Poll Everywhere Test

Linear regression is

A) Parametric

B) Non-parametric

K-NN is

A) Parametric

B) Non-parametric

- There are lots of office hours!!!!

Decision Trees
and Information Theory

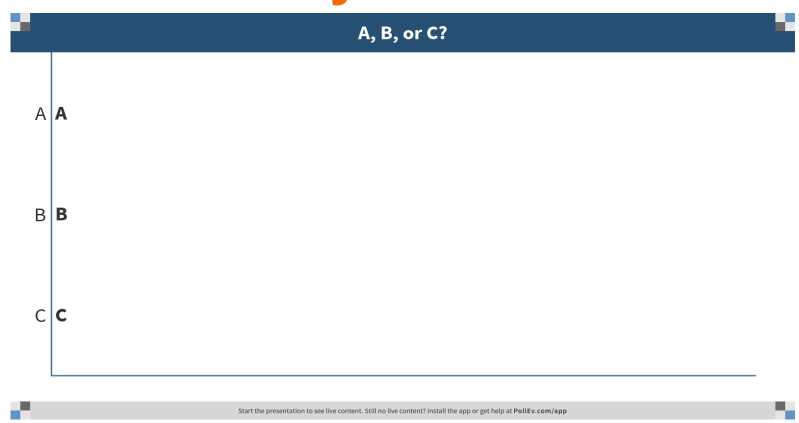
Lyle Ungar
University of Pennsylvania

What symptom tells you most about the disease?

S1	S2	S3	D
y	n	n	y
n	y	y	y
n	y	n	n
n	n	n	n
y	y	n	y

A) S1
B) S2
C) S3

Why?



What symptom tells you most about the disease?

S1/D

	y	**n**
y	2	0
n	1	2

S2/D

	y	**n**
y	2	1
n	1	1

s3/D

	y	**n**
y	1	0
n	2	2

A) S1
B) S2
C) S3

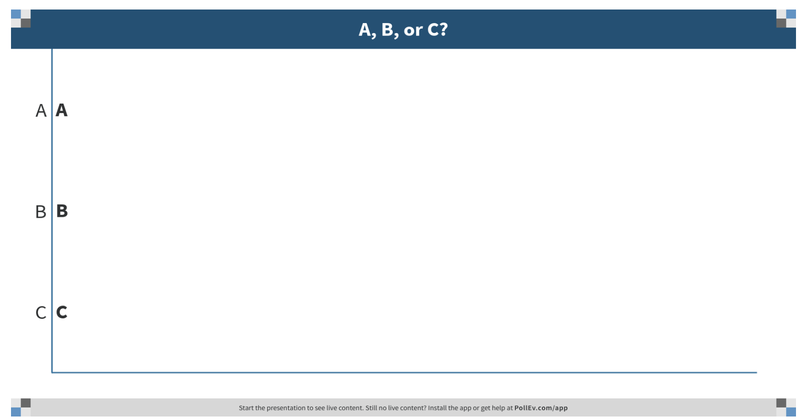
Why?

If you know S1=n, what symptom tells you most about the disease?

S1	S2	S3	D
y	n	n	y
n	y	y	y
n	y	n	n
n	n	n	n
y	y	n	y

A) S1
B) S2
C) S3

Why?



Resulting decision tree

```
      S1
   y/   \n
  D      S3
       y/   \n
       D     ~ D
```

The key question: what criterion to use do decide which question to ask?

Entropy and Information Gain

Andrew W. Moore

Carnegie Mellon University

www.cs.cmu.edu/~awm

awm@cs.cmu.edu

412-268-7599

modified by
Lyle Ungar

Bits

You observe a set of independent random samples of X

You see that X has four possible values

P(X=A) = 1/4	P(X=B) = 1/4	P(X=C) = 1/4	P(X=D) = 1/4

So you might see: BAACBADCDADDDA…

You transmit data over a binary serial link. You can encode each reading with two bits (e.g. A = 00, B = 01, C = 10, D = 11)

0100001001001110110011111100…

Fewer Bits

Someone tells you that the probabilities are not equal

P(X=A) = 1/2	P(X=B) = 1/4	P(X=C) = 1/8	P(X=D) = 1/8

It is possible to invent a coding for your transmission that only uses 1.75 bits on average per symbol. How?

