CS 260: Foundations of Data Science

Prof. Thao Nguyen

Fall 2024

HAVERFORD COLLEGE

Materials by Sara Mathieson

Admin

- **Lab 6** grades & feedback posted on Moodle
- **Lab 8** posted (due next Monday 11/18)
- **Midterm 2** will be handed out next Monday
  - Take in a 3-hour block of your choice
  - Due the following Monday (11/25) at the beginning of class
- **Wednesday & Monday**: review sessions

Outline for today

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

Outline for today

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

The bootstrap: Resampling

Data, ***X*** = [2, 3, 4, 8, 0, 6, 1, 10, 2, 4]

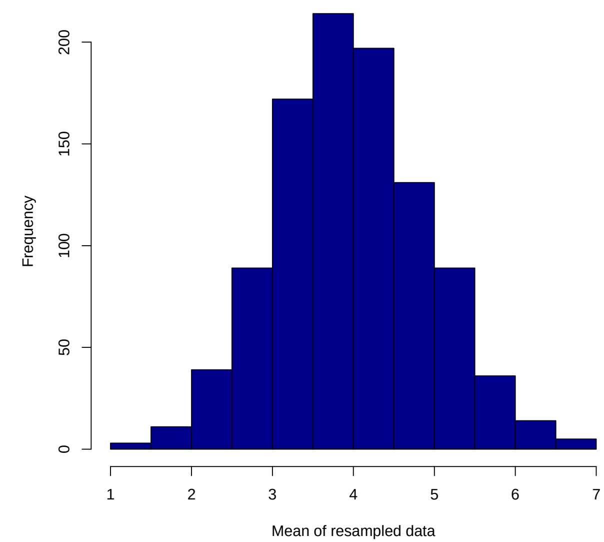
Compute Mean

Resample, with replacement, T times

Resampled data	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)"

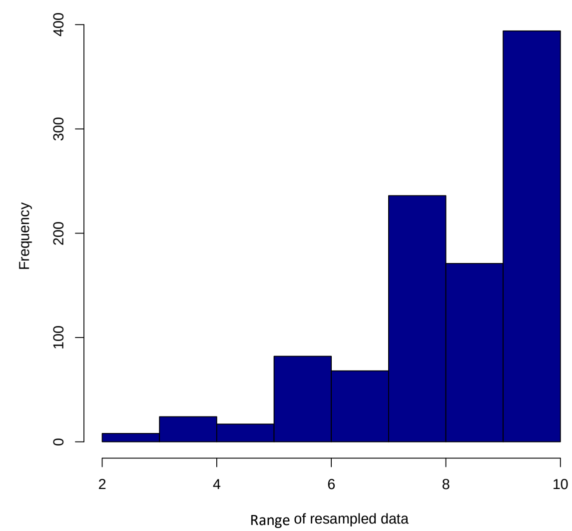
Data, ***X*** = [2, 3, 4, 8, 0, 6, 1, 10, 2, 4]

Compute Range

Resample, with replacement, T times

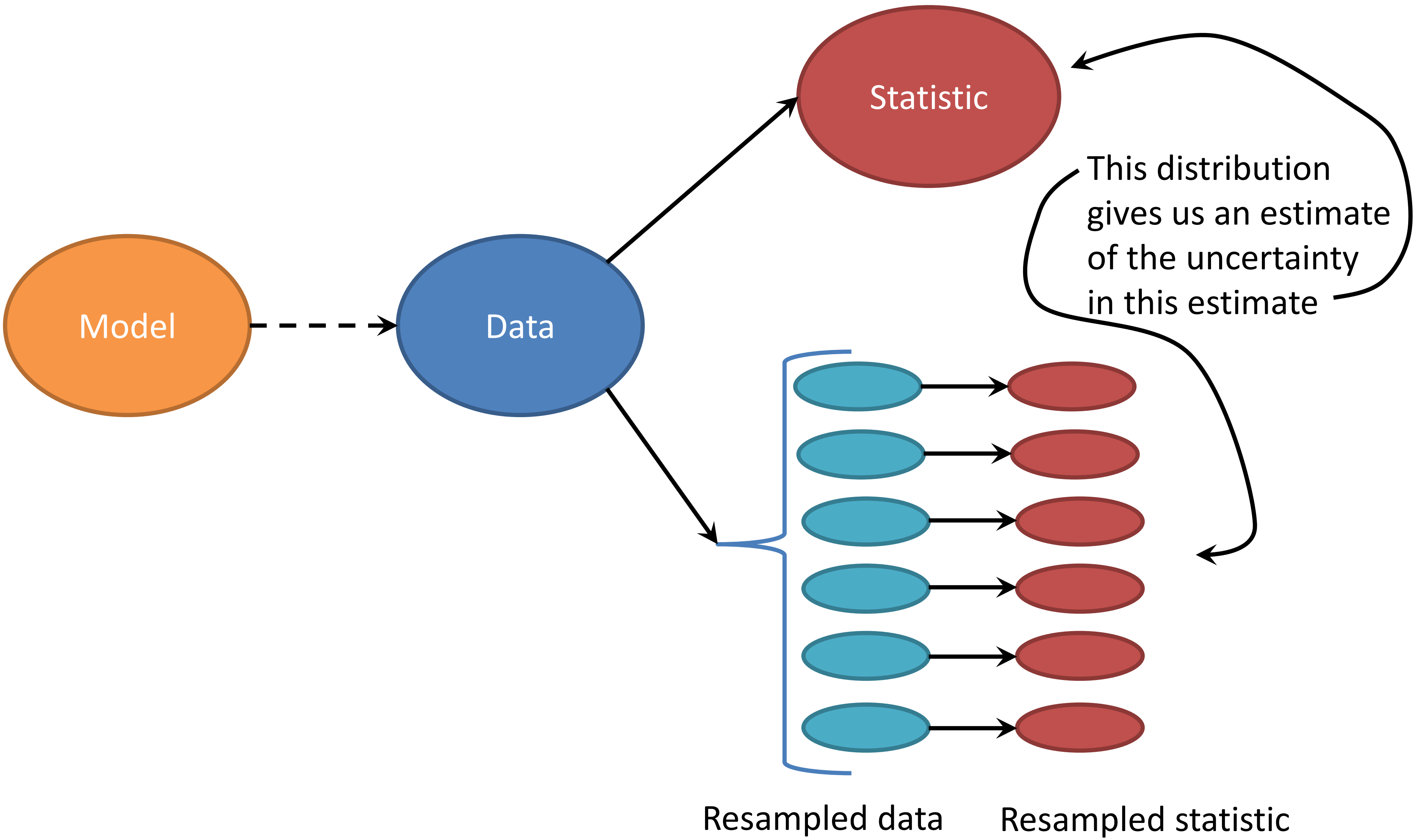
Resampled data		Range
1 8 2 4 6 10 1 1 1 8	→	9
1 0 1 6 4 1 4 2 1 2	→	6
8 1 6 2 6 4 2 4 10 2	→	9
8 3 4 2 10 8 10 8 8 1	→	8
6 4 6 4 6 4 2 4 3 4 0	→	6
…	→	…
…	→	…

