

Exploring decision trees

Lab 1.3

Build and explore

- Build a decision tree for US_census_1994.arff.
- Examine attributes and their combinations
- Make the tree readable (by leaving only age, work class and education, for example)
- Experiment with different combinations of attributes
- Find 5 rules which you find interesting (If all rules seem trivial, find the 5 rules with the highest predictive power)

Possible treatment of highly branching attributes

The screenshot shows the Weka Explorer interface. The 'Preprocess' tab is active, and the 'Work class' attribute is selected. The 'Attributes' list on the left shows 10 attributes, with 'Work class' selected. The 'Selected attribute' panel on the right displays the following information:

Name: Work class
Missing: 1836 (6%)
Distinct: 8
Type: Nominal
Unique: 0 (0%)

| No. | Label | Count |
|-----|------------------|-------|
| 1 | State-gov | 1298 |
| 2 | Self-emp-not-inc | 2541 |
| 3 | Private | 22696 |
| 4 | Federal-gov | 960 |
| 5 | Local-gov | 2093 |
| 6 | Self-emp-inc | 1116 |
| 7 | Without-pay | 14 |

The 'Class' dropdown is set to 'Income (Nom)'. Below the table, a bar chart visualizes the distribution of the 'Work class' attribute across the 'Income' class. The bars are stacked with blue and red colors, representing different income levels. The counts for each bar are: 1298, 2541, 22696, 960, 2093, 1116, 14, and 7.

Status: OK

Convert each attribute value to a binary field

The screenshot shows the Weka Explorer interface. The 'Filter' dropdown menu is set to 'NominalToBinary -R first-last', which is circled in red. The 'Current relation' section shows 'Relation: census_1994_US-weka.filters.unsupervised.attribute.NominalToBinary' and 'Instances: 32561'. The 'Attributes' list on the left shows 'Age' selected. The 'Selected attribute' section for 'Age' shows 'Name: Age', 'Type: Numeric', 'Missing: 0 (0%)', 'Distinct: 16', and 'Unique: 0 (0%)'. A table below shows statistics for 'Age':

| Statistic | Value |
|-----------|--------|
| Minimum | 15 |
| Maximum | 90 |
| Mean | 36.58 |
| StdDev | 13.756 |

The 'Class: Income (Nom)' dropdown is set to 'Income (Nom)'. A histogram at the bottom right shows the distribution of 'Age' values, with bars colored in blue and red. The x-axis is labeled with 15, 52.5, and 90. The status bar at the bottom shows 'OK' and a 'Log' button.

Filters: unsupervised :attribute: NominalToBinary. Apply

...and remove all the binary attributes except private

The screenshot shows the Weka Explorer application window. The 'Preprocess' tab is active, and the 'NominalToBinary -R first-last' filter is applied. The 'Attributes' list on the left shows several binary attributes selected with checkmarks, including 'Work class=Never-worked'. The 'Remove' button at the bottom of the attributes list is circled in red. The 'Selected attribute' panel on the right shows statistics for 'Work class=Never-worked', including a mean of 0 and a standard deviation of 0.015. The 'Class' dropdown is set to 'Income (Nom)', and a visualization of the distribution is shown below it.

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter: Choose **NominalToBinary -R first-last** Apply

Current relation: Relation: census_1994_U5-weka.filters.unsupervised.attribute.Nomin...
Instances: 32561 | Attributes: 100

Attributes: All | None | Invert | Pattern

| No. | Name |
|-----|--|
| 1 | <input type="checkbox"/> Age |
| 2 | <input checked="" type="checkbox"/> Work class=State-gov |
| 3 | <input checked="" type="checkbox"/> Work class=Self-emp-not-inc |
| 4 | <input type="checkbox"/> Work class=Private |
| 5 | <input checked="" type="checkbox"/> Work class=Federal-gov |
| 6 | <input checked="" type="checkbox"/> Work class=Local-gov |
| 7 | <input checked="" type="checkbox"/> Work class=Self-emp-inc |
| 8 | <input checked="" type="checkbox"/> Work class=Without-pay |
| 9 | <input checked="" type="checkbox"/> Work class=Never-worked |
| 10 | <input type="checkbox"/> Education=Preschool |
| 11 | <input type="checkbox"/> Education=1st-4th |
| 12 | <input type="checkbox"/> Education=5th-6th |

Remove

Selected attribute: Name: Work class=Never-worked | Type: Numeric
Missing: 1836 (6%) | Distinct: 2 | Unique: 0 (0%)

| Statistic | Value |
|-----------|-------|
| Minimum | 0 |
| Maximum | 1 |
| Mean | 0 |
| StdDev | 0.015 |

Class: Income (Nom) | Visualize All

0 (0.09, 0.093]

0 | 0.5 | 1

Status: OK | Log | x 0