CS 260: Foundations of Data Science

Prof. Sara Mathieson

Fall 2023

HAVERFORD COLLEGE

Admin

- Lab 5 due **Wednesday** (tomorrow)
- Lab 6 posted tomorrow
- **Midterm 1** returned today

- **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

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)

Outline for October 24

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

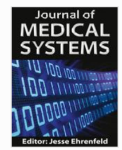
Outline for October 24

- 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!

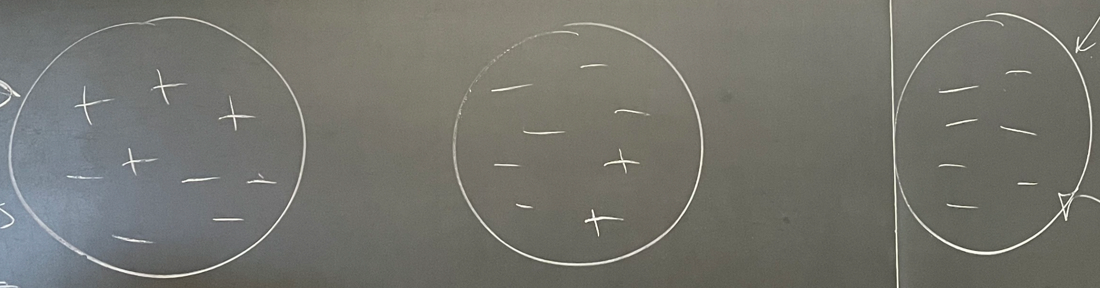
How do we choose the best feature?

- 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 : avg # of bits needed to send one datapoint

posinous + edible mushrooms



all edible

0 entropy (i.e. grocery store)

(1 bit)

high entropy

low entropy (what we want)

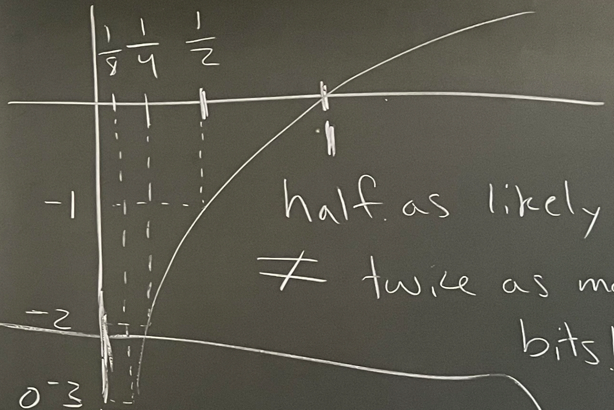
send information

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

junior junior senior/junior

soph?

$5.5 \Rightarrow \boxed{101.1}$



half as likely ≠ twice as many bits!

binary

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

decimal point!

$\frac{1}{2}$

$\frac{1}{4}$

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

$\Rightarrow \boxed{101}$ in binary

fixed len encoding

fixed len encoding	year	prob (p)	idea	cummulative prob	in binary	ceil round up $\lceil -\log_2(p) \rceil$	variable Shannon len encoding
00	senior	0.5	~~0~~	0	0.000...	1	0
01	junior	0.25	~~1~~	0.5	0.100...	2	10
10	soph	0.125	~~01~~	0.75	0.110...	3	110
11	first	0.125	~~10~~	0.875	0.1110...	3	111

works!

sort

prob (p): sum to 1

$\lceil -\log_2(p) \rceil$: # bits

no code is the prefix of another

Entropy

$$H(y) = -\sum_{c \in \text{vals}(y)} p(y=c) \overbrace{\log_2(p(y=c))}^{\#\text{ bits}}$$

label

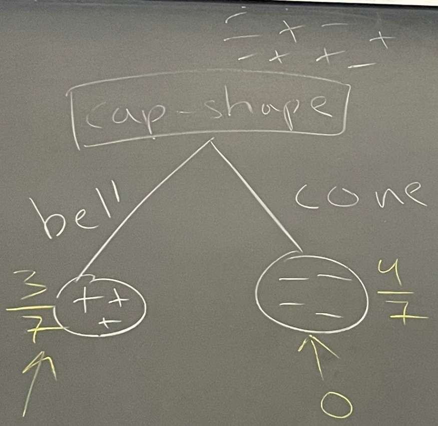
$$H(\text{year}) = \frac{1}{2}\cdot 1 + \frac{1}{4}\cdot 2 + \frac{1}{8}\cdot 3 + \frac{1}{8}\cdot 3$$

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

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

Outline for October 24

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



$$H(y \mid x = \text{bell}) = -\left(\cancel{0 \cdot \log 0}^{\,0} + \cancel{1 \cdot \log 1}^{\,0}\right)$$

$$= 0$$

Conditional entropy

$$H(y \mid \underset{\substack{\uparrow \\ \text{ex} \\ \text{cap-shape}}}{x}) = \sum_{v \in vals(x)} \overbrace{p(x=v)\underbrace{H(y \mid x=v)}_{?}}^{\text{weighted avg}} = \frac{3}{7} \cdot 0 + \frac{4}{7} \cdot 0 = 0$$

same

$$\underbrace{H(y \mid x=v)}_{\text{single feature value}} = -\sum_{c \in vals(y)} \underset{\substack{\uparrow \\ y=0}}{p}(y=c \mid \underset{\substack{\uparrow \\ \text{cap-shape = bell}}}{x=v}) \log_2 p(y=c \mid x=v)$$

all

=0

$\boxed{\text{Handout 13}}$

all values/leaves

Information gain

$$G(Y,X) = \underbrace{H(Y) - \underbrace{H(Y|X)}_{\text{want low}}}_{\text{want high}}$$

$\boxed{\text{Want the feature that maximizes info gain}}$

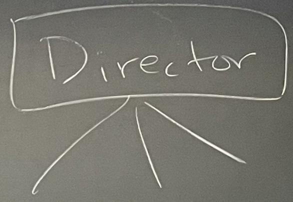
Handout 13

Handout 13

for Lab6
$\Rightarrow$ no rounding

1. $P(L_i = \text{yes}) = \frac{6}{9} = \frac{2}{3}$
2. $H(L_i) = -\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right) \approx 0.92$
3. $Gain(L_i, T) = 0.92 - 0.61 = 0.31$

$0.92 - 0.61 \ldots$



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

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

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

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

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

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

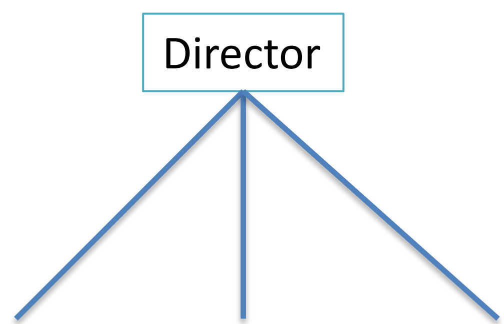
$\mathrm{H(Li \mid F)} = 0.85$

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

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

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

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



Start of the tree

Outline for October 24

- 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 70%: please meet with me
- Below 60%: not passing

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

Midterm solutions
not posted online

Outline for October 24

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