H2O in KNIME: Integrating High Performance Machine Learning

Jo-Fai Chow (H2O.ai),
Marten Pfannenschmidt (KNIME),
Christian Dietz (KNIME)
H2O for High Performance Machine Learning
222 of the 500 Fortune companies love H2O.

8 of top 10 banks

7 of top 10 insurance companies

4 of top 10 healthcare companies

Companies using H2O.ai

2015: 3,810
2016: 6,427
TODAY: 12,622

H2O.ai Users

2015: 38,257
2016: 54,163
TODAY: 129,406
Gartner 2018 Magic Quadrant for Data Science and Machine Learning Platforms

@KNIME gained the ability to run @H2O.ai algorithms, so these two may be viewed as complementary, not competitors
#Ecosystem #OpenSource

@srisatish

3:32 PM - 2 Mar 2018
High Level Architecture

H₂O Compute Engine

- Load Data
- Exploratory & Descriptive Analysis
- Supervised & Unsupervised Modeling
- Predict

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

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

Production Scoring Environment

- HDFS
- S3
- NFS
- Local
- SQL

- H₂O Flow
- Python
- JSON
- Java
- Scala
- KNIME

H₂O.ai
High Level Architecture

H₂O Compute Engine

- Load Data
- Distributed In-Memory
- Loss-less Compression
- Exploratory & Descriptive Analysis
- Feature Engineering & Selection
- Supervised & Unsupervised Modeling
- Model Evaluation & Selection
- Predict
- Data & Model Storage

Production Scoring Environment

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

Multiple Interfaces

Languages: R, Python, JSON, H₂O Flow, Java, Scala, KNIME

Data Storage: HDFS, S3, NFS, Local, SQL

Model Export: Plain Old Java Object

Production Scoring Environment

Languages: Spark, Kafka, Storm, KNIME
Algorithms Overview

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
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
H₂O on a Single Machine
H₂O on a Multi-Node Cluster
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
Performance Benchmark (GBM)
by Szilard Pafka (https://github.com/szilard/benchm-ml)

- **H₂O Gradient Boosting Machine (GBM)** is the fastest algorithm in the test of 10M Samples.
- Area Under Curve (AUC)
  - Comparable with xgboost
  - Better than R/Spark MLLib
- Benchmark results for other algorithms (Generalized Linear Model, Random Forest etc.) are available in Szilard’s GitHub repository.
H2O in KNIME
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
Customer prediction with H2O in KNIME
Welcome to Kaggle Competitions
Challenge yourself with real-world machine learning problems

Get the data  Build a model  Make a submission
The competition

Recruit Restaurant Visitor Forecasting

Predict how many future visitors a restaurant will receive

Provided data:
- Number of visitors
- Reservations
- Store information
- Calendar date info

<table>
<thead>
<tr>
<th>Date</th>
<th>Store ID</th>
<th>Visitors</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016-01-01</td>
<td>ba937bf13d40fb24</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2017-04-22</td>
<td>324f7c39a8410e7c</td>
<td>216</td>
</tr>
<tr>
<td>2017-04-23</td>
<td>e8ed9335d0c38333</td>
<td>?</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2017-05-31</td>
<td>8f13ef0f5e8c64dd</td>
<td>?</td>
</tr>
</tbody>
</table>
Solving a Kaggle competition with KNIME and H2O

Data preparation → Model training → Model optimization → Model evaluation → Deployment
Solving a Kaggle competition with KNIME and H2O

Data preparation → Model training → Model optimization → Model evaluation → Deployment

Diagram:
- Data preparation
- H2O Table to Frame
- Generalized Linear Model
- Concatenate (Optional in)
- Element Selector
- Cell To Model
- H2O Model to MOJO
- H2O MOJO Writer
- H2O Local Context
- Random Forest
- Gradient Boosting Machine
- H2O MOJO Predictor (Regression)
Data preparation with KNIME Nodes
The blended dataset

