EDAN20
Language Technology
http://cs.lth.se/edan20/
Chapter 4: Topics in Information Theory and Machine Learning

Pierre Nugues

Lund University
Pierre.Nugues@cs.lth.se
http://cs.lth.se/pierre_nugues/

September 3, 2015

Entropy

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

$$\begin{aligned} H(X) &= -\sum_{x \in X} p(x) \log_2 p(x). \\ &= -p(A) \log_2 p(A) - p(B) \log_2 p(B) - \ldots \\ &\quad -p(Z) \log_2 p(Z) - p(\grave{A}) \log_2 p(\grave{A}) - \ldots \\ &\quad -p(\ddot{Y}) \log_2 p(\ddot{Y}) - p(blanks) \log_2 p(blanks). \end{aligned}$$

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 $H(p) \leq H(p, m)$.

	Entropy	Cross entropy	Difference
Salammbô, chapters 1-14, training set	4.39481	4.39481	0.0
Salammbô, chapter 15, test set	4.34937	4.36074	0.01137
Notre Dame de Paris, test set	4.43696	4.45507	0.01811
Nineteen Eighty-Four, test set	4.35922	4.82012	0.46090

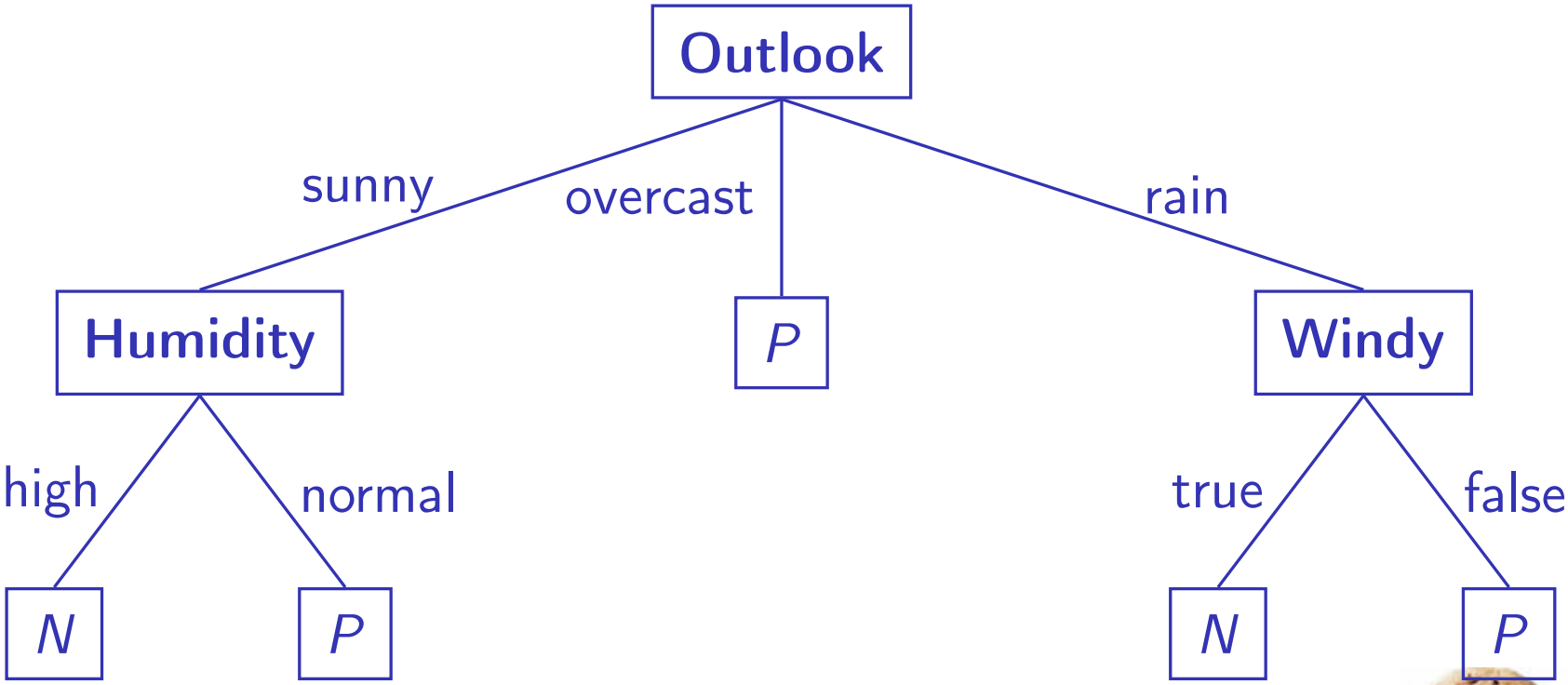
Entropy, Decision Trees, and Classification

