H2O AutoML Roadmap 2016.10

Raymond Peck

Director of Product Engineering, H2O.ai

rpeck@h2o.ai

What Will We Cover?

- What is AutoML?
- What is the roadmap for H2O AutoML?

What is AutoML?

H2O AutoML automates parts of data preparation and model training in order to help both Machine Learning / Data Science experts and complete novices.

Other AutoML projects concentrate on novices.

Outside AutoML Projects

- auto-sklearn
- AutoCompete
- TPOT
- DataRobot
- Automatic Statistician
- BigML
- et al...

Who is the Target Audience?

- "Big green button" for novice users such as software developers and business analysts;
- Iterative, interactive use and controls for expert users:
  - Machine Learning experts
  - Descriptive Data Scientists

What Are the Pieces?

- data cleaning
- feature engineering / feature generation
- feature selection
  - for both the original and generated features
- model hyperparameter tuning
- automatic smart ensemble generation

Prior Work @ H2O

- ensembles (stacking), from Erin LeDell
- random hyperparameter search with automatic stopping, from Raymond Peck
- some dataset characterization and feature engineering, from Spencer Aiello
- hyperopt Bayesian hyperparameter optimization, from Abhishek Malali

Current Work

- random hyperparameter search with parameter values based on open datasets
- moving ensembles into the back end
- working on basic metalearning for hyperparameter vectors, starting with 140 OpenML datasets

Future Work

- feature selection
- feature engineering for IID data
- Bayesian hyperparameter search with warm start
- feature engineering for non-IID data, e.g. time series
- iterate w/ larger datasets that are typical for our customers
- distribution guesser for regression

How Do We Evaluate Our Work?

- public datasets from
  - OpenML
  - ChaLearn AutoML challenge
  - Kaggle
- our own Data Scientists' work with customer datasets
- customer feedback (soon)

Data Cleaning

- outlier analysis (with user feedback)
- sentinel value detection
  - as a side-effect of outlier analysis
  - type-based heuristics (e.g., 999999, 1970.01.01)
- identifier detection (e.g., customer ID)
- smart imputation

Feature Generation

We will be using several techniques including:

- type-based heuristics
    - date/time expansion
    - log and other transforms of numerics
- interactions (product, ratio, etc)
- feature generation with Deep Learning deepfeatures()
- clustering

Feature Selection

We will be evaluating several techniques including:

- Mutual Information (non-linear correlation)
- variable importance from GBM and Deep Learning
- PCA
- GLM with Elastic Net / LASSO

Perhaps different selectors for initial data and transforms / interactions to trade off speed and the detection of non-linear relationships.

Hyperparameter Tuning

- currently do random hyperparameter search with metric-based smart stopping
  - hyperparameter values taken from hand-tuning 140 OpenML datasets
- soon adding simple "nearest neighbors" warm start (basic metalearning)
- then adding Bayesian hyperparameter optimization
  - possibly integrating hyperopt into the back end

Automatic Smart Ensemble Generation

- currently adding Erin LeDell's stacking / SuperLearner into the back end
- initially, ensemble top N models from hyperparameter searches
- optional "use original features"
- smarter ensemble generation for faster scoring, less overfitting:
  - greedy ensemble creation
  - ensemble models with uncorrelated residuals

Possible Futures

- try to predict accuracy from dataset metadata
- training time prediction
- scoring time prediction
- multiple concurrent H2O clusters for speed
- freeze/thaw model training
- outlier analysis with user feedback
- residuals analysis with user feedback
- composite models using pre-clustering step