CS 260: Foundations of Data Science

Prof. Thao Nguyen

Fall 2026



HAVERFORD
COLLEGE

Materials by Sara Mathieson

Admin

- **Project proposal feedback** given over email
- **Lab 8** due tonight at midnight
- **Lab this week:** midterm 2 review
  - Please attend the 9am lab

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
  - Algorithmic Bias and Disparate Impact
  - 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
  - Algorithmic Bias and Disparate Impact
  - Logistic regression and cross entropy

The bootstrap: Resampling

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

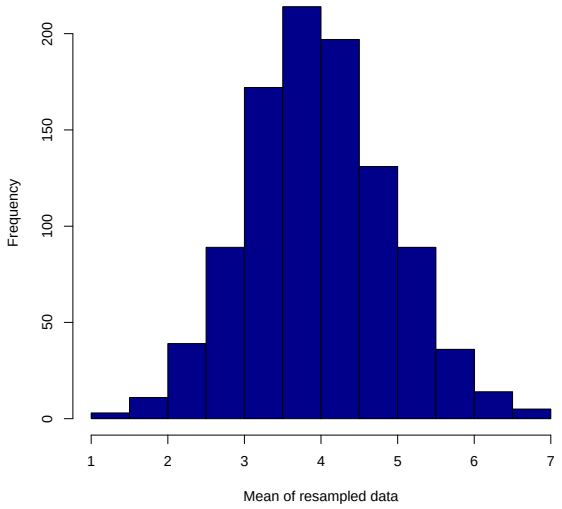
Resample, with replacement, T times

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 6 3	6.2
6 4 6 4 6 4 2 4 3 4	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)"

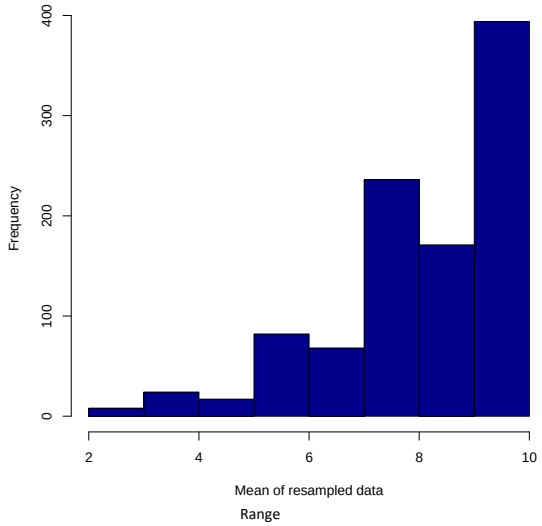
Data: ***X*** = [2, 3, 4, 8, 0, 6, 1, 10, 2, 4]

Compute Range

Resample, with replacement, T times

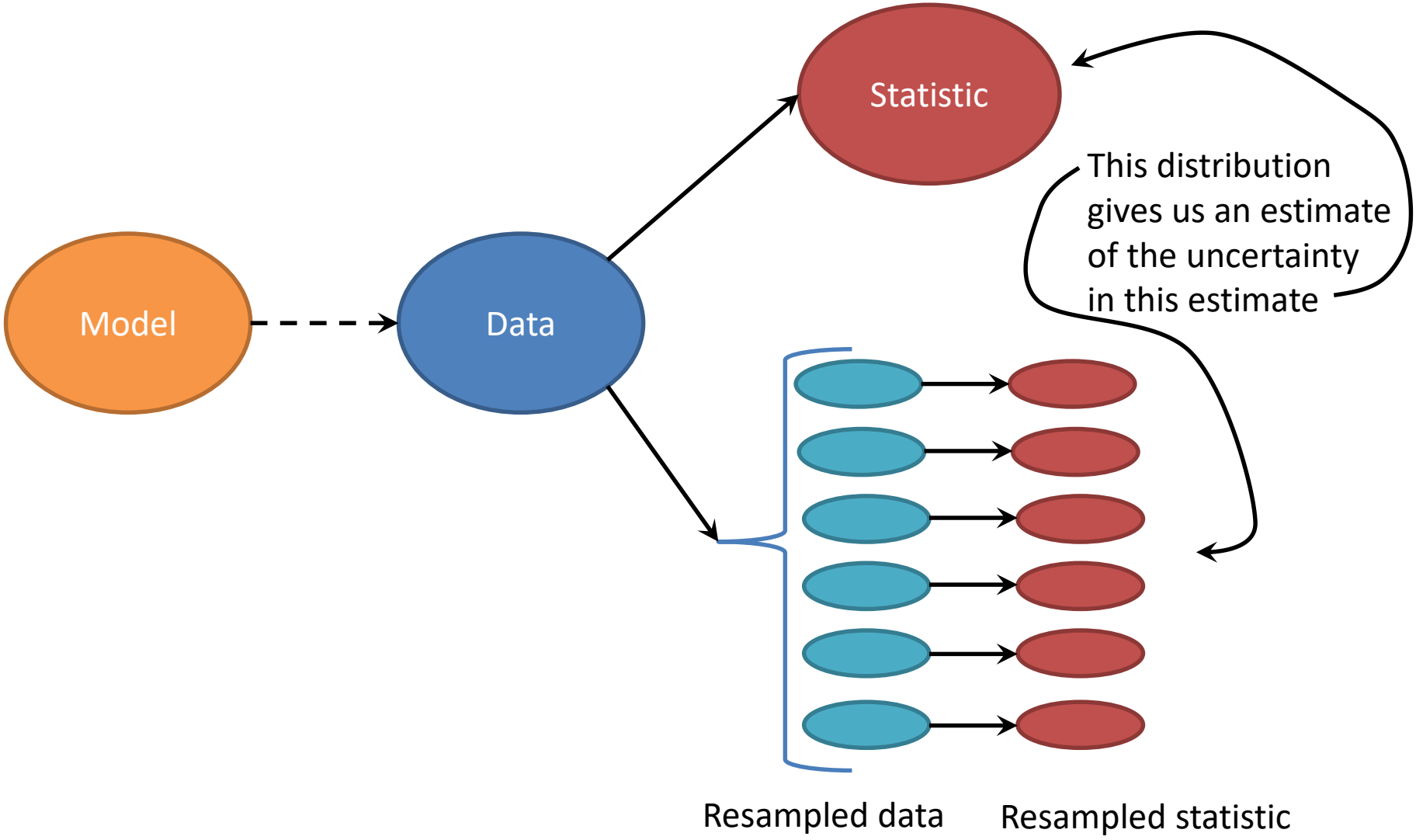
1 8 2 4 6 10 1 1 1 8	9
1 0 1 6 4 1 4 2 1 2	6
8 1 6 2 6 4 2 4 10 2	9
8 3 4 2 10 8 10 8 6 3	8
6 4 6 4 6 4 2 4 3 4	4
…	…
…	…

