Machine Learning Principles

Class7 : Sept. 25

Linear Classification I: Linear Discriminant Analysis

Instructor: Diana Kim

Today's Lecture

1. What is Classification Problem?

    - Discriminant Functions

    - Linear Classification

2. MAP optimal classification

    - Gaussian Discriminant Analysis (GDA)

3. Generative vs. Discriminative Modeling (to estimate posterior $\mathrm{P}(C_k|x)$)

    - Introduction to Logistic Regression

- Classification Problem

[1] Classification Problem (regression vs. classification)

- In regression, learning the function f
  to predict continuous y (real value)
  given the value of M dimensional input data $(x_1, x_{2,}, \dots .\, x_{m,})$

$$y = f(x_1, x_{2,}, \dots .\, x_{m.})$$

(functional relation between x and y)

[2] Classification Problem (regression vs. classification)

- In classification, learning a set of function f_k $(k = 1, \dots, K)$
  to predict a class C_k (category/ discrete)
  given the value of M dimensional input data $(x_1, x_{2,}, \dots . x_{m,})$

$$C_k = \arg\max_k f_k(x_1, x_2, \dots x_m)$$

(class decision based on scores by the functions)

[3] Classification Problem (discriminant functions)

- we call the set of function $f_k \quad (k = 1, \dots, K)$
  "discriminant functions". It can be a linear / non-linear functions
  defined over the data domain (scalar, 2d, 3d,…)

- classification ML algorithm is about how to learn the functions.

[4] Classification Problem (class decision: scalar example)

- based on a set of discriminant functions (learned by an ML algorithm) we can classify an input data point.

[5] Classification Problem (class decision: 2d example)

ex] $f_1(x_1, x_2) = x_2 - x_1 - 1$

$f_2(x_1, x_2) = x_2 + x_1 - 1$

$f_3(x_1, x_2) = x_2$

three discriminant functions.

Q: assign a class for the data points?

data points	f_1	f_2	f_3	class
$(-2,0)$				
$(0,0)$				
$(2,0)$				

- Linear Classification

  when the decision boundary/surface is a hyperplane.

[1] Linear Classification (decision boundary)

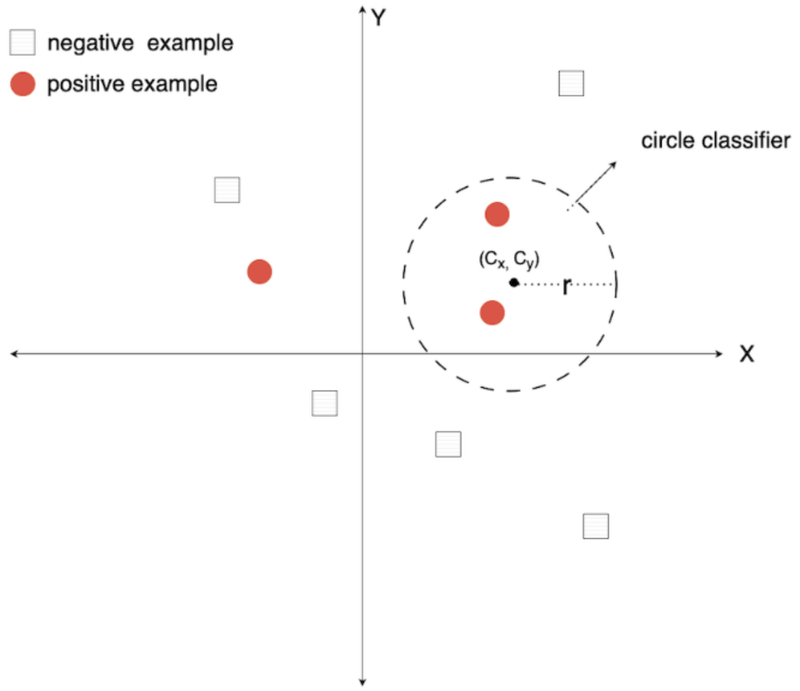
- decision boundary is a hypersurface
  that separates different classes of data.

- in the example
  what is the decision boundary?

- Q: the decision boundary for the functions below?

$$\begin{cases} f_1(x_1, x_2) = x_1^2 + x_2^2 + 3x_1 + 4x_2 + 3 \\ f_2(x_1, x_2) = x_1^2 + x_2^2 + 8x_1 + 7x_2 + 5 \end{cases}$$

$$\begin{cases} \delta(x, y) = +1 & (x - c_x)^2 + (y - c_y)^2 \leq r^2 \\ \delta(x, y) = -1 & (x - c_x)^2 + (y - c_y)^2 > r^2 \end{cases}$$



[2] Linear Classification (linear decision boundary)

We call classification is linear,
when the decision boundary is defined by a linear hyperplane.
However, the discriminative functions can be non-linear as long as their decision boundary remains linear.

ex] the previous example defines a linear classifier.

$$\begin{cases} f_1(x_1, x_2) = x_1^2 + x_2^2 + 3x_1 + 4x_2 + 3 \\ f_2(x_1, x_2) = x_1^2 + x_2^2 + 8x_1 + 7x_2 + 5 \end{cases}$$

[3] Linear Classification (the example of classification $K = 3$)

ex] define the decision regions for classification by the discriminative function below.

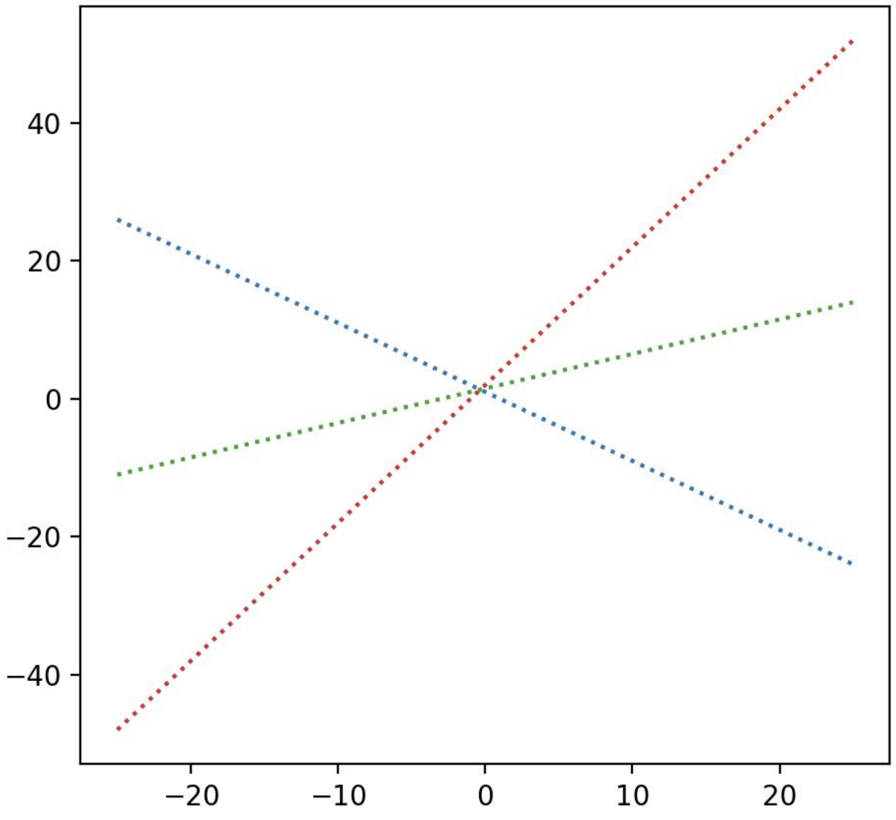
$$\begin{cases} f_1(x_1, x_2) = x_1 + x_2 + 3 \\ f_2(x_1, x_2) = 2x_1 + 2x_2 + 2 \\ f_3(x_1, x_2) = 3x_2 \end{cases}$$

Recitation Problem!

