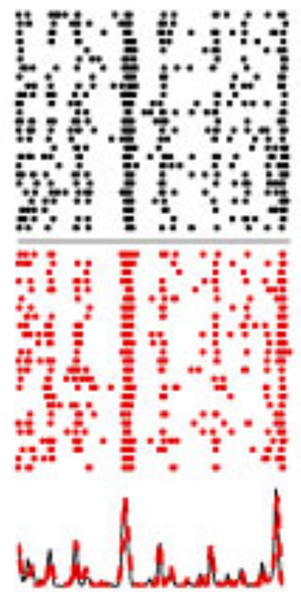
Information Theory part 1: entropy

NEU 314, Fall 2021
Lecture 17



Jonathan Pillow

practice problem

Consider the following model describing how a single neuron responds to houses and faces, which is given by a pair of conditional distributions:

# spikes	**0**	**1**	**2**	**3**	**4**
P(# spikes I "house")	0.1	0.1	0.3	0.4	0.1
P(# spikes I "face")	0.1	0.4	0.3	0.2	0

Furthermore, suppose P(house) = 0.4, P(face) = 0.6

1) What is the joint distribution P(# spikes, stimulus)?
2) What is the marginal distribution P(# spikes)?

bonus warmup problems for today:

- log(ab) = ?
- log (1/a) = ?

Information Theory

A mathematical theory of communication,
Claude Shannon 1948

Entropy

yes/no questions needed, on average, to determine the value of a random variable

motivating example #1: I'm thinking of a # between 1 and 8. How many Y/N questions do you need to guess it?

1 2 3 4 5 6 7 8

Strategy 1:

Q1: is it 1?

Q2: is it 2?

⋮

Q2: is it 8?

- worst case: 8 questions
- average case (assuming uniform): 4 questions

Can we do better???

motivating example #1: I'm thinking of a # between 1 and 8.
How many Y/N questions do you need to guess it?

1 2 3 4 5 6 7 8

Strategy 2: Q1: is it >4?

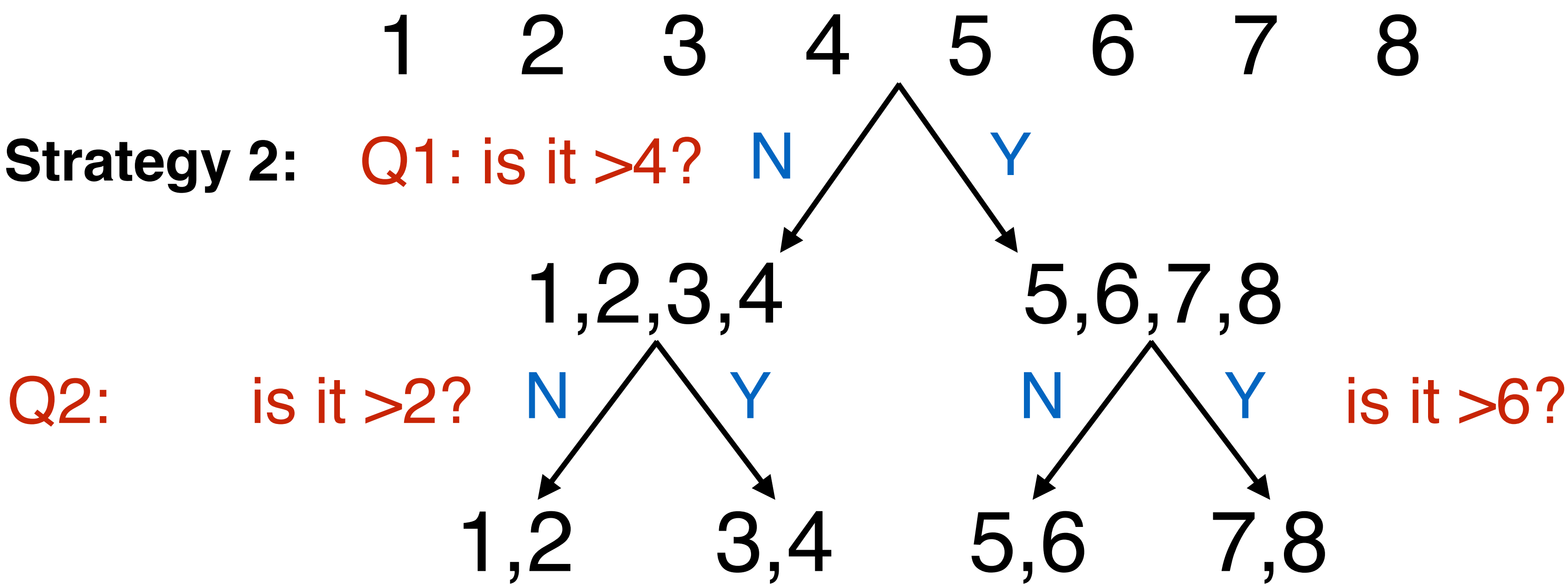
motivating example #1: I’m thinking of a # between 1 and 8. How many Y/N questions do you need to guess it?

