Demo: SAS® In-Memory Statistics

Soren Johansen
Advisory Services Analytics
What is it?

• Provides a single interactive programming environment to perform:
  • Analytical data preparation
  • Variable transformations
  • Exploratory analysis
  • Statistical modeling and machine learning
  • Integrated modeling comparison and scoring

• Takes advantage of distributed in-memory computing optimized for analytical workloads
Benefits for data scientist

• All required components for successful data science (analytics lifecycle) at your fingertips in one integrated programming environment
• Support the experimental, iterative nature of analytical model development
• Faster time to insight – nimble and agile
• Analyze data sets of any size and any structure
• Explore and compare different modeling approaches to build the optimal champion model fast
• Better collaboration by empowering multiple users to work on the same data concurrently
• Maintain even distribution of workload among the nodes in a distributed computing environment
Benefits for IT

- Utilize compute resources more efficiently
  - Load data once for entire exploration and modeling process
  - Multiple user concurrency
- Improve data governance by eliminating data duplication
- Utilize distributed computing environment for high-performance data science
- An in-memory architecture eliminates costly data movement
- Easier maintenance with web-based clients
Data access and preparation

- Access structured and unstructured data
- In-memory data compression
- Aggregation techniques for transactional data
- Data filtering, including outliers
- Join tables and promote tables
- Dynamic group-by operations without sorting data
- Data step coding
- Group filtering, partitioning and data ordering within partitions
- Compute columns/variables on-the-fly to experiment with multiple scenarios faster
Data exploration

- Variety of analytics graphs
- Descriptive statistics, variable distribution, multi-dimensional summarizations
- Create and transform variables
- Dynamic group-by summarization
Model development

• Create models for different business problems – classification, predictions, recommendations and text analytics – in one single programming environment

• Combine structured and unstructured data for predictive analytics

• Explore/utilize multiple modeling approaches using SAS' advanced analytical and machine-learning algorithms

• Test model relevancy and accuracy fast with interactive programming

• Identify champion model quickly with large selection of model assessment statistics

• Dynamic model segmentation – create models instantly for different segments in your data, such as customer segments, product segments, geographical regions etc.
Model scoring

• Generation of SAS data step code to execute model on new data
• Integration with SAS® Model Manager for model management and model performance monitoring
Data Manipulation
• Aggregate
• Compute
• Update
• Append
• Set
• Schema
• DeleteRows
• DropTables
• PurgeTempTables

Data Exploration
• Boxplot
• Corr
• Crosstab
• Distinct
• Fetch
• Frequency
• Histogram
• KDE
• MDSummary
• Percentile
• Summary
• TopK

Predictive Modeling
• Decision Tree
• Forecast
• Gen Linear Model
• Linear Regression
• Logistic Regression
• Random Forests
• Neural Networks

Descriptive Modeling
• Association
• Path Analysis
• Clustering (k-means)
• Clustering (DBSCAN)

Evaluation, Deployment
• Assess
  • Misclassification matrix
  • Lift, ROC, Concordance
• Score
• Training / Validation

Utilities
• Where
• GroupBy
• TableInfo, ColumnInfo, ServerInfo
• Partition, Balance
• Store, Replay, Free
• Table, Promote

Text Analytics
• Parsing
• SVD
• Topic generation
• Document projection

Recommendation Systems
• Association
• Clustering
• kNN
• SVD
• Ensemble

HDFS I/O
• Sasiola
• Sashdat
• Anyfile Reader
Hypergroups

• Provide a number of analytics on data in which values in two or more columns are used to form a graph (vertices and edges)

• Three forms for structural analysis are available:
  • Identification of hypergroups. Hypergroups are sets of vertices which are completely disconnected from each other
  • Identification of strongly connected subgraphs using the Label Propagation algorithm
  • Identification of strongly connected subgraphs using the Graph Partitioning algorithm

• Determine graph layout in 2D or 3D space by Force-directed algorithms

• Calculate multiple vertex centrality measures based on shortest path and layout
Hypergroups – applications

Grouping – fuzzy names, address, telephone #, ...