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



Slide: Iain Mathieson

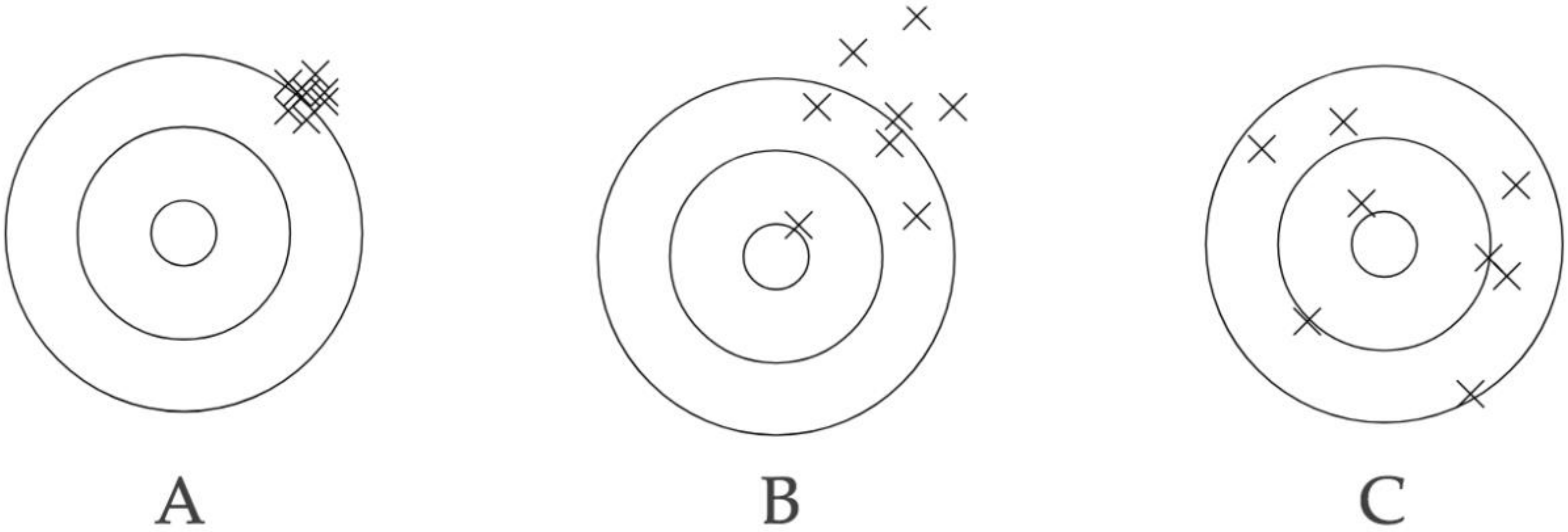
The bootstrap: Resampling



Slide: Iain Mathieson

Bagging (Bootstrap Aggregation)