1 2 3 4 5 6 7 8

Strategy 2: Q1: is it >4? N Y

1,2,3,4 5,6,7,8

motivating example #1: I'm thinking of a # between 1 and 8. How many Y/N questions do you need to guess it?



motivating example #1: I'm thinking of a # between 1 and 8.
How many Y/N questions do you need to guess it?

1 2 3 4 5 6 7 8

Strategy 2: Q1: is it >4? N Y

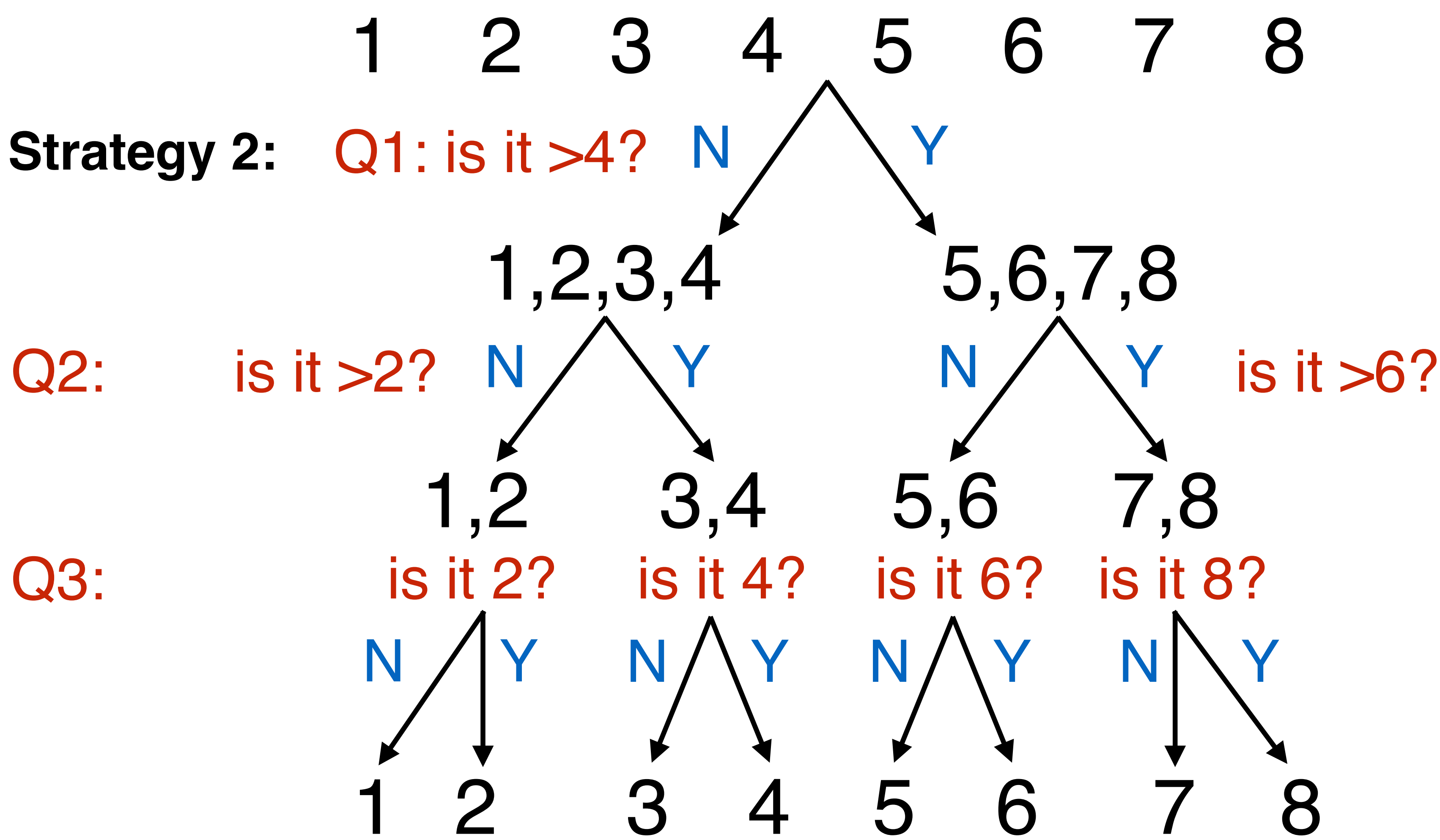
1,2,3,4 5,6,7,8

Q2: is it >2? N Y N Y is it >6?

