Linear Classification 1 - LDA

Julian Karch

Universiteit Leiden
The Netherlands

Overview

生成式

- Intro Generative Classification + Naïve Bayes
- Linear Discriminant Analysis 判别式
- Quadratic Discriminant Analysis and Regularized LDA

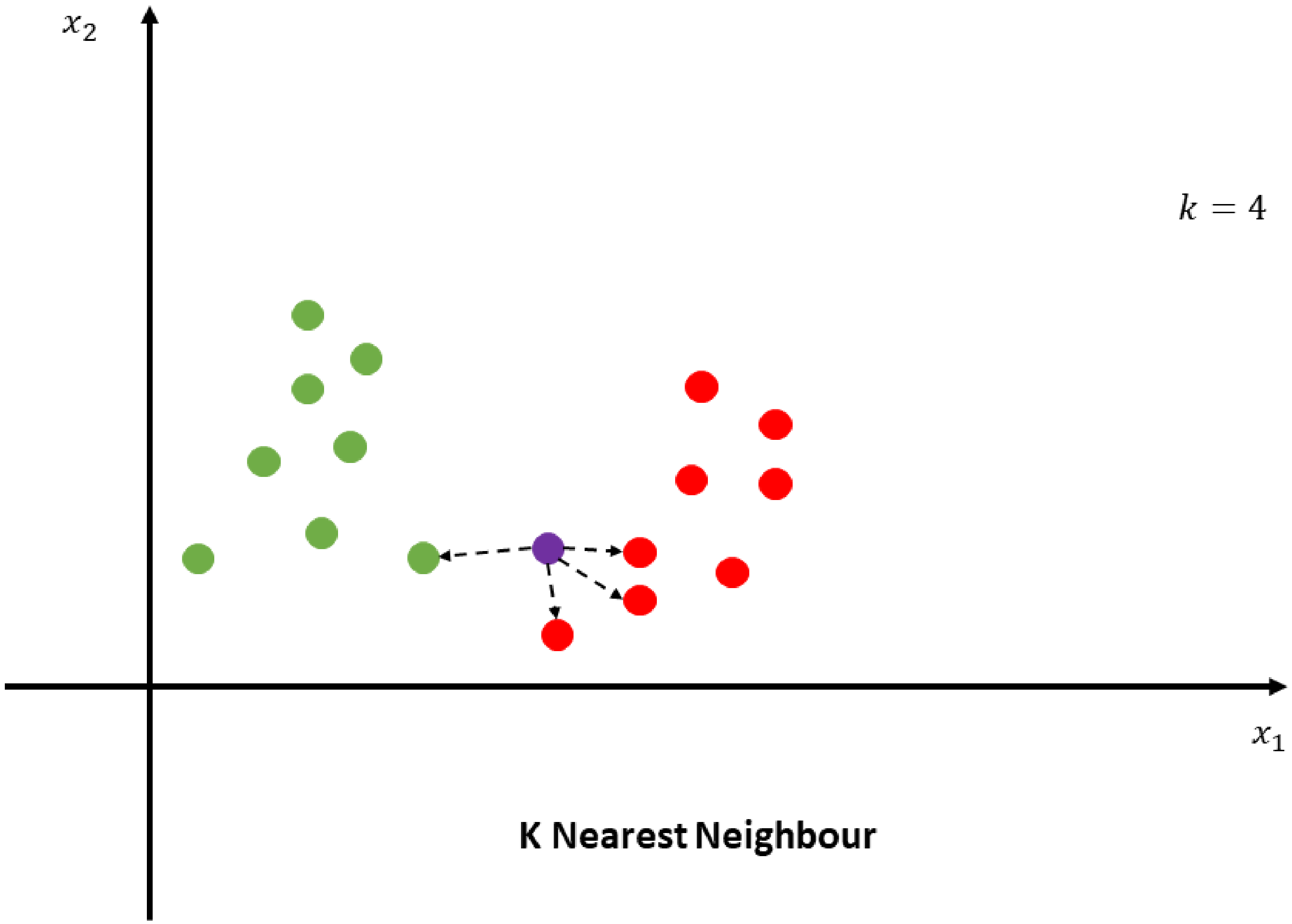
Intro + Naïve Bayes

Universiteit Leiden
The Netherlands

Classification

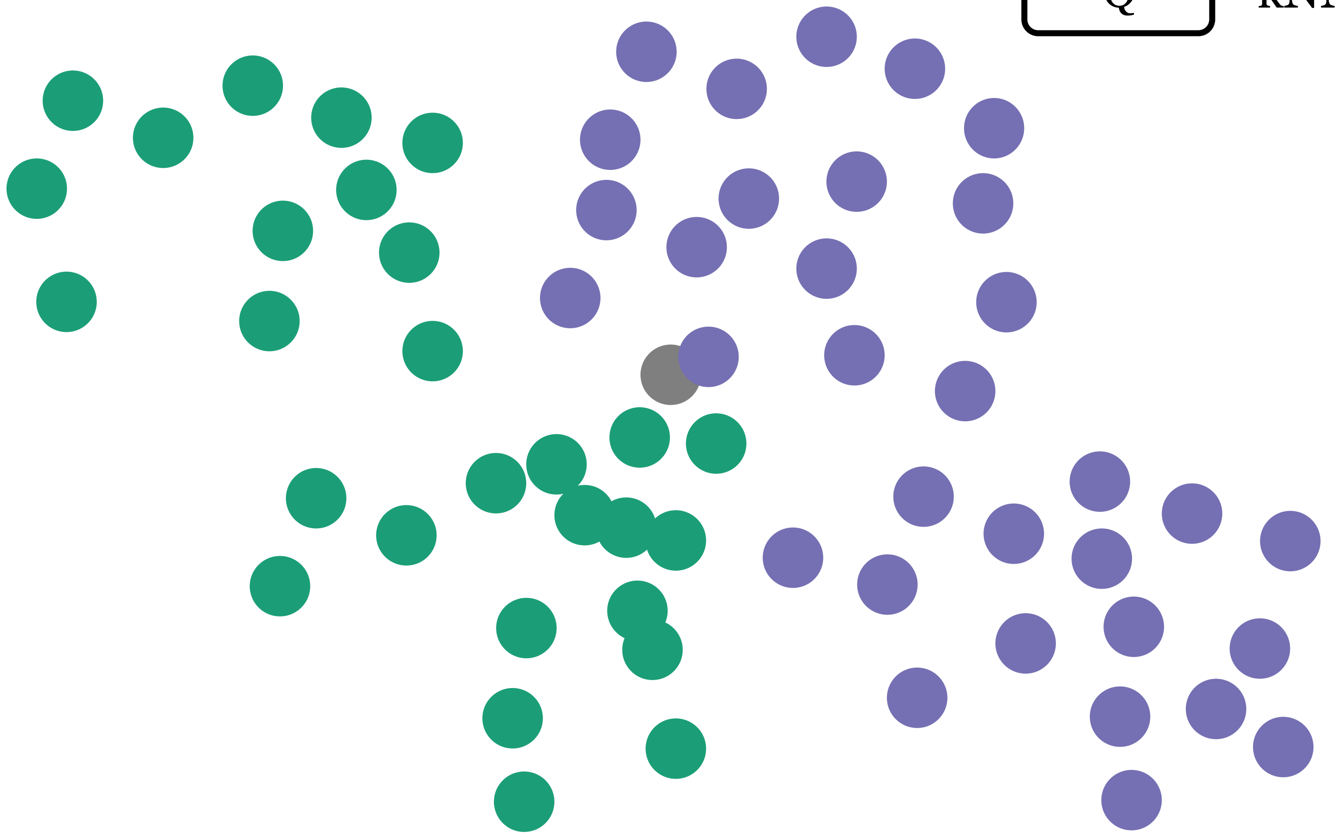
- Last week regression: $\mathrm{Y} \in \mathbb{R}, \hat{Y} = f(X)$
- This week classification: $\mathrm{Y} \in \{1, \dots, \mathrm{K}\}, \hat{Y} = C(X)$

