Linear Classification 1 - LDA

Julian Karch
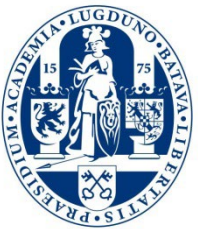
Overview

- Intro Generative Classification + Naïve Bayes

- Linear Discriminant Analysis

- Quadratic Discriminant Analysis and Regularized LDA
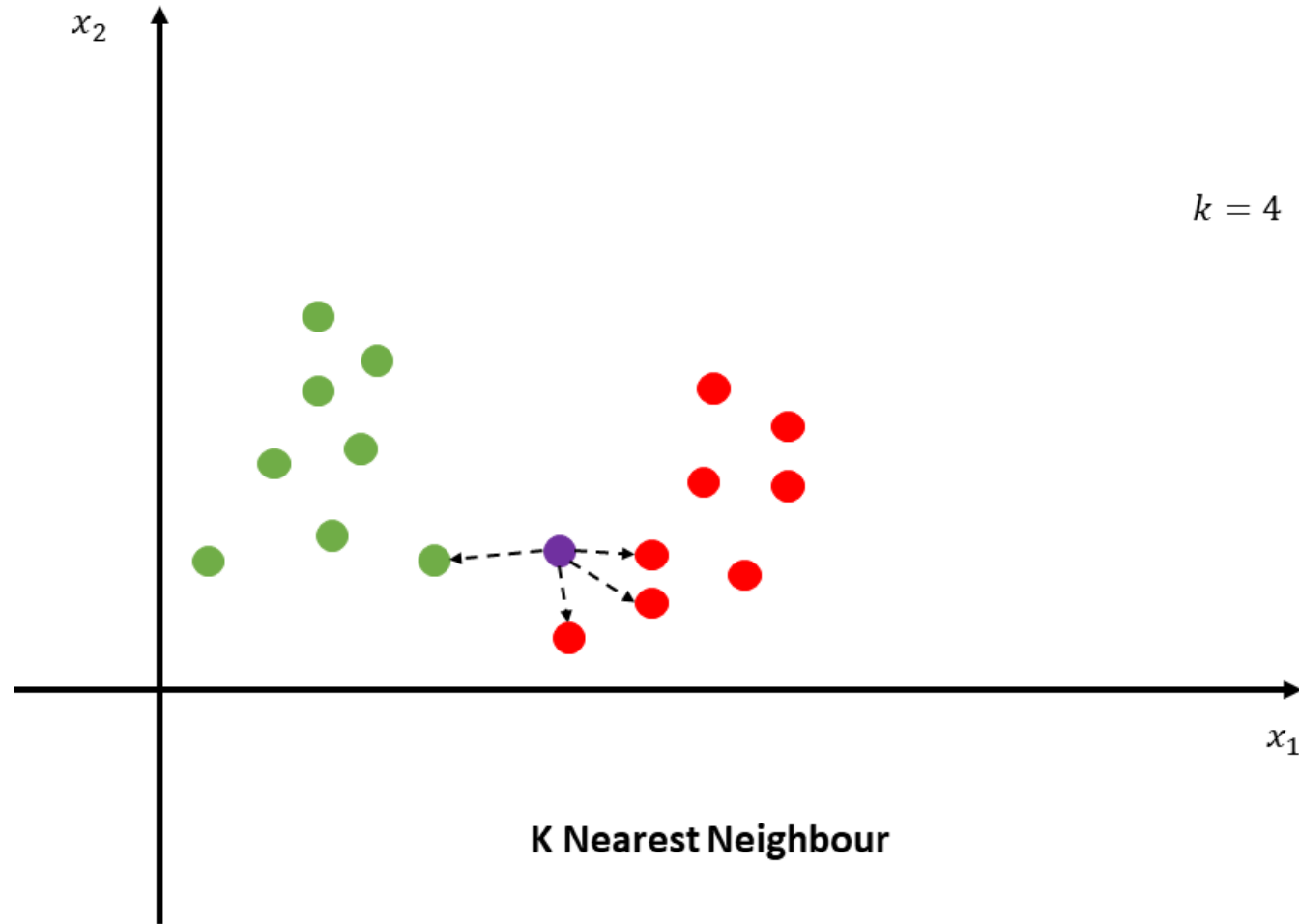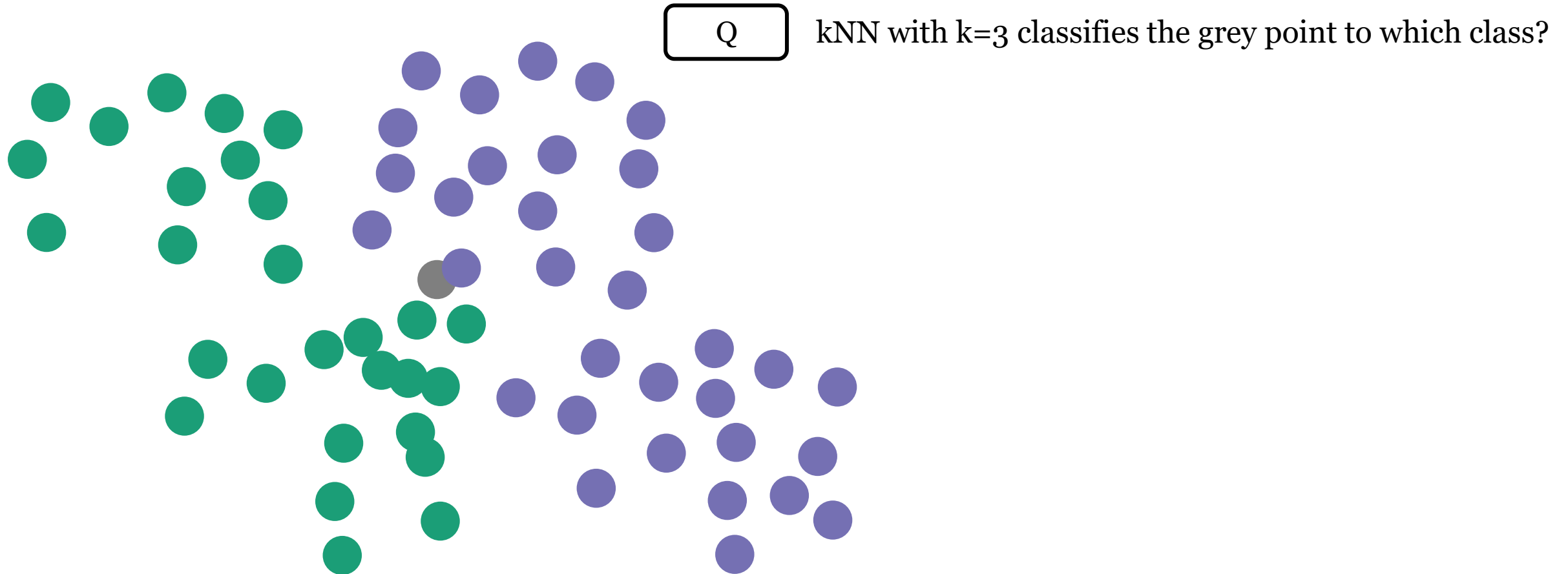
Intro + Naïve Bayes