Fewer Bits

Someone tells you that the probabilities are not equal

$P(X=A) = 1/2$	$P(X=B) = 1/4$	$P(X=C) = 1/8$	$P(X=D) = 1/8$

It is possible to invent a coding for your transmission that only uses 1.75 bits on average per symbol. How?

A	0
B	10
C	110
D	111

(This is just one of several ways)

Fewer Bits

Suppose there are three equally likely values…

P(X=A) = 1/3	P(X=B) = 1/3	P(X=C) = 1/3

Here's a naïve coding, costing 2 bits per symbol

A	00
B	01
C	10

Can you think of a coding that only needs1.6 bits per symbol on average?

In theory, it can in fact be done with 1.58496 bits per symbol.

General Case: Entropy

Suppose X can have one of *m* values… $V_1, V_2, \ldots V_m$

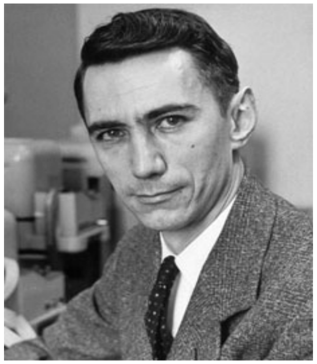
$P(X=V_1) = p_1$	$P(X=V_2) = p_2$	….	$P(X=V_m) = p_m$

What's the smallest possible number of bits, on average, per symbol, needed to transmit a stream of symbols drawn from X's distribution?

It is

$$H(X) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \ldots - p_m \log_2 p_m$$

$$= -\sum_{j=1}^{m} p_j \log_2 p_j$$



H(X) = The entropy of X

- **"High Entropy" means X is from a uniform (boring) distribution**
- **"Low Entropy" means X is from varied (peaks and valleys) distribution**

General Case

Suppose X can have one of *m* values… $V_1, V_2, \ldots V_m$

$P(X=V_1) = p_1$	$P(X=V_2) = p_2$	….	$P(X=V_m) = p_m$

What's the smallest possible number of bits, on average, per symbol, needed to transmit a stream of symbols drawn from X's distribution? It's

$$H(X) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \ldots - p_m \log_2 p_m$$

$$= -\sum_{j=1}^{m} p_j \log_2 p_j$$

A histogram of the frequency distribution of values of X would be flat

A histogram of the frequency distribution of values of X would have many lows and one or two highs

H(X) = The entropy of X

- **"High Entropy" means X is from a uniform (boring) distribution**
- **"Low Entropy" means X is from varied (peaks and valleys) distribution**

General Case

Suppose X can have one of *m* values… $V_1, V_2, \ldots V_m$

$P(X=V_1) = p_1$	$P(X=V_2) = p_2$	….	$P(X=V_m) = p_m$

What's the smallest possible number of bits, on average, per symbol, needed to transmit a stream of symbols drawn from X's distribution? It's

$$H(X) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \ldots - p_m \log_2 p_m$$

$$= -\sum_{j=1}^{m} p_j \log_2 p_j$$

A histogram of the frequency distribution of values of X would be flat

..and so the values sampled from it would be all over the place

A histogram of the frequency distribution of values of X would have many lows and one or two highs

..and so the values sampled from it would be more predictable

H(X) = The entropy of X

- **"High Entropy" means X is from a uniform (boring) distribution**
- **"Low Entropy" means X is from varied (peaks and valleys) distribution**

Entropy in a nut-shell

Low Entropy

High Entropy

Entropy in a nut-shell

Low Entro

..the values (locations of soup) sampled entirely from within the soup bowl

High Ent

..the values (locations of soup) unpredictable... almost uniformly sampled throughout our dining room

Why does entropy have this form?

