CS 260: Foundations of Data Science

Prof. Sara Mathieson

Fall 2021

HAVERFORD
COLLEGE

Admin

Video on if possible!

- **Note-taker**: Ryan
- **Lab 8** due Thursday Nov 11
- **Exam** goes out on Thursday
  - Pickup outside my office
- **Office hours TODAY: 3:30-5pm** (same zoom link)
  - Join the next available breakout room
  - You should be able to see where I am
- All **project proposals** have been reviewed
  - *Please send a title ASAP if you didn't before!*

Outline for November 9

- Recap Bootstrapping
- Bagging and Random forests
- Midterm 2 Review
  - Revisit confusion matrices
  - PCA (linear transformation + interpretation
  - Naïve Bayes
  - Central Limit Theorem
  - Entropy vs. classification error
  - Logistic regression and cross entropy

Probably Thurs

Outline for November 9

- Recap Bootstrapping
- Bagging and Random forests
- Midterm 2 Review
  - Revisit confusion matrices
  - PCA (linear transformation + interpretation
  - Naïve Bayes
  - Central Limit Theorem
  - Entropy vs. classification error
  - Logistic regression and cross entropy

The bootstrap: Resampling

Data, $\boldsymbol{X}$ = [2, 3, 4, 8, 0, 6, 1, 10, 2, 4]

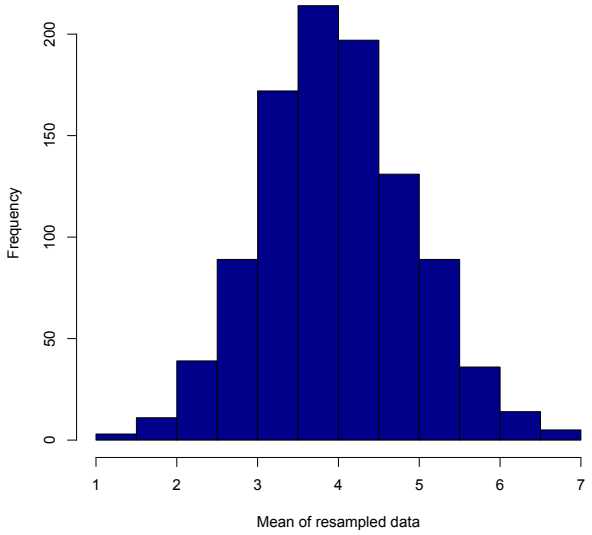
Resample, with replacement, T times

Compute Mean

1 8 2 4 6 10 1 1 1 8	4.2
1 0 1 6 4 1 4 2 1 2	2.2
8 1 6 2 6 4 2 4 10 2	4.5
8 3 4 2 10 8 10 8 8 1	6.2
6 4 6 4 6 4 2 4 3 4 0	4.3
…	…
…	…

Use the means from the resampled data to estimate the distribution!

95% of the means are between 2.3 and 5.9 (T=1000)



Slide: Iain Mathieson

The bootstrap: Resampling

"Estimate the range (Max—Min)"

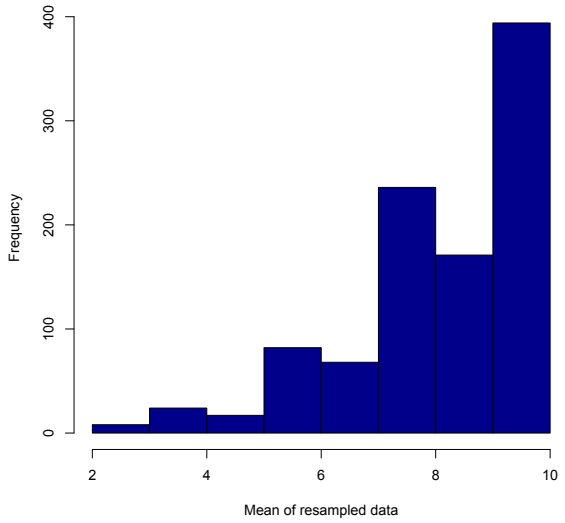
Data, $\boldsymbol{X}$ = [2, 3, 4, 8, 0, 6, 1, 10, 2, 4]