[4] Linear Classification (the example of classification $K = 3$)

ex] define the decision regions for classification by the discriminative function below.

$$\begin{cases} f_1(x_1, x_2) = x_1 + x_2 + 3 \\ f_2(x_1, x_2) = 2x_1 + 2x_2 + 2 \\ f_3(x_1, x_2) = 3x_2 \end{cases}$$



[5] Linear Classification

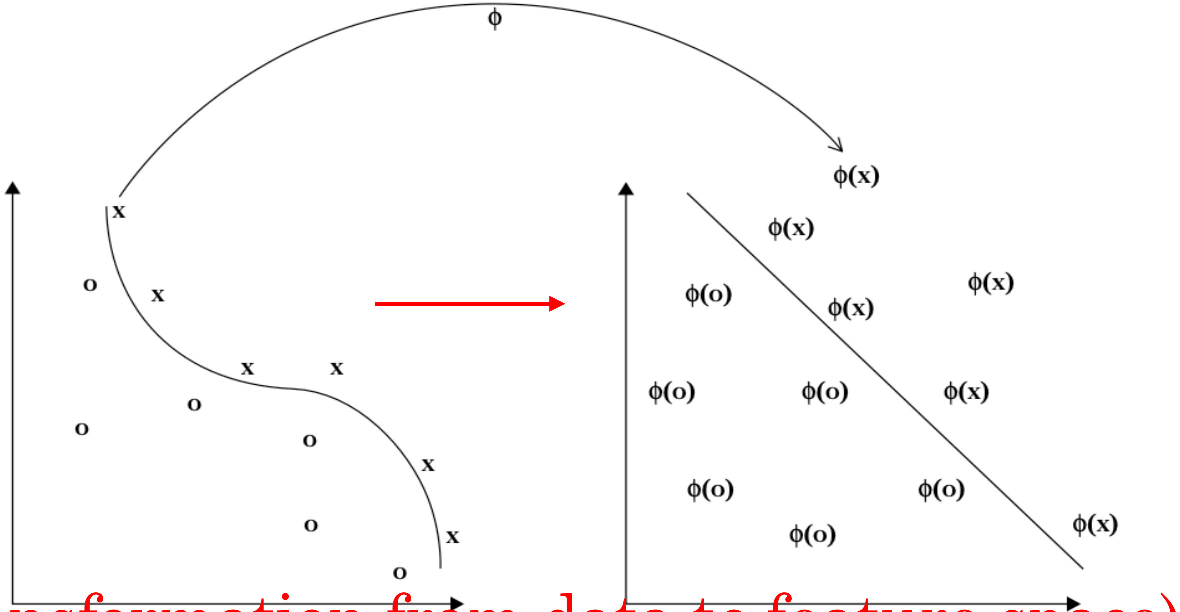
Q: isn't too simplistic to classify high dimensional data samples
in the real world using a linear classifier?
a linear classifier is useful: ease of implementation, interpretability, etc.

how could we make the linear classifier work?

[5] Linear Classification (feature engineering)

- For linear classification,
  feature engineering will be needed to make the data points linearly separable like.

From Kernel Methods for Pattern Analysis by John Shawe-Talyor



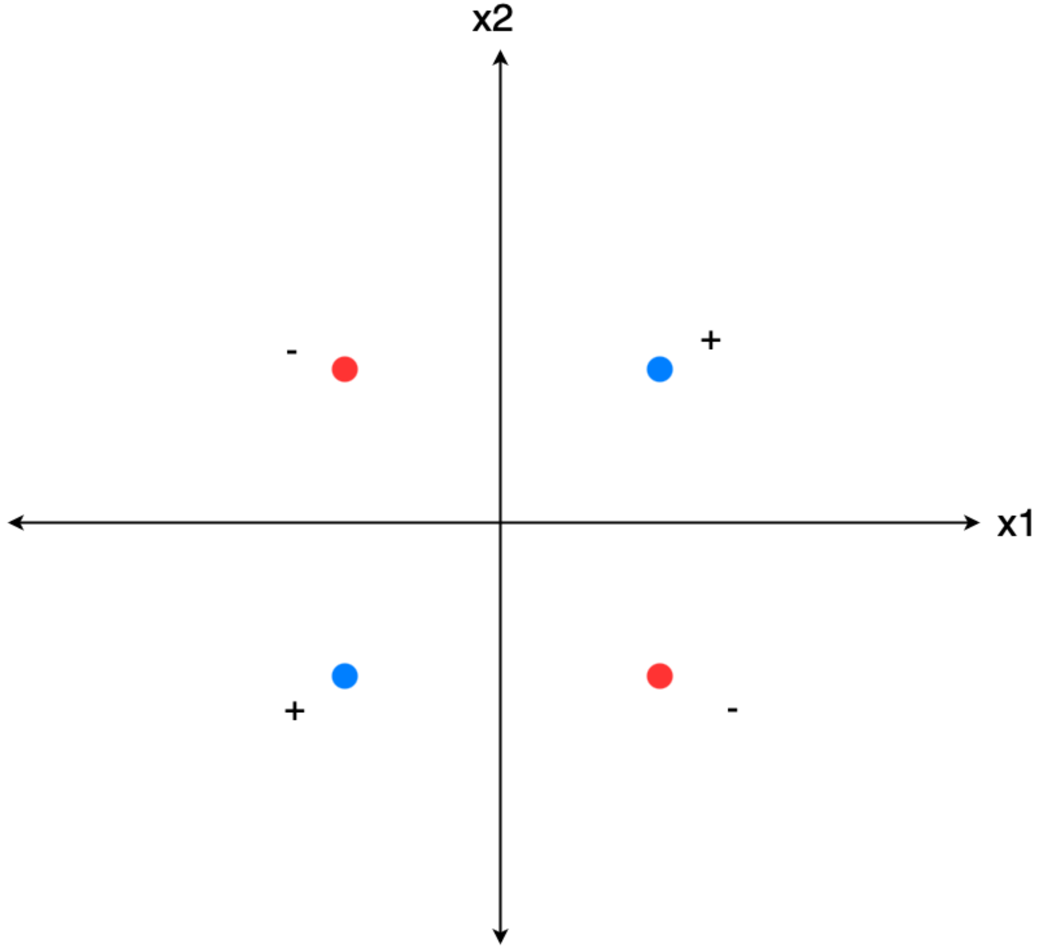
(transformation from data to feature space)

Fig. 2.1. The function ϕ embeds the data into a feature space where the nonlinear pattern now appears linear. The kernel computes inner products in the feature space directly from the inputs.

- given data, we can separate the data samples by hyperplanes.
- feature engineering makes it possible.

