Poll Everywhere Test

Linear regression is
A) Parametric
B) Non-parametric

K-NN is
A) Parametric
B) Non-parametric

- HW0 due today
- HW1 released today
- 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?**

<table>
<thead>
<tr>
<th></th>
<th>S1/D</th>
<th>S2/D</th>
<th>s3/D</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>n</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>n</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

A) S1
B) S2
C) S3

**Why?**
If you know $S_1=n$, what symptom tells you most about the disease?

<table>
<thead>
<tr>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>n</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>n</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>n</td>
<td>y</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>y</td>
<td>y</td>
<td>n</td>
<td>y</td>
</tr>
</tbody>
</table>

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

Copyright © 2001, 2003, Andrew W. Moore
Bits

You observe a set of independent random samples of X

You see that X has four possible values

\[
\begin{array}{|c|c|c|c|}
\hline
X=A & X=B & X=C & X=D \\
\hline
P(X=A) = 1/4 & P(X=B) = 1/4 & P(X=C) = 1/4 & P(X=D) = 1/4 \\
\hline
\end{array}
\]

So you might see: BAACBADCADDAD…

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)

010000100100111011001111111100…

Copyright © 2001, 2003, Andrew W. Moore
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

<table>
<thead>
<tr>
<th>P(X=A) = 1/2</th>
<th>P(X=B) = 1/4</th>
<th>P(X=C) = 1/8</th>
<th>P(X=D) = 1/8</th>
</tr>
</thead>
</table>

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

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>10</td>
</tr>
<tr>
<td>C</td>
<td>110</td>
</tr>
<tr>
<td>D</td>
<td>111</td>
</tr>
</tbody>
</table>

(This is just one of several ways)
Fewer Bits

Suppose there are three equally likely values...

<table>
<thead>
<tr>
<th>X</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1/3</td>
</tr>
<tr>
<td>B</td>
<td>1/3</td>
</tr>
<tr>
<td>C</td>
<td>1/3</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>X</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>00</td>
</tr>
<tr>
<td>B</td>
<td>01</td>
</tr>
<tr>
<td>C</td>
<td>10</td>
</tr>
</tbody>
</table>

Can you think of a coding that only needs 1.6 bits per symbol on average?

In theory, it can in fact be done with 1.58496 bits per symbol.
Suppose $X$ can have one of $m$ values... $V_1, V_2, \ldots, V_m$.

| $P(X=V_1) = p_1$ | $P(X=V_2) = p_2$ | $\ldots$ | $P(X=V_m) = p_m$ |

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

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

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

$H(X) = \text{The entropy of } X$

- “High Entropy” means $X$ is from a uniform (boring) distribution
- “Low Entropy” means $X$ is from varied (peaks and valleys) distribution
Suppose X can have one of $m$ values... $V_1, V_2, \ldots, V_m$.

<table>
<thead>
<tr>
<th>$P(X=V_1) = p_1$</th>
<th>$P(X=V_2) = p_2$</th>
<th>$\ldots$</th>
<th>$P(X=V_m) = p_m$</th>
</tr>
</thead>
</table>

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

It's

$$H(X) = \sum_{j=1}^{m} p_j \log_2 p_j$$

$H(X)$ = 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

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

| $P(X=V_1) = p_1$ | $P(X=V_2) = p_2$ | $\ldots$ | $P(X=V_m) = p_m$ |

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

$$H(X) = -\sum_{j=1}^{m} p_j \log p_j$$

$H(X)$ = 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

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
Entropy in a nut-shell

Low Entropy

High Entropy

Copyright © 2001, 2003, Andrew W. Moore
Entropy in a nut-shell

Low Entropy

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

High Entropy

..the values (locations of soup) unpredictable... almost uniformly sampled throughout our dining room
Why does entropy have this form?

\[ H(X) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \cdots - p_m \log_2 p_m \]

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

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 \( \frac{1}{2} \)
C) \( \frac{1}{2} \)
D) between \( \frac{1}{2} \) 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 $X$

- $X = \text{College Major}$
- $Y = \text{Likes “Gladiator”}$

Let’s assume this reflects the true probabilities

- $P(\text{LikeG} = \text{Yes}) = 0.5$
- $P(\text{Major} = \text{Math} \& \text{LikeG} = \text{No}) = 0.25$
- $P(\text{Major} = \text{Math}) = 0.5$
- $P(\text{LikeG} = \text{Yes} | \text{Major} = \text{History}) = 0$

Note:
- $H(X) = 1.5$
- $H(Y) = 1$
Specific Conditional Entropy $H(Y|X=v)$

$X = \text{College Major}$
$Y = \text{Likes “Gladiator”}$

*Definition of Specific Conditional Entropy:*

$H(Y | X=v) = \text{The entropy of } Y \text{ among only those records in which } X \text{ has value } v$
Definition of Specific Conditional Entropy:

\[ H(Y | X=v) = \text{The entropy of } Y \text{ among only those records in which } X \text{ has value } v \]

