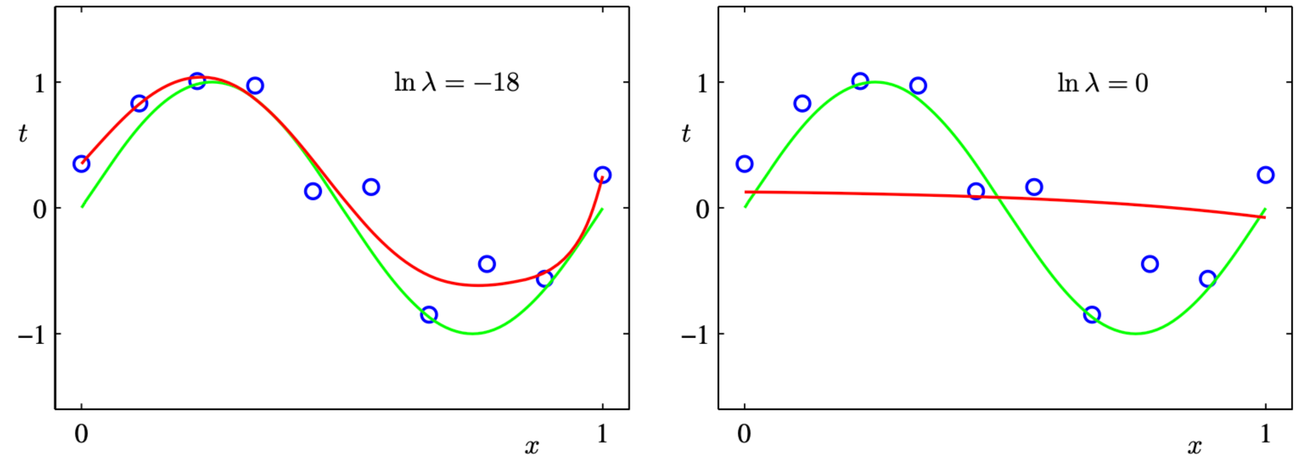
Trade off between Variance and Bias according to λ in Ridge Regression (Complexity)