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)

Object	Attributes				Class
	Outlook	Temperature	Humidity	Windy	
1	Sunny	Hot	High	False	*N*
2	Sunny	Hot	High	True	*N*
3	Overcast	Hot	High	False	*P*
4	Rain	Mild	High	False	*P*
5	Rain	Cool	Normal	False	*P*
6	Rain	Cool	Normal	True	*N*
7	Overcast	Cool	Normal	True	*P*
8	Sunny	Mild	High	False	*N*
9	Sunny	Cool	Normal	False	*P*
10	Rain	Mild	Normal	False	*P*
11	Sunny	Mild	Normal	True	*P*
12	Overcast	Mild	High	True	*P*
13	Overcast	Hot	Normal	False	*P*
14	Rain	Mild	High	True	*N*

Classifying Objects with Decision Trees. After Quinlan (1986)

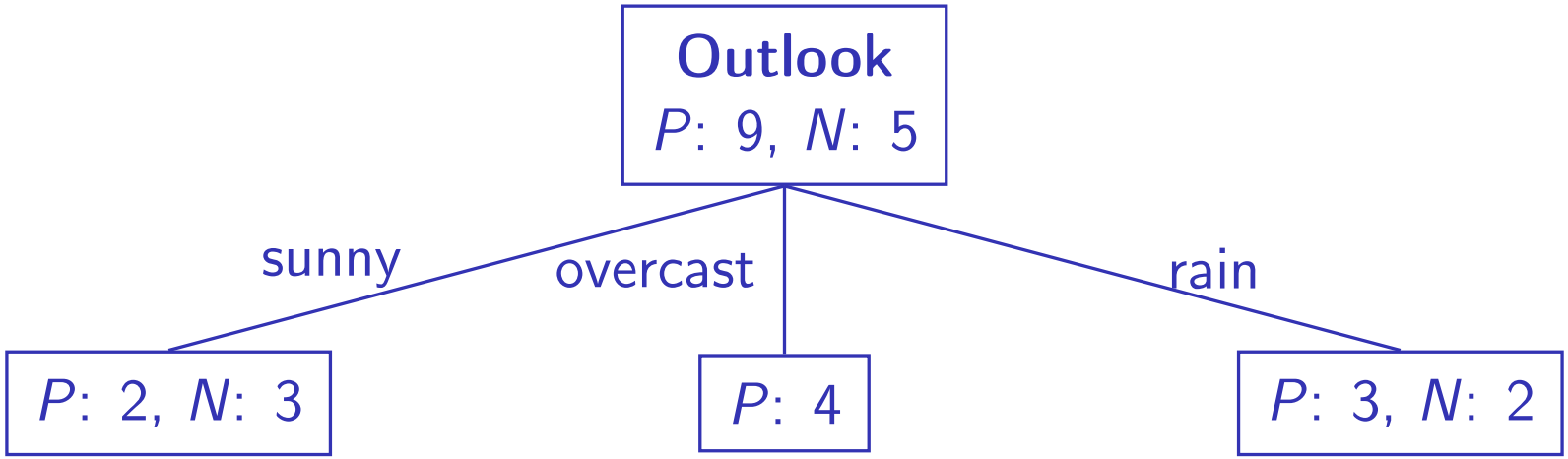


Decision Trees and Classification

Each object is defined by an attribute vector (or feature vector) $\{A_1, A_2, ..., A_v\}$
Each object belongs to a class $\{C_1, C_2, ..., C_n\}$
The attributes of the examples are:
$\{Outlook, Temperature, Humidity, Windy\}$ 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 $\{Sunny, Rain, Overcast\}$
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

ID3 (Quinlan, 1986)

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:

