Information Theory and Decision Trees
Entropy

- For a discrete random variable $X$
- Entropy: $H(X) = -\sum p(x)\log_2(p(x))$
- The “average number of bits needed each symbol of $X$”
Example

• X generates 4 possible symbols \{a, b, c, d\}
  – \( P(X = 'a') = 0.5 \)
  – \( P(X = 'b') = 0.25 \)
  – \( P(X = 'c') = 0.125 \)
  – \( P(X = 'd') = 0.125 \)

\[
\text{abaabadabcaadabbaad} \ldots
\]

Naïve Encoding:

\[
\begin{array}{c|c}
\text{I.I.D. X Generator} & \text{abaabadabcaadabbaad} \\
\end{array}
\]

\[
\begin{array}{c|c}
\text{Naïve Encoding:} & 00 \\
\text{a} & 10 \\
\text{b} & 01 \\
\text{c} & 11 \\
\text{d} & \\
\end{array}
\]
Naïve Encoding

- X generates 4 possible symbols \{a, b, c, d\}
  - \(P(X = 'a') = 0.5\)
  - \(P(X = 'b') = 0.25\)
  - \(P(X = 'c') = 0.125\)
  - \(P(X = 'd') = 0.125\)

Average cost of encoding each symbol?

Length of Encoding symbol ‘a’ \* Probability of symbol ‘a’
Length of Encoding symbol ‘b’ \* Probability of symbol ‘b’
... + ... + ... +

\[
= 2 \times 0.5 + 2 \times 0.25 + 2 \times 0.125 + 2 \times 0.125
\]

= 2 bits
We Can Do Better

- X generates 4 possible symbols \{a, b, c, d\}
  - \( P(X = 'a') = 0.5 \)
  - \( P(X = 'b') = 0.25 \)
  - \( P(X = 'c') = 0.125 \)
  - \( P(X = 'd') = 0.125 \)

Better Encoding:

**Average cost of encoding each symbol?**

Length of Encoding symbol ‘a’ * Probability of symbol ‘a’ +

\[
1 \times 0.5 + 2 \times 0.25 + \ldots + 3 \times 0.125 + \ldots
\]

= 1.25 bits

This Encoding is Optimal
We Can Do Better

• X generates 4 possible symbols \{a, b, c, d\}
  - \(P(X = 'a') = 0.5\)
  - \(P(X = 'b') = 0.25\)
  - \(P(X = 'c') = 0.125\)
  - \(P(X = 'd') = 0.125\)

**Better Encoding:**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Encoding</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>

**Average cost of encoding each symbol?**

Length of Encoding symbol ‘a’ * Probability of symbol ‘a’

Length of Encoding symbol ‘b’ * Probability of symbol ‘b’

= 1.25 bits

**But observe that:**

\[-\log(0.5) = 1\]
\[-\log(0.25) = 2 \ldots\ etc\]

Length of Encoding of symbol x = -\log(x)

Average Cost of encoding a symbol = \(-\sum p(x)\log_2(p(x))\) = \(H(X)\)
More Generally

• X generates 4 possible symbols \{a,b,c,d\}
  – P(X = ‘a’) = 0.21
  – P(X = ‘b’) = 0.14
  – P(X = ‘c’) = 0.52
  – P(X = ‘d’) = 0.13

What is the optimal encoding for this?

It is possible to attain an average of H(X) bits per symbol, but quite complicated: See *Arithmetic Compression*
Conditional Entropy

- \( H(X|Y) = \sum_y p(y) \sum H(X|Y = y) \)
  
  \[ = \sum_y p(y) \sum_x p(x|y) \log p(x|y) \]

Intuition: Average # bits needed to Encode Y, if X is already transmitted
Information Gain

- $H(X) = - \sum p(x) \log_2(p(x))$
- $H(X|Y) = \sum_y p(y) \sum H(X|Y = y)$
  $$= \sum_y p(y) \sum_x p(x|y) \log p(x|y)$$

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

$H(X)$: # bits to encode $X$
$H(X|Y)$: # bits to encode $X$ if $Y$ is already transmitted
$IG(X,Y)$: The “number of bits of information” gained about $X$, by knowing $Y$

**Counter Intuitive Property:** $IG(X, Y) = IG(Y, X)$
Decision Tree

Where are you?

Are you outdoors?

No

Do you see a computer?

