CS 260: Foundations of Data Science

Prof. Sara Mathieson

Fall 2021

HAVERFORD COLLEGE

Admin

- Lab 5 due **Thursday**
- TA hours **now Mon, Tues, Thurs**
  - Monday office hours now over zoom only
- Office hours TODAY **3:30-5pm in H204**
- **Note-taker:** Matt

- **No class on Thursday** (I'm "at" a conference)
  - Watch video lecture instead (can come to class)
  - Do Handout 14 and come ready to discuss on Tues!
  - Note-taker: Victor
  - There will still be lab!

Lab 5 suggestions

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)

Outline for October 19

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

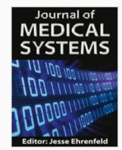
Outline for October 19

- 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



Journal of Medical Systems
October 2002, Volume 26, Issue 5, pp 445–463 | Cite as

Decision Trees: An Overview and Their Use in Medicine

Authors | Authors and affiliations

Vili Podgorelec, Peter Kokol, Bruno Stiglic, Ivan Rozman

- Credit risk analysis

Computational Economics
April 2000, Volume 15, Issue 1–2, pp 107–143 | Cite as

Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications

Authors | Authors and affiliations

J. Galindo, P. Tamayo

- 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!

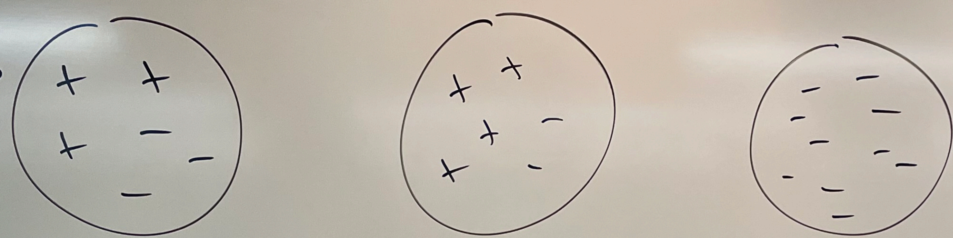
How do we choose the best feature?

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

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

Idea of entropy

edible & poisonous

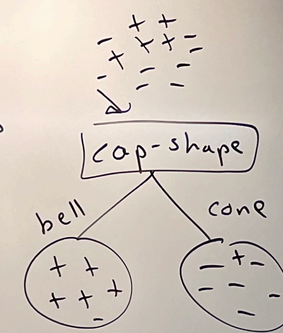


all edible mushrooms

high entropy ⟷ low entropy

$\boxed{\text{avg \# of bits needed to send one datapoint}}$

0/1



reduced entropy

PLEASE LEAVE COMPUTERS ON

Shannon encoding

year	prob (p)	~~idea~~	cumulative prob	binary	$\lceil -\log_2(p) \rceil$	Shannon encoding
senior	0.5	~~0~~	0	0.000…	1	0
junior	0.25	~~1~~	0.5	0.100…	2	10
soph	0.125	~~01~~	0.75	0.110…	3	110
first	0.125	~~10~~	0.875	0.1110	3	111

sort

proportions of each class year

Ambiguous

bits

unambiguous.

variable length encoding

110 110 10 0 0 111 0

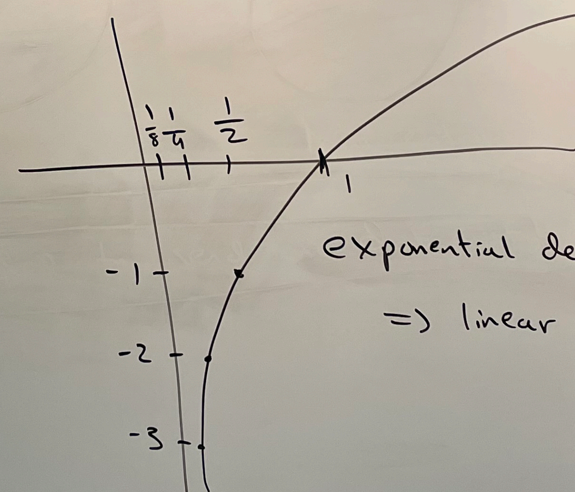
$$\ldots \; \square \cdot 2^{2} + \square \cdot 2^{1} + \square \cdot 2^{0} + \square \cdot 2^{-1} + \square \cdot 2^{-2} \ldots$$