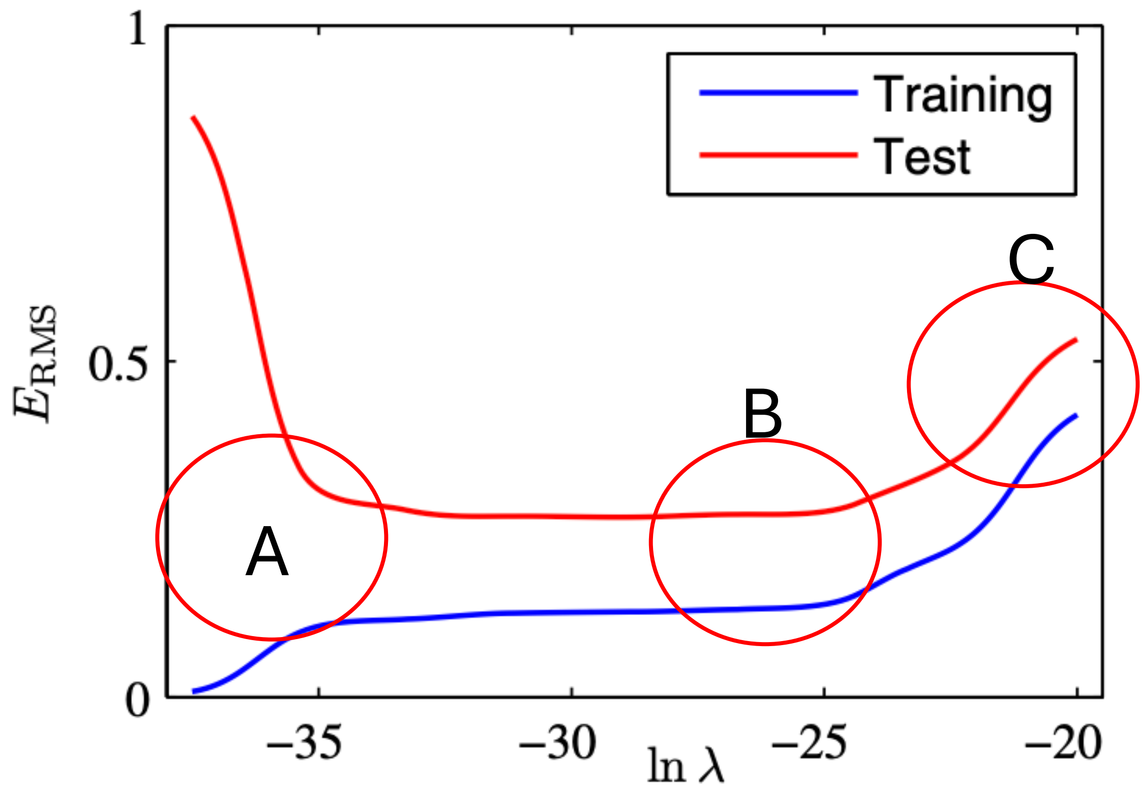


High Complexity

Low Complexity

Overfitting can be avoided by adopting Bayesian Approach





- Ridge regression regulates complexity.
- We need to choose regularization parameter:
- Q: which λ *would you choose A or B*?

+ In the lecture, I mentioned that the performance gap between train a
  so the answer was B instead of A (both test performance are similar)
+ But I realized that this may not be true (we only consider test error).
  Let me revisit this problem again and open for discussion in the next lecture.
  We need to choose the lambda showing the lowest **test** performance.

Classification Problem

How can we solve a classification problem?
How can we assign an input x to a certain class C_k?

- Regression: Learning $y = f(x)$ “$the\ functional\ relation\ beween\ x\ and\ y$
- Classification: Learning the discriminant functions $f_k(x)$

$$C_k = \arg\max_k f_k(x)$$

: by using discriminant functions, we can assign a class to x

To classify inputs to three groups,
How many discriminant functions do we need?

Example) three discriminant functions in the feature space of $\mathbb{R}^2$

1. $f_1(x_1, x_2) = x_2 - x_1 - 1$
2. $f_2(x_1, x_2) = x_2 + x_1 - 1$
3. $f_3(x_1, x_2) = x_2$

$$C_k = \arg\max_k f_k(x)$$

Q: Based on the discriminant functions above, assign a class for the points?

	f_1	f_2	f_3	class inference
$(-2,0)$	1	-3	0	1
$(0,0)$	-1	-1	0	3
$(2,0)$	-3	1	0	2

Example) computing the decision boundaries/ Splitting the feature space

1. $f_1(x_1, x_2) = x_2 - x_1 - 1$
2. $f_2(x_1, x_2) = x_2 + x_1 - 1$
3. $f_3(x_1, x_2) = x_2$

$$C_k = \arg\max_k f_k(x)$$

Q: Based on the discriminant functions above, assign a class for the points?

	f_1	f_2	f_3
$(-2,0)$	1	-3	0
$(0,0)$	-1	-1	0
$(2,0)$	-3	1	0

Example) computing the decision regions

1. $f_1(x_1, x_2) = x_2 - x_1 - 1$
2. $f_2(x_1, x_2) = x_2 + x_1 - 1$
3. $f_3(x_1, x_2) = x_2$

- $R_1 : x_1 < -1$
- $R_2 : x_1 \geq 1$
- $R_3 : -1 \leq x_1 < 1$

Today, we are going to focus on **binary classification**. (+ / −)
To classify inputs by two groups,
How many discriminant functions do we need?

