Module 4: Classification
Part 2

TMA4268 Statistical Learning V2023

Daesoo Lee,

Department of Mathematical Sciences, NTNU

NTNU
Norwegian University of Science and Technology

01/02/2023

Recap

Recap

Two approaches to estimate $\Pr(Y = k \mid X = x)$

Diagnostic Paradigm

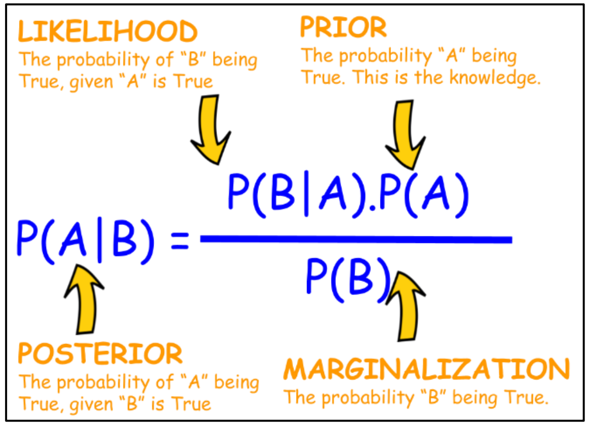
- *Directly* estimating $\Pr(Y = k \mid X = x)$
- *e.g.*, Logistic regression, KNN classification

Sampling Paradigm

- *Indirectly* estimating $\Pr(Y = k \mid X = x)$
  by modeling the likelihood $\Pr(X = x \mid Y = k)$ and the prior $\Pr(Y = k)$.

$$\Pr(Y = k \mid X = x) \propto \Pr(X = x \mid Y = k)\Pr(Y = k)$$

Bayes Theorem



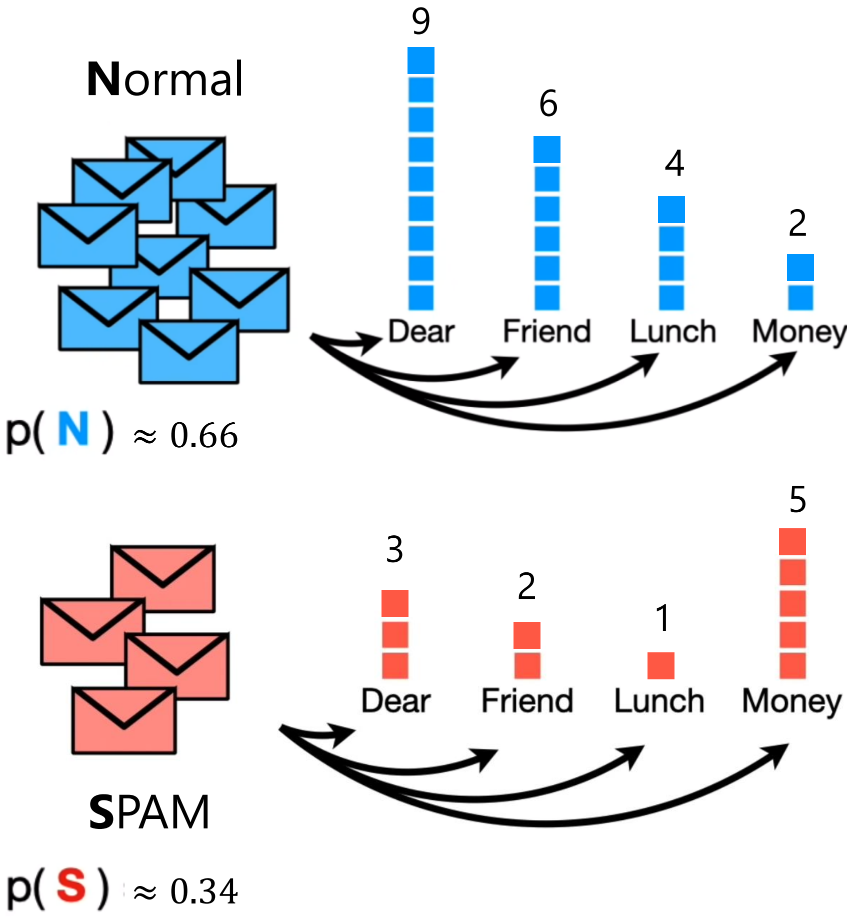
Remember Naïve Bayes Classifier?

$$p(Y_k|\boldsymbol{x}) \propto p(x_1|Y_k)p(x_2|Y_k)\cdots p\left(x_p\middle|Y_k\right)p(Y_k)$$

The Bayes classifier

$$p(Y_k|\boldsymbol{x}) \propto p(x_1|Y_k)p(x_2|Y_k)\cdots p(x_p|Y_k)p(Y_k)$$

Example: Naïve Bayes Classifier



Total num of emails = 32

p(**Dear** | **N**) ≈ 0.43
p(**Friend** | **N**) ≈ 0.28
p(**Lunch** | **N**) ≈ 0.19
p(**Money** | **N**) ≈ 0.1

p(**Dear** | **S**) ≈ 0.27
p(**Friend** | **S**) ≈ 0.18
p(**Lunch** | **S**) ≈ 0.09
p(**Money** | **S**) ≈ 0.46

Email: "Dear Friend"

$p(\text{N} \mid \text{"Dear Friend"})$
$\propto p(\text{Dear}|\text{N})p(\text{Friend}|\text{N})p(\text{N})$
$\propto 0.43 \cdot 0.28 \cdot 0.66 = 0.07$

$p(\text{S} \mid \text{"Dear Friend"})$
$\propto p(\text{Dear}|\text{S})p(\text{Friend}|\text{S})\ p(\text{S})$
$\propto 0.27 \cdot 0.18 \cdot 0.34 = 0.017$

Email: "Lunch Money Money Money"

$p(\text{N} \mid \text{"Lunch Money Money Money"})$
$\propto p(\text{Lunch}|\text{N})p(\text{Money}|\text{N})^3 p(\text{N})$
$\propto 0.19 \cdot 0.1^3 \cdot 0.66 = 0.00013$

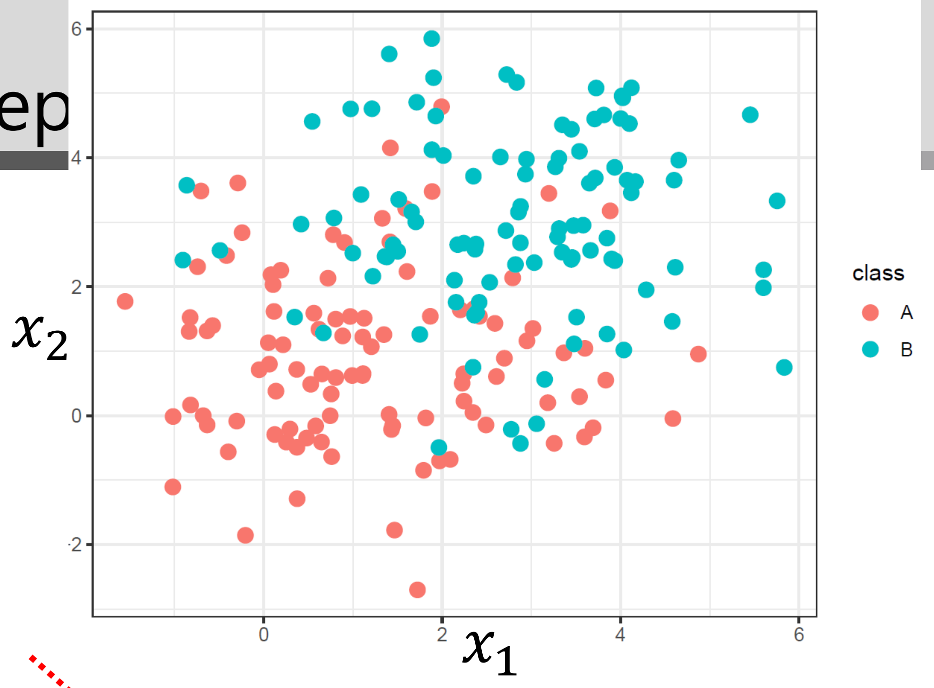
$p(\text{S} \mid \text{"Lunch Money Money Money"})$
$\propto p(\text{Lunch}|\text{S})p(\text{Money}|\text{S})^3 p(\text{S})$
$\propto 0.09 \cdot 0.46^3 \cdot 0.34 = 0.003$

https://youtu.be/O2L2Uv9pdDA

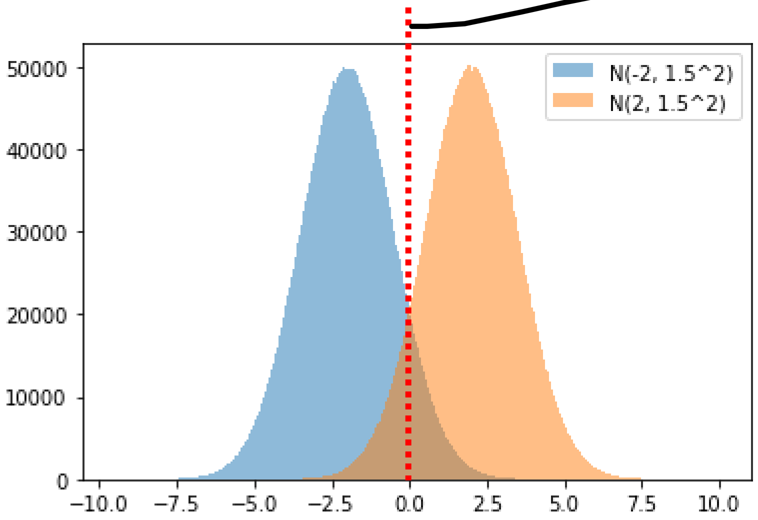
Discriminant Analysis
High-Level Concept

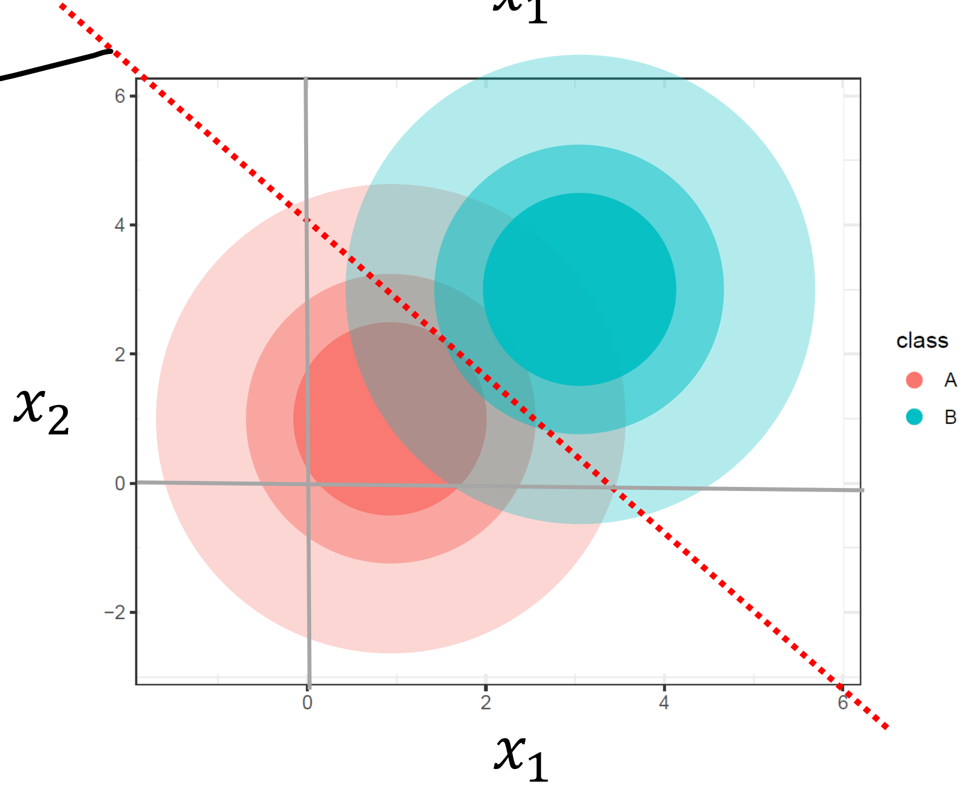
Discriminant Analysis: High-Level Concep

What's the *Discriminant Analysis* method?