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



Slide: Iain Mathieson

The bootstrap: Resampling

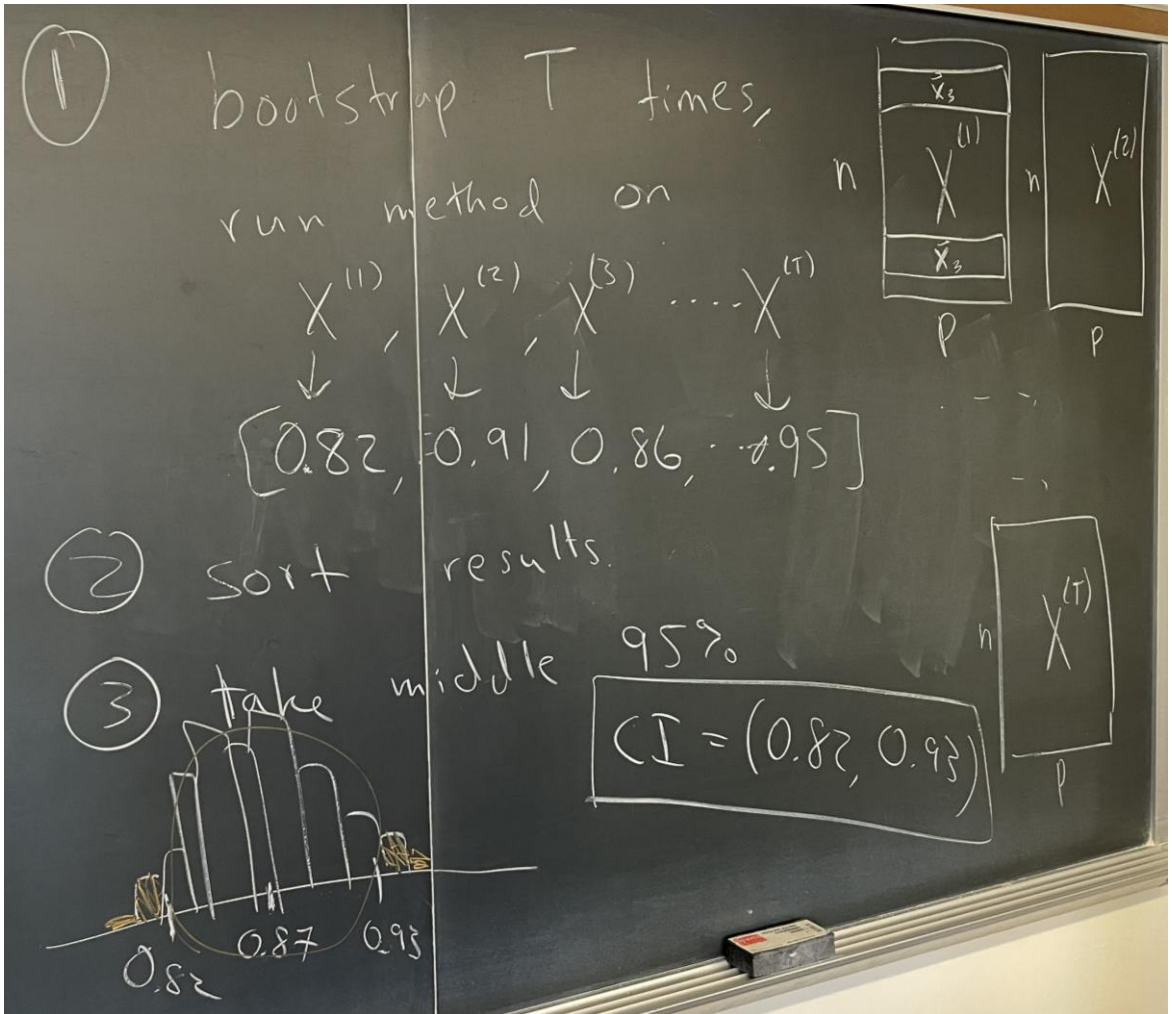


Slide: Iain Mathieson

Bootstrap example

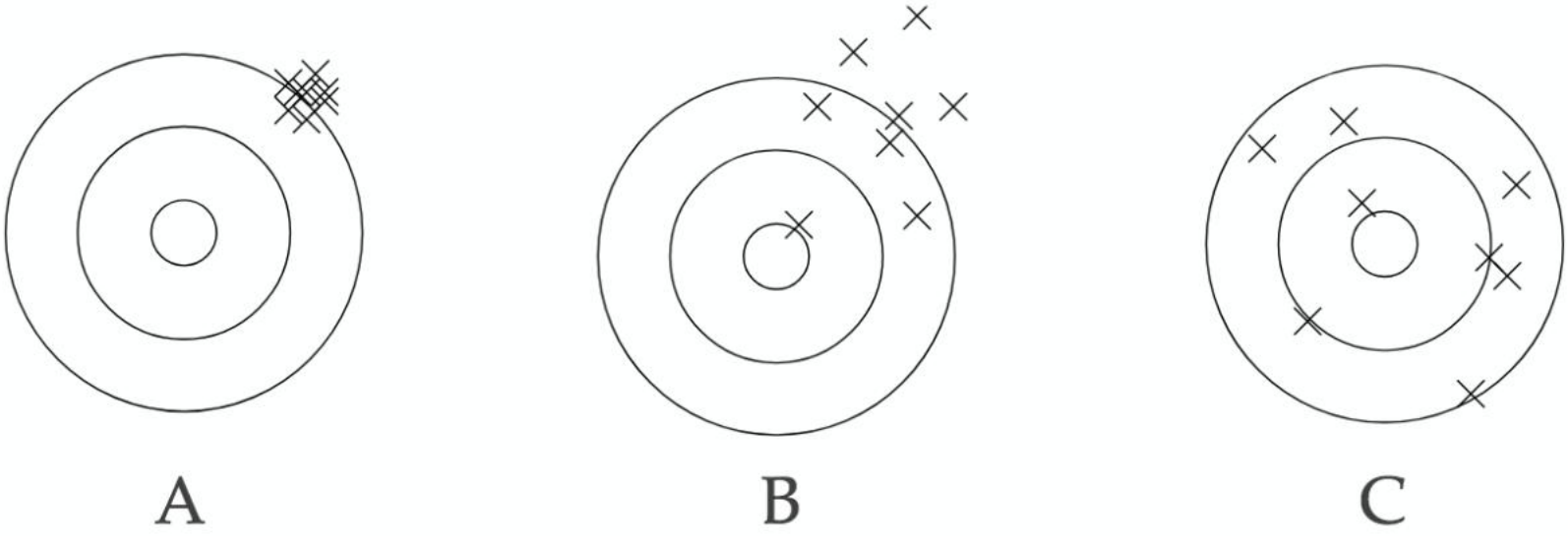
Setup: you obtain 0.87 accuracy on a test dataset using a new algorithm

Goal: find a 95% confidence interval for your estimate



Bagging (Bootstrap Aggregation)