K Nearest Neighbors



K Nearest Neighbour

K Nearest Neighbors



Q

kNN with k=3 classifies the grey point to which class?

Expected Prediction Error Regression

- Last week: For regression, we want f that minimizes
$$MSE_{pred} = \mathbb{E}[(f(X) - Y)^2]$$
- More general: $\mathrm{EPE} = \mathbb{E}[L(f(X), Y)]$
- L is a **loss function** that quantifies how "bad" wrong predictions are
- We get MSE_{pred} by using $L_{sq}\left(\hat{Y}, Y\right) = (\hat{Y} - Y)^2$
- Another option: $L_{abs}\left(\hat{Y}, Y\right) = |\hat{Y} - Y|$

Expected Prediction Error Classification

$$\text{EPE} = \mathbb{E}[L(C(X), Y)]$$

- Most popular loss is 0-1 loss

$L_{01}(C(\mathrm{X}), y) = 0$ if $y = C(x)$, otherwise $L_{01}(C(\mathrm{X}), y) = 1$

Posterior Probability

- $P(\text{no rain}|\text{data}) = 80\%$
- More generally: $P(Y|X)$ is called *posterior probability*

Q Would you act like it would rain tomorrow?

Bayes Optimal Classification

- We minimize EPE by assigning the class that has the highest posterior probability
- This is called the *Bayes classifier* and is optimal in the sense that it minimizes EPE (for 0-1 loss) among **all** classifiers
- Note that sometimes 0-1 loss is not what we want
  - Example: falsely predicting “no rain” can be more costly than falsely prediction “rain”

From Data to Posterior

Universiteit Leiden
The Netherlands

Corona Test

- Corona Test X (positive or negative), Health Status Y (corona or healthy)

sensitivity test

- $P(X = \text{positive}|Y = \text{Corona}) = 0.9$
- $P(Y = \text{Corona}) = 0.02, P(Y = \text{Healthy}) = 0.98$
- $P(X = \text{positive}|Y = \text{Healthy}) = 0.05$

求 $P(Y=\text{Corona}|X=\text{positive}), P(Y=\text{Healthy}|X=\text{pos})$

Q

- **What is $P(X = \text{positive}, Y = \text{Corona})$?** $0.9 \times 0.02 = 0.018$
- **What is $P(X = \text{positive}, Y = \text{Healthy})$?** $0.05 \times 0.98 = 0.049$
- **What would you classify (assume 0-1 loss)?**

Bayes Rule

$P(Y=\text{Corona} \mid X=\text{pos}) = \frac{P(X=\text{pos}, Y=\text{Corona})}{P(X=\text{pos})}$

$P(Y=\text{Healthy} \mid X=\text{pos}) = \frac{P(X=\text{pos}, Y=\text{Healthy})}{P(X=\text{pos})}$

- Want: $P(Y = \text{Corona} \mid X = \text{positive})$
- $P(Y = \text{Corona}|X = \text{positive}) = \frac{P(X=positive, Y=Corona)}{P(X=positive)}$
- $P(X = positive)$ does not dependent on Y

- → Assigning to class y for which $P(X = positive, Y = \text{y})$ is highest is equivalent to assigning for which $P(Y = \text{y}|\ X = \text{positive})$ is highest

Bayes Rule

- $P(X = \text{positive}) = P(X = \text{positive}, Y = \text{Corona}) +$
$$P(X = \text{positive}, Y = \text{Healthy})$$
- $P(Y = \text{Healthy} | X = positive) = \frac{.049}{.018+.049} = 0.73$

Intermediate Summary

- Posterior probabilities $P(Y|X)$ enable easy classification
- Bayes rule allows obtaining posterior probabilities from prior probabilities $P(Y)$ and likelihoods $P(X|Y)$

