CS 260: Foundations of Data Science

Prof. Thao Nguyen

Fall 2024

HAVERFORD
COLLEGE

Materials by Sara Mathieson

Admin

- **Lab 5** due Tuesday (tomorrow)
- **Lab 6** posted, due next Monday (Oct 28)
- **Midterm 1** returned today

Outline for today

- Entropy and Shannon encoding

- Information gain for selecting features

- Go over Midterm 1

- Continuous features (if time)

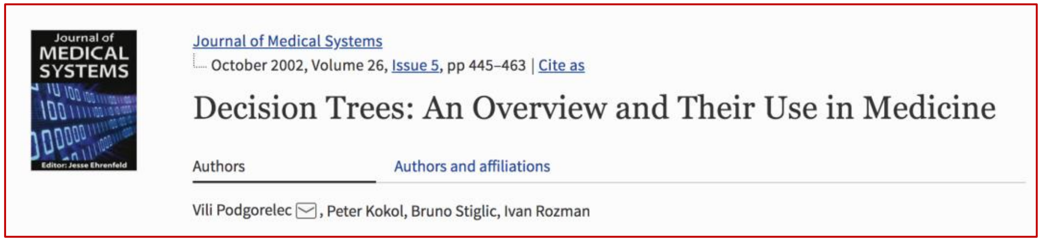
Outline for today

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

Applications of Decision Trees

Examples

- Medical diagnostics



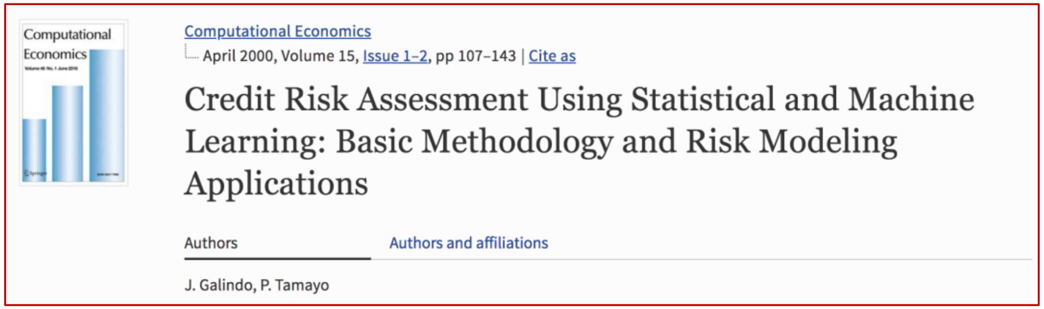
Journal of Medical Systems
October 2002, Volume 26, Issue 5, pp 445–463 | Cite as