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

The information gain is defined as $I_{before} - I_{after}$

Example

$$I_{before}(p,n) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.940.$$

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

$$\begin{aligned} I(p_1,n_1) &= -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971. \\ I(p_2,n_2) &= 0. \\ I(p_3,n_3) &= -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.971. \end{aligned}$$

The gain is $0.940 - 0.694 = 0.246$, the highest possible among the attributes

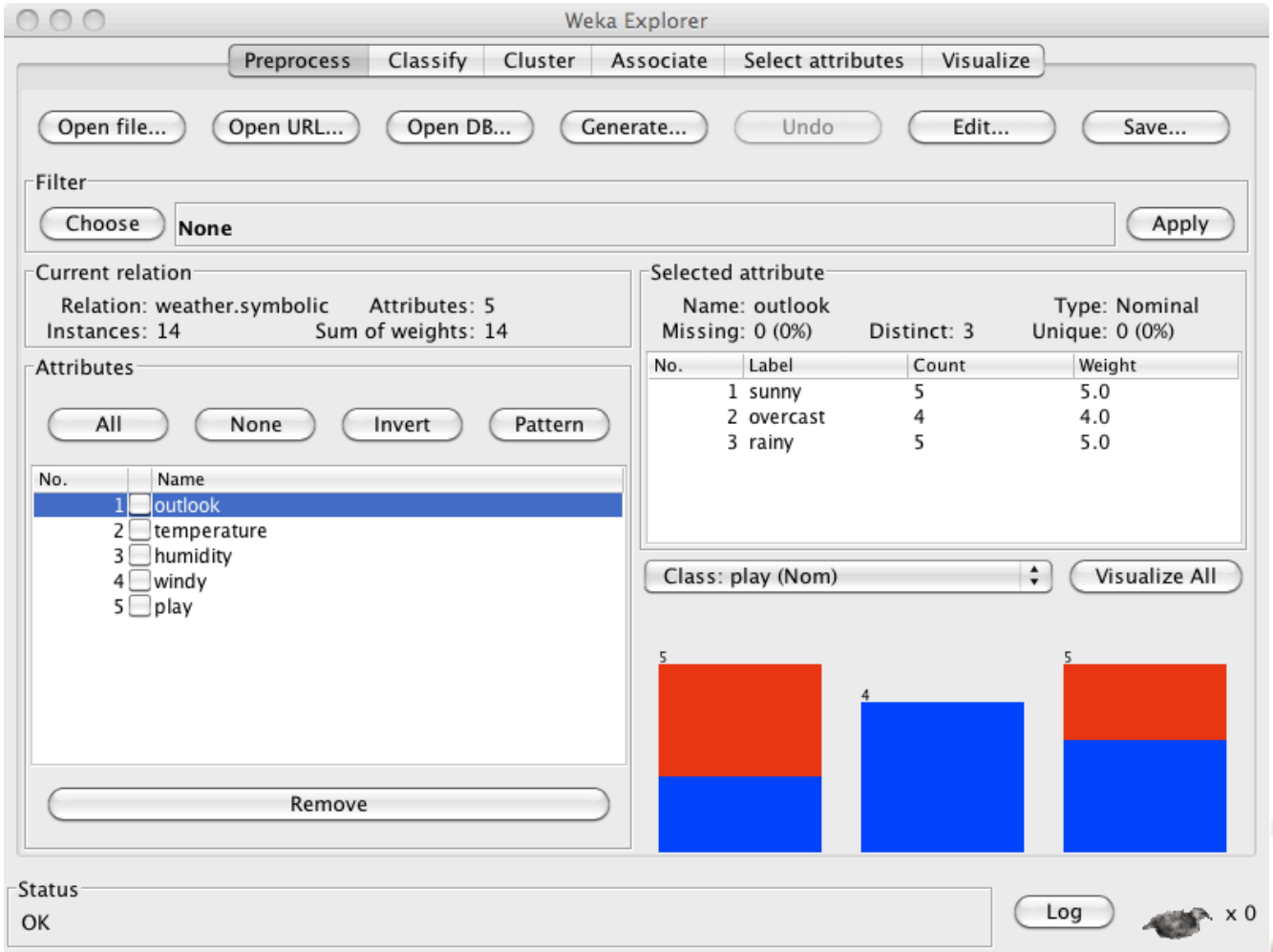
Other Supervised Machine-Learning Algorithms

