Language Processing with Perl and Prolog
Chapter 4: Topics in Information Theory and Machine Learning

Pierre Nugues

Lund University
Pierre.Nugues@cs.lth.se
http://cs.lth.se/pierre_nugues/
Information theory models a text as a sequence of symbols. Let $x_1, x_2, ..., x_N$ be a discrete set of $N$ symbols representing the characters. The information content of a symbol is defined as

$$I(x_i) = - \log_2 p(x_i) = \log_2 \frac{1}{p(x_i)},$$

Entropy, the average information content, is defined as:

$$H(X) = - \sum_{x \in X} p(x) \log_2 p(x),$$

By convention: $0 \log_2 0 = 0.$
Entropy of a Text

The entropy of the text is

\[ H(X) = - \sum_{x \in X} p(x) \log_2 p(x). \]

\[ = -p(A) \log_2 p(A) - p(B) \log_2 p(B) - ... \\
- p(Z) \log_2 p(Z) - p(\hat{A}) \log_2 p(\hat{A}) - ... \\
- p(\check{Y}) \log_2 p(\check{Y}) - p(\text{blanks}) \log_2 p(\text{blanks}). \]

Entropy of Gustave Flaubert’s *Salammbô* in French is \( H(X) = 4.39 \).
Cross-Entropy

The cross entropy of $m$ on $p$ is defined as:

$$H(p, m) = - \sum_{x \in X} p(x) \log_2 m(x).$$

We have the inequality $H(p) \leq H(p, m)$.

<table>
<thead>
<tr>
<th>Book</th>
<th>Entropy</th>
<th>Cross entropy</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Salammbô</em>, chapters 1-14, training set</td>
<td>4.39481</td>
<td>4.39481</td>
<td>0.0</td>
</tr>
<tr>
<td><em>Salammbô</em>, chapter 15, test set</td>
<td>4.34937</td>
<td>4.36074</td>
<td>0.01137</td>
</tr>
<tr>
<td><em>Notre Dame de Paris</em>, test set</td>
<td>4.43696</td>
<td>4.45507</td>
<td>0.01811</td>
</tr>
<tr>
<td><em>Nineteen Eighty-Four</em>, test set</td>
<td>4.35922</td>
<td>4.82012</td>
<td>0.46090</td>
</tr>
</tbody>
</table>
Decision trees are useful devices to classify objects into a set of classes. Entropy can help us learn automatically decision trees from a set of data. The algorithm is one of the simplest machine-learning techniques to obtain a classifier.

There are many other machine-learning algorithms, which can be classified along two lines: supervised and unsupervised. Supervised algorithms need a training set. Their performance is measured against a test set. We can also use $N$-fold cross validation, where the test set is selected randomly from the training set $N$ times, usually 10.
Objects, Classes, and Attributes. After Quinlan (1986)

<table>
<thead>
<tr>
<th>Object</th>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Windy</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>False</td>
<td>N</td>
</tr>
<tr>
<td>2</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>True</td>
<td>N</td>
</tr>
<tr>
<td>3</td>
<td>Overcast</td>
<td>Hot</td>
<td>High</td>
<td>False</td>
<td>P</td>
</tr>
<tr>
<td>4</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>False</td>
<td>P</td>
</tr>
<tr>
<td>5</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>False</td>
<td>P</td>
</tr>
<tr>
<td>6</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>True</td>
<td>N</td>
</tr>
<tr>
<td>7</td>
<td>Overcast</td>
<td>Cool</td>
<td>Normal</td>
<td>True</td>
<td>P</td>
</tr>
<tr>
<td>8</td>
<td>Sunny</td>
<td>Mild</td>
<td>High</td>
<td>False</td>
<td>N</td>
</tr>
<tr>
<td>9</td>
<td>Sunny</td>
<td>Cool</td>
<td>Normal</td>
<td>False</td>
<td>P</td>
</tr>
<tr>
<td>10</td>
<td>Rain</td>
<td>Mild</td>
<td>Normal</td>
<td>False</td>
<td>P</td>
</tr>
<tr>
<td>11</td>
<td>Sunny</td>
<td>Mild</td>
<td>Normal</td>
<td>True</td>
<td>P</td>
</tr>
<tr>
<td>12</td>
<td>Overcast</td>
<td>Mild</td>
<td>High</td>
<td>True</td>
<td>P</td>
</tr>
<tr>
<td>13</td>
<td>Overcast</td>
<td>Hot</td>
<td>Normal</td>
<td>False</td>
<td>P</td>
</tr>
<tr>
<td>14</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>True</td>
<td>N</td>
</tr>
</tbody>
</table>
Classifying Objects with Decision Trees. After Quinlan (1986)
Decision Trees and Classification

Each object is defined by an attribute vector (or feature vector) \( \{A_1, A_2, \ldots, A_v\} \)

Each object belongs to a class \( \{C_1, C_2, \ldots, C_n\} \)

The attributes of the examples are:
\( \{\text{Outlook}, \text{Temperature}, \text{Humidity}, \text{Windy}\} \) and the classes are: \( \{N, P\} \).

The nodes of the tree are the attributes.

Each attribute has a set of possible values. The values of Outlook are
\( \{\text{Sunny}, \text{Rain}, \text{Overcast}\} \)

The branches correspond to the values of each attribute.

The optimal tree corresponds to a minimal number of tests.
ID3 (Quinlan, 1986)

Each attribute scatters the set into as many subsets as there are values for this attribute.

At each decision point, the “best” attribute has the maximal separation power, the maximal information gain.

Outlook

\[ \begin{align*}
\text{P}: & 9, \text{N}: 5 \\
\text{sunny} & : P: 2, N: 3 \\
\text{overcast} & : P: 4 \\
\text{rain} & : P: 3, N: 2
\end{align*} \]
The entropy of a two-class set $p$ and $n$ is:

$$I(p, n) = -\frac{p}{p+n} \log_2 \frac{p}{p+n} - \frac{n}{p+n} \log_2 \frac{n}{p+n}.$$ 

The weighted average of all the nodes below an attribute is:

$$\sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I\left(\frac{p_i}{p_i + n_i}, \frac{n_i}{p_i + n_i}\right).$$

The information gain is defined as $I_{\text{before}} - I_{\text{after}}$.
Example

\[ l_{\text{before}}(p, n) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.940. \]

Outlook has three values: sunny, overcast, and rain.

\[ I(p_1, n_1) = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} = 0.971. \]
\[ I(p_2, n_2) = 0. \]
\[ I(p_3, n_3) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0.971. \]

The gain is 0.940 – 0.694 = 0.246, the highest possible among the attributes.
Other Supervised Machine-Learning Algorithms

Linear classifiers:

1. Perceptron
2. Logistic regression
3. Support vector machines
Weka: A powerful collection of machine-learning algorithms
Running ID3

The Weka Explorer is a tool used for data mining and machine learning. It provides a user-friendly interface for selecting and applying machine learning algorithms, such as ID3, to data sets. The screenshot shows the ID3 classifier being run on a data set, with options for cross-validation and a specified number of folds. The output includes performance metrics like the percentage of correctly classified instances, kappa statistic, mean absolute error, and more. The detailed accuracy by class is also displayed.
Storing Quinlan’s data set in Weka’s attribute-relation file format (ARFF):

http://weka.wikispaces.com/ARFF:

@relation weather.symbolic

@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,hot,high,FALSE,no
sunny,hot,high,TRUE,no
overcast,hot,high,FALSE,yes
rainy,mild,high,FALSE,yes
rainy,cool,normal,FALSE,yes
rainy,cool,normal,TRUE,no
overcast,cool,normal,TRUE,yes