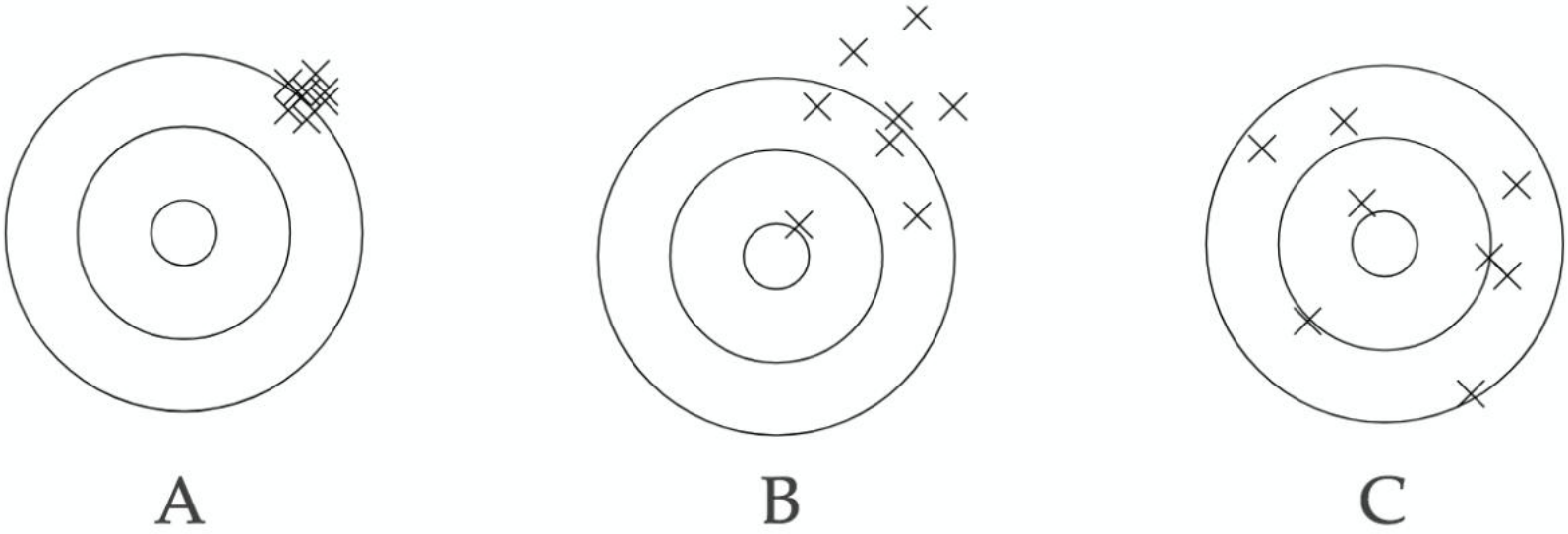
Motivation: bias and variance



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

Example from Ameet Soni

Motivation: bias and variance

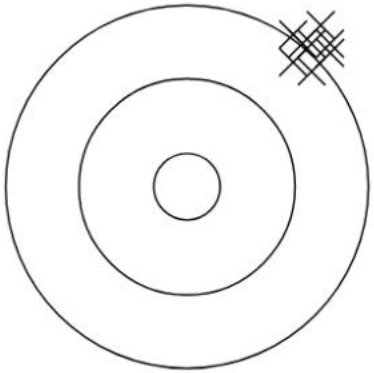


Variance: low
Bias: high

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

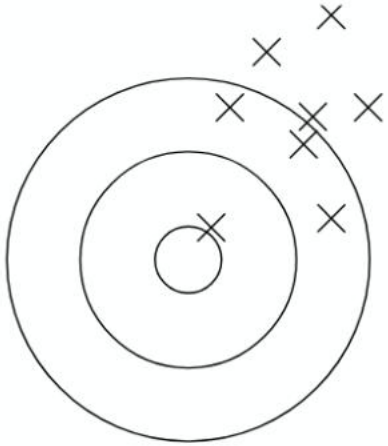
Example from Ameet Soni

Motivation: bias and variance



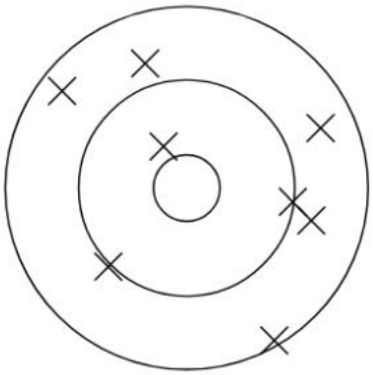
A

Variance: low
Bias: high



B

Variance: high
Bias: high

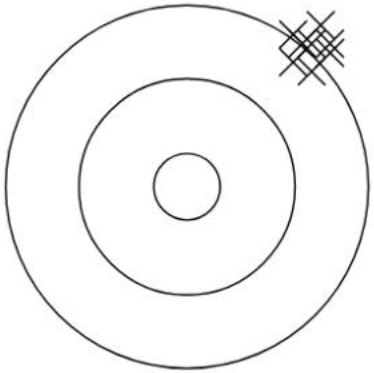


C

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

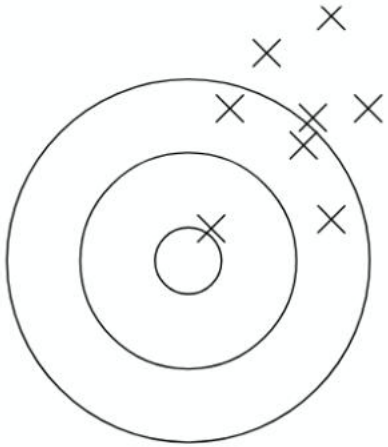
Example from Ameet Soni

Motivation: bias and variance



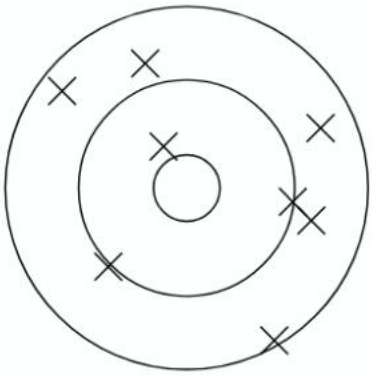
A

Variance: low
Bias: high



B

Variance: high
Bias: high



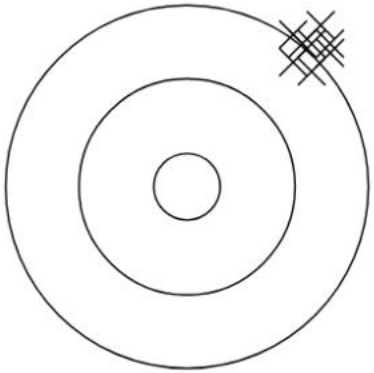
C

Variance: high
Bias: low

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

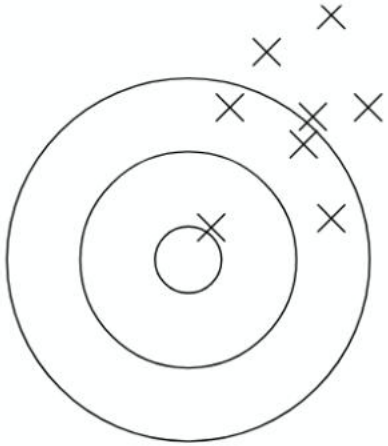
Example from Ameet Soni

Motivation: bias and variance



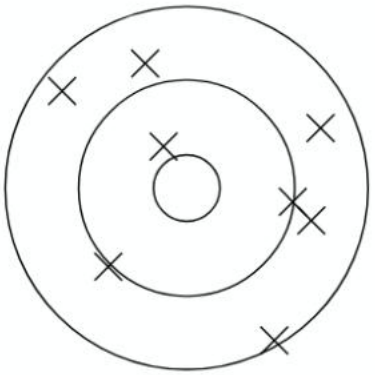
A

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

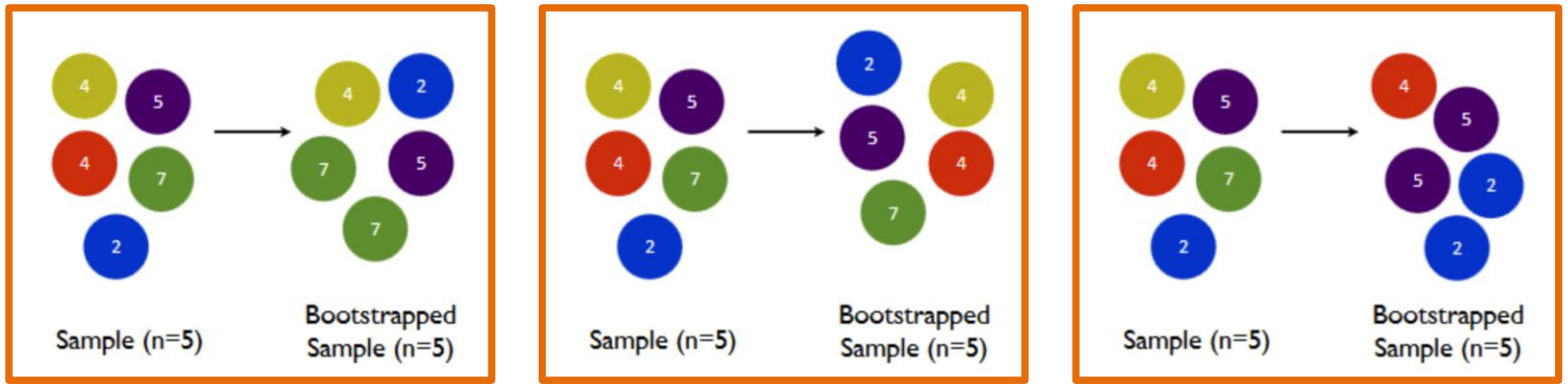
Ensemble Idea

- 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):

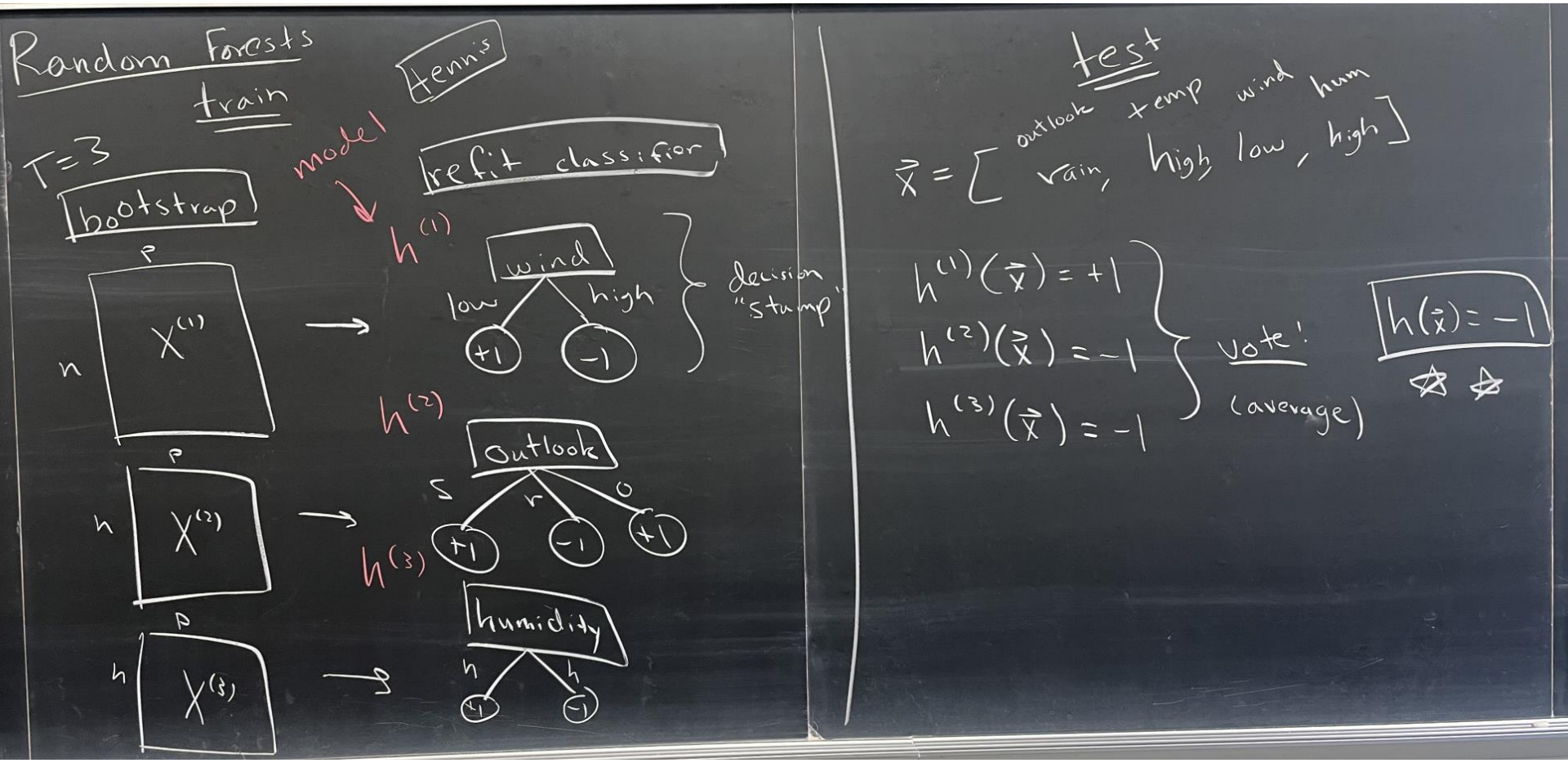
* create bootstrap sample $X^{(t)}$ of size n from training data

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

Test:

