CS 260: Foundations of Data Science

Prof. Sara Mathieson

Spring 2025

HAVERFORD
COLLEGE

Admin

- Lab 5 due **Wednesday** (tomorrow)
- Lab 6 posted
- **Midterm 1** returned today
- No office hours on Wednesday (tomorrow)

- **Lab today**: Lab 5 implementation advice and check-ins
  - If you're *completely* finished, don't need to attend, but please email me
  - Otherwise will check in about Lab 5 or start Lab 6

Lab 5 implementation

Partition contains:

- Features dictionary F:

  F = {**age**: [Senior, Middle-age, Mid-adult, Young-adult, Child], **workclass**: [Private, Local-gov…] … }

- List of Examples
  - Each example contains

    features = {**age**: Senior, **workclass**: Private … }

    label = 1 (Female)

defaultdict

```
from collections import defaultdict

# Create a defaultdict with int as the default factory (default value is
count_dict = defaultdict(int)

# Increment the count for some items
items = ['apple', 'banana', 'apple', 'orange', 'banana', 'apple']
for item in items:
    count_dict[item] += 1

# Print the counts
print(count_dict)
# Expected output: defaultdict(<class 'int'>, {'apple': 3, 'banana': 2,

# Access a non-existent key
print(count_dict['grape'])
# Expected output: 0 (because int() returns 0)
```

Credit: Gemini

Outline

- Entropy and Shannon encoding
- Information gain for selecting features
- Go over Midterm 1
- Continuous features (if time)

Outline

- Entropy and Shannon encoding
- Information gain for selecting features
- Go over Midterm 1
- Continuous features (if time)

Decision Trees use entropy to select best features

Examples

- Medical diagnostics

- Credit risk analysis

- Modeling calendar scheduling preferences

Decision Trees in Chemistry reactions

- Example of decision trees in practice
- Use decision trees to interpret another ML algorithm (SVMs)

Machine-learning-assisted materials discovery using failed experiments

Paul Raccuglia, Katherine C. Elbert, Philip D. F. Adler, Casey Falk, Malia B. Wenny, Aurelio Mollo, Matthias Zeller, Sorelle A. Friedler ✉, Joshua Schrier ✉ & Alexander J. Norquist ✉

Nature **533**, 73–76 (05 May 2016) | Download Citation ⭳

Optional Reading!

How do we choose the best feature?

- Single feature model + evaluate with a ROC curve **(Lab 4)**

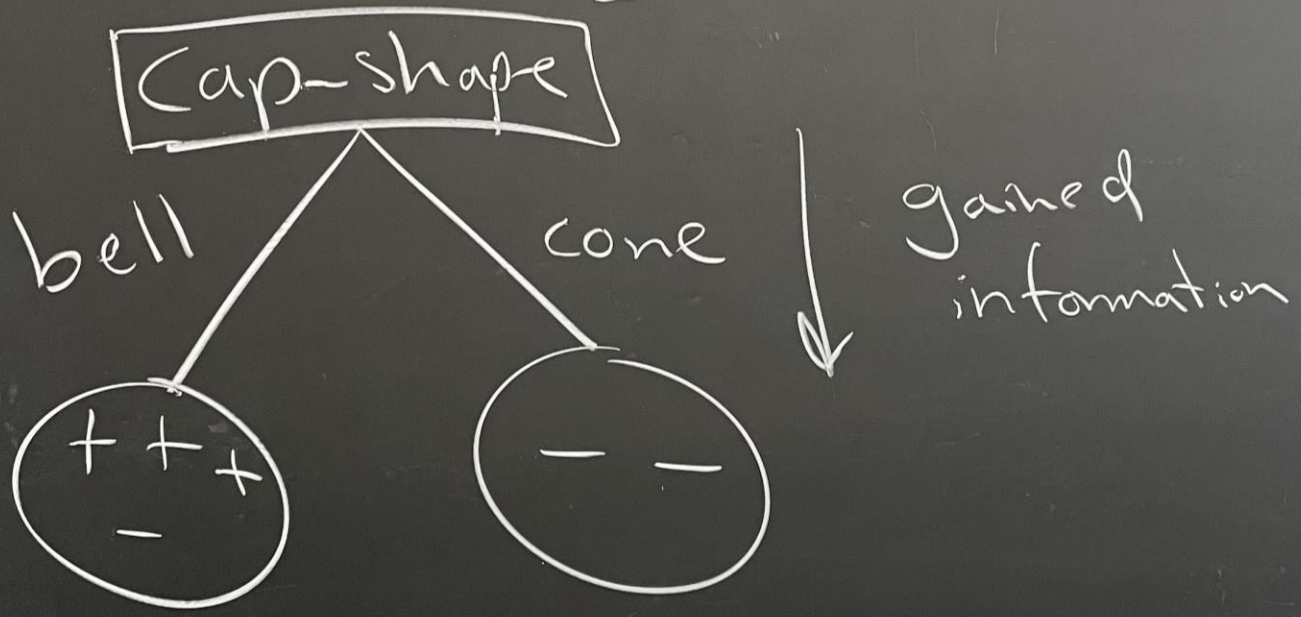
- What feature gives us the most information about the label? **(Lab 6)**

