Automatic Machine Learning in R

UseR! Brussels

July 2017

Erin LeDell Ph.D.
H2O.ai

What is H2O?

H2O.ai, the company

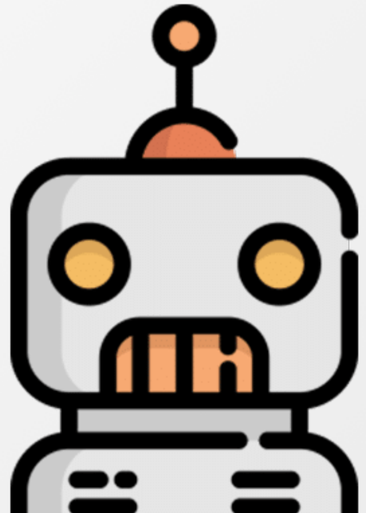
- Founded in 2012
- Advised by Stanford Professors Hastie, Tibshirani & Boyd
- Headquarters: Mountain View, California, USA

H2O, the platform

- Open Source Software (Apache 2.0 Licensed)
- R, Python, Scala, Java and Web Interfaces
- Distributed Machine Learning Algorithms for Big Data

Agenda

- Intro to Automatic Machine Learning (AutoML)
- Random Grid Search & Stacked Ensembles
- H2O's AutoML from R



Aspects of Automatic Machine Learning

Data Preprocessing

- Imputation, one-hot encoding, standardization
- Feature selection and/or feature extraction (e.g. PCA)
- Count/Label/Target encoding of categorical features

Model Generation

- Cartesian grid search or random grid search
- Bayesian Hyperparameter Optimization
- Individual models can be tuned using a validation set

Ensembles

- Ensembles often out-perform individual models
- Stacking / Super Learning (Wolpert, Breiman)
- Ensemble Selection (Caruana)

H2O AutoML (first release)

Data Preprocessing

- Imputation, one-hot encoding, standardization
- Feature selection and/or feature extraction (e.g. PCA)
- Count/Label/Target encoding of categorical features

Model Generation

- Cartesian grid search or random grid search
- Bayesian Hyperparameter Optimization
- Individual models can be tuned using a validation set

Ensembles

- Ensembles often out-perform individual models:
- Stacking / Super Learning (Wolpert, Breiman)
- Ensemble Selection (Caruana)

Random Grid Search & Stacking

- Random Grid Search combined with Stacked Ensembles is a powerful combination.

- Ensembles perform particularly well if the models they are based on (1) are individually strong, and (2) make uncorrelated errors.

- Stacking uses a second-level metalearning algorithm to find the optimal combination of base learners.

h2o R package

- A collection of distributed implementations of machine learning algos (GBM, RF, DNN, K-Means, GLM, etc.)
- CV, grid search, model eval & vis, deployment
- Computations are performed in highly optimized Java code in the H2O Cluster, initiated by REST calls from R.

H2O AutoML

- Basic data pre-processing (as in all H2O algos).
- Trains a random grid of GBMs, DNNs, GLMs, etc. using a carefully chosen parameter space; individual models are tuned using a validation set.
- A Stacked Ensemble is trained using all models.
- Returns a sorted "Leaderboard" of all models.

Available in H2O 3.12 & Bleeding Edge (not on CRAN yet)
https://h2o.ai/download

H2O AutoML

```
library(h2o)
h2o.init()

train <- h2o.importFile("train.csv")

aml <- h2o.automl(y = "response_colname",
                  training_frame = train,
                  max_runtime_secs = 600)
lb <- aml@leaderboard
```

H2O AutoML Leaderboard

model_id	auc	logloss
StackedEnsemble_0_AutoML_20170605_212658	0.776164	0.564872
GBM_grid_0_AutoML_20170605_212658_model_2	0.75355	0.587546
DRF_0_AutoML_20170605_212658	0.738885	0.611997
GBM_grid_0_AutoML_20170605_212658_model_0	0.735078	0.630062
GBM_grid_0_AutoML_20170605_212658_model_1	0.730645	0.67458
XRT_0_AutoML_20170605_212658	0.728358	0.629296
GLM_grid_0_AutoML_20170605_212658_model_1	0.685216	0.635137
GLM_grid_0_AutoML_20170605_212658_model_0	0.685216	0.635137

Example Leaderboard for binary classification

H2O Resources

- Documentation: http://docs.h2o.ai
- Tutorials: https://github.com/h2oai/h2o-tutorials
- Slidedecks: https://github.com/h2oai/h2o-meetups
- Video Presentations: https://www.youtube.com/user/0xdata
- Events & Meetups: http://h2o.ai/events

@ledell on Github, Twitter
erin@h2o.ai