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

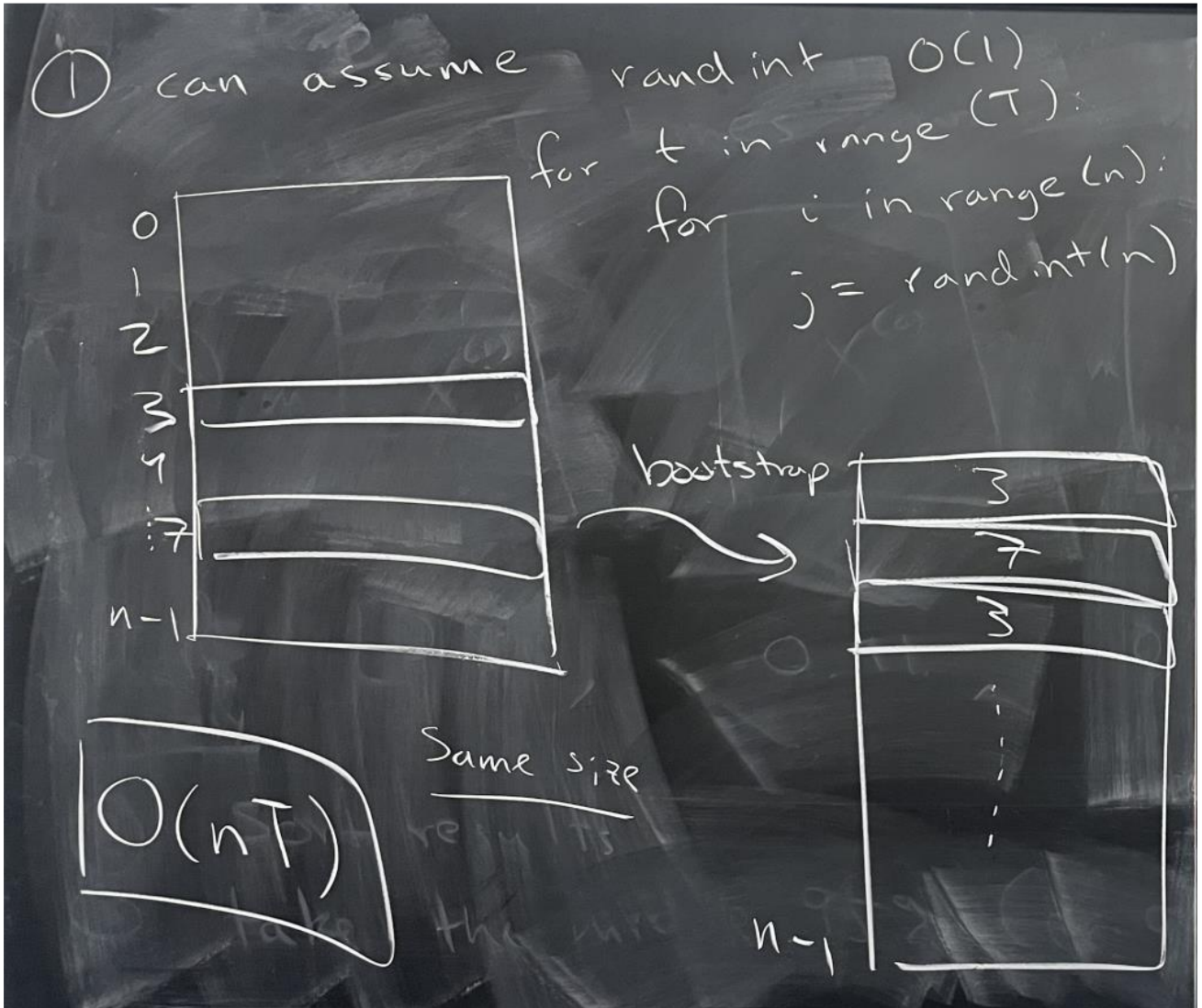
Random Forests



Handout 18

Handout 18

1. Bootstrap



Handout 18

2. Bagging

Unordered

- $n = 2 \Rightarrow \{n_1, n_1\},\ \{n_1, n_2\},\ \{n_2, n_2\}$ 3 sets
- $n = 3$
  - $\{n_1, n_1, n_1\} \Rightarrow$ 3 sets
  - $\{n_1, n_2, n_3\} \Rightarrow$ 1 set
  - $\{n_1, n_1, n_2\} \Rightarrow$ 6 sets

  } 10 sets

Ordered: n^n sets

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
  - Algorithmic Bias and Disparate Impact
  - Logistic regression and cross entropy

Confusion matrix with more classes

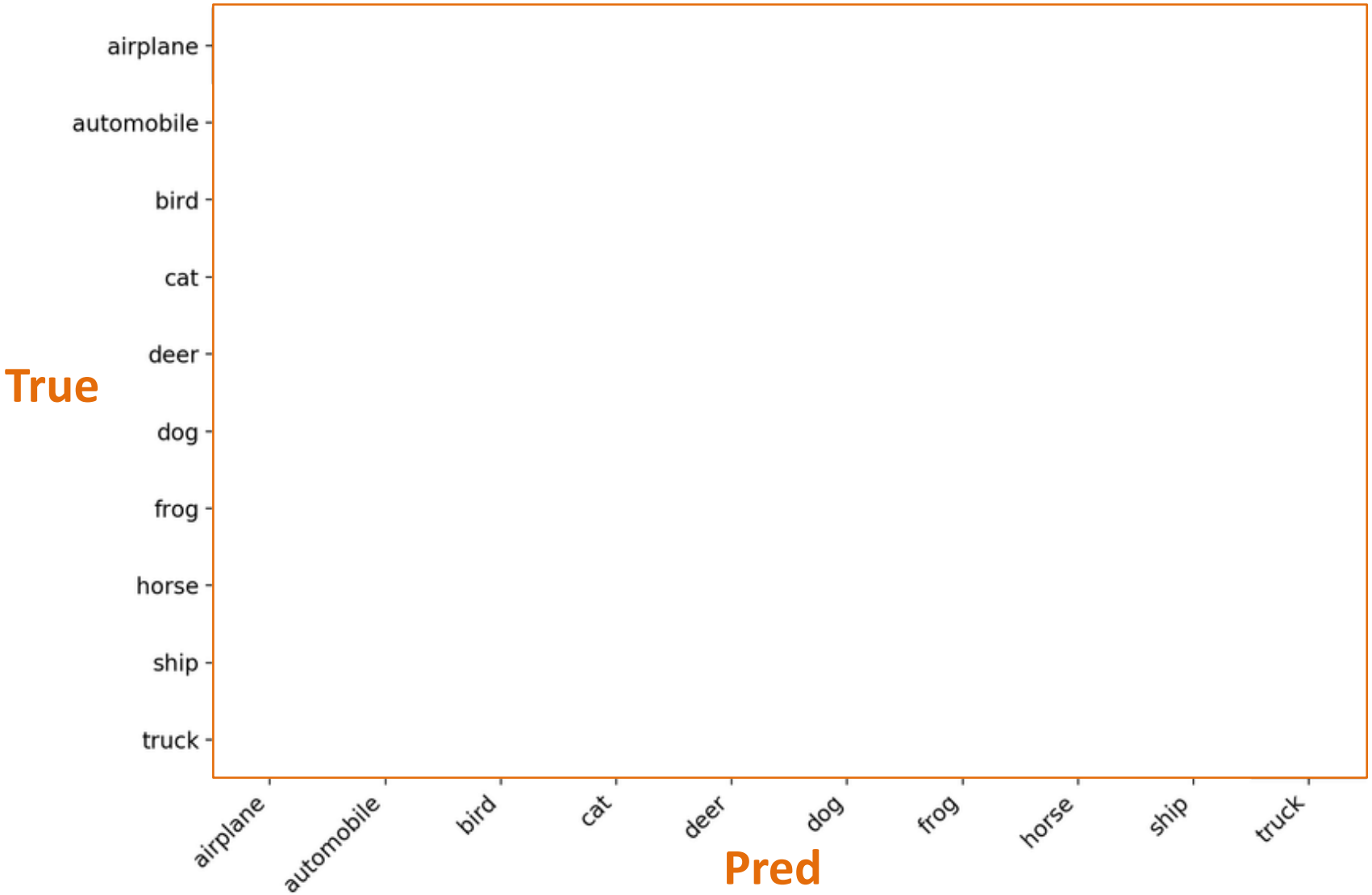


Figure by: Qun Liu (confusion matrix on cifar-10 dataset)