$$H(X) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \ldots - p_m \log_2 p_m$$
$$= -\sum_{j=1}^{m} p_j \log_2 p_j$$

Entropy is the expected value of the information content (surprise) of the message $log_2 p_j$

If an event is certain, the entropy is

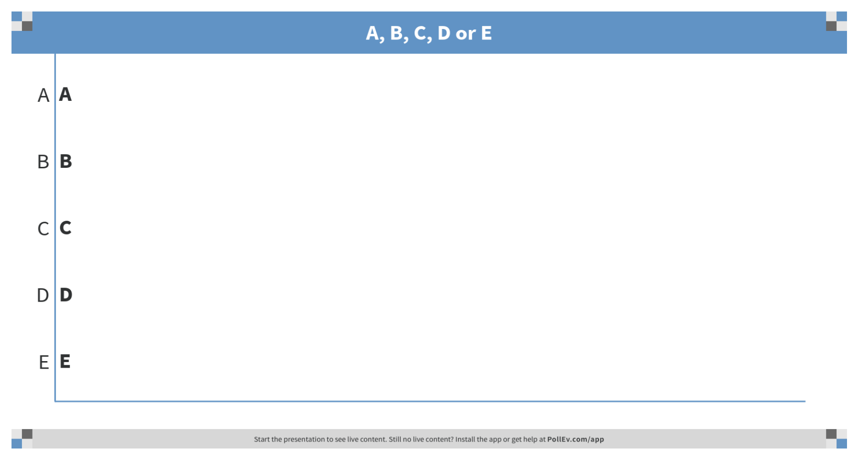
A) 0

B) between 0 and ½

C) ½

D) between ½ and 1

E) 1

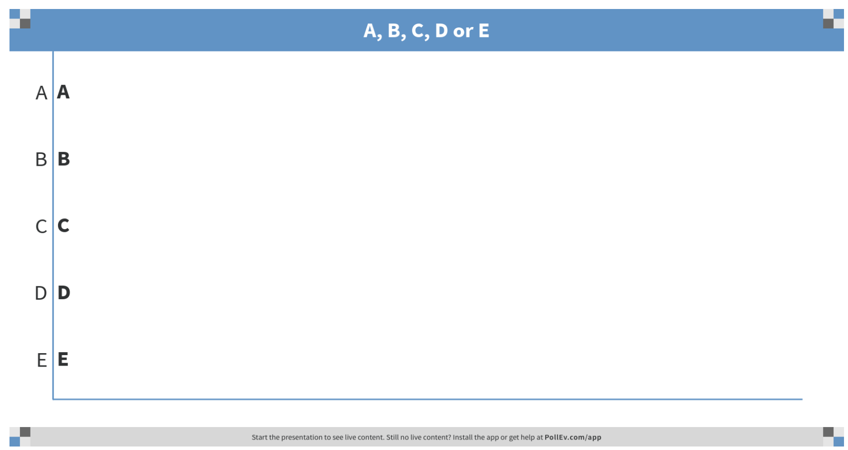


Why does entropy have this form?

$$H(X) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \ldots - p_m \log_2 p_m$$
$$= -\sum_{j=1}^{m} p_j \log_2 p_j$$

If two events are equally likely, the entropy is

A) 0
B) between 0 and ½
C) ½
D) between ½ and 1
E) 1



Specific Conditional Entropy H(Y|X=v)

Suppose I'm trying to predict output Y and I have input X

X = College Major

Y = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

Let's assume this reflects the true probabilities

e.g. From this data we estimate

- *P(LikeG = Yes) = 0.5*
- *P(Major = Math & LikeG = No) = 0.25*
- *P(Major = Math) = 0.5*
- *P(LikeG = Yes | Major = History) = 0*

Note:

- *H(X) = 1.5*
- *H(Y) = 1*

Specific Conditional Entropy H(Y|X=v)

X = College Major

Y = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

Definition of Specific Conditional Entropy:

$H(Y|X=v)$ = The entropy of Y among only those records in which X has value v

Specific Conditional Entropy H(Y|X=v)

