Scalable Machine Learning for Massive Astronomical Datasets

Nicholas M. Ball
Data Scientist (and former astronomy postdoc)
Skytree, Inc.

nick@skytree.net
Outline

What is Skytree and Why is it Interesting
Some Results on Large Astronomy Datasets
Linking to Exascale Radio Astronomy
Machine Learning

- Finding useful patterns within data
- Supervised: Known examples, predict new examples with model
- Unsupervised: unknown data structure, find new patterns, anomalies
- E.g., neural nets, decision trees, support vector machine, K-means, etc.
Technology Landscape

Data warehousing
Hadoop
Dedicated hardware

Production Grade
Machine Learning

Large Scale

Netezza (IBM)
Greenplum (EMC)
Vertica (HP)
Aster Data (Teradata)
Oracle Exadata
SAP HANA

Tableau
Spotfire
Qlikview
Cognos (Oracle)
Business Objects (SAP)

Relational databases
Spreadsheets

Small Scale

Oracle
Microsoft
IBM

Tableau
Spotfire
Qlikview
Cognos (Oracle)
Business Objects (SAP)

Basic Analytics

Statistical Packages
Math packages

Advanced Analytics

Matlab
R (open source)
Weka (open source)
SAS
SPSS (IBM)

“The machine learning company”
Skytree People

EXECUTIVE TEAM

**Martin Hack, CEO & Co-Founder**
Sun, GreenBorder (Google)

**Prof. Alexander Gray, PhD, CTO & Co-Founder**
National Expert on Large-Scale, Fast ML Algorithms

**Prof. Leland Wilkinson, PhD, VP Data Visualization**
Creator of Grammar of Graphics, SYSTAT (SPSS/IBM)

**Tim Marsland, PhD, VP Engineering**
Sun Fellow, CTO Software, Apple, Oracle

**Burke Kaltenberger, VP Worldwide Sales**
Infochimps/CSC, MapR, ParAccel

**Jin H. Kim, PhD, VP Marketing**
Tom Sawyer Software, Vitria, Mentor Graphics

TECHNICAL ADVISORY BOARD

**Prof. Michael Jordan**, UC Berkeley: machine learning ‘godfather’

**Prof. David Patterson**, UC Berkeley: systems (inventor RISC, RAID)

**Prof. Pat Hanrahan**, Stanford: data visualization (Tableau, Pixar)

**Prof. James Demmel**, UC Berkeley: high-performance computing (LAPACK)
Company Foundation

ACADEMIC MEMBERS

INVESTORS
USVP
JAVELIN
UPS
SAMSUNG
IQT
How to get the *highest predictive accuracy*?

**Skytree’s Key Differentiators**

1. **Breadth of Accurate Methods**: more types of advanced methods and options (thus higher chance of having best model type available)

2. **Speed/Scalability**: more data, test more parameters in the time available

3. **Automation/Ease of Use**: shorter time to accurate models and insights, more people in the organization can use it

Unlike previous systems, Skytree is designed from the ground up for this.
Skytree has invented ways to reduce the complexity of ML methods from $O(N^2)$ and $O(N^3)$ to $O(N)$ or $O(N \log N)$. 

**Common Machine Learning Use Cases**

- **Predictions**
  - Classification
  - Regression

- **Anomaly Discovery**
  - Clustering
  - Density Estimation

- **Recommendations**
  - Dimension Reduction
  - Multi-dimensional Querying

**Machine Learning Methods (Partial List)**

- Random Decision Forests
- Gradient Boosting Machines
- Nearest Neighbor
- Kernel Density Estimation (KDE)
- Decision Tree
- Linear Regression
- Support Vector Machine (SVM)
- 2-point Correlation
- Range Search
- Logit Regression
- Singular Value Decomposition

**Algorithms – Skytree’s Technology Breakthrough**

**Ease of Use**

**Completeness of Functionality**

**Speed and Scalability**

**Predictive Accuracy = Business Value**

Skytree’s product: High-performance ML software

- Predictive Accuracy = Business Value
- Distributed
- Streaming
Skytree’s Speed

**K-Means Clustering**
(time in seconds—shorter is better)

- **WEKA**: 262,440 seconds
- **R**: 24,456 seconds
- **SKYTREE**: 698 seconds
- **SKYTREE APPROX**: 271 seconds