Confusion matrix with more classes

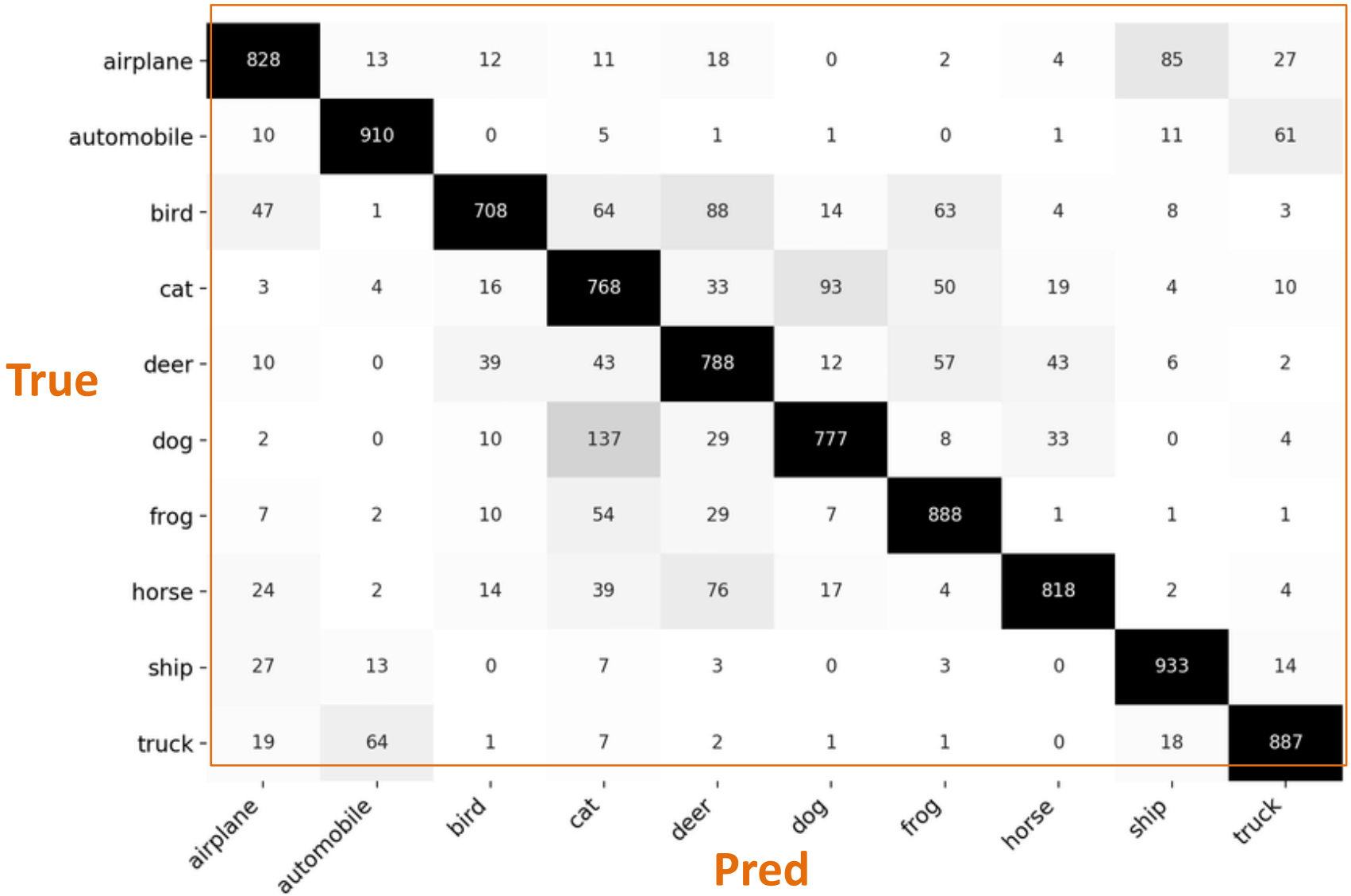


Figure by: Qun Liu (confusion matrix on cifar-10 dataset)

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

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

Outline for today

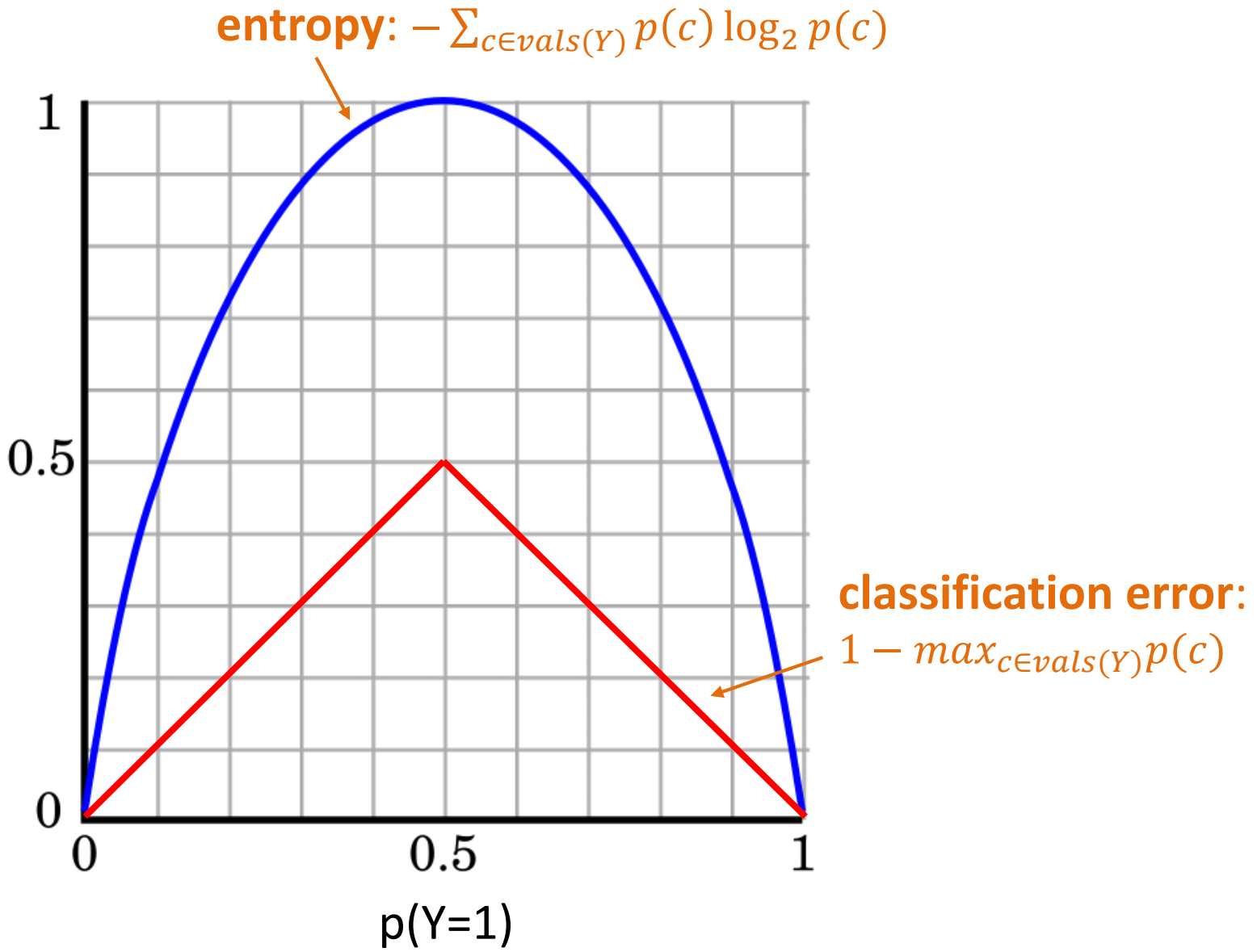
- Bootstrap, Bagging and Random forests
- Midterm 2 Review
  - Revisit confusion matrices
  - Entropy vs. classification error
  - Central Limit Theorem
  - PCA (linear transformation + interpretation)
  - Naïve Bayes
  - Algorithmic Bias and Disparate Impact
  - Logistic regression and cross entropy

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 13)