Information Theory

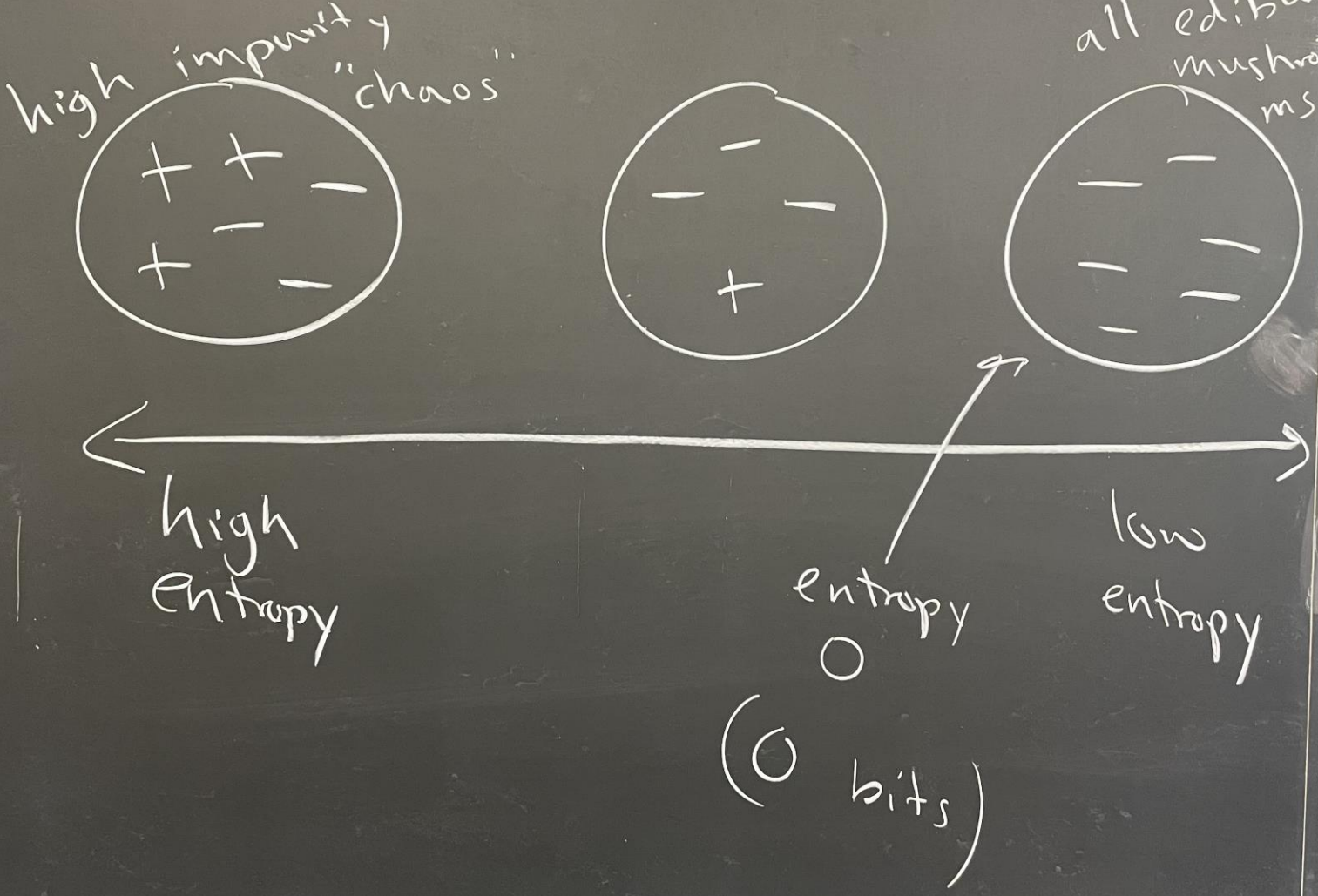
How do we choose the best feature?

