GBM & Random Forest in H2O

Mark Landry

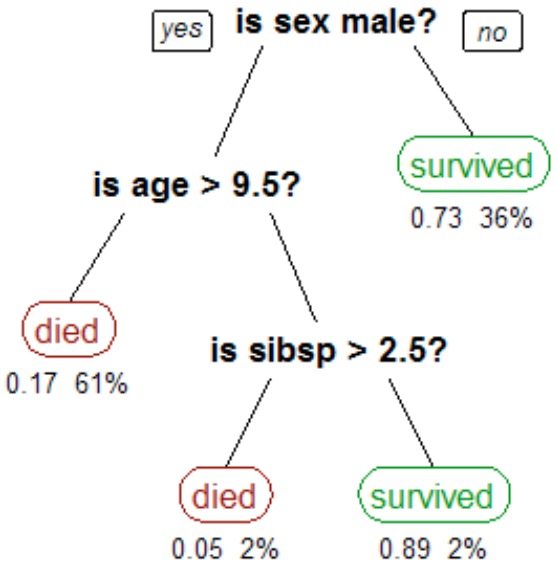
Presentation Outline

- Algorithm Background
  - Decision Trees
  - Random Forest
  - Gradient Boosted Machines (GBM)
- H2O Implementations
  - Code examples
  - Description of parameters and general usage

Decision Trees: Concept

- Separate the data according to a series of questions
  - Age > 9.5?
- The questions are found automatically to optimize separation of the data point by the "target"

Example decision tree:
Predicting survival of Titanic passengers



Source: wikimedia CART tree Titanic survivors

Decision Trees: Practical Use

Strengths

- Non linear
- Robust to correlated features
- Robust to feature distributions
- Robust to missing values
- Simple to comprehend
- Fast to train
- Fast to score

Weaknesses

- Poor accuracy
- Cannot project
- Inefficiently fits linear relationships

Improved Decision Trees: Ensembles

Random Forest

- Bootstrap aggregation (bagging)
- Fit many trees against different samples of the data and average together

GBM

- Boosting
- Fits consecutive trees where each solves for the net error of the prior trees

Random Forest

Conceptual

- Combine multiple decision trees, each fit to a random sample of the original data
- Randomly samples
  - Rows
  - Columns
- Reduce variance, with minimal increase in bias

Practical

- Strengths
  - Easy to use
    - Few parameters
    - Well-established default values for parameters
  - Robust
  - Competitive accuracy on most data sets
- Weaknesses
  - Slow to score
  - Lack of transparency

Gradient Boosted Machines (GBM)

Conceptual

- Boosting: ensemble of weak learners*
- Fits consecutive trees where each solves for the net loss of the prior trees
- Results of new trees are applied partially to the entire solution

Practical

- Strengths
  - Often best possible model
  - Robust
  - Directly optimizes cost function
- Weaknesses
  - Overfits
    - Need to find proper stopping point
  - Sensitive to noise and extreme values
  - Several hyper-parameters
  - Lack of transparency

* the notion of “weak” is being challenged in practice

H2O WORLD

Trees in H2O

- Individual tree fitting is performed in parallel
- Shared histograms calculate cut-points
- Greedy search of histogram bins, optimizing squared error

Explore Further through Examples

I have H2O Installed

I have R installed

I have the H2O World data sets

H2O WORLD