Linear classifiers:

1. Perceptron
2. Logistic regression
3. Support vector machines

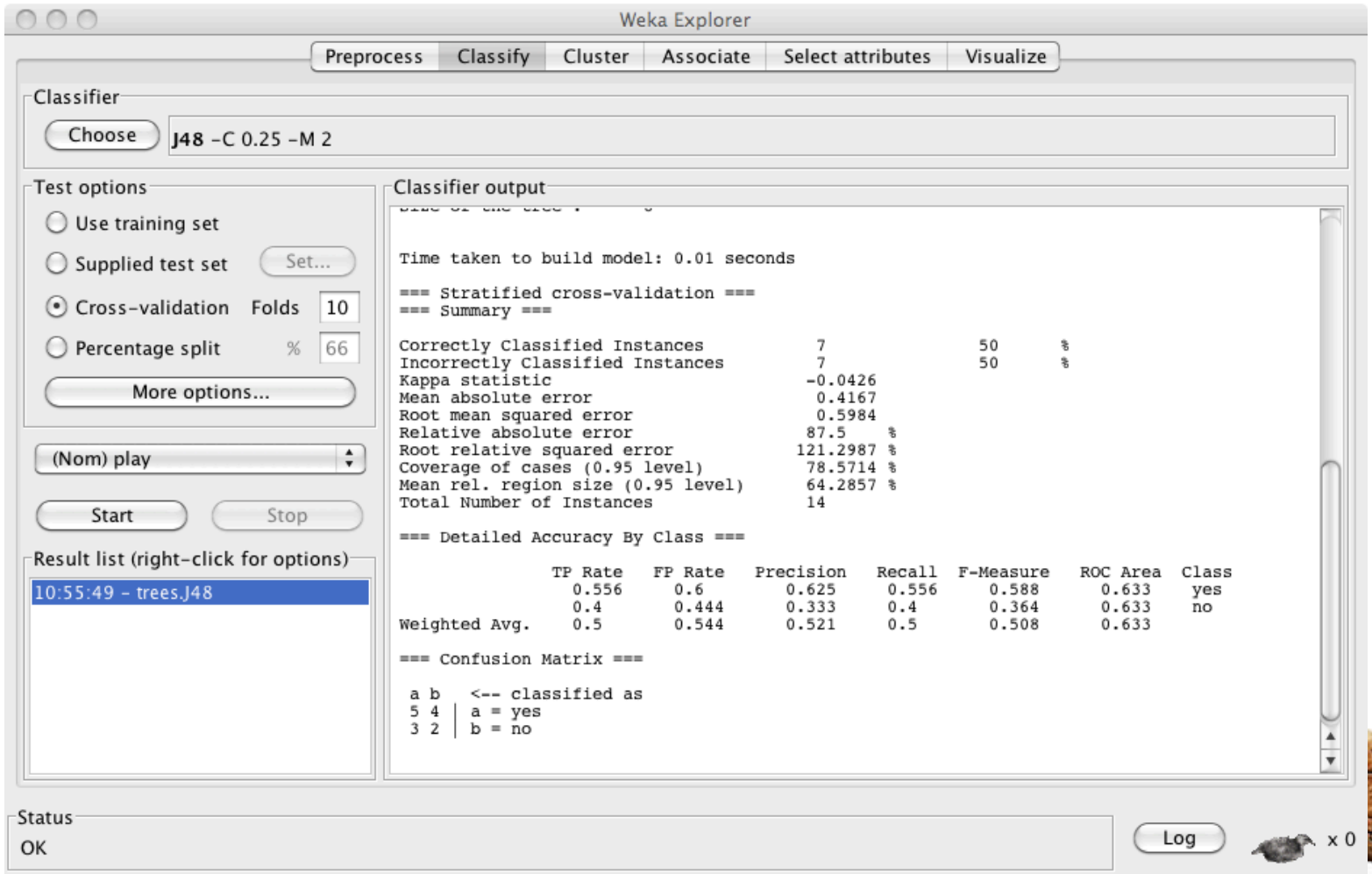
The Weka Toolkit

Weka: A powerful collection of machine-learning algorithms
http://www.cs.waikato.ac.nz/ml/weka/.



The Weka Toolkit

Running ID3



ARFF: The Weka Data Format

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
```