CS 260: Foundations of Data Science

Prof. Sara Mathieson

Spring 2025

HAVERFORD COLLEGE

Admin

- **Lab 8** due **TONIGHT**!
- Review today and Tuesday (in class and lab)
- **Exam** next Thursday (week from today)
- **Sophomore CS majors meeting TODAY**
  - 3:30-4:30pm in H109

Outline

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

Outline

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

Confusion matrix with more classes

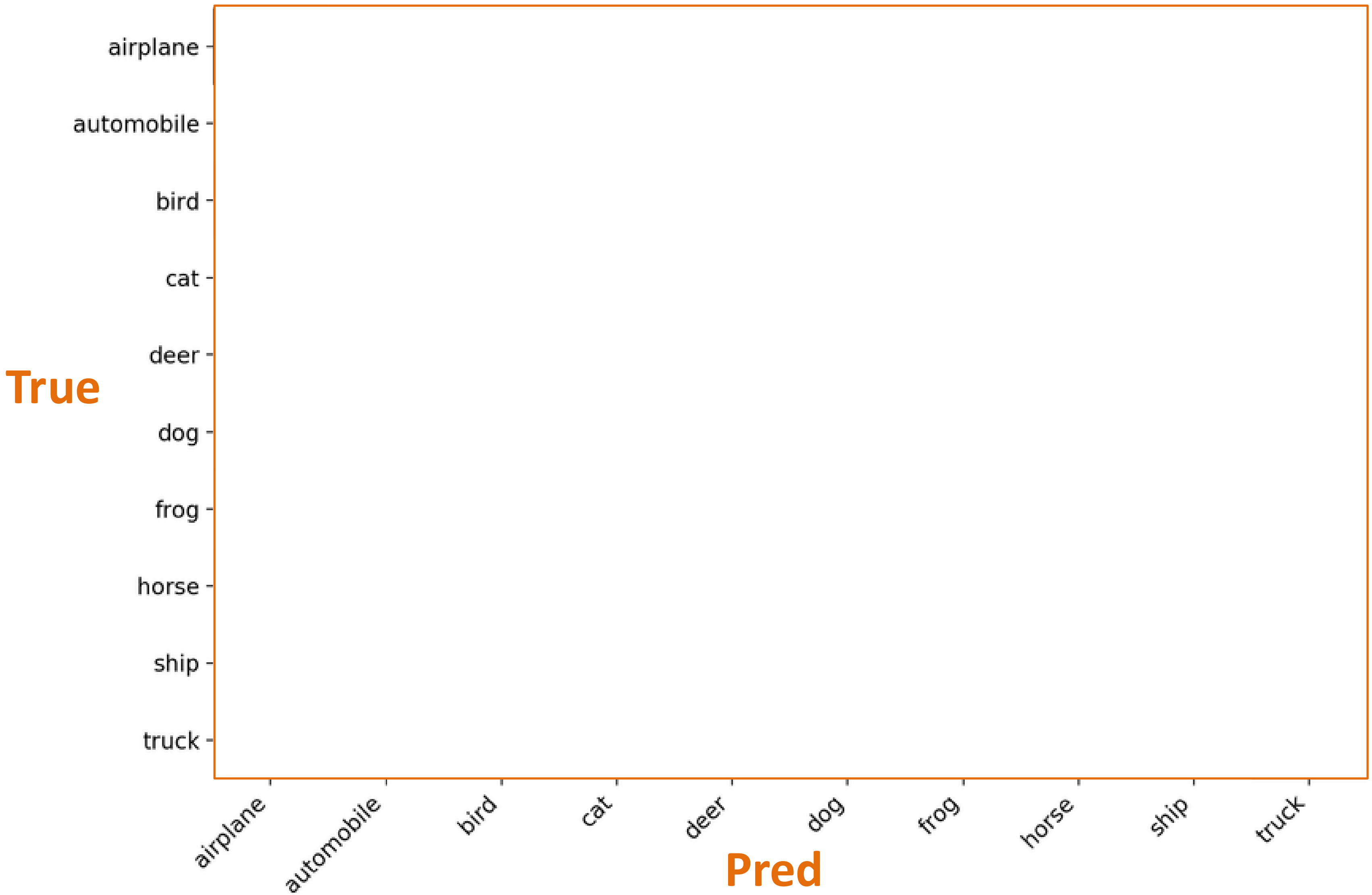


Figure by: Qun Liu (confusion matrix on cifar-10 dataset)

Confusion matrix with more classes

True	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

Pred

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

\# classes (K)