Decision
boundary

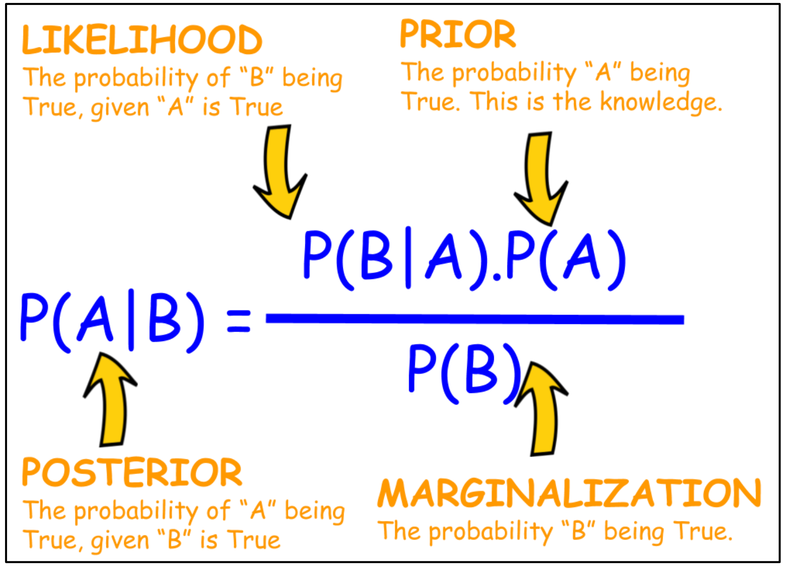




Discriminant Analysis

Bayes Theorem

- Let's reformulate it a bit with different notations.
- $\Pr(Y = k|X = x)$
  $= p_k(x)$
  $= \frac{\Pr(X=x|Y=k)\ \Pr(Y=k)}{\Pr(X=x)}$
  $= \frac{f_k(x)\pi_k}{f(x)} = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$



Discriminant Analysis

$$\Pr(Y = k|X = x) = p_k(x) = \frac{\Pr(X=x|Y=k)\Pr(Y=k)}{\Pr(X=x)} = \frac{f_k(x)\pi_k}{f(x)} = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

Discriminant Analysis belongs to *Sampling Paradigm*

- *Sampling Paradigm*
  Indirectly estimating $\Pr(Y = k \mid X = x)$ using the Bayes theorem

Example

Blue is modeled by $N(-2, 1.5^2)$

Orange is modeled by $N(2, 1.5^2)$

Discriminant Analysis

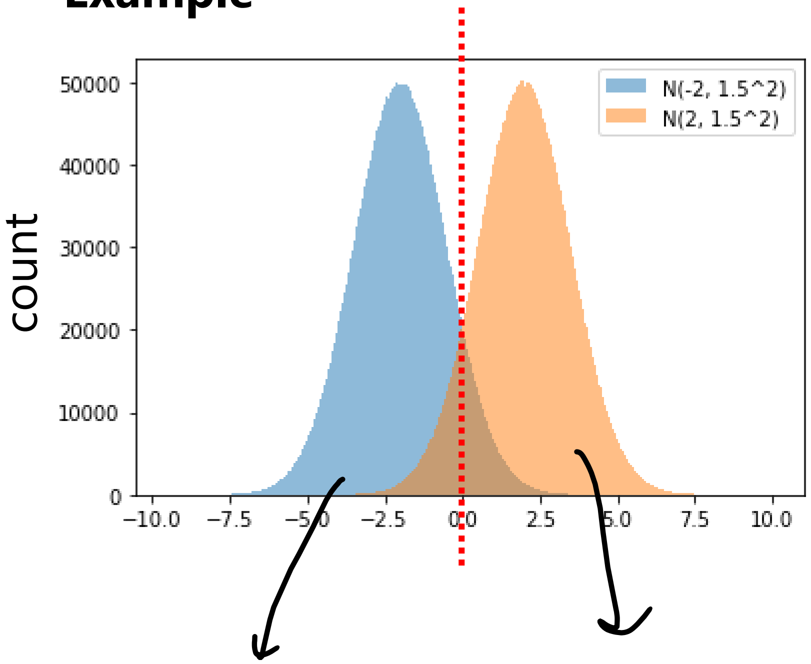
$$
\begin{aligned}
&\Pr(Y = k|X = x) \\
&= p_k(x) \\
&= \frac{\Pr(X=x|Y=k)\Pr(Y=k)}{\Pr(X=x)} \\
&= \frac{f_k(x)\pi_k}{f(x)} = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}
\end{aligned}
$$

Discriminant Analysis belongs to *Sampling Paradigm*

- *Sampling Paradigm*

  Indirectly estimating $\Pr(Y = k \mid X = x)$ using the Bayes theorem

Example



10,000,000 samples

10,000,000 samples

Blue is modeled by $N(-2, 1.5^2)$

Orange is modeled by $N(2, 1.5^2)$

Then, where's the classification boundary line?

How do we calculate it?

$$\Pr(Y = blue|X = x) = \Pr(Y = orange|X = x)$$

$$\frac{f_{blue}(x)\pi_{blue}}{f(x)} = \frac{f_{orange}(x)\pi_{orange}}{f(x)}$$

$$f_{blue}(x)\pi_{blue} = f_{orange}(x)\pi_{orange}$$

$$f_{blue}(x)\ 0.5 = f_{orange}(x)\ 0.5$$

$$f_{blue}(x) = f_{orange}(x)$$

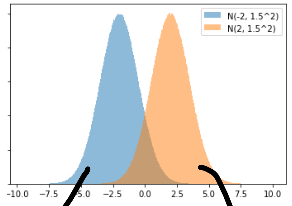
Discriminant Analysis - LDA

$$\Pr(Y = k|X = x) = p_k(x) = \frac{\Pr(X=x|Y=k)\Pr(Y=k)}{\Pr(X=x)} = \frac{f_k(x)\pi_k}{f(x)} = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

Linear discriminant analysis (LDA) when $p = 1$

- The class-conditional distributions $f_k(X)$ are assumed normal for $k = 1, \dots, K$, that is

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}$$



$f_{blue}(x)$ $f_{orange}(x)$

- LDA has the following *modeling assumption*:
  All classes have the same standard deviation $\sigma_k = \sigma$
  In other words, all $f_k(x)$ is modeled by $N(\mu_k, \sigma^2)$.

- Also, note that $\sum_{k=1}^{K} \pi_k = 1$, meaning $\Pr(Y = blue) + \Pr(Y = orange) = 1$.

- Then, how to make the classification prediction?

$$\Pr(Y = k|X = x) = p_k(x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\,\pi_l} = \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2} \pi_k}{\sum_{l=1}^{K} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_l}{\sigma}\right)^2} \pi_l}$$

- Classification is done by $\mathrm{argmax}_k(p_k(x))$

Discriminant Analysis - LDA

$$\Pr(Y = k|X = x)$$
$$= p_k(x)$$
$$= \frac{\Pr(X=x|Y=k)\Pr(Y=k)}{\Pr(X=x)}$$
$$= \frac{f_k(x)\pi_k}{f(x)} = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

Linear discriminant analysis (LDA) when $p = 1$

- Classification is done by $\mathrm{argmax}_k(p_k(x))$

$$\Pr(Y = k|X = x) = p_k(x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\,\pi_l} = \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2} \pi_k}{\sum_{l=1}^{K} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_l}{\sigma}\right)^2} \pi_l}$$

- We can 1) apply *log*, and 2) discard terms that don't depend on k, to make $p_k(x)$ simpler:

$$\log(p_k(x)) = \log\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2} \pi_k\right) - \log\left(\sum_{l=1}^{K} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2} \pi_l\right)$$

$$= \cancel{\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)} + \log\left(e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}\right) + \log(\pi_k) - \cancel{\log\left(\sum_{l=1}^{K} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_l}{\sigma}\right)^2} \pi_l\right)}$$

$$= -\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2 + \log(\pi_k) = -\frac{1}{2}\frac{x^2 - 2x\mu_k - \mu_k^2}{\sigma^2} + \log(\pi_k) = \cancel{-\frac{1}{2}\frac{x^2}{\sigma^2}} + \frac{1}{2}\frac{2x\mu_k}{\sigma^2} + \frac{1}{2}\frac{\mu_k^2}{\sigma^2} + \log(\pi_k)$$

$$= \frac{x\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) = \delta_k(x) \quad \cdots \quad \textit{discriminant score}$$

Discriminant Analysis - LDA

$$\Pr(Y = k|X = x)$$
$$= p_k(x)$$
$$= \frac{\Pr(X=x|Y=k)\,\Pr(Y=k)}{\Pr(X=x)}$$
$$= \frac{f_k(x)\pi_k}{f(x)} = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

Linear discriminant analysis (LDA) when $p = 1$

- Classification is originally done by $\mathrm{argmax}_k(p_k(x))$

$$\Pr(Y = k|X = x) = p_k(x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\,\pi_l} = \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2} \pi_k}{\sum_{l=1}^{K} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_l}{\sigma}\right)^2} \pi_l}$$

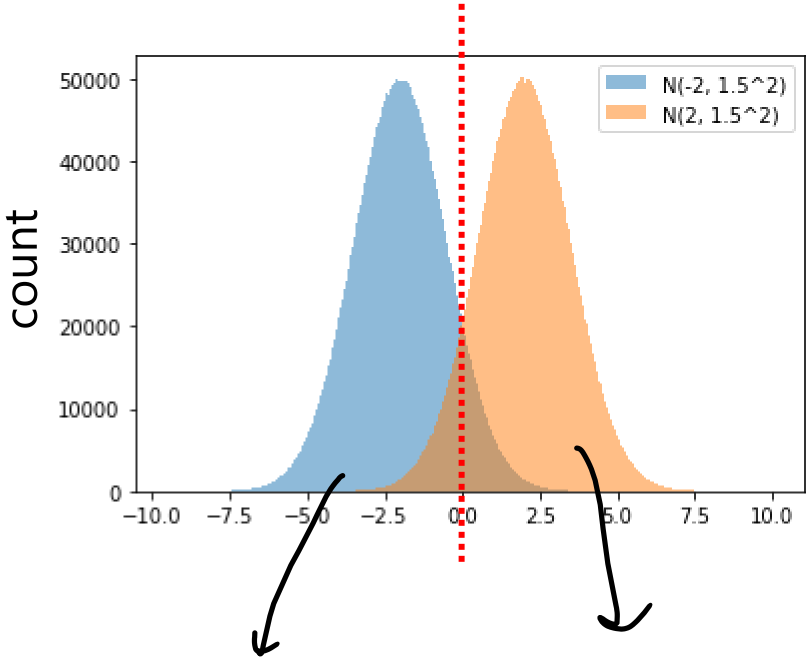
- Classification can also be performed by $\mathrm{argmax}_k\left(\delta_k(x)\right)$

$$\delta_k(x) = \frac{x\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) \qquad \cdots \textit{discriminant score}$$

- *i.e.*, $\mathrm{argmax}_k(p_k(x)) = \mathrm{argmax}_k\left(\delta_k(x)\right)$

Discriminant Analysis - LDA

Back to Example



10,000,000 samples

10,000,000 samples

Blue is modeled by $N(-2, 1.5^2)$

Orange is modeled by $N(2, 1.5^2)$

- $\pi_{blue} = \pi_{orange} = 0.5$

Let's do some classification predictions:

- when $x = -1$

$$\delta_k(x) = \begin{cases} \delta_{blue}(x) = \dfrac{-1\cdot(-2)}{1.5^2} - \dfrac{(-2)^2}{2\cdot 1.5^2} + \log(0.5) \approx -0.69 \\ \delta_{orange}(x) = \dfrac{-1\cdot(2)}{1.5^2} - \dfrac{(2)^2}{2\cdot 1.5^2} + \log(0.5) \approx -2.47 \end{cases}$$