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

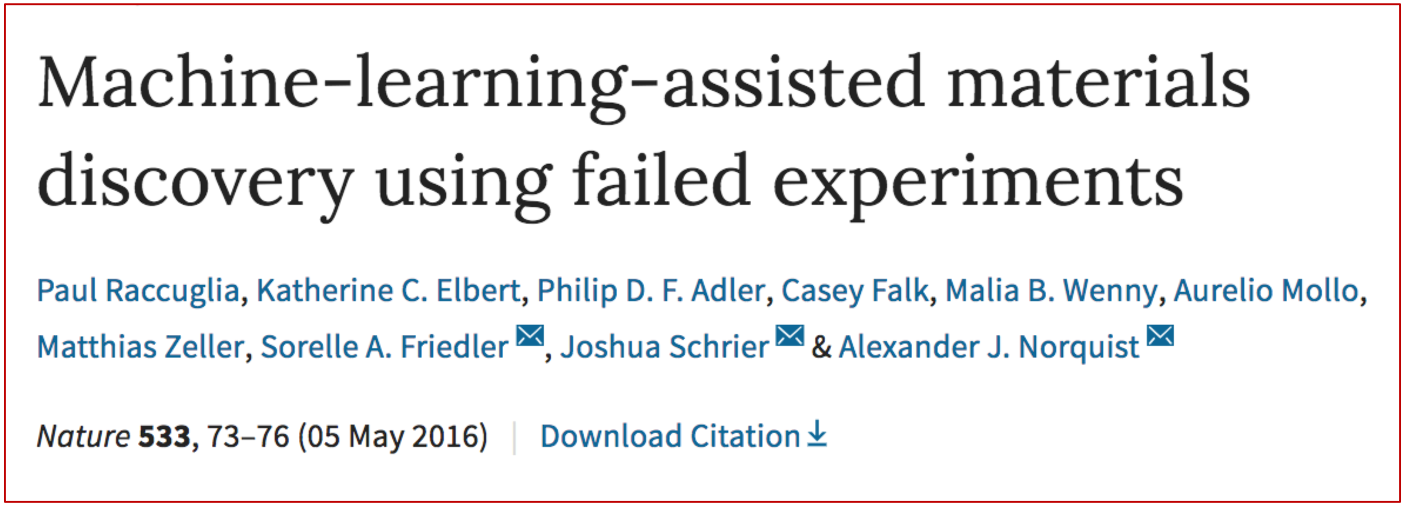
Authors Authors and affiliations

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

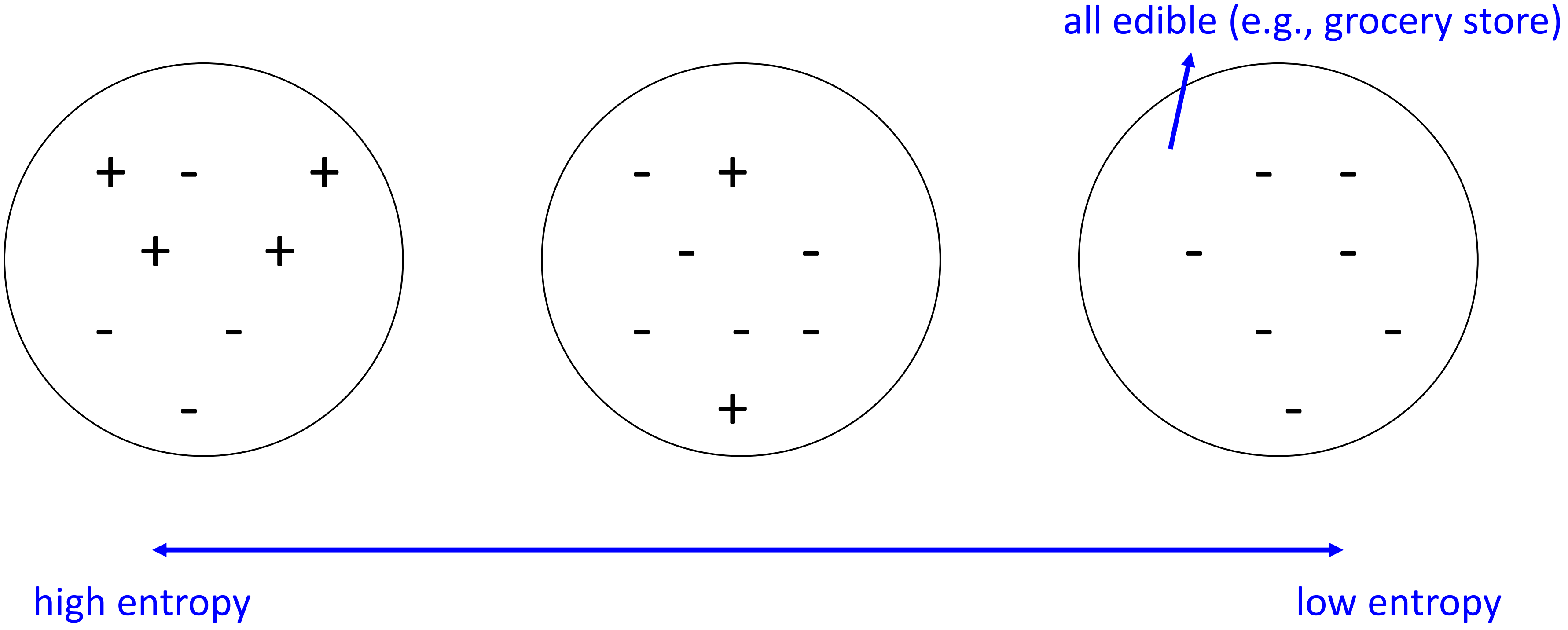
- Average # of bits needed to send one datapoint

Poisonous & edible mushrooms

Idea of Entropy

- Average # of bits needed to send one datapoint

Poisonous & edible mushrooms



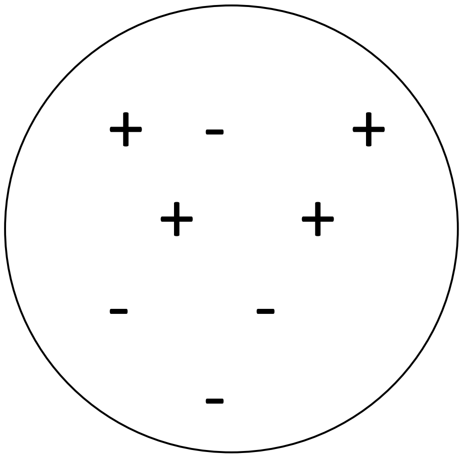
all edible (e.g., grocery store)

high entropy

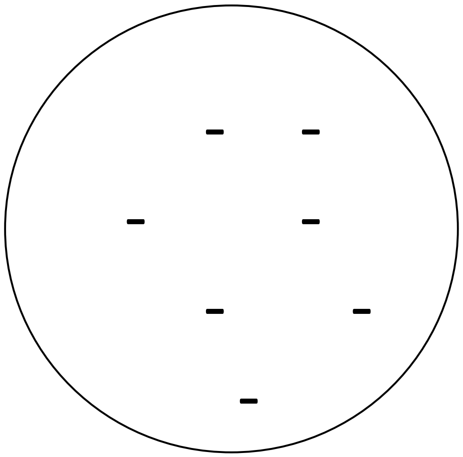
low entropy

Entropy

$$H(\underset{\overset{\uparrow}{\text{label}}}{y}) = -\sum_{c \in vals(y)} p(y = c) \underbrace{\log_2(p(y = c))}_{\text{\# of bits}}$$



H(y) = 1



H(y) = 0

Encoding data

Class year	Fixed-length encoding
senior	00
junior	01
sophomore	10
first year	11

Works!

