H2O.ai

H2O: Algorithms Roadmap

February 1, 2015

H2O Algorithms Roadmap

Introduction

H2O is an open source math & machine learning engine for big data that brings distribution and parallelism to powerful algorithms while keeping the widely used languages such as R, Spark, Python, Java, and JSON as an API. H2O brings and elegant Lego-like infrastructure that brings fine-grained parallelism to math over simple distributed arrays.

H2O.ai brings a breadth of algorithms with the goal of being useful and relevant to our data science and algorithm users. Currently implemented and soon-to-be implemented algorithms are listed in this document.

Also, data characteristics influence some or most of the algorithm implementations.
sparse datasets, unbalanced asymmetric data and streaming (larger than memory) data make unique demands for each of the algorithms.

Finally, advanced tooling that enables parameter search in a given algorithm makes it easy for data scientists to iterate a given algorithm for best figure of merit.

Glossary

Data Science	The art of discovering insights from data
GLM	Generalization of Linear Regression techniques with different family and link functions
Decision Trees	A decision support tool that uses a tree-like graph or model of decisions and their possible consequences
Sampling	Technique of using smaller part of data for modeling.

Algorithm Roadmap

✔ = currently implemented feature

DATA CHARACTERISTICS

Data characteristics play a great role in algorithmic performance and implementation architecture.

- Sparse Data ✔
- Unbalanced Datasets ✔
- Very Large Data Modeling ✔
- Streaming Data
    - Scoring ✔
    - Modeling

STATISTICAL ANALYSIS

- GLM ✔
- Distributions: Gaussian, Binomial, Poisson, Gamma, Tweedie ✔
- Naïve Bayes ✔
- Cox Proportional Hazards ✔
- GLM Multinomial Regression
- Support Vector Machines

ENSEMBLES

- Distributed Random Forest ✔
- Gradient Boosting Machine ✔
- Distributed Trees ✔
- R Package- Ensembles Functions ✔

DEEP NEURAL NETWORKS

- Deep Learning ✔
- Auto-encoder ✔
- Anomaly Detection ✔
- Deep Features ✔
- Feed-Forward Neural Network ✔
- Convolutional and Pooling Layers
- GPU Support
- Stacked auto-encoders
- Recurrent Neural Nets for NLP

SOLVERS & OPTIMIZATION

- Generalized ADMM Solver ✔
- L-BFGS (Quasi Newton Method) ✔
- Ordinary Least-Square Solver ✔
- Stochastic Gradient Descent ✔
- MCMC

CLUSTERING

- K-Means ✔
- K-Nearest Neighbors
    - Singular Value Decomposition
- Locality Sensitive Hashing ✔

DIMENSTIONALITY REDUCTION

- Principal Component Analysis ✔

RECOMMENDATION

- Collaborative Filtering
    - Alternating Least Squares

TIME-SERIES

- ARIMA, ARMA Modeling
- Forecasting

DATA MUNGING

- Plyr ✔
- Integrated R-Environment ✔
- Slice, Log Transform ✔
- Anonymizing / Obfuscating (for personalized or confidential data)

GENERAL

- Weights for GBM, etc.
- Cost Function Specification

Version	Algorithms
1.0	GLM, Random Forest, K-Means, R
2.0	GLM-Categorical, GBM, PCA, Summary, Deep Learning, Python
3.0	ADMM, GLM-Sparse, SVM,
4.0	Adhoc Analytics, Unbalanced & Streaming Data