CS 260: Foundations of Data Science

Prof. Sara Mathieson

Fall 2023

HAVERFORD
COLLEGE

Admin

- **Lab 8** due tomorrow!
  - In lab today: Lab 8 check-ins (and/or project questions)
- **Exam** goes out today
  - Take in a 3 hour block of your choice
  - Due next Tuesday (Nov 21)
- **Extra credit** opportunity (will be posted on Piazza)
  - Create a video of one of the handouts

Outline for November 14

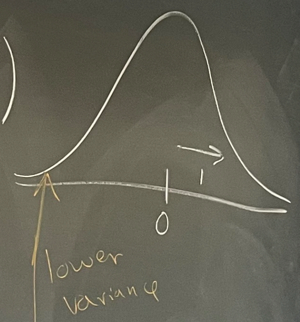
- t-tests
- Bootstrap, Bagging and Random forests
- Midterm 2 Review
  - Revisit confusion matrices
  - PCA (linear transformation + interpretation)
  - Naïve Bayes
  - Central Limit Theorem
  - Entropy vs. classification error
  - Logistic regression and cross entropy

Outline for November 14

- t-tests
- Bootstrap, 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

CLT:

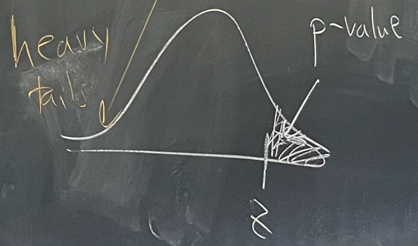
$$z = \left(\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}\right) \sim N(0,1)$$



don't know σ^2?

$\Rightarrow$ use sample variance

$$z = \sqrt{n}\left(\frac{\bar{X}_n-\mu}{S}\right) \sim t\text{-distribution}$$



$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

Outline for November 14

- t-tests
- Bootstrap, 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, ***X*** = [2, 3, 4, 8, 0, 6, 1, 10, 2, 4]

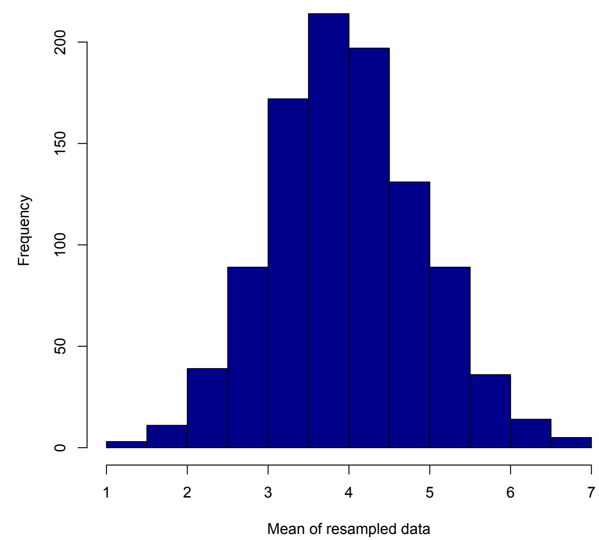
Compute Mean

Resample, with replacement, T times

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, ***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