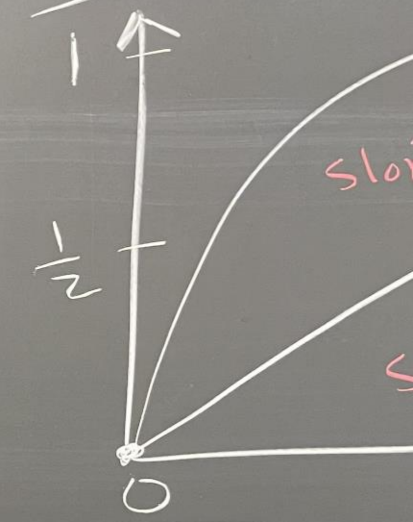
```
cm = np.zeros((K, K))
for ex in test:
    true = ex.label
    pred = model.classify(ex.features)
    cm[true, pred] += 1
```

true: i.e. 3

pred: i.e. 1

cm[3, 1]

entropy vs



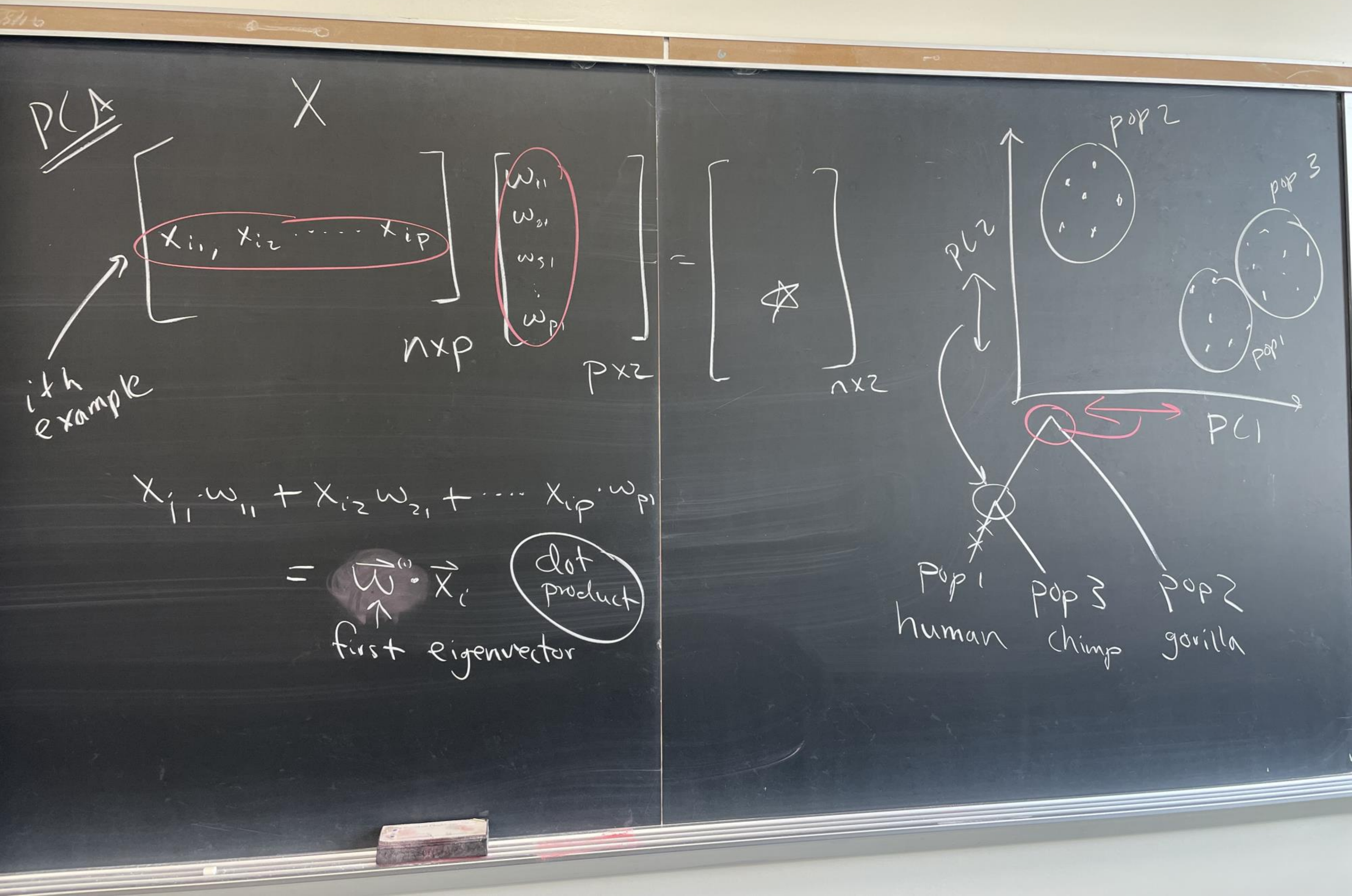
info gain

From the study guide

6. Data Visualization, Dimensionality Reduction, and Unsupervised Learning

- 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
- Basic idea of tSNE as an alternative to PCA
- ~~Idea of clustering as a form of unsupervised learning, for example K-means~~

