Information Entropy: Illustrating Example

Andrew Kusiak
Intelligent Systems Laboratory
2139 Seamans Center
The University of Iowa
Iowa City, Iowa 52242 - 1527

andrew-kusiak@uiowa.edu
http://www.icaen.uiowa.edu/~ankusiak
Tel: 319 - 335 5934        Fax: 319-335 5669

Etymology of “Entropy”

Origin in chemistry

- Entropy = randomness

Information theory

- Amount of uncertainty

Shannon Entropy

- \( S \) = final probability space composed of two disjoint events \( E_1 \) and \( E_2 \) with probability \( p_1 = p \) and \( p_2 = 1 - p \), respectively.
- The Shannon entropy is defined as

\[
H(S) = H(p_1, p_2) = - p \log p - (1 - p) \log (1 - p)
\]

Single term only

Definitions

Information content

\[
I(s_1, s_2, \ldots, s_m) = S \sum_{i=1}^{m} \frac{s_i}{S} \log \frac{s_i}{S}
\]

Entropy

\[
E(A) = \sum_{i=1}^{m} \frac{s_j + \ldots + s_m}{S} I(s_1, \ldots, s_m)
\]

Information gain

\[
\text{Gain}(A) = I(s_1, s_2, \ldots, s_m) - E(A)
\]
### Case 1

**Example**

<table>
<thead>
<tr>
<th>No.</th>
<th>F1</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Blue</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Blue</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Blue</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>Blue</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>Red</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>Red</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>Red</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>Red</td>
<td>2</td>
</tr>
</tbody>
</table>

For Blue \(D_{1B} = 4, D_{2B} = 0\)

\[
I(D_{1B}, D_{2B}) = -4/4 \log_2 (4/4) = 0
\]

For Red \(D_{1R} = 0, D_{2R} = 4\)

\[
I(D_{1R}, D_{2R}) = -4/4 \log_2 (4/4) = 0
\]

\[
E(F1) = 4/8 I(D_{1B}, D_{2B}) + 4/8 I(D_{1R}, D_{2R}) = 0
\]

\[
\text{Gain} (F1) = I(D_1, D_2) - E(F1) = 1
\]

\[
E(A) = \sum_{x=1}^{m} \frac{-e^{-\frac{4}{4}}}{x} \log_2 \left( \frac{e^{-\frac{4}{4}}}{x} \right)
\]

### Case 2

**Example**

\[
I(D_1, D_2, D_3) = -\frac{2}{8} \log_2 (2/8) - 3/8 \log_2 (3/8) - 3/8 \log_2 (3/8) = 1.56
\]

For Blue \(D_{1B} = 2, D_{2B} = 2, D_{3B} = 0\)

\[
I(D_{1B}, D_{2B}) = -\frac{2}{4} \log_2 (2/4) - \frac{2}{4} \log_2 (2/4) = 0
\]

For Red \(D_{1R} = 0, D_{2R} = 1, D_{3R} = 3\)

\[
I(D_{2R}, D_{3R}) = -\frac{1}{4} \log_2 (1/4) - \frac{3}{4} \log_2 (3/4) = 0.81
\]

\[
E(F1) = 4/8 I(D_{1B}, D_{2B}) + 4/8 I(D_{2R}, D_{3R}) = 0.905
\]

\[
\text{Gain} (F1) = I(D_1, D_2, D_3) - E(F1) = 0.655
\]

### Case 3

**Example**

\[
I(D_1, D_2, D_3) = -1/8 \log_2 (1/8) - 3/8 \log_2 (3/8) - 4/8 \log_2 (4/8) = 1.41
\]

For Blue \(D_{1B} = 1, D_{2B} = 3, D_{3B} = 0\)

\[
I(D_{1B}, D_{2B}) = -1/4 \log_2 (1/4) - 3/4 \log_2 (3/4) = 0.81
\]

For Red \(D_{1R} = 0, D_{2R} = 0, D_{3R} = 4\)

\[
I(D_{1R}, D_{2R}) = -4/4 \log_2 (4/4) = 0
\]

\[
E(F1) = 4/8 I(D_{1R}, D_{2R}) + 4/8 I(D_{3R}) = 0.41
\]

\[
\text{Gain} (F1) = I(D_1, D_2, D_3) - E(F1) = 1
\]

### Case 4

**Example**

\[
I(D_1, D_2, D_3) = -\frac{2}{8} \log_2 (2/8) - 3/8 \log_2 (3/8) - 3/8 \log_2 (3/8) = 1.56
\]

For Blue \(D_{1B} = 2, D_{2B} = 0, D_{3B} = 0\)

\[
I(D_{1B}, D_{2B}) = -2/2 \log_2 (2/2) = 0
\]

