CS 260: Foundations of Data Science

Prof. Sara Mathieson

Fall 2023

HAVERFORD COLLEGE

Write one midterm question or topic on your notecard

Will answer later in class today or on Piazza

Admin

- **Exam** due in class on Tuesday
  - Do not open the exam until you’re ready to start
  - Time limit: **3 hours**
  - Resources: hand-written study sheet, calculator

- **First candidate talk on Monday!**
  - 4pm tea
  - 4:15pm talk (H109)

Outline for November 16

- Midterm 2 Review
  - Entropy vs. classification error
  - PCA
  - Naïve Bayes
  - Central Limit Theorem
  - Logistic regression and cross entropy

Outline for November 16

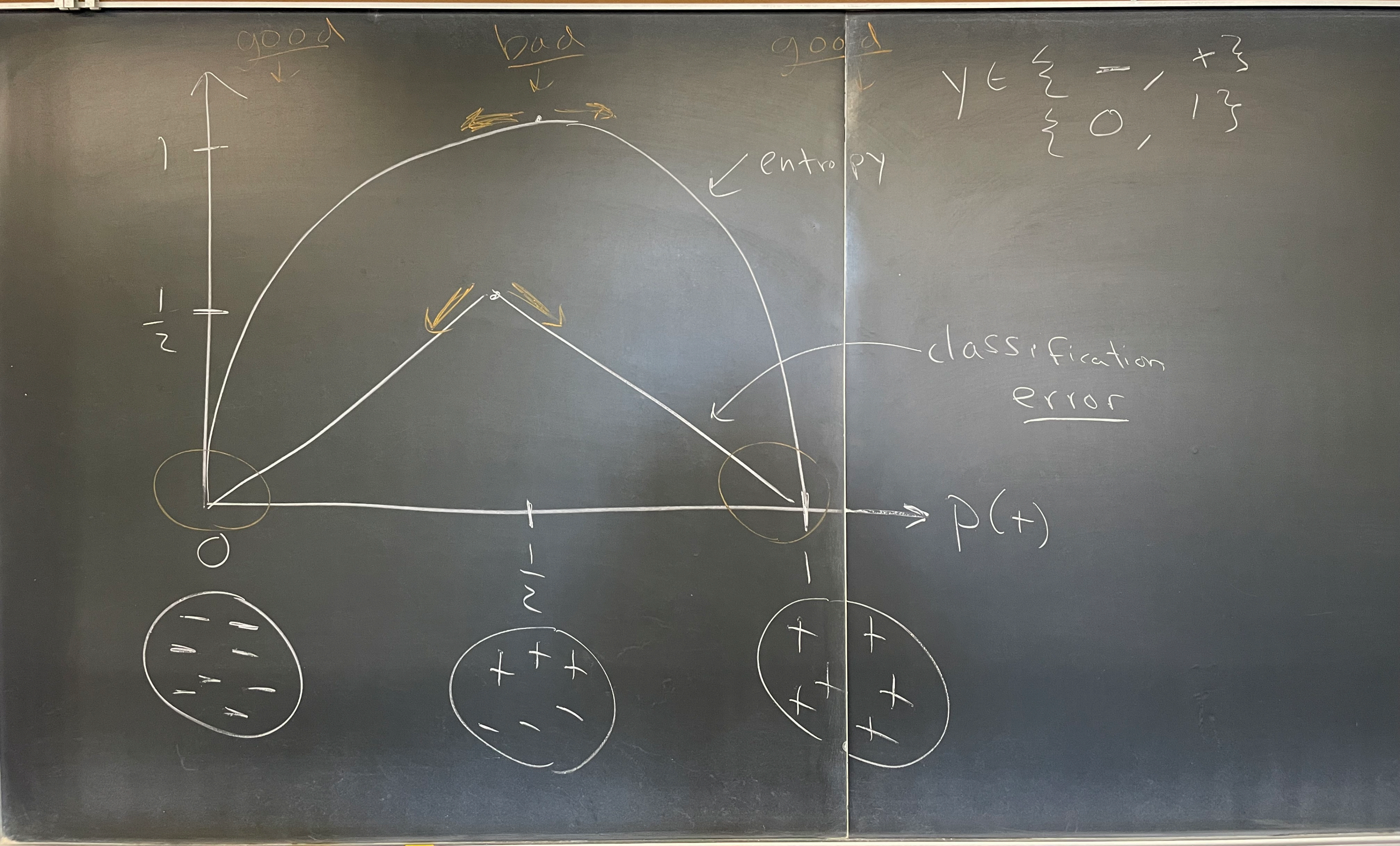
- Midterm 2 Review
  - Entropy vs. classification error
  - PCA
  - Naïve Bayes
  - Central Limit Theorem
  - Logistic regression and cross entropy