PCA creates linear combinations of features



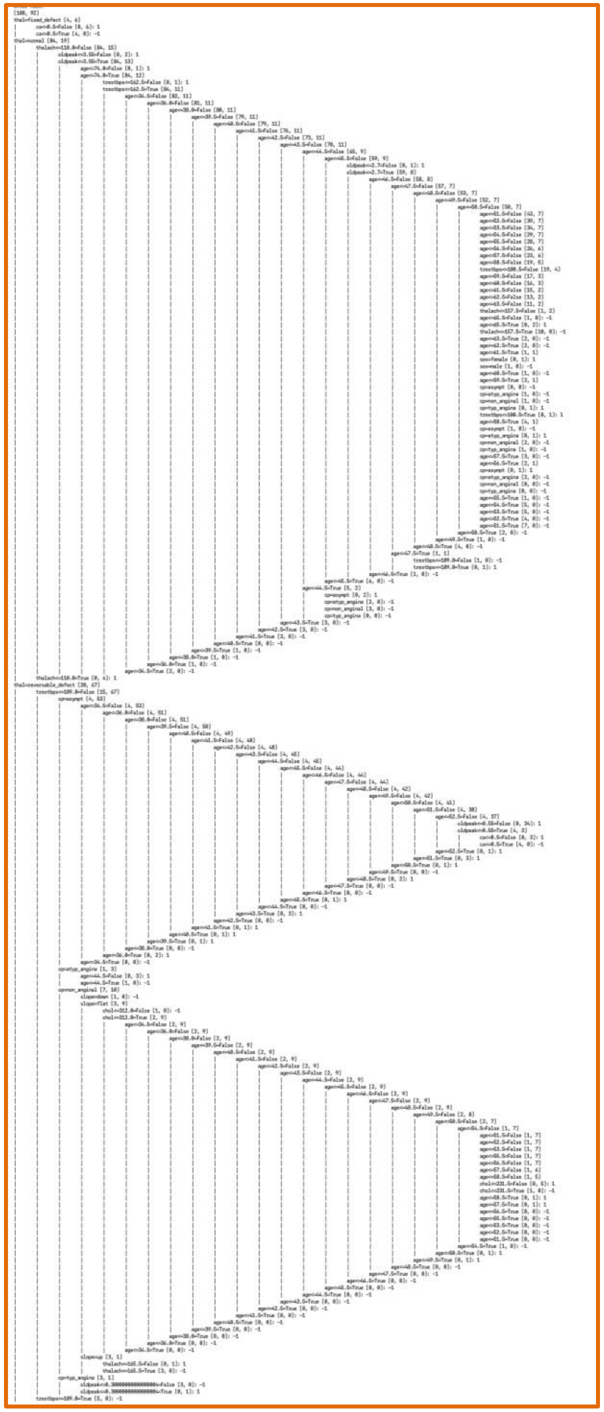
From the study guide

4. Information Theory

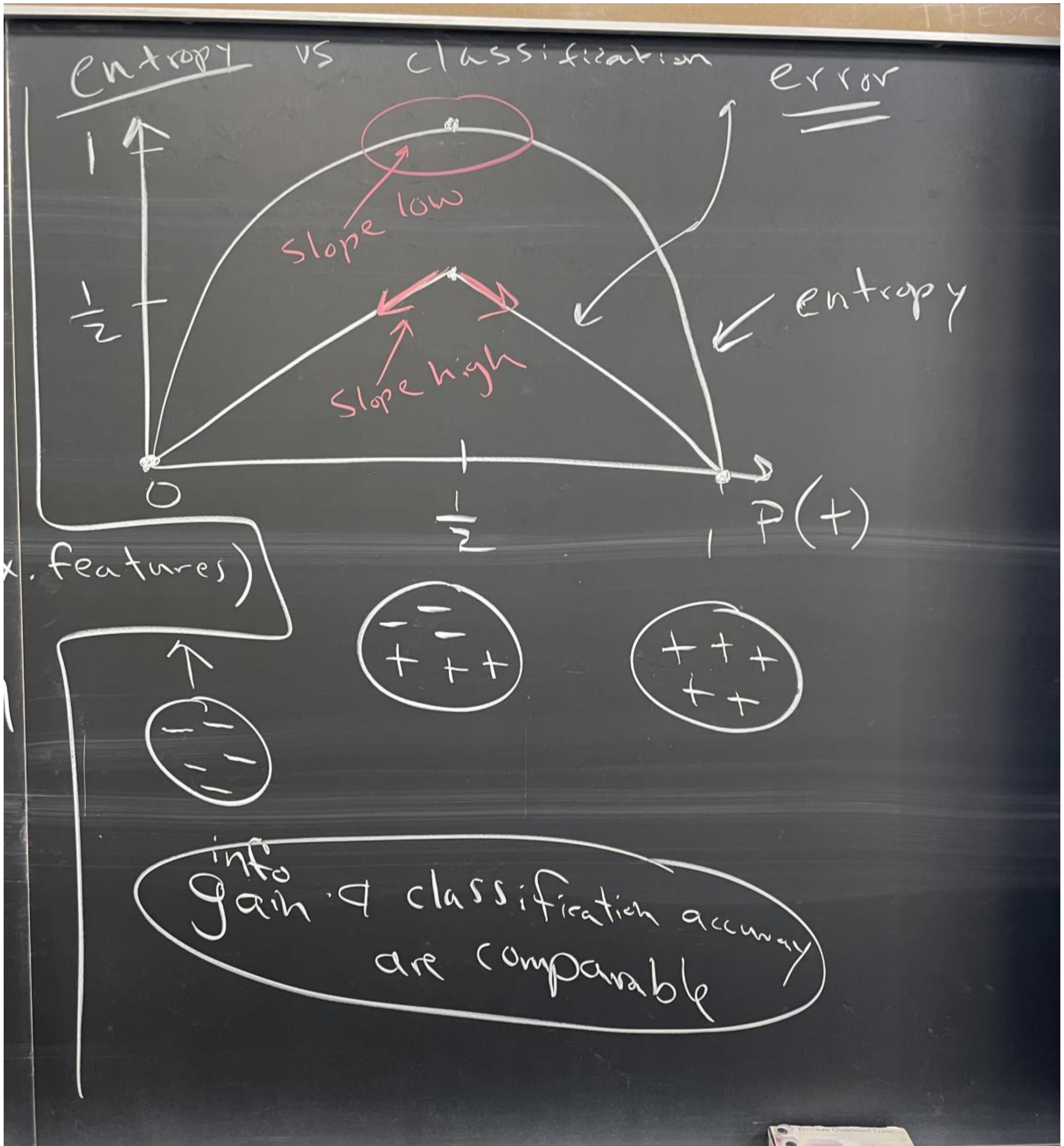
- Conceptual idea of entropy as well as formal definition
- Shannon encoding (and decoding), plus how to use entropy to compute average number of bits needed to send one piece of information
- Use of conditional entropy and information gain to choose best features
- Comparison with classification accuracy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

