Integrating high-performance machine learning: H2O and KNIME

Mark Landry (H2O), Christian Dietz (KNIME)
H2O: in-memory machine learning platform designed for speed on distributed systems.
High Level Architecture

H₂O.ai

H₂O Compute Engine

Load Data
Exploratory & Descriptive Analysis
Supervised & Unsupervised Modeling
Predict

Distributed In-Memory
Feature Engineering & Selection
Model Evaluation & Selection
Data & Model Storage

Loss-less Compression

Data Prep Export: Plain Old Java Object
Model Export: Plain Old Java Object

Production Scoring Environment

H₂O.ai
Distributed Algorithms

Advantageous Foundation

- Foundation for In-Memory Distributed Algorithm Calculation - Distributed Data Frames and columnar compression
- All algorithms are distributed in H2O: GBM, GLM, DRF, Deep Learning and more. Fine-grained map-reduce iterations.
- Only enterprise-grade, open-source distributed algorithms in the market

User Benefits

- “Out-of-box” functionalities for all algorithms (NO MORE SCRIPTING) and uniform interface across all languages: R, Python, Java
- Designed for all sizes of data sets, especially large data
- Highly optimized Java code for model exports
- In-house expertise for all algorithms
Scientific Advisory Council

Dr. Trevor Hastie
- John A. Overdeck Professor of Mathematics, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Co-author with John Chambers, *Statistical Models in S*
- Co-author, *Generalized Additive Models*

Dr. Robert Tibshirani
- Professor of Statistics and Health Research and Policy, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Author, *Regression Shrinkage and Selection via the Lasso*
- Co-author, *An Introduction to the Bootstrap*

Dr. Steven Boyd
- Professor of Electrical Engineering and Computer Science, Stanford University
- PhD in Electrical Engineering and Computer Science, UC Berkeley
- Co-author, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*
- Co-author, *Linear Matrix Inequalities in System and Control Theory*
- Co-author, *Convex Optimization*
Machine Learning Benchmarks
(https://github.com/szilard/benchm-ml)

Gradient Boosting Machine Benchmark
(also available for GLM and Random Forest)
**H2O Algorithms**

**Supervised Learning**

- **Statistical Analysis**
  - Generalized Linear Models: Binomial, Gaussian, Gamma, Poisson and Tweedie
  - Naïve Bayes

- **Ensembles**
  - Distributed Random Forest: Classification or regression models
  - Gradient Boosting Machine: Produces an ensemble of decision trees with increasing refined approximations

- **Deep Neural Networks**
  - Deep learning: Create multi-layer feed forward neural networks starting with an input layer followed by multiple layers of nonlinear transformations

**Unsupervised Learning**

- **Clustering**
  - K-means: Partitions observations into k clusters/groups of the same spatial size. Automatically detect optimal k

- **Dimensionality Reduction**
  - Principal Component Analysis: Linearly transforms correlated variables to independent components
  - Generalized Low Rank Models: extend the idea of PCA to handle arbitrary data consisting of numerical, Boolean, categorical, and missing data

- **Anomaly Detection**
  - Autoencoders: Find outliers using a nonlinear dimensionality reduction using deep learning
H2O in KNIME

Live Demo
H2O in KNIME

• Offer our users high-performance machine learning algorithms from H2O in KNIME

• Allow to mix & match with other KNIME functionality
  – Data wrangling KNIME Analytics Platform functionality
  – KNIME Big-Data Connectors
  – Text Mining, Image Processing, Cheminformatics, ...
  – and more!
H2O in KNIME

Live Demo
H2O in KNIME – Cross Validation
H2O in KNIME – Cross Validation
H2O in KNIME – Cross Validation

Collected results - 0:19 - Loop End

<table>
<thead>
<tr>
<th>Row ID</th>
<th>D Log Loss</th>
<th>D Mean</th>
<th>D Mean Error</th>
<th>D R2</th>
<th>D RMSE</th>
<th>D Error</th>
<th>D Error</th>
<th>D Accuracy</th>
<th>D Max Precision</th>
<th>D Pr (AUC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistics#0</td>
<td>0.658</td>
<td>0.497</td>
<td>0.233</td>
<td>0.045</td>
<td>0.483</td>
<td>1,964,007</td>
<td>0.421</td>
<td>0.579</td>
<td>0.989</td>
<td>0.682</td>
</tr>
<tr>
<td>Statistics#1</td>
<td>0.658</td>
<td>0.497</td>
<td>0.233</td>
<td>0.045</td>
<td>0.483</td>
<td>1,964,577</td>
<td>0.421</td>
<td>0.579</td>
<td>0.99</td>
<td>0.682</td>
</tr>
<tr>
<td>Statistics#2</td>
<td>0.658</td>
<td>0.497</td>
<td>0.233</td>
<td>0.045</td>
<td>0.483</td>
<td>1,964,234</td>
<td>0.421</td>
<td>0.579</td>
<td>0.99</td>
<td>0.682</td>
</tr>
<tr>
<td>Statistics#3</td>
<td>0.658</td>
<td>0.497</td>
<td>0.233</td>
<td>0.045</td>
<td>0.483</td>
<td>1,964,604</td>
<td>0.421</td>
<td>0.579</td>
<td>0.99</td>
<td>0.682</td>
</tr>
<tr>
<td>Statistics#4</td>
<td>0.658</td>
<td>0.496</td>
<td>0.233</td>
<td>0.045</td>
<td>0.483</td>
<td>1,964,437</td>
<td>0.421</td>
<td>0.579</td>
<td>0.989</td>
<td>0.682</td>
</tr>
</tbody>
</table>
H2O in KNIME – Parameter Optimization

Dialog - 0:61 - Parameter Optimization Loop Start (Co... - 

- **File**

- **Parameter Manager Selection**
  - **Memory Policy**
  - **Standard settings**
  - **Flow Variables**

- **Parameter**
  - **Start value**
  - **Stop value**
  - **Step size**
  - **Integer?**

- **Search strategy**
  - **BruteForce**

- **Add new parameter**

- **OK**
  - **Apply**
  - **Cancel**

- **Cross Joiner**
  - **Loop End**
  - **Collect results**

- **File Reader**

- **H2O Table to Frame**

- **H2O Local Context**

- **Start local H2O context**

© 2017 KNIME AG. All Rights Reserved.
H2O in KNIME – Parameter Optimization
H2O in KNIME – Nodes in KNIME 3.4
H2O in KNIME – What’s cooking?
H2O in KNIME – What’s cooking?

H2O MOJO Reader
Read back MOJO

File Reader
Read data

H2O MOJO Predictor
(Classification)

Apply classifier as MOJO
Thank you!