1,2 3,4 5,6 7,8

Q3: is it 2? is it 4? is it 6? is it 8?

motivating example #1: I'm thinking of a # between 1 and 8. How many Y/N questions do you need to guess it?



- average: 3 questions

motivating example #1: I'm thinking of a # between 1 and 8.
How many Y/N questions do you need to guess it?

1 2 3 4 5 6 7 8

General result:

2^N options $\Longrightarrow$ N binary questions

Therefore: # questions = $\log_2(K)$

possibilities (unknowns)

motivating example #2: how many Y/N questions needed?

P(X)	0	0	0	1/8	1/8	1/8	1/8	1/2
	1	2	3	4	5	6	7	8

- 3 questions would still suffice
- But can we do better?

motivating example #2: how many Y/N questions needed?

P(X)	0	0	0	1/8	1/8	1/8	1/8	1/2
	1	2	3	4	5	6	7	8

Strategy: Q1: is it 8?

N → 4,5,6,7

Y → 8 (done!)

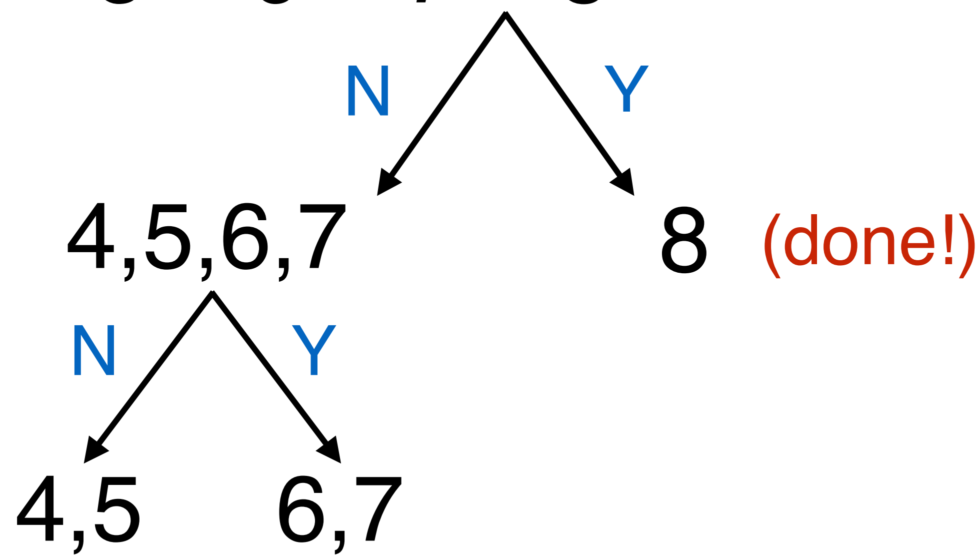
motivating example #2: how many Y/N questions needed?

P(X)	0	0	0	1/8	1/8	1/8	1/8	1/2
	1	2	3	4	5	6	7	8

Strategy:

Q1: is it 8?

Q2: is it >5?



motivating example #2: how many Y/N questions needed?

P(X)	0	0	0	1/8	1/8	1/8	1/8	1/2
	1	2	3	4	5	6	7	8

Strategy:

Q1: is it 8?

N: 4,5,6,7 | Y: 8 (done!)

Q2: is it >5?

N: 4,5 | Y: 6,7

Q3: is it 5? 4,5 | 6,7 is it 7?

