Classification

Prof. Eric A. Suess

February 22, 2021

Introduction

Today we will begin to discuss Classification algorithms using Decision Trees and Rules.

We will learn about the C5.0, 1R, and RIPPER algorithms.

- C5.0 Rulequest Research
- C5.0 Max Kuhn UseR! 2013

Decision Trees

Decision Tree learners build a model in the form of a **tree structure**. Similar to a flowchart. **Decision nodes** indicate a decision to be made on an attribute/feature/variable. These split the tree into **branches**. Ending in the **leaf nodes**.

The tree begins with the data in the the **root node**.

Decision Trees

The path that each example/observation/record takes through the tree funnels it to a leaf node which assigns it to a predicted class.

Decision trees are very **transparent**.

Decision trees can be used with **almost any kind of data**.

However, if there are a large number of categorical features with a large number of categories or if there are a large number of numerical features, then decision trees may lead to very complex trees, which may not be so useful.

Divide and conquer

Decision trees are build using **recursive partitioning**, also known as **divide and conquer**. Splits the data in to smaller and smaller subsets of similar classes.

The first step is to find the most predictive feature of the target class. The next steps proceed with the next most predictive feature. In the end a stopping criterion is used.

Example

Movie scripts.

Predict if

- mainstream hit
- critic's choice
- box office bust

Use two features of past data

- budget
- A-list celebrities

Example

The splits of the data

- number of A-list celebrities too low
- high budget

See page 123/131.

C5.0

J. Ross Quinlan C5.0 Rulequest Research

A single threaded version of C5.0 has been release as open source software, it is included in R and other software, such as weka.

Choosing a best split

If a segment of data is a single class, it is considered **pure**.

C5.0 uses **entropy** for measuring purity.

The entropy of a sample of data indicates how mixed the class values are.

- Entropy of 0 indicates the sample is homogeneous.
- Entropy of 1 indicates the sample has the maximum amount of disorder.

Entropy

The definition of entropy is:

$Entropy(S) = \sum_{i=1}^{c} -p_i log_2(p_i)$

where

- S is the segment of data
- c is the number of class levels
- p_i is the proportion of values falling into class level i

Information Grain

To choose the feature to split on, information gain is used.

$InfoGain(F) = Entropy(S_1) - Entropy(S_2)$

See the paper An Empirical Comaprison of Selection Measures for Decision-Tree Induction for a discussion of other criterion that can be used.

Pruning

We do not want overly complicated trees. Smaller trees may be better for understanding and generalizing.

- pre-pruning
- post-pruning

Bank Loans

Next time we will work with the bank loans data example to develop a decision tree.