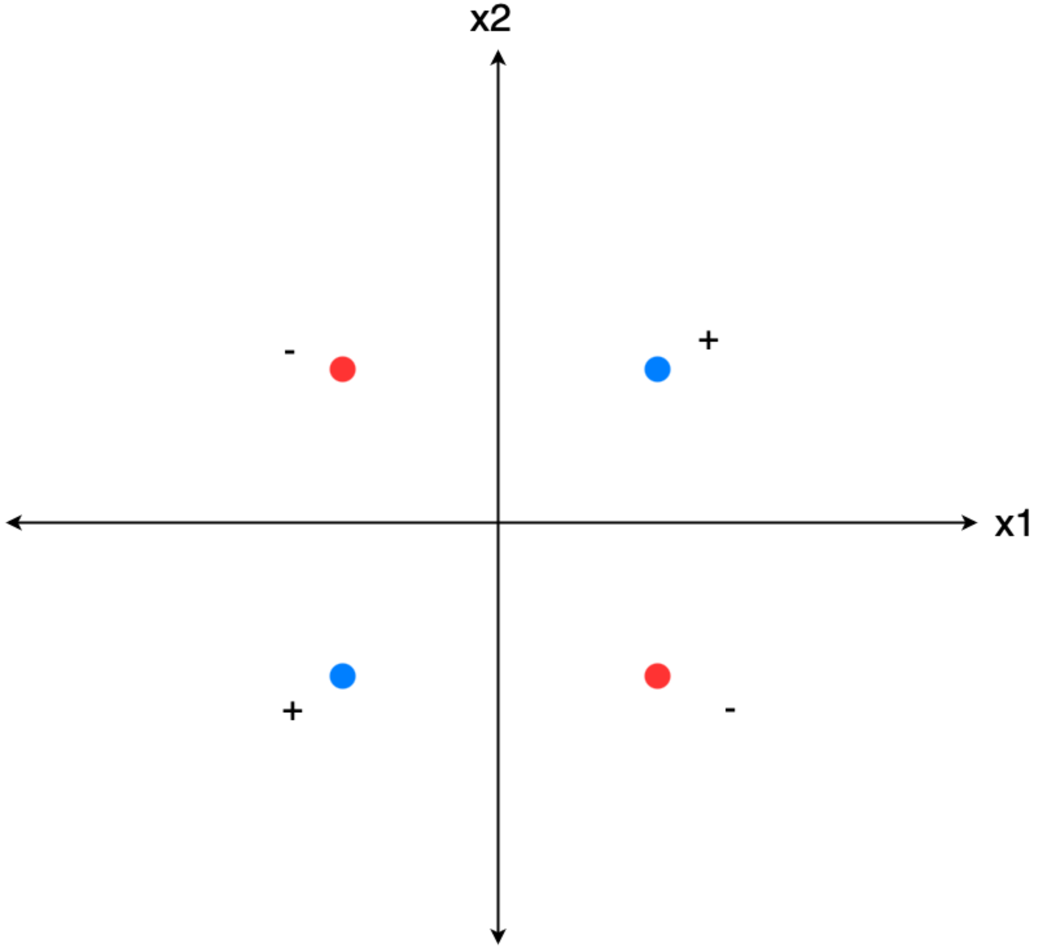
[6] Linear Classification (XOR problem)

Q: how would you create $\mathbf{X}_3$ to make the feature space to be linearly separable?



[7] Linear Classification (XOR problem)

Q: how would you create $\mathbf{X}_3$ to make the feature space to be linearly separable?



- ans: $(x_1, x_2) \rightarrow (x_1, x_2, x_1 \cdot x_2)$

$x_1 \cdot x_2$ is added to the feature space,
then the space becomes linearly separable.

MAP classification (optimal)

Q: given data $\vec{x}$, what probability would you use for classification?

[1] MAP Classification (posterior as discriminant functions)

Q: could we use the posterior probability as a discriminant function?

[a posterior for class K given data point $\vec{x}$]

$$P(C_k|\vec{x}) = \frac{P(\vec{x}|C_k)P(C_k)}{P(\vec{x})}$$

- the posterior is a function of $\vec{x}$
- by $argmax_{C_k} P(C_k|\vec{x})$ we could decide which class is most probable given a data point $\vec{x}$.
- ML classification algorithms aim to learn posterior.

[2] MAP Classification (MAP is an optimal decision rule)

- [Expected Error / Error probability]
  where y_o and y_1 are the decision region for class $K = 0$ and 1

$$
\begin{aligned}
E[R] = \pi_0 \cdot E[R|C_0] + \pi_1 E[R|C_1] &= \pi_1 \cdot \int_{y_0} f(y|C_1)dy + \pi_0 \cdot \int_{y_1} f(y|C_0)dy \\
&= \pi_1 \cdot \int_{y_0} f(y|C_1)dy + \pi_0 \cdot (1 - \int_{y_0} f(y|C_0)dy) \\
&= \pi_0 + \int_{y_0} \pi_1 \cdot f(y|C_1) - \pi_0 \cdot f(y|C_0)dy
\end{aligned}
$$

Q: to minimize $E[R]$, the decision region y_o and y_1 ?

[3] MAP Classification (MAP is an optimal decision rule)

Q: to minimize $E[R]$, the decision region y_o and y_1 ?

- [optimal decision rule]

$$\text{if } \pi_1 \cdot f(y|C_1) - \pi_0 \cdot f(y|C_0) < 0 \text{ then } y_0$$
$$\text{else if } \pi_1 \cdot f(y|C_1) - \pi_0 \cdot f(y|C_0) \geq 0 \text{ then } y_1$$

- [the optimal rule is MAP]

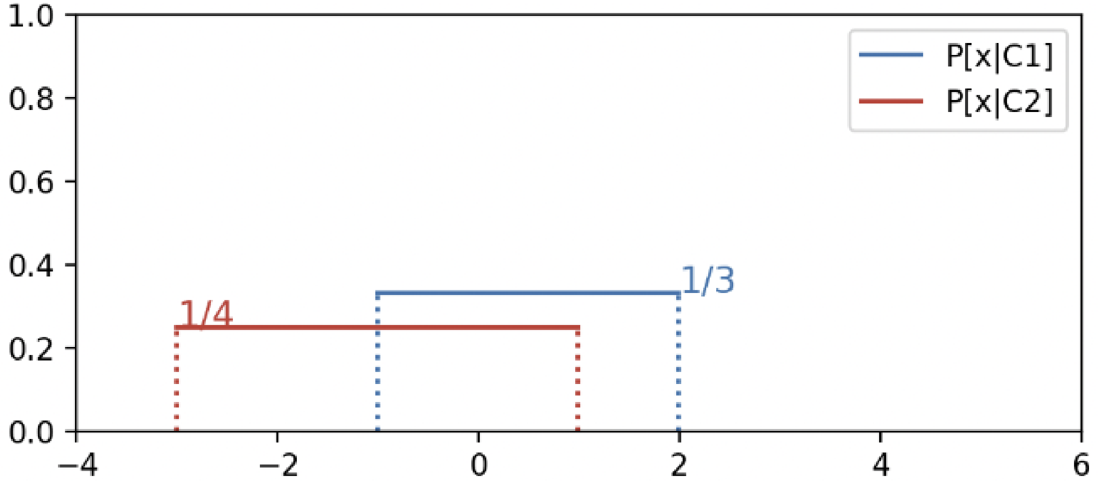
$$\frac{p(y|C_1)}{P(y|C_0)} \overset{C_1}{\underset{C_0}{\gtreqless}} \frac{\pi_0}{\pi_1} \longleftrightarrow p(x|C_1)\pi_1 \overset{C_1}{\underset{C_2}{\gtreqless}} p(x|C_0)\, \pi_0$$

[4] MAP Classification (example)

2. Suppose you classify a sample x using the MAP rule.

MAP rule:

$$\mathcal{K}^* = \arg\max_k P[C_k|x] \propto P[x|C_k]P[C_k]$$



- if prior probabilities are uniform then MAP rule becomes comparison $P[x|C_k]$

2.1 Classify the sample $x = 0$ when $P[C_1] = P[C_2] = \frac{1}{2}$. Use the two conditional densities above.

- **G**aussian **D**iscriminant **A**nalysis (**GDA**)
  - first classification algorithm
  - MAP approach, assuming P$[x|C_k] \sim N(\mu_k, \Sigma_k)$

[1] Gaussian Discriminant Analysis (MAP)

$$\mathrm{P}(x|C_1) \cdot \pi_1 \underset{C_2}{\overset{C_1}{\gtreqless}} P(x|C_0) \cdot \pi_0$$

[Gaussian density]

Q: what statistics do we need to learn for GDA?

[2] Gaussian Discriminant Analysis (anisotropic: decision boundary)

- $\Sigma_0 = \Sigma_1$

$$P[\mathcal{C}_0] \cdot \frac{1}{\sqrt{2\pi|\Sigma_0|}} \exp -\frac{1}{2}(x-\mu_0)^t \Sigma_0^{-1}(x-\mu_0) \underset{C_1}{\overset{C_0}{\gtrless}} P[\mathcal{C}_1] \cdot \frac{1}{\sqrt{2\pi|\Sigma_1|}} \exp -\frac{1}{2}(x-\mu_1)^t \Sigma_1^{-1}(x-\mu_1)$$

$$\Updownarrow$$

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

$$\Updownarrow$$

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

- two linear discriminant functions!
- linear decision boundary!