N: 4 | Y: 5 | N: 6 | Y: 7

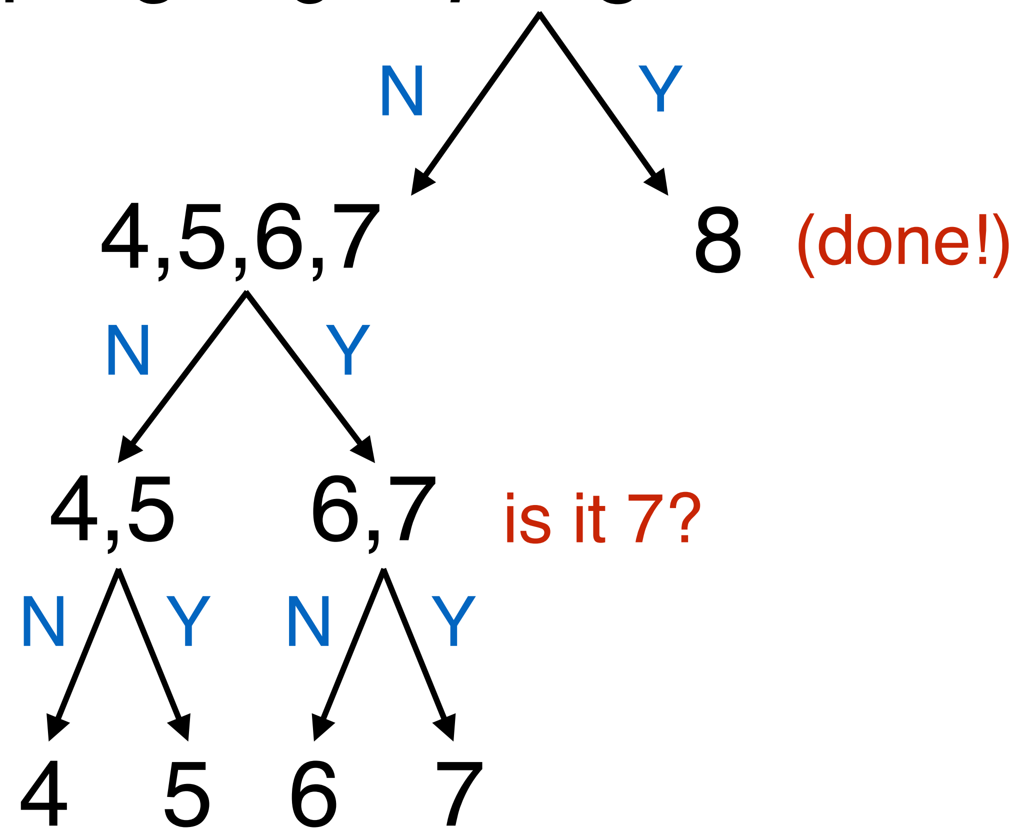
motivating example #2: how many Y/N questions needed?

P(X)	0	0	0	1/8	1/8	1/8	1/8	1/2
	1	2	3	4	5	6	7	8

Strategy: Q1: is it 8?

Q2: is it >5?

Q3: is it 5?



- what is the average # of questions?

1/2 (1 question) + 1/2 (3 questions) = 1/2 + 3/2 = 2 questions on average!

8 4,5,6,7

motivating example #2: how many Y/N questions needed?

P(X)	0	0	0	1/8	1/8	1/8	1/8	1/2
	1	2	3	4	5	6	7	8

General results:

- optimal strategy: divide probability in half w/ each question
- need N questions to identify options with probability $1/2^N$
- thus: $\log_2(K)$ questions for options with probability 1/K
- or: $\underbrace{-\log_2(p)}_{\text{code length}}$ questions for options with probability p

(using: $\log(K) = -\log(1/K) = -\log(p)$)

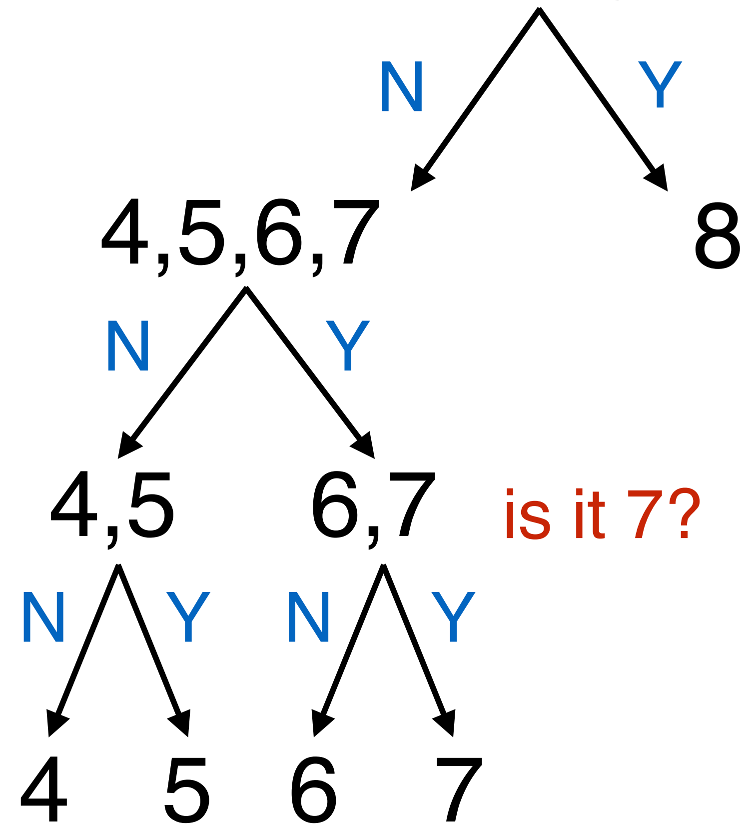
P(X)	0	0	0	1/8	1/8	1/8	1/8	1/2
	1	2	3	4	5	6	7	8

