

Description of and Grammar for Mark's File Format

COSC-288

January 31, 2021

Introduction

Since machine-learning algorithms operate on examples, we need a way to get examples into our programs. One way is to read the data directly from a database. Another is to read the information from a file. Even if the data were in a database, it is often convenient to put at least a portion of the information in a file for development and experimentation.

For information stored in a file, we need a simple format for specifying examples, attributes, domains, and values. Based on other commonly used, but more complicated formats, I devised a simple format that should suit our purposes. A grammar for the format follows, but you should be able to parse files by reading tokens, or by reading lines and processing tokens. An example of the the Bikes data set follows the grammar.

Grammar

$\langle dataset \rangle$	\rightarrow	$\langle header \rangle \langle attributes \rangle \langle examples \rangle$
$\langle header \rangle$	\rightarrow	'@dataset' $\langle identifier \rangle$
$\langle attributes \rangle$	\rightarrow	$\langle attribute \rangle \langle attribute \rangle \langle attribute-list \rangle$
$\langle attribute-list \rangle$	\rightarrow	$\langle attribute \rangle \langle attribute-list \rangle \mid \epsilon$
$\langle attribute \rangle$	\rightarrow	'@attribute' $\langle identifier \rangle \langle declaration \rangle$
$\langle declaration \rangle$	\rightarrow	'numeric' $\mid \langle nominal-values \rangle$
$\langle examples \rangle$	\rightarrow	'@examples' $\langle example-list \rangle$
$\langle example-list \rangle$	\rightarrow	$\langle example \rangle \langle example-list \rangle \mid \epsilon$
$\langle example \rangle$	\rightarrow	$\langle attribute-value \rangle \langle example \rangle \mid \epsilon$
$\langle attribute-value \rangle$	\rightarrow	$\langle identifier \rangle \mid \langle number \rangle$
$\langle nominal-values \rangle$	\rightarrow	$\langle identifier \rangle \langle identifier \rangle \langle identifier-list \rangle$
$\langle identifier-list \rangle$	\rightarrow	$\langle identifier \rangle \langle identifier-list \rangle \mid \epsilon$
$\langle identifier \rangle$	\rightarrow	$\langle non-whitespace-character \rangle \langle non-whitespace-characters \rangle$
$\langle non-whitespace-characters \rangle$	\rightarrow	$\langle non-whitespace-character \rangle \langle non-whitespace-characters \rangle \mid \epsilon$
$\langle non-whitespace-character \rangle$	\rightarrow	'a', ..., 'z', 'A', ..., 'Z', '0', ..., '9', '-', ..., '+'
$\langle number \rangle$	\rightarrow	$-\infty, \dots, +\infty$

Example

```
@dataset bikes
```

```
@attribute make trek bridgestone cannondale nishiki garyfisher
```

```
@attribute tires knobby treads
```

```
@attribute bars straight curved
```

```
@attribute bottle y n
```

```
@attribute weight numeric
```

```
@attribute type mountain hybrid
```

```
@examples
```

```
trek knobby straight y 250.3 mountain
```

```
bridgestone treads straight y 200 hybrid
```

```
cannondale knobby curved n 222.9 mountain
```

```
nishiki treads curved y 190 hybrid
```

```
trek treads straight y 196.8 hybrid
```