y → cat
Dog
f(x | c)
x

x → { 0, 1 }

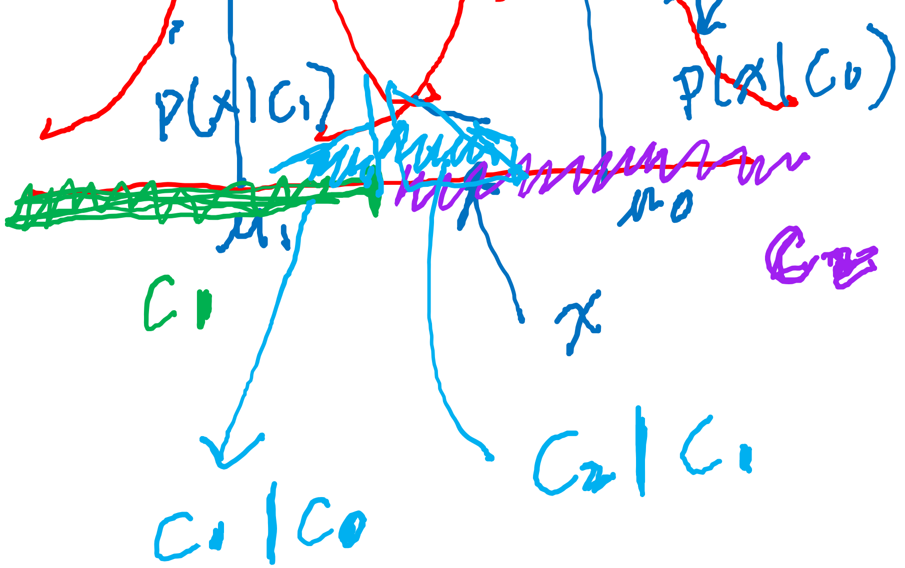
Today, we are going to focus on binary classification.
To classify inputs by two groups,
How many discriminant functions do we need?

We want to design the two discriminant functions $f_0(\vec{x})$ and $f_1(\vec{x})$.
Once we have them, we can decide/inference the class for feature input $\vec{x}$.

classification can be represented hypothesis problem
X is generated from which conditional density?

$$\begin{cases} \mathcal{H}_0 : & \vec{x} \sim f(\vec{x}|C_0) \\ \mathcal{H}_1 : & \vec{x} \sim f(\vec{x}|C_1) \end{cases}$$

$$f_0(\vec{x}) \underset{H_1}{\overset{H_0}{\gtreqless}} f_1(\vec{x})$$



Binary Classification Problem & Decision Regions / Boundaries

Binary Classification Problem

$$\begin{cases} \mathcal{H}_0 : & \vec{x} \sim f(\vec{x}|C_0) \\ \mathcal{H}_1 : & \vec{x} \sim f(\vec{x}|C_1) \end{cases}$$

we need to learn these discriminant functions.

$$f_0(\vec{x}) \underset{H_1}{\overset{H_0}{\gtrless}} f_1(\vec{x})$$

Decision Regions and Decision Boundaries

- $R_0 : f_0(\vec{x}) \geq f_1(\vec{x})$
- $R_1 : f_0(\vec{x}) < f_1(\vec{x})$

$$f_1(\vec{x}) \overset{?}{=} f_0(\vec{x})$$

$$\boxed{f_1(\vec{x}) - f_0(\vec{x}) = 0}$$

Hyperplane in the feature space

Q: dimension?

- MAP is a possible way
  forming the discriminant functions.

$P(H_0 | x)$

$P(H_1 | x)$

$P[H_0 \text{ True}]$

$$\overbrace{\underbrace{p(x|H_0)}_{\text{likelihood}}\,\underbrace{\pi_0}_{\text{prior}}}^{\text{Posterior}} \underset{H_1}{\overset{H_0}{\gtrless}} p(x|H_1)\,\pi_1$$

$f_0(x)$

$f_1(x)$

- MAP Rule Minimizes Expected Classification Error E[R]

$$E[R] = \pi_0 \cdot \underbrace{E[R|H_0]} + \pi_1 E[R|H_1] = \pi_1 \cdot \int_{y_0} f(y|H_1)dy + \pi_0 \cdot \int_{y_1} f(y|H_0)dy$$

P[H0 = True]

$$= \pi_1 \cdot \int_{y_0} f(y|H_1)dy + \pi_0 \cdot \underline{(1 - \int_{y_0} f(y|H_0)dy)}$$

$$= \pi_0 + \int_{y_0} \pi_1 \cdot f(y|H_1) - \pi_0 \cdot f(y|H_0)dy$$

Q: How should we set the decision rule for y_0?

for y, if $\pi_1 \cdot f(y|H_1) - \pi_0 \cdot f(y|H_0) < 0$ then detect as H_0 R_0
else if $\pi_1 \cdot f(y|H_1) - \pi_0 \cdot f(y|H_0) \geq 0$ then detect as H_1 R_1

$$\frac{p(y|H_1)}{P(y|H_0)} \gtrless \frac{\pi_0}{\pi_1}$$

"MAP rule"

Gaussian **D**iscriminant **A**nalysis (**GDA**)

(μ, Σ)

$$\underbrace{p(x|H_0)\,\pi_0}_{p(H_0|x)} \underset{H_1}{\overset{H_0}{\gtreqless}} \underbrace{p(x|H_1)\,\pi_1}_{p(H_1|x)}$$

Gaussian!