- when $x = 0$

$$\delta_k(x) = \begin{cases} \delta_{blue}(x) = \dfrac{0\cdot(-2)}{1.5^2} - \dfrac{(-2)^2}{2\cdot 1.5^2} + \log(0.5) \approx -1.58 \\ \delta_{orange}(x) = \dfrac{0\cdot(2)}{1.5^2} - \dfrac{(2)^2}{2\cdot 1.5^2} + \log(0.5) \approx -1.58 \end{cases}$$

- when $x = 1$

$$\delta_k(x) = \begin{cases} \delta_{blue}(x) = \dfrac{1\cdot(-2)}{1.5^2} - \dfrac{(-2)^2}{2\cdot 1.5^2} + \log(0.5) \approx -2.47 \\ \delta_{orange}(x) = \dfrac{1\cdot(2)}{1.5^2} - \dfrac{(2)^2}{2\cdot 1.5^2} + \log(0.5) \approx -0.69 \end{cases}$$

$$\begin{aligned} &\Pr(Y = k|X = x) \\ &= p_k(x) \\ &= \frac{\Pr(X=x|Y=k)\Pr(Y=k)}{\Pr(X=x)} \\ &= \frac{f_k(x)\pi_k}{f(x)} = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l} \end{aligned}$$

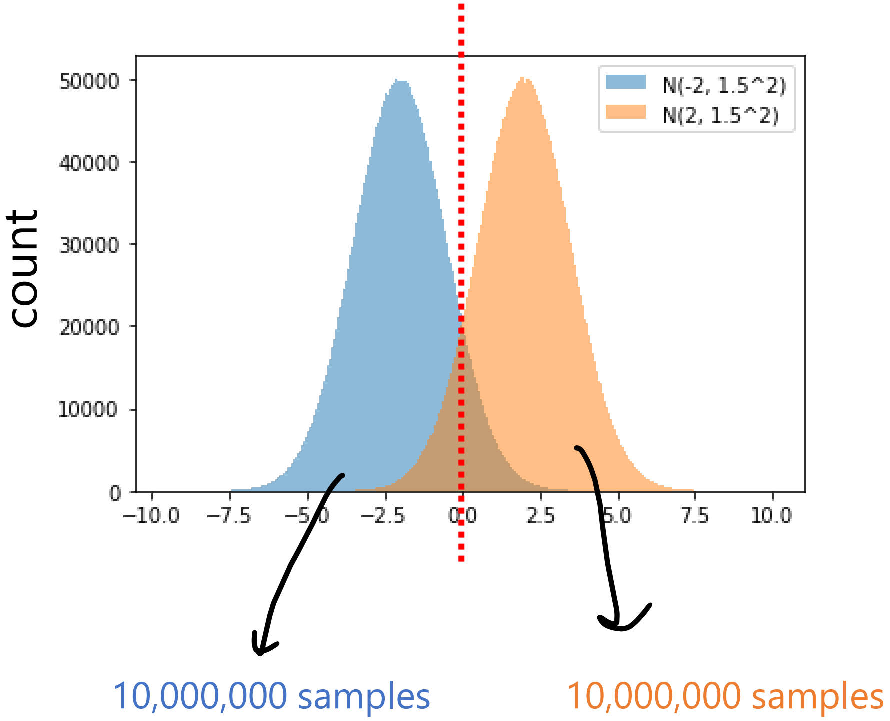
$$\boxed{\delta_k(x) = \frac{x\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)}$$

Discriminant Analysis - LDA

$$\delta_k(x) = \frac{x\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

$$\Pr(Y = k|X = x) = p_k(x) = \frac{\Pr(X=x|Y=k)\Pr(Y=k)}{\Pr(X=x)} = \frac{f_k(x)\pi_k}{f(x)} = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

Back to Example



The analytic solution for the boundary line:

$$\delta_{blue}(x) = \delta_{orange}(x)$$

$$\frac{x\mu_{blue}}{\sigma^2} - \frac{\mu_{blue}^2}{2\sigma^2} + \log(\pi_{blue}) = \frac{x\mu_{orange}}{\sigma^2} - \frac{\mu_{orange}^2}{2\sigma^2} + \log(\pi_{orange})$$

$$\frac{x\mu_{blue}}{\sigma^2} - \frac{\mu_{blue}^2}{2\sigma^2} = \frac{x\mu_{orange}}{\sigma^2} - \frac{\mu_{orange}^2}{2\sigma^2}$$

$$x\mu_{blue} - \frac{\mu_{blue}^2}{2} = x\mu_{orange} - \frac{\mu_{orange}^2}{2}$$

$$x\mu_{blue} - x\mu_{orange} = \frac{\mu_{blue}^2}{2} - \frac{\mu_{orange}^2}{2}$$

$$x(\mu_{blue} - \mu_{orange}) = \frac{\mu_{blue}^2 - \mu_{orange}^2}{2} = \frac{(\mu_{blue} - \mu_{orange})(\mu_{blue} + \mu_{orange})}{2}$$

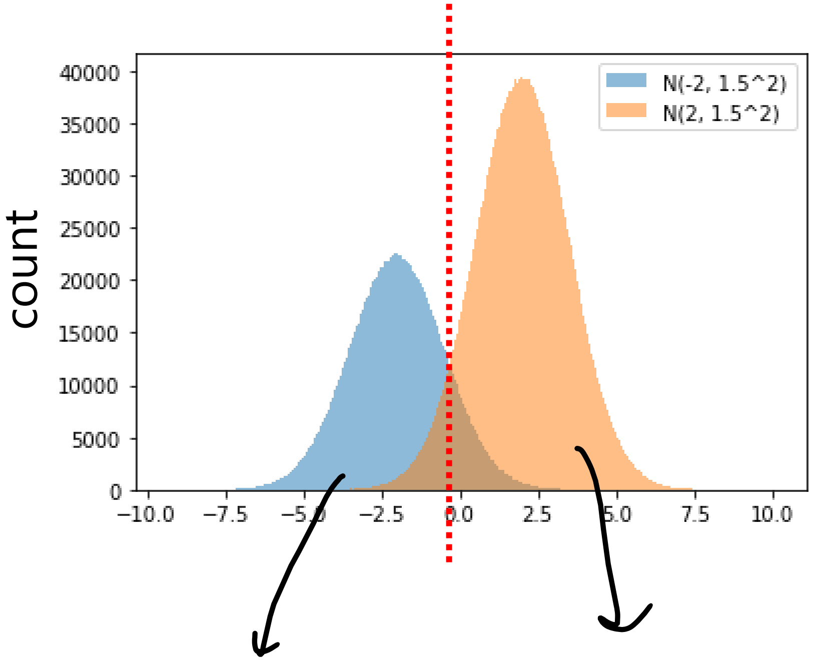
$$\therefore x = \frac{(\mu_{blue} + \mu_{orange})}{2}$$

Discriminant Analysis - LDA

$$\Pr(Y = k|X = x) = p_k(x) = \frac{\Pr(X=x|Y=k)\Pr(Y=k)}{\Pr(X=x)} = \frac{f_k(x)\pi_k}{f(x)} = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

$$\delta_k(x) = \frac{x\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

Another Example



3,000,000 samples

7,000,000 samples

Blue is modeled by $N(-2, 1.5^2)$

Orange is modeled by $N(2, 1.5^2)$

- $\pi_{blue} = 0.3,\ \pi_{orange} = 0.7$

Let's do some classification predictions:

- when $x = -1$

$$\delta_k(x) = \begin{cases} \delta_{blue}(x) = \dfrac{-1 \cdot (-2)}{1.5^2} - \dfrac{(-2)^2}{2 \cdot 1.5^2} + \log(0.3) \approx -1.2 \\ \delta_{orange}(x) = \dfrac{-1 \cdot (2)}{1.5^2} - \dfrac{(2)^2}{2 \cdot 1.5^2} + \log(0.7) \approx -2.13 \end{cases}$$

- when $x = 0$

$$\delta_k(x) = \begin{cases} \delta_{blue}(x) = \dfrac{0 \cdot (-2)}{1.5^2} - \dfrac{(-2)^2}{2 \cdot 1.5^2} + \log(0.3) \approx -2.09 \\ \delta_{orange}(x) = \dfrac{0 \cdot (2)}{1.5^2} - \dfrac{(2)^2}{2 \cdot 1.5^2} + \log(0.7) \approx -1.25 \end{cases}$$

- when $x = 1$

$$\delta_k(x) = \begin{cases} \delta_{blue}(x) = \dfrac{1 \cdot (-2)}{1.5^2} - \dfrac{(-2)^2}{2 \cdot 1.5^2} + \log(0.3) \approx -2.98 \\ \delta_{orange}(x) = \dfrac{1 \cdot (2)}{1.5^2} - \dfrac{(2)^2}{2 \cdot 1.5^2} + \log(0.7) \approx -0.36 \end{cases}$$

Discriminant Analysis - LDA

$$\Pr(Y = k|X = x)$$
$$= p_k(x)$$
$$= \frac{\Pr(X=x|Y=k)\Pr(Y=k)}{\Pr(X=x)}$$
$$= \frac{f_k(x)\pi_k}{f(x)} = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

$$\delta_k(x) = \frac{x\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

In real life, we often don't know μ_k, σ, π_k

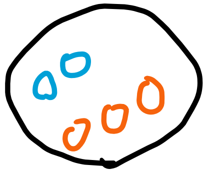
- In the previous examples,
  we knew the true distributions, therefore do know not μ_k, σ, π_k.
  But typically, we don't know them. We only have a training dataset
- **Idea:** We can *estimate the parameters* given the training dataset: $\hat{\mu}_k, \hat{\sigma}, \hat{\pi}_k$

Discriminant Analysis - LDA

Estimation of the Parameters $\hat{\mu}_k, \hat{\sigma}, \hat{\pi}_k$

- $\hat{\pi}_k = \frac{n_k}{n}$

  where n is a total number of observations and n_k is a number of observations that belong to class k.



- $\pi_{blue} = 2/5$
- $\pi_{orange} = 3/5$

- $\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$

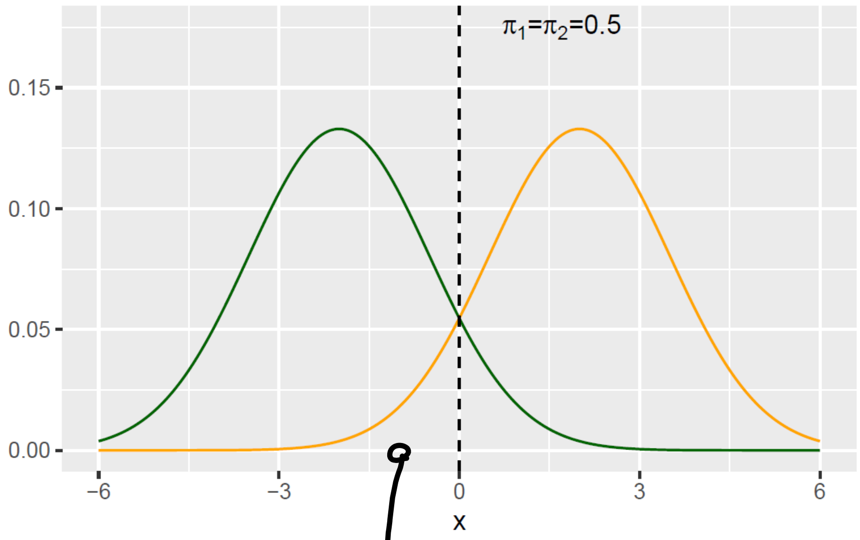
- $\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$

$$\Pr(Y = k|X = x) = p_k(x) = \frac{\Pr(X=x|Y=k)\,\Pr(Y=k)}{\Pr(X=x)} = \frac{f_k(x)\pi_k}{f(x)} = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

$$\delta_k(x) = \frac{x\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

Performance Assessment: Bayes Error Rate (BER)

- Bayes Error Rate (BER) is defined as $1 - \mathrm{E}\left[\max_k \Pr(Y = k \mid X)\right]$



$$\begin{cases} p(Y = green \mid X = x) = 0.7 \\ p(Y = yellow \mid X = x) = 0.3 \end{cases}$$