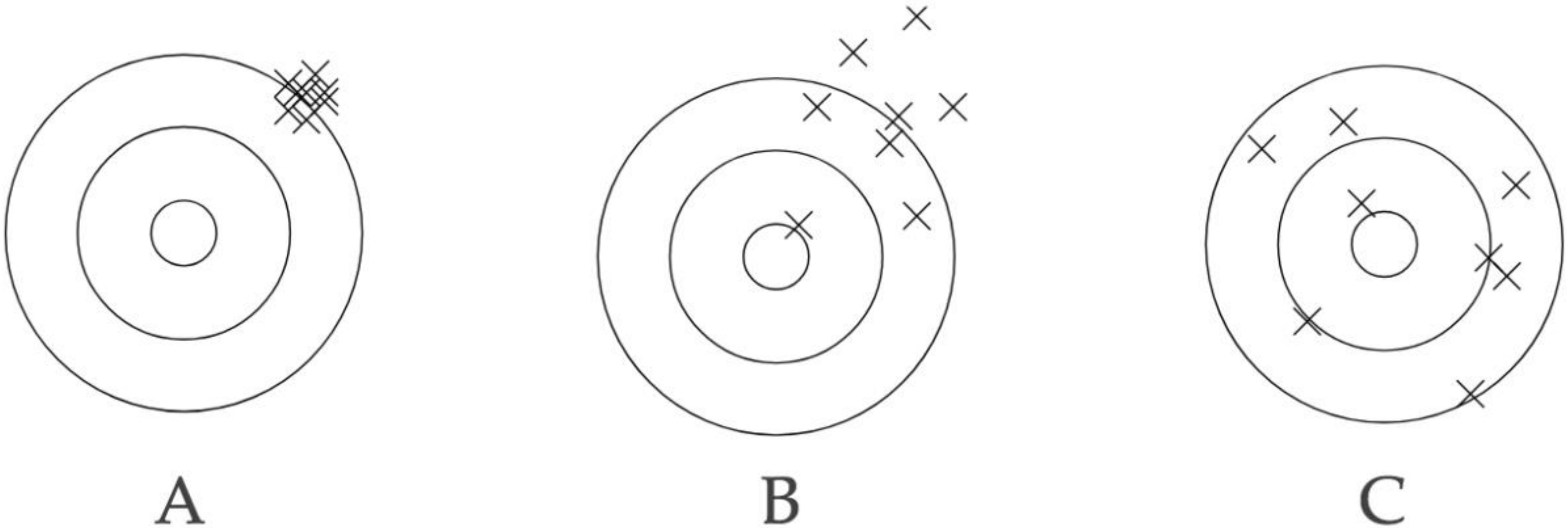
Motivation: bias and variance



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

Example from Ameet Soni

Motivation: bias and variance

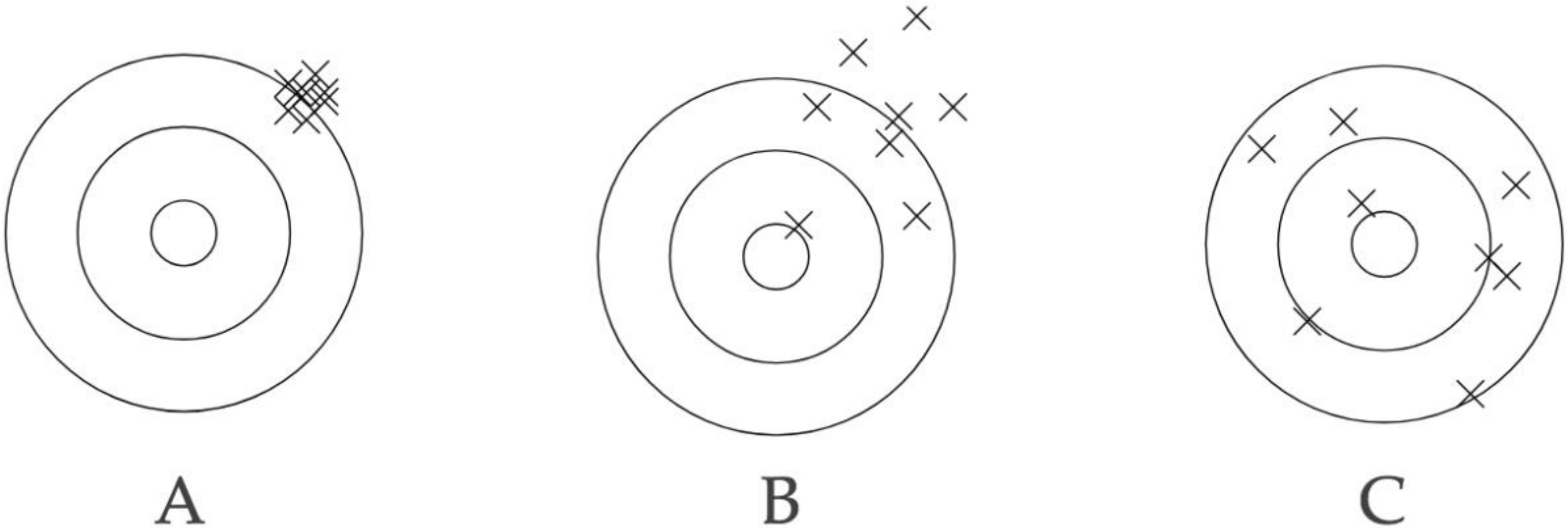


Variance: low
Bias: high

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

Example from Ameet Soni

Motivation: bias and variance

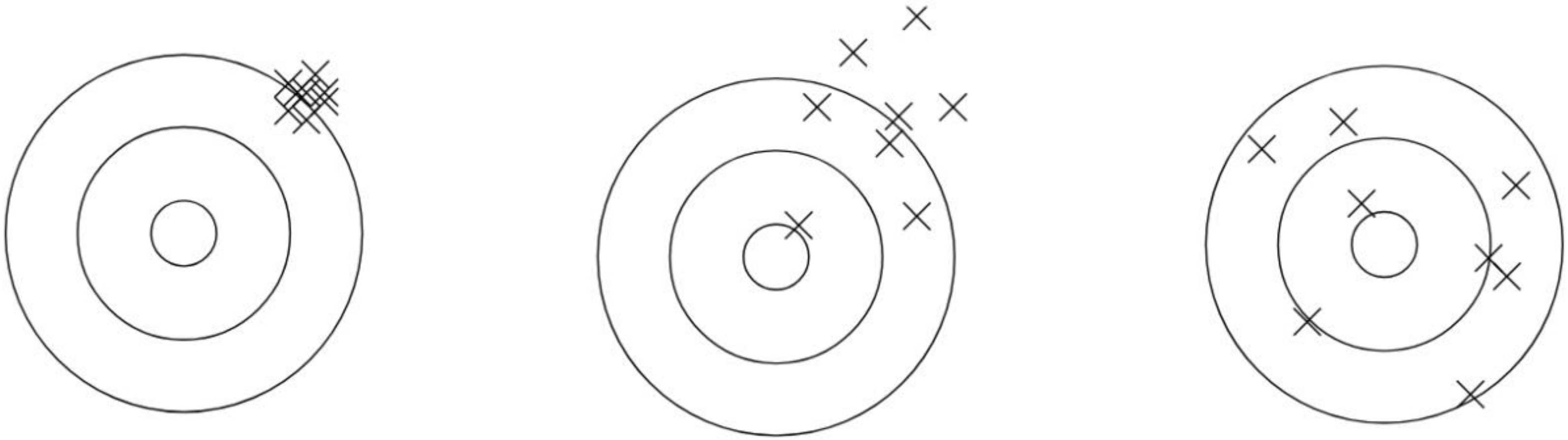


Variance:	low	Variance:	high		
Bias:	high	Bias:	high		

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

Example from Ameet Soni

Motivation: bias and variance



A

Variance: low
Bias: high

B

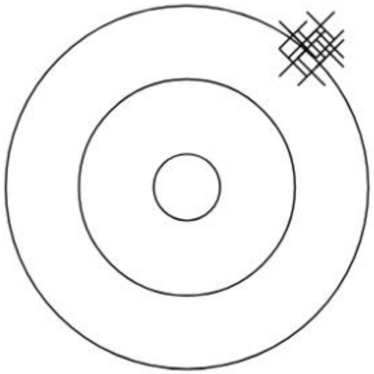
Variance: high
Bias: high

C

Variance: high
Bias: low

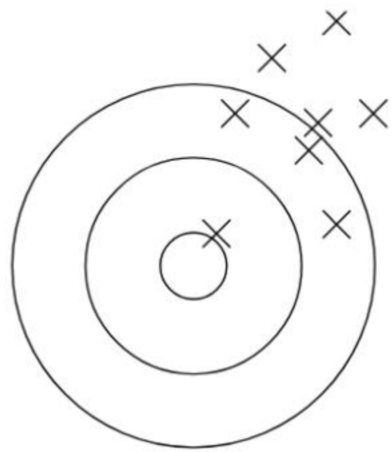
Label each picture with variance (high or low) and bias (high or low)