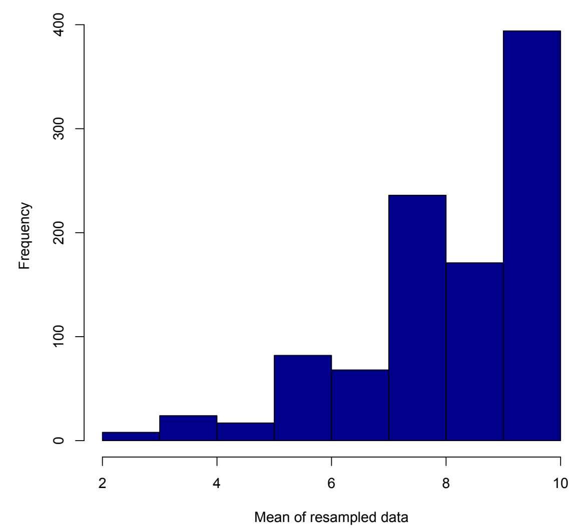
8 3 4 2 10 8 10 8 8 1 → 8

6 4 6 4 6 4 2 4 3 4 0 → 6

… → …

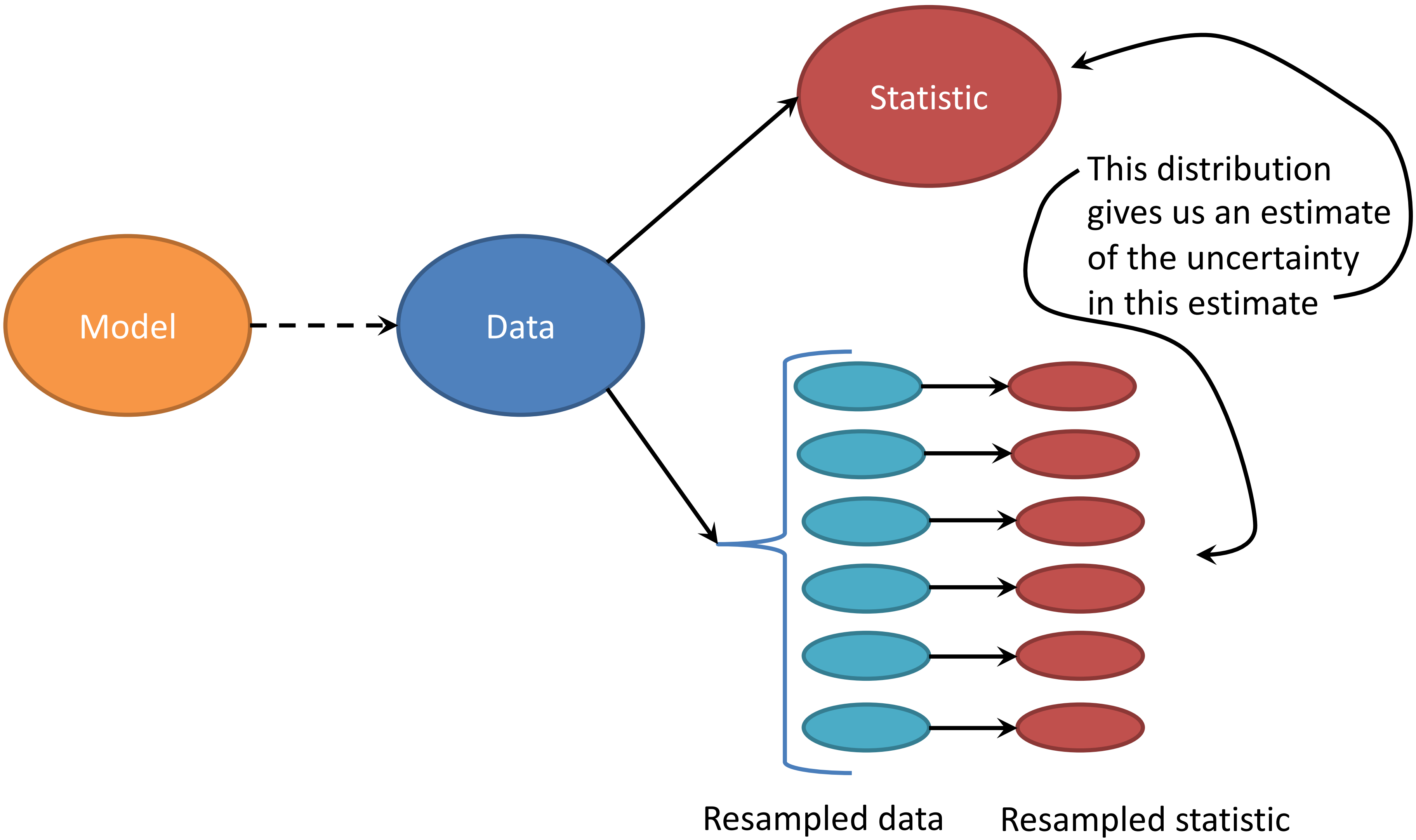
… → …

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



Slide: Iain Mathieson

The bootstrap: Resampling