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
```



Entropy vs. classification error



Midterm Practice Exam: pg 1 and 2

On your own for ~15 min, then partners

① step 1: get the data ✓ $O(1)$

step 2: subtract off mean $\boxed{O(np)}$

$O(np) + O(np)$ $\left[\,\text{(columns)}\,\right]_{n \times p}$

step 3: cov of each pair of features

$$\sum_{i=1}^{n} (x_{if} - \bar{f})(x_{ig} - \bar{g}) \; \Big\} \; O(n)$$

(2 features: f, g; means $\bar{f}, \bar{g} \to 0$)

p^2 pairs of features $\Rightarrow \boxed{O(p^2 n)}$

$$A = \begin{bmatrix} 0 & & \\ 0 & & \\ 0 & & \end{bmatrix}_{p \times p}$$

step 4: eigenvalues & eigenvectors of A

$\Rightarrow$ cubic $\boxed{O(p^3)}$

Step 5: transform data

$$\underbrace{(n \times p)}_{X} \times \underbrace{(p \times r)}_{W_r} = \underbrace{(n \times r)}_{T_r} \rightarrow \boxed{O(npr)}$$

step 6: plot!

Overall runtime of PCA

$O(\cancel{np}) + O(np^2) + O(p^3) + \cancel{O(npr)}$

$r \leq p$

$\Rightarrow \boxed{O(np^2 + p^3)}$

② (a) $E[Y] = \sum_y y\,p(y)$

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

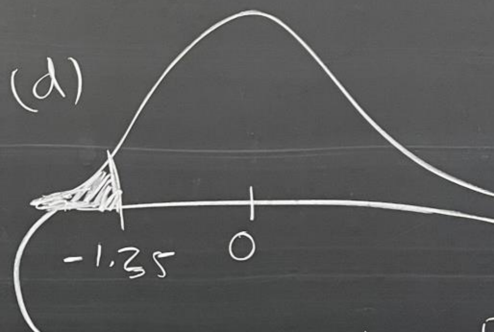
$= \boxed{2.125} = \mu$

(b) $Var(Y) = E[(Y-\mu)^2]$

$= (0 - 2.125)^2 \cdot \frac{1}{8} + \ldots = \boxed{1.109} = \sigma^2$

(c) $Z = \dfrac{\bar{Y}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \dfrac{1.9 - 2.125}{\sqrt{\frac{1.109}{40}}} = \boxed{-1.35}$

(d)



area = p-value = $\boxed{0.08833} > 0.05$

fail to reject H_0

③ (a) age ⇒ $\boxed{\frac{15}{99}}$

oldpeak = $\boxed{\frac{13}{99}}$ } choose as best feature

(b) $H(Y) = -\left(\underbrace{\frac{84}{99}\log_2\frac{84}{99}}_{Y=-1} + \underbrace{\frac{15}{99}\log_2\frac{15}{99}}_{Y=+1}\right)$

$= 0.61$

(c) H(Y | oldpeak

H(Y

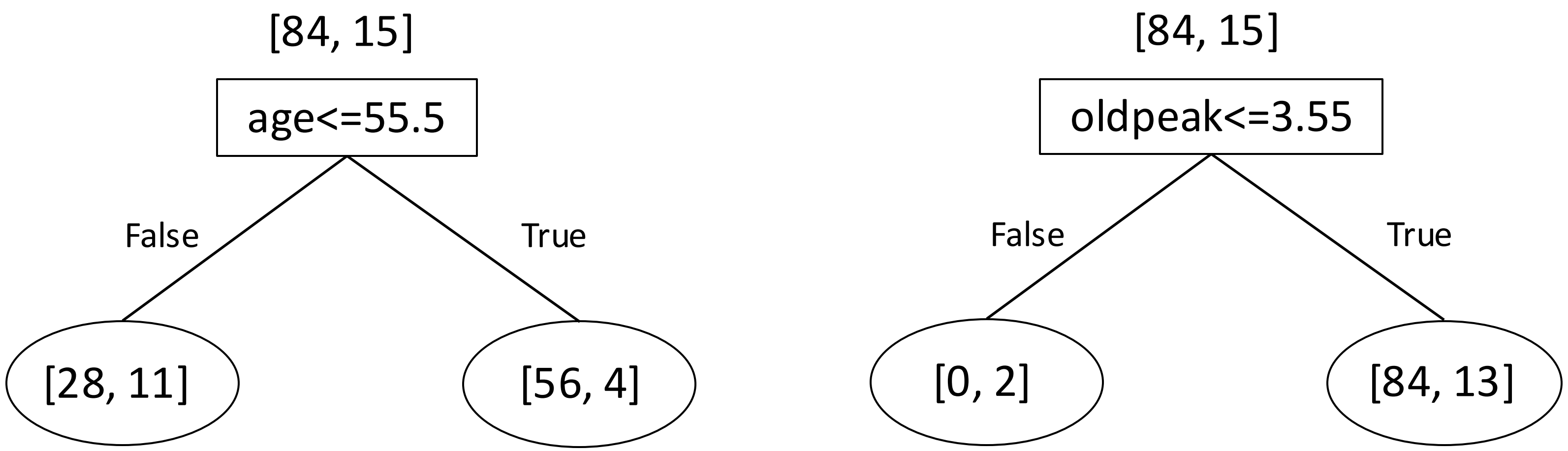
=

(C) $H(Y \mid \text{oldpeak}) = \frac{2}{99} H(Y \mid \text{oldpeak} = F) + \frac{97}{99} H(Y \mid \text{oldpeak} = T)$

(the term $H(Y \mid \text{oldpeak} = F)$ is marked with an arrow $\to 0$)

$$H(Y \mid \text{oldpeak} = T) = -\left(\frac{84}{97}\log_2\frac{84}{97} + \frac{13}{97}\log_2\frac{13}{97}\right)$$

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!