Motivation: bias and variance



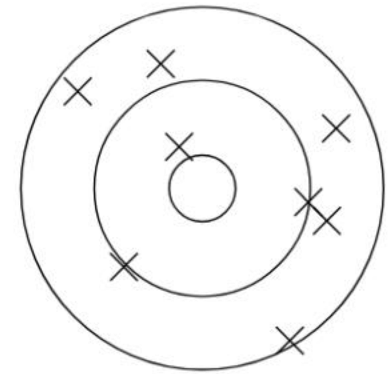
A

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

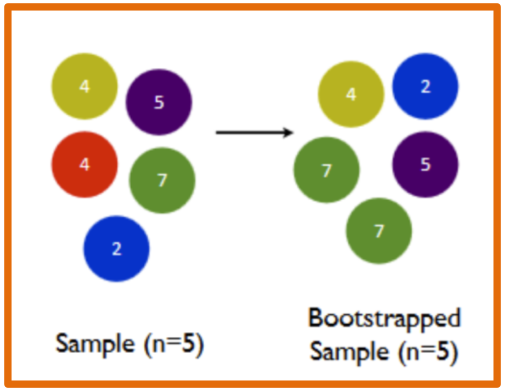
Ensemble Idea

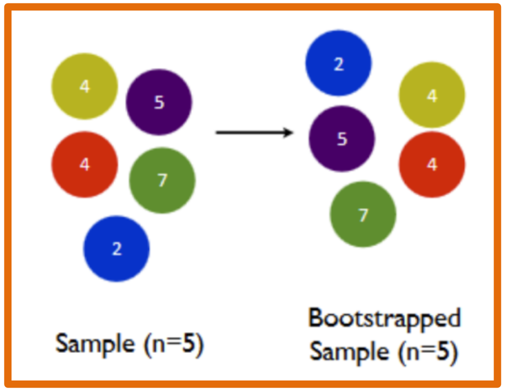
- Average the results from several models with **high variance** and **low bias**
  - Important that models be diverse (don't want them to be wrong in the same ways)

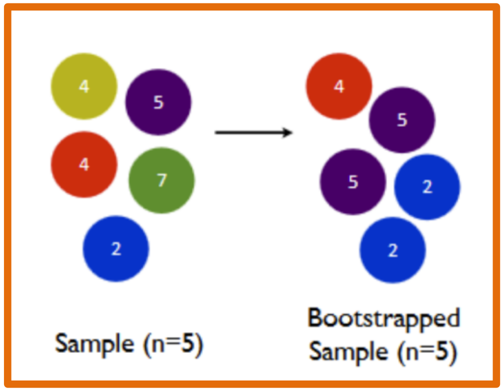
- If n observations each have variance s^2, then the mean of the observations has variance s^2/n (reduce variance by averaging!)

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

Modified from Ameet Soni

Bagging (Bootstrap Aggregation)

Train:

for t in range(T):

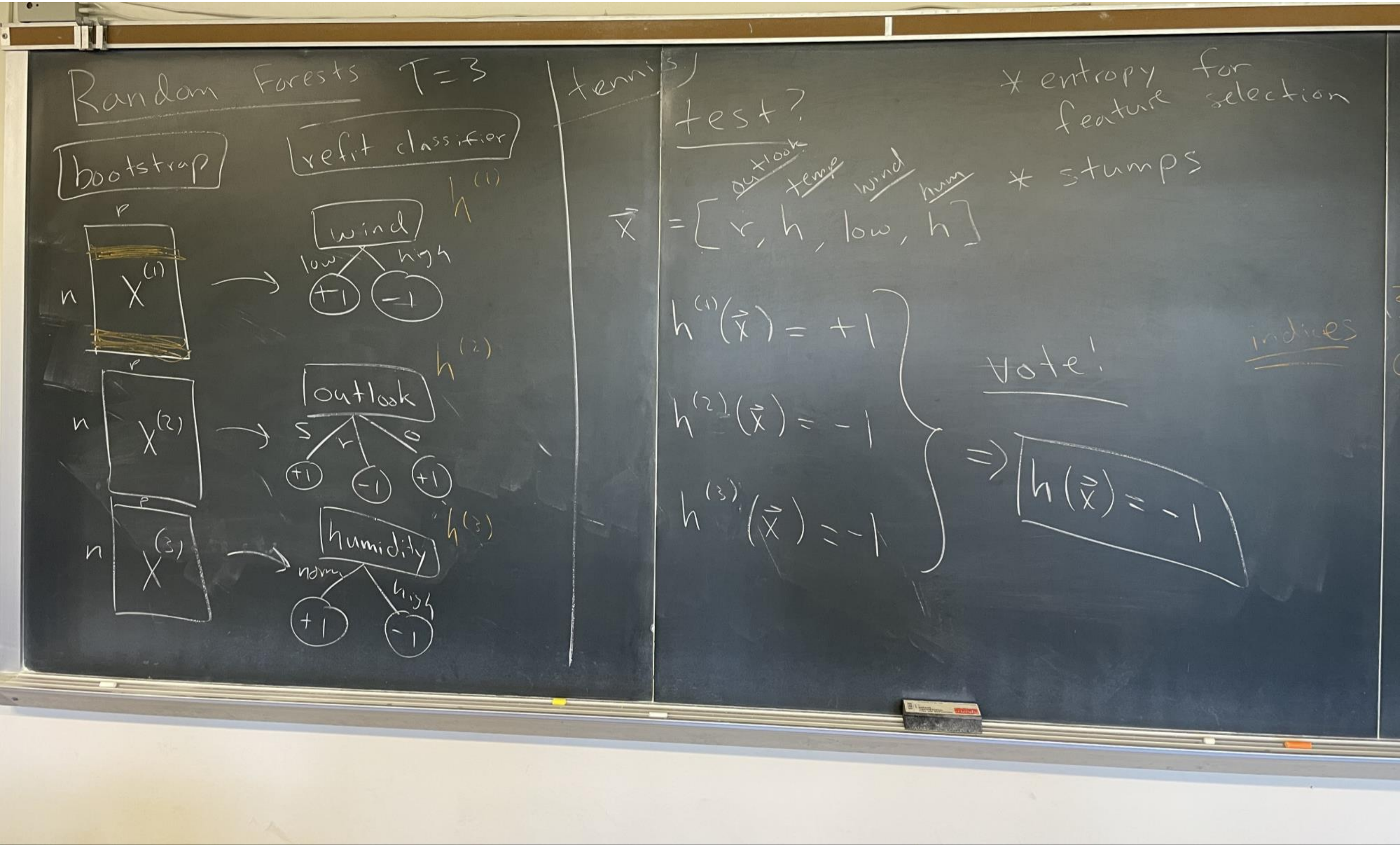
* create bootstrap sample $X^{(t)}$ of size n from training data