From the study guide

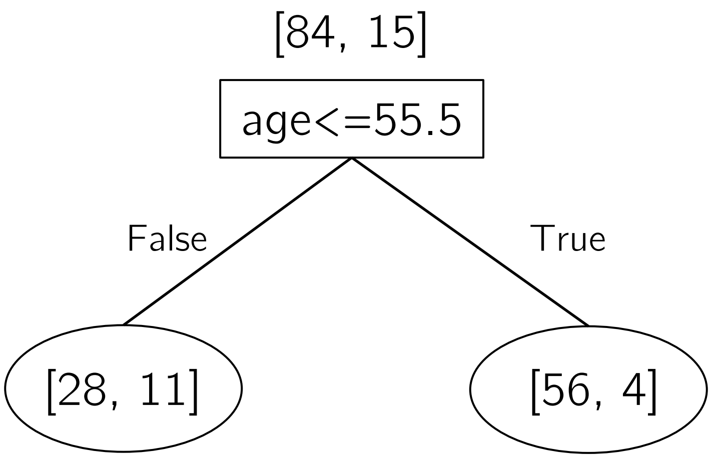
4. Information Theory

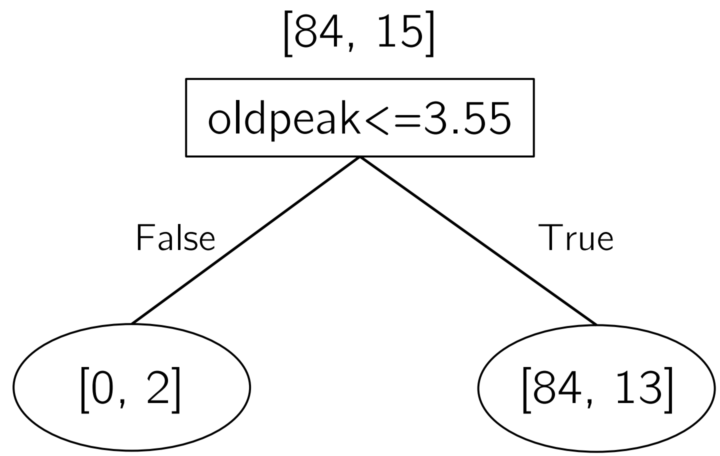
- Conceptual idea of entropy as well as formal definition
- Shannon encoding (and decoding), plus how to use entropy to compute average number of bits needed to send one piece of information
- Use of conditional entropy and information gain to choose best features
- Comparison with classification accuracy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

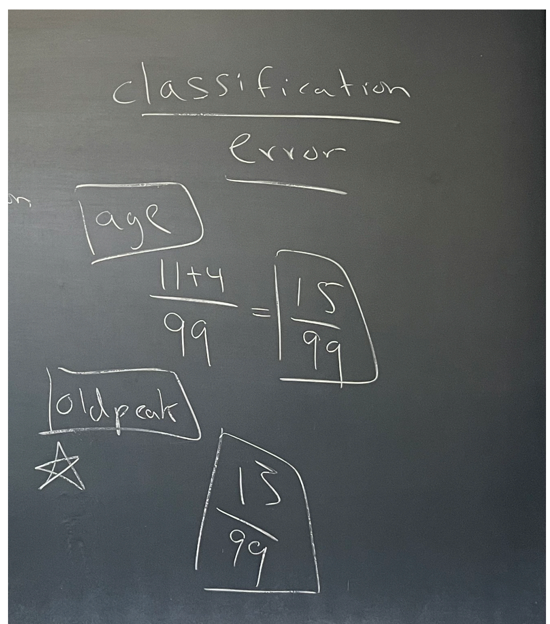
Entropy vs. classification error



One feature models (decision stumps): information gain vs. classification error







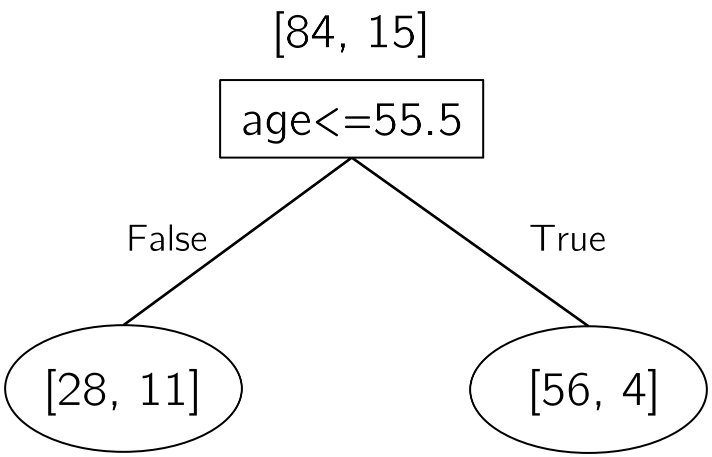
$$H(Y) = -\sum_{c \in vals(y)} p(y=c)\log_2 p(y=c) \qquad y \in \{-1, +1\}$$