**Example:**

- \( H(Y|X=\text{Math}) = 1 \)
- \( H(Y|X=\text{History}) = 0 \)
- \( H(Y|X=CS) = 0 \)
Conditional Entropy $H(Y|X)$

$X = \text{College Major}$

$Y = \text{Likes “Gladiator”}$

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math</td>
<td>Yes</td>
</tr>
<tr>
<td>History</td>
<td>No</td>
</tr>
<tr>
<td>CS</td>
<td>Yes</td>
</tr>
<tr>
<td>Math</td>
<td>No</td>
</tr>
<tr>
<td>Math</td>
<td>No</td>
</tr>
<tr>
<td>CS</td>
<td>Yes</td>
</tr>
<tr>
<td>History</td>
<td>No</td>
</tr>
<tr>
<td>Math</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Definition of Conditional Entropy:

$H(Y \mid X) = \text{The average specific conditional entropy of } Y$

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

$= \text{Expected number of bits to transmit } Y \text{ if both sides will know the value of } X$

$= \sum_j \text{Prob}(X=v_j) \cdot H(Y \mid X = v_j)$
Conditional Entropy

X = College Major
Y = Likes “Gladiator”

Definition of Conditional Entropy:

\[ H(Y|X) = \text{The average conditional entropy of } Y \]

\[ = \sum_j \text{Prob}(X=v_j) \cdot H(Y | X = v_j) \]

Example:

| \( v_j \) | \text{Prob}(X=v_j) | H(Y | X = v_j) |
|---|---|---|
| Math | 0.5 | 1 |
| History | 0.25 | 0 |
| CS | 0.25 | 0 |

\[ H(Y|X) = 0.5 \times 1 + 0.25 \times 0 + 0.25 \times 0 = 0.5 \]
Information Gain

X = College Major
Y = Likes “Gladiator”

Definition of Information Gain:

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

\[ IG(Y|X) = H(Y) - H(Y | X) \]

Example:

- H(Y) = 1
- H(Y|X) = 0.5
- Thus \( IG(Y|X) = 1 - 0.5 = 0.5 \)
Information Gain Example

wealth values: poor rich

| gender | Female | 14423 | 1769 | H(wealth | gender = Female) = 0.497654 |
|--------|--------|-------|------|--------------------------------|
|        | Male   | 22732 | 9918 | H(wealth | gender = Male) = 0.885847    |

H(wealth) = 0.793844  H(wealth|gender) = 0.757154

IG(wealth|gender) = 0.0366896
Another example

<table>
<thead>
<tr>
<th>agegroup</th>
<th>poor</th>
<th>rich</th>
</tr>
</thead>
<tbody>
<tr>
<td>10s</td>
<td>2507</td>
<td>3</td>
</tr>
<tr>
<td>20s</td>
<td>11262</td>
<td>743</td>
</tr>
<tr>
<td>30s</td>
<td>9468</td>
<td>3461</td>
</tr>
<tr>
<td>40s</td>
<td>6738</td>
<td>3986</td>
</tr>
<tr>
<td>50s</td>
<td>4110</td>
<td>2509</td>
</tr>
<tr>
<td>60s</td>
<td>2245</td>
<td>809</td>
</tr>
<tr>
<td>70s</td>
<td>668</td>
<td>147</td>
</tr>
<tr>
<td>80s</td>
<td>115</td>
<td>16</td>
</tr>
<tr>
<td>90s</td>
<td>42</td>
<td>13</td>
</tr>
</tbody>
</table>

\[ H(\text{wealth} \mid \text{agegroup} = 10s) = 0.0133271 \]
\[ H(\text{wealth} \mid \text{agegroup} = 20s) = 0.334906 \]
\[ H(\text{wealth} \mid \text{agegroup} = 30s) = 0.838134 \]
\[ H(\text{wealth} \mid \text{agegroup} = 40s) = 0.951961 \]
\[ H(\text{wealth} \mid \text{agegroup} = 50s) = 0.957376 \]
\[ H(\text{wealth} \mid \text{agegroup} = 60s) = 0.834049 \]
\[ H(\text{wealth} \mid \text{agegroup} = 70s) = 0.680882 \]
\[ H(\text{wealth} \mid \text{agegroup} = 80s) = 0.535474 \]
\[ H(\text{wealth} \mid \text{agegroup} = 90s) = 0.788941 \]

\[ \text{H(wealth)} = 0.793844 \]
\[ \text{H(wealth}\mid\text{agegroup)} = 0.709463 \]
\[ \text{IG(wealth}\mid\text{agegroup)} = 0.0843813 \]
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?