**All Nearest Neighbors Query**
(time in seconds—shorter is better)

- **WEKA**: 72,000+ seconds
- **R**: 2,272 seconds
- **SKYTREE**: 9 seconds
- **SKYTREE INDEX**: 4.2 seconds

**Support Vector Machine Classification**
(time in seconds—shorter is better)

- **WEKA**: 187,982 seconds
- **R**: 3,351 seconds
- **SKYTREE**: 1,326 seconds

---

*Benchmarks were performed on the Amazon Elastic Compute Cloud. The systems had the following specifications: 1.7GB of memory – 5 EC2 compute units, 2 virtual cores with 2.5 EC2 compute units each, 350GB of instance storage 32-bit platform, Ubuntu Server ver. 10.04.

Data sets: Sloan Digital Sky Survey (SDSS) public data. K-Means Clustering – 1 million records; Support Vector Machine Classification – 384,000 records; All Nearest Neighbor – 1 million records.
Scalability: Multiple Machines to Process More Data

- Weak scalability

<table>
<thead>
<tr>
<th></th>
<th>Strong Scalability</th>
<th>Weak Scalability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Size</td>
<td>Constant</td>
<td>Increase by (&lt;y&gt;)</td>
</tr>
<tr>
<td>No. of Cores</td>
<td>Increase by (&lt;z&gt;)</td>
<td>Increase by (&lt;y&gt;)</td>
</tr>
<tr>
<td>Processing Time</td>
<td>Decrease by (&lt;z&gt;)</td>
<td>Constant</td>
</tr>
</tbody>
</table>
Speed: Multiple Machines for Higher Performance

Strong scaling

Data:
- 64 nodes (1024 cores)
Skytree Deployment

**Big Data Sources**
- Flat files
- Data Warehouse
- RDBMS
- NoSQL
- Hadoop

**Outputs**
- Business reports
- Systems monitoring
- Client application

Skytree Server
Enterprise Machine Learning

ETL, Data Sources ➔ Skytree Server ➔ Visualization of Results

Skytree Server runs on these Platforms:
- **HADOOP**
  - Cloudera, Hortonworks, MapR, Apache™ Hadoop®
- **DISTRIBUTED**
  - x86
- **STANDALONE WORKSTATION**
  - x86

Flexible Delivery

Show both Modeling & Production systems

ON PREMISE ➔ Flexible Delivery ➔ CLOUD
Skytree Server Real-Time Scoring

Real-time scoring with trained models

Client/Server communication model
(Skytree Server loads the model and performs scoring, Client streams queries and receives scores back)

Streaming via TCP sockets
(Additional option, e.g., “--port 5678”)

Low Latency: Round-trip times of
< 0.1 ms ( > 10k points / sec )
(Example: GBT, 32 numerical features, same rack, 10GbE)
What is Skytree and Why is it Interesting
Some Results on Large Astronomy Datasets
Linking to Exascale Radio Astronomy
$N^2 \rightarrow N$: Nearest Neighbors

Skytree all nearest neighbors (nn) on 470,992,970 2MASS objects

The lines show $y = a + (bx)^n$

$n \sim 1$ is linear scaling

$n \sim 2$ would be naïve scaling
Outliers

• Many ways to define outliers
• Use multiple methods -> more robust results
• We run:
  • KDE: points with low density
  • K-means: high clustercentric distances
  • NN: large neighbor distances

• We run them on the 2MASS dataset, for all 470,992,970 objects
Weak Scaling

Skytree K-means on 1,231,051,050 SDSS DR10 objects

```
SKYTREE (r) - THE MACHINE LEARNING COMPANY (tm) - http://www.skytree.net/
Copyright (c) 2010-2013, Skytree Inc.
Use of this software is subject to license terms.

Release: Skytree Server 12.10.0
Source ID: c7a38fa
Source Date: Fri Mar 28 17:19:09 2014 -0400
Compiler: 4.4.7 20120313 (Red Hat 4.4.7-4)
Compile Date: Mar 28 2014 14:44:29

Local time: 2014-Mar-31 12:35:26
Username: nick
Hostname: node05.skytree.internal
System: CentOS release 6.4 (Final)
Processor: Intel(R) Xeon(R) CPU E5-2650 0 @ 2.00GHz
# CPU Cores: 32 (hyperthreading)
Total Memory: 126.00 GB
Free Memory: 76.64 GB
System Load: 86.81 %

Working directory: /shared_homes/nick/Skytree/data/sdss/dr10/dr10csv/features
Command-line arguments:
skytree-server kmeans --references_in sdss_all_features_cModelMag_raw.csv --centroids_out
sdss_all_features_cModelMag_raw_centroids.out --k_clusters 2 --memberships_out
sdss_all_features_cModelMag_raw_memberships.out --algorithm=lloyds --iterations=10
--hosts=node05,node06,node07,node08,node09
```
Weak Scaling