Classification

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

K Nearest Neighbors

$k = 4$

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

- L is a **loss function** that quantifies how "bad" wrong predictions are

- We get MSE_{pred} by using $L_{sq}(\hat{Y}, Y) = (\hat{Y} - Y)^2$

- Another option: $L_{abs}(\hat{Y}, Y) = |\hat{Y} - Y|$

Expected Prediction Error Classification

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

- Most popular loss is 0-1 loss

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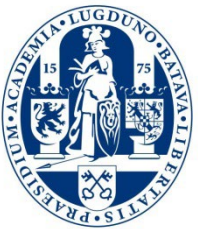
Posterior Probability

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

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

Q

- **What is $P(X = \text{positive}, Y = \text{Corona})$?**
- **What is $P(X = \text{positive}, Y = \text{Healthy})$?**
- **What would you classify (assume 0-1 loss)?**

Bayes Rule

- Want: $P(Y = \text{Corona} \mid X = \text{positive})$

- $P(Y = \text{Corona} \mid X = \text{positive}) = \dfrac{P(X=positive, Y=Corona)}{P(X=positive)}$

- $P(X = positive)$ does not dependent on Y

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

Bayes Rule

- $P(X = \text{positive}) = P(X = \text{positive}, Y = \text{Corona}) +$
$$P(X = \text{positive}, Y = \text{Healthy})$$