X = College Major

Y = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

Definition of Specific Conditional Entropy:

$H(Y\,|X=v)$ = The entropy of Y among only those records in which X has value v

Example:

- $H(Y|X=Math) = 1$
- $H(Y|X=History) = 0$
- $H(Y|X=CS) = 0$

Conditional Entropy H(Y|X)

X = College Major

Y = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

Definition of Conditional Entropy:

$H(Y|X)$ = The average specific conditional entropy of Y

If you choose a record at random what will be the conditional entropy of Y, conditioned on that row's value of X

= Expected number of bits to transmit Y if both sides will know the value of X

$$= \Sigma_j \, Prob(X=v_j) \, H(Y \mid X = v_j)$$

Conditional Entropy

X = College Major

Y = Likes "Gladiator"

X	Y
Math	**Yes**
History	**No**
CS	**Yes**
Math	**No**
Math	**No**
CS	**Yes**
History	**No**
Math	**Yes**

Definition of Conditional Entropy:

$H(Y|X)$ = The average conditional entropy of Y

$$= \Sigma_j Prob(X=v_j)\ H(Y \mid X = v_j)$$

Example:

v_j	$Prob(X=v_j)$	$H(Y \mid X = v_j)$
Math	0.5	1
History	0.25	0
CS	0.25	0

$$H(Y|X) = 0.5 * 1 + 0.25 * 0 + 0.25 * 0 = 0.5$$

Information Gain

X = College Major

Y = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

Definition of Information Gain:

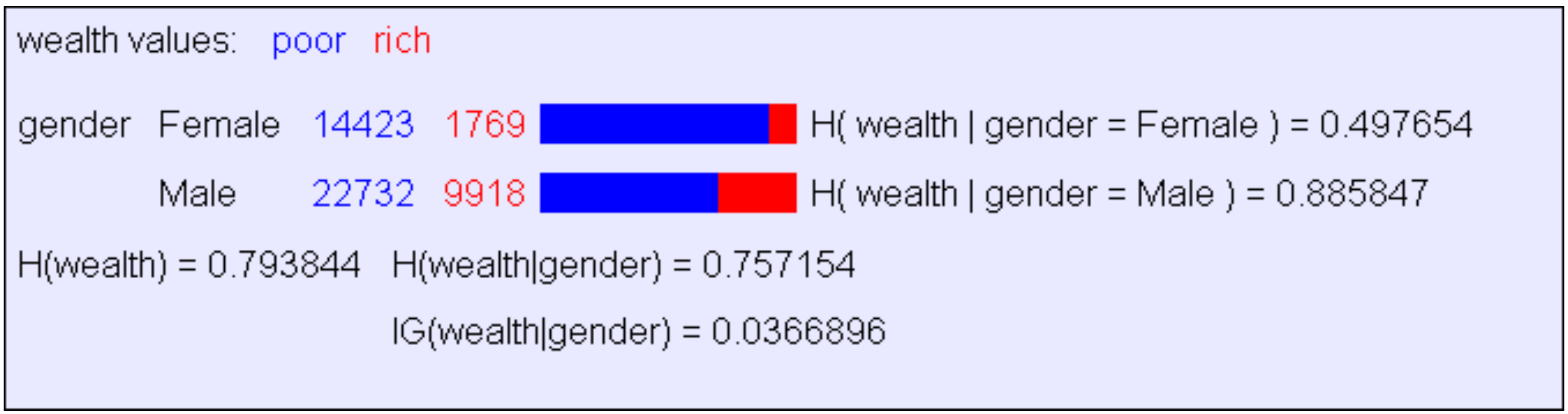
$IG(Y|X)$ = I must transmit Y. How many bits on average would it save me if both ends of the line knew X?