[3] Gaussian Discriminant Analysis (anisotropic: decision boundary)

- $\Sigma_0 \neq \Sigma_1$

$$P[\mathcal{C}_0] \cdot \frac{1}{\sqrt{2\pi|\Sigma_0|}} \exp -\frac{1}{2}(x-\mu_0)^t \Sigma_0^{-1}(x-\mu_0) \underset{C_1}{\overset{C_0}{\gtrless}} P[\mathcal{C}_1] \cdot \frac{1}{\sqrt{2\pi|\Sigma_1|}} \exp -\frac{1}{2}(x-\mu_1)^t \Sigma_1^{-1}(x-\mu_1)$$

$$\Big\Updownarrow$$

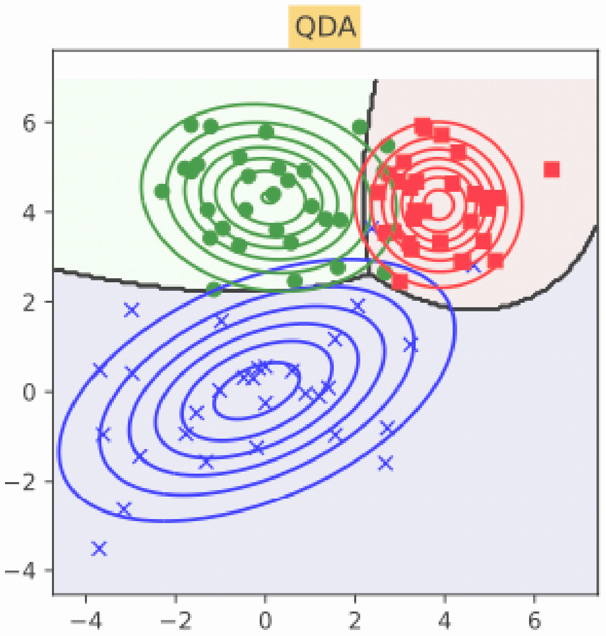
$$\ln P[\mathcal{C}_0] + \ln \frac{1}{\sqrt{2\pi|\Sigma_0|}} - \frac{1}{2}(x-\mu_0)^t \Sigma_0^{-1}(x-\mu_0) \gtrless \ln P[\mathcal{C}_1] + \ln \frac{1}{\sqrt{2\pi|\Sigma_1|}} - \frac{1}{2}(x-\mu_1)^t \Sigma_1^{-1}(x-\mu_1)$$

- two quadratic discriminant functions!
- quadratic decision boundary!

[4] Gaussian Discriminant Analysis (textbook figures)

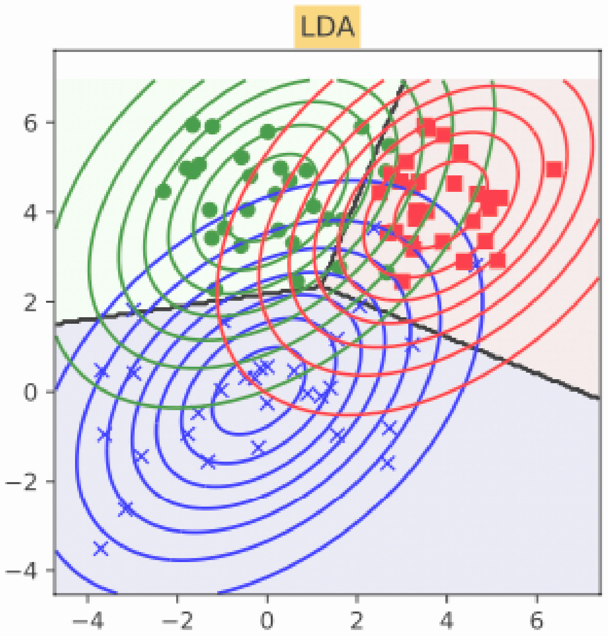
From Murphy Figure 9.2

[Quadratic Discriminant Analysis]



$\Sigma_0 \neq \Sigma_1$

[Linear Discriminant Analysis]