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

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)
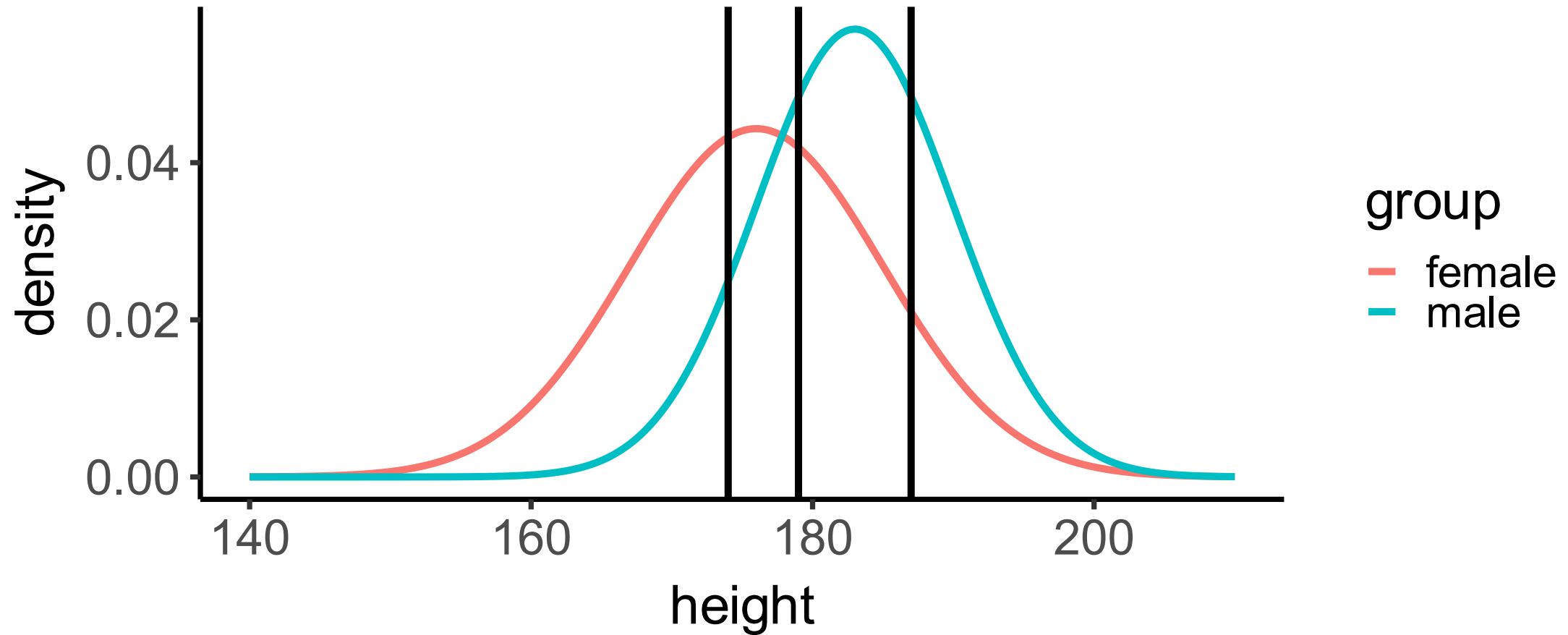
Posterior (Prob. of group given height)
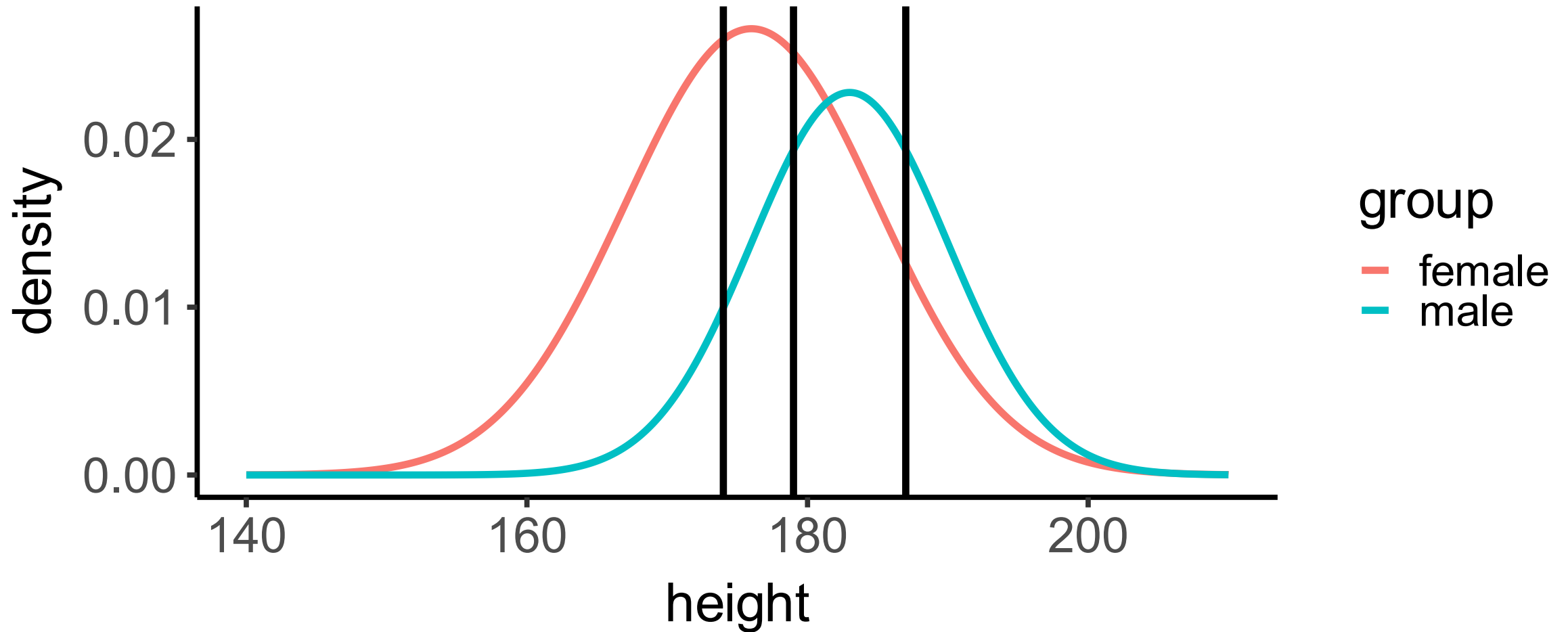
Multiple Features: Naïve Bayes