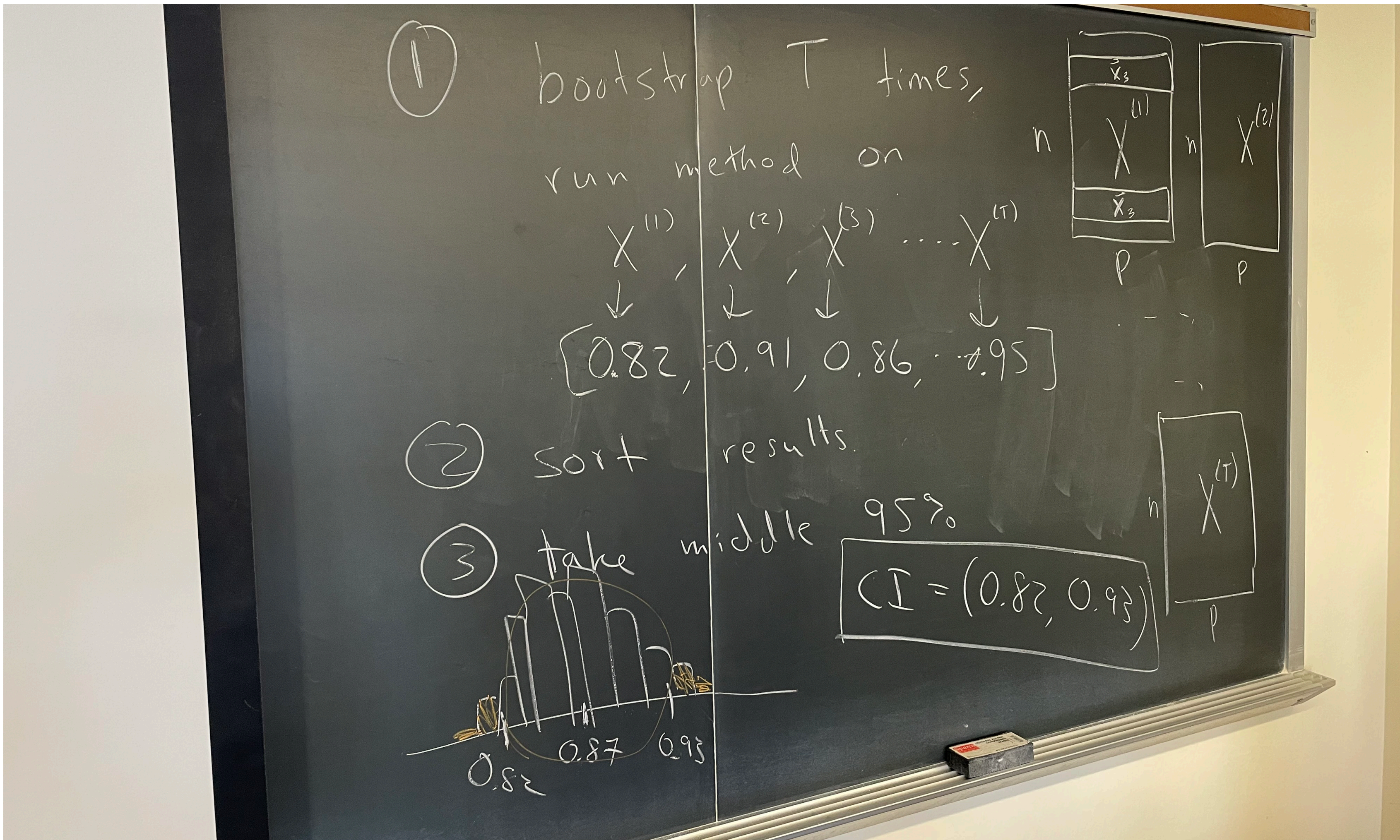


Slide: Iain Mathieson

Bootstrap example

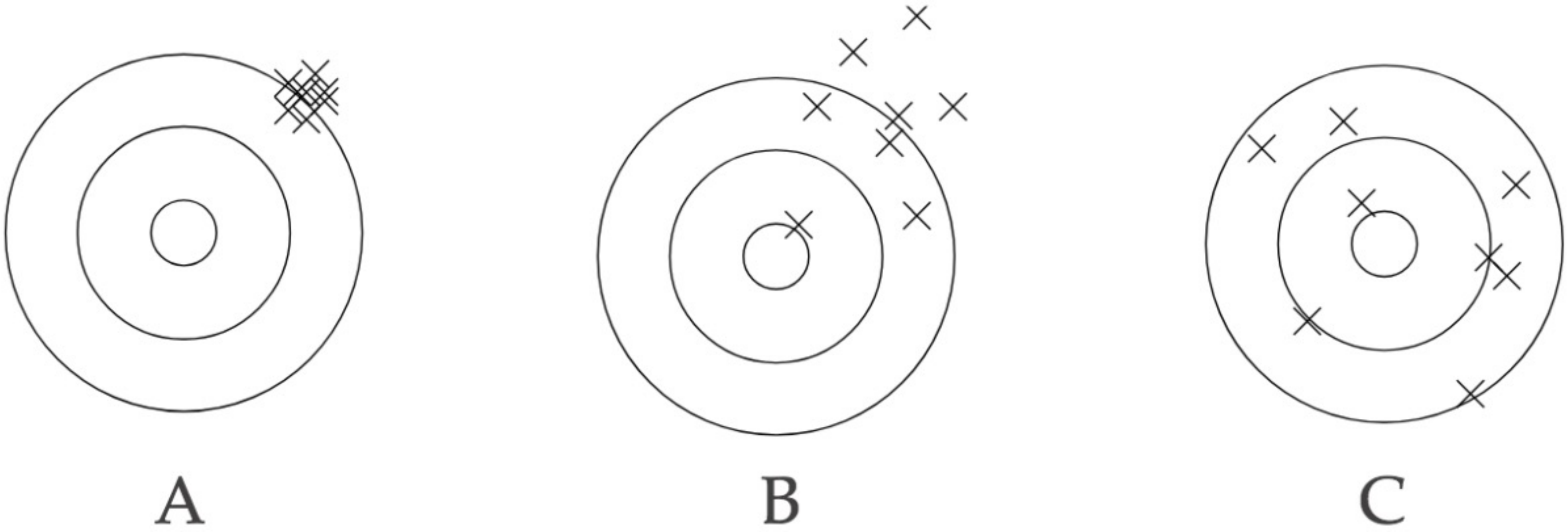
Setup: you obtain 0.87 accuracy on a test dataset using a new algorithm

Goal: find a 95% confidence interval for your estimate



Bagging (Bootstrap Aggregation)