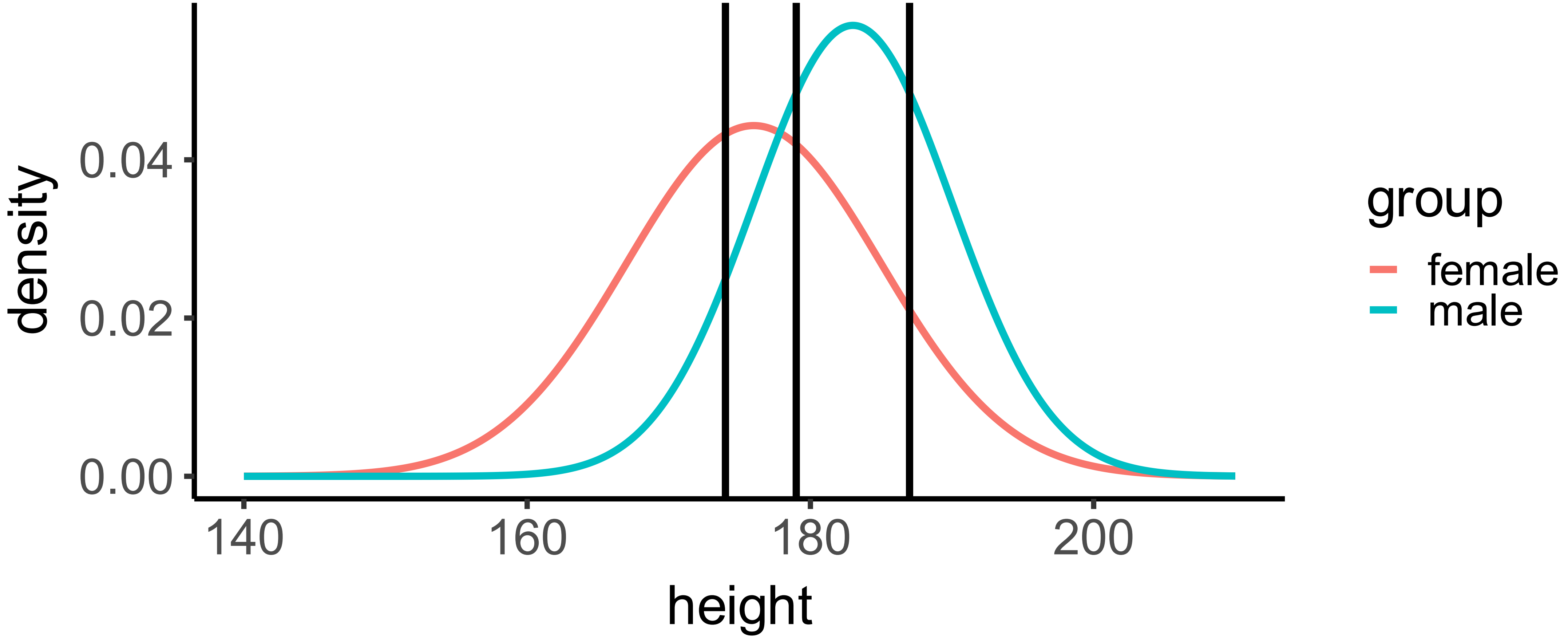
$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

- Note that this optimality requires that we know the likelihood ($P(X|Y)$) and prior probabilites [$P(Y)$]
- In pratice, we especially virtually never know the likelihood $P(X|Y)$

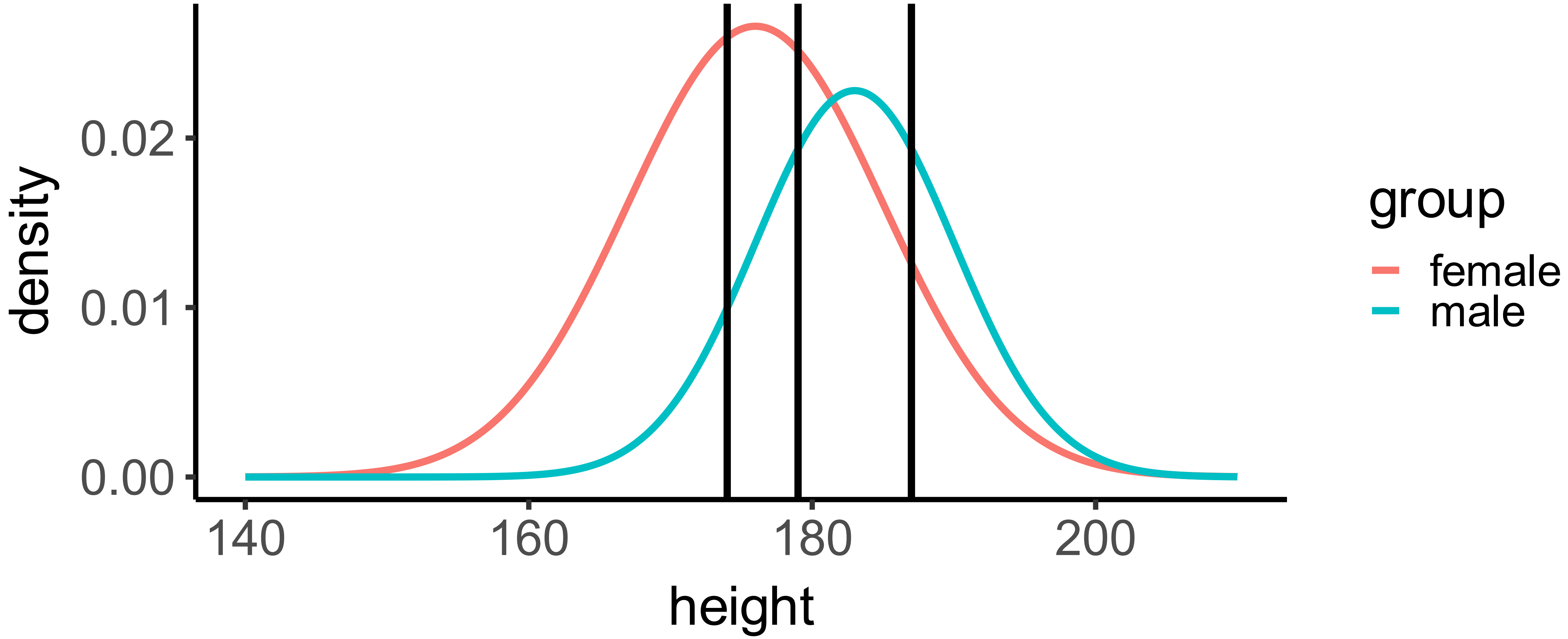
Estimating Prior and Likelihood

- Example: Classify male and female based on their heights (have training set)
- Use relative frequencies for estimating prior probabilies $P(Y = Female)$ and $P(Y = Male)$ (assume $P(Y = Female) = 60\%$)
- For likelihood, make assumption: Within groups heights are normally distributed
- Via standard estimation: $P(X|Y = Male) = N(183, \sigma^2 = 49)$, $P(X|Y = Female) = N(176{,}81)$

Likelihood (Prob. of height given group)



Posterior (Prob. of group given height)



Multiple Features: Naïve Bayes

- If we have multiple features, we need a model for the joint distribution
- The easiest / naïve approach is to assume that all features are independent
- Formally, naïve Bayes assumes: $P(X = x|Y = k) = \prod_{i=1}^{p} P(X_j = x_j|Y = k)$
- For each feature, a model is required
- Typically those models are parametric distribution families (Gaussian, uniform etc.)
- For those families, parameters are estimated per class (class-specific mean and variance, for example)