Entropy vs. classification error



Splitting nodes based on entropy

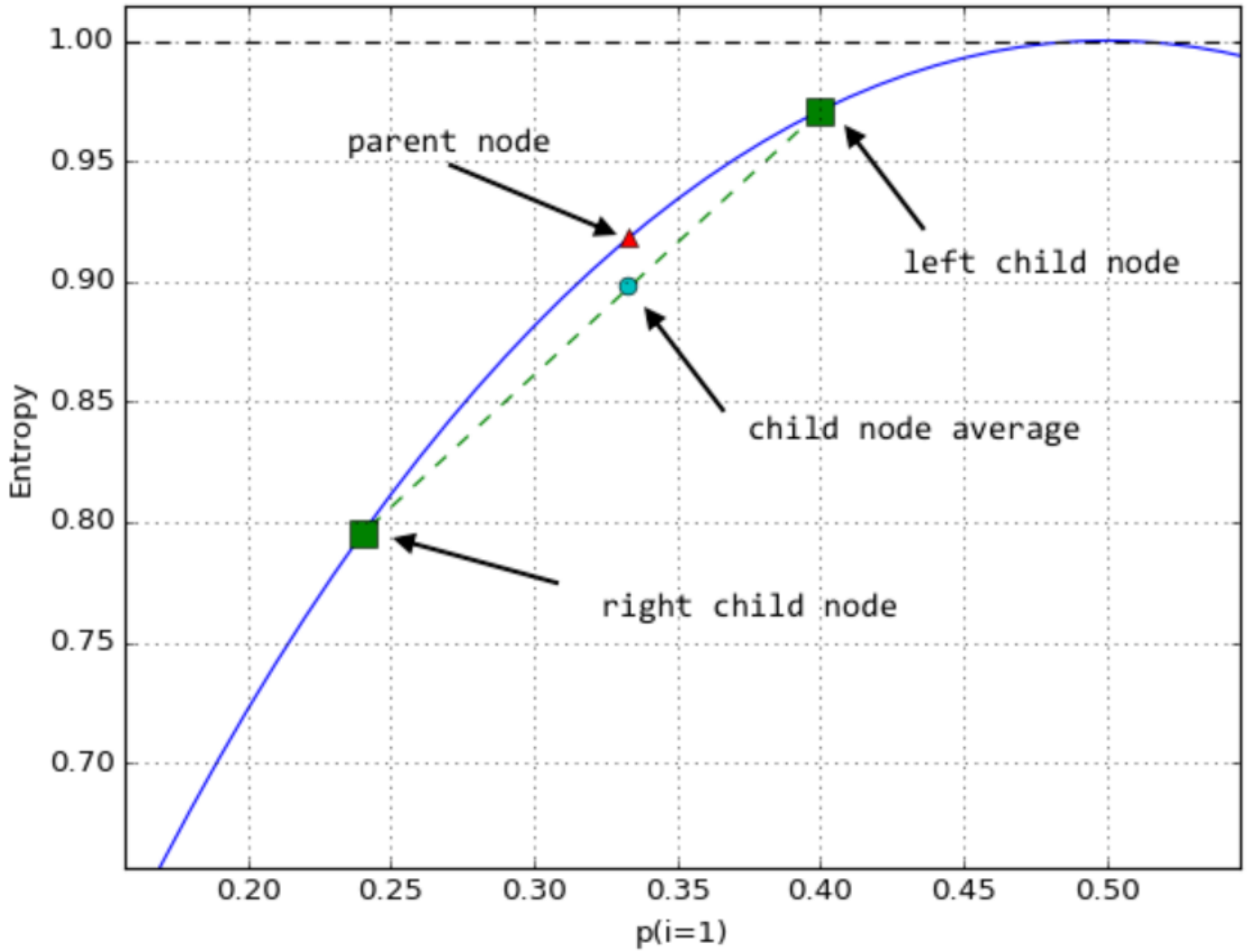
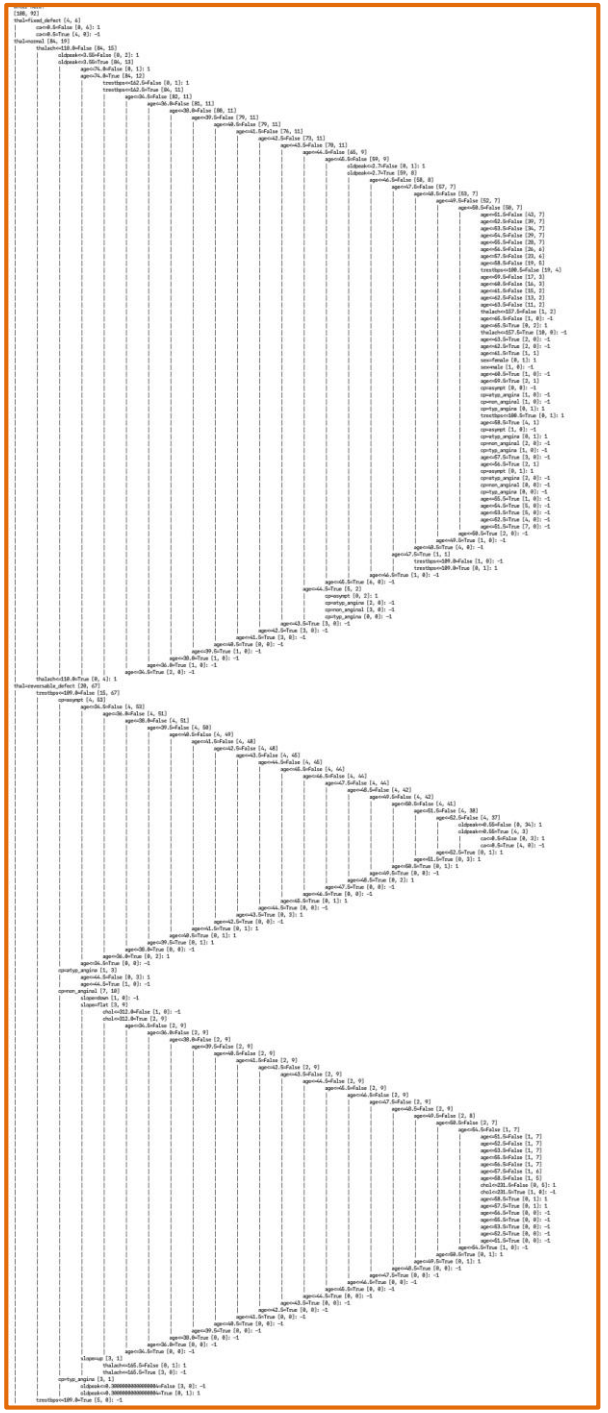


Figure by Sebastian Raschka

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
  - Algorithmic Bias and Disparate Impact
  - 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"

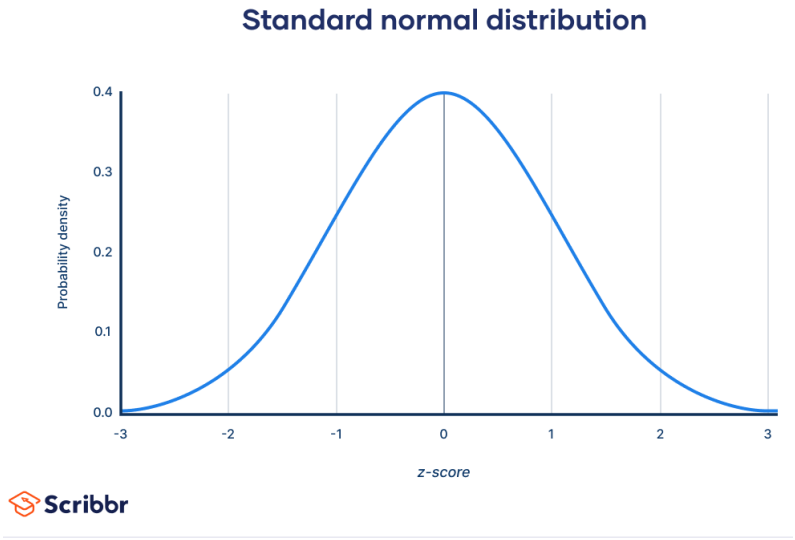
Central Limit Theorem

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

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

mean variance

is a standard normal distribution $N(0,1)$.



p-value

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

- Estimated by:
  - Integrating pdf based on test statistic
  - N_e/T (T: # trials ran, N_e: # times observed extreme result)

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

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
  - Algorithmic Bias and Disparate Impact
  - 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

Principal Component Analysis (PCA)

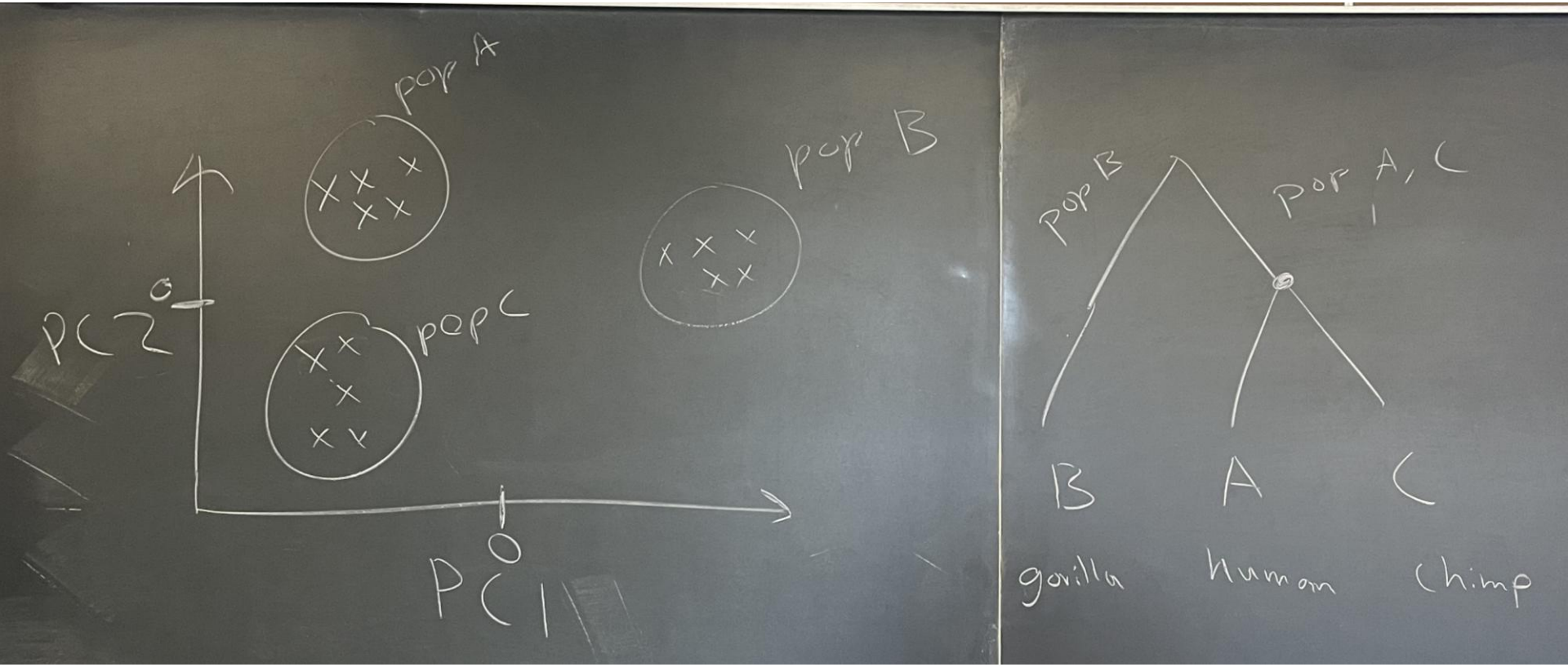
- Transforms *p*-dimensional data so that the new first dimension explains as much of the variation as possible, the new second explains as much of the remaining variation as possible, and so on