Then, $\max_k \Pr(Y = k \mid X) = 0.7$

Therefore, BER is $1 - 0.7 = 0.3$

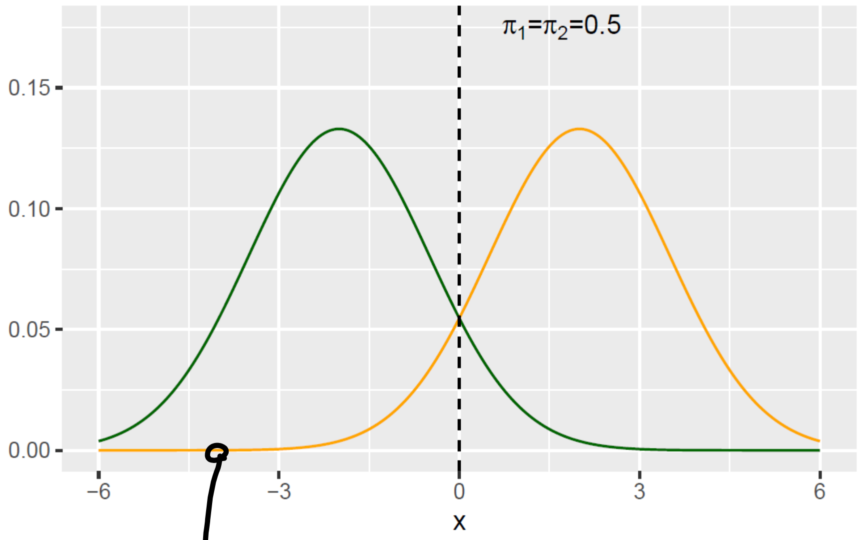
Discriminant Analysis - LDA

$$\Pr(Y = k|X = x) = p_k(x) = \frac{\Pr(X=x|Y=k)\,\Pr(Y=k)}{\Pr(X=x)} = \frac{f_k(x)\pi_k}{f(x)} = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

$$\delta_k(x) = \frac{x\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

Performance Assessment: Bayes Error Rate (BER)

- Bayes Error Rate (BER) is defined as $1 - \mathrm{E}\left[\max_k \Pr(Y = k \mid X)\right]$



$$\begin{cases} p(Y = green \mid X = x) = 1.0 \\ p(Y = yellow \mid X = x) = 0.0 \end{cases}$$

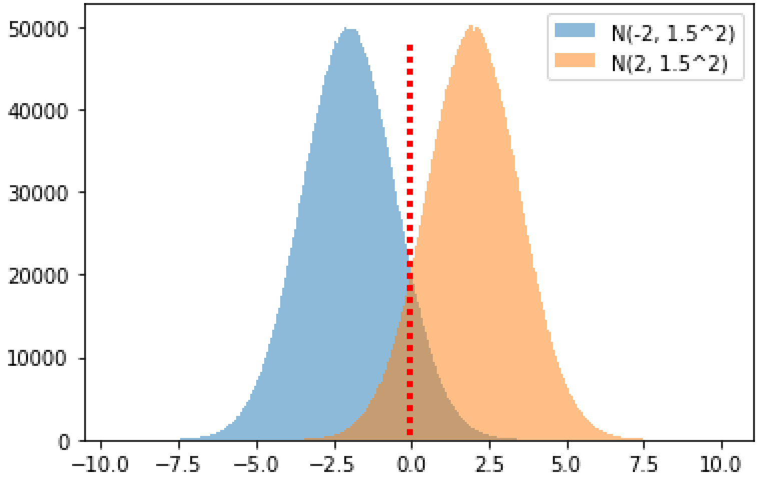
Then, $\max_k \Pr(Y = k \mid X) = 1.0$. Therefore, BER is $1 - 1.0 = 0$.

- We do this for all x since we have expectation E.

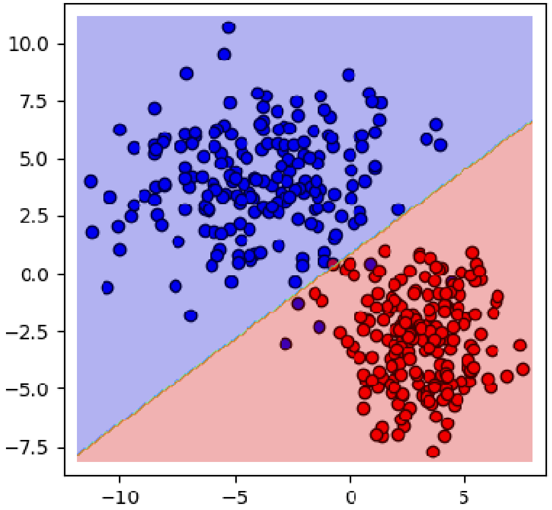
Multivariate LDA

Multivariate LDA

LDA



Multivariate LDA



$\Pr(Y = k|X = x)$

$$= p_k(x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\,\pi_l} = \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2} \pi_k}{\sum_{l=1}^{K} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_l}{\sigma}\right)^2} \pi_l}$$

$$\log(p_k(x)) \propto \delta_k(x) = \frac{x\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

Replace the univariate normal distribution with the multivariate normal distribution:

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

Then, we get

$$\delta_k(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^T\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \log(\pi_k)$$

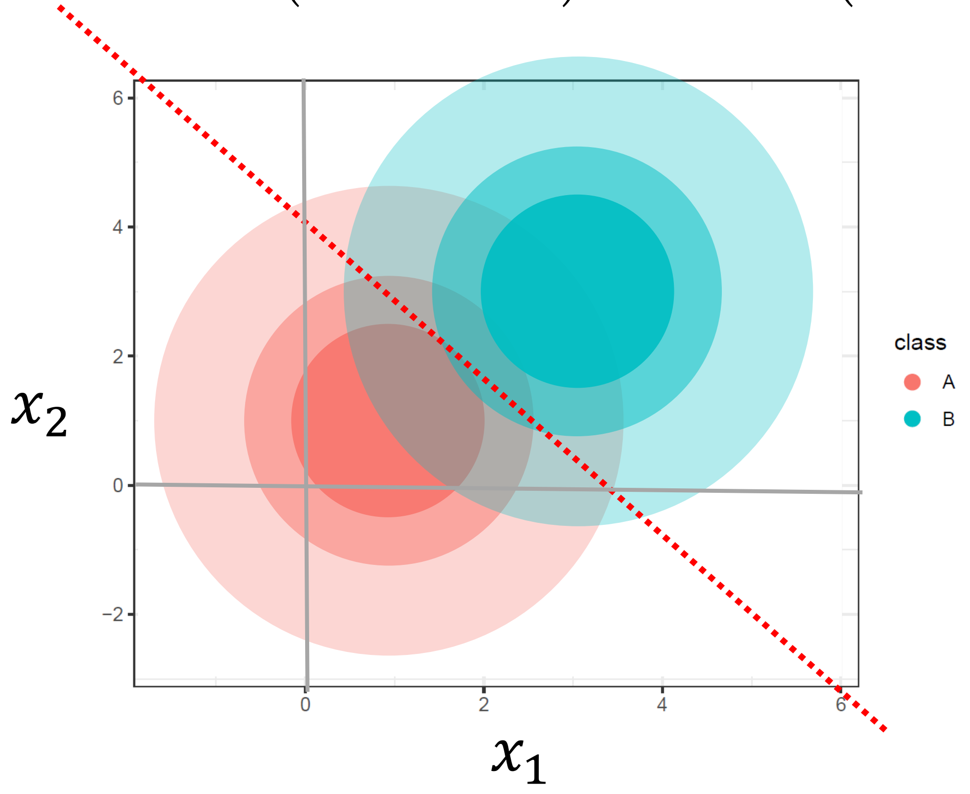
Multivariate LDA

Example

Discriminant score

$$\delta_k(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^T\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \log(\pi_k)$$

- We now have two normal distributions in 2-dimensional space
  - $x_{red} \sim N\left((1,1)^T, \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}\right)$ $\quad x_{blue} \sim N\left((3,3)^T, \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}\right)$



We can analytically find the boundary line by solving

$$\delta_{red}(\boldsymbol{x}) = \delta_{blue}(\boldsymbol{x})$$

$$\boldsymbol{x}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{red} - \frac{1}{2}\boldsymbol{\mu}_{red}^T\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{red} + \log(\pi_{red}) = \boldsymbol{x}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{blue} - \frac{1}{2}\boldsymbol{\mu}_{blue}^T\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{blue} + \log(\pi_{blue})$$

Then, it gives

$$x_2 = 4 - x_1$$

Multivariate LDA

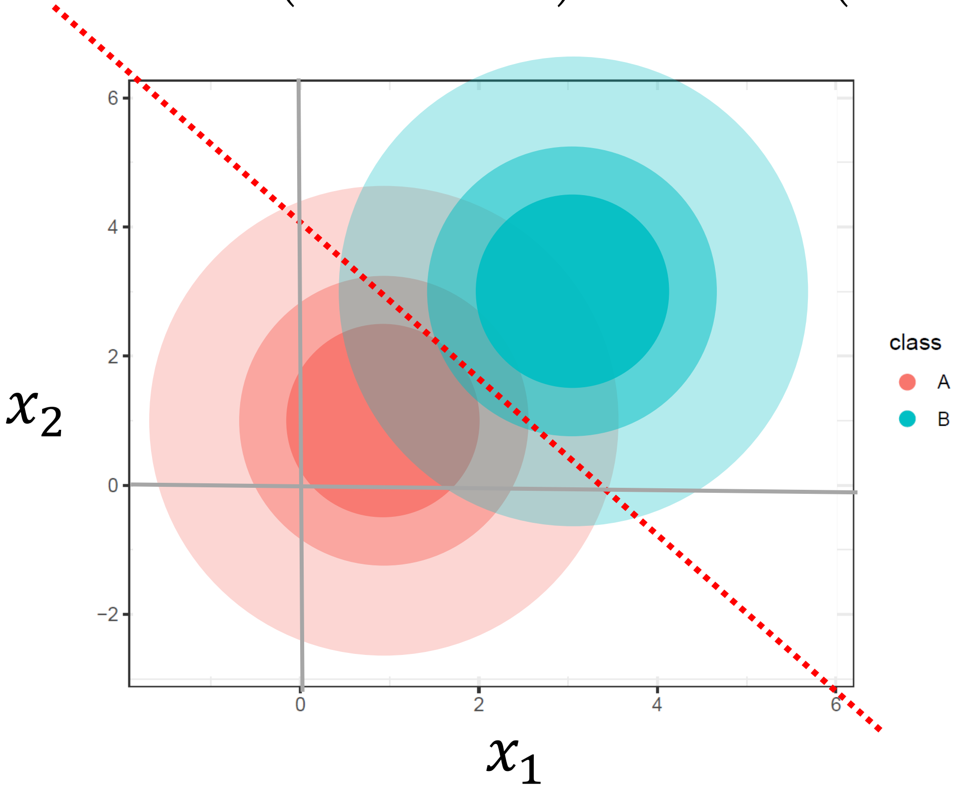
Discriminant score

$$\delta_k(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^T\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \log(\pi_k)$$

Example

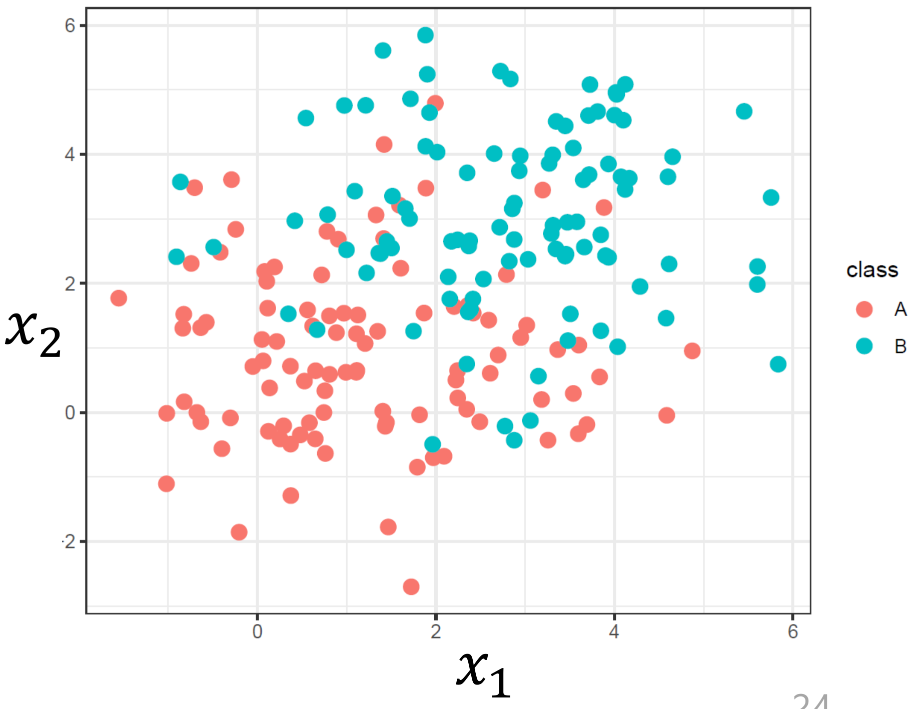
- We now have two normal distributions in 2-dimensional space

$$x_{red} \sim N\left((1,1)^T, \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}\right) \qquad x_{blue} \sim N\left((3,3)^T, \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}\right)$$



In real life, we do not know the true distribution.