Lab 4

− + = edible
+ + = poisonous
+ − =



Idea of entropy : avg # of bits needed to send one data pt



How to find # bits?

year	prob (p)	idea	Cummutative prob	in binary!	# digits to take $\lceil -\log_2(p) \rceil$	code (Shannon)
senior	0.5	0	0	0.000…	1	0
junior	0.25	1	0.5	0.1000…	2	10
soph	0.125	01	0.75	0.1100…	3	110
first-year	0.125	10	0.875	0.1110…	3	111
Sorted high → low			1			

xample first?

decode!

1 1 0 1 1 0 1 0 0 0 1 1 1 0

junior junior?

instead prefix encoding

no code is the prefix of another code

nary

$\cdots \square \cdot 2^2 + \square \cdot 2^1 + \square \cdot 2^0 + \square \cdot 2^{-1} + \square \cdot 2^{-2} \cdots$

$5 = 1 \cdot 4 + 0 \cdot 2 + 1 \cdot 1$

$\Rightarrow \boxed{101}$ in binary

5.5 $\Rightarrow$ 101.1

exponential decay

$\frac{1}{8}$ $\frac{1}{4}$ $\frac{1}{2}$ 1

linear increase

-1

-2

-3

half as likely $\neq$ twice as many bits

Outline

- Entropy and Shannon encoding
- Information gain for selecting features
- Go over Midterm 1
- Continuous features (if time)

entropy

$$H(Y) = -\sum_{c \in \text{vals}(Y)} \underbrace{p(y=c)}_{\text{prob or freq of } c} \underbrace{\log_2 p(y=c)}_{\text{\# bits for } c}$$

$$H(\text{year}) = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 3$$

$$= \boxed{1.75} \quad \text{avg \# bits}$$

not ~~$\frac{1+2+3+3}{4} = 2.25$~~

Conditional entropy

$$H(Y|X) = \sum_{v \in vals(X)} p(X=v)\, H(Y|X=v)$$

label | feature

?

Conditional entropy of one feature value (leaf)

$$H(Y|X=v) = -\sum_{c \in vals(Y)} p(Y=c|X=v) \log_2 p(Y=c|X=v)$$

ex

$H(Y \mid \underbrace{\text{cap-shape}}_{X} = \underbrace{\text{bell}}_{v})$

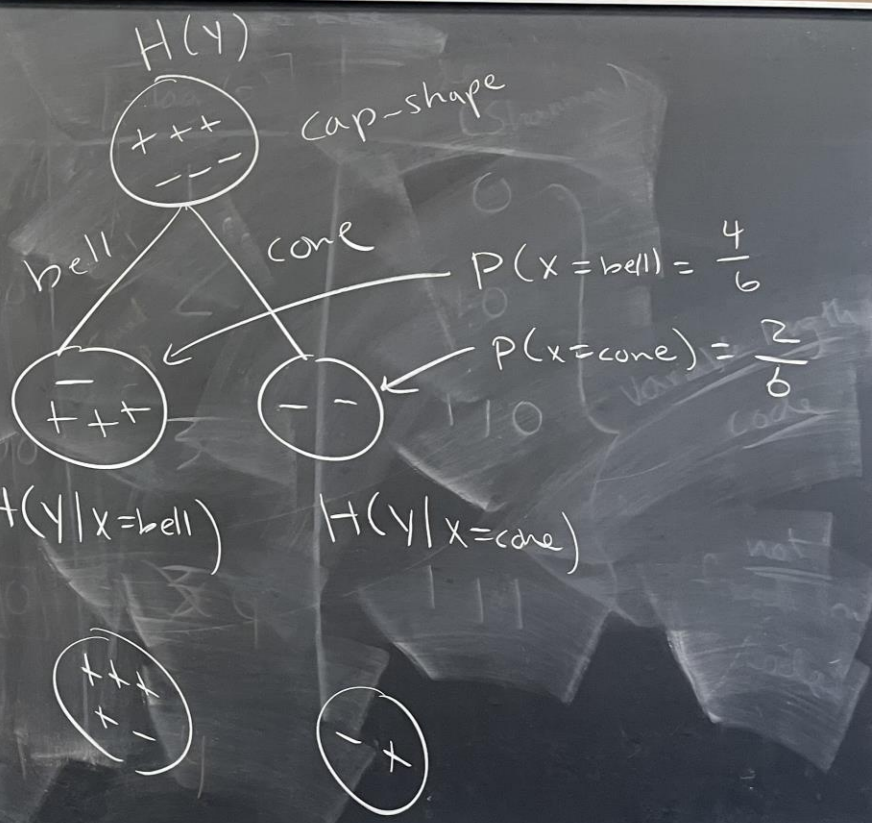
$= -\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}\right)$