Compute Range

Resample, with replacement, T times

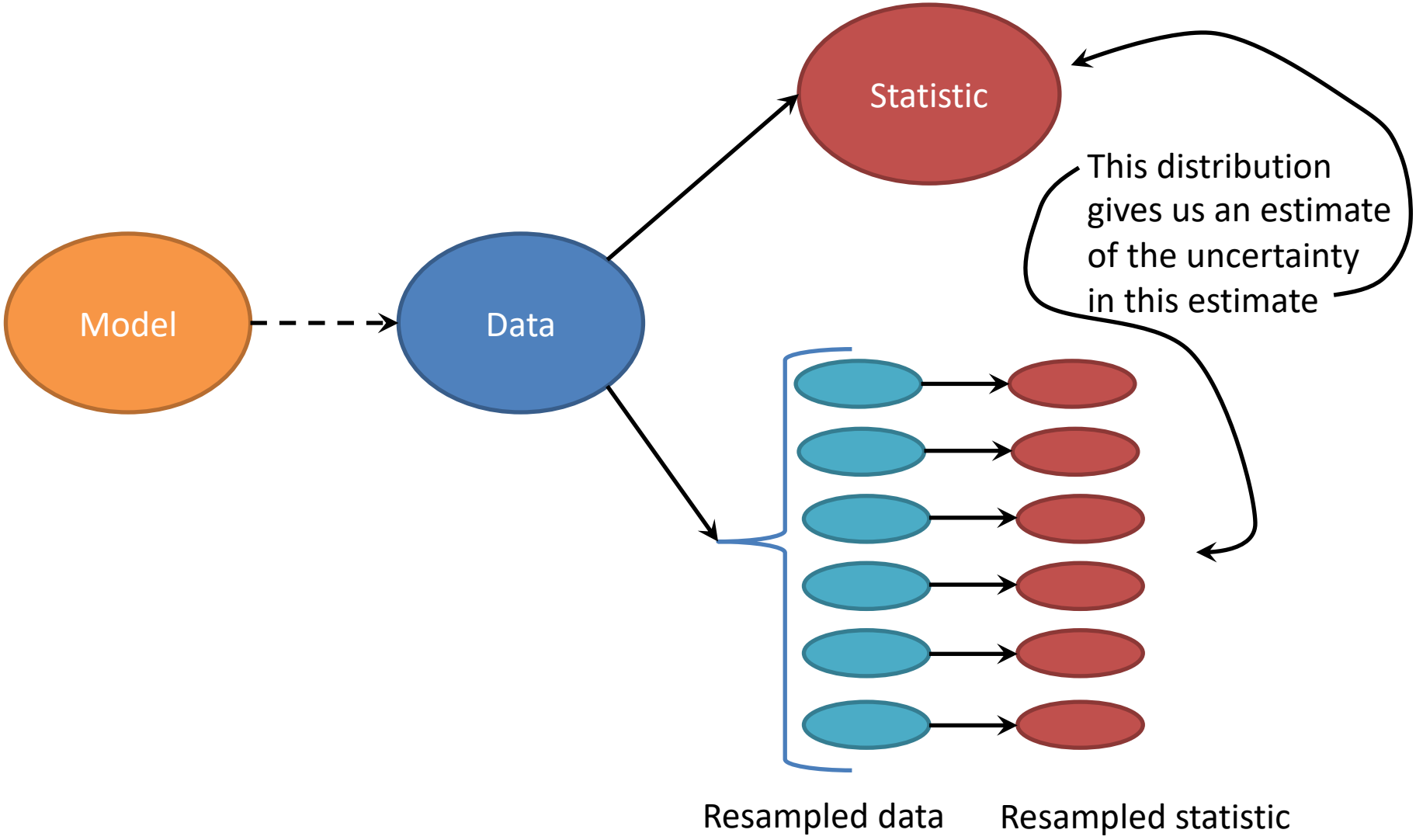
1 8 2 4 6 10 1 1 1 8	9
1 0 1 6 4 1 4 2 1 2	6
8 1 6 2 6 4 2 4 10 2	9
8 3 4 2 10 8 10 8 8 1	8
6 4 6 4 6 4 2 4 3 4 0	6
…	…
…	…

Use the ranges from the resampled data to estimate the distribution!



Slide: Iain Mathieson

The bootstrap: Resampling



Slide: Iain Mathieson

Bootstrap example

Setup: you obtain 0.87 accuracy on a test dataset using a new algorithm

Goal: find a 95% confidence interval for your estimate

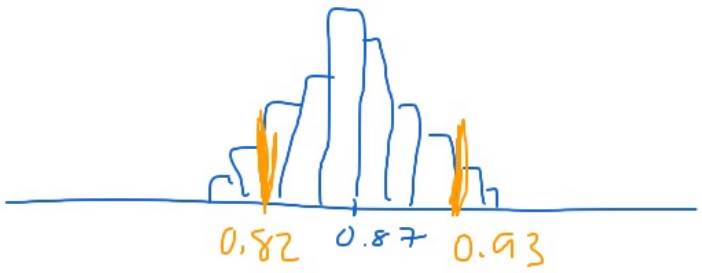
(1) bootstrap T times, run our method on each dataset => record result

[0.82, 0.91, 0.86, 0.95 ----] ← T trials

(2) sort results

(3) take the middle 95%

CI = (0.82, 0.93)