We just have a training set.



Parameter Estimation & Classification with Multivariate LDA

- *In real life, we do not know the true distribution. We just have a training set.*

- To make the classification prediction, we first estimate $\widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}, \hat{\pi}_k$

$$\widehat{\boldsymbol{\mu}}_k = \frac{1}{n_k} \sum_{i: y_i = k} \boldsymbol{X}_i$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n - K} \sum_{k=1}^{K} \sum_{i: y_i = k} (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_k)(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_k)^T$$

$$\hat{\pi}_k = \frac{n_k}{n}$$

where n is a total number of observations and n_k is a number of observations that belong to class k.

- Then, we can compute the discriminative score $\delta_k(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^T\, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \log(\pi_k)$

- And the classification is performed by $\mathrm{argmax}_k \delta_k(\boldsymbol{x})$

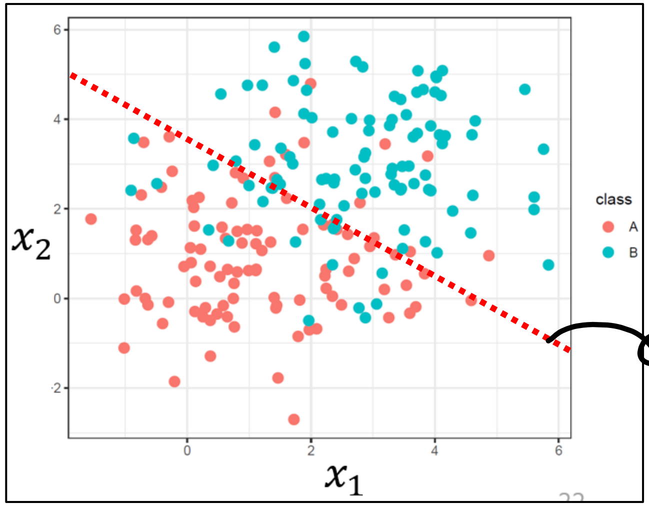
Multivariate LDA

LDA in R

```
r.lda <- lda(class ~ X1 + X2, df)
r.pred <- predict(r.lda, df)$class
table(real = df$class, predicted = r.pred)
```

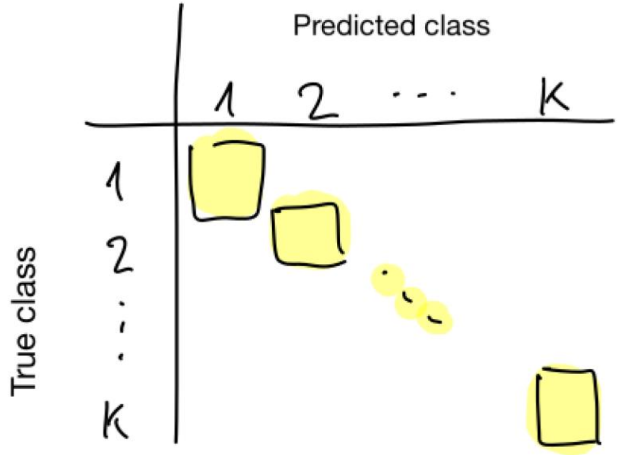
```
##       predicted
## real   A   B
##      A 87  13
##      B 18  82
```

Confusion matrix
(important performance metric)



Estimated decision boundary

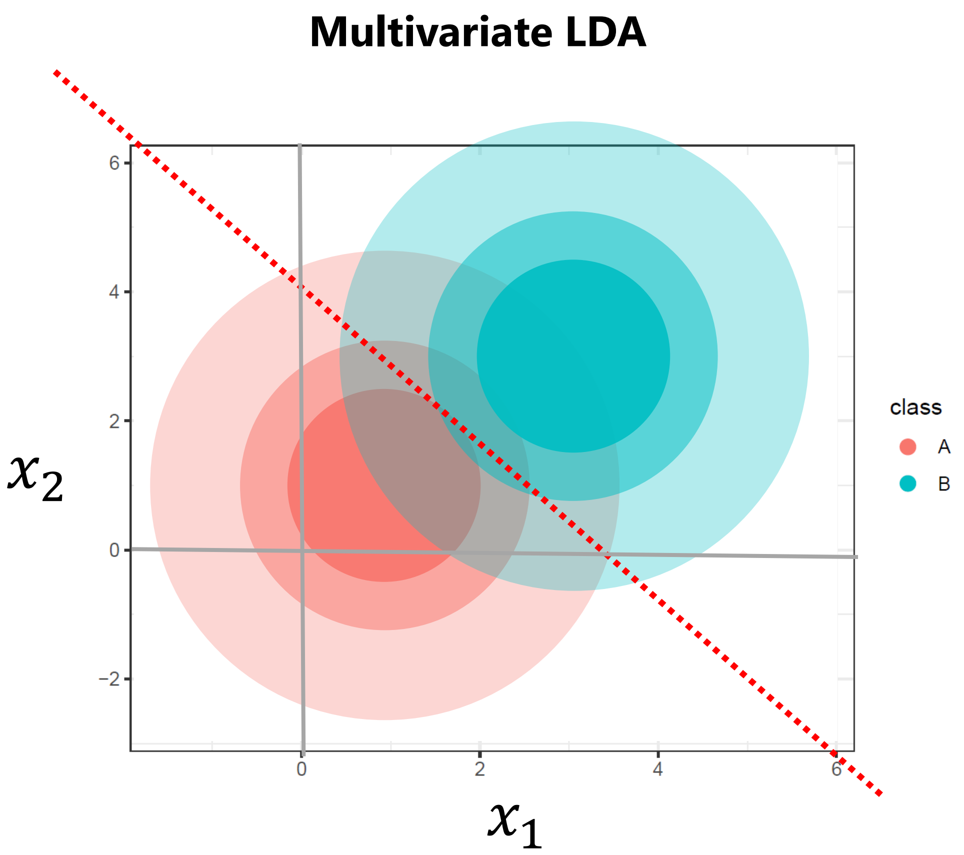
General confusion matrix forms:

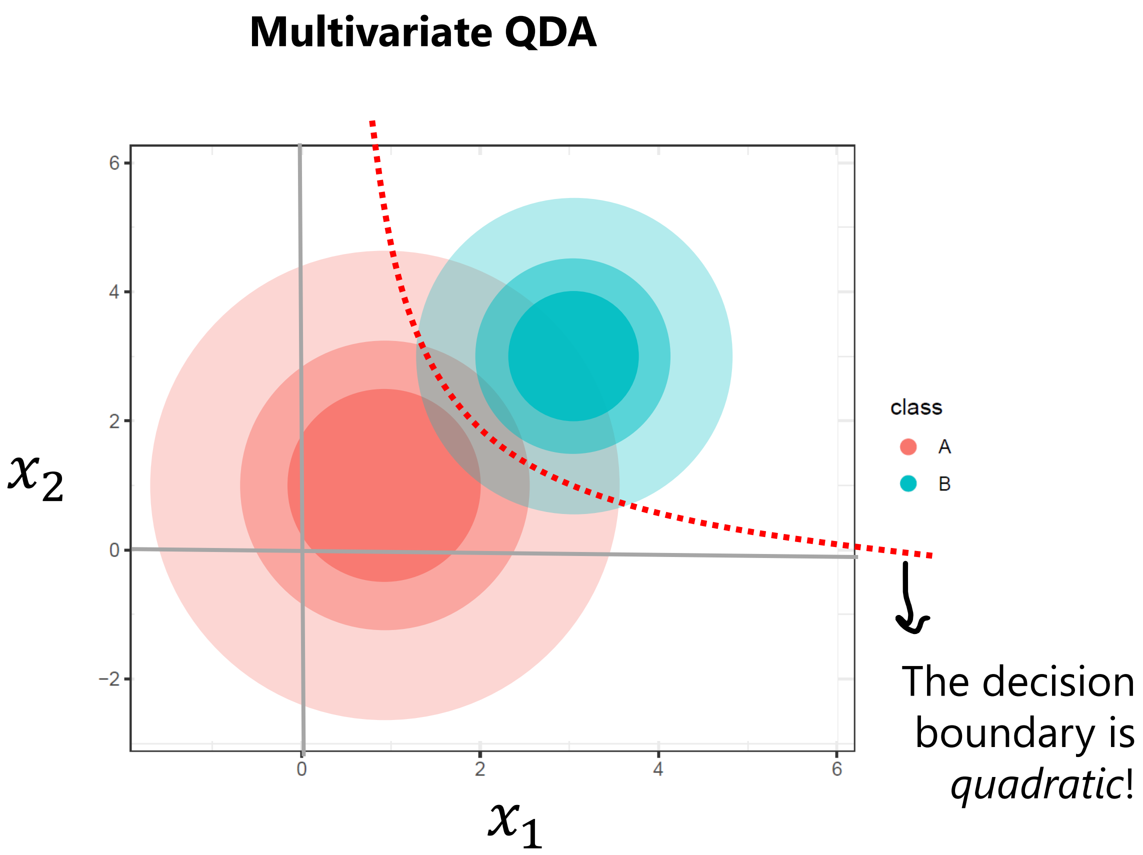


Quadratic Discriminant Analysis (QDA)

Quadratic Discriminant Analysis (QDA)

- In LDA, we assumed that $\Sigma_k = \Sigma$ for all classes (modeling assumption).
- In QDA, we allow different covariance matrices Σ_k for each class.

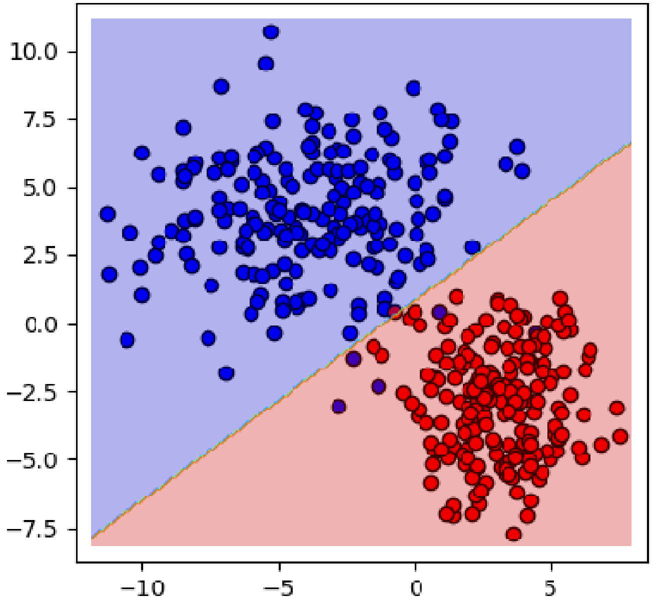




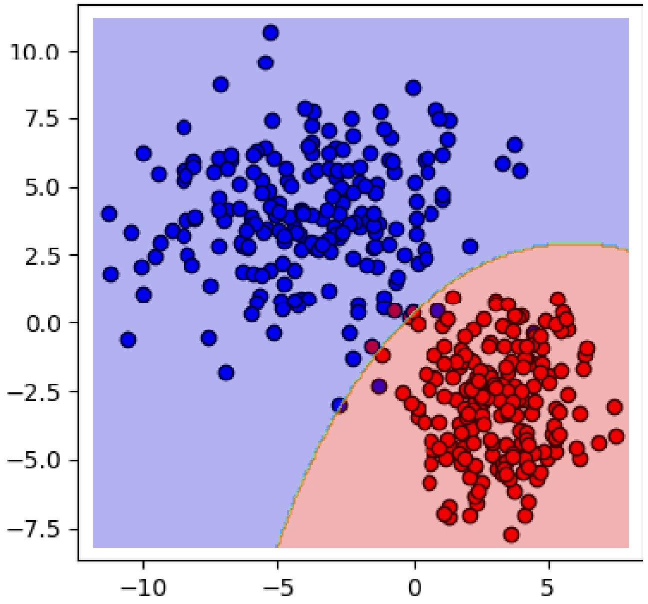
Quadratic Discriminant Analysis (QDA)

- In LDA, we assumed that $\Sigma_k = \Sigma$ for all classes.
- In QDA, we allow different covariance matrices Σ_k for each class.