- 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)
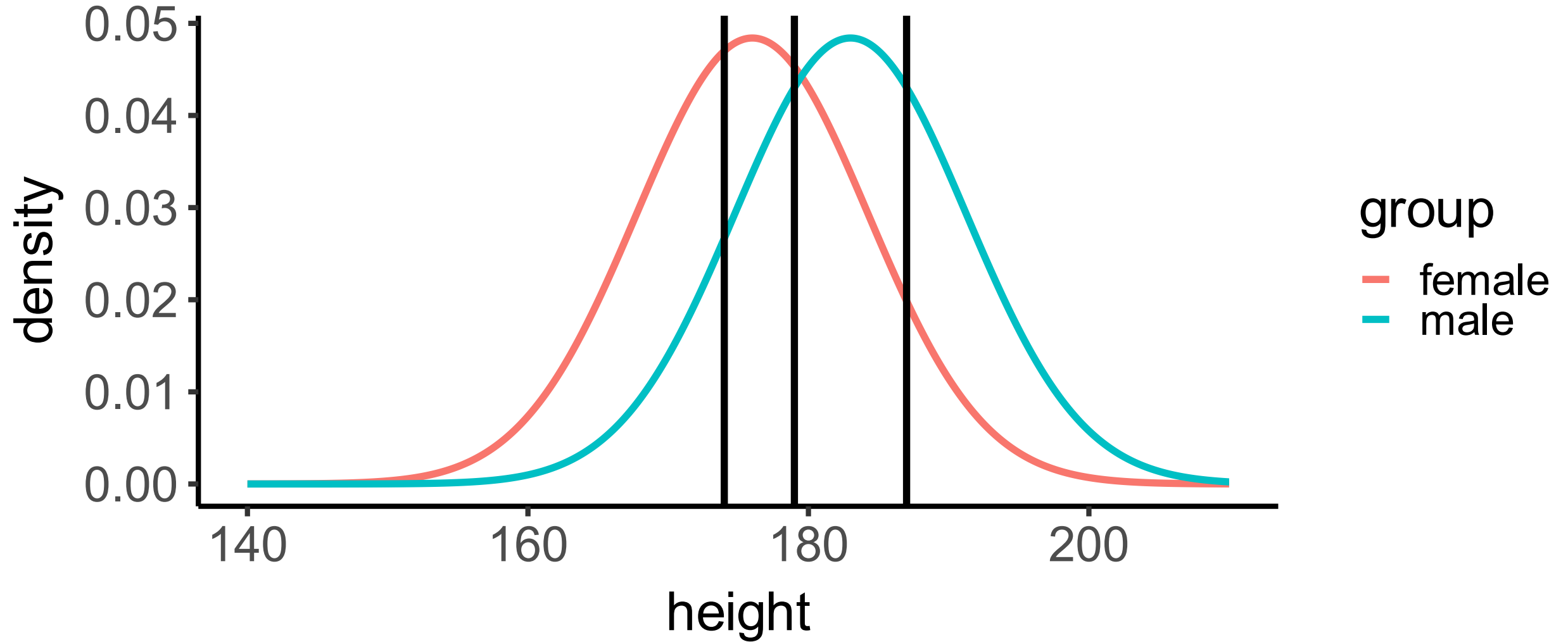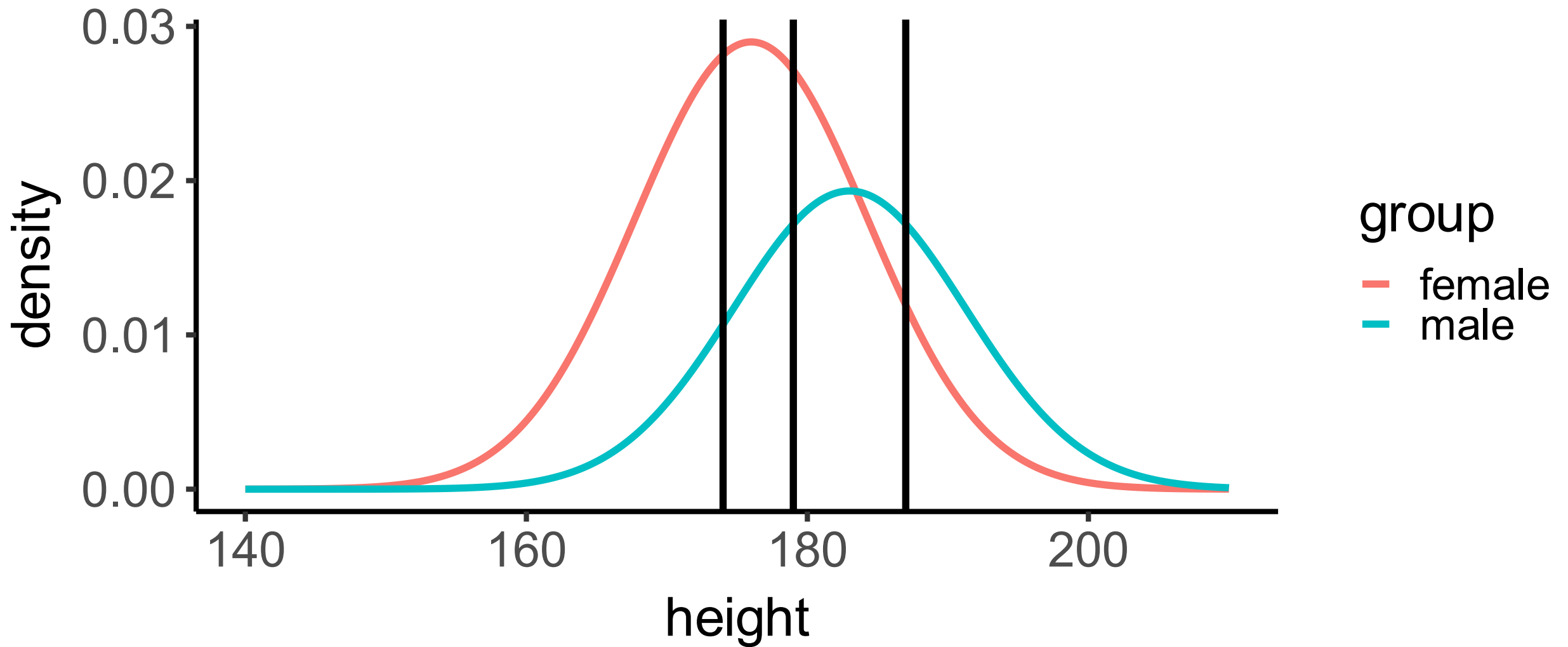
Linear Discriminant Analysis