No

You are lost

Yes

You are at work

Yes

You are lost
**ID3**

**Heuristic:** Pick the variable which provides the most Information Gain about Y

<table>
<thead>
<tr>
<th>Outdoors</th>
<th>Computer</th>
<th>Lost</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>2</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>2</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>5</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>1</td>
</tr>
</tbody>
</table>
**ID3**

**Heuristic:** Pick the variable which provides the most Information Gain about Y

<table>
<thead>
<tr>
<th>X1</th>
<th>X2</th>
<th>Y</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>2</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>2</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>5</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ IG(X1,Y) = H(Y) - H(Y | X1) \]

\[ H(Y) = - \frac{(5/10)}{2} \log(5/10) - \frac{5/10}{2} \log(5/10) = 1 \]

\[ H(Y | X1) = P(X1=T) H(Y | X1=T) + P(X1=F) H(Y | X1=F) \]

\[ = (4/10 \times 1 \log(1) - 0 \log(0)) + (6/10 \times - \frac{5/6}{2} \log(5/6) - \frac{1/6}{2} \log(1/6)) = 0.3900 \]

\[ IG(X1, Y) = 0.6100 \]
**Heuristic:** Pick the variable which provides the most Information Gain about Y

<table>
<thead>
<tr>
<th></th>
<th>X1</th>
<th>X2</th>
<th>Y</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>2</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>2</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>5</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>1</td>
</tr>
</tbody>
</table>

\[
IG(X2, Y) = H(Y) - H(Y|X2)
\]

\[
H(Y) = - \frac{5}{10} \log\left(\frac{5}{10}\right) - \frac{5}{10} \log\left(\frac{5}{10}\right) = 1
\]

\[
H(Y|X2) = P(X2=T) H(Y|X2=T) + P(X2=F) H(Y|X2=F) = \frac{7}{10} \cdot \left(\frac{7}{10} \log\left(\frac{7}{10}\right) + \frac{3}{10} \log\left(\frac{3}{10}\right)\right)
\]

\[
= 0.8797
\]

\[
IG(X2, Y) = 0.1203
\]
ID3

**Heuristic:** Pick the variable which provides the most Information Gain about $Y$

<table>
<thead>
<tr>
<th>Outdoors</th>
<th>Computer</th>
<th>Lost</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>2</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>2</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>5</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>1</td>
</tr>
</tbody>
</table>

$\text{IG}(X_1, Y) = 0.6100$
$\text{IG}(X_2, Y) = 0.1203$
**Heuristic:** Pick the variable which provides the most Information Gain about Y

<table>
<thead>
<tr>
<th>Outdoors</th>
<th>Computer</th>
<th>Lost</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>2</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>2</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>5</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>1</td>
</tr>
</tbody>
</table>

IG(Lost, Outdoors) = 0.6100
IG(Lost, Computer) = 0.1203
**ID3**

**Heuristic:** Pick the variable which provides the most Information Gain about Y. Recurse on the branches.

<table>
<thead>
<tr>
<th>Outdoors</th>
<th>Computer</th>
<th>Lost</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>2</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>2</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>5</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>1</td>
</tr>
</tbody>
</table>
**Heuristic:** Pick the variable which provides the most Information Gain about Y
Recurse on the branches

<table>
<thead>
<tr>
<th>Outdoors</th>
<th>Computer</th>
<th>Lost</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>2</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>2</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>5</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>1</td>
</tr>
</tbody>
</table>
Chi Square Pruning

- Decision Trees tend to overfit
- Pruning Necessary
- Bottom Up Pruning
Chi Square Pruning

1. Build Complete Tree
2. Consider each “leaf” decision and perform the chi-square test (label vs split variable)
Chi Square Pruning

# of instances entering this decision: \( s \)
# of + instances entering this decision: \( p \)
# of - instances entering this decision: \( n \)

Hypothesis: \( X \) is uncorrelated with the decision
Chi Square Pruning

# of instances entering this decision: s
# of + instances entering this decision: p
# of - instances entering this decision: n

\[
\begin{align*}
\text{\# instances here: } s_f \\
\text{\# of + instances here: } p_f \\
\text{\# of - instances here: } n_f \\
\text{\# instances here: } s_t \\
\text{\# of - instances here: } p_t \\
\text{\# of - instances here: } n_t
\end{align*}
\]

Hypothesis: X is uncorrelated with the decision

Then \( p_f \) should be “close” to \( (s_f \times p/s) \)
And \( p_t \) should be “close” to \( (s_t \times p/s) \)

