Model Management: Ensemble Edition

Bay Area R Users Group

Strata+Hadoop WORLD

San Jose, CA March 2016

Erin LeDell Ph.D.
Machine Learning Scientist
H2O.ai

Introduction

- Statistician & Machine Learning Scientist at H2O.ai in Mountain View, California, USA
- Ph.D. in Biostatistics with Designated Emphasis in Computational Science and Engineering from UC Berkeley (focus on Machine Learning)
- Worked as a data scientist at several startups

Ensemble Learning

In statistics and machine learning, ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained by any of the constituent algorithms.

– Wikipedia (2016)

Common Types of Ensemble Methods

Bagging

- Reduces variance and increases accuracy
- Robust against outliers or noisy data
- Often used with Decision Trees (i.e. Random Forest)

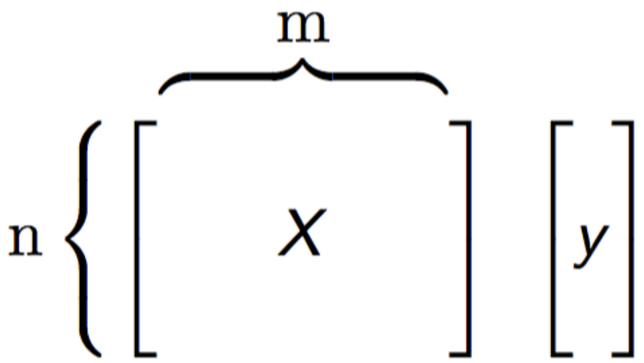
Boosting

- Also reduces variance and increases accuracy
- Not robust against outliers or noisy data
- Flexible – can be used with any loss function

Stacking / Super Learning

- Used to ensemble a diverse group of strong learners
- Involves training a second-level machine learning algorithm called a "metalearner" to learn the optimal combination of the base learners

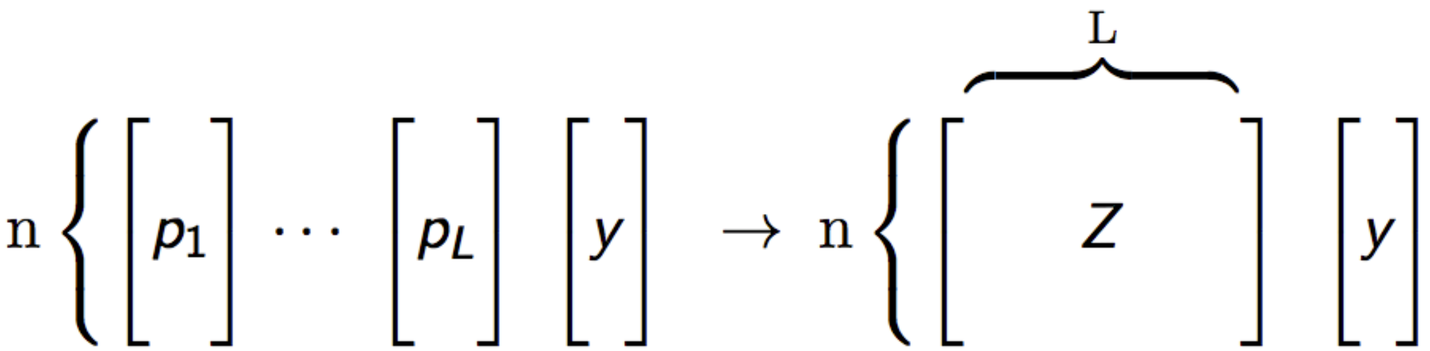
The Super Learner Algorithm



"Level-zero" data

- Start with design matrix, X, and response, y
- Specify L base learners (with model params)
- Specify a metalearner (just another algorithm)
- Perform k-fold CV on each of the L learners

The Super Learner Algorithm



"Level-one" data

- Collect the predicted values from k-fold CV that was performed on each of the L base learners
- Column-bind these prediction vectors together to form a new design matrix, Z
- Train the metalearner using Z, y

Super Learning vs. Parameter Tuning/Search

- A common task in machine learning is to perform model selection by specifying a number of models with different parameters.
- An example of this is Grid Search or Random Search.
- The first phase of the Super Learner algorithm is computationally equivalent to performing model selection via cross-validation.
- The latter phase of the Super Learner algorithm (the metalearning step) is just training another single model (no CV).
- With Super Learner, your computation does not go to waste!

Super Learner R Software Overview

SuperLearner

- Original Super Learner R implementation (2010).
- Comes with support for many existing machine learning R packages and can be customized to wrap any other.

subsemble

- Implements the Subsemble algorithm for combining models trained on partitions of the data, a variant of Super Learning.
- Like SuperLearner, can be used with any R algorithm.

h2oEnsemble

- H2O Ensemble implements the standard Super Learner algorithm using H2O distributed algorithms.
- Includes functions for automatically creating diverse ensembles.

H2O Ensemble R Package

Branch: **master** **h2o-3** / h2o-r / **ensemble** / +

ledell Update h2oEnsemble README — Latest commit 4824ede a minute ago

..		
demos	Added save/load functions to h2oEnsemble	8 days ago
h2oEnsemble-package	Optimized predict.h2o.ensemble function	2 days ago
README.md	Update h2oEnsemble README	a minute ago
SuperLearner_wrappers.R	Added h2o-3 version of h2oEnsemble package	4 months ago
create_h2o_wrappers.R	Added example to h2o-r/ensemble/create_h2o_wrappers.R	4 months ago
example_twoClass_higgs.R	Updated higgs example in h2oEnsemble	5 days ago

README.md

H2O Ensemble

The `h2oEnsemble` R package provides functionality to create ensembles from the base learning algorithms that are accessible via the `h2o` R package (H2O version 3.0 and above). This type of ensemble learning is called "super learning", "stacked regression" or "stacking." The Super Learner algorithm learns the optimal combination of the base learner fits. In a 2007 article titled, "Super Learner," it was shown that the super learner ensemble represents an asymptotically optimal system for learning.

H2O Ensemble R Interface

Example

```
library(h2oEnsemble)  #Install from GitHub

learner <- c("h2o.randomForest.1",
             "h2o.deeplearning.1",
             "h2o.deeplearning.2")

metalearner <- "h2o.glm.wrapper"

family <- "binomial"
```

H2O Ensemble R Interface

Example

```
fit <- h2o.ensemble(x = x, y = y, training_frame = train,
                    family = family,
                    learner = learner,
                    metalearner = metalearner)

pred <- predict(fit, test)
```

Stacking with Random Grids

New H2O Ensemble function in v0.1.8:

h2o.stack

http://tinyurl.com/h2o-randomgrid-stack-demo

Strata San Jose Exclusive!!

H2O Ensemble Resources

H2O Ensemble training guide:
http://tinyurl.com/learn-h2o-ensemble

H2O Ensemble homepage on Github:
http://tinyurl.com/github-h2o-ensemble

H2O Ensemble R Demos:
http://tinyurl.com/h2o-ensemble-demos

Thank you!

@ledell on Github, Twitter

erin@h2o.ai

http://www.stat.berkeley.edu/~ledell