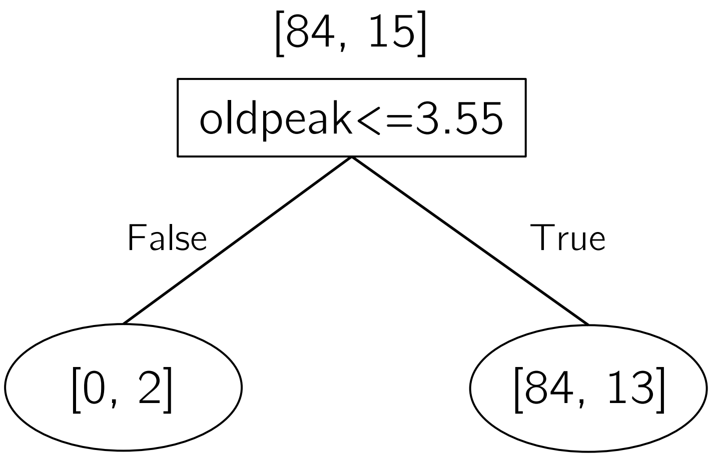
$$H(Y) = -\left(\frac{84}{99}\log_2\frac{84}{99} + \frac{15}{99}\log_2\frac{15}{99}\right) = 0.61$$

$$H(Y|oldpeak) = \frac{2}{99}\cancel{H(Y|oldpeak=F)}^{\to 0} + \frac{97}{99}H(Y|oldpeak=T)$$

$$H(Y|oldpeak=T) = -\left(\frac{84}{97}\log\frac{84}{97} + \frac{13}{97}\log\frac{13}{97}\right)$$

One feature models (decision stumps): information gain vs. classification error





$H(Y) = 0.6136190195993708$

$H(Y|age<=55.5) = 0.5522480910534322$

$H(Y|oldpeak<=3.55) = 0.5568804630596093$

=> Age feature produces more information gain!

