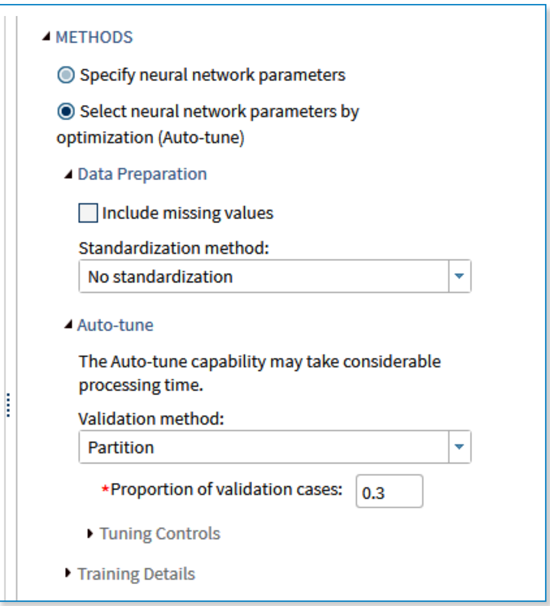
SAS® Viya™ Data Mining and Machine Learning

Everything needed to solve your most complex problems within a single, integrated in-memory environment



◢ METHODS

○ Specify neural network parameters

● Select neural network parameters by optimization (Auto-tune)

◢ Data Preparation

□ Include missing values

Standardization method: No standardization

◢ Auto-tune

The Auto-tune capability may take considerable processing time.

Validation method: Partition

*Proportion of validation cases: 0.3

▸ Tuning Controls

▸ Training Details

What does SAS® Viya™ Data Mining and Machine Learning do?

SAS® Viya™ Data Mining and Machine Learning combines data wrangling, data exploration, visualization, feature engineering, and modern statistical, data mining and machine-learning techniques all in a single, scalable in-memory processing environment. This provides faster, more accurate answers to complex business problems, increased deployment flexibility and one easy-to-administer and fluid IT environment.

Why is SAS® Viya™ Data Mining and Machine Learning important?

It enables data scientists and others to solve previously unfeasible business problems by removing barriers created by data sizes, data diversity, limited analytical depth and computational bottlenecks. Dramatic performance gains and innovative algorithms mean greater productivity and faster, more creative answers to your most complex problems.

For whom is SAS® Viya™ Data Mining and Machine Learning designed?

It is designed for those who want to use a powerful and customizable in-memory programming language to analyze large, complicated data and uncover new insights faster. This includes data scientists, experienced statisticians, data miners, engineers, researchers and scientists.

Data volumes continue to grow. Highly skilled data scientists and analytical professionals are in short supply. Organizations struggle to find timely answers to increasingly complex problems.

Whether it's analyzing every transaction to identify emerging fraud patterns, analyzing growing amounts of social media chatter to improve customer experience or producing an accurate and fast recommendation system for predicting next-best offers, sophisticated machine-learning software gives organizations a way to solve their most important issues.

SAS Viya Data Mining and Machine Learning addresses all of the steps necessary to turn raw data into new insights. From a single, integrated in-memory environment, your data scientists can access and prepare data, perform exploratory analysis, build and compare machine-learning models, and create score code for implementing predictive models, more quickly than ever before.

Benefits

- **Boost the productivity of your data scientists.** Time to value is essential for the success of analytical projects in businesses. SAS Viya Data Mining and Machine Learning enables data scientists to get highly accurate results quicker.
- **Solve complex analytical problems faster.** This solution takes advantage of SAS Viya, a new in-memory architecture, to deliver predictive modeling and machine-learning capabilities for breakthrough performance. In-memory data persistence avoids multiple data loading required during iterative analyses. Analytical model processing time is measured in seconds or minutes rather than hours so you can find solutions to difficult problems faster than ever.
- **Quickly explore multiple approaches to find optimal solutions.** The distributed analytical engine with superior performance and the feature-rich building blocks for your machine-learning pipeline enable you to quickly and easily explore multiple approaches. Automated tuning lets you test different scenarios in an integrated environment to find the best performing model and provide answers with high levels of accuracy and confidence.
- **Overcome big data analytics challenges.** Apply modern machine-learning techniques to large volumes of structured and unstructured textual data to derive previously unknown insights.
- **Quickly deploy your predictive models with automated generation of SAS score code.** Shorten the time to value even more with easy-to-implement score code that is automatically generated in multiple programming languages for all your machine-learning models.
- **Use graphical interfaces for common machine learning tasks.** Intuitive graphical interfaces are part of the web-based programming environment and allow for the easy configuration of common machine-learning tasks. The associated SAS code is automatically generated for later batch runs and automation. Users can share data sources and code-snippets in this environment for better collaboration.