decimal point

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

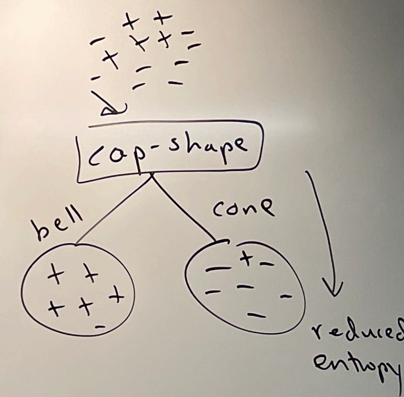
$$5 = 1 \cdot 2^{2} + 0 \cdot 2^{1} + 1 \cdot 2^{0}$$

↳ in binary => 101



exponential decay

=> linear function



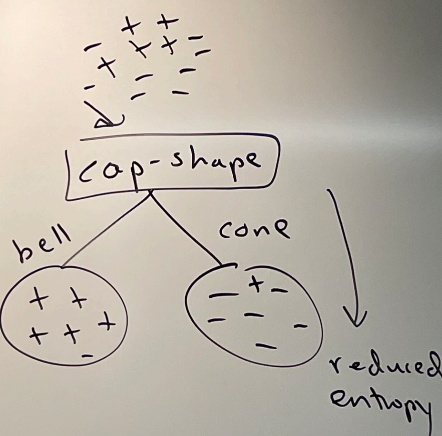
Entropy

$$H(y) = -\sum_{c \in vals(y)} \overbrace{P(y=c)}^{\text{prob of } c} \underbrace{\log_2 P(y=c)}_{\text{\# bits for class } c}$$

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

$$= \boxed{1.75 \text{ bits!}}$$

$$\cancel{\frac{1+2+3+3}{4} = 2.25}$$



Outline for October 19

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

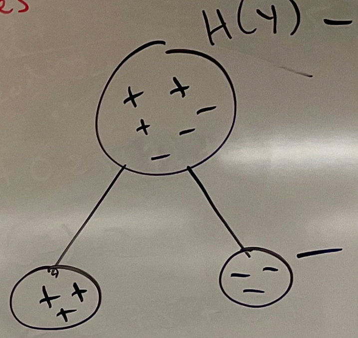
PLEASE LEAVE COMPUTERS ON

Conditional entropy

$$H(Y|X) = \sum_{v \in vals(X)} \overbrace{P(X=v)}^{\frac{\#\ \text{examples with } v}{\text{total}}} \underbrace{H(Y|X=v)}_{?}$$

label   one feature

weighted avg of entropies of the leaves



cond. entropy of one value:

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

$$\rightarrow -\tfrac{1}{2}\log\tfrac{1}{2} - \tfrac{1}{2}\log\tfrac{1}{2}$$

$$= \tfrac{1}{2} + \tfrac{1}{2} = 1$$

$$\rightarrow H(y|x) = 0$$

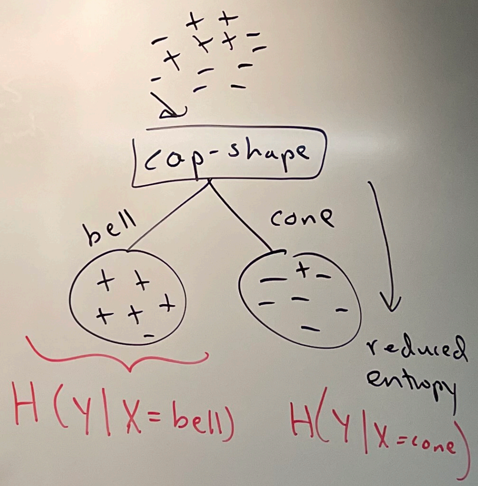
Handout 13

Information Gain

$$G(Y,X) = H(Y) - H(Y|X)$$

how much entropy did feature X remove from the system?