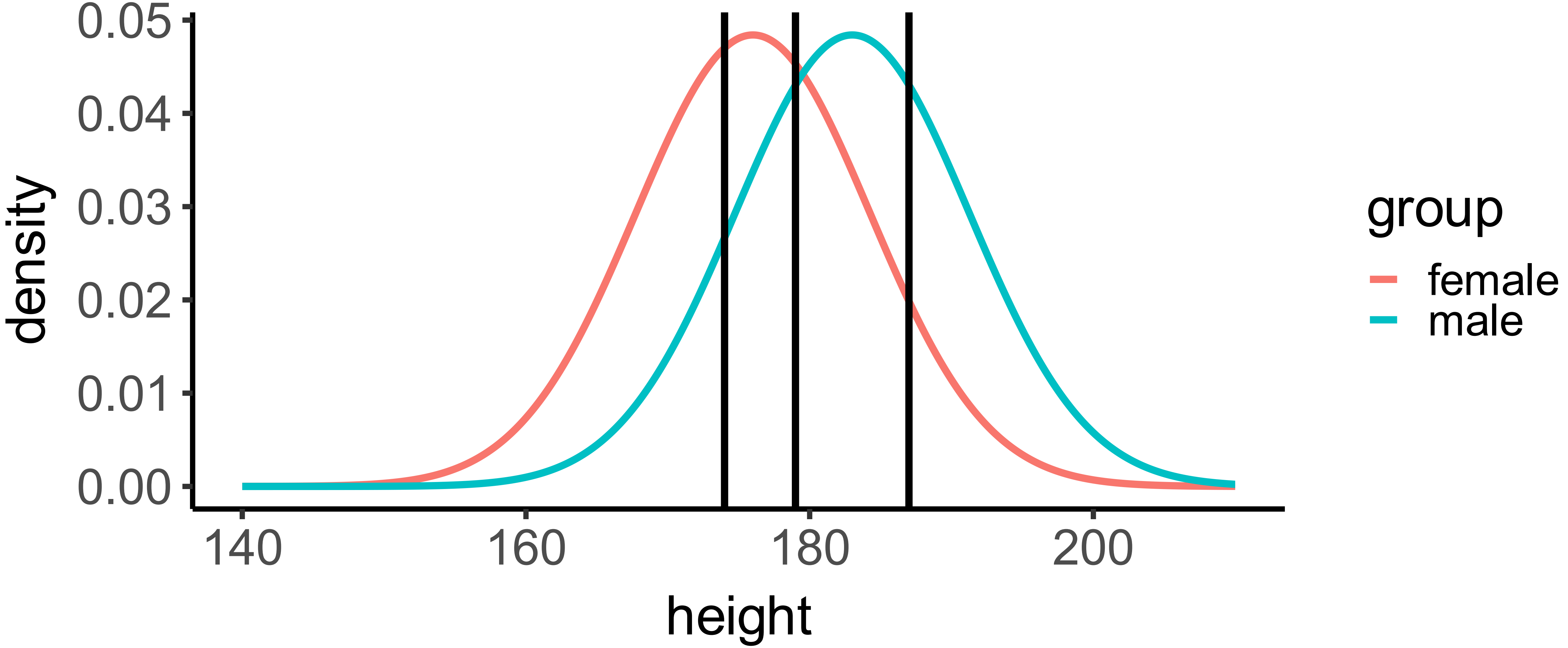
Linear Discriminant Analysis

Universiteit Leiden
The Netherlands

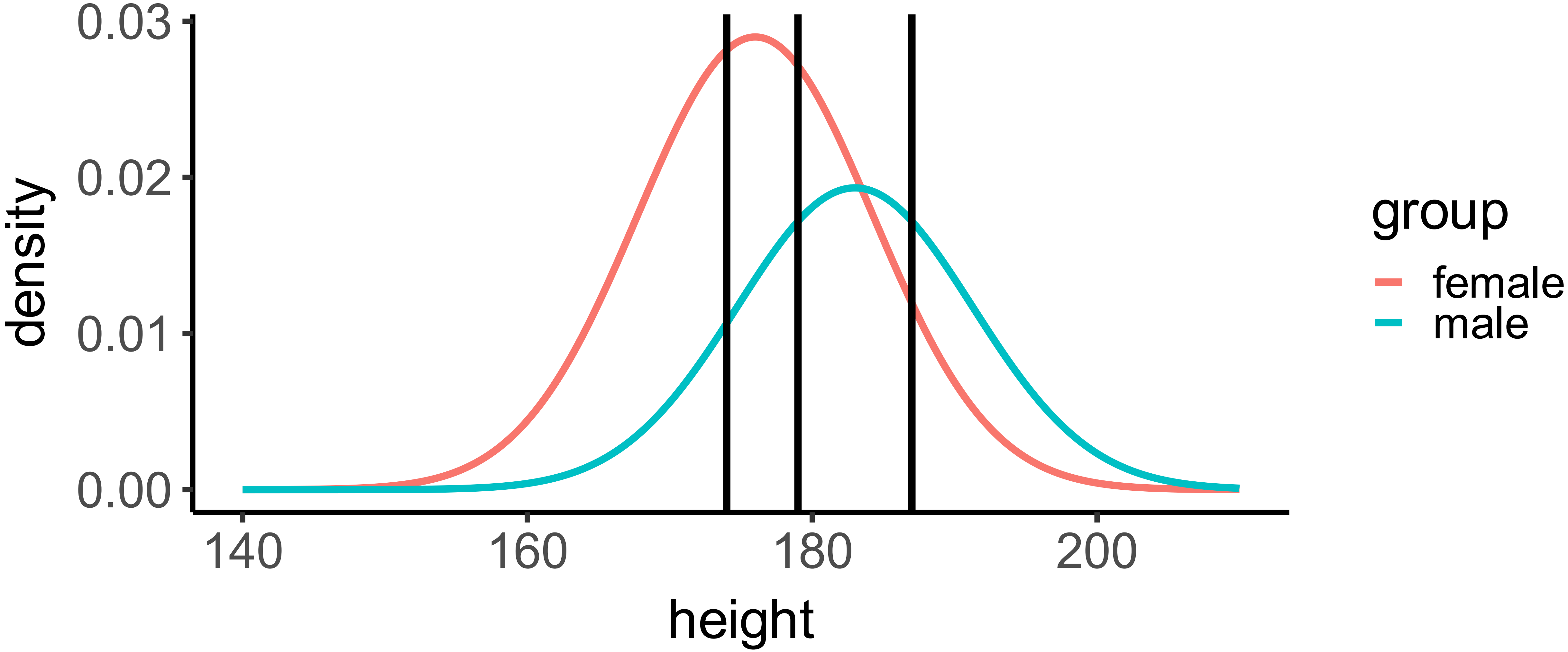
LDA p=1 $k=1$

- The same as the height example, but assumes equal variances across classes

LDA Likelihood



LDA Posterior