Multivariate LDA



Multivariate QDA



Multivariate LDA

Univariate LDA

$\Pr(Y = k|X = x)$

$$= p_k(x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\,\pi_l} = \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2} \pi_k}{\sum_{l=1}^{K} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_l}{\sigma}\right)^2} \pi_l}$$

$$\log(p_k(x)) \propto \delta_k(x) = \frac{x\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

Univariate QDA

$\Pr(Y = k|X = x)$

$$= p_k(x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\,\pi_l} = \frac{\frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2} \pi_k}{\sum_{l=1}^{K} \frac{1}{\sqrt{2\pi}\sigma_l} e^{-\frac{1}{2}\left(\frac{x-\mu_l}{\sigma_l}\right)^2} \pi_l}$$

$$\log(p_k(x)) \propto \delta_k(x) = \frac{x\mu_k}{\sigma_k^2} - \frac{\mu_k^2}{2\sigma_k^2} + \log(\pi_k)$$

Quadratic Discriminant Analysis (QDA)

Multivariate QDA

- $\Pr(Y = k|X = \boldsymbol{x}) = p_k(\boldsymbol{x}) = \frac{f_k(\boldsymbol{x})\pi_k}{\sum_{l=1}^{K} f_l(\boldsymbol{x})\pi_l}$

  where $f_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k{}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\right)$

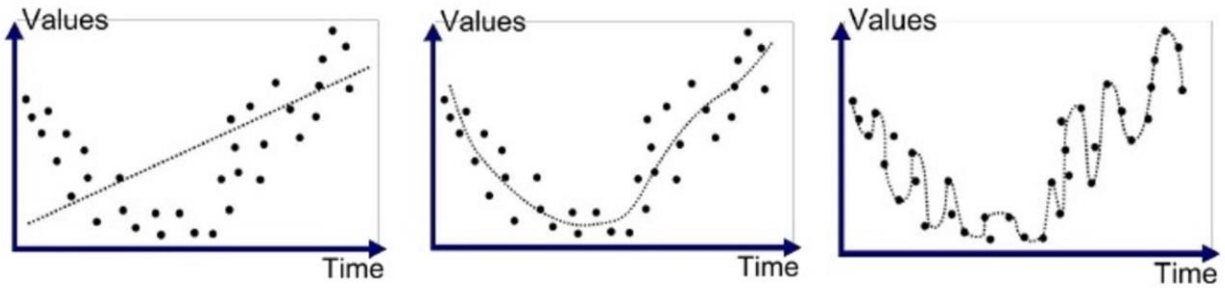
- $\delta_k(x) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k{}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k) - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| + \log \pi_k$

  $= -\frac{1}{2}\boldsymbol{x}^T \boldsymbol{\Sigma}_k{}^{-1}\boldsymbol{x} + \boldsymbol{x}^T \boldsymbol{\Sigma}_k{}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^T \boldsymbol{\Sigma}_k{}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| + \log \pi_k$

Quadratic Discriminant Analysis (QDA)

LDA vs QDA

- QDA is more flexible than LDA, as it allows for class-specific covariance matrices Σ_k.
- Should we always prefer QDA to LDA?

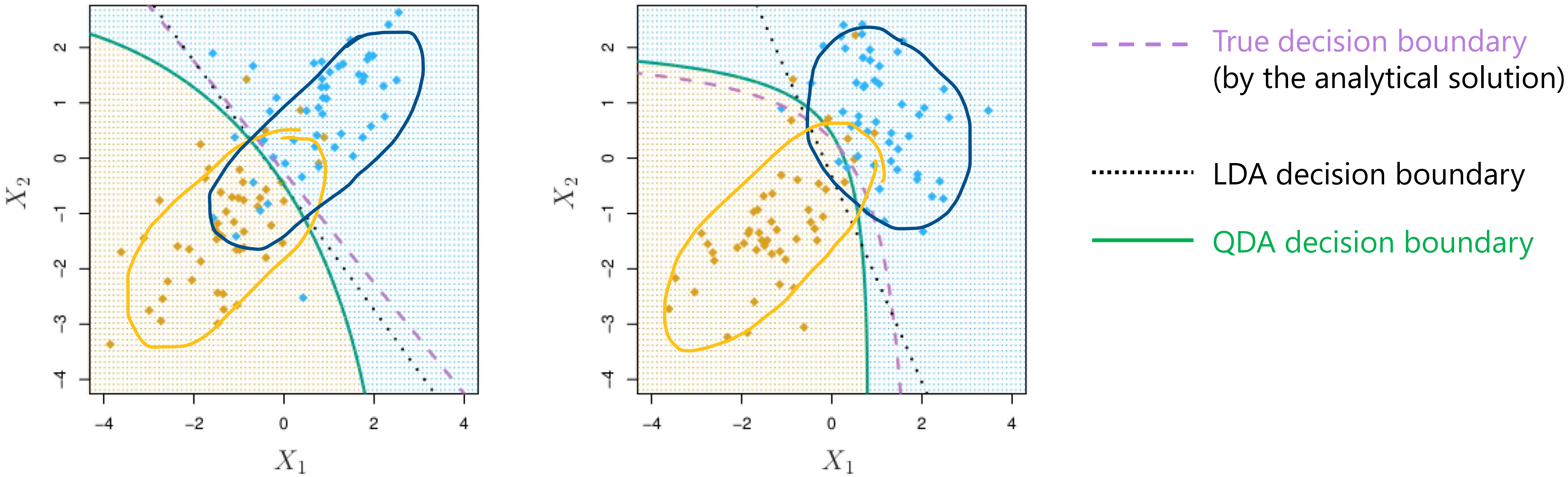


High Bias
Low Variance

Low Bias
High Variance

LDA vs QDA

- QDA is more flexible than LDA, as it allows for class-specific covariance matrices Σ_k.
- Should we always prefer QDA to LDA?



Example w/ LDA, QDA

Iris dataset

The `iris` flower data set was introduced by the British statistician and biologist Ronald Fisher in 1936.

- **Three plant species**: {setosa, virginica, versicolor}.
- **Four features**: `Sepal.Length`, `Sepal.Width`, `Petal.Length` and `Petal.Width`.

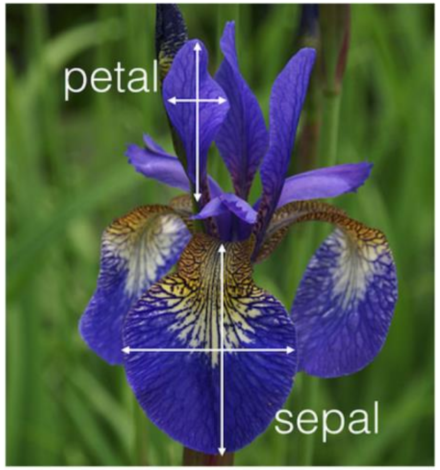


Figure 1: Iris plant with sepal and petal leaves

http://blog.kaggle.com/2015/04/22/scikit-learn-video-3-machine-learning-first-steps-with-the-iris-dataset/

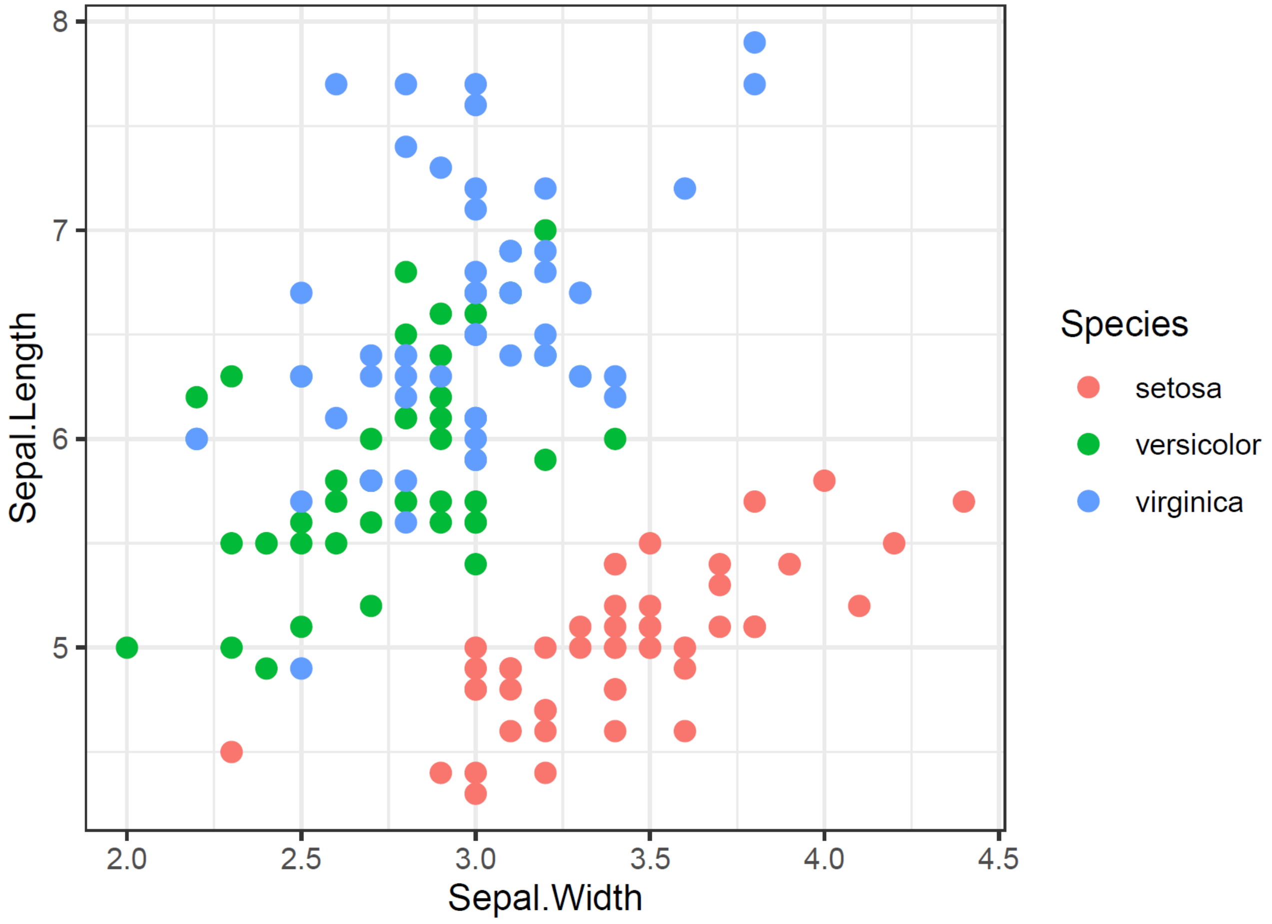
Example: Classification of iris plants

We will use `sepal width` and `sepal length` to build a classificator. We have 50 observations from each class.

```
attach(iris)
head(iris)
```