Prior!

Q: Decision Boundaries for the Two Possible Cases

$x \in C_0 \rightarrow \Sigma_0$

$x \in C_1 \rightarrow \Sigma_1$

- $\Sigma_0 \neq \Sigma_1$
- ~~$\Sigma_0 = \Sigma_1$~~

$$\underbrace{\frac{1}{\sqrt{2\pi|\Sigma_0|}} \exp -\frac{1}{2}(x-\mu_0)^t \Sigma_0^{-1}(x-\mu_0) \cdot P[\mathcal{H}_0 = 0]}_{PC_1} \underset{H_1}{\overset{H_0}{\gtrless}} \underbrace{\frac{1}{\sqrt{2\pi|\Sigma_1|}} \exp -\frac{1}{2}(x-\mu_1)^t \Sigma_1^{-1}(x-\mu_1) \cdot P[\mathcal{H}_0 = 1]}_{PC_2}$$

$$\Updownarrow$$

$$\underbrace{\ln P[\mathcal{H}_0] + \ln \frac{1}{\sqrt{2\pi|\Sigma_0|}} - \frac{1}{2}(x-\mu_0)^t \Sigma_0^{-1}(x-\mu_0)}_{f_0(x)} \gtrless \underbrace{\ln P[\mathcal{H}_1] + \ln \frac{1}{\sqrt{2\pi|\Sigma_1|}} - \frac{1}{2}(x-\mu_1)^t \Sigma_1^{-1}(x-\mu_1)}_{f_1(x)}$$

- Quadratic discriminant functions!
- Quadratic decision boundary!

Q: Decision Boundaries for the Two Possible Cases

- $\Sigma_0 = \Sigma_1$

$$\frac{1}{\sqrt{2\pi|\Sigma_0|}}\exp -\frac{1}{2}(x-\mu_0)^t\Sigma_0^{-1}(x-\mu_0)\cdot P[\mathcal{H}_0=0] \underset{H_1}{\overset{H_0}{\gtrless}} \frac{1}{\sqrt{2\pi|\Sigma_1|}}\exp -\frac{1}{2}(x-\mu_1)^t\Sigma_1^{-1}(x-\mu_1)\cdot P[\mathcal{H}_0=1]$$

$\cancel{x^T\Sigma_0^{-1}x}$ $\cancel{x^T\Sigma_1^{-1}x}$

$\Updownarrow$

$$\ln P[\mathcal{H}_0] + \cancel{\ln\frac{1}{\sqrt{2\pi|\Sigma_0|}}} - \frac{1}{2}(x-\mu_0)^t\Sigma_0^{-1}(x-\mu_0) \gtrless \ln P[\mathcal{H}_1] + \cancel{\ln\frac{1}{\sqrt{2\pi|\Sigma_1|}}} - \frac{1}{2}(x-\mu_1)^t\Sigma_1^{-1}(x-\mu_1)$$