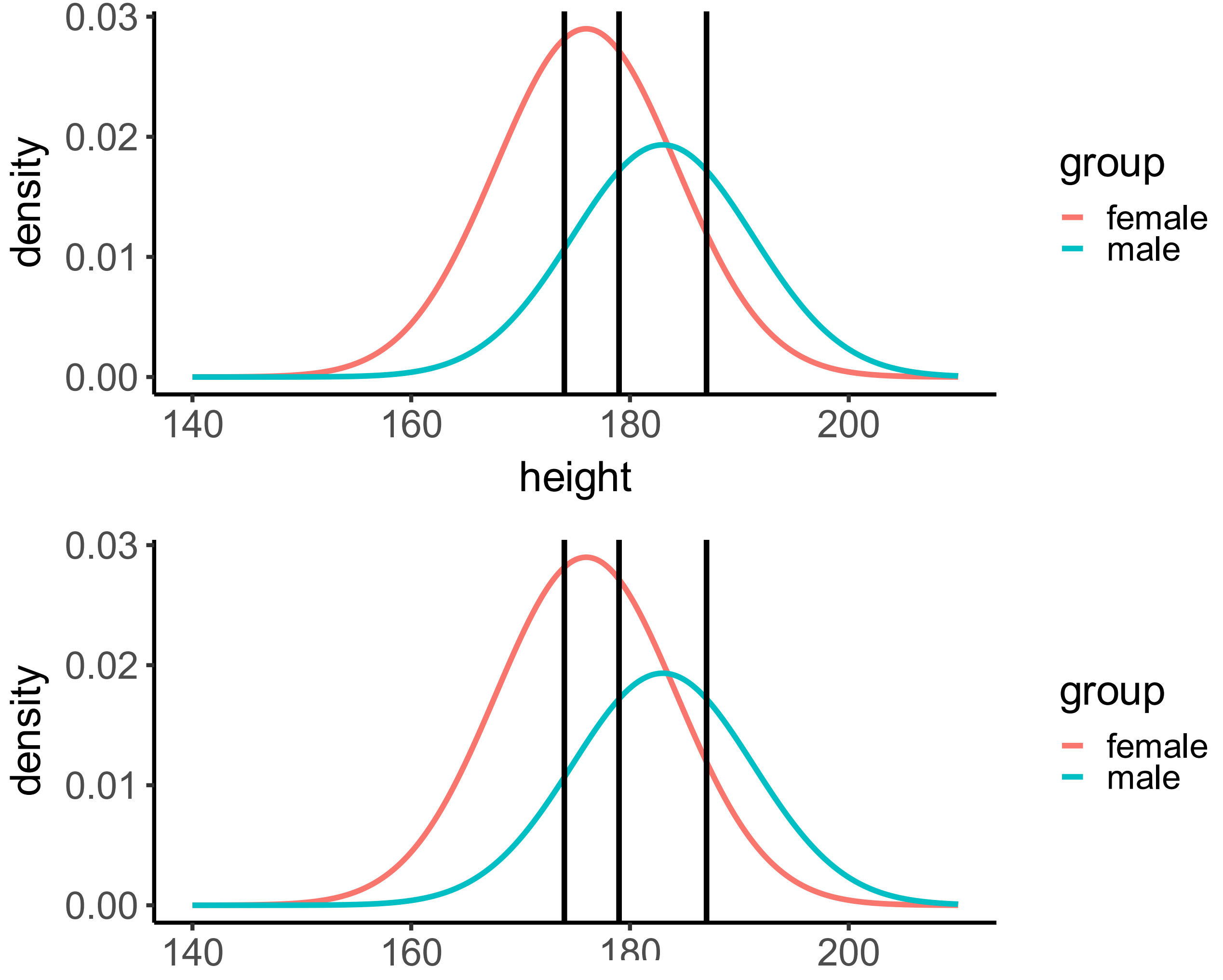


LDA p>1 $k>1$

- Idea: Fit a univariate LDA per dimension

weight

LDA p>1



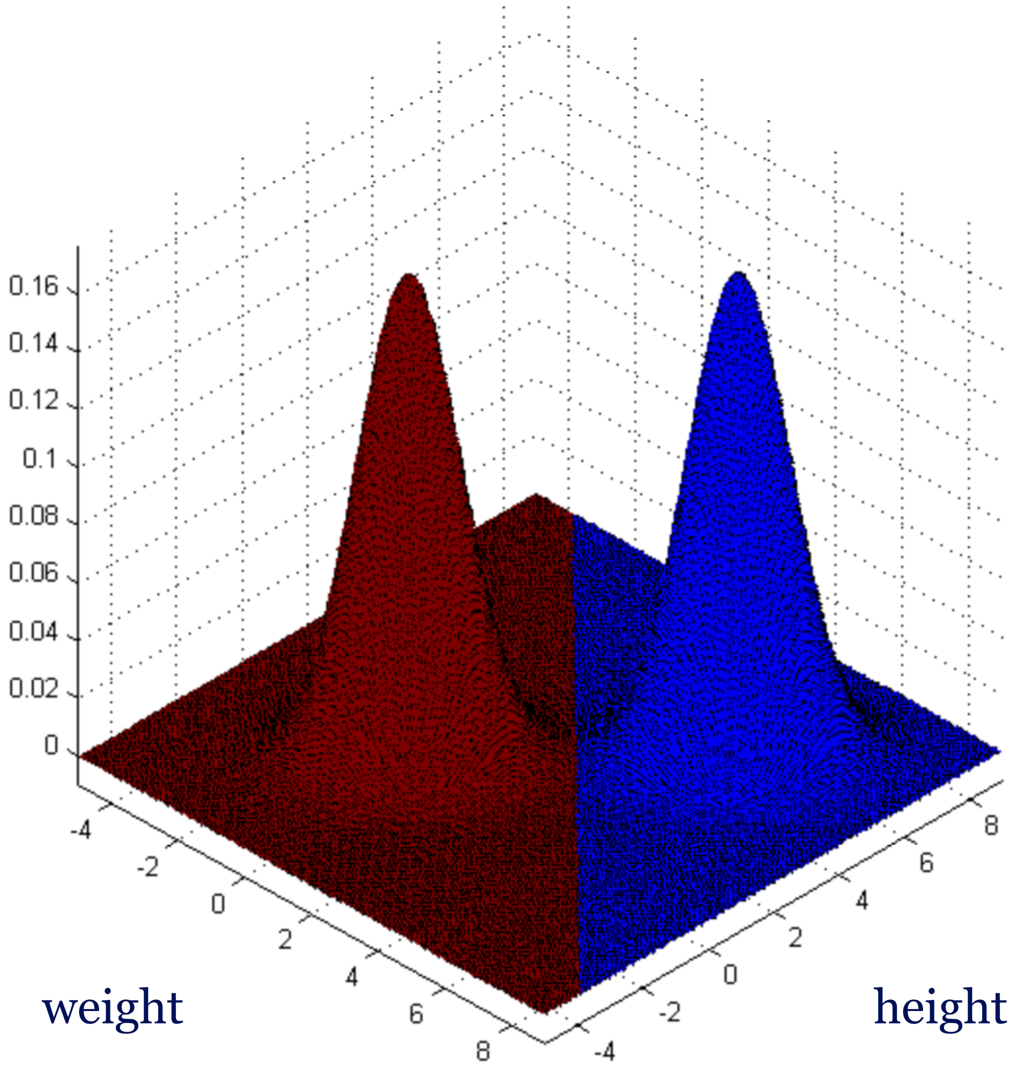
weight

LDA p>1

- Assumes independence of features: often heavily violated
- Fit **multivariate** Gaussian to each class
- **Assume equal covariances**