LDA p=1

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

LDA Likelihood

LDA Posterior

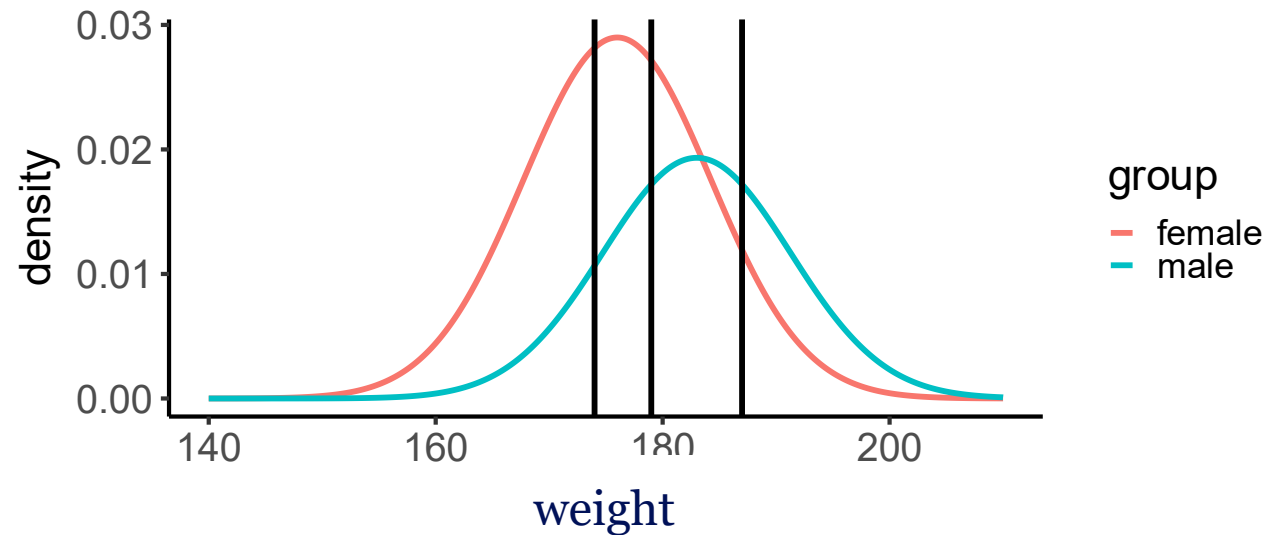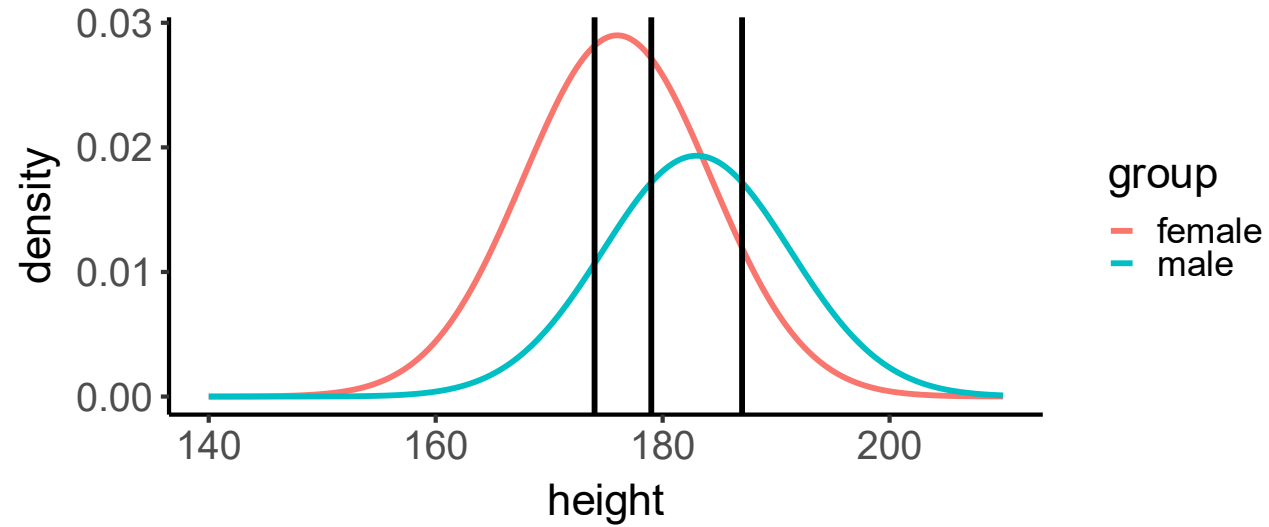# LDA p>1