$\approx \boxed{0.811}$

Information Gain

$G(X, Y) = H(Y) - H(Y \mid X)$

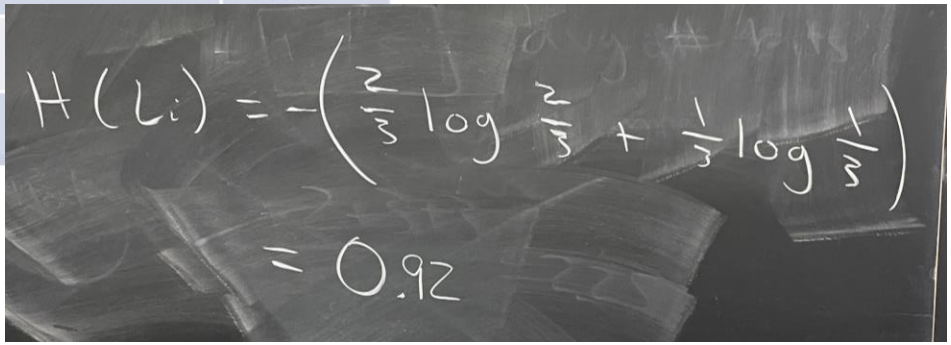
Want high — Want low



Handout 13

Handout 13

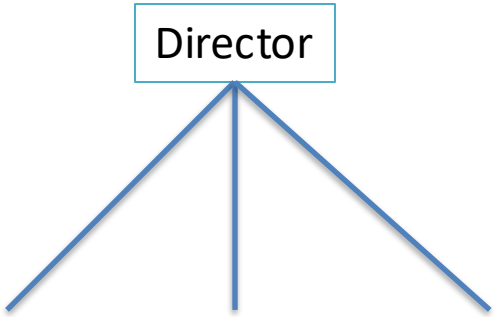
Movie	Type	Length	Director	Famous actors	Liked?
m1	Comedy	Short	Adamson	No	Yes
m2	Animated	Short	Lasseter	No	No
m3	Drama	Medium	Adamson	No	Yes
m4	Animated	Long	Lasseter	Yes	No
m5	Comedy	Long	Lasseter	Yes	No
m6	Drama	Medium	Singer	Yes	Yes
m7	Animated	Short	Singer	No	Yes
m8	Comedy	Long	Adamson	Yes	
m9	Drama	Medium	Lasseter	No	



$P(\text{Li} = \text{yes}) =$ 2/3
$H(\text{Li}) =$ 0.92

$H(\text{Li} \mid \text{T}) = 0.61$
$H(\text{Li} \mid \text{Le}) = 0.61$
$H(\text{Li} \mid \text{D}) = 0.36$ MIN ENTROPY
$H(\text{Li} \mid \text{F}) = 0.85$

$\text{Gain}(\text{Li}, \text{T}) =$ 0.92 – 0.61 = 0.31
$\text{Gain}(\text{Li}, \text{Le}) =$ 0.92 – 0.61 = 0.31
$\text{Gain}(\text{Li}, \text{D}) =$ 0.92 – 0.36 = 0.56 MAX INFO GAIN
$\text{Gain}(\text{Li}, \text{F}) =$ 0.92 – 0.85 = 0.07



Start of the tree

Outline

- Entropy and Shannon encoding
- Information gain for selecting features
- Go over Midterm 1
- Continuous features (if time)

Midterm 1 Grades

Median: 88

- 88-100% A
- 78-87% B
- 68-77% C
- 58-67% D
- Below 58%: not passing, please meet with me

- Note: as per the syllabus, you must pass at least one exam to pass the course
- Any questions about the exam: bring to me within one week

Quote of the week

“Enjoy the little things, for one day you may look back and realize they were the big things.”

--Robert Breault