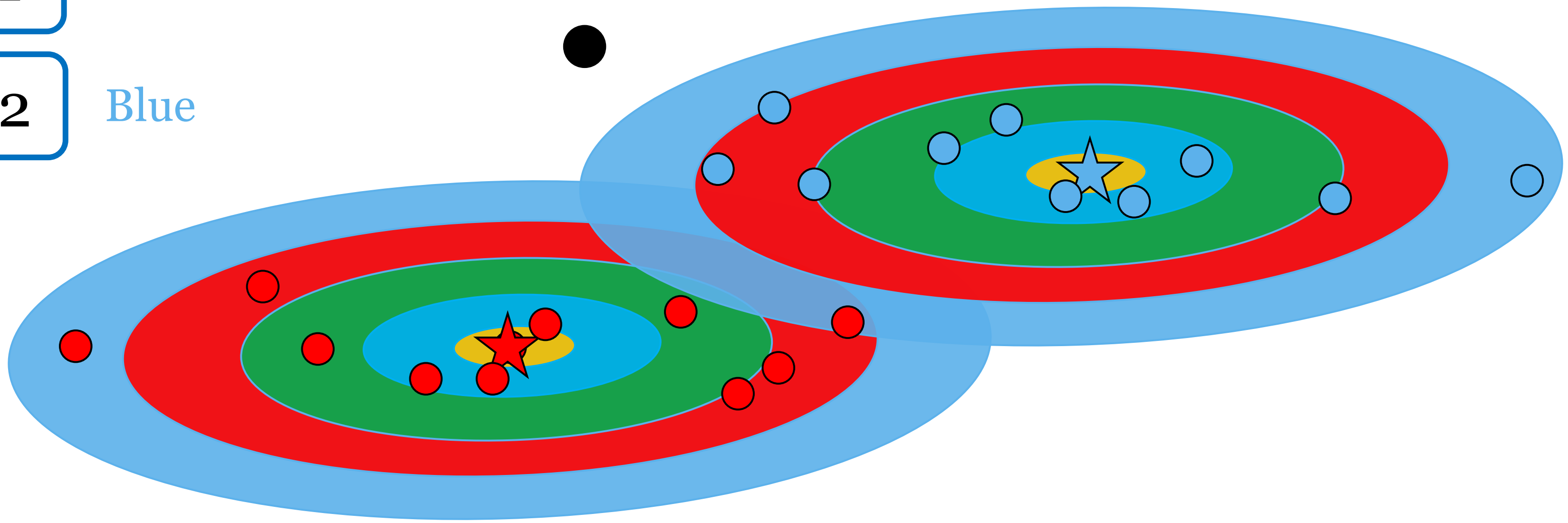
weight

LDA p>1


0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
-4
-2
0
2
4
6
8
8
6
4
2
0
-2
-4
weight
height

Linear Discriminant Analysis

Q	Which color does LDA classify this point to?
1	RED
2	Blue



Linear Classifier

- x is feature vector with $x = [x_1, x_2, \ldots, x_p]^\top$
- A classifier is called linear if its decision function is of the form

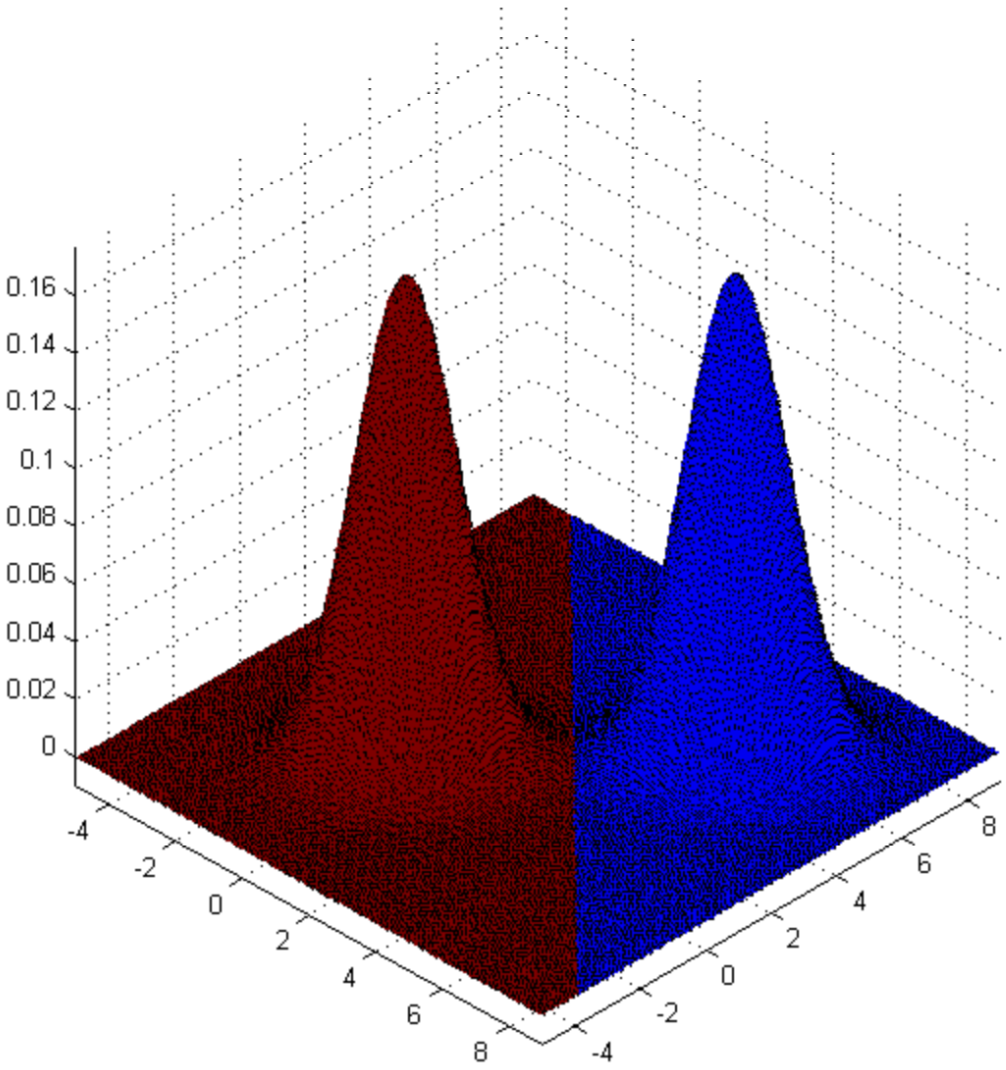
$$C(x) = \text{sign}\left(\beta_0 + \sum_{i=1}^{p} \beta_j x_j\right) = \begin{cases} 1, & \beta_0 + \sum \beta_i x_i \geq 0 \\ 0, & \beta_0 + \sum \beta_i x_i < 0 \end{cases}$$