* 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 today

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

Confusion matrix with more classes

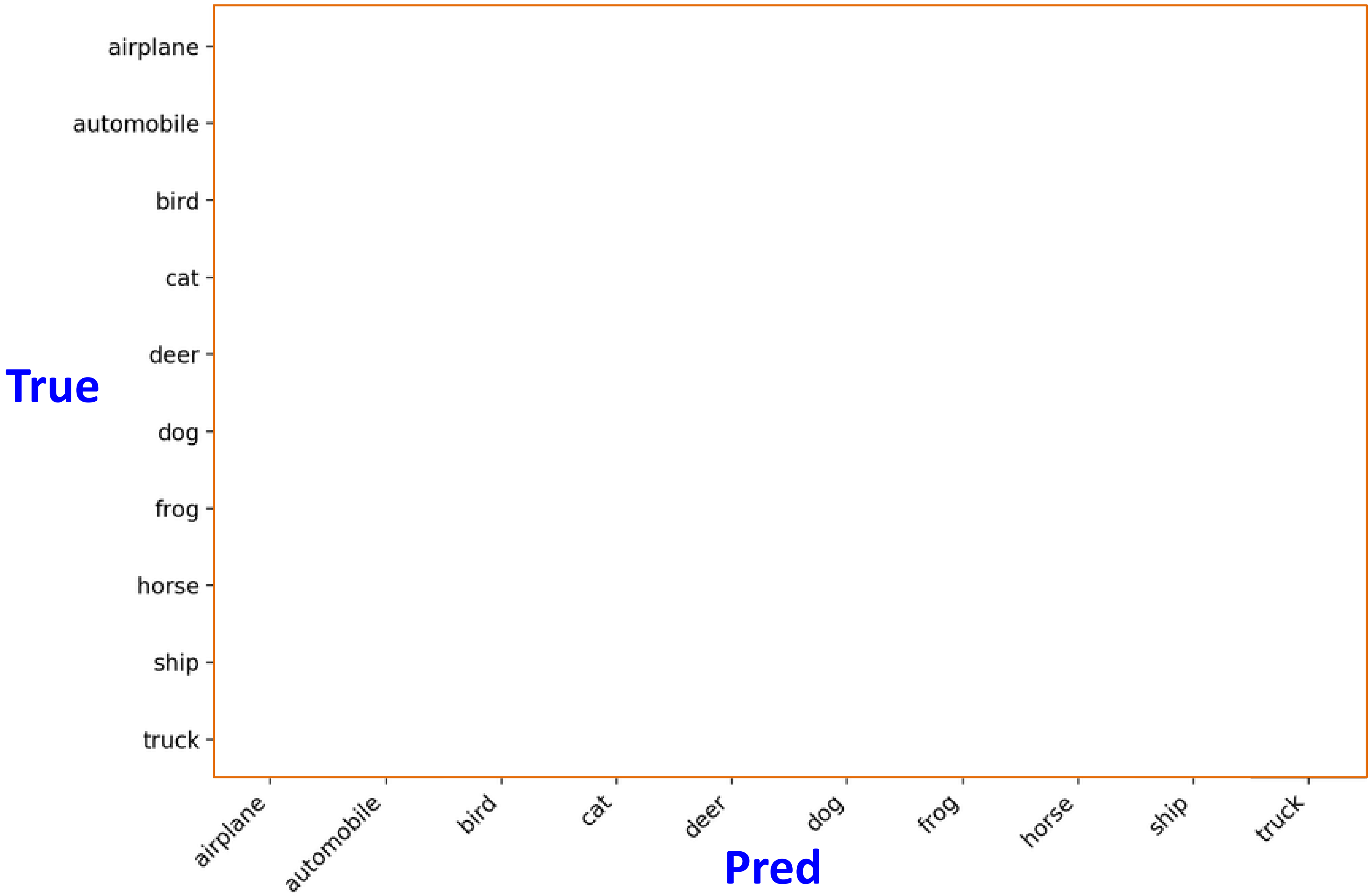


Figure by: Qun Liu (confusion matrix on cifar-10 dataset)

Confusion matrix with more classes

True \ Pred	airplane	automobile	bird	cat	deer	dog	frog	horse	ship	truck
airplane	828	13	12	11	18	0	2	4	85	27
automobile	10	910	0	5	1	1	0	1	11	61
bird	47	1	708	64	88	14	63	4	8	3
cat	3	4	16	768	33	93	50	19	4	10
deer	10	0	39	43	788	12	57	43	6	2
dog	2	0	10	137	29	777	8	33	0	4
frog	7	2	10	54	29	7	888	1	1	1
horse	24	2	14	39	76	17	4	818	2	4
ship	27	13	0	7	3	0	3	0	933	14
truck	19	64	1	7	2	1	1	0	18	887

