Information Gain

Andrew W. Moore
Professor
School of Computer Science
Carnegie Mellon University

www.cs.cmu.edu/~awm
awm@cs.cmu.edu
412-268-7599

Note to other teachers and users of these slides. Andrew would be delighted if you found this source material useful in giving your own lectures. Feel free to use these slides verbatim, or to modify them to fit your own needs. PowerPoint originals are available. If you make use of a significant portion of these slides in your own lecture, please include this message, or the following link to the source repository of Andrew’s tutorials: http://www.cs.cmu.edu/~awm/tutorials. Comments and corrections gratefully received.

Information Gain

Andrew W. Moore
Professor
School of Computer Science
Carnegie Mellon University

www.cs.cmu.edu/~awm
awm@cs.cmu.edu
412-268-7599

Copyright © 2001, 2003, Andrew W. Moore

Information Gain: Slide 2

Bits

You are watching 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: BAACBACDADDDA...
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)

010001001001011010011111100...
Fewer Bits
Someone tells you that the probabilities are not equal

\[ P(X=A) = \frac{1}{2} \quad P(X=B) = \frac{1}{4} \quad P(X=C) = \frac{1}{8} \quad P(X=D) = \frac{1}{8} \]

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

(This is just one of several ways)
Fewer Bits

Suppose there are three equally likely values...

\[
\begin{align*}
P(X=A) &= 1/3 \\
P(X=B) &= 1/3 \\
P(X=C) &= 1/3
\end{align*}
\]

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

<table>
<thead>
<tr>
<th>Symbol</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 would need only 1.6 bits per symbol on average?

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

---

General Case

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

\[
P(X=V_j) = p_j
\]

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

\[
H(X) = - \sum_{j=1}^{m} p_j \log_2 p_j
\]

\( H(X) \) = 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
Suppose X can have one of \( m \) values... \( V_1, V_2, \ldots V_m \)

What's the smallest possible number of bits, on average, per symbol, \( X \)'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 have many lows and one or two highs

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

..and so the values sampled from it would be more predictable
Entropy in a nut-shell

Low Entropy

High Entropy

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

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

E.G. From this data we estimate

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

Note:

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

Specific Conditional Entropy $H(Y|X=v)$

Definition of Specific Conditional Entropy:

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

<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>
Specific Conditional Entropy $H(Y|X=v)$

**X = College Major**
**Y = Likes “Gladiator”**

**Definition of Specific Conditional Entropy:**

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

**Example:**
- $H(Y|X=\text{Math}) = 1$
- $H(Y|X=\text{History}) = 0$
- $H(Y|X=\text{CS}) = 0$

<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>

Conditional Entropy $H(Y|X)$

**X = College Major**
**Y = Likes “Gladiator”**

**Definition of Conditional Entropy:**

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

= if you choose a record at random what will be the conditional entropy of $Y_i$ conditioned on that row’s value of $X$

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

$$= \sum_j Prob(X=v_j) \cdot H(Y | X = v_j)$$
Conditional Entropy

**Definition of Conditional Entropy:**

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

\[ = \sum \text{Prob}(X = v_j) \cdot H(Y \mid X = v_j) \]

**Example:**

<table>
<thead>
<tr>
<th>( v_j )</th>
<th>( \text{Prob}(X = v_j) )</th>
<th>( H(Y \mid X = v_j) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>History</td>
<td>0.25</td>
<td>0</td>
</tr>
<tr>
<td>CS</td>
<td>0.25</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ H(Y \mid X) = 0.5 \times 1 + 0.25 \times 0 + 0.25 \times 0 = 0.5 \]

Information Gain

**Definition of Information Gain:**

\[ IG(Y \mid X) = \text{I must transmit } Y. \]

How many bits on average would it save me if both ends of the line knew \( X \)?

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

**Example:**

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

<table>
<thead>
<tr>
<th>gender</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>14423</td>
<td>1750</td>
<td>H(\text{wealth</td>
</tr>
<tr>
<td>Male</td>
<td>22732</td>
<td>5918</td>
<td>H(\text{wealth</td>
</tr>
</tbody>
</table>

\(H(\text{wealth}) = 0.793844\), \(H(\text{wealth | gender}) = 0.757154\), \(\text{IG(wealth | gender)} = 0.0365896\)

Another example

<table>
<thead>
<tr>
<th>agegroup</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10s</td>
<td>2507</td>
<td>3</td>
<td>H(\text{wealth</td>
</tr>
<tr>
<td>20s</td>
<td>11262</td>
<td>743</td>
<td>H(\text{wealth</td>
</tr>
<tr>
<td>30s</td>
<td>9488</td>
<td>3461</td>
<td>H(\text{wealth</td>
</tr>
<tr>
<td>40s</td>
<td>6738</td>
<td>3986</td>
<td>H(\text{wealth</td>
</tr>
<tr>
<td>50s</td>
<td>4110</td>
<td>2500</td>
<td>H(\text{wealth</td>
</tr>
<tr>
<td>60s</td>
<td>2245</td>
<td>809</td>
<td>H(\text{wealth</td>
</tr>
<tr>
<td>70s</td>
<td>668</td>
<td>147</td>
<td>H(\text{wealth</td>
</tr>
<tr>
<td>80s</td>
<td>115</td>
<td>16</td>
<td>H(\text{wealth</td>
</tr>
<tr>
<td>90s</td>
<td>42</td>
<td>13</td>
<td>H(\text{wealth</td>
</tr>
</tbody>
</table>

\(H(\text{wealth}) = 0.793844\), \(H(\text{wealth | agegroup}) = 0.709483\), \(\text{IG(wealth | agegroup)} = 0.0843013\)
Relative Information Gain

X = College Major
Y = Likes “Gladiator”

Definition of Relative Information Gain:

\[ RIG(Y \mid X) = \text{I must transmit } Y, \text{ what fraction of the bits on average would it save me if both ends of the line knew } X? \]

\[ RIG(Y \mid X) = H(Y) - H(Y \mid X) / H(Y) \]

Example:

- \( H(Y \mid X) = 0.5 \)
- \( H(Y) = 1 \)
- Thus \( IG(Y \mid X) = (1 - 0.5) / 1 = 0.5 \)

<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>

What is Information Gain used for?

Suppose you are trying to predict whether someone is going live past 80 years. From historical data you might find...

- \( IG(\text{LongLife} \mid \text{HairColor}) = 0.01 \)
- \( IG(\text{LongLife} \mid \text{Smoker}) = 0.2 \)
- \( IG(\text{LongLife} \mid \text{Gender}) = 0.25 \)
- \( IG(\text{LongLife} \mid \text{LastDigitOfSSN}) = 0.00001 \)

IG tells you how interesting a 2-d contingency table is going to be.