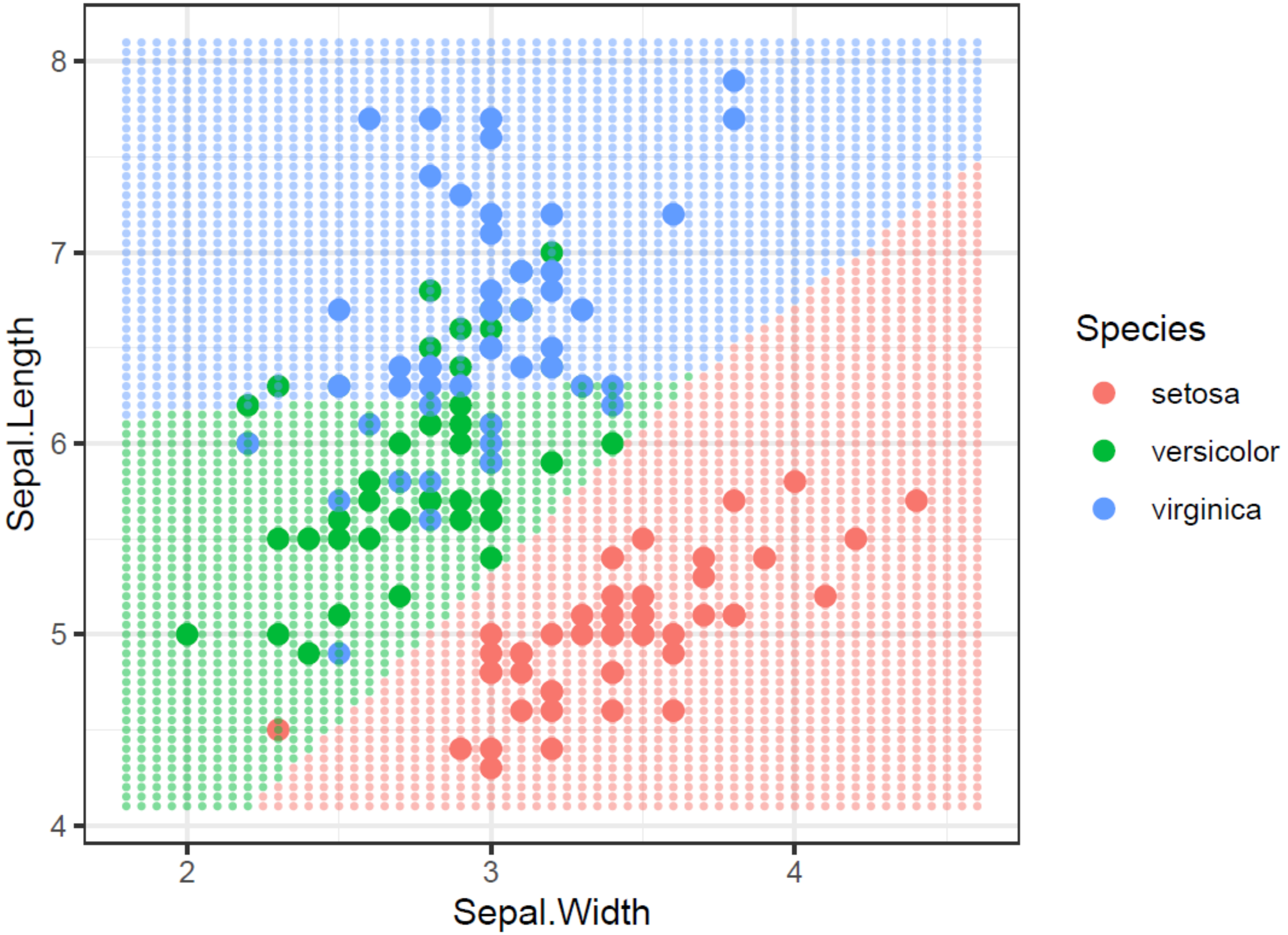
```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

Example w/ LDA, QDA



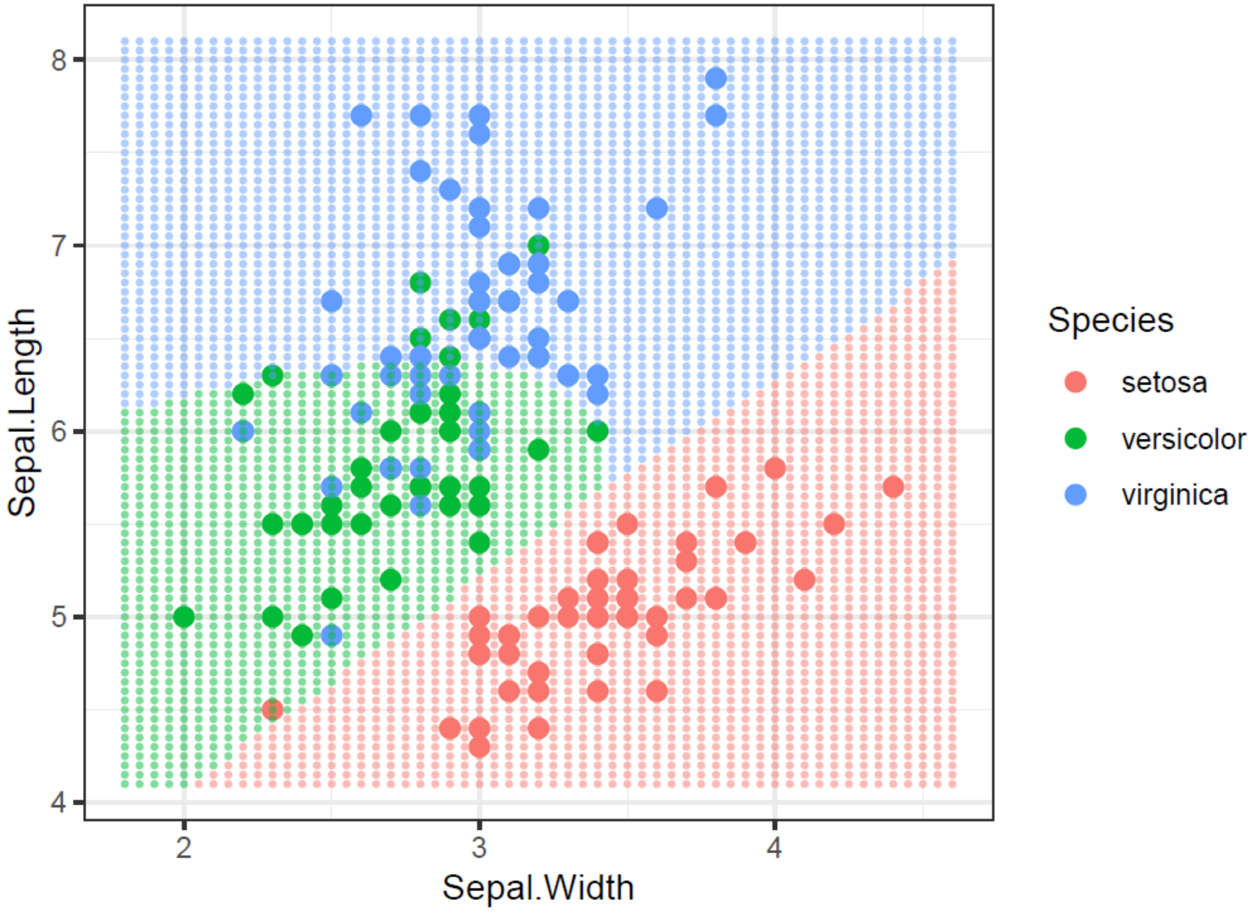
Example w/ LDA, QDA

```
```{r, echo=T}
library(MASS)
iris_lda = lda(Species~Sepal.Length+Sepal.Width, data=iris)
```
```



Example w/ LDA, QDA

```
```{r, echo=T}
iris_qda = qda(Species~Sepal.Length + Sepal.Width, data=iris)
```
```



Example w/ LDA, QDA

Compare LDA and QDA by their prediction accuracies

- Let's create $\boldsymbol{X}_{train}$ and $\boldsymbol{X}_{test}$

```
set.seed(1)
train = sample(1:150, 75)

iris_train = iris[train, ]
iris_test = iris[-train, ]
```

- Fit LDA and QDA on $\boldsymbol{X}_{train}$

```
iris_lda2 = lda(Species~Sepal.Length + Sepal.Width, data=iris_train)

iris_qda2 = qda(Species~Sepal.Length + Sepal.Width, data=iris_train)
```

Example w/ LDA, QDA

Test LDA and QDA on X_{test}

LDA training error: $\frac{14}{75} = 0.19$

```
table(predict(iris_lda2, newdata = iris_train)$class, iris_train$Species)
```

```
##
##              setosa versicolor virginica
##   setosa         27          0         0
##   versicolor      1         15         8
##   virginica       0          5        19
```

LDA test error: $\frac{19}{75} = 0.26$.

```
iris_lda2_predict = predict(iris_lda2, newdata = iris_test)
table(iris_lda2_predict$class, iris$Species[-train])
```

```
##
##              setosa versicolor virginica
##   setosa         22          0         0
##   versicolor      0         22        11
##   virginica       0          8        12
```

QDA training error: $\frac{13}{75} = 0.17$.

```
table(predict(iris_qda2, newdata = iris_train)$class, iris_train$Species)
```

```
##
##              setosa versicolor virginica
##   setosa         28          0         0
##   versicolor      0         16         9
##   virginica       0          4        18
```

QDA test error: $\frac{24}{75} = 0.32$.

```
iris_qda2_predict = predict(iris_qda2, newdata = iris_test)
table(iris_qda2_predict$class, iris$Species[-train])
```