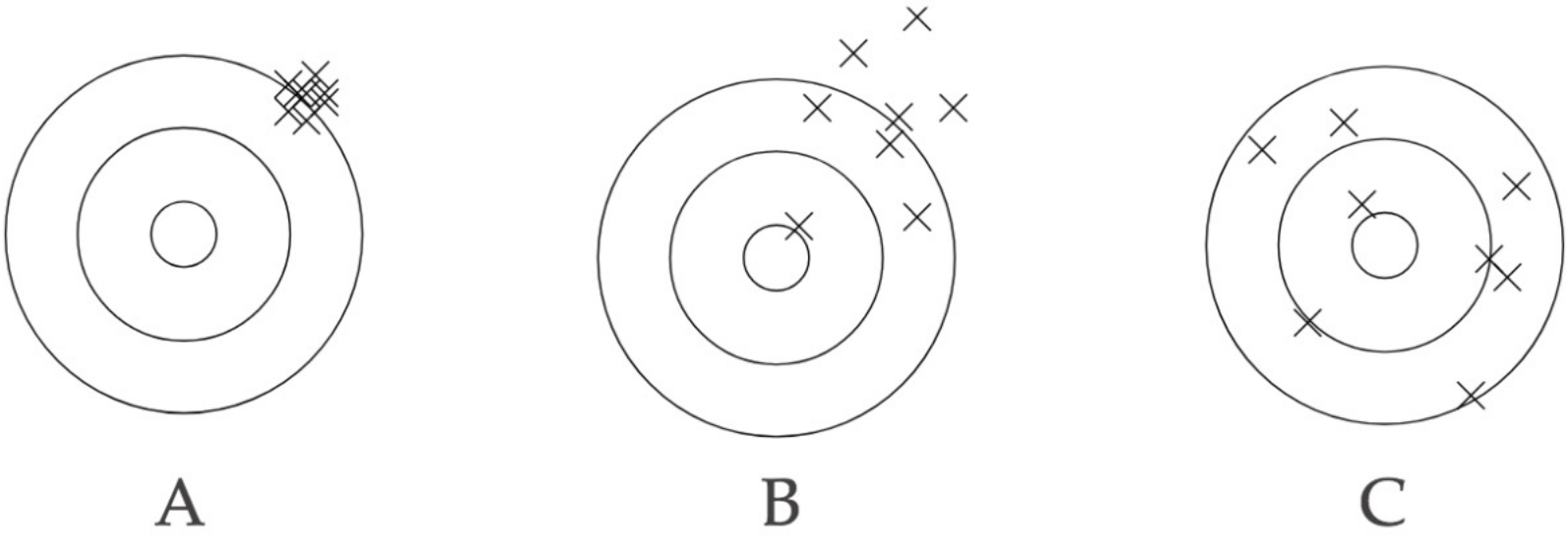
T = 1000

take middle 950

Outline for November 9

- Recap Bootstrapping
- Bagging and Random forests
- Midterm 2 Review
  - Revisit confusion matrices
  - PCA (linear transformation + interpretation
  - Naïve Bayes
  - Central Limit Theorem
  - Entropy vs. classification error
  - Logistic regression and cross entropy

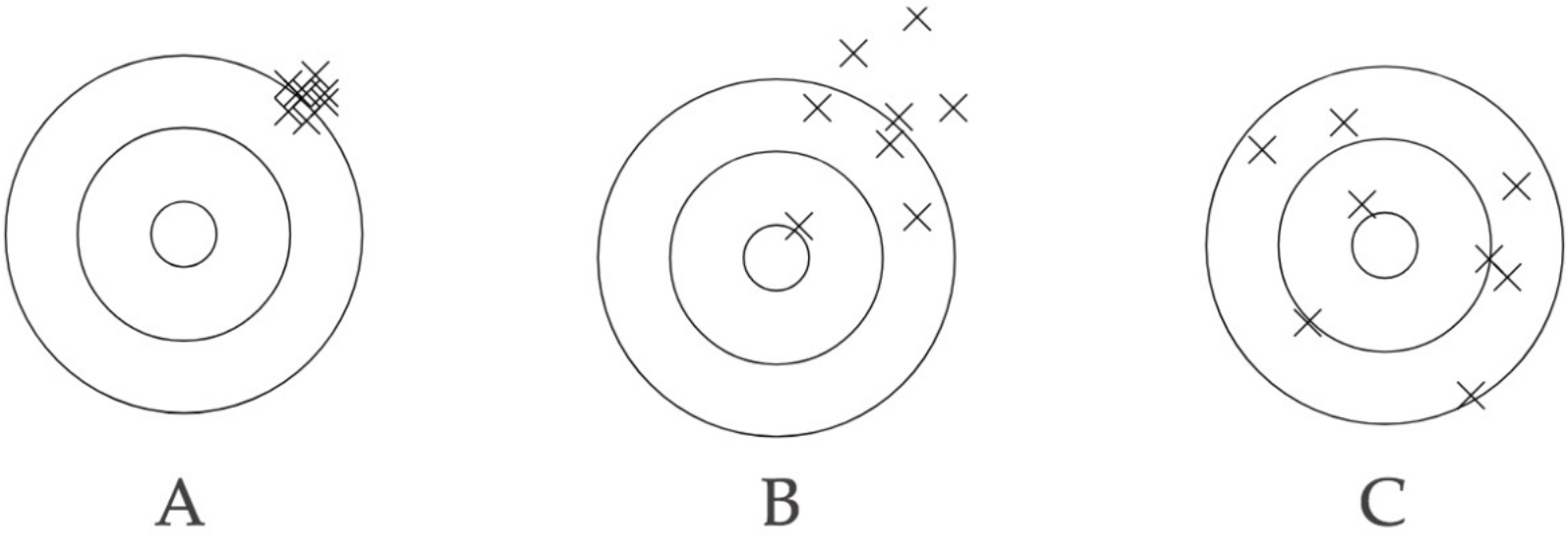
Motivation: bias and variance



Label each picture with variance (high or low) and bias (high or low)

Example from Ameet Soni

Motivation: bias and variance

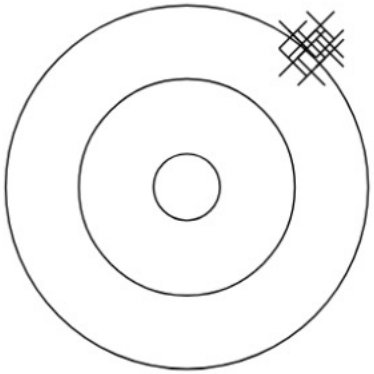


Variance: low
Bias: high

Label each picture with variance (high or low) and bias (high or low)