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

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

Outline for today

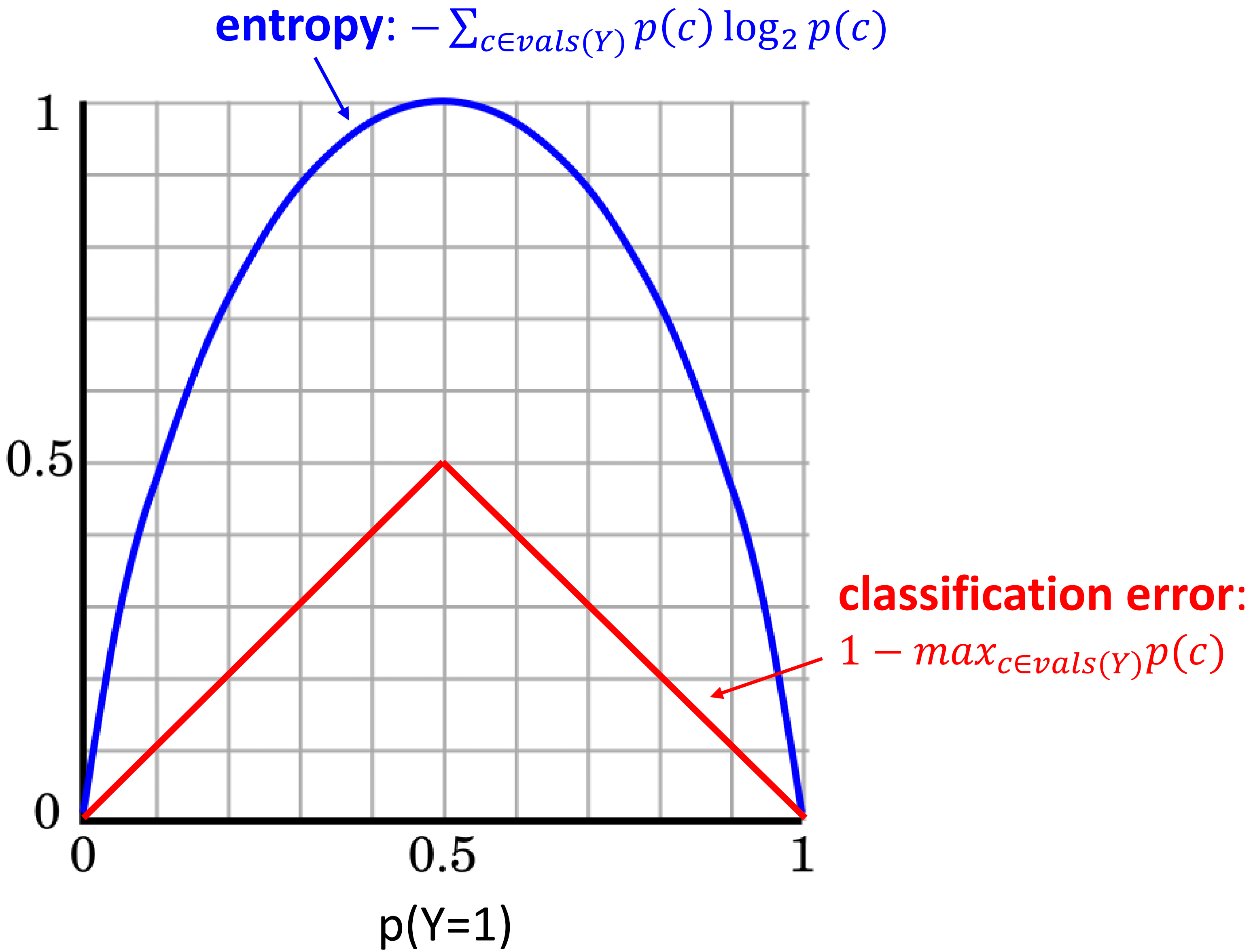
- Bootstrap, Bagging and Random forests
- Midterm 2 Review
  - Revisit confusion matrices
  - Entropy vs. classification error
  - Central Limit Theorem
  - PCA (linear transformation + interpretation)
  - Naïve Bayes
  - Logistic regression and cross entropy

From the study guide

4. Information Theory

- Conceptual idea of entropy as well as formal definition
- Shannon encoding (and decoding), plus how to use entropy to compute average number of bits needed to send one piece of information
- Use of conditional entropy and information gain to choose best features
- Comparison with classification accuracy as a way to choose best features
- How to transform continuous features into binary features? (see Handout 14)