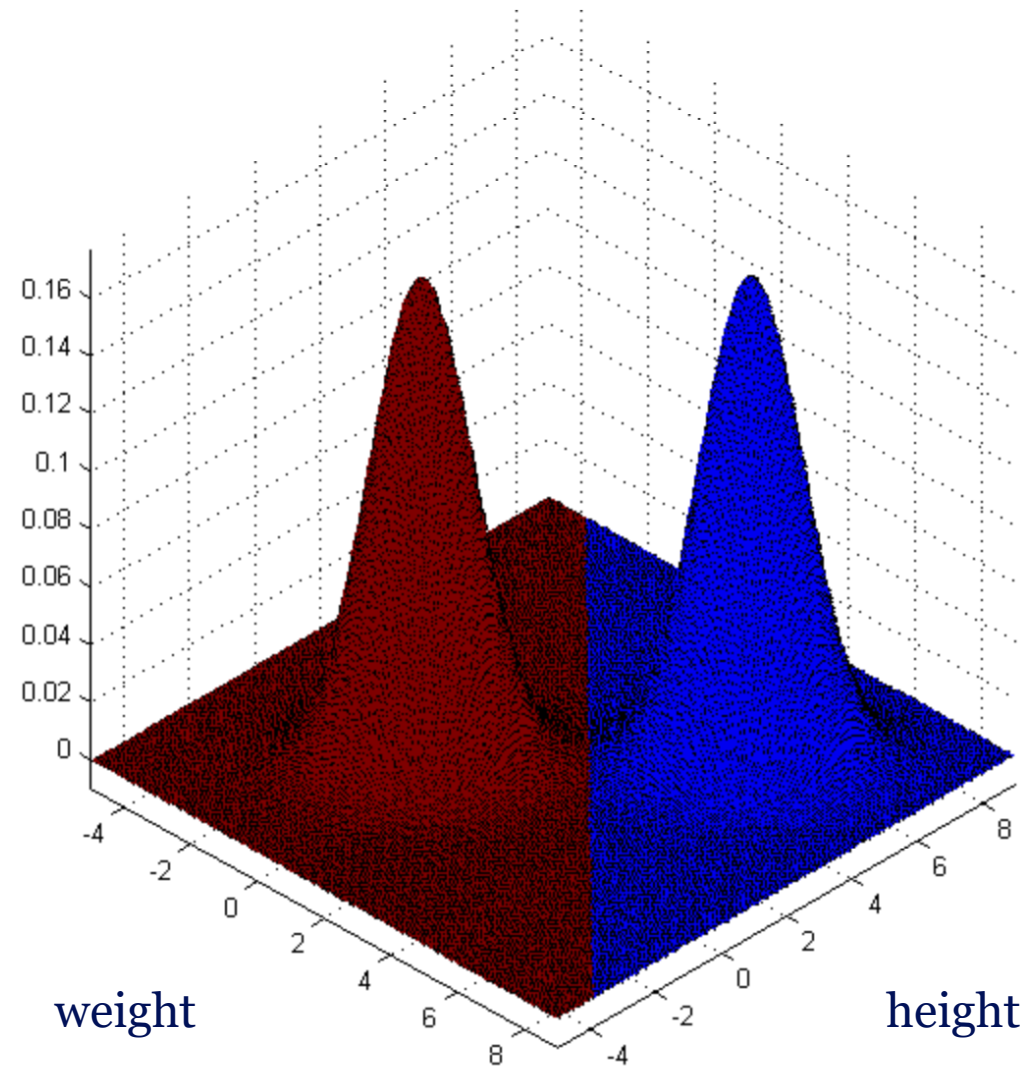- Idea: Fit a univariate LDA per dimension

