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
• 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)
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.
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.
• 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
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
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

<table>
<thead>
<tr>
<th>model_id</th>
<th>auc</th>
<th>logloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>StackedEnsemble_0_AutoML_20170605_212658</td>
<td>0.776164</td>
<td>0.564872</td>
</tr>
<tr>
<td>GBM_grid_0_AutoML_20170605_212658_model_2</td>
<td>0.75355</td>
<td>0.587546</td>
</tr>
<tr>
<td>DRF_0_AutoML_20170605_212658</td>
<td>0.738885</td>
<td>0.611997</td>
</tr>
<tr>
<td>GBM_grid_0_AutoML_20170605_212658_model_0</td>
<td>0.735078</td>
<td>0.630062</td>
</tr>
<tr>
<td>GBM_grid_0_AutoML_20170605_212658_model_1</td>
<td>0.730645</td>
<td>0.67458</td>
</tr>
<tr>
<td>XRT_0_AutoML_20170605_212658</td>
<td>0.728358</td>
<td>0.629296</td>
</tr>
<tr>
<td>GLM_grid_0_AutoML_20170605_212658_model_1</td>
<td>0.685216</td>
<td>0.635137</td>
</tr>
<tr>
<td>GLM_grid_0_AutoML_20170605_212658_model_0</td>
<td>0.685216</td>
<td>0.635137</td>
</tr>
</tbody>
</table>

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