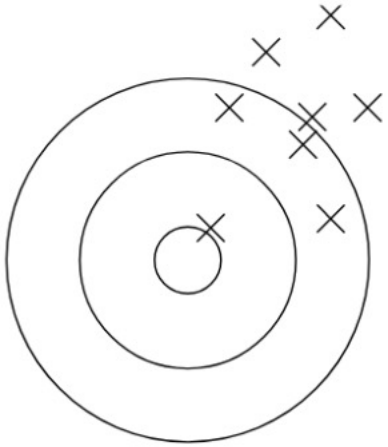
Example from Ameet Soni

Motivation: bias and variance



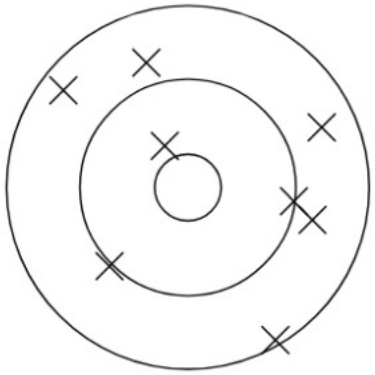
A

Variance: low
Bias: high



B

Variance: high
Bias: high

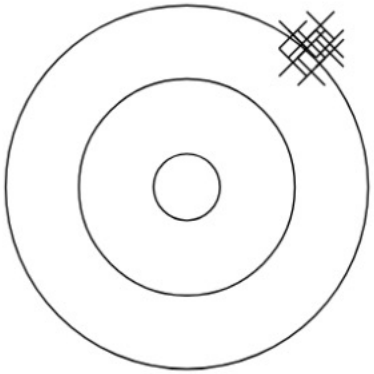


C

Label each picture with variance (high or low) and bias (high or low)

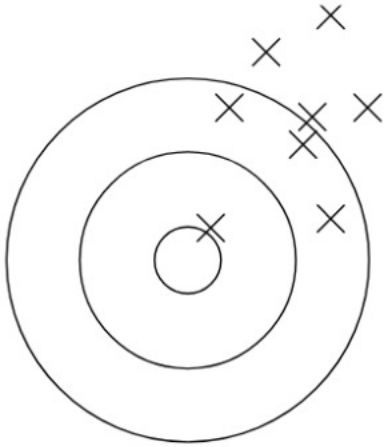
Example from Ameet Soni

Motivation: bias and variance



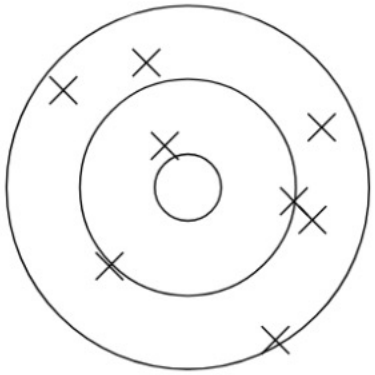
A

Variance: low
Bias: high



B

Variance: high
Bias: high



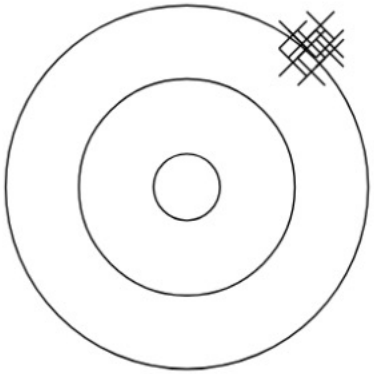
C

Variance: high
Bias: low

Label each picture with variance (high or low) and bias (high or low)

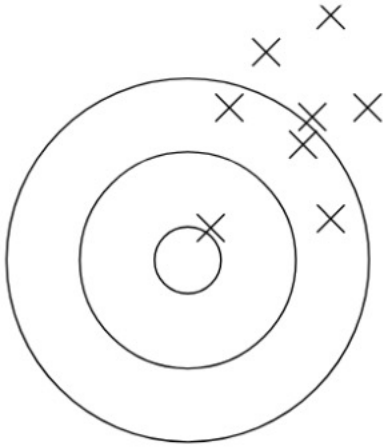
Example from Ameet Soni

Motivation: bias and variance



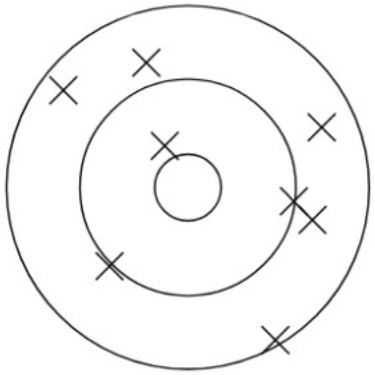
A

Variance: low
Bias: high



B

Variance: high
Bias: high



C

Variance: high
Bias: low

This is the type of classifier we want to average!

Label each picture with variance (high or low) and bias (high or low)

Example from Ameet Soni

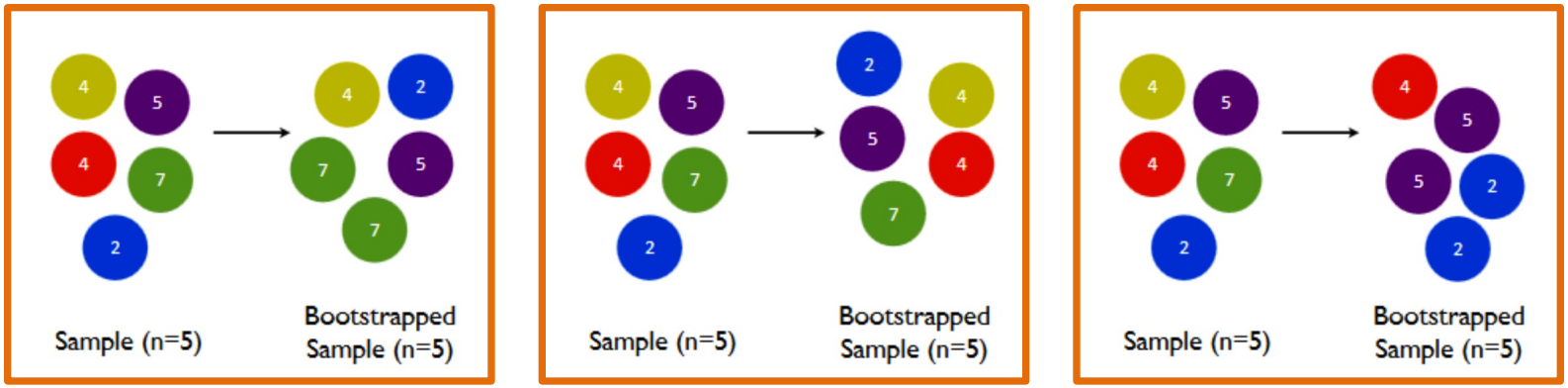
Ensemble Idea

- Average the results from several models with **high variance** and **low bias**
  - Important that models be diverse (don't want them to be wrong in the same ways)

- If n observations each have variance s^2, then the mean of the observations has variance s^2/n (reduce variance by averaging!)