For Red \(D_{1R} = 0, D_{2R} = 3, D_{3R} = 0\)

\[
I(D_{1R}, D_{2R}) = -3/3 \log_2 (3/3) = 0
\]

For Green \(D_{1G} = 0, D_{2G} = 0, D_{3G} = 3\)

\[
I(D_{1G}, D_{2G}) = -3/3 \log_2 (3/3) = 0
\]

\[
E(F1) = 2/8 I(D_{1B}, D_{2B}) + 3/8 I(D_{2R}, D_{3R}) = 0.905
\]

\[
\text{Gain} (F1) = I(D_1, D_2, D_3) - E(F1) = 0.655
\]
Example

<table>
<thead>
<tr>
<th>No.</th>
<th>FI</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Blue</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Green</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Green</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>Green</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>Green</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>Green</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>Red</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>Red</td>
<td>2</td>
</tr>
</tbody>
</table>

For Blue
\[ I(D_{1B}) = -\frac{1}{1} \log_2 \left( \frac{1}{1} \right) = 0 \]
For Red
\[ I(D_{2R}) = -\frac{2}{2} \log_2 (\frac{2}{2}) = 0 \]
For Green
\[ I(D_{1G}, D_{3G}, D_{4G}) = -\frac{1}{5} \log_2 (\frac{1}{5}) - \frac{2}{5} \log_2 (\frac{2}{5}) - \frac{2}{5} \log_2 (\frac{2}{5}) = 1.52 \]

\[ I(D_1, D_2, D_3) = \frac{-2}{8} \log_2 (\frac{2}{8}) \]

\[ E(F_1) = \frac{1}{8} I(D_{1B}) + \frac{2}{8} I(D_{2R}) + \frac{5}{8} I(D_{1G}, D_{3G}, D_{4G}) = 0.95 \]

\[ \text{Gain (F1)} = I(D_1, D_2, D_3) - E(F_1) = 1.05 \]

Summary

The higher the information gain, the more relevant is the observed feature (parameter) to the decision.

The lower the entropy, the more relevant is the feature to the decision.
Entropy: A Measure of Homogeneity

Set S of N objects

Information Content of S

\[ I(S) = - p_+ \log_2 p_- - p_- \log_2 p_+ \]

\[ p_+ = \frac{n_+}{N} \quad p_- = \frac{n_-}{N} \]

Given set S of 14 examples

- 9 positive examples
- 5 negative examples

\[ S = [9+, \ 5-] \]

The information content I

\[ I(S) = - \frac{9}{14} \log_2(\frac{9}{14}) - \frac{5}{14} \log_2(\frac{5}{14}) = \]

\[ = 0.940 \]

Which Feature to Select? Information Gain Used in C4.5?

Expected reduction in entropy caused by the use of feature A

\[ \text{Gain}(A) = I(S) - \sum_{v \in \text{Values}(A)} \frac{\text{card}(S_v)}{\text{card}(S)} \cdot I(S_v) \]

\( S_v \) - a subset of S for which A assumes value v

Example

\[ S = [9+, \ 5-] \]

\[ \text{Gain}(\text{wind}) = I(S) - \frac{8}{14} \cdot E(S_{\text{weak}}) - \frac{6}{14} \cdot E(S_{\text{strong}}) = \]

\[ = 0.940 - \frac{8}{14} \cdot 0.811 - \frac{6}{14} \cdot 1.0 = 0.048 \]
Feature Selection

**feature** wind
Gain(wind) = 0.048

**feature** outlook
Gain(outlook) = 0.246

**feature** humidity
Gain(humidity) = 0.151

**feature** temperature
Gain(temperature) = 0.029

Continue to Evaluate

Calculate information gain for Sunny node, then redo feature selection

**feature** wind
Gain(wind) = 0.020

**feature** humidity
Gain(humidity) = 0.971

**feature** temperature
Gain(temperature) = 0.571

Continue to Evaluate

Calculate information gain for Rain node, then redo feature selection

**feature** wind
Gain(wind) = 0.971

**feature** humidity
Gain(humidity) = 0.020

**feature** temperature
Gain(temperature) = 0.020
Complete Decision Tree

Overcast
Sunny
Humidity
High Yes No
Low Yes No
Wind
Strong Yes No
Weak Yes No

From Decision Trees to Rules

If Outlook = Overcast
OR
Outlook = Sunny AND Humidity = Normal
OR
Outlook = Rain AND Wind = Weak
THEN Play tennis

Decision Trees: Key Characteristics

- Complete space of finite discrete-valued functions
- Maintaining a single hypothesis
- No backtracking in search
- All training examples used at each step

Avoiding Overfitting the Data

Accuracy

Size of tree

Training data set

Testing data set
Reference