- Defining $\tilde{x} = [1, x]^\top$ this can more compactly be expressed as

$$C(x) = \text{sign}\left(\beta^\top \tilde{x}\right)$$

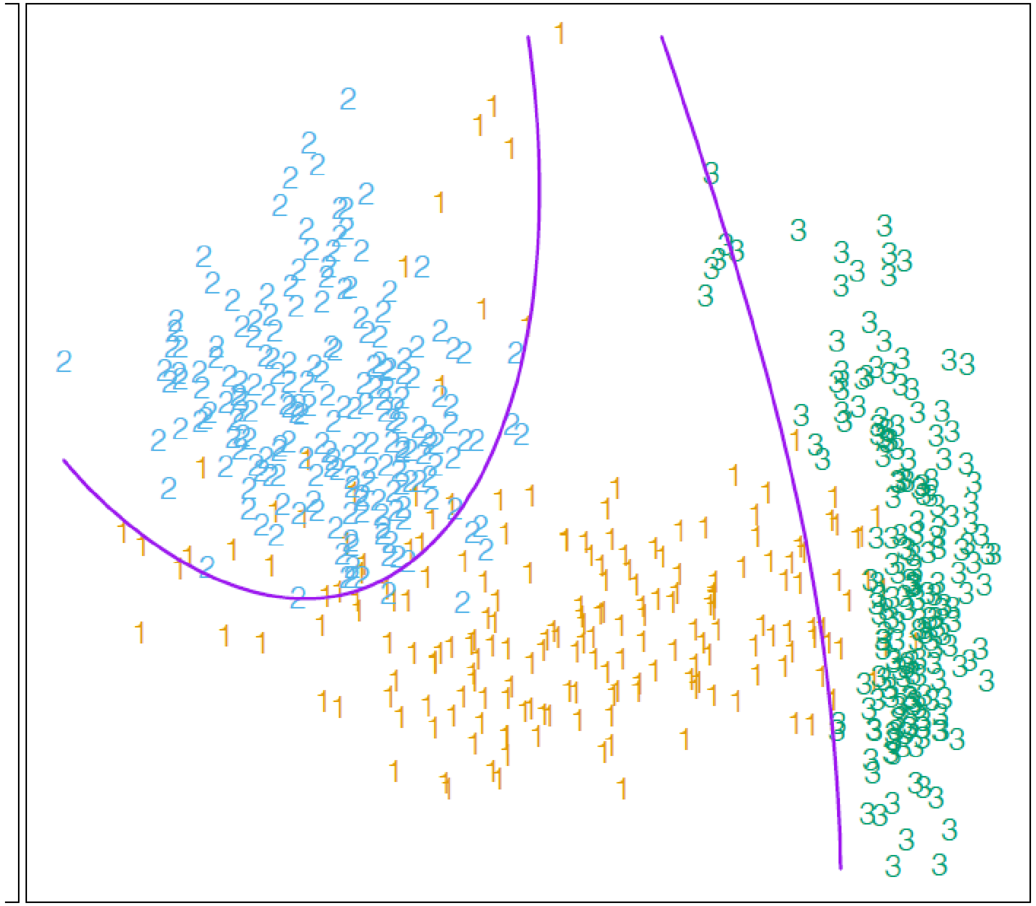
- It follows directly that the decision boundary is linear (linear / hyper-plane)

Intuition LDA Linear



Quadratic Discriminant Analysis

- Drop the assumption of equal covariance matrices
- Leads to quadratric decision boundaries
- Decreases bias at the cost of increased variance



To focus in book (for LDA)

- Formulas for LDA and QDA
- Discriminant function

Bias and Variance

- LDA generally low variance, high bias
- If within-class distribution normal and same covariance → no bias
- QDA higher variance but lower bias

Performance Metrics

Universiteit Leiden
The Netherlands

Problems with Accuracy

- Accuracy = $\frac{\#classifier\ correct}{\#total\ predictions}$
- Want to predict patient has Corona yes/no (positive / negative)
- Assume: at any given time 2% of the tested people are positive
- $C(x) =$ "$negative$" has accuracy 98%
- For unbalanced classes, accuracy tends to be misleading

Confusion Matrix

- TP: Number of samples that are *correctly* classified as positive
- TN: Number of samples that are *correctly* classified as negative
- FP: Number of samples that are *incorrectly* classified as positive
- FN: Number of samples that are *incorrectly* classified as negative

	Predicted Positive	**Predicted Negative**
Actually Positive	True Positive (TP)	False Negative (FN)
Actually Negative	False Positive (FP)	True Negative (TN)

Hours of Fun With 4 Numbers

precision rate $\frac{TP}{TP+FP}$

- Accuracy: $\frac{\#classifier\ correct}{\#total\ predictions} = \frac{TP+TN}{TP+FP+TN+FN}$

"The probability of correctly classifying a random person"

- Sensitivity: accuracy within the positive class: $\frac{TP}{TP+FN}$ recall rate positive 查全率

"The probability of the test to be positive if the patient has Corona"

- Specificity: accuracy within the negative class: $\frac{TN}{TN+FP}$ recall rate negative

- "The probability of the test to be negative if the patient is healthy"

Hours of Fun With 4 Numbers

- Positive Predictive Value: accuracy within the positive predictions

$$\frac{TP}{TP+FP}$$ precision rate 查准率

"Probability of having Corona after having been tested positive"

- Negative Predictive Value: accuracy within the negative predictions:

$$\frac{TN}{TN+FN}$$ precision rate negative

"Probability of being healthy after having been tested negative"