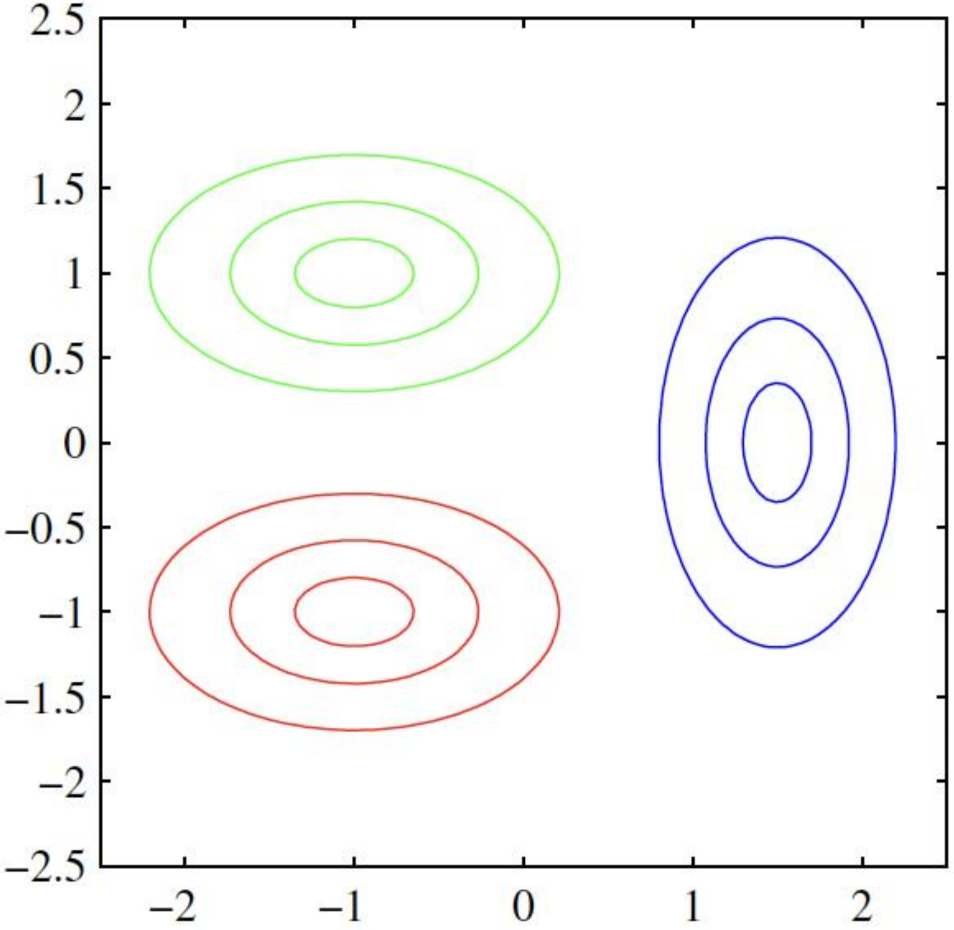


$\Sigma_0 = \Sigma_1$

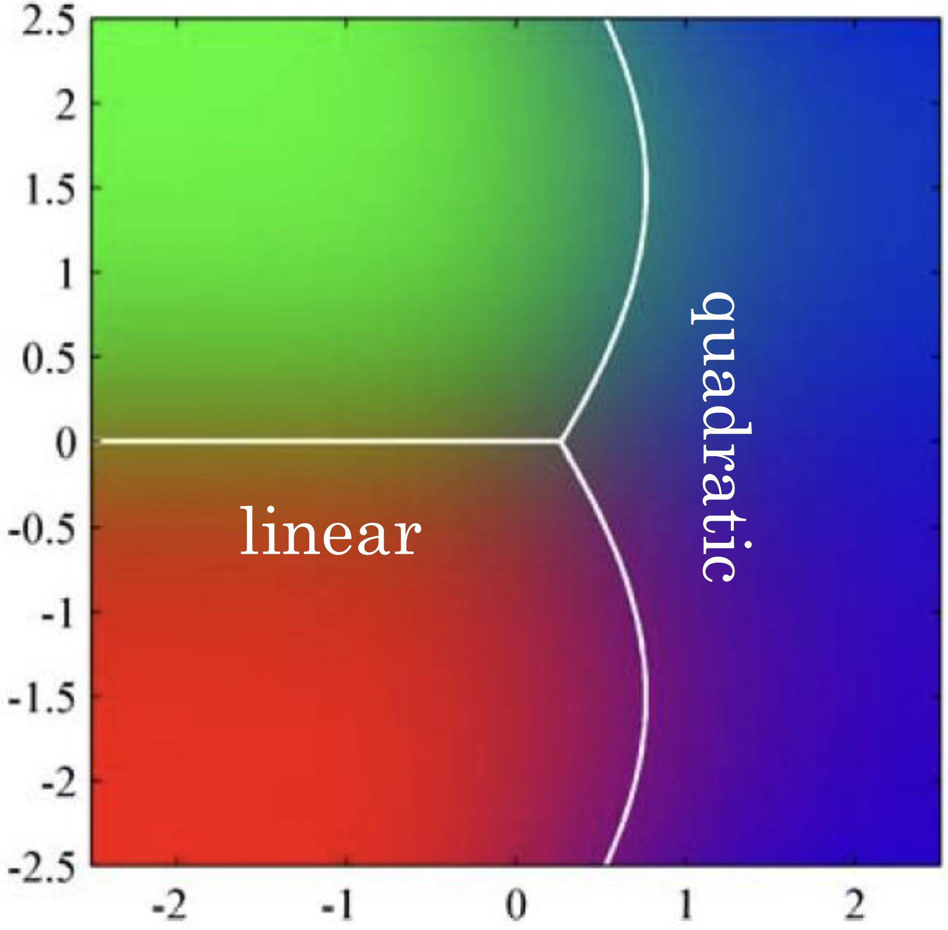
- Gaussian Discriminant Analysis (GDA)
  becomes a linear classifier as assuming tied covariance.

[5] Gaussian Discriminant Analysis (textbook figures)

From



[class conditional Gaussian densities]



[RGB representation for posterior]

- Gaussian Discriminant Analysis (case example $\Sigma_0 = \Sigma_1$)

[6] Gaussian Discriminant Analysis (case example I)

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

$$\ln P[\mathcal{C}_0] + \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(x-\mu_0)^2 \underset{C_1}{\overset{C_0}{\gtrless}} \ln P[\mathcal{C}_1] + \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(x-\mu_1)^2$$

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

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

$$\boxed{x \overset{C_0}{\gtrless} \frac{1}{2}(\mu_0 + \mu_1)}$$

[binary classification decision rule]

($\mu_0 \geq 0$)

[7] Gaussian Discriminant Analysis (case example II)

- feature vector
- $\Sigma_0 = \Sigma_1 = \sigma^2 I$, isotropic
- $P[\mathcal{C}_0] = P[\mathcal{C}_1]$

$$\ln P[\mathcal{C}_0] + \ln \frac{1}{\sqrt{2\pi\sigma^N}} + -\frac{1}{2\sigma^2}(x-\mu_0)^t(x-\mu_0) \underset{C_1}{\overset{C_0}{\gtrless}} \ln P[\mathcal{C}_1] + \ln \frac{1}{\sqrt{2\pi\sigma^N}} + -\frac{1}{2\sigma^2}(x-\mu_1)^t(x-\mu_1)$$

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

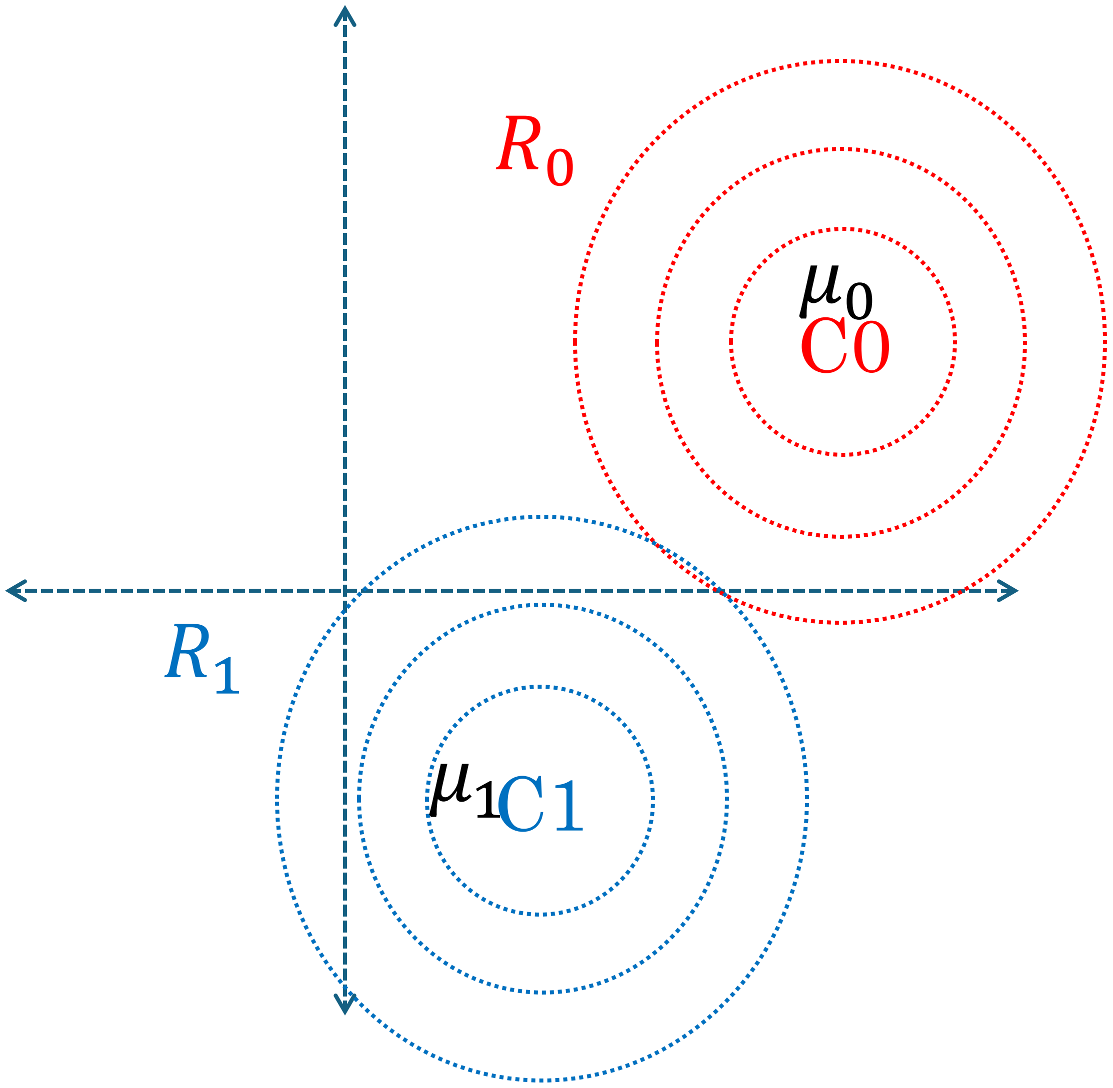
$$\boxed{(\mu_0 - \mu_1)^t x} \gtrless \frac{1}{2}(\mu_0 - \mu_1)^t(\mu_0 + \mu_1)$$