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
```

```
[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]
|   |   oldpeak<=3.55=False [0, 2]: 1
|   |   oldpeak<=3.55=True [84, 13]
|   |   |   age<=74.0=False [0, 1]: 1
|   |   |   age<=74.0=True [84, 12]
|   |   |   |   trestbps<=162.5=False [0, 1]: 1
|   |   |   |   trestbps<=162.5=True [84, 11]
|   |   |   |   |   age<=34.5=False [82, 11]
|   |   |   |   |   |   age<=36.0=False [81, 11]
|   |   |   |   |   |   |   age<=38.0=False [80, 11]
|   |   |   |   |   |   |   |   age<=39.5=False [79, 11]
|   |   |   |   |   |   |   |   |   age<=40.5=False [79, 11]
|   |   |   |   |   |   |   |   |   |   age<=41.5=False [76, 11]
|   |   |   |   |   |   |   |   |   |   |   age<=42.5=False [73, 11]
|   |   |   |   |   |   |   |   |   |   |   |   age<=43.5=False [70, 11]
|   |   |   |   |   |   |   |   |   |   |   |   |   age<=44.5=False [65, 9]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=45.5=False [59, 9]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   oldpeak<=2.7=False [0, 1]: 1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   oldpeak<=2.7=True [59, 8]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=46.5=False [58, 8]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=47.5=False [57, 7]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=48.5=False [53, 7]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=49.5=False [52, 7]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=50.5=False [50, 7]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=51.5=False [43, 7]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=52.5=False [39, 7]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=53.5=False [34, 7]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=54.5=False [29, 7]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=55.5=False [28, 7]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=56.5=False [26, 6]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=57.5=False [23, 6]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=58.5=False [19, 5]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   trestbps<=100.5=False [19, 4]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=59.5=False [17, 3]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=60.5=False [16, 3]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=61.5=False [15, 2]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=62.5=False [13, 2]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=63.5=False [11, 2]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   thalach<=157.5=False [1, 2]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=65.5=False [1, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=65.5=True [0, 2]: 1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   thalach<=157.5=True [10, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=63.5=True [2, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=62.5=True [2, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=61.5=True [1, 1]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   sex=female [0, 1]: 1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   sex=male [1, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=60.5=True [1, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=59.5=True [2, 1]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   cp=asympt [0, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   cp=atyp_angina [1, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   cp=non_anginal [1, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   cp=typ_angina [0, 1]: 1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   trestbps<=100.5=True [0, 1]: 1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=58.5=True [4, 1]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   cp=asympt [1, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   cp=atyp_angina [0, 1]: 1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   cp=non_anginal [2, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   cp=typ_angina [1, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=57.5=True [3, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=56.5=True [2, 1]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   cp=asympt [0, 1]: 1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   cp=atyp_angina [2, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   cp=non_anginal [0, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   cp=typ_angina [0, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=55.5=True [1, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=54.5=True [5, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=53.5=True [5, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=52.5=True [4, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=51.5=True [7, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=50.5=True [2, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=49.5=True [1, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=48.5=True [4, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=47.5=True [1, 1]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   trestbps<=109.0=False [1, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   trestbps<=109.0=True [0, 1]: 1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=46.5=True [1, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=45.5=True [6, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   age<=44.5=True [5, 2]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   cp=asympt [0, 2]: 1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   cp=atyp_angina [2, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   cp=non_anginal [3, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   cp=typ_angina [0, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   age<=43.5=True [3, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   age<=42.5=True [3, 0]: -1