$\Updownarrow$

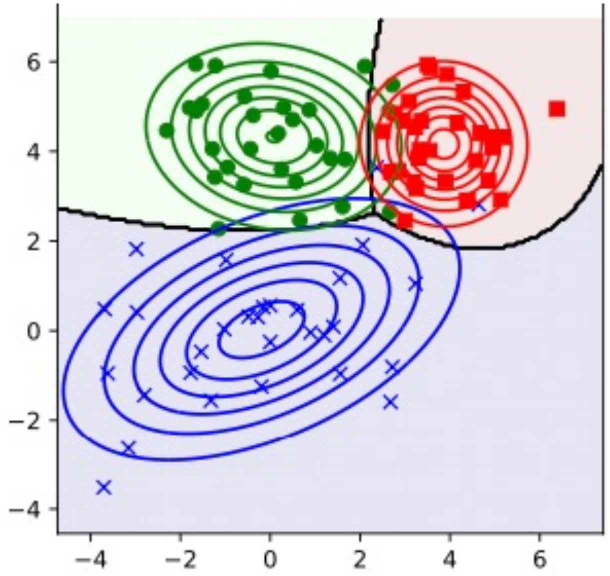
$$\ln P[\mathcal{H}_0] + \mu_0^t\Sigma_0^{-1}x - \frac{1}{2}\mu_0^t\Sigma_0^{-1}\mu_0^t \gtrless \ln P[\mathcal{H}_1] + \mu_1^t\Sigma_1^{-1}x - \frac{1}{2}\mu_1^t\Sigma_1^{-1}\mu_1^t$$

$g_0(x)$ $g_1(x)$

- Linear discriminant functions!
- Linear decision boundary!

From Murphy Figure 9.2

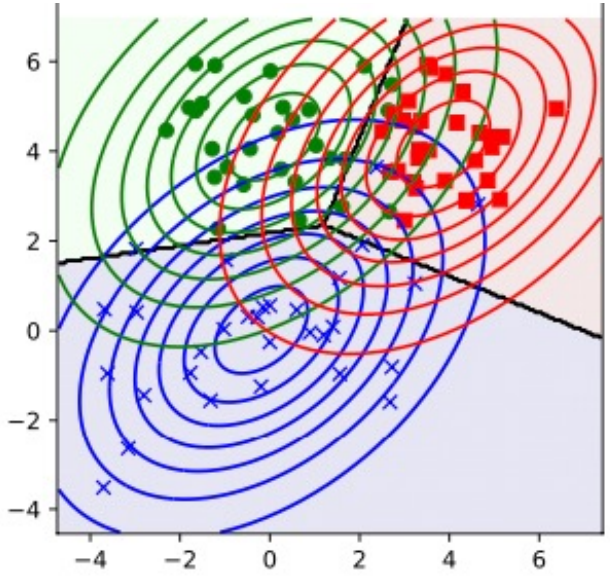
[Quadratic Discriminant Analysis]



$$\Sigma_0 \neq \Sigma_1$$

(Unconstrained Covariance)

[Linear Discriminant Analysis]



$$\Sigma_0 = \Sigma_1$$

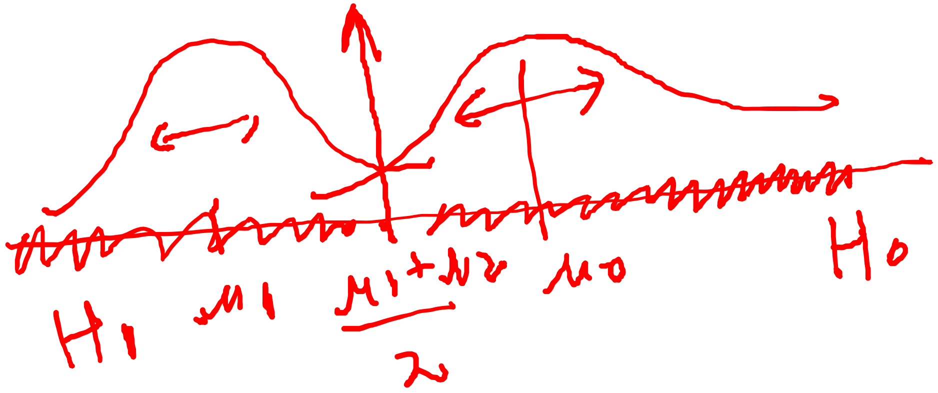
(Tied Covariance)

QDA becomes LDA as assuming tied Covariance.

case 1]

- scalar feature
- $\sigma_0 = \sigma_1 = \sigma$
- $P[\mathcal{H}_0] = P[\mathcal{H}_1]$

x



f_0 f_1

$$\ln P[\mathcal{H}_0] + \ln\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(x-\mu_0)^2 \underset{H_1}{\overset{H_0}{\gtrless}} \ln P[\mathcal{H}_1] + \ln\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(x-\mu_1)^2$$