```
##
##              setosa versicolor virginica
##   setosa         22          0         0
##   versicolor      0         18        12
##   virginica       0         12        11
```

Summary of Classification Methods

Summary of Classification Methods

Diagnostic Paradigm: directly estimate $\Pr(Y = k \mid X = x)$

- Logistic regression
- KNN (K Nearest Neighbors)

Sampling Paradigm: indirectly estimate $\Pr(\mathrm{Y} = \mathrm{k} \mid \mathrm{X} = \mathrm{x})$ using Bayes Theorem

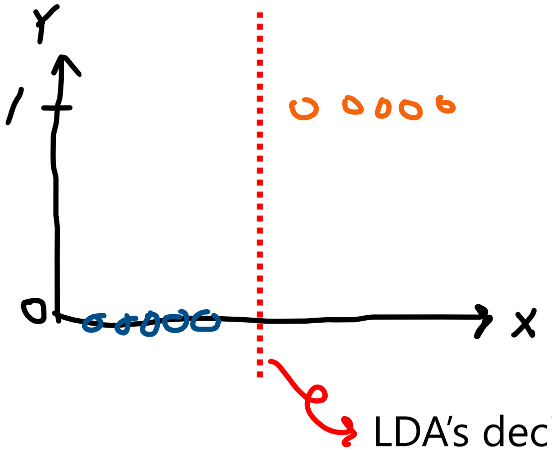
- Naïve Bayes
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)

Which classification method is the best?

Which classification method is the best?

Advantages of Discriminant Analysis

- Linear Discriminant Analysis (LDA) is more stable than logistic regression *when the classes are well-separated*.



LDA's decision boundary

- Moreover, LDA is more popular for *multi-class classification*.

  (Logistic regression is essentially for binary classification)

Which classification method is the best?

Linearity Property

- Assume a binary classification problem with one covariate
- Recall that logistic regression can be written:

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$$

- For a two-class problem, one can show that for LDA:

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = \cdots = \beta_0 + \beta_1 x$$

thus the same linear form. The difference is in how the parameters are estimated.

Which classification method is the best?

LDA vs Logistic regression

- In practice, the results are often very similar, but
  - LDA is more “more available” in the multi-class setting.
  - if the class conditional distributions are multivariate normal, then LDA (or QDA) is preferred.
  - Logistic regression makes no assumptions about the covariates and is therefore to be preferred in many practical applications.
  - In medicine, for two-class problems, logistic regression is often preferred (for interpretability) and (always) together with ROC and AUC (for model comparison).

and KNN?

- KNN is used when the class boundaries are non-linear.
- Yet, remember the curse of dimensionality (when p is large)!

Which classification method is the best?

The answer is "It depends!"

- Logistic regression is very popular for binary classification.
- LDA is useful when n is small, or the classes are well separated, and Gaussian assumptions are reasonable. Also when $K > 2$.
- Naïve Bayes is useful when p is very large.
- KNN is non-parametric, thus no assumptions about the decision boundary nor the distribution of the variables. Expected to work better than LDA and logistic regression when boundary is very non-linear.
    - Caveats:
      1) no interpretation of the effect of the covariates is possible.
      2) curse of dimensionality.
- Please read Section 4.5 of the coursebook (James et al. 2013)

Two-class problems: Sensitivity, Specificity

Two-class problems: Sensitivity, Specificity

Predicted Class

		Positive	**Negative**		
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP + FN)}$	*True Positive Rate*
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN + FP)}$	*True Negative Rate*
		Precision $\frac{TP}{(TP + FP)}$	**Negative Predictive Value** $\frac{TN}{(TN + FN)}$	**Accuracy** $\frac{TP + TN}{(TP + TN + FP + FN)}$	

Two-class problems: Sensitivity, Specificity

Example

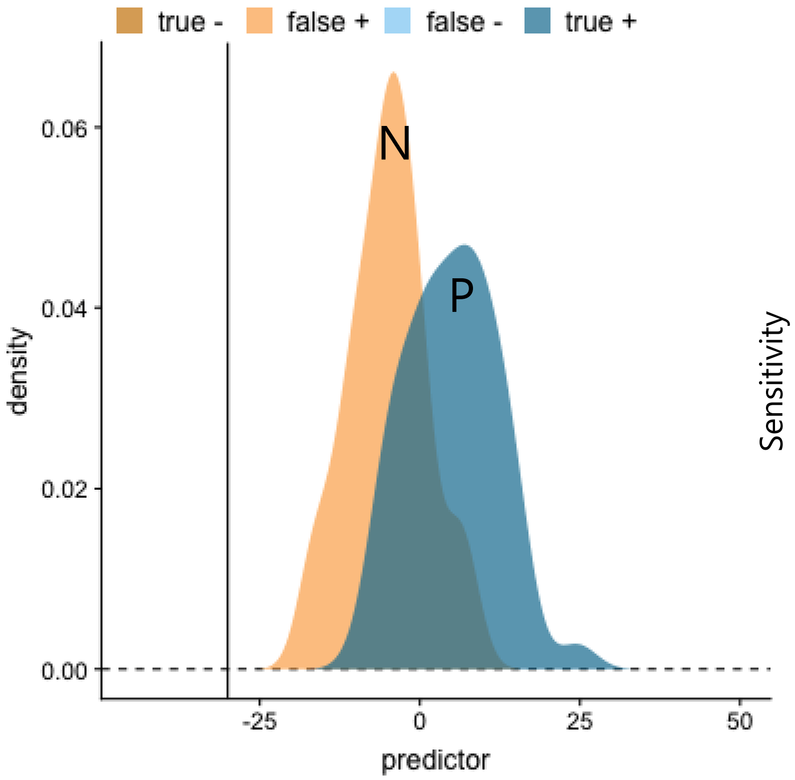
		Predicted Class	
		Spam	**Non-Spam**
Actual Class	**Spam**	TP=45	FN=20
	Non-Spam	FP=5	TN=30

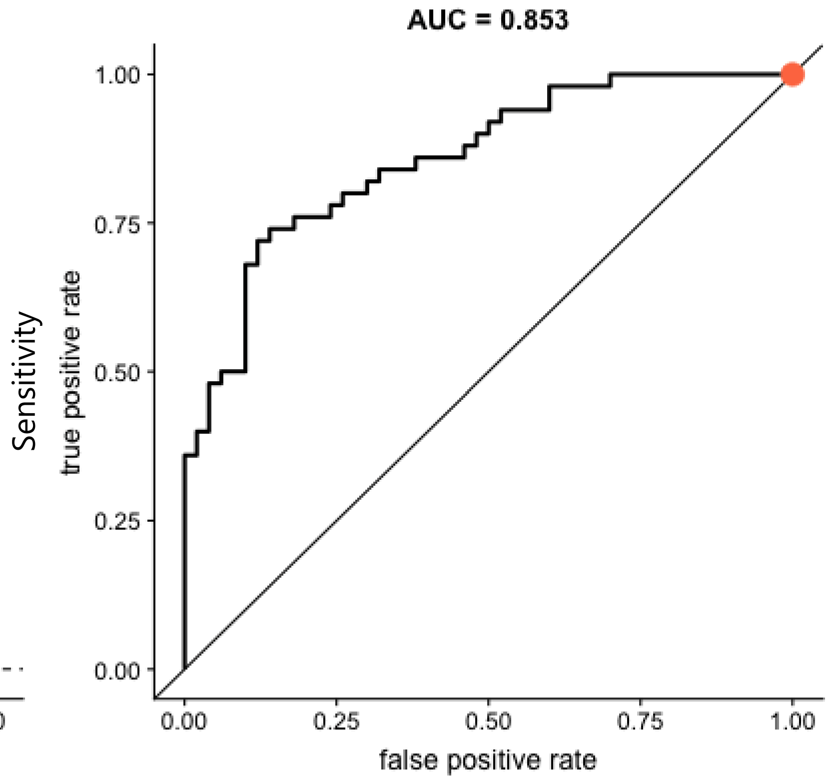
$$\text{Test error} = \frac{20+5}{45+20+5+30} = 0.25$$

$$\text{Sensitivity } (\textit{True Positive Rate}) = \frac{45}{45+20}$$

$$\text{Specificity } (\textit{True Negative Rate}) = \frac{30}{5+30}$$

ROC curve & AUC

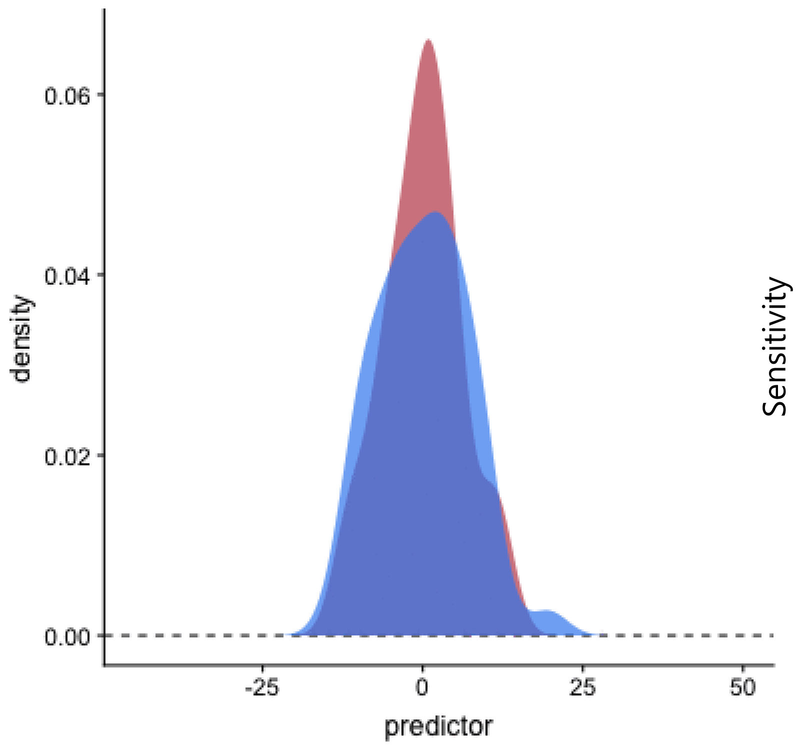




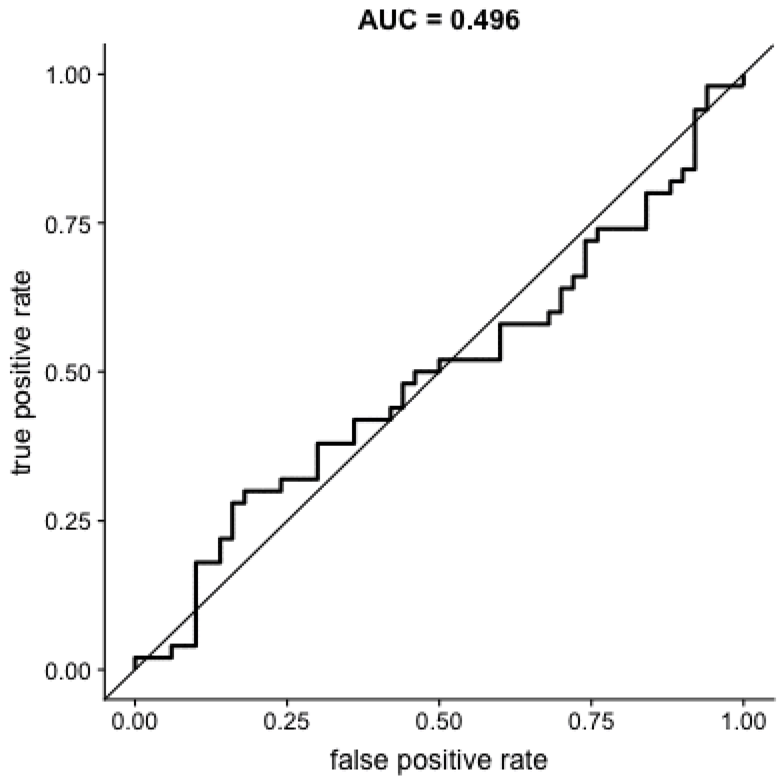
1-Specificity
= 1 - TrueNegativeRate

Two-class problems: Sensitivity, Specificity

ROC curve & AUC



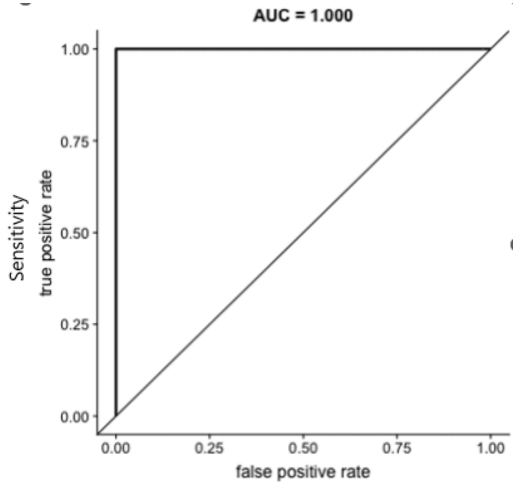
Sensitivity



1-Specificity
= 1 - TrueNegativeRate

ROC curve and AUC

- ROC (receiver operating characteristics) curve gives a graphical display of the sensitivity against (1-specificity), as the threshold value (decision boundary) is moved from 0 to 1.
- An ideal classifier will give:



- The AUC ranges between 0 and 1.
  - 0 denotes completely-wrong, 1 denotes completely-correct.
- AUC is useful for comparing the performance of different classifiers.

Calendar week	Module	Date	Weekday	Topic	Resp. Instructor
2	**Module 1**	09.01.23	Monday	Introduction	Steffi
		12.01.23	Thursday	R-course	Self-study
		12.01.23	Thursday	R-course	Self-study
3	**Module 2**	16.01.23	Monday	Statistical learning 1	Daesoo
		19.01.23	Thursday	Statistical learning 2	Daesoo
		19.01.23	Thursday	Statistical learning RecEx	Emma
4	**Module 3**	23.01.23	Monday	Linear regression 1	Daesoo
		26.01.23	Thursday	Linear regression 2	Daesoo
		26.01.23	Thursday	Linear regression RecEx	Kenneth
5	**Module 4**	30.01.23	Monday	Classification 1	Daesoo
		02.02.23	Thursday	Classification 2	Daesoo
		02.02.23	Thursday	Classification RecEx	Emma
6	**Module 5**	06.02.23	Monday	Resampling 1	Steffi
		09.02.23	Thursday	Resampling 2	Steffi / Emma (Rmd and ggplot)
		09.02.23	Thursday	Resampling RecEx	Kenneth
7		13.02.23	Monday	Compulsory Exercise 1	Emma and Kenneth
		16.02.23	Thursday	Compulsory Exercise 1	All
		16.02.23	Thursday	Compulsory Exercise 1	All
8	**Module 6**	20.02.23	Monday	Model Selection, Regularization 1	Steffi
		23.02.23	Thursday	Model Selection, Regularization 2	Steffi
		23.02.23	Thursday	Model Selection, Regularization RecEx	Daesoo
9	**Module 7**	27.02.23	Monday	Moving Beyond Linearity	Steffi


The End!