Entropy vs. classification error



Splitting nodes based on entropy

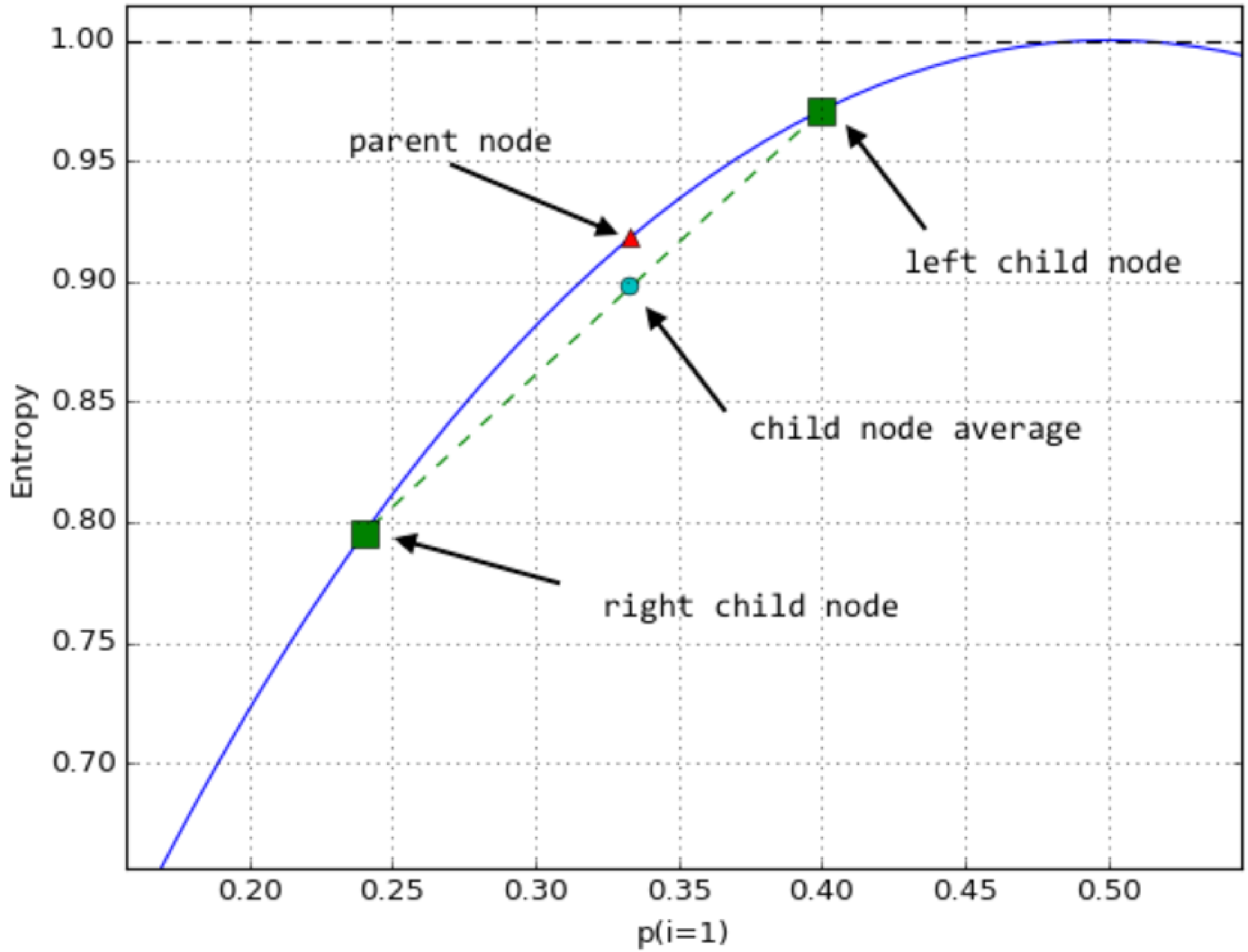
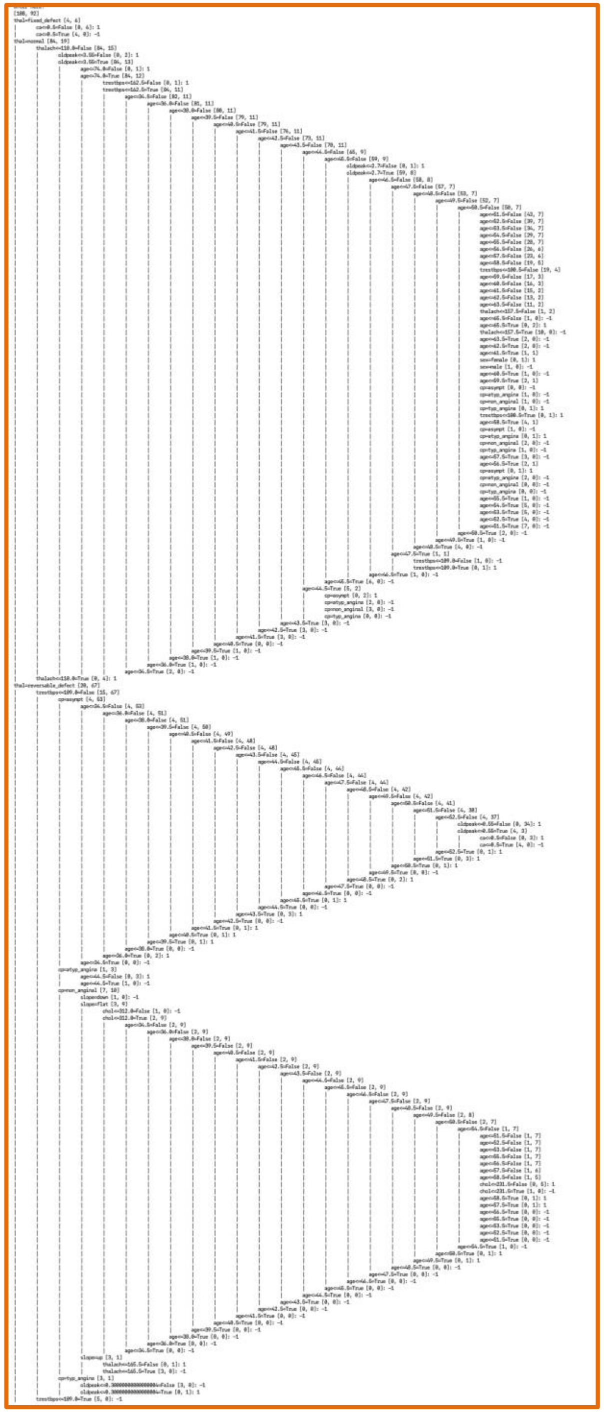


