Ensemble Learning and the Heritage Health Prize

Jonathan Stroud, Igii Enverga, Tiffany Silverstein, Brian Song, and Taylor Rogers

iCAMP 2012
University of California, Irvine
Advisors: Dr. Max Welling, Dr. Alexander Ihler, Sungjin Ahn, and Qiang Liu

December 4, 2013
Outline

- Heritage Health Prize
  - Data
  - Evaluation
- Our Approach
  - Motivation
  - Models
  - Blending
- Results
The Heritage Health Prize

Goal: Identify patients who will be admitted to a hospital within the next year, using historical claims data.[1]

1,659 teams

Improve Healthcare, Win $3,000,000.

- Goal: Identify patients who will be admitted to a hospital within the next year, using historical claims data.[1]
- 1,659 teams
Purpose

▶ Reduce cost of unnecessary hospital admissions per year
▶ Identify at-risk patients earlier
Evaluation

Root Mean Squared Logarithmic Error (RMSLE)

\[ \varepsilon = \sqrt{\frac{1}{n} \sum_{i}^{n} [\log(p_i + 1) - \log(a_i + 1)]^2} \]

Threshold: \( \varepsilon \leq 0.4 \)
Our Approach

- Motivation
- Individual Models
- Optimized Ensemble
Blending

Blend several predictors to create a more accurate predictor.
Motivated by solution to the Netflix Prize [3]
Prediction Models

- Preprocessing: Feature Selection
- K-Nearest Neighbors
- Random Forests
- Gradient Boosting Machines
- Logistic Regression
- Support Vector Regression
- Neural Networks
Feature Selection

- Used Market Makers method [2]
- Reduced each patient to vector of 139 features
Kitchen Sink

- Generate large number of features
- Need method for eliminating useless features
  - Compare distributions of feature in test and training sets
  - Basis pursuit
Basis Pursuit

- Find best generated features (reduces error the most)
- Add to feature set
- Repeat until enough useful features are found
K-Nearest Neighbors

- Weighted average of closest neighbors
- Very slow

Neighbors: $k = 1000$
RMSLE: 0.475197
(996th place)
Decision Trees

![Plot of data points and decision tree]

Stroud, Enverga, Silverstein, Song, and Rogers
Ensemble Learning
Random Forests

RMSLE: 0.464918
(469th place)
Gradient Boosting Machines

Trees = 8000
Shrinkage = 0.002
Depth = 7
Minimum Observations = 100
RMSLE: 0.462998
(325th place)
Logistic Regression

Optimized with gradient descent

RMSLE: 0.466726
(580th place)
Support Vector Regression

\[ \varepsilon = .02 \]

RMSLE: 0.467152

(629th place)
Number of hidden neurons = 7
Number of cycles = 3000
RMSLE: 0.465705
(511th place)
Individual Predictors (Summary)

- K-Nearest Neighbors 0.475197 (996th place)
- Support Vector Regression 0.467152 (629th place)
- Logistic Regression 0.466726 (580th place)
- Neural Networks 0.465705 (511th place)
- Random Forests 0.464918 (469th place)
- Gradient Boosting Machines 0.462998 (325th place)
The Blending Equation

X as a combination of predictors

\[ \tilde{X} = Xw \]

Minimize cost function

\[ C = \frac{1}{n} \sum_{i=1}^{N} (Y_i - \tilde{X}_i)^2 \]
The Blending Equation

Optimizing predictors’ weights

\[ w_c = (Y^T X)(X^T X)^{-1} \]

\[ Y^T X = \sum_i X_{ic}^2 + \sum_{i=1}^{n} (X_i - Y_i)^2 - \sum_{i=1}^{n} Y_i^2 \]

\[ Y^T X = \sum_i X_{ic}^2 + n\varepsilon_0^2 - n\varepsilon_c^2 \]

Y = Actual values (Unknown)
X = Our predictions (Known)
\[ \varepsilon = \text{Feedback (Known)} \]
Blending Results

RMSLE: 0.461432
88th place on the final milestone leaderboard
Results (Summary)

- K-Nearest Neighbors 0.475197 (996th place)
- Support Vector Regression 0.467152 (629th place)
- Logistic Regression 0.466726 (580th place)
- Neural Networks 0.465705 (511th place)
- Random Forests 0.464918 (469th place)
- Gradient Boosting Machines 0.462998 (325th place)
- Blending 0.461432 (88th place)
Observations and Problems

- Fewer repeated classifiers worked better
- Overfitting based on feedback
- Test and Training data were inconsistent
Future Work

- Optimizing Blending Equation with Regularization Constant
  \[ w_c = (Y^T X)(X^T X + \lambda I)^{-1} \]
- More predictors
- Adjust for changes over time
Questions
Heritage provider network health prize, 2012. 

David Vogel Phil Brierley and Randy Axelrod. 
Market makers - milestone 1 description. 
September 2011.

Andreas Töscher and Michael Jahrer. 
The bigchaos solution to the netflix grand prize. 
September 2009.