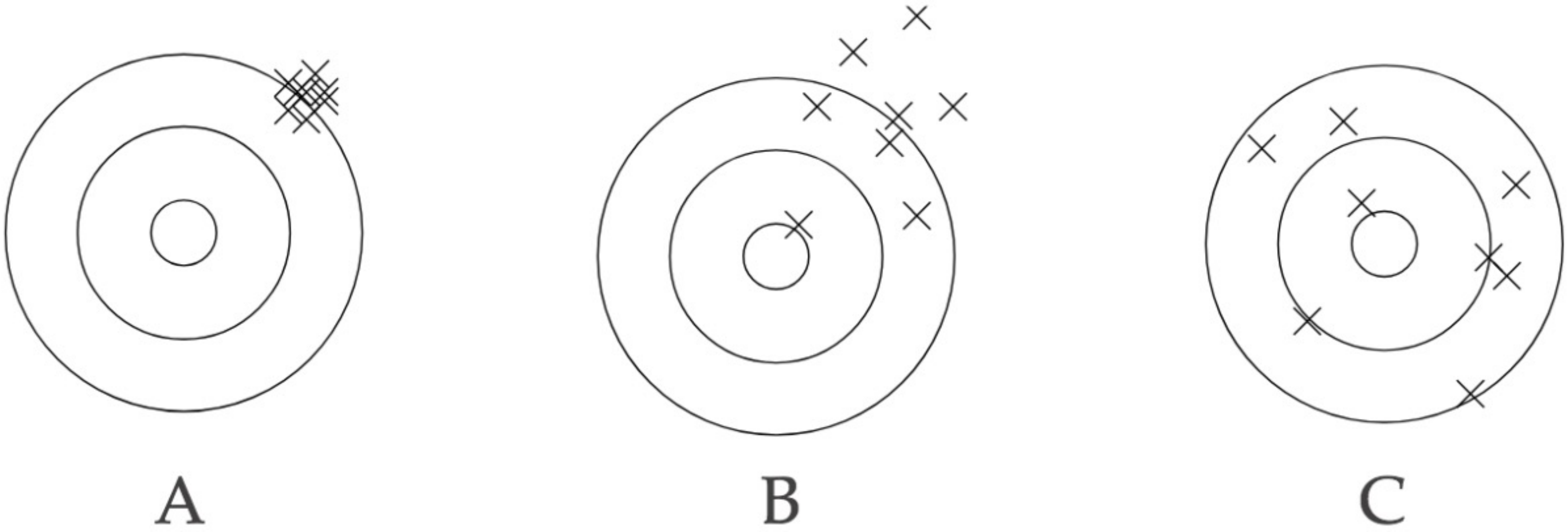
Motivation: bias and variance



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

Example from Ameet Soni

Motivation: bias and variance

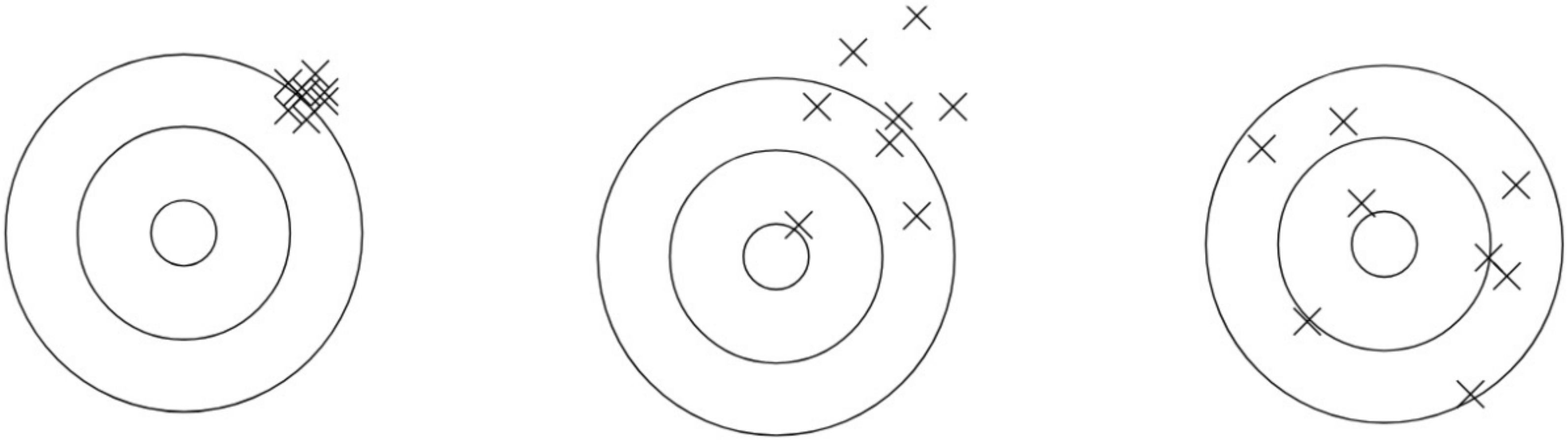


Variance: low
Bias: high

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

Example from Ameet Soni

Motivation: bias and variance



A B C

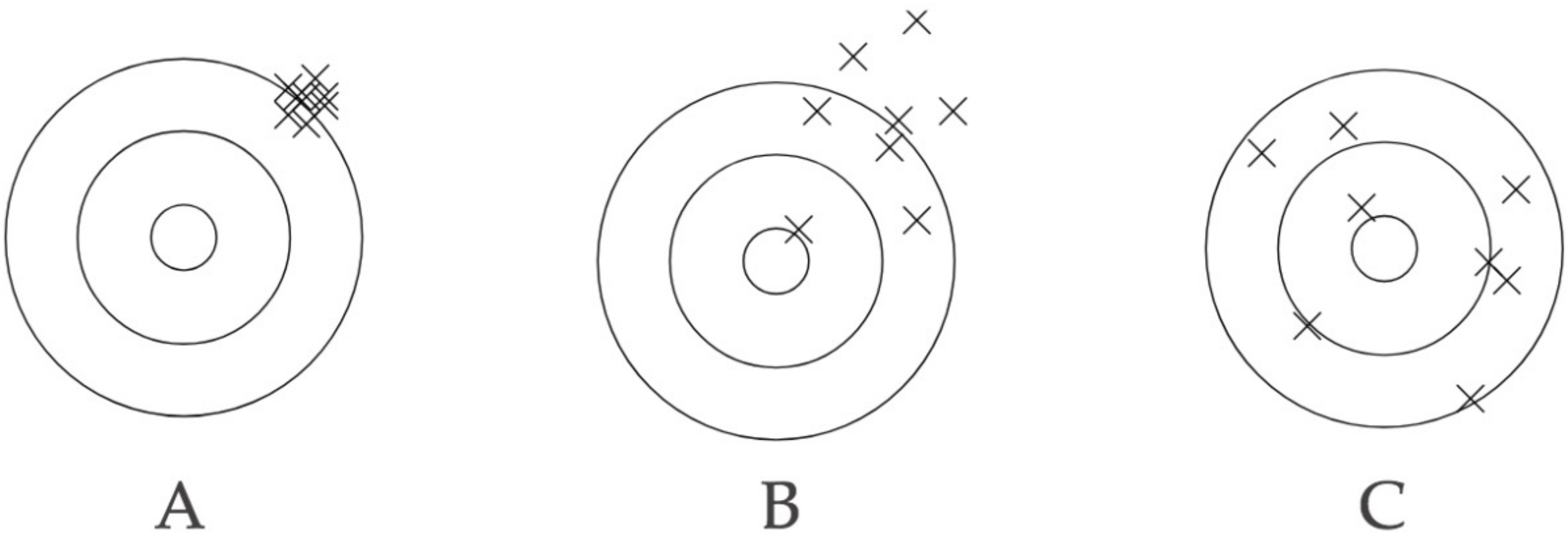
Variance: low
Bias: high

Variance: high
Bias: high

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