- PCA is a linear transformation

- Typically, we look at the first few dimensions of the transformed data as a means of dimensionality reduction and visualization

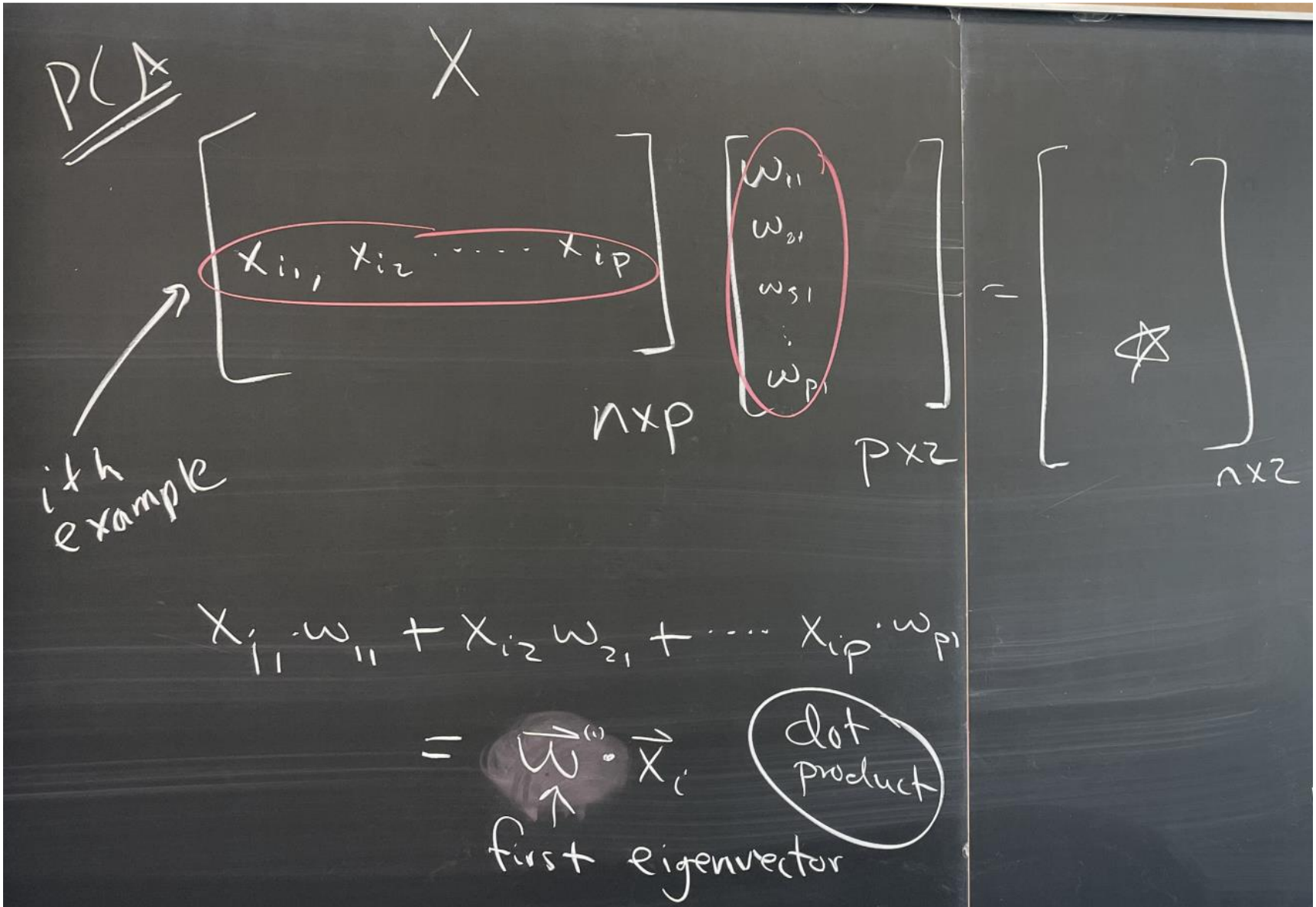
- PCA is often used for:
  - Data visualization
  - Infer qualitative relationships between groups

PCA “classic” genetics example

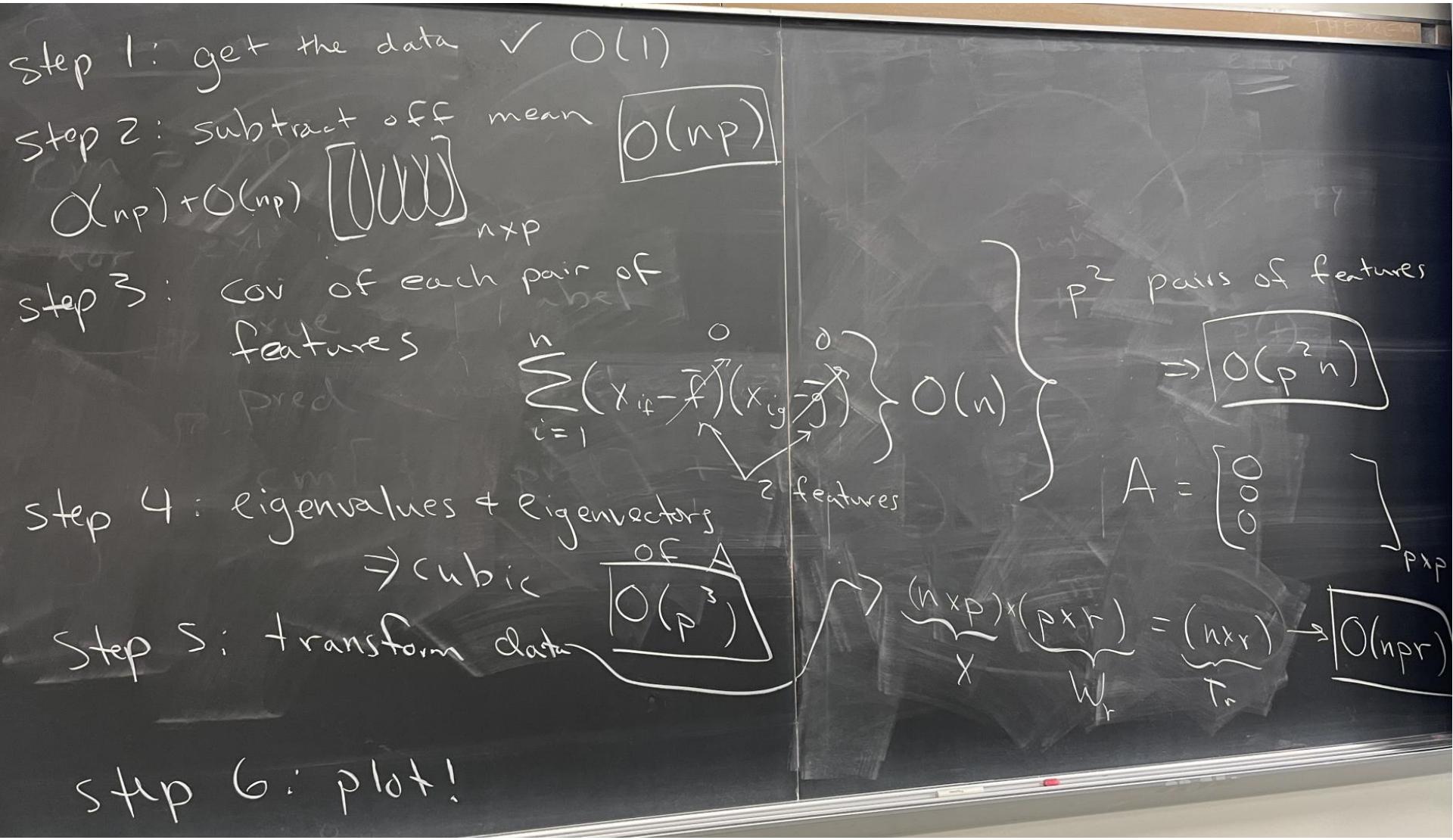


Handout 18

Handout 18


PCA
X
$x_{i1}, x_{i2} \cdots x_{ip}$
ith example
n x p
w_{11} w_{21} w_{31} w_{p1}
p x 2
n x 2
$x_{i1} \cdot w_{11} + x_{i2} w_{21} + \cdots x_{ip} \cdot w_{p1}$
$= \vec{w}^{(1)} \cdot \vec{x}_i$
dot product
first eigenvector

Handout 18


step 1: get the data ✓ O(1)
step 2: subtract off mean
O(np) + O(np)
O(np)
n × p
step 3: cov of each pair of features
$\sum_{i=1}^{n}(x_{if}-\bar{f})(x_{ig}-\bar{g})$
O(n)
p^2 pairs of features
$\Rightarrow O(p^2 n)$
2 features
step 4: eigenvalues & eigenvectors of A
$\Rightarrow$ cubic
$O(p^3)$
$A = \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix}_{p \times p}$
step 5: transform data
$(n \times p) \times (p \times r) = (n \times r)$
X
W_r
T_r
$\rightarrow O(npr)$
step 6: plot!