Similarly for \( n_f \) and \( n_t \)
Chi Square Pruning

Consider the X2 split

<table>
<thead>
<tr>
<th>Variable Assignment</th>
<th>Real Counts</th>
<th>Expected Counts (S_{x2} * p / S)</th>
</tr>
</thead>
<tbody>
<tr>
<td>X2 = F</td>
<td>1</td>
<td>1/6</td>
</tr>
<tr>
<td>X2 = T</td>
<td>0</td>
<td>5/6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable Assignment</th>
<th>Real Counts</th>
<th>Expected Counts (S_{x2} * n / S)</th>
</tr>
</thead>
<tbody>
<tr>
<td>X2 = F</td>
<td>0</td>
<td>5/6</td>
</tr>
<tr>
<td>X2 = T</td>
<td>5</td>
<td>25/6</td>
</tr>
</tbody>
</table>
**Chi Square Pruning**

**Y = Lost**

<table>
<thead>
<tr>
<th>Variable Assignment</th>
<th>Real Counts</th>
<th>Expected Counts ($S_{X2} \times p / S$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X2 = F$</td>
<td>1</td>
<td>1/6</td>
</tr>
<tr>
<td>$X2 = T$</td>
<td>0</td>
<td>5/6</td>
</tr>
</tbody>
</table>

**Y = Not Lost**

<table>
<thead>
<tr>
<th>Variable Assignment</th>
<th>Real Counts</th>
<th>Expected Counts ($S_{X2} \times n / S$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X2 = F$</td>
<td>0</td>
<td>5/6</td>
</tr>
<tr>
<td>$X2 = T$</td>
<td>5</td>
<td>25/6</td>
</tr>
</tbody>
</table>

If uncorrelated, I expect the Real Counts to be close to Expected Counts.

Need some kind of measure of “deviation”

\[
C = \sum_{X2} \frac{(\text{Real Count}_{\text{lost}} - \text{Expected Count}_{\text{lost}})^2}{\text{Expected Count}_{\text{lost}}} + \frac{(\text{Real Count}_{\text{notlost}} - \text{Expected Count}_{\text{notlost}})^2}{\text{Expected Count}_{\text{notlost}}}
\]

\[
c \sim \chi^2((\text{num Y labels} - 1) \times (\text{num X2 labels} - 1))
\]

\[
c \sim \chi^2(1)
\]
**Chi Square Pruning**

\[
C = \sum X_2 \left( \frac{\text{Real Count}_{\text{lost}} - \text{Expected Count}_{\text{lost}}}{\text{Expected Count}_{\text{lost}}} \right)^2 + \left( \frac{\text{Real Count}_{\text{notlost}} - \text{Expected Count}_{\text{notlost}}}{\text{Expected Count}_{\text{notlost}}} \right)^2
\]

Intuitively, the smaller C is, the more likely they are uncorrelated.

\[
c = 6
\]

\[
c \sim \chi^2(1)
\]

If X2 and Y are uncorrelated, 
\(P(C \geq c)\) is the “probability” that we see such large deviations “by chance”.

We define “maxPChance” as the “worst chance we are willing to accept”

*(Coin Flip Example: we believe coin is unbiased. Then out of 1000 flips, How many “heads” do you want to see before you stop believing coin is unbiased?)*
**Chi Square Pruning**

\[ c = \sum X_2 \left( \frac{\text{Real Count}_{\text{lost}} - \text{Expected Count}_{\text{lost}}}{\text{Expected Count}_{\text{lost}}} \right)^2 + \left( \frac{\text{Real Count}_{\text{notlost}} - \text{Expected Count}_{\text{notlost}}}{\text{Expected Count}_{\text{notlost}}} \right)^2 \]

Intuitively, the smaller \( C \) is, the more likely they are uncorrelated.

\[ c = 6 \]

\[ C \sim \chi^2(1) \]

Let maxPchance = 0.05

We only stop believing that the splits are “by chance” if the probability of getting a deviation larger than \( c \) is < 0.05.
Let $\text{maxPchance} = 0.05$

We only stop believing that the splits are “by chance” if the probability of getting a deviation larger than $c$ is $< 0.05$.

Look at cdf. $P(C \leq 3.8415) = 0.95$
$P(C > 3.8415) = 0.05$

If $c \leq 3.8415$ we believe the split is “by chance” and prune the decision
If $c > 3.8415$ we do not believe the split is “by chance”
Applet

Play With Applet