|   |   |   |   |   |   |   |   |   |   age<=41.5=True [3, 0]: -1
|   |   |   |   |   |   |   |   |   age<=40.5=True [0, 0]: -1
|   |   |   |   |   |   |   |   age<=39.5=True [1, 0]: -1
|   |   |   |   |   |   |   age<=38.0=True [1, 0]: -1
|   |   |   |   |   |   age<=36.0=True [1, 0]: -1
|   |   |   |   |   age<=34.5=True [2, 0]: -1
|   thalach<=110.0=True [0, 4]: 1
thal=reversable_defect [20, 67]
|   trestbps<=109.0=False [15, 67]
|   |   cp=asympt [4, 53]
|   |   |   age<=34.5=False [4, 53]
|   |   |   |   age<=36.0=False [4, 51]
|   |   |   |   |   age<=38.0=False [4, 51]
|   |   |   |   |   |   age<=39.5=False [4, 50]
|   |   |   |   |   |   |   age<=40.5=False [4, 49]
|   |   |   |   |   |   |   |   age<=41.5=False [4, 48]
|   |   |   |   |   |   |   |   |   age<=42.5=False [4, 48]
|   |   |   |   |   |   |   |   |   |   age<=43.5=False [4, 46]
|   |   |   |   |   |   |   |   |   |   |   age<=44.5=False [4, 46]
|   |   |   |   |   |   |   |   |   |   |   |   age<=45.5=False [4, 44]
|   |   |   |   |   |   |   |   |   |   |   |   |   age<=46.5=False [4, 44]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=47.5=False [4, 44]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=48.5=False [4, 42]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=49.5=False [4, 42]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=50.5=False [4, 41]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=51.5=False [4, 38]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=52.5=False [4, 37]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   oldpeak<=0.55=False [0, 34]: 1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   oldpeak<=0.55=True [4, 3]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   ca<=0.5=False [0, 3]: 1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   ca<=0.5=True [4, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=52.5=True [0, 1]: 1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=51.5=True [0, 3]: 1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=50.5=True [0, 1]: 1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=49.5=True [0, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=48.5=True [0, 2]: 1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=47.5=True [0, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   age<=46.5=True [0, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   age<=45.5=True [0, 1]: 1
|   |   |   |   |   |   |   |   |   |   |   age<=44.5=True [0, 0]: -1
|   |   |   |   |   |   |   |   |   |   age<=43.5=True [0, 3]: 1
|   |   |   |   |   |   |   |   |   age<=42.5=True [0, 0]: -1
|   |   |   |   |   |   |   |   age<=41.5=True [0, 1]: 1
|   |   |   |   |   |   |   age<=40.5=True [0, 1]: 1
|   |   |   |   |   |   age<=39.5=True [0, 1]: 1
|   |   |   |   |   age<=38.0=True [0, 0]: -1
|   |   |   |   age<=36.0=True [0, 2]: 1
|   |   |   age<=34.5=True [0, 0]: -1
|   |   cp=atyp_angina [1, 3]
|   |   |   age<=44.5=False [0, 3]: 1