Bagging Algorithm

- Bagging = Bootstrap Aggregation [Brieman, 1996]
- *Bootstrap* (randomly sample with replacement) original data to create many different training sets
- Run base learning algorithm on each new data set independently



Desmond Ong, Stanford

Bagging (Bootstrap Aggregation)

Train:

for t in range(T):

* create bootstrap sample $X^{(t)}$ of size n from training data

* train on $X^{(t)}$ to get model $h^{(t)}$

Test:

for each test example, the T classifiers **vote** on the label

Random Forests

tennis

$T = 3$

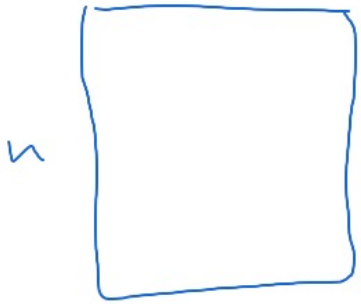
bootstrap

refit classifier

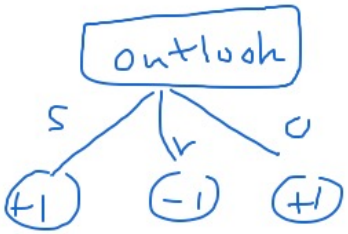
$X^{(1)}$

$h^{(1)}$

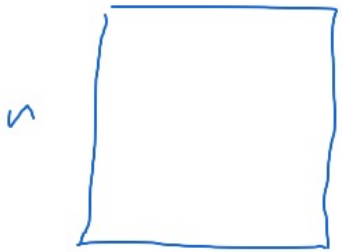
$X^{(2)}$



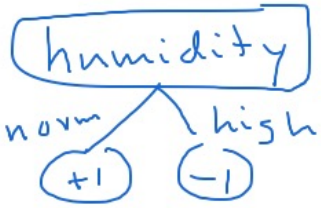
$h^{(2)}$



$X^{(3)}$



$h^{(3)}$



test

$\vec{x} = [r, h, low, high]$ (out, temp, wind, hum)

what do you predict?

$h^{(1)}(\vec{x}) = +1$

$h^{(2)}(\vec{x}) = -1$

$h^{(3)}(\vec{x}) = -1$

vote

$h(\vec{x}) = -1$

Outline for November 9

- Recap Bootstrapping
- Bagging and Random forests
- Midterm 2 Review
  - Revisit confusion matrices
  - PCA (linear transformation + interpretation
  - Naïve Bayes
  - Central Limit Theorem
  - Entropy vs. classification error
  - Logistic regression and cross entropy