Example from Ameet Soni

Motivation: bias and variance

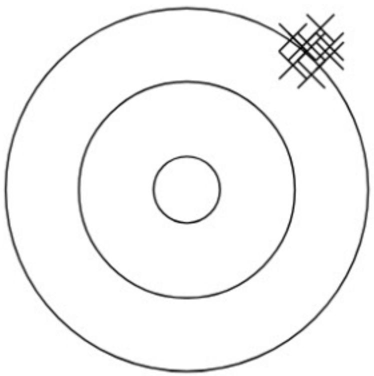


	A	B	C
Variance:	low	high	high
Bias:	high	high	low

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

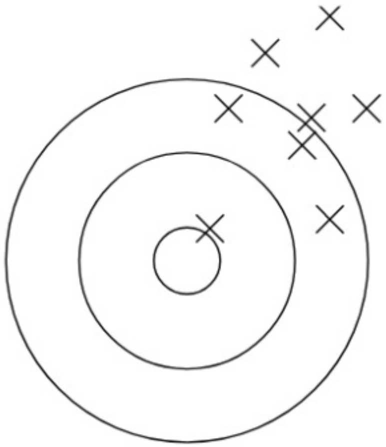
Example from Ameet Soni

Motivation: bias and variance



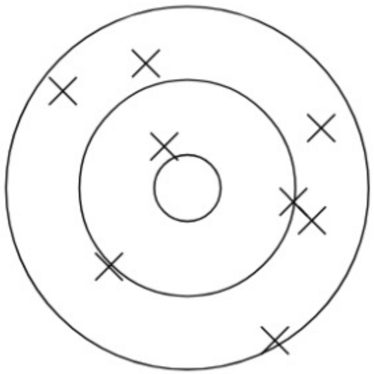
A

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

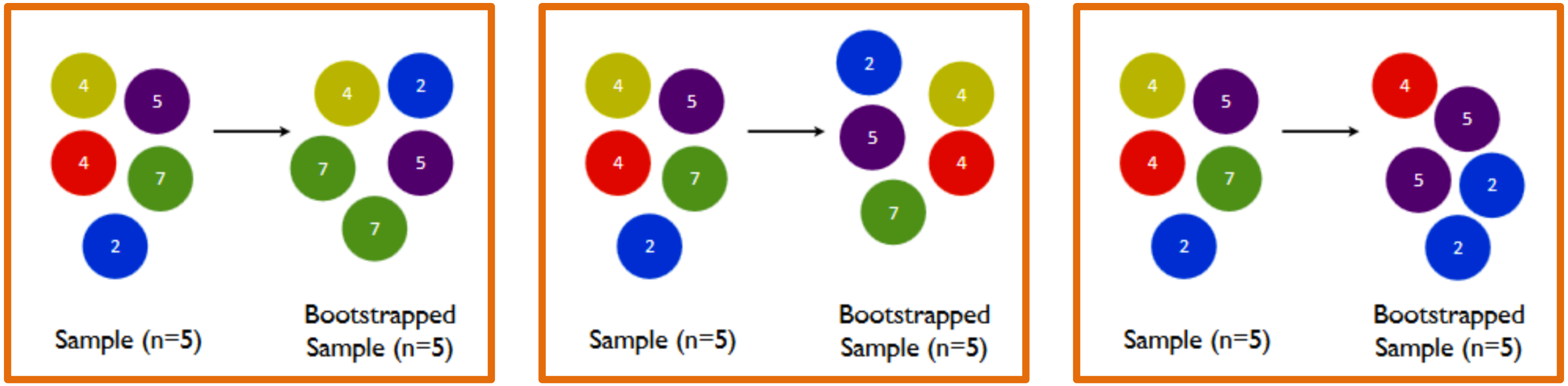
Ensemble Idea

- 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):