Figure by Sebastian Raschka

Decision trees from entropy (info gain) vs. classification error!

```
[108, 92]
thal=fixed_defect [4, 6]
|       ca<=0.5=False [0, 6]: 1
|       ca<=0.5=True [4, 0]: -1
thal=normal [84, 19]
|       thalach<=110.0=False [84, 15]
|       |       age<=55.5=False [28, 11]
|       |       |       chol<=248.5=False [14, 10]
|       |       |       |       sex=female [13, 3]
|       |       |       |       |       cp=asympt [3, 3]
|       |       |       |       |       |       age<=57.5=False [1, 3]
|       |       |       |       |       |       |       chol<=337.5=False [1, 0]: -1
|       |       |       |       |       |       |       chol<=337.5=True [0, 3]: 1
|       |       |       |       |       |       age<=57.5=True [2, 0]: -1
|       |       |       |       |       cp=atyp_angina [2, 0]: -1
|       |       |       |       |       cp=non_anginal [7, 0]: -1
|       |       |       |       |       cp=typ_angina [1, 0]: -1
|       |       |       |       sex=male [1, 7]
|       |       |       |       |       age<=65.5=False [1, 2]
|       |       |       |       |       |       age<=66.5=False [0, 2]: 1
|       |       |       |       |       |       age<=66.5=True [1, 0]: -1
|       |       |       |       |       age<=65.5=True [0, 5]: 1
|       |       |       chol<=248.5=True [14, 1]
|       |       |       |       oldpeak<=2.7=False [0, 1]: 1
|       |       |       |       oldpeak<=2.7=True [14, 0]: -1
|       |       age<=55.5=True [56, 4]
|       |       |       trestbps<=113.5=False [47, 1]
|       |       |       |       oldpeak<=3.55=False [0, 1]: 1
|       |       |       |       oldpeak<=3.55=True [47, 0]: -1
|       |       |       trestbps<=113.5=True [9, 3]
|       |       |       |       oldpeak<=0.05=False [6, 0]: -1
|       |       |       |       oldpeak<=0.05=True [3, 3]
|       |       |       |       |       cp=asympt [0, 2]: 1
|       |       |       |       |       cp=atyp_angina [2, 0]: -1
|       |       |       |       |       cp=non_anginal [1, 1]
|       |       |       |       |       |       age<=41.5=False [0, 1]: 1
|       |       |       |       |       |       age<=41.5=True [1, 0]: -1
|       |       |       |       |       cp=typ_angina [0, 0]: -1
|       thalach<=110.0=True [0, 4]: 1
thal=reversable_defect [20, 67]
|       cp=asympt [5, 53]
|       |       oldpeak<=0.55=False [0, 43]: 1
|       |       oldpeak<=0.55=True [5, 10]
|       |       |       chol<=237.5=False [0, 8]: 1
|       |       |       chol<=237.5=True [5, 2]
|       |       |       |       chol<=179.5=False [4, 0]: -1
|       |       |       |       chol<=179.5=True [1, 2]
|       |       |       |       |       age<=59.5=False [1, 0]: -1
|       |       |       |       |       age<=59.5=True [0, 2]: 1
|       cp=atyp_angina [3, 3]
|       |       age<=46.5=False [1, 3]
|       |       |       trestbps<=109.0=False [0, 3]: 1
|       |       |       trestbps<=109.0=True [1, 0]: -1
|       |       age<=46.5=True [2, 0]: -1
|       cp=non_anginal [9, 10]
|       |       oldpeak<=1.85=False [0, 5]: 1
|       |       oldpeak<=1.85=True [9, 5]
|       |       |       trestbps<=121.0=False [3, 5]
|       |       |       |       chol<=232.5=False [0, 4]: 1
|       |       |       |       chol<=232.5=True [3, 1]
|       |       |       |       |       trestbps<=128.5=False [3, 0]: -1
|       |       |       |       |       trestbps<=128.5=True [0, 1]: 1
|       |       |       trestbps<=121.0=True [6, 0]: -1
|       cp=typ_angina [3, 1]
|       |       oldpeak<=0.30000000000000004=False [3, 0]: -1
|       |       oldpeak<=0.30000000000000004=True [0, 1]: 1
```



Outline for today

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

From the study guide

7. Statistics

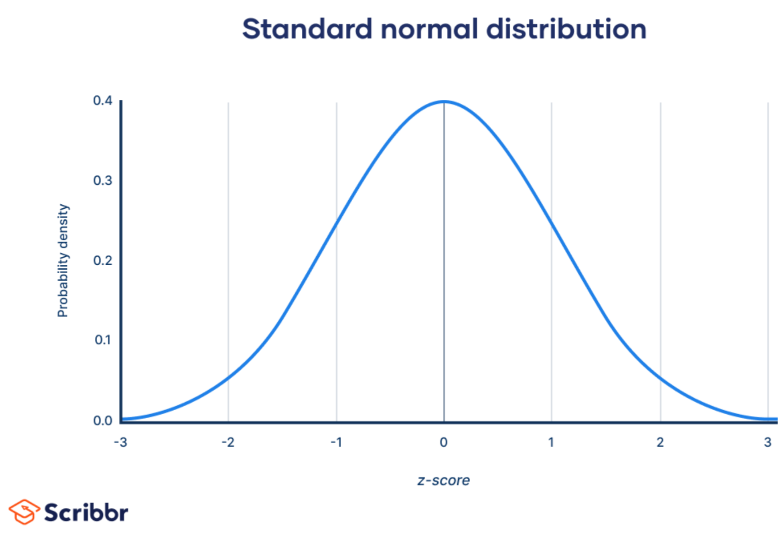
- Motivation for studying statistics and hypothesis testing
- Probability distributions (discrete vs. continuous)
- Computing (theoretical) expected value and variance for discrete distributions
- Sample mean and sample variance
- Central limit theorem (CLT) and application in cases where the mean/variance are known
- Computation and interpretation of Z-scores and p-values
- Null vs. alternative hypotheses; when to reject the null hypothesis; significance level α
- Using randomized trials and permutation testing to obtain more precise p-values
- Idea of a t-test as a way to test differences in means (not details)
- Bootstrap: sampling from our data with replacement (usually keeping n the same)
- How to use bootstrapping to obtain confidence intervals
- Bagging (Bootstrap Aggregation): create a classifier for each bootstrapped training dataset
- Idea of using an ensemble of classifiers (ideally with low bias) to reduce variance
- To test, let each classifier in the ensemble "vote"

Central Limit Theorem

If $X_1, X_2, \ldots, X_n$ are samples from a population with expected value μ and finite variance σ^2, and $\overline{X_n}$ is the sample mean, then

$$Z = \lim_{n\to\infty} \left(\frac{\overline{X_n} - \mu}{\sigma / \sqrt{n}} \right)$$

is a standard normal distribution $N(\overset{\text{mean}}{0},\overset{\text{variance}}{1})$.



p-value

- Probability of observing a result as or more extreme than ours *under the null hypothesis*

- Estimated by:
  - Integrating $pdf = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ based on test statistic
  - N_e/T (T: # trials ran, N_e: # times observed extreme result)

- Usually compare with $\alpha = 0.05$ (significance level)

Bootstrap demo

See video tutorial on Piazza!