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

[8] Gaussian Discriminant Analysis (case example II)

ex] draw the decision boundary

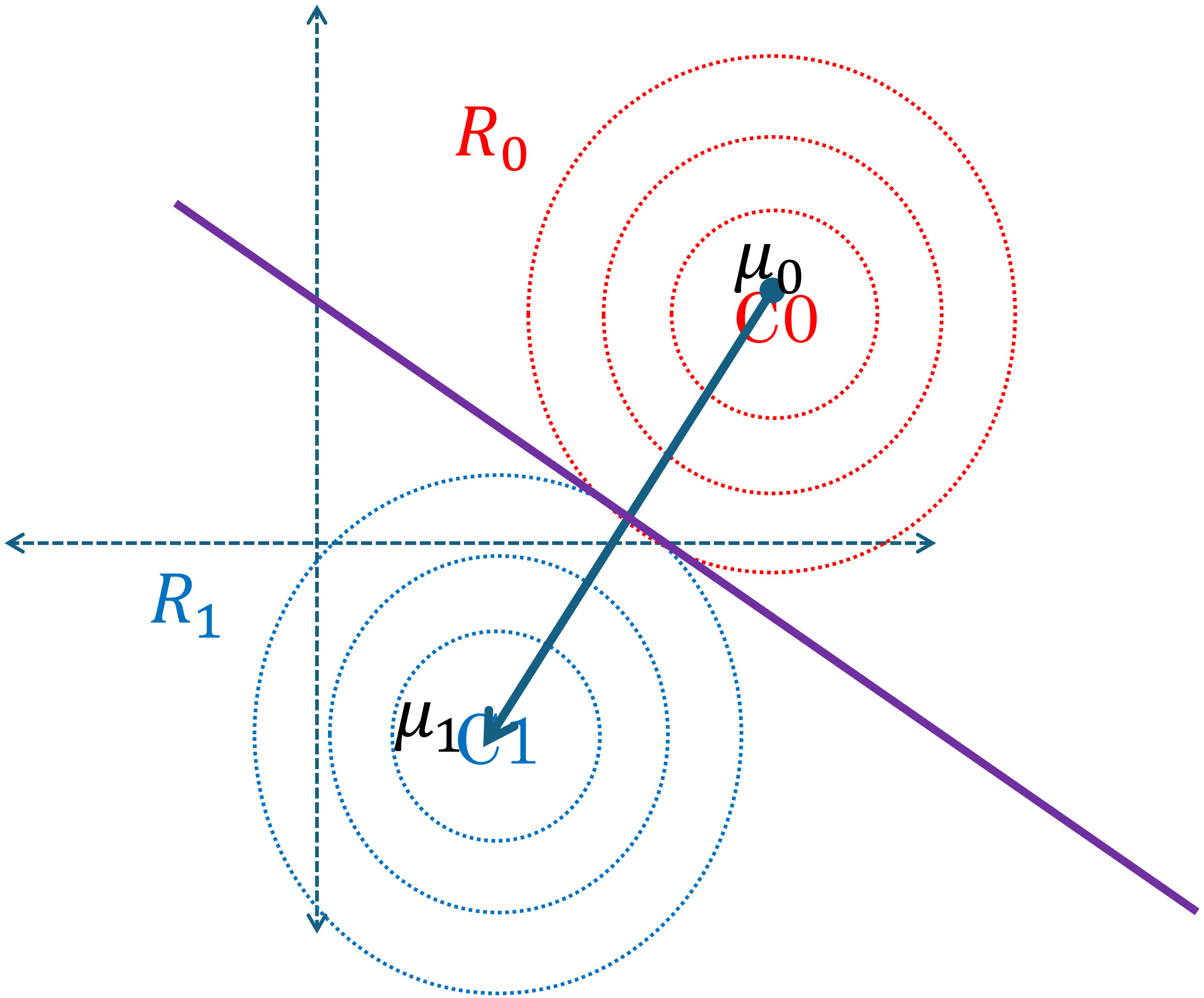
- feature vector
- $\Sigma_0 = \Sigma_1 = \sigma^2 I$, isotropic
- $P[\mathcal{C}_0] = P[\mathcal{C}_1]$



[8] Gaussian Discriminant Analysis (case example II)

ex] draw the decision boundary

- feature vector
- $\Sigma_0 = \Sigma_1 = \sigma^2 I$, isotropic
- $P[\mathcal{C}_0] = P[\mathcal{C}_1]$



[9] Gaussian Discriminant Analysis (case example III)

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

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

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

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

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

[10] Gaussian Discriminant Analysis (case example III interpretation)

Q: what is the meaning of this? $(\mu_0 - \mu_1)^t E\Lambda^{-1}E^t x \gtrless \frac{1}{2}\mu_0^t E\Lambda^{-1}E^t\mu_0 - \frac{1}{2}\mu_1^t E\Lambda^{-1}E^t\mu_1$

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

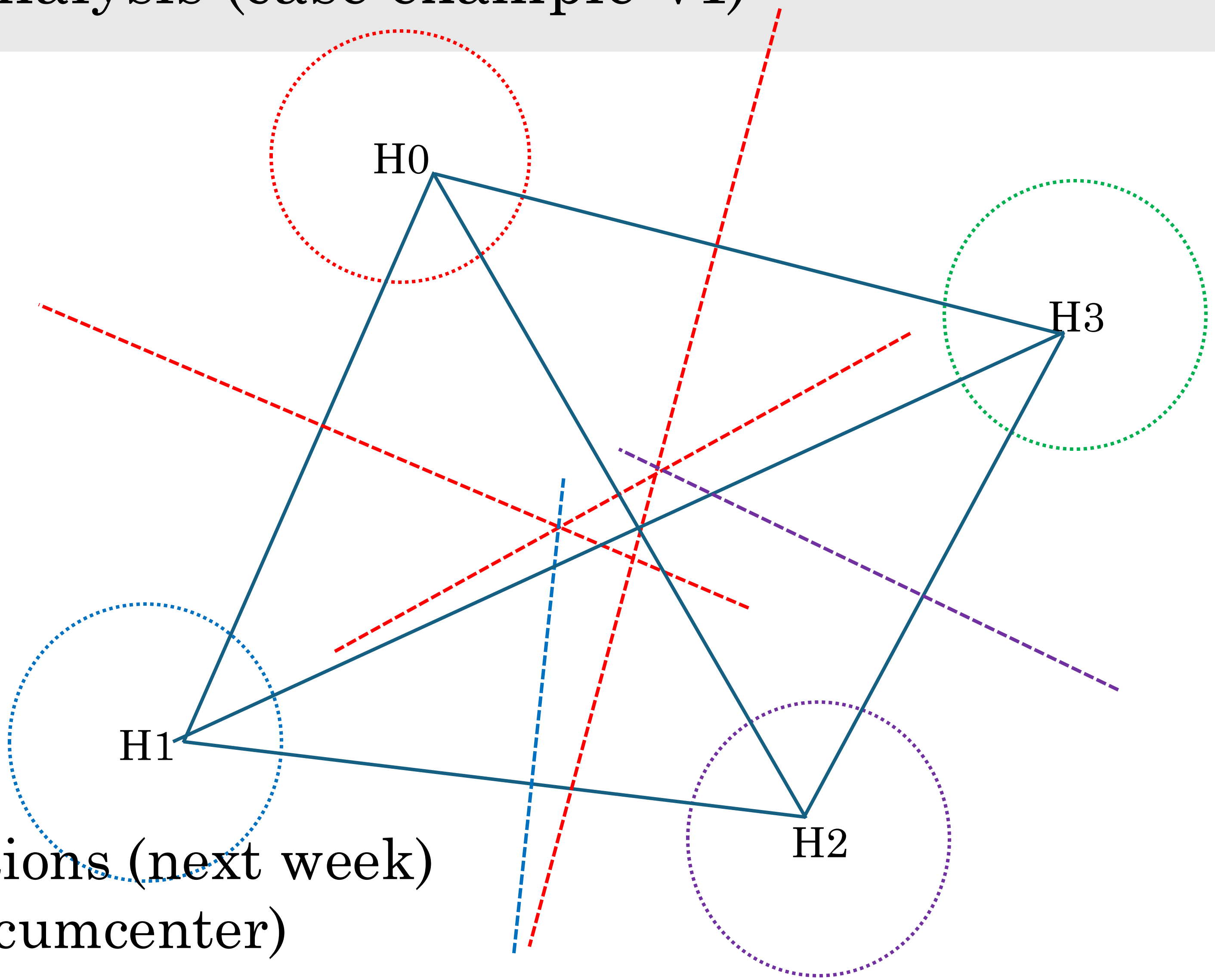
$$[\Lambda^{-1/2}E^t(\mu_0 - \mu_1)]^t[\Lambda^{-1/2}E^t]x \gtrless \frac{1}{2}[\Lambda^{-1/2}E^t(\mu_0 - \mu_1)]^t[\Lambda^{-1/2}E^t](\mu_0 + \mu_1)$$