weight

LDA p>1

LDA p>1

- Assumes independence of features: often heavily violated

- Fit **multivariate** Gaussian to each class

- **Assume equal covariances**

weight

LDA p>1

weight · height

Linear Discriminant Analysis

| Q | Which color does LDA classify this point to? |

| 1 | RED |

| 2 | Blue |
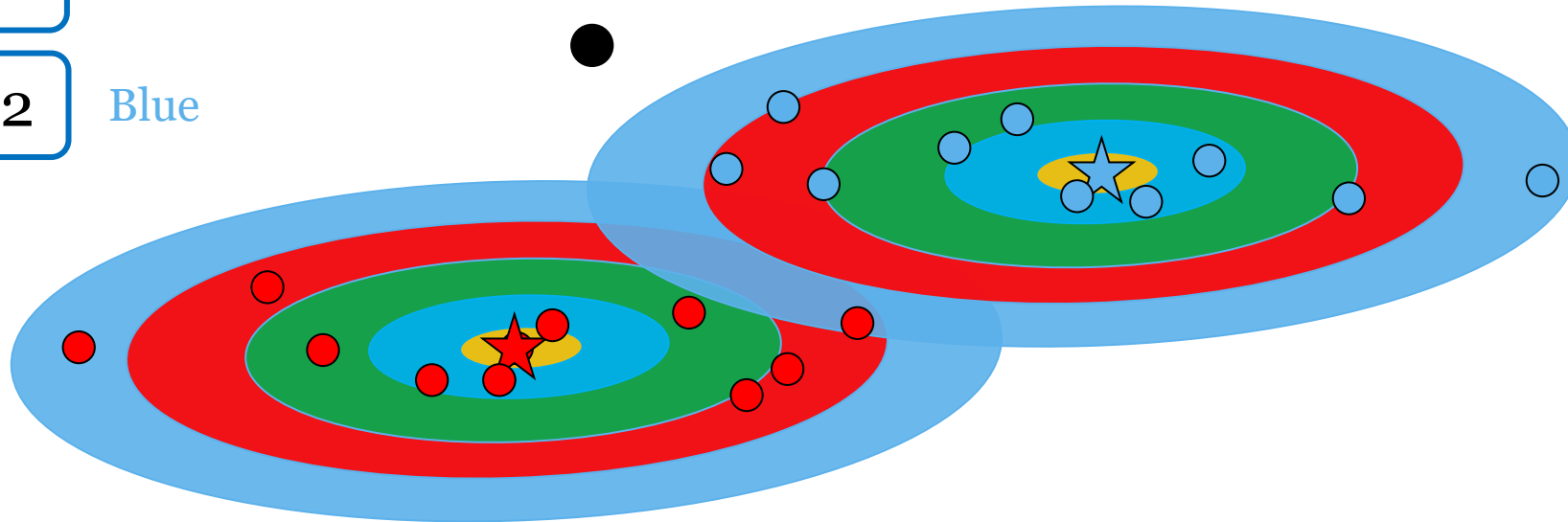
Linear Classifier

- 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

$$C(x) = \text{sign}\left(\beta_0 + \sum_{i=1}^{p} \beta_j x_j\right)$$

- 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
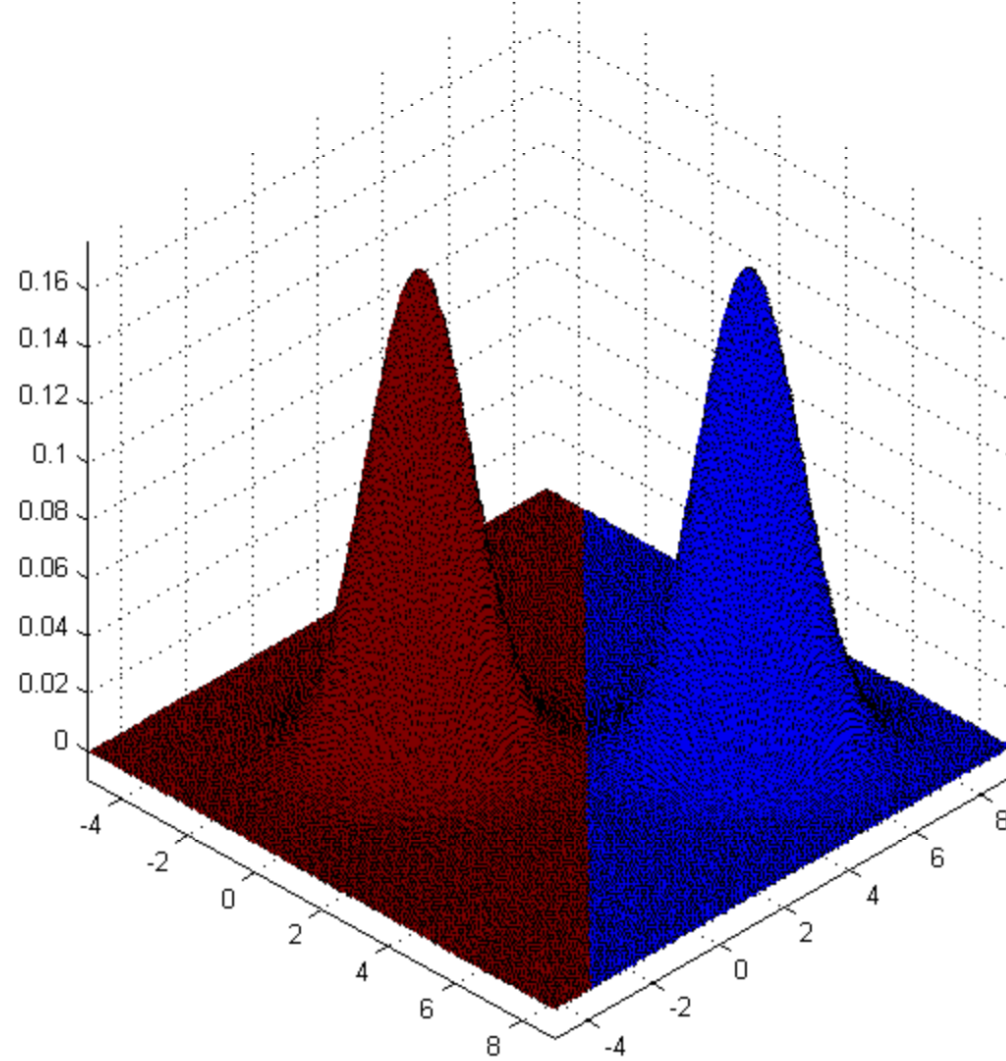
Quadratic Discriminant Analysis

- Drop the assumption of equal covariance matrices

- Leads to quadratric decision boundaries

- Decreases bias at the cost of increased variance
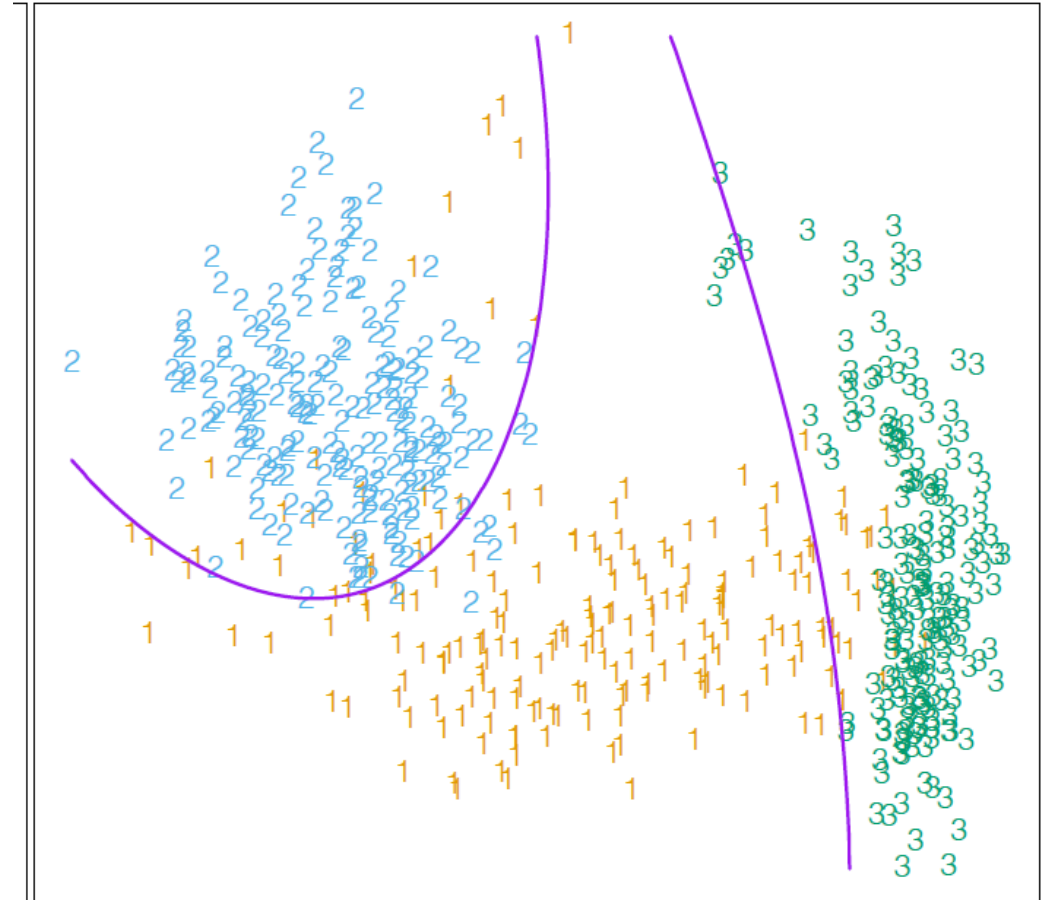
To focus in book (for LDA)

