Ensemble Methods

Background and Binary Decision Trees

Schedule

9:00 – 9:15	Background On Ensemble Method
9:15 – 10:15	Binary Decision Trees
10:15 – 10:45	rxDTree for Big Data
10:45 – 11:00	Break
11:00 – 12:00	Intro to Bagging and Boosting
12:00 – 1:00	Lunch
1:00 – 2:00	Gradient Boosting and gbm Package
2:00 – 2:15	Intro to Random Forest
2:15 – 4:00	Random Forest and rxDForest for Big Dat

Learning Objectives

1. Basic Principles of Ensemble Methods
2. Binary Decision Trees
    A. Training
    B. Overfitting
    C. Parameters for Controlling Fit
    D. Using R-package rpart

3. Best Practices in Machine Learning
    A. Procedures for Measuring and Controlling Overfit
    B. General Procedure: Train and Test
4. Training Trees on Big Data - rxDTree

What Are Ensemble Methods?

- Combine Hordes of Independent Models
- Crowdsourcing for Machines
- If models are independent and classify better than 50/50, then probability of error decreases.

Use Different Classification Algo?

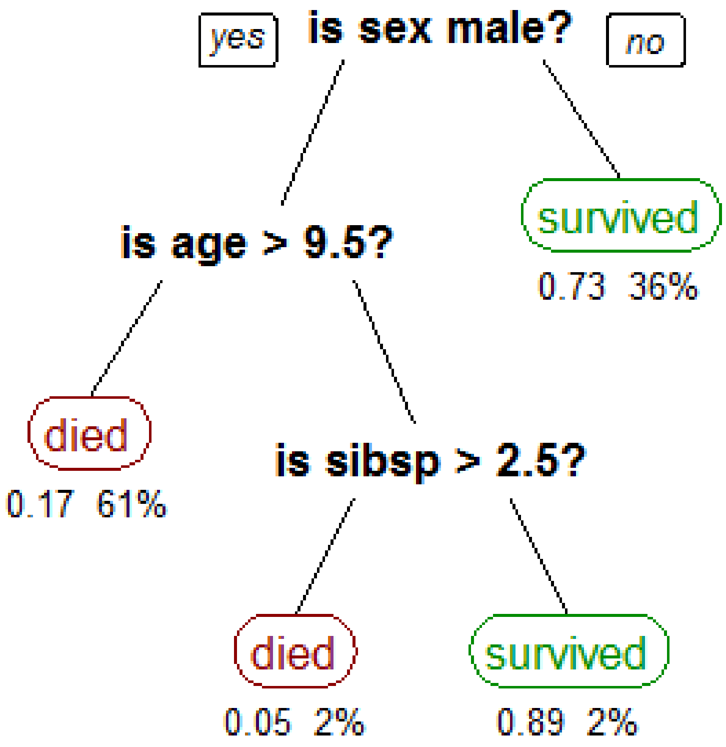
- Netflix Prize – Aggregate several different models within teams and between teams
- Q: How many different classification methods can we think of?

- http://www.cbcb.umd.edu/~hcorrada/PracticalML/pdf/lectures/EnsembleMethods.pdf

Many Different Problems

- Most used methods use one algo (base learner) on many different variants of the same problem.
- Binary Decision Trees are the usual choice for base learners.

Binary Decision Tree



A tree showing survival of passengers on the Titanic ("sibsp" is the number of spouses or siblings aboard). The figures under the leaves show the probability of survival and the percentage of observations in the leaf.

http://en.wikipedia.org/wiki/File:CART_tree_titanic_survivors.png

Training a Binary Decision Tree

- R Script, Section 1
- Q: What criterion is used to choose split point?
- More Complicated Trees – Section 2
- Q1: How to Calculate 2nd Split Point?
- Q2: Many machine learning algorithms have a “complexity parameter”. What are they for BDT?

Controlling Overfit

- R Script – Section 3
- Rpart control object –
  - maxdepth – max # of splits
  - cp – minimum improvement factor
  - minsplit – minimum node size

ML Best Practice

- Overfitting means a model is too complicated for the amount of training data available
- Performance on new, previously unseen data is usually all that matters.
- Procedure: Simulate unseen data by holding some data out from the training set.
- R Script – Section 4

BDT for Classification

- For regression, splits were selected to minimize sum squared error.
- For classification use misclassification error
- R Script Section 5

Building Trees on Big Data

- Split point determination drives computation
- Revolution rxDTree uses histogram to approximate split points selection (similar to Google PLANET)
- rxDTree call is very similar to rpart

Revolution Analytics rxDTree

- rxDTree(Rformula, data=, maxdepth=, cp=, xVal=)
  - Rformula - R formula language object
  - data – Data frame
  - maxdepth – maximum tree depth
  - cp – minimum improvement to split node
  - xVal – number of cross-val folds (default=2)

rxDTree

- Also Control of
  - Min node size to split
  - Granularity of histogram
- Includes cross-validation for tuning

Wrap Up

- Review training objectives for the section
- Look at Schedule for next session