



Ensemble Methods

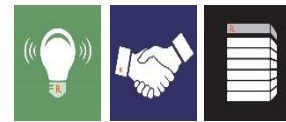
Random Forests



Learning Objectives

Random Forests - **Learning Objectives**

1. Understand Principles of RF Algorithm
2. Facility with R package randomforest
3. Facility with Revolution Analytics rxDForest
 - A. Application to different problem types
 - B. Parameters for controlling training
 - C. Interpreting Results



Random Forest

- Name coined by Leo Brieman in 2001 paper
- Combinations Brieman's "Bootstrap Aggregating" and Ho's "Random Decision Forests".
- How does RF construct multitude of somewhat independent problems?



Random Forest Algorithm

- Function approximation problem: Approximate a vector of responses Y , using a matrix of predictors X .
- Build collection of tree models $T_i()$ and average (or vote).



Building Different Trees

- Each tree is built as follows:
 - Take a random subset of the rows of Y and X (Brieman's bagging)
 - During training restrict each level of the tree to a random subset of the attributes (Ho's random decision forest).
- R-script RandomForest.R – Section 1.



Random Forest Package

- Training function – `randomForest()`
- Plot training progress – `plot()`
- Understanding model – `importance()`, `varImpPlot()`, `$proximity`, `varUsed()`, `partialPlot()`



Simple Specification

- Using R formula language- `randomForest(label~., data=dataset)`
- Using separate label and attribute files – `randomForest(x, y, data=dataset)`
- If y is factor variable, then classification problem is assumed, otherwise regression
- R-script section Random Forest 2.



Random Forests on Big Data

- Why?
 - Fitting complex function requires many degrees of freedom
 - Resolving many degrees of freedom requires lots of data
- How?
 - Revolution Analytics rxDForest()



Using rxDForest

- Simple to use
- `rxDForests(Rformula, data=, nTree=, mtry=, cp=)`
 - Rformula – R formula object
 - data – R Data frame
 - nTree – number of trees in ensemble
 - mtry – number of attributes used in trees
 - cp – minimum improvement for splitting



Using rxDForest

- Regression trees are used if labels are numeric
- Decision trees are used if labels are factor
- R script: Section 3.