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
  - Algorithmic Bias and Disparate Impact
  - Logistic regression and cross entropy

From the study guide

1. Probability and Bayesian Models

   - Probability basics including joint probability, conditional probability, Bayes rule
   - Other terms: marginalization, independence, conditional independence
   - Bayesian models: posterior, prior, likelihood, evidence
   - Examples of when you might use a Bayesian model (e.g., email spam, trisomy detection)
   - Idea of using marginalization to compute the evidence (see Handout 10)

2. Naive Bayes

   - Derivation of the Naive Bayes model for $p(y = k|\vec{x})$ (via the Naive Bayes assumption)
   - How do we estimate the probabilities of a Naive Bayes model?
   - Laplace counts (motivation, application details)
   - How can we predict the label of a new example after fitting a Naive Bayes model?
   - What types of features/label do we currently require for Naive Bayes?
   - How Naive Bayes can be implemented using dictionaries in Python

Bayes' Theorem

- P(A,B) = P(A|B)P(B)
- P(A,B) = P(B|A)P(A)

always true!

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Independence

- Independence: $P(A,B) = P(A)P(B)$

not always true!

- Conditional independence: $P(A|B,C) = P(A|C)$

Naïve Bayes assumption

Naïve Bayes

Model

$$p(y = k|\vec{x}) \propto p(y = k) \prod_{j=1}^{p} p(x_j|y = k)$$

Classification: $\hat{y} = argmax_{k=1,\ldots,K}\, p(y = k|\vec{x})$

Estimating $p(y = k)$ & $p(x_j|y = k)$

of examples with label k

- $\theta_k = \frac{N_k+1}{n+K}$

# of classes for y

# of examples with feature j = value v and class label k

- $\theta_{k,j,v} = \frac{N_{k,j,v}+1}{N_k+|f_j|}$

# of feature values for feature j

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
  - Algorithmic Bias and Disparate Impact
  - Logistic regression and cross entropy

From the study guide

3. Algorithmic Bias and Disparate Impact

- Sample size disparity and how it can impact results
- May need different models for different groups, so a single model is not possible
- General idea that training on past data will recapitulate historical biases
- Problem setup/notation for redundant encoding of features (X, Y, C)
- Definitions of: direct vs. indirect discrimination, disparate impact
- Idea of training a classifier to predict X (protected) from Y to detect disparate impact

How can we tell if an algorithm is biased?

D: dataset with attributes X, Y

* X is protected
* Y is unprotected (other features)

Goal: determine outcome C (hired, admitted, etc)

Direct discrimination: C = f(X)

* Female instrumentalist not hired for orchestra
* Some ethnic groups not allowed to eat at a restaurant

"Certifying and Removing Disparate Impact" Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, Suresh Venkatasubramanian

How can we tell if an algorithm is biased?

D: dataset with attributes X, Y

* X is protected
* Y is unprotected (other features)

Goal: determine outcome C (hired, admitted, etc)

Indirect discrimination: C = f(Y)

* but strong correlation between X and Y
* Ex: housing loans
* Ex: programming experience

"Certifying and Removing Disparate Impact" Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, Suresh Venkatasubramanian

Disparate Impact

features

- X: protected attributes
- Y: other attributes
- C: binary outcome $\in \{0,1\}$

0 minority group
1 majority group

not hired hired

Legal definition

If $P(C = 1|X = 0) < 0.8 * P(C = 1|X = 1)$

$\Rightarrow$ disparate impact

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
  - Algorithmic Bias and Disparate Impact
  - Logistic regression and cross entropy

From the study guide

5. Logistic Regression

- Motivation for logistic regression; our model is a logistic function that takes in $\vec{w} \cdot \vec{x}$
- Logistic regression creates a *linear* decision boundary (compute/visualize for $p = 1$)
- In logistic regression our cost is the negative log likelihood (don't need to derive)
- Intuition/visualization of the cost function (and relationship to cross entropy)
- Stochastic gradient descent (SGD) for logistic regression, relationship to linear regression
- Interpretation of the weights as feature importance

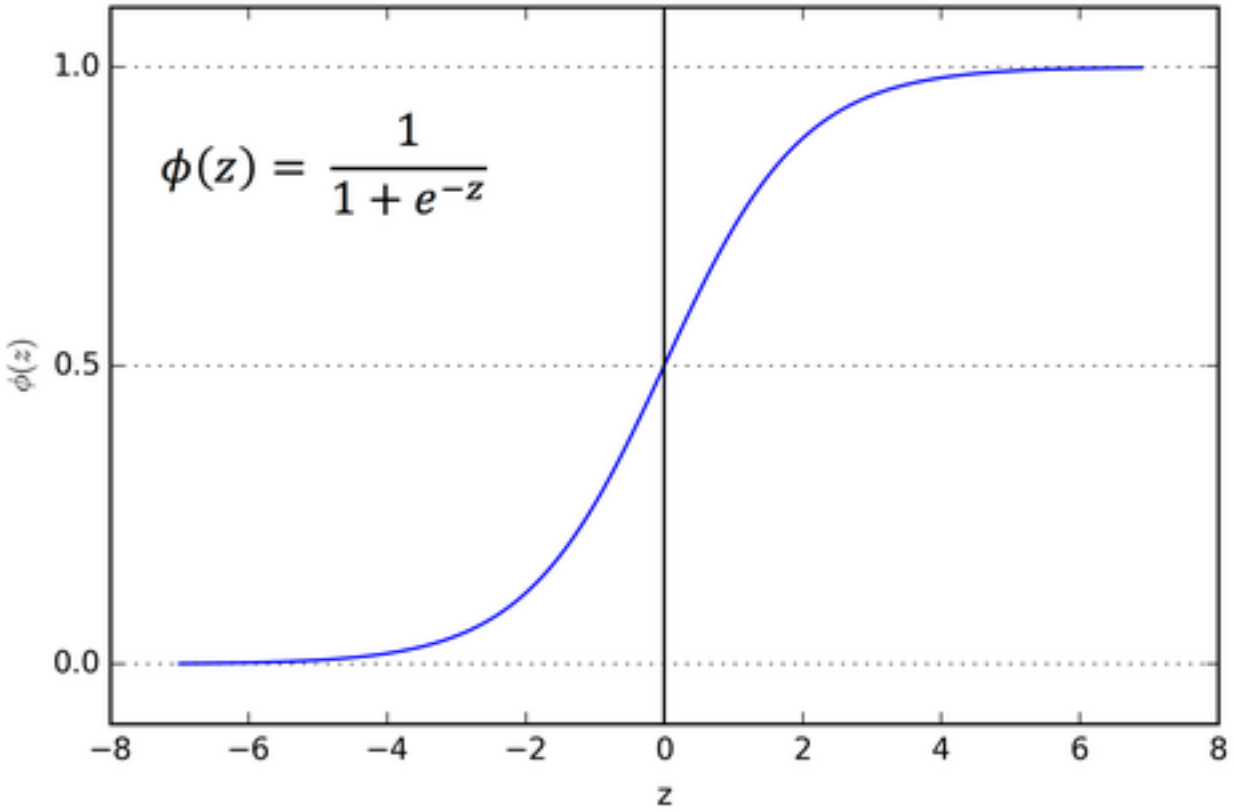
3 important pieces to SGD

- Hypothesis function (prediction)

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = p(y = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w}\cdot\boldsymbol{x}}}$$

Logistic (sigmoid) function

Transforms a continuous real number into a range of (0, 1)



Logistic Regression

- Binary classification $y \in \{0,1\}$
- Model will be

$$h_{\vec{w}}(\vec{x}) = p(y = 1|\vec{x}) = \frac{1}{1 + e^{-\vec{w}\cdot\vec{x}}}$$

3 important pieces to SGD

- Hypothesis function (prediction)

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = p(y = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w}\cdot\boldsymbol{x}}}$$

- Cost function (want to minimize)

$$J(\boldsymbol{w}) = -\sum_{i=1}^{n} y_i \log h_{\boldsymbol{w}}(\boldsymbol{x_i}) + (1 - y_i) \log(1 - h_{\boldsymbol{w}}(\boldsymbol{x_i}))$$

Stochastic Gradient Descent for Logistic Regression (binary classification)

set $\vec{w} = \vec{0}$

while cost $J(\vec{w})$ is still changing:

shuffle data points

for i = 1,…,n:

$$\vec{w} \leftarrow \vec{w} - \alpha \underbrace{\nabla J_{\vec{x_i}}(\vec{w})}$$

derivative of $J(\vec{w})$ wrt x_i

store $J(\vec{w})$

3 important pieces to SGD

- Hypothesis function (prediction)

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = p(y = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w}\cdot\boldsymbol{x}}}$$

- Cost function (want to minimize)

$$J(\boldsymbol{w}) = -\sum_{i=1}^{n} y_i \log h_{\boldsymbol{w}}(\boldsymbol{x_i}) + (1 - y_i) \log(1 - h_{\boldsymbol{w}}(\boldsymbol{x_i}))$$

- Gradient of cost wrt single data point x_i

$$\nabla J_{\boldsymbol{x}_i}(\boldsymbol{w}) = (h_{\boldsymbol{w}}(\boldsymbol{x_i}) - y_i)\boldsymbol{x_i}$$

For each method/approach, is X continuous or discrete? What about y?

- Linear regression
- Polynomial regression
- Decision trees/stumps
- ROC curve as an evaluation metric
- Naïve Bayes
- Logistic regression
- Entropy and information gain
- PCA

Think about offline!