Strategy:

Q1: is it 8?

Q2: is it >5?

Q3: is it 5? is it 7?



code:

8: Y
7: NYY
6: NYN
5: NNY
4: NNN

- or: $\underbrace{-\log_2(p)}_{\textbf{code length}}$ questions for options with probability p

Entropy

$$H(x) = \boxed{-}\boxed{\sum_x p(x)}\boxed{\log p(x)}$$

averaged over p(x)

questions for x

- number of "yes/no" questions needed to identify x (on average)

exercises

Compute the entropy:

X	1	2	3	4	5	6	7	8
$P_1(X)$	1/4	0	0	1/2	1/4	0	0	0
$P_2(X)$	1/16	0	1/16	1/4	1/8	1/4	0	1/4
$P_1(X)$	0	0	0	0	0	0	1	0

Entropy

$$H(x) = \boxed{-}\boxed{\sum_x p(x)}\boxed{\log p(x)}$$

averaged over p(x)

questions for x

• number of “yes/no” questions needed to identify x (on average)

for distribution on K bins,

• maximum entropy = log K (achieved by uniform dist)
• minimum entropy = 0 (achieved by all probability in 1 bin)

What about when the probabilities aren't powers of 2?

X:	A	B
$P_1(X)$	1/3	2/3

formula still applies:

$$H(X) = - P(A) \log P(A) - P(B) \log P(B)$$

$$= -1/3 \log(1/3) - 2/3 \log(2/3)$$

≈ 0.91 questions "on average"

But how could you achieve that?

ANSWER: consider longer blocks of symbols

What about when the probabilities aren't powers of 2?

X:	A	B
$P_1(X)$	1/3	2/3

formula still applies:

$H(X) = -P(A)\log P(A) - P(B)\log P(B)$

$= -1/3 \log(1/3) - 2/3 \log(2/3)$

≈ 0.91 questions "on average"

AA	AB	BA	BB
1/9	2/9	2/9	4/9

$= 0.9444$ questions / symbol

Shannon showed: converges to entropy as you make the blocks longer

entropy: alternate derivation

Shannon: wanted a “surprise” function $h(\,\cdot\,)$ that had two properties:

- decreasing function $p(X)$
- the surprise of independent variables adds:

$$h(p(X, Y)) = h(p(X)) + h(p(Y))$$

if

$$p(X, Y) = p(X)p(Y)$$

Only function that has this property: $h(p) = -\log(p)$

- <u>entropy</u> = “average surprise” for values from P(X)

Conditional Entropy

$$H(x|y) \quad = \quad -\boxed{\sum_y p(y)}\boxed{\sum_x p(x|y)\log p(x|y)}$$

averaged over p(y)

entropy of x given some fixed value of y

Conditional Entropy

$$H(x|y) = -\boxed{\sum_y p(y)}\boxed{\sum_x p(x|y)\log p(x|y)}$$

averaged over p(y)

entropy of x given some fixed value of y

$$= -\sum_{x,y} p(x,y)\log p(x|y)$$

$$= H(x) \quad \text{if} \quad P(x,y) = P(x)P(y)$$

"On average, how uncertain are you about *x* if you know *y*?"

exercise

Compute the conditional entropy:

X	1	2	3	4	5	6	7	8
P(X|Y=0)	1/4	0	0	1/2	1/4	0	0	0
P(X|Y=1)	0	0	0	0	0	0	1	0

Y	0	1
P(Y)	2/3	1/3