Confusion matrix with more classes

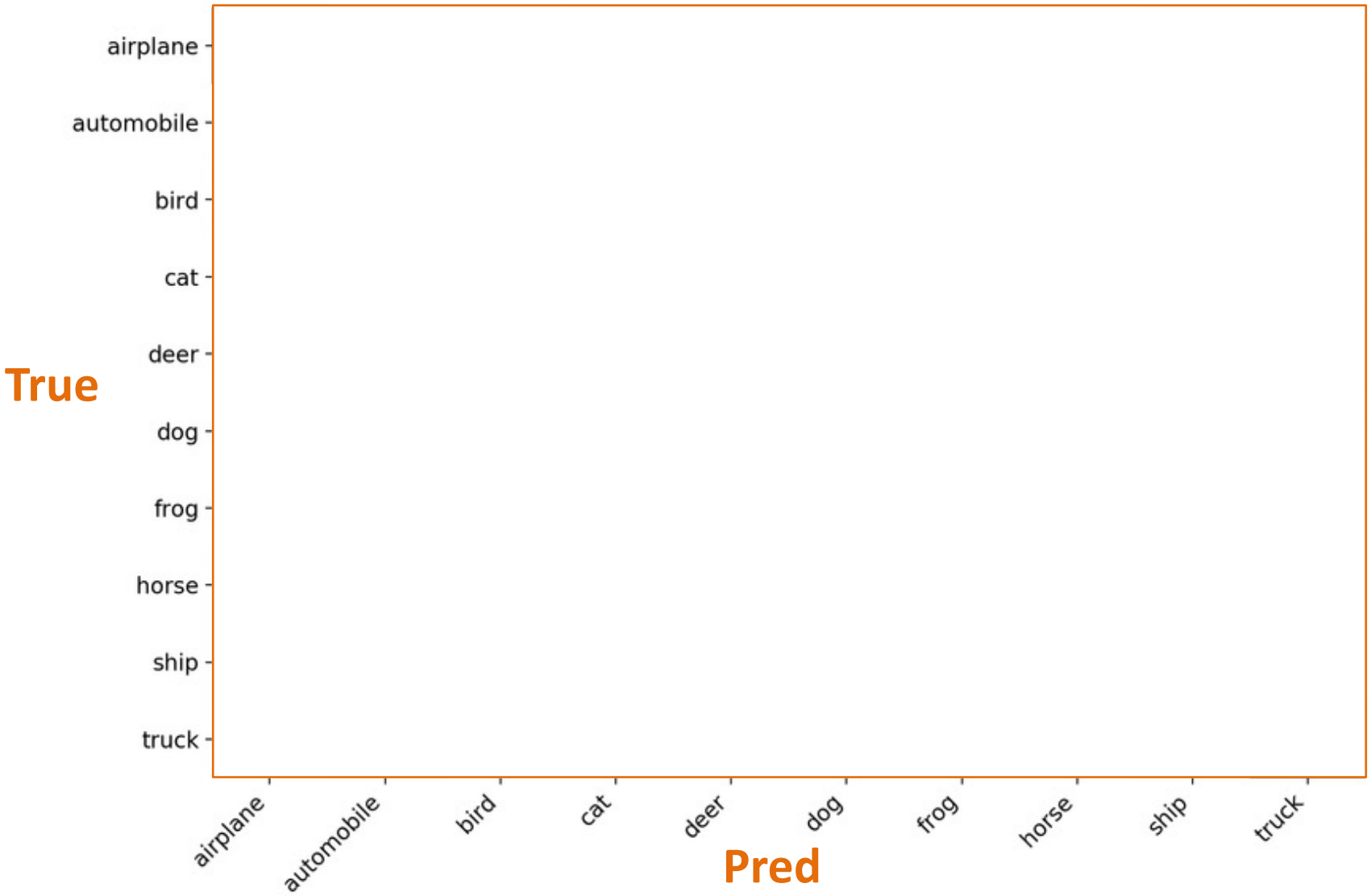


Figure by: Qun Liu (confusion matrix on cifar-10 dataset)

Confusion matrix with more classes

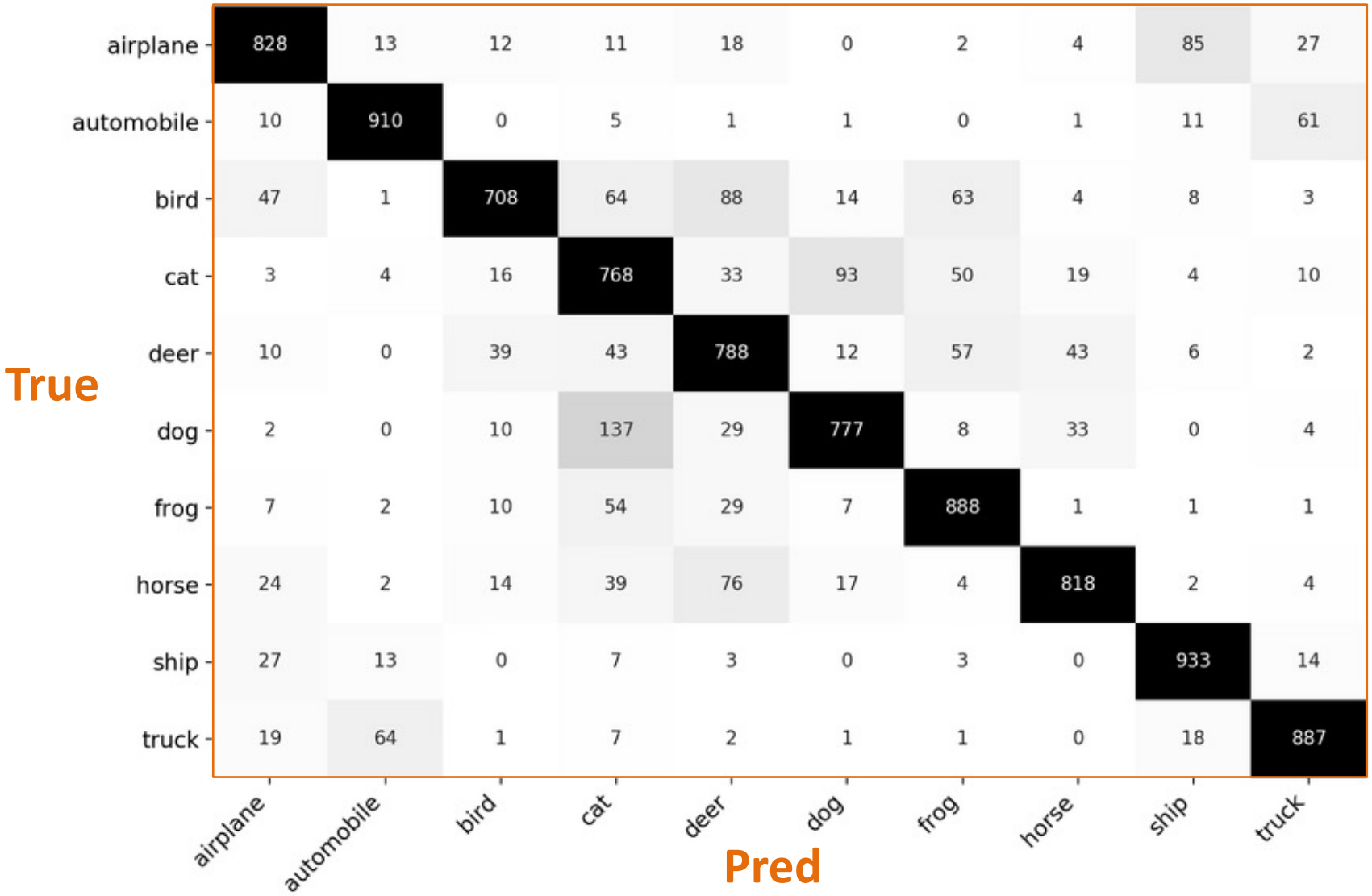
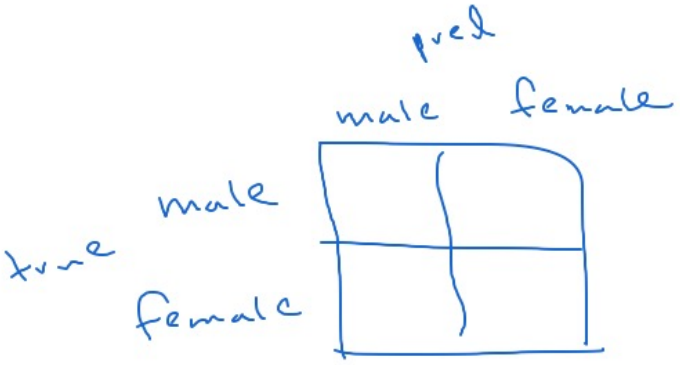