|   |   |   age<=44.5=True [1, 0]: -1
|   |   cp=non_anginal [7, 10]
|   |   |   slope=down [1, 0]: -1
|   |   |   slope=flat [3, 9]
|   |   |   |   chol<=312.0=False [1, 0]: -1
|   |   |   |   chol<=312.0=True [2, 9]
|   |   |   |   |   age<=34.5=False [2, 9]
|   |   |   |   |   |   age<=36.0=False [2, 9]
|   |   |   |   |   |   |   age<=38.0=False [2, 9]
|   |   |   |   |   |   |   |   age<=39.5=False [2, 9]
|   |   |   |   |   |   |   |   |   age<=40.5=False [2, 9]
|   |   |   |   |   |   |   |   |   |   age<=41.5=False [2, 9]
|   |   |   |   |   |   |   |   |   |   |   age<=42.5=False [2, 9]
|   |   |   |   |   |   |   |   |   |   |   |   age<=43.5=False [2, 9]
|   |   |   |   |   |   |   |   |   |   |   |   |   age<=44.5=False [2, 9]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=45.5=False [2, 9]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=46.5=False [2, 9]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=47.5=False [2, 9]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=48.5=False [2, 9]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=49.5=False [2, 8]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=50.5=False [2, 7]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=54.5=False [1, 7]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=51.5=False [1, 7]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=52.5=False [1, 7]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=53.5=False [1, 7]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=55.5=False [1, 7]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=56.5=False [1, 7]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=57.5=False [1, 6]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=58.5=False [1, 5]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   chol<=231.5=False [0, 5]: 1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   chol<=231.5=True [1, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=58.5=True [0, 1]: 1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=57.5=True [0, 1]: 1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=56.5=True [0, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=55.5=True [0, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=53.5=True [0, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=52.5=True [0, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=51.5=True [0, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=54.5=True [1, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=50.5=True [0, 1]: 1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=49.5=True [0, 1]: 1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=48.5=True [0, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=47.5=True [0, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=46.5=True [0, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   age<=45.5=True [0, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   |   age<=44.5=True [0, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   |   age<=43.5=True [0, 0]: -1