$\updownarrow$

$$\ln P[\mathcal{H}_0] + \ln\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(x-\mu_0)^2 \gtrless \ln P[\mathcal{H}_1] + \ln\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(x-\mu_1)^2$$

$$\frac{1}{\sigma^2}x\mu_0 - \frac{1}{2\sigma^2}\mu_0^2 \gtrless \frac{1}{\sigma^2}x\mu_1 - \frac{1}{2\sigma^2}\mu_1^2$$

$$x(\mu_0-\mu_1) \gtrless \frac{1}{2}(\mu_0^2-\mu_1^2)$$

$$\boxed{x \underset{H_1}{\overset{H_0}{\gtrless}} \frac{1}{2}(\mu_0+\mu_1)}$$

WLOG if $(\mu_0 > \mu_1)$

Binary classification decision rule!

case 2]

- feature vector
- $\boxed{\Sigma_0 = \Sigma_1 = \sigma I, \text{ isotropic}}$
- $P[\mathcal{H}_0] = P[\mathcal{H}_1]$

$$\ln P[\mathcal{H}_0] + \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(x-\mu_0)^2 \underset{H_1}{\overset{H_0}{\gtrless}} \ln P[\mathcal{H}_1] + \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(x-\mu_1)^2$$

$\updownarrow$

$$\ln P[\mathcal{H}_0] + \ln \frac{1}{\sqrt{2\pi\sigma^N}} + -\frac{1}{2\sigma^2}(x-\mu_0)^t(x-\mu_0) \gtrless \ln P[\mathcal{H}_1] + \ln \frac{1}{\sqrt{2\pi\sigma^N}} + -\frac{1}{2\sigma^2}(x-\mu_1)^t(x-\mu_1)$$

$$\frac{1}{\sigma^2}\mu_0^t x - \frac{1}{2\sigma^2}\mu_0^t\mu_0 \gtrless \frac{1}{\sigma^2}\mu_1^t x - \frac{1}{2\sigma^2}\mu_1^t\mu_1$$

$$\boxed{(\mu_0-\mu_1)^t x \gtrless \frac{1}{2}(\mu_0-\mu_1)^t(\mu_0+\mu_1)}$$

projection to (μ_0 - μ_1)
then the decision rule same as the scalar case.

case 3]

- feature vector
- $\Sigma_0 = \Sigma_1 = \Sigma$, anisotropic
- $P[\mathcal{H}_0] = P[\mathcal{H}_1]$

Q: projection to $(\mu'_0$ - $\mu'_1)$
then the decision rule same as the scalar case?

$$\ln P[\mathcal{H}_0] + \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(x-\mu_0)^2 \gtrless \ln P[\mathcal{H}_1] + \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(x-\mu_1)^2$$

$$\Updownarrow$$

$$\ln P[\mathcal{H}_0] + \ln \frac{1}{\sqrt{2\pi|\Sigma|}} - \frac{1}{2}(x-\mu_0)^t\Sigma^{-1}(x-\mu_0) \gtrless \ln P[\mathcal{H}_1] + \ln \frac{1}{\sqrt{2\pi|\Sigma|}} - \frac{1}{2}(x-\mu_1)^t\Sigma^{-1}(x-\mu_1)$$

$$\mu_0^t\Sigma^{-1}x - \frac{1}{2}\mu_0^t\Sigma^{-1}\mu_0 \gtrless \mu_1^t\Sigma^{-1}x - \frac{1}{2}\mu_1^t\Sigma^{-1}\mu_1$$

$$(\mu_0^t - \mu_1^t)\Sigma^{-1}x \gtrless \frac{1}{2}\mu_0^t\Sigma^{-1}\mu_0 - \frac{1}{2}\mu_1^t\Sigma^{-1}\mu_1$$

$$\boxed{(\mu_0 - \mu_1)^t E\Lambda^{-1}E^t x \gtrless \frac{1}{2}\mu_0^t E\Lambda^{-1}E^t\mu_0 - \frac{1}{2}\mu_1^t E\Lambda^{-1}E^t\mu_1}$$