Figure by: Qun Liu (confusion matrix on cifar-10 dataset)

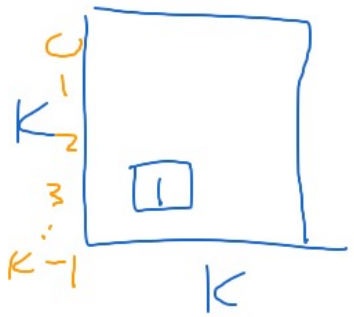
Confusion matrices with just two classes don't have to be "positive" and "negative"



no pos/neg

no ROC curve

Confusion matrices without hard-coding




```
cm = np.zeros((K, K))

for ex in test:
    true = ex.label
    pred = model.classify(ex.features)
    cm[true, pred] += 1
```

indices: 3, 1

Outline for November 9

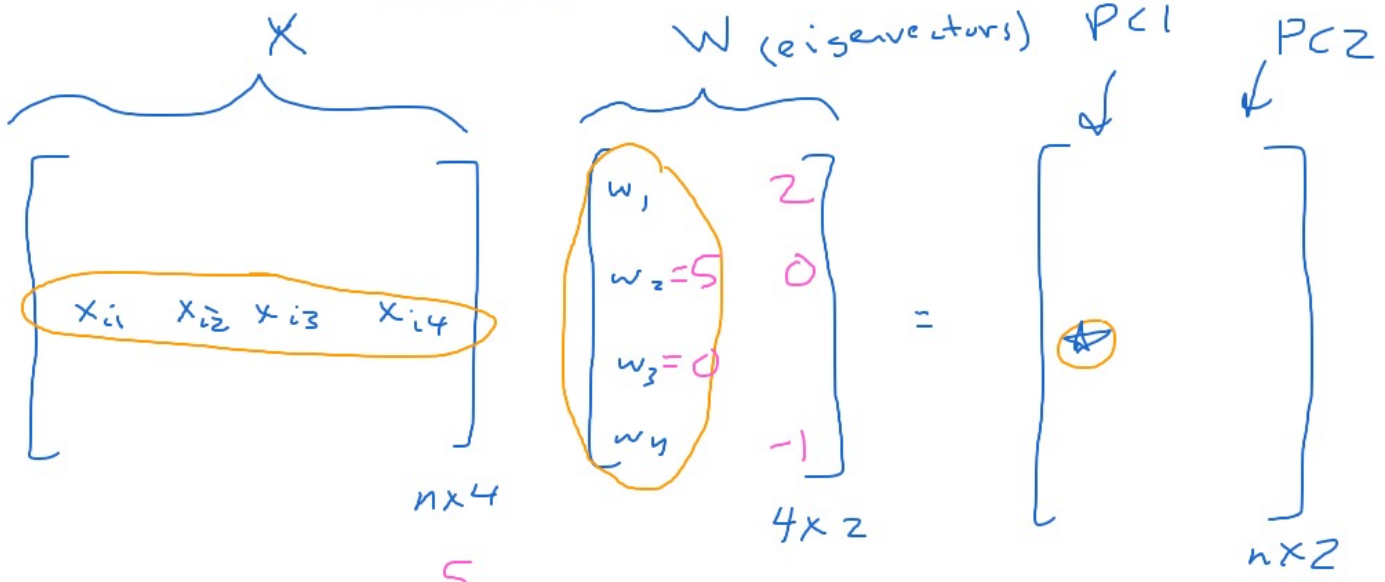
- Recap Bootstrapping
- Bagging and Random forests
- Midterm 2 Review
  - Revisit confusion matrices
  - PCA (linear transformation + interpretation
  - Naïve Bayes
  - Central Limit Theorem
  - Entropy vs. classification error
  - Logistic regression and cross entropy

From the study guide

6. Data Visualization

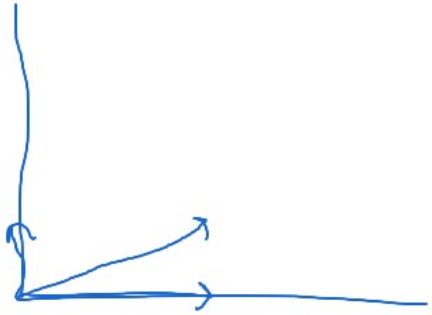
- Best ways of visualizing discrete vs. continuous data
- How to choose colors; idea of sequential, diverging, or qualitative color schemes
- How to make color schemes color-blind and black/white printing friendly
- Idea of principal component analysis (PCA) as a way to accomplish dimensionality reduction
- Using dimensionality reduction to visualize high-dimensional data
- Details of the PCA algorithm (except computing eigenvalues and eigenvectors)
- Runtime of PCA
- Genealogical interpretation of PCA plots for genetic data

PCA creates linear combinations of features



$$\bigstar = w_1 x_{i1} + \underbrace{w_2}_{5} x_{i2} + \cancel{w_3 x_{i3}} + w_4 x_{i4} = \vec{w} \cdot \vec{x}_i$$

linear model!

