H2O Algorithms Roadmap

0xdata
Date: July, 13 2013

Abstract:
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.

Data Characteristics:

1. Sparse Datasets
2. Unbalanced Asymmetric
3. Streaming Data (Larger than Memory)

Simple "legos" of Statistics:

1. Summarization
2. Histograms, Percentiles
3. Univariate feature Filtering (t-Tests, ratios of t-Tests; f-test)
4. Logarithm transformation

Regression & Classification:

5. GLM, Generalized Linear Modeling [Completeness to be R replacement.]
    a. Backwards-forwards feature selection using BIC/AIC
    b. Text-book GLMNet
    c. Feature Generation, Variable Importance
    d. k-folds, test-trains
    e. Handling NAs.
    f. Multinomial Regression
6. PCA, Principal Components Analysis
7. Bayesian Regression
8. Hierarchical Bayes Regression
9. BUGS (http://www.openbugs.info/Examples/Seeds.html)
10. Support Vector Machine (SVM)

Decision Trees:

11. GBM, Gradient Boosting Machine
12. DRF, Distributed Random Forest

Neural Networks

13. Multi-Layer Perceptron
14. Auto-encoder
15. Restricted Boltzmann Machines

Clustering:
16. K-Means (DEMO), hclust() – Sparse Data.
17. K-Nearest-Neighbors
18. Locality Sensitive Hashing.
19. Dimensionality Reduction
20. Topic clustering / LDA
21. Singlular Value Decomposition(SVD)

Markov Chains:
22. Hidden Markov Models

Time-Series: (TBD)
23. ARIMA library(forecast) from R.
24. GARCH

Solvers & Optimization
25. ADMM Solver
26. L-BFGS (quasi-Newton method)
27. Ordinary Least-Squares Solver Method.
28. Stochastic Gradient Descent
29. MCMC [non-trivial]

Data Munging:
30. plyr
31. Integrated R Environment.

Reference:

1. Anthes, Gary. "Deep Learning Comes of Age." Communications of the ACM (June, 2013). ACM.
2. Baldi, Pierre, and Kurt Hornik. "Neural Networks and Principal Component Analysis: Learning From Examples Without Local Minima." Neural networks 2.1 (1989): 53-58.
3. Elith, Jane, John R Leathwick, and Trevor Hastie. "A Working Guide to Boosted Regression Trees." Journal of Animal Ecology 77.4 (2008): 802-813. Google Scholar.
4. Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning: Springer, 2009. 337-387. Google Scholar.
5. Krizhevsky, Alex, Ilya Sutskever, and Geoff Hinton. Advances in Neural Information Processing Systems 25. N.p.: n.p., 2012.
6. http://plyr.had.co.nz/