Hours of Fun With 4 Numbers Continued

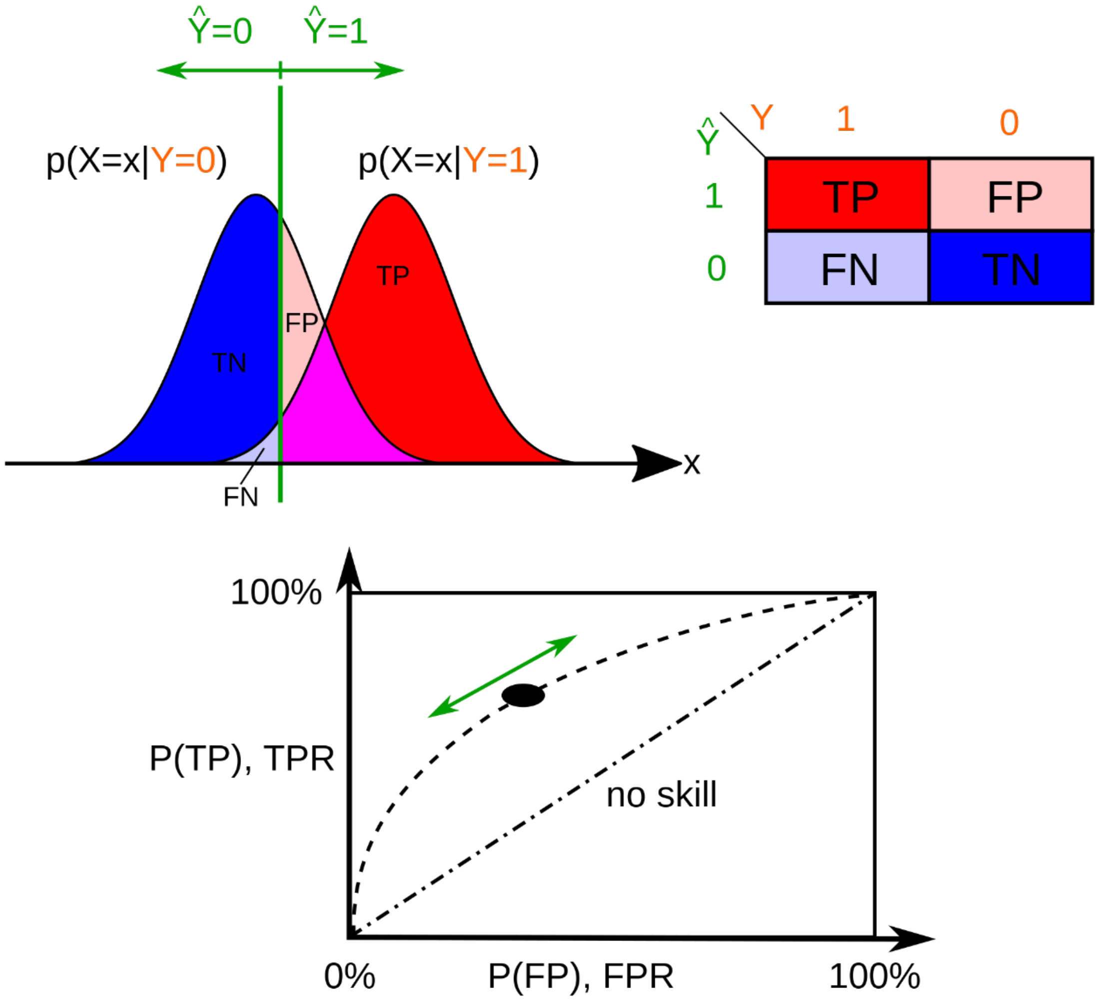
- True positive rate (TPR) = sensitivity = Recall, True negative rate (TNR) = specificity
- Balanced accuracy = $\frac{TPR+TNR}{2} = \frac{sensitivity + specificity}{2}$ = average of within class accuracies
- Positive Predictive Value = Precision
- F-Measure = harmonic mean of precision and recall $\frac{2}{\frac{1}{precision}+\frac{1}{recall}}$
- Harmonic mean is closer to the smaller value
  - $H(0.5,0.5) = 0.50$
  - $H(0.5,0.6) = 0.54$
  - $H(0.5,1) = 0.67$

TP vs FP Tradeoff

- True positive rate (TPR)= accuracy within the positive class: $\frac{TP}{TP+FN}$
  - "The probability of the test to be positive if the patient has Corona"
- True negative rate (TNR)=accuracy within the negative class: $\frac{TN}{TN+FP}$
  - "The probability of the test to be positive if the patient has Corona"
- False positive rate (FPR)=inaccuracy within the negative class:

$$1 - TNR$$

ROC Curve



Ŷ=0
Ŷ=1
p(X=x|Y=0)
p(X=x|Y=1)
TN
FP
TP
FN
x
Y
Ŷ
1
0
1
0
TP
FP
FN
TN
100%
P(TP), TPR
no skill
0%
P(FP), FPR
100%

ROC Curve

- Shows TPR and FPR for every threshold
- Optimally $TPR = 1$ and $FPR = 0$
  - Corresponds to horizontal line at 1
- Random guessing → diagonal
- Above the diagonal → better than random guessing
- Below the diagonal → worse than random guess
- AUC

When to Use What

- Have to distinguish: Evaluation vs. Optimization
- For evaluation, I recommend using many different methods, as quantify many different, important aspects of performance
- For selection: Select based on what is most important

Performance Measures for Selection

- If every misclassification is equally “bad” and the distribution of classes is roughly equal or the same between training and test → Accuracy
  - Example: Is a car red or orange?
- Same conditions but unequal class sizes and likely difference between training and test → Balanced Accuracy
- If misclassifying the positive class is much worse than the negative class → TPR 查全率 $\frac{TP}{TP+FN}$
  - Example: Corona test

Performance Measures for Selection

- If misclassifying the negative class is worse than the negative class → FPR $\frac{FP}{FP+TN}$
  - Example: Spam detection
- If misclassifying the positive class is worse, you also want to consider the negative, and unbalanced class → F1-score
  - AUC: When you do not care about a particular threshold but the performance at many thresholds
  - Example: Evaluation of probabilistic classifier

Question Performance Metrics

Q Consider a classifier that always predicts the positive class. What is it's TPR and what it's FPR rate?

100%

50

1

预

		正 100	反 0
真	正 50	50 TP	0 FN
	反 50	50 FP	0 TN

Hours of Fun with 4 Numbers (the end)

- Many more measures
  - https://en.wikipedia.org/wiki/Evaluation_of_binary_classifiers