12:35:26 [INFO] Module specified: kmeans
12:35:26 [INFO] System status:

<table>
<thead>
<tr>
<th>#</th>
<th>Hostname</th>
<th>Avail. Cores</th>
<th>Threads</th>
<th>Load (%)</th>
<th>Mem. (GB)</th>
<th>Free Mem. (GB)</th>
<th>Network Lat. (us)</th>
<th>Network B/W (MB/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>node05.skytree.</td>
<td>16</td>
<td>32</td>
<td>86.81</td>
<td>126.00</td>
<td>76.62</td>
<td>origin</td>
<td>origin</td>
</tr>
<tr>
<td>1</td>
<td>node06.skytree.</td>
<td>16</td>
<td>32</td>
<td>3.25</td>
<td>126.00</td>
<td>84.26</td>
<td>4.82</td>
<td>219.88</td>
</tr>
<tr>
<td>2</td>
<td>node07.skytree.</td>
<td>16</td>
<td>32</td>
<td>3.19</td>
<td>126.00</td>
<td>85.14</td>
<td>6.46</td>
<td>282.10</td>
</tr>
<tr>
<td>3</td>
<td>node08.skytree.</td>
<td>16</td>
<td>32</td>
<td>3.78</td>
<td>126.00</td>
<td>85.48</td>
<td>9.80</td>
<td>1061.63</td>
</tr>
<tr>
<td>4</td>
<td>node09.skytree.</td>
<td>16</td>
<td>32</td>
<td>3.19</td>
<td>126.00</td>
<td>84.28</td>
<td>7.35</td>
<td>891.34</td>
</tr>
</tbody>
</table>

12:35:26 [INFO] Full hostnames:

<table>
<thead>
<tr>
<th>#</th>
<th>Hostname</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>node05.skytree.internal</td>
</tr>
<tr>
<td>1</td>
<td>node06.skytree.internal</td>
</tr>
<tr>
<td>2</td>
<td>node07.skytree.internal</td>
</tr>
<tr>
<td>3</td>
<td>node08.skytree.internal</td>
</tr>
<tr>
<td>4</td>
<td>node09.skytree.internal</td>
</tr>
</tbody>
</table>

12:35:26 [INFO] Loading references from sdss_all_features_cModelMag_raw.csv
12:51:21 [INFO] k-means algorithm is lloyds
12:52:05 [INFO] ***** batch restart=0, distortion=1.32818e+06, iterations=4
12:52:05 [INFO] Lowest distortion found = 1.32818e+06
12:52:05 [INFO] Writing cluster centroids to sdss_all_features_cModelMag_raw_centroids.out
12:52:05 [INFO] Writing cluster memberships to sdss_all_features_cModelMag_raw_memberships.out
12:55:30 [INFO] Total time taken: 1204.88 seconds (20 minutes, 4.88 seconds)
What is Skytree and Why is it Interesting
Some Results on Large Astronomy Datasets
Linking to Exascale Radio Astronomy
Uses of Machine Learning in Astronomy

- Object detection
- Classification
- Distances
- Time series
- Dimension reduction
- Complex parts of simulations
- Data “triage”
Why Skytree ML is Interesting for Exascale Radio Astronomy

- Data complexity: Skytree has ML already, better than what any group could write themselves
- Designed to be the ML engine within a larger dataflow
- Data velocity: Fastest ML available, real-time streaming
- Company academic background, esp. astronomy
- Higher predictive accuracy for given data
- Best of both worlds: academia + industry
Conclusions

• Large astronomy data requires advanced analysis
• For exascale, both offline, and online (what to retain)
• One approach to advanced analysis is machine learning
• Machine learning is Skytree’s *raison d’être*
• Showed results for 0.5 billion and 1.2 billion objects for assorted machine learning methods, including *weak scaling*
• Potential for collaboration

nick@skytrees.net

Thanks!