|   |   |   |   |   |   |   |   |   |   |   age<=42.5=True [0, 0]: -1
|   |   |   |   |   |   |   |   |   |   age<=41.5=True [0, 0]: -1
|   |   |   |   |   |   |   |   |   age<=40.5=True [0, 0]: -1
|   |   |   |   |   |   |   |   age<=39.5=True [0, 0]: -1
|   |   |   |   |   |   |   age<=38.0=True [0, 0]: -1
|   |   |   |   |   |   age<=36.0=True [0, 0]: -1
|   |   |   |   |   age<=34.5=True [0, 0]: -1
|   |   |   slope=up [3, 1]
|   |   |   |   thalach<=165.5=False [0, 1]: 1
|   |   |   |   thalach<=165.5=True [3, 0]: -1
|   |   cp=typ_angina [3, 1]
|   |   |   oldpeak<=0.30000000000000004=False [3, 0]: -1
|   |   |   oldpeak<=0.30000000000000004=True [0, 1]: 1
|   trestbps<=109.0=True [5, 0]: -1
```

Outline for November 16

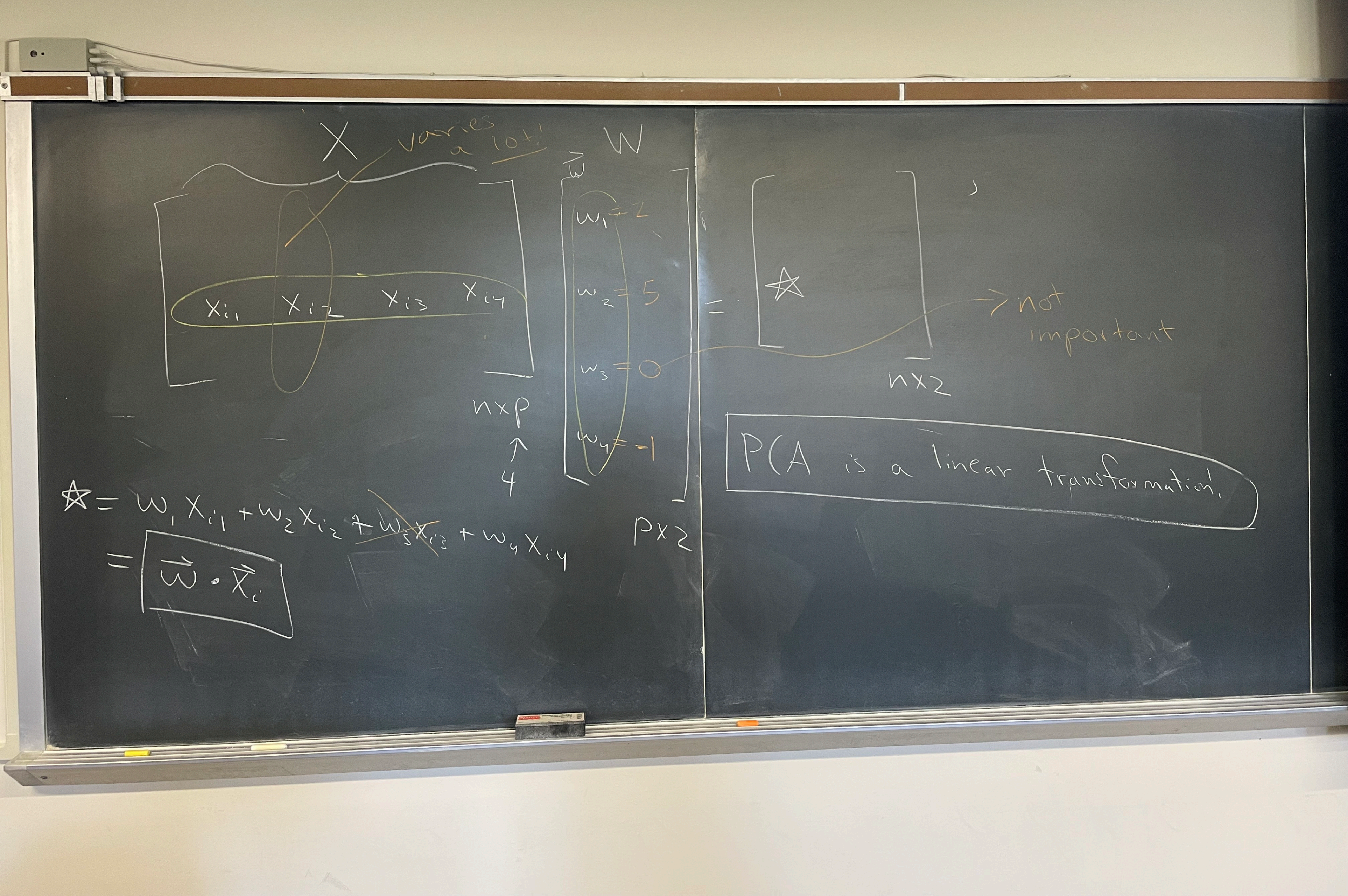
- Midterm 2 Review
  - Entropy vs. classification error
  - PCA
  - Naïve Bayes
  - Central Limit Theorem
  - Logistic regression and cross entropy

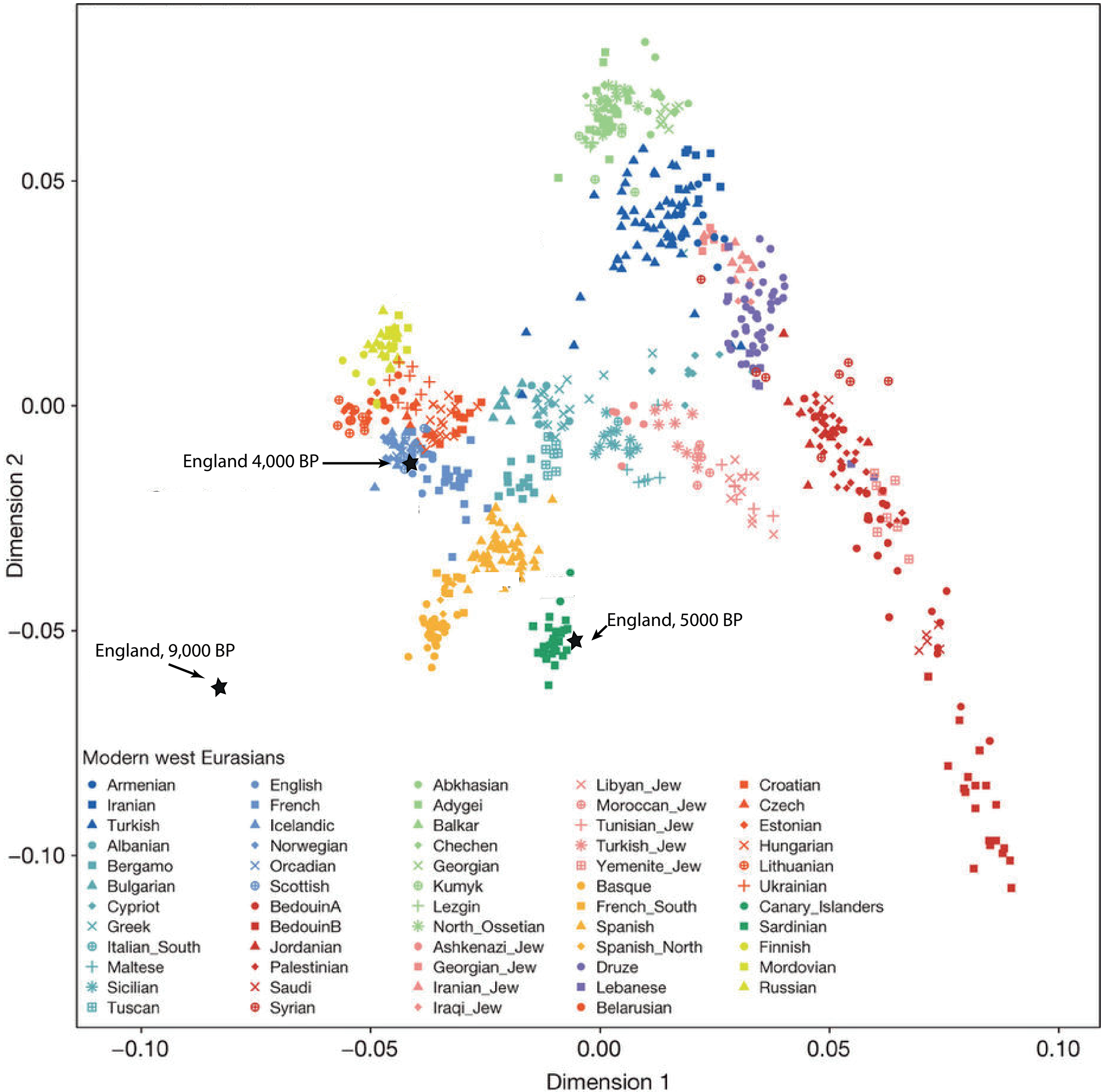
From the study guide

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

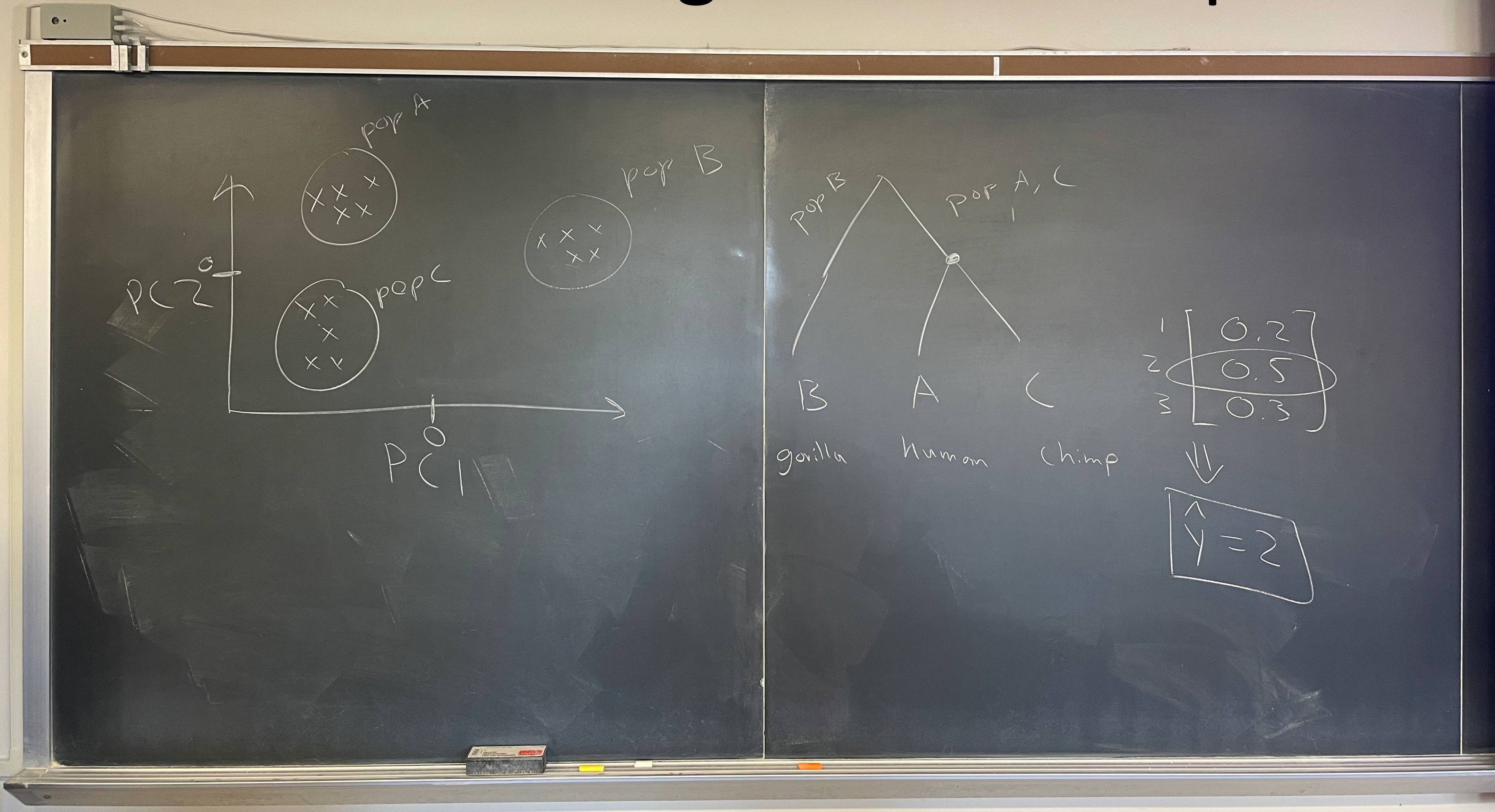
PCA creates linear combinations of features





England 4,000 BP
England, 5000 BP
England, 9,000 BP
Dimension 1
Dimension 2
0.05
0.00
−0.05
−0.10
−0.10
−0.05
0.00
0.05
0.10
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

PCA "classic" genetics example



Outline for November 16

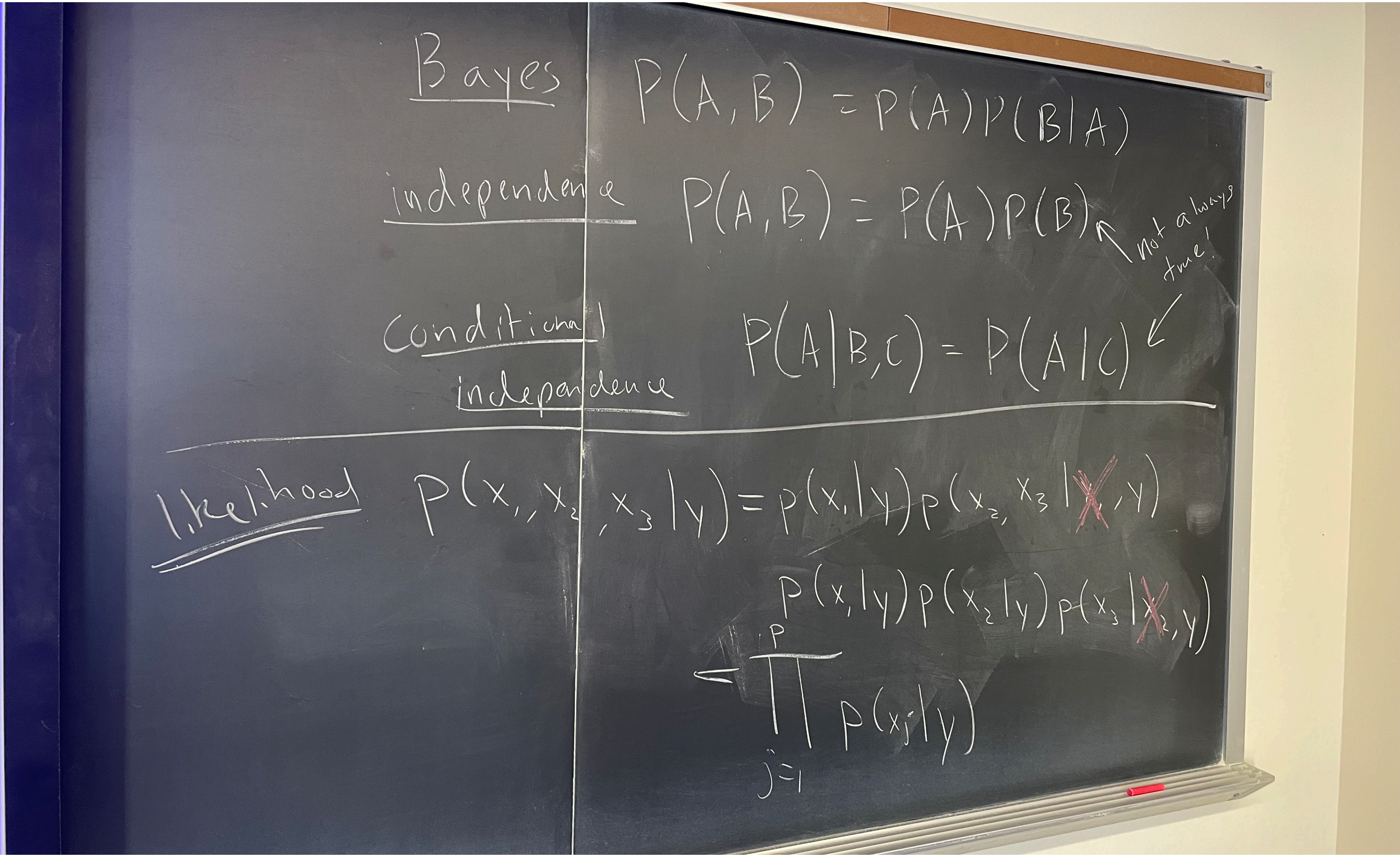
- Midterm 2 Review
  - Entropy vs. classification error
  - PCA
  - Naïve Bayes
  - Central Limit Theorem
  - Logistic regression and cross entropy

From the study guide

2. Naive Bayes

- Bayes rule in data science: identify and explain the evidence, prior, posterior, likelihood.
- 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

Naïve Bayes assumption



Naïve Bayes Model

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

Naïve Bayes Prediction

$$\hat{y} = \underset{k \in \{1,2,\cdots,K\}}{\arg\max}\ p(y = k) \prod_{j=1}^{p} p(x_j|y = k).$$

Estimating prior: p(y=k)

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

Estimating likelihood: p(x_j=v | y=k)

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

Outline for November 16

- Midterm 2 Review
  - Entropy vs. classification error
  - PCA
  - Naïve Bayes
  - Central Limit Theorem
  - 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"

Bootstrap demo

See video tutorial on Piazza!

Outline for November 16

- Midterm 2 Review
  - Entropy vs. classification error
  - PCA
  - Naïve Bayes
  - Central Limit Theorem
  - Logistic regression and cross entropy

From the study guide

5. <u>Logistic Regression</u>

- 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 (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)
- Idea of SGD for logistic regression, relationship to linear regression

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!

Notecards: will post responses on Piazza!