0xdata

H2O : Algorithms Roadmap

July 17, 2013

H_2O 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 of R and JSON as an API. H2O brings and elegant lego-like infrastructure that brings fine-grained parallelism to math over simple distributed arrays.

0xdata is bringing a breadth of Algorithms in H2O with the goal of being useful and relevant to our data science and algorithms users. Towards here's our roadmap for high-scale and fast implementations of Math, Machine Learning and Statistical algorithms.

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.
A summary list of Algorithms and Solvers that were hand picked from our early customer interactions.

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.

DATA CHARACTERISTICS

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

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

REGRESSIONS

Generalized Linear Modeling is the sliced bread of Algorithms.

- GLMNet,
- Distributions: Gaussian, Binomial, Poisson, Gamma, Tweedie
- Bayesian Regression
- Multinomial Regression

CLASSIFICATIONS

- Distributed Random Forest
- Gradient Boosting Machine
- Distributed Trees

NEURAL NETWORKS

- Multi-Layer Perceptron
- Auto-encoder
- Restricted Boltzmann Machines

SOLVERS & OPTIMIZATION

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

CLUSTERING

- K-Means, K-Nearest Neighbors
- Locality Sensitive Hashing
- Dimensionality reduction
- Singular Value Decomposition

TIME-SERIES

- ARIMA, ARMA Modeling
- Forecasting

DATA MUNGING

- plyr
- Integrated R-Environment
- Slice, Log Transform

Version	Algorithms
1.0	GLM, Random Forest, K-Means
2.0	GLM-Categorical, GBM, Perceptron
3.0	GLM-Sparse, RBM, SVM