$$\underbrace{[\Lambda^{-1/2}E^t(\mu_0 - \mu_1)]^t}_{\text{(2) projection to } (\mu'_0 - \mu'_1)}\underbrace{[\Lambda^{\frac{-1}{2}}E^t]}_{\text{(1) rotation and scaling (whitening without centering)}}x \gtrless \frac{1}{2}(\mu_0 - \mu_1)^t E\Lambda^{-1}E^t(\mu_0 + \mu_1)$$

- still we can derive a scalar decision rule by
  (1) decorrelation and compute new μ'_0 and μ'_1
  (2) projection to new $(\mu'_0 - \mu'_1)$

[11] Gaussian Discriminant Analysis (case example VI)

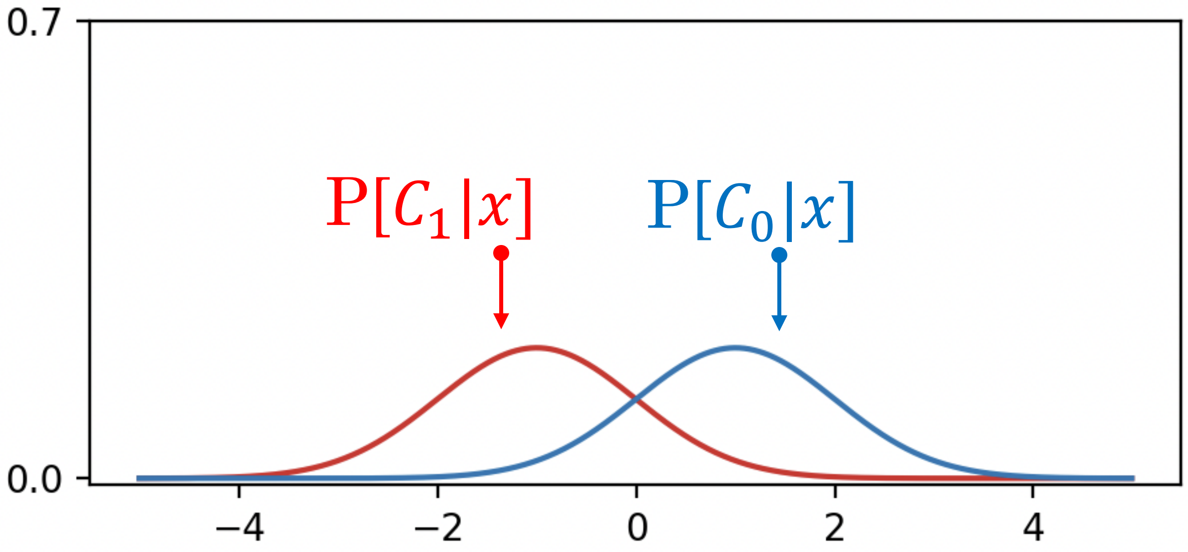
- feature vector
- $\Sigma_0 = \Sigma_1 = \Sigma_3 = \Sigma_4 = \sigma^2 I$, isotropic
- $P[\mathcal{C}_0] = P[\mathcal{C}_1] = P[\mathcal{C}_2] = P[\mathcal{C}_4]$



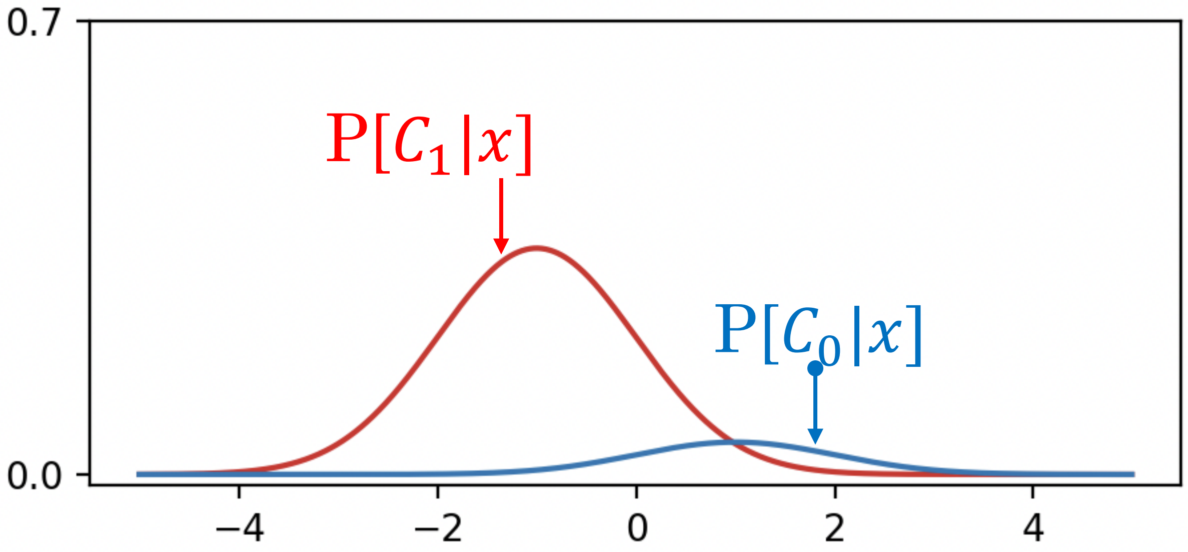
- this will be discussed in recitations (next week)
- the lines meet at one point (circumcenter)

[12] Gaussian Discriminant Analysis (prior effect)

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



- scalar feature
- $\sigma_0 = \sigma_1 = \sigma$
- $P[\mathcal{C}_0] = 1/4$
- $P[\mathcal{C}_1] = 3/4$ [prior is not uniform]



- Generative vs. Discriminative Classifier

:both aim to learn $\mathrm{P}[C_k|x]$ but in different ways.

$$\arg\max_k P(C_k|x) \propto P(x, C_k) = P(x|C_k) \cdot P(C_k)$$

- Generative Classification

  - Gaussian Discriminant Analysis
  - Naïve Bayes

[1] Generative vs. Discriminative Classifier (generative modeling)

- Generative classifier learns posterior through prior and likelihood.

$$\arg\max_k P(C_k|x) \propto P(x, C_k) = \underbrace{P(x|C_k)}_{\text{[likelihood]}} \cdot \underbrace{P(C_k)}_{\text{[prior]}}$$

- in the training stage, we learn likelihood and prior first so multiply them to compute posterior.
- Q: how to learn the prior and likelihood?

- **GDA is generative modeling**.
  as we train the discriminant functions, we estimate prior, mean, covariance for each class to define posterior probability.

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

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

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

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

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

- Naïve Bayes is one example of GDA.

[HW problem #4]

[posterior] [likelihood] [prior]

$$P[D=+|G=g,B=b] = \frac{P[D=+,G=g,B=b]}{P[G=g,B=b]} = \frac{P[G=g,B=b|D=+]\cdot P[D=+]}{P[G=g,B=b]}$$

[**the likelihood is conditionally independent.]