$\Rightarrow$ want the feature w/ max info gain!



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	Yes
m9	Drama	Medium	Lasseter	No	Yes

$\mathrm{P}(\mathrm{Li} = \mathrm{yes}) =$ 2/3

$\mathrm{H}(\mathrm{Li}) =$ 0.92

$\mathrm{H}(\mathrm{Li} \mid \mathrm{T}) = 0.61$

$\mathrm{H}(\mathrm{Li} \mid \mathrm{Le}) = 0.61$

$\boxed{\mathrm{H}(\mathrm{Li} \mid \mathrm{D}) = 0.36}$ MIN ENTROPY

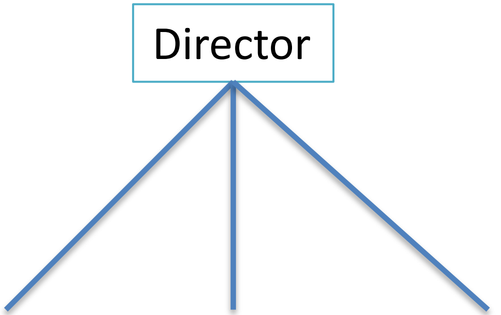
$\mathrm{H}(\mathrm{Li} \mid \mathrm{F}) = 0.85$

$\mathrm{Gain}(\mathrm{Li}, \mathrm{T}) =$ 0.92 – 0.61 = 0.31

$\mathrm{Gain}(\mathrm{Li}, \mathrm{Le}) =$ 0.92 – 0.61 = 0.31

$\boxed{\mathrm{Gain}(\mathrm{Li}, \mathrm{D}) = 0.92 - 0.36 = 0.56}$ MAX INFO GAIN

$\mathrm{Gain}(\mathrm{Li}, \mathrm{F}) =$ 0.92 – 0.85 = 0.07



Start of the tree

PLEASE LEAVE COMPUTERS ON

① $P(Li = Yes) = \frac{2}{3}$

real code need an integer

② $H(Li) = -\frac{2}{3}\log_2\frac{2}{3} - \underbrace{\frac{1}{3}\log_2\frac{1}{3}} = 0.92$

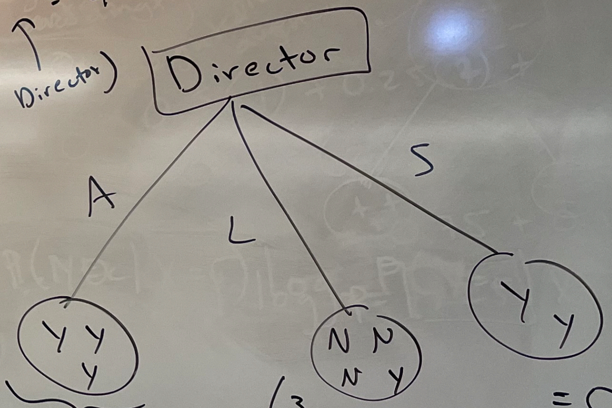
③ $Gain(Li, T) = 0.92 - 0.61 = 0.31$

$= 0.31$

$0.92 - 0.36 = \boxed{0.56}$

$= 0.07$

$H(Y|\text{Director}) = \frac{3}{9}\cdot 0 + \frac{4}{9}\cdot H(Y|X=L) + \frac{2}{9}\cdot 0$



$H(Y|X=A) = 0$

$-\left(\frac{3}{4}\log\frac{3}{4} + \frac{1}{4}\log\frac{1}{4}\right)\cdot\frac{4}{9}$

≈ 0.36

$= 0$

Outline for October 19

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

Midterm 1 Grades

- 90-100% A
- 80-89% B
- 70-79% C
- Below 60%: not passing (please meet with me)

- Average: 85

- Any questions about the exam: bring to me within one week

Midterm 1

(not posted online)

Outline for October 19

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

Next time!