Decision Trees ID3

A Python implementation

Daniel Pettersson[1] Otto Nordander[2] Pierre Nugues[3]

[1]Department of Computer Science
Lunds University

[2]Department of Computer Science
Lunds University

[3]Department of Computer Science
Lunds University
Supervisor

EDAN70, 2017

Outline

Decision trees

- Easy to explain.
- More closely relates to human decision-making than other machine learning approaches.
- Trees can be displayed in an easy to understand manner.
- Gives a basic understanding of data.
- Often less accurate predictions but very fast.

Scikit-learn

- Very popular toolbox for machine learning.
- Scriptable and easy to integrate (fit, predict).
- Written with NumPy SciPy.
- No support for decision tree with nominal values.

ID3

Features of ID3

- Categorical values
- Entropy/Information gain
- Nodes can have several children

Current state

Scikit learn

- Scikit learn
  - CART
  - Only numerical, no nominal values.
  - No post-pruning.

Integration and API

Scikit learn

Scikit learn linear regression

```
regr = linear_model.LinearRegression()
regr.fit(data, target)
regr.predict(data_test)
```

Scikit learn decision tree

```
clf = tree.DecisionTreeClassifier()
clf.fit(data, target)
clf.predict(data_test)
```

Our decision tree

```
clf = id3.Id3Estimator()
clf.fit(data, target)
clf.predict(data_test)
```

Scikit-learn-contrib

- Compatible with Scikit-learn.
- Enforcing standards on code and documentation.
- Deploy to PyPI.

Extensions

ID3 and our extensions

Gain ratio

$$IV(Ex, a) = -\sum_{v \in values(a)} \frac{|\{x \in Ex | value(x, a) = v\}|}{|Ex|} \cdot \log_2 \left(\frac{|\{x \in Ex | value(x, a) = v\}|}{|Ex|} \right) \tag{1}$$

$$IGR(Ex, a) = IG / IV \tag{2}$$

Where IG is Information Gain and IV is Intrinsic Value. Gain ratio is used in place of information gain to reduce bias towards features that have many possible values.

Extensions

ID3 and our extensions

- Pre-pruning
  - Min samples split
  - Max depth
- Post-pruning
  - Split data into test and training.
  - Transform feature nodes to classifying nodes.
  - If new test error is lower keep the transformation (prune).

Extensions

ID3 and our extensions

- Numerical values.
- Multiclass classification.
- Reuse features.

Current state of our work

Demo and Usage

- Demo...
- Github - https://github.com/svaante/decision-tree-id3/
- PyPI - pip install decision-tree-id3
- Scikit-learn-contrib