- Discriminative Classification
  - Logistic Regression
  - Deep Convolutional Neural Net

[1] Generative vs. Discriminative Classifier (discriminative modeling)

Discriminative Classification (w.o learning prior/likelihood)

: can directly model the posterior $P(C_k|x)$ with a linear function of feature map $w^t\phi(x) + b$?
without learning likelihood / prior?

[2] Generative vs. Discriminative Classifier (discriminative modeling)

Q: how can we directly model the posterior $P(C_k|x)$?

[3] Generative vs. Discriminative Classifier (logistic sigmoid)

$$P[C_1|x] = \frac{P[x|C_1]P[C_1]}{P[x|C_1]P[C_1] + P[x|C_0]P[C_0]}$$

$$P[C_1|x] = \frac{1}{1 + \frac{P[x|C_0]P[C_0]}{P[x|C_1]P[C_1]}}$$

$$P[C_1|x] = \frac{1}{1 + \exp\left(\ln \frac{P[x|C_0]P[C_0]}{P[x|C_1]P[C_1]}\right)}$$

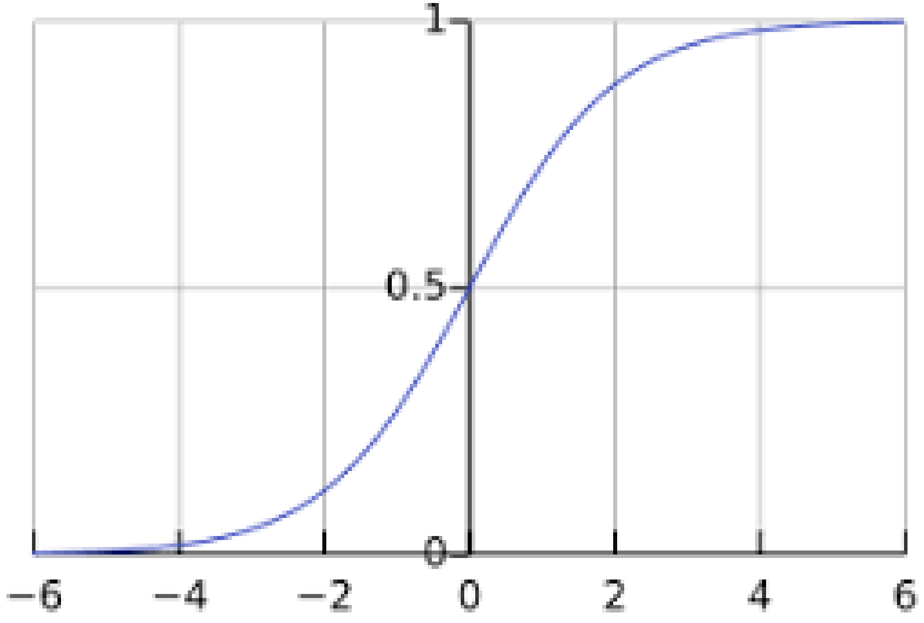
$$P[C_1|x] = \frac{1}{1 + \exp\left(\boxed{- \ln \frac{P[x|C_1]P[C_1]}{P[x|C_0]P[C_0]}}\right)}$$

In Gaussian modeling, this was simplified into a linear decision rule: $w^t x + b \lesseqgtr 0$ like

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

[4] Generative vs. Discriminative Classifier (logistic sigmoid)

[logistic sigmoid function]



$$\sigma(x) = \frac{1}{1 + exp^{-x}}$$

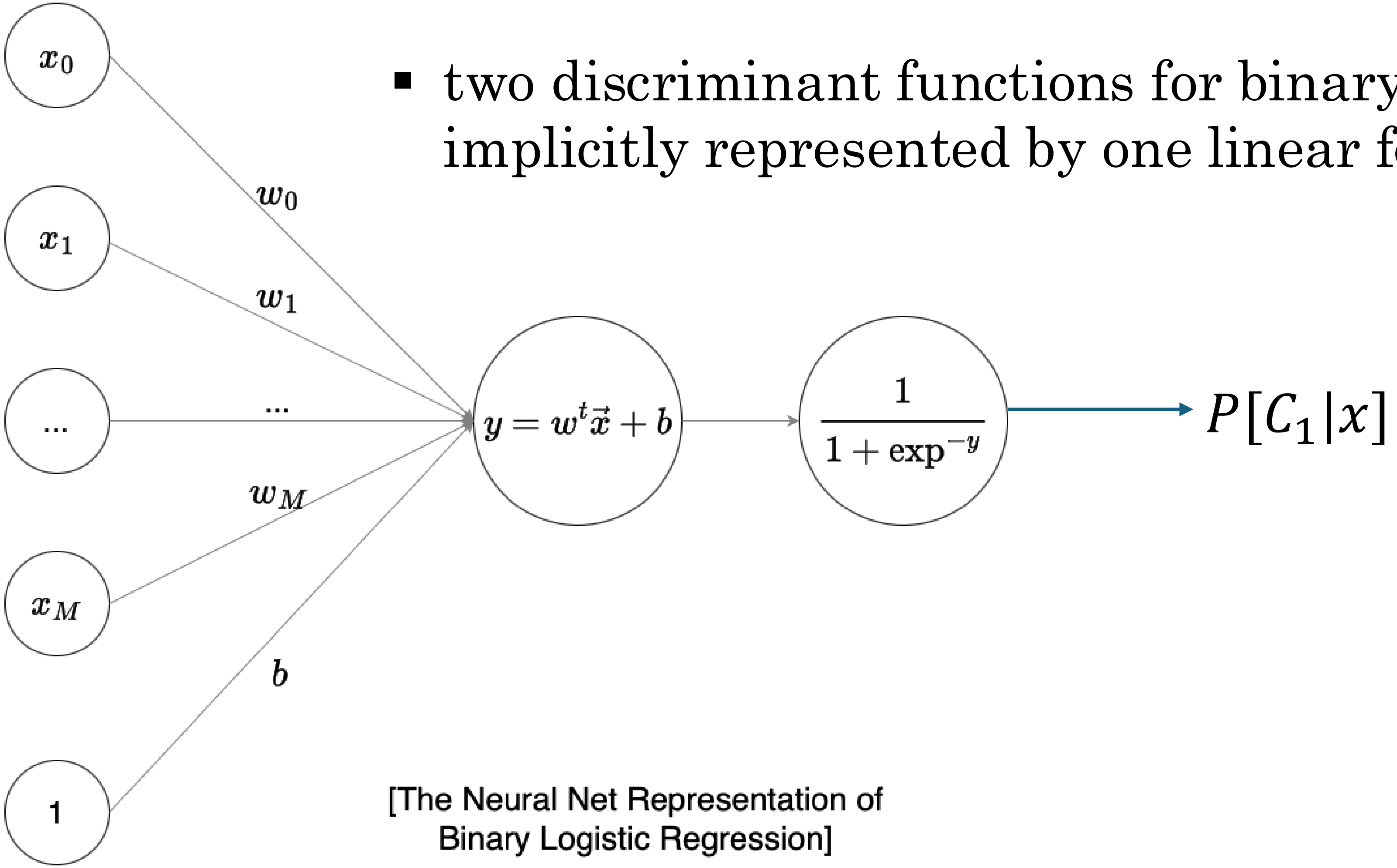
$$P[C_1|\vec{x}] = \frac{1}{1 + \exp\left(-\vec{w}^t\vec{x} - b\right)}$$

$$P[C_0|\vec{x}] = \frac{\exp\left(-\vec{w}^t\vec{x} - b\right)}{1 + \exp\left(-\vec{w}^t\vec{x} - b\right)}$$

$$** P[C_0|x] = 1 - P[C_1|x] **$$

[5] Generative vs. Discriminative Classifier (binary logistic regression)

- two discriminant functions for binary classification are implicitly represented by one linear formula $w^t x + b$



[The Neural Net Representation of Binary Logistic Regression]