$IG(Y|X) = H(Y) - H(Y|X)$

Example:

- $H(Y) = 1$
- $H(Y|X) = 0.5$
- Thus $IG(Y|X) = 1 – 0.5 = 0.5$

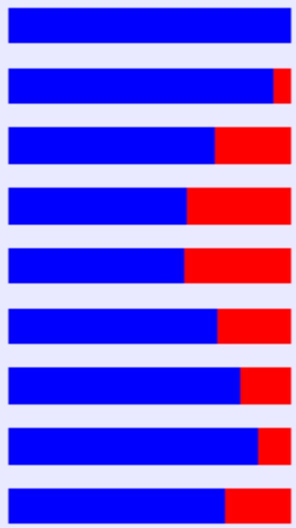
Information Gain Example



Another example

wealth values: poor rich

		poor	rich	
agegroup	10s	2507	3	H(wealth | agegroup = 10s) = 0.0133271
	20s	11262	743	H(wealth | agegroup = 20s) = 0.334906
	30s	9468	3461	H(wealth | agegroup = 30s) = 0.838134
	40s	6738	3986	H(wealth | agegroup = 40s) = 0.951961
	50s	4110	2509	H(wealth | agegroup = 50s) = 0.957376
	60s	2245	809	H(wealth | agegroup = 60s) = 0.834049
	70s	668	147	H(wealth | agegroup = 70s) = 0.680882
	80s	115	16	H(wealth | agegroup = 80s) = 0.535474
	90s	42	13	H(wealth | agegroup = 90s) = 0.788941



H(wealth) = 0.793844 H(wealth|agegroup) = 0.709463

IG(wealth|agegroup) = 0.0843813

What is Information Gain used for?

If you are going to collect information from someone (e.g. asking questions sequentially in a decision tree), the “best” question is the one with the highest information gain.

Information gain is useful for model selection (later!)

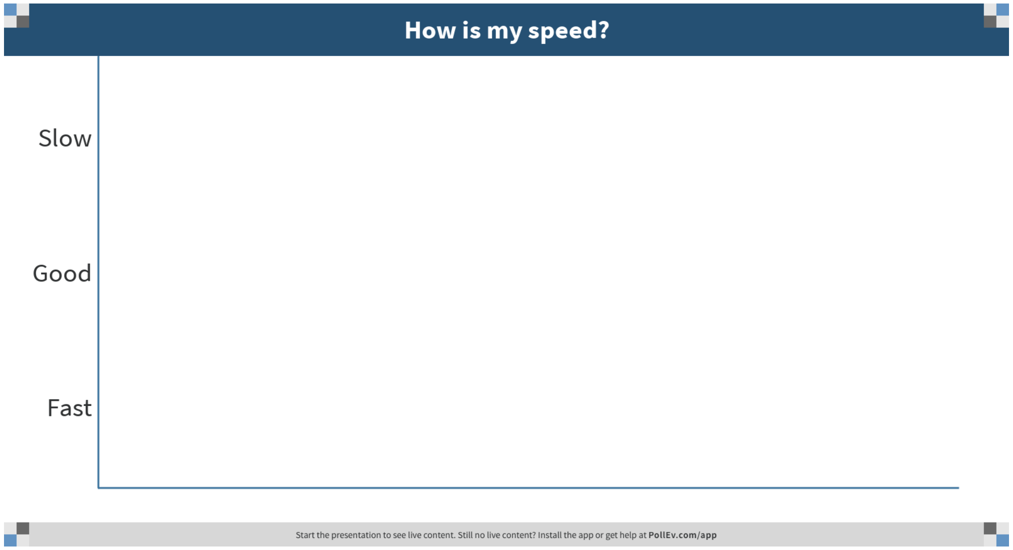
What question did we not ask (or answer) about decision trees?

What you should know

- **Entropy**
- **Information Gain**
- **The standard decision tree algorithm**
  - Recursive partition trees
  - Also called: ID3/C4.5/CART/CHAID

How is my speed?

- A) Slow
- B) Good
- C) Fast



What one thing

- Do you like about the course so far?
- Would you improve about the course so far?

What one thing …

Start the presentation to activate live content

If you see this message in presentation mode, install the add-in or get help at PollEv.com/app