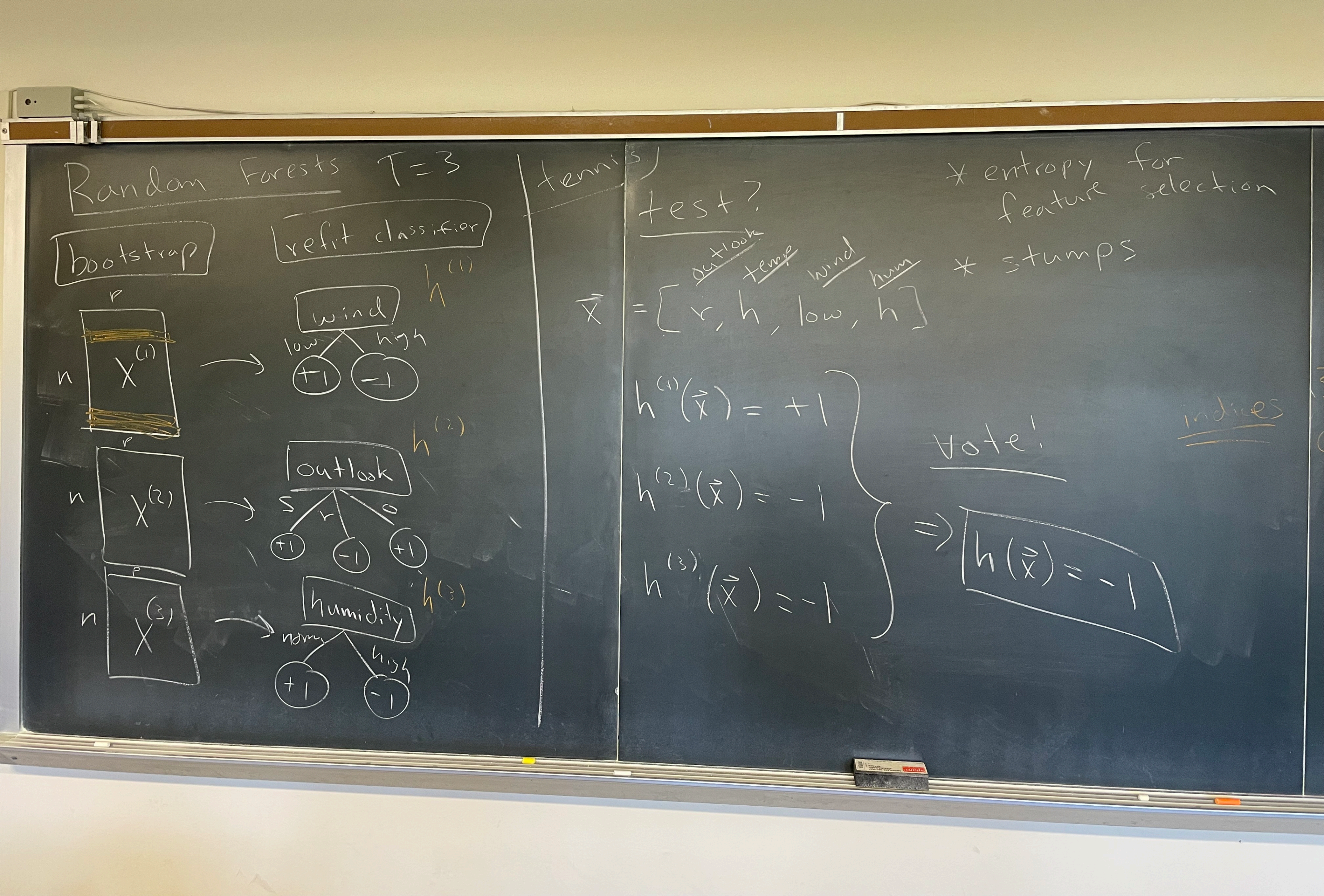
* create bootstrap sample $X^{(t)}$ of size n from training data

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

Test:

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

Random Forests



Outline for November 14

- t-tests
- Bootstrap, 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

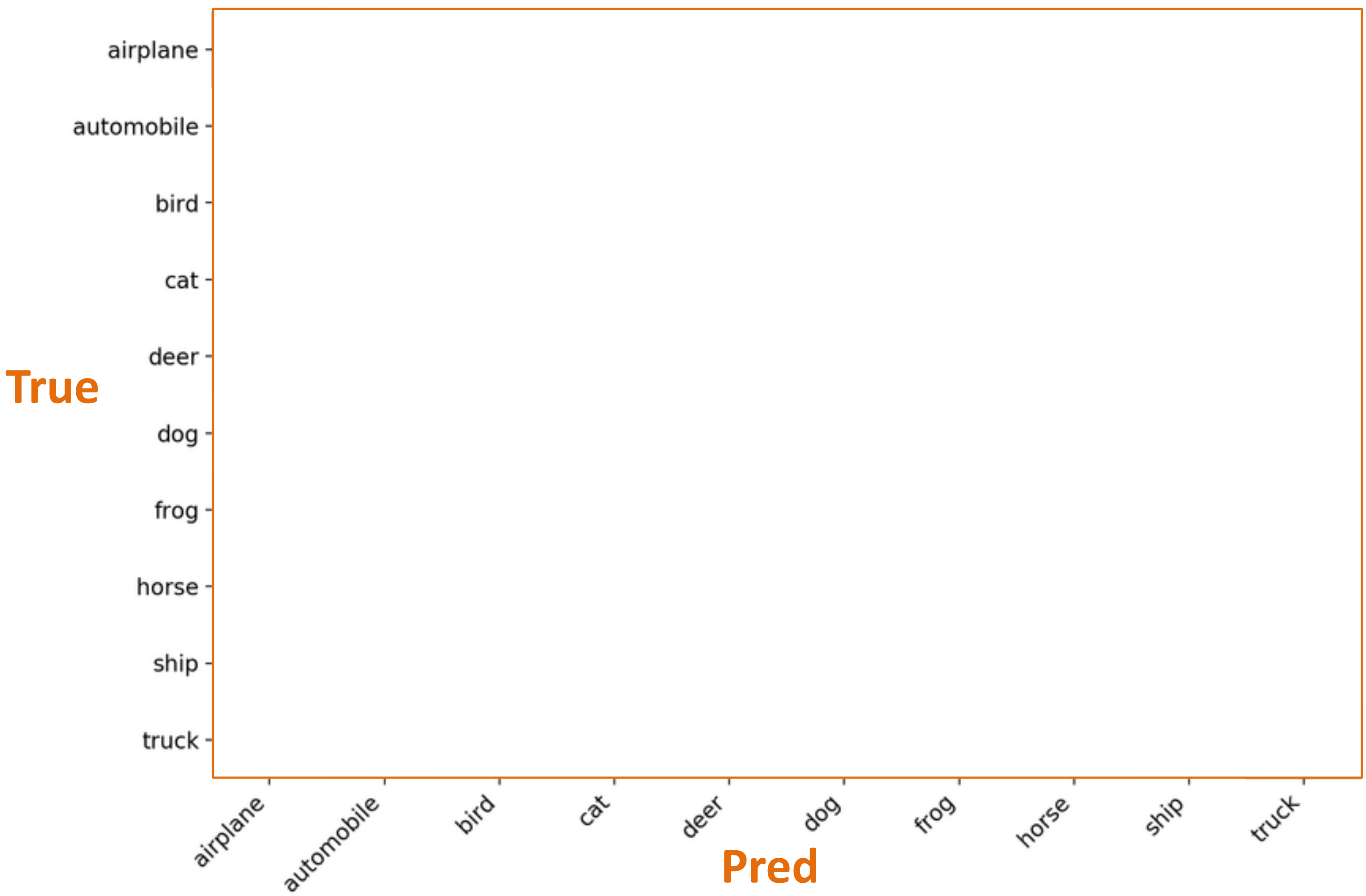


Figure by: Qun Liu (confusion matrix on cifar-10 dataset)