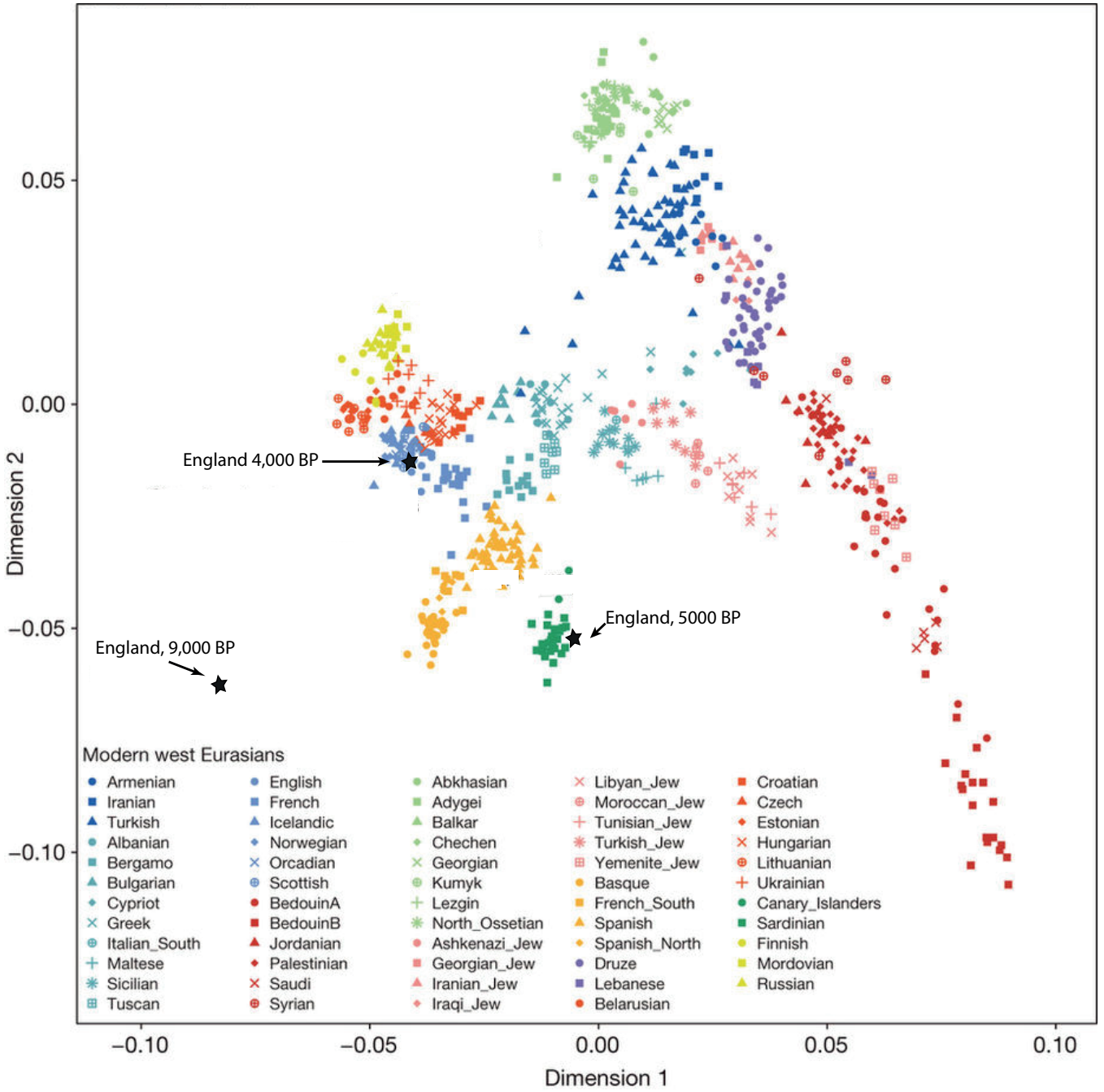


Lab 7 solutions

(not posted online)

Handout 19

(see Piazza)


0.05
0.00
−0.05
−0.10
Dimension 2
England 4,000 BP
England, 5000 BP
England, 9,000 BP
Modern west Eurasians
Armenian
Iranian
Turkish
Albanian
Bergamo
Bulgarian
Cypriot
Greek
Italian_South
Maltese
Sicilian
Tuscan
English
French
Icelandic
Norwegian
Orcadian
Scottish
BedouinA
BedouinB
Jordanian
Palestinian
Saudi
Syrian
Abkhasian
Adygei
Balkar
Chechen
Georgian
Kumyk
Lezgin
North_Ossetian
Ashkenazi_Jew
Georgian_Jew
Iranian_Jew
Iraqi_Jew
Libyan_Jew
Moroccan_Jew
Tunisian_Jew
Turkish_Jew
Yemenite_Jew
Basque
French_South
Spanish
Spanish_North
Druze
Lebanese
Belarusian
Croatian
Czech
Estonian
Hungarian
Lithuanian
Ukrainian
Canary_Islanders
Sardinian
Finnish
Mordovian
Russian
−0.10
−0.05
0.00
0.05
0.10
Dimension 1

Outline for November 9

- Recap Bootstrapping
- Bagging and Random forests
- Midterm 2 Review
  - Revisit confusion matrices
  - PCA (linear transformation + interpretation
  - Naïve Bayes
  - Central Limit Theorem
  - Entropy vs. classification error
  - Logistic regression and cross entropy