Overview

Currently, statisticians and data scientists often use multiple programming languages or products to manage the different tasks of data mining and machine-learning pipelines. With SAS Viya Data Mining and Machine Learning, they can work in a single intuitive programming environment with access to a fast in-memory processing engine optimized for analytical workloads. Data scientists will improve time to value for their advanced analytics projects by being able to test many modeling scenarios quickly and select the champion predictive model with high confidence.

A flexible, web-based programming environment

SAS Studio provides a web-based interface for the most common machine-learning steps – from data prep to model building, assessment and scoring. You choose whether to program your projects in SAS code or use intuitive graphical interfaces for the most common tasks of a machine-learning pipeline. Each task also generates SAS code behind the scenes for later batch runs, editing and automation.

Highly scalable, in-memory analytical processing

SAS Viya Data Mining and Machine Learning takes advantage of the next generation of SAS In-Memory Analytics, which is optimized for multipass analytical computations. The analytics processing engine provides a

Key Features

Interactive, distributed, in-memory web-based programming environment

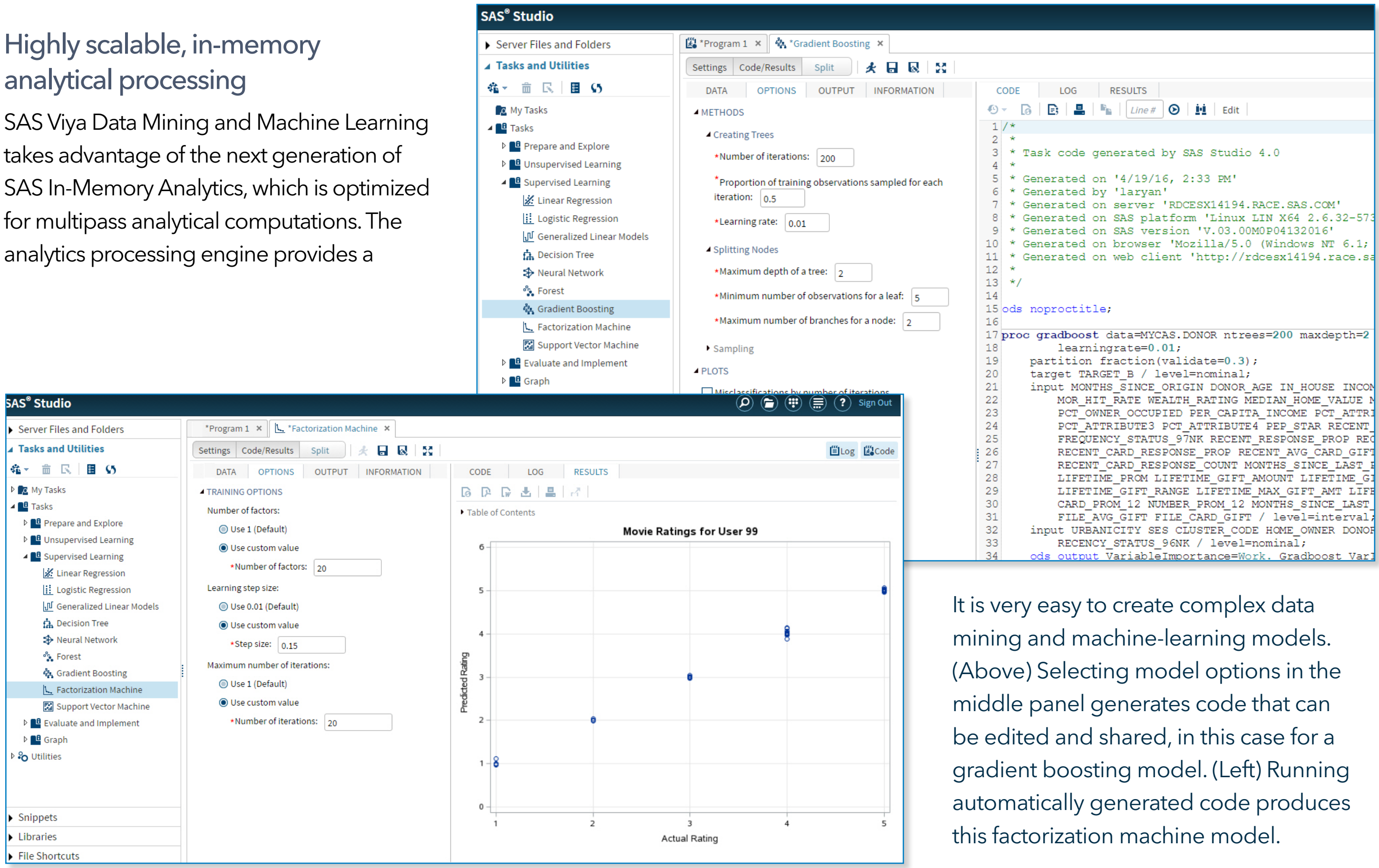
- Low maintenance web-based interface (SAS Studio) for programmers