Confusion matrix with more classes

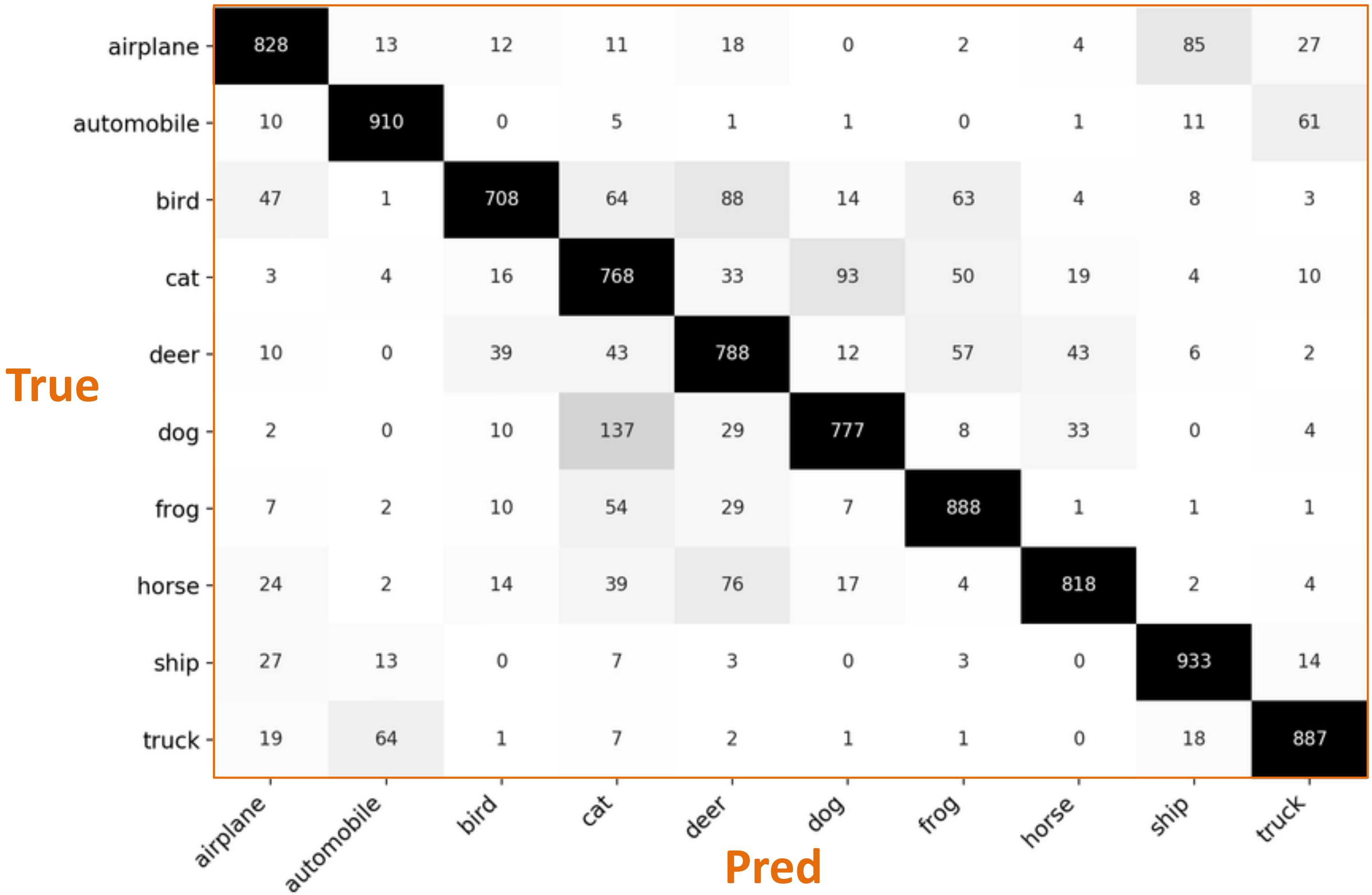
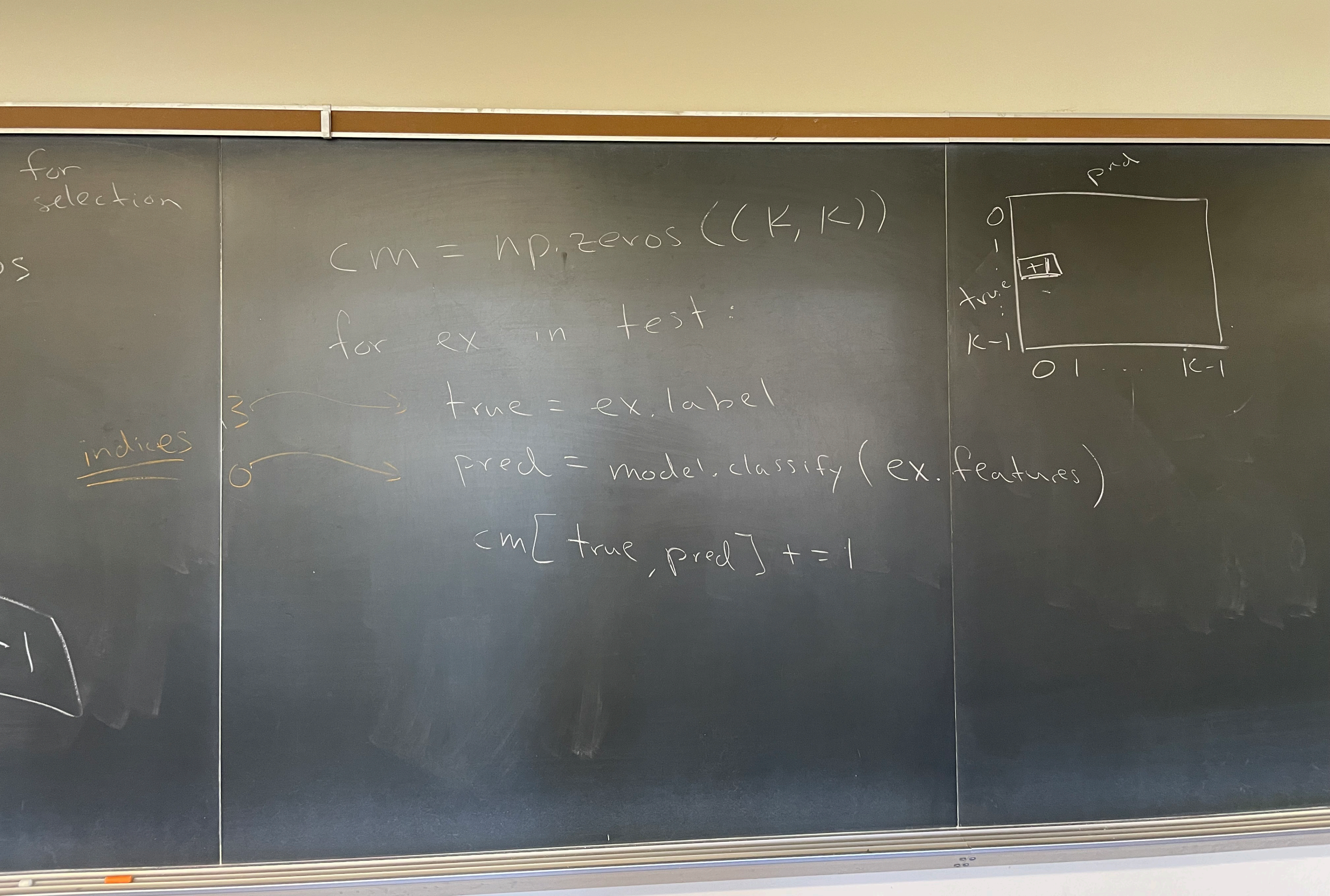


