

Binary Decision Trees

Binary Decision trees work for regression and classification problems
They will handle attributes that are real numbers or factor variables
They also serve as the building blocks for ensemble methods

Some code examples will demonstrate how they work

Run tree.R Section 1 – Intro

Here are some more complicated trees

Run tree.R Section 2 – more complicated trees

Training a Regression Tree

First step – Find best split point on best attribute.

- How to measure "best"?

- Sum Square Error in Children

- We saw this in the simple code example

- If we have more than one attribute, pick the best one.

What about non-numeric (categorical) attributes?

- Take every possible subset

- Breiman's Theorem

Training a Classification Tree

Can't use sum square error

Use "impurity" measures instead.

Misclassification Error, Entropy and Gini

ESL Ch 9.2

Binary Decision Trees – Regularization

What parameters are available for controlling the over-fitting of a tree?

Tree depth, smallest node to be split, smallest improvement in fit

Run tree.R – Section 3 regularizing a tree

Any of these (and combinations of them) serves to regularize a binary tree.

A more concrete problem

Run tree.R – Section 4 – multidimensional problem

Tree Exercise

Here's a stretch problem.

Binary decision trees also work for classification problems.

Use `rpart()` to predict classification for the iris data set.

Follow the same process as you just saw for the concrete data just make sure that "Species" is a factor variable.

How will you measure performance? (Suggest miss-classification)