Data used for model training

<table>
<thead>
<tr>
<th>Date</th>
<th>Store ID</th>
<th>Visitors</th>
<th>Mean visitors</th>
<th>Area</th>
<th>Reservations</th>
<th>Holiday</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016-01-01</td>
<td>ba937bf13d40fb24</td>
<td>28</td>
<td>23.8</td>
<td>Area1</td>
<td>15</td>
<td>False</td>
<td>Bar</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2017-04-22</td>
<td>324f7c39a8410e7c</td>
<td>216</td>
<td>45.3</td>
<td>Area2</td>
<td>22</td>
<td>True</td>
<td>Karaoke</td>
</tr>
</tbody>
</table>

Data used for Kaggle prediction

<table>
<thead>
<tr>
<th>Date</th>
<th>Store ID</th>
<th>Visitors</th>
<th>Mean visitors</th>
<th>Area</th>
<th>Reservations</th>
<th>Holiday</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>2017-04-23</td>
<td>e8ed9335d0c38333</td>
<td>?</td>
<td>12.5</td>
<td>Area1</td>
<td>10</td>
<td>False</td>
<td>Italian</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2017-05-31</td>
<td>8f13ef0f5e8c64dd</td>
<td>?</td>
<td>54.1</td>
<td>Area7</td>
<td>20</td>
<td>False</td>
<td>Izakaya</td>
</tr>
</tbody>
</table>
Solving a Kaggle competition with KNIME and H2O

Data preparation → Model training → Model optimization → Model evaluation → Deployment

Diagram showing the workflow with nodes such as Random Forest, Concatenate, H2O Local Context, H2O MOJO Writer, etc.
Solving a Kaggle competition with KNIME and H2O

Data preparation → Model training → Model optimization → Model evaluation → Deployment
Modeling with the new H2O nodes

Parameter Optimization
- Loop Start
- H2O Random Forest Learner (Regression)
- H2O Cross Validation
- H2O Predictor (Regression)
- H2O Regression Scorer

Combine with other KNIME Nodes as controllers.

H2O Random Forest Learner (Regression)

- Select all search hits
- Include: 
  - air_store_id
  - id
  - air_area_name
  - shortAreaName
  - hpg_store_id

- Exclude: 
  - Manual Selection
  - Wildcard/Regex Selection
  - Type Selection

- Column(s): Search

Tree Options
- Number of levels (tree depth): 10

Forest Options
- Number of models: 100
- Use static random seed: 1520267662085

"CFG_NUMOFTREES" and "CFG_MAXDEPTH" are controlled by variables.
Modeling with the new H2O nodes
Modeling with the new H2O nodes
Modeling with the new H2O nodes
Solving a Kaggle competition with KNIME and H2O

Data preparation → Model training → Model optimization → Model evaluation → Deployment

Diagram showing the workflow with nodes for data preparation, H2O Table to Frame, Random Forest, Generalized Linear Model, Concatenate (Optional in), Element Selector, Cell To Model, H2O Model to MOJO, H2O MOJO Writer, H2O Local Context, and Gradient Boosting Machine.
Solving a Kaggle competition with KNIME and H2O

Model evaluation

Deployment

Data preparation

Model training

Model optimization

Model evaluation

KNIME and H2O

Recruit Restaurant Visitor Forecasting

Predict how many future visitors a restaurant will receive

© 2018 KNIME AG. All Rights Reserved.
What’s Cooking?
What’s Cooking: Scoring with H2O MOJOs on Spark

- Scoring using MOJOs
What’s Cooking: Scoring with H2O MOJOs on Spark

- New: Scoring using MOJOs on Spark
What’s Cooking: H2O Sparkling Water in KNIME
What’s Cooking: H2O Sparkling Water in KNIME
What’s Cooking: H2O Sparkling Water in KNIME
What’s Cooking: H2O Sparkling Water in KNIME
What’s Cooking: H2O Sparkling Water in KNIME