Figure by: Qun Liu (confusion matrix on cifar-10 dataset)

Confusion matrices with just two classes don’t have to be “positive” and “negative”

- Example: male and female
  - No “positive” and “negative” class
  - ROC curve not appropriate

Confusion matrices without hard-coding



Handout 19

Handout 19

unordered

① $n=2$, $\{x_1, x_2\}$ $n=3$

$\{x_1, x_1\}$

$\{x_2, x_2\}$ $\boxed{3}$

$\{x_1, x_2\}$

$\{1, 1, 1\} \Rightarrow 3$

$\{1, 2, 3\} \Rightarrow 1$

$\{1, 1, 2\} \Rightarrow 6$

$\boxed{10}$

$\boxed{\text{Ordered} \Rightarrow n^n}$

② $E[Y] = \sum_y y\,p(y)$

(a)

$= 0 \cdot \frac{1}{8} + 1 \cdot \frac{1}{8} + 2 \cdot \frac{1}{4} + 3$

$= \boxed{2.125}$

(b) $Var(Y) = \sum_y (y - \mu)^2$

$= (0 - 2.125)^2 \cdot \frac{1}{8} + \ldots$

$= \boxed{1.109}$

(2) $E[Y] = \sum_y y\,p(y)$

(a)

$= 0 \cdot \frac{1}{8} + 1 \cdot \frac{1}{8} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{2}$

$= \boxed{2.125}$

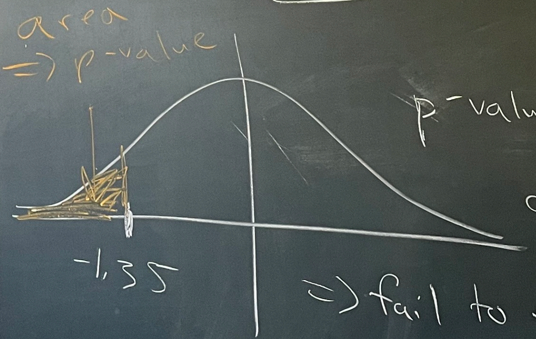
(b) $Var(Y) = \sum_y (y-\mu)^2 p(y)$

$= (0 - 2.125)^2 \cdot \frac{1}{8} + \ldots\ldots$

$= \boxed{1.109}$

(c) $\dfrac{\bar{Y}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \dfrac{1.9 - 2.125}{\sqrt{\frac{1.109}{40}}}$

$\boxed{z = -1.35}$



p-value = 0.08…

$\alpha = 0.05$

⇒ fail to reject H_0

Outline for November 14

- t-tests
- Bootstrap, 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

Next time!