Encoding data

Class year	Prob (p)
senior	0.5
junior	0.25
sophomore	0.125
first year	0.125

Idea: Use fewer bits to encode values that appear more often

Shannon Encoding

Class year	Prob (p)	Cumulative prob	Cumulative prob in binary
senior	0.5	0	0.000…
junior	0.25	0.5	0.100…
sophomore	0.125	0.75	0.110…
first year	0.125	0.875	0.111…

sort highest
to lowest

Decimal to binary conversion

- Multiply the decimal point number with 2
- Take note of the number *before* the decimal point in the result
- Multiply the result's value *after and including* the decimal point with 2
- Repeat until the result is 1
- Place the numbers we noted down after the decimal point in the order we got them

Shannon Encoding

ceiling
(round up)

Class year	Prob (p)	Cumulative prob	Binary	$\lceil -log_2 p \rceil$	Encoding
senior	0.5	0	0.000…	1	0
junior	0.25	0.5	0.100…	2	10
sophomore	0.125	0.75	0.110…	3	110
first year	0.125	0.875	0.111…	3	111

sort highest
to lowest

# of bits to use from
the binary form

$$H(class\ year) = 0.5 * 1 + 0.25 * 2 + 0.125 * 3 + 0.125 * 3 = 1.75$$

Outline for today

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

Conditional Entropy

- Quantifies the amount of information needed to describe the outcome of Y given X

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

feature
e.g., cap shape

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

single feature value
e.g., cap shape = bell

Information Gain

- Reduction in entropy/uncertainty given some information

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

- Select the feature that maximizes the information 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

$\mathrm{P(Li = yes)} =$ 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

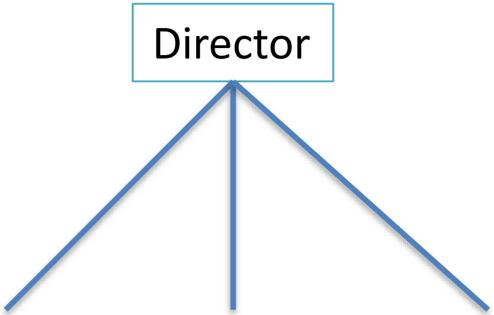
$\mathrm{H(Li \mid F)} = 0.85$

$\mathrm{Gain(Li,\ T)} =$ 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 today

- 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 today

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

Continuous Features

(do this for the TRAIN only!)

X	Y
10	Y
7	Y
8	N
3	Y
7	N
12	Y
2	Y

1) Sort examples based on given feature

2	3	7	7	8	10	12
Y	Y	Y	N	N	Y	Y

Continuous Features

(do this for the TRAIN only!)

X	Y
10	Y
7	Y
8	N
3	Y
7	N
12	Y
2	Y

1) Sort examples based on given feature

2	3	7	7	8	10	12
Y	Y	Y	N	N	Y	Y

2) Different label with same feature value, collapse to “None”

2	3	7	8	10	12
Y	Y	None	N	Y	Y

Continuous Features

(do this for the TRAIN only!)

X	Y
10	Y
7	Y
8	N
3	Y
7	N
12	Y
2	Y

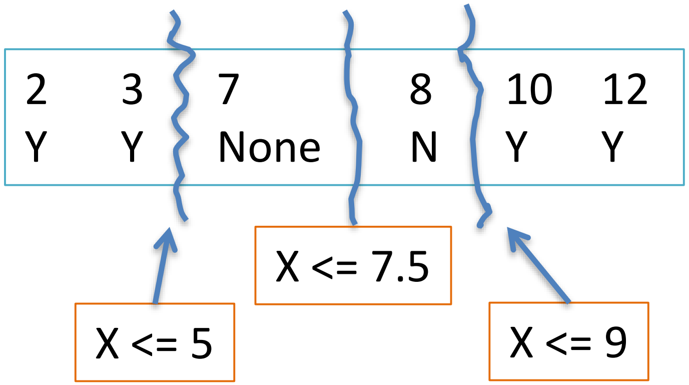
1) Sort examples based on given feature

2	3	7	7	8	10	12
Y	Y	Y	N	N	Y	Y

2) Different label with same feature value, collapse to "None"

2	3	7	8	10	12
Y	Y	None	N	Y	Y

3) Whenever label changes, make a feature (use avg)



2	3	7	8	10	12
Y	Y	None	N	Y	Y

X <= 5

X <= 7.5

X <= 9