- 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

Problems with Accuracy

- Accuracy $= \dfrac{\#classifier\ correct}{\#total\ predictions}$

- 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

- 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: $\frac{TP}{TP+FN}$

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

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

- "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

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

"Probability of having Corona after having been tested positive"

- Negative Predictive Value: accuracy within the negative predictions:

$$\frac{TN}{TN+FN}$$

"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 $\dfrac{2}{\frac{1}{precision}+\frac{1}{recall}}$

- Harmonic mean is closer to the smaller value
  - $H(0.5, 0.5) = 0.5o$
  - $H(0.5, 0.6) = 0.54$
  - $H(0.5, 1) = 0.67$

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
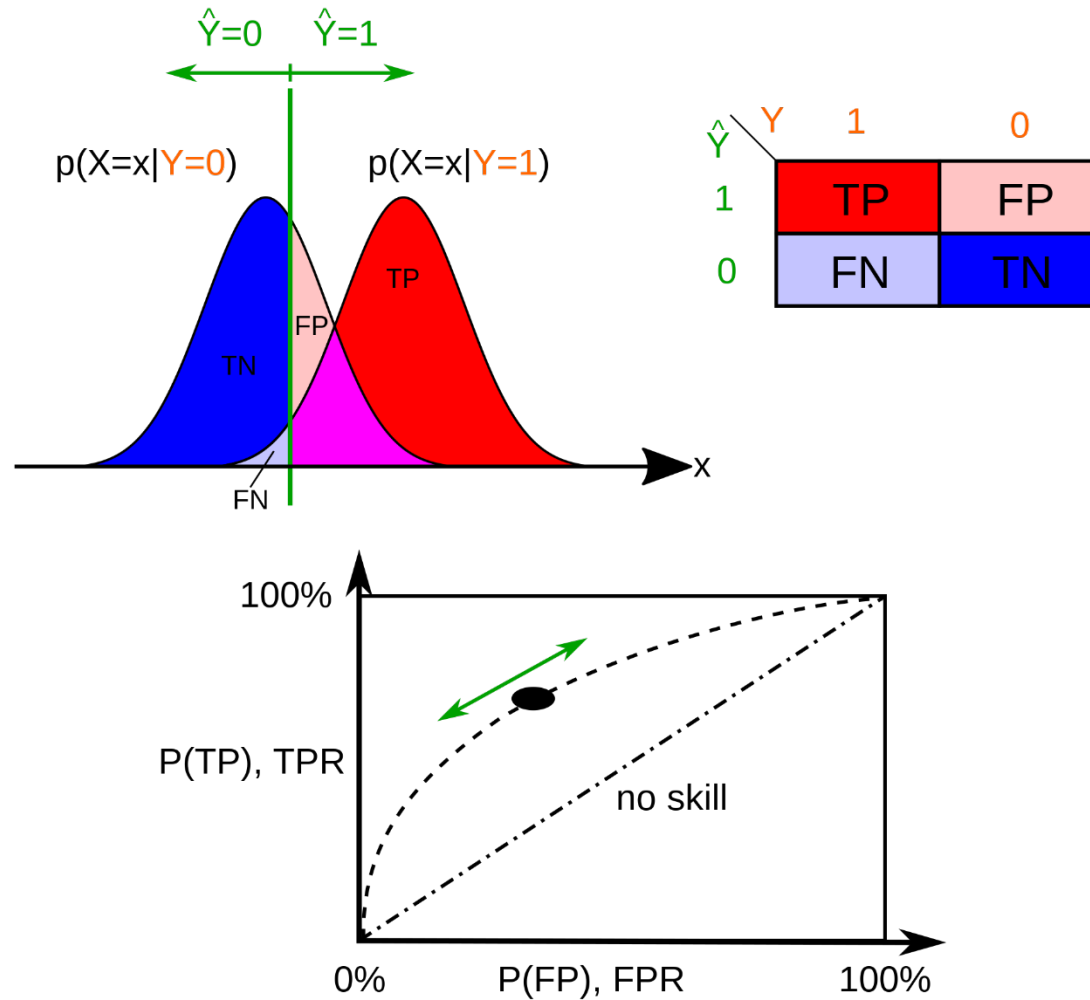
ROC Curve

- Shows TPR and FPR for every threshold
- Optimally $TPR = 1$ and $FPR = 0$

  -Corresponds to horizontal line at 1
- Random guessing $\rightarrow$ diagonal
- Above the diagonal $\rightarrow$ better than random guessing
- Below the diagonal $\rightarrow$ 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

  - Example: Corona test

Performance Measures for Selection

- If misclassifying the negative class is worse than the negative class → FPR

  - 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?

Hours of Fun with 4 Numbers (the end)

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