Customer Segmentation

Fraud

Subsetting big data

Entity
DEMO AF SAS® IN-MEMORY STATISTICS
Demo agenda

- Proc Recommend
- Data manipulation & Model Building
- Hypergroups
Hypergroups

- Facebook ‘friend’ relationships
  - Data description
    - 4,039 nodes (people) and 88,234 edges
    - An edge exists between two people if they ‘friend’ each other
  - [http://snap.stanford.edu/data/egonets-Facebook.html](http://snap.stanford.edu/data/egonets-Facebook.html)
- 2D and 3D layouts were produced by HYPERGROUP where coordinates were determined by Force-directed algorithms
- Each color represents a community. About 50 communities were detected
*demo_short_with_ODS_10032013_v3_unplUSE.sas*

```
proc imstat;
   proc imstat;
   proc imstat;
```

```
%MERGE=
  table lasr.cardatap;
  schema carinfo (modelkey=modelkey / prefix=info) / mode=table;
run;
```

```
41table lasr.carinfo;
42fetch / fset=E to=5 format;
43run;
44/* --- EXPLORE ---*/
45table lasr.cardatap;
46distinct _all_ / save=tab;
47run;
* the name list in LIST1 is automatically created;
* no need to type in variable names one by one;
48store tabl(where='NDistinct le 5, 1)=list1;
49/*
50frequ frequency list1;
51crosstab wheelTypeID*wheeltype;
52proc mean;
53ods output frequency=work.freq;
54run;
55/*
56table lasr.cardatap(sampnames=avgOdo);
57summary avgOdo / temppvavgOdo tempexpr=avgOdo * veho.do /
58groupby=badbuy partition;
59run;
60/*
61ods graphics / height=10 width=12in;
```

```
proc sgpanel data=frqg;
   panelby column / novname columns=0 rows=1 uniscale=row;
   vbar formattedvalue / datalabel;
run;
```

```
70proc imstat;
    %MERGE=
    table lasr.cardatap;
    schema carinfo (modelkey=modelkey / prefix=info) / mode=table;
run;
```

```
76/* --- CLEAN UP ---*/
77table lasr_i_templast;
78where info_nationality ne 'NULL';
80run;
81/*
82proc imstat;
83/*
84compute delta_l = vehcoost - mmreacquisitionretailaverageprice;
```

```
```
proc imstat;
  table lasr.movlenth_med;
    fetch / format;
  table lasr.movieprofile;
    fetch / format;
  table lasr.userprofile;
    fetch / format;
run;

proc recommend port=&myport recom = rs.movielens;
  * Start a new project;
  add rs.movielens / item = itemid user = userid
            rating = rating;
  * Add tables;
  addtable lasr.movlenth_med / recom = rs.movielens
            type = rating
            vars = (itemid userid rating);
  addtable lasr.movieprofile / recom = rs.movielens
            type = item;
  addtable lasr.userprofile / recom = rs.movielens
            type = user;
run;

*Method -- KNN seed=1234;
method knn / label = "knn" k = 20 positive
similarity = pc seed = 1234;
predict / method = knn label = "knn" Num = 5
  users = ("1", "33", "478", "2035");
run;

*Method -- svd LRECS -1 with factors 20:
/*--- ASSESS ---*/
table lasr. & _template ;
asess _ILINK_ / y = &response event = 'bad' nbins=10 step=0.05;
ods output ROCInfo = roc LiftInfo = lift;
run;

/*--- TREE ---*/
table lasr.cardata2;
decisiontree isbadbuy
/ nbins=40 maxlevel=10 greedy gain prune
   input = ( &list1)
nominal = ( &list1)
   code=(file="/home/&user/scorecode2.sas" comment replace)
run;

/*--- RANDOM FOREST ---*/
table lasr.cardata2;
randomwoods isbadbuy
/ ntree=10 nbins=40 m=6 maxlevel=10 bootstrap=0.7 greedy gain
leafsize=100
   input = ( &list1)
nominal = ( &list1)
   code=(file="/home/sasdemo/scorecode2.sas" comment replace)
run;

quit;
ods graphics / height=6in width=6in;

proc sgplot data=lift,
series x = depth y = Cumlift / markers markerattrs=(symbol=circlefilled);
series x = depth y = CumliftBest;
yaxis label = ' grid;
run;
data roc;
set roc;
one_minus_Specificity = 1 - Specificity;
proc sort data=roc;
by one minus Specificity;
Key capabilities

- **SPEED**
  - Proven state-of-the-art statistical algorithms and machine-learning techniques
  - Memory and data efficient for a significant reduction of data latency to rapidly analyze large and complex data

- **PRECISION**
  - Highly scalable, in-memory environment grows easily as needed

- **INTERACTIVE**
  - Multi-user interactive analytics environment for increased productivity
Key takeaways

• **Faster** time to insight
• **Interactive** analytics
• **Seamless integration** of
  • Data access and preparation
  • Data exploration
  • Model development
  • Model scoring
• Proven and **state-of-the-art** analytical and machine-learning algorithms
• **Scalable**
• **Concurrent** access to data loaded in-memory in a **multi-user** environment
For more information

• Learn more about SAS and Hadoop: www.sas.com/hadoop

• Learn more about SAS® In-Memory Statistics: www.sas.com/in-mem

• Read the TDWI Checklist Report, ‘Eight Considerations for Utilizing Big Data Analytics with Hadoop’: www.sas.com/consider-hadoop

• Big Data Matters Webinar Series: www.sas.com/bigdatamatters

Follow us on Twitter: @sasanalytics
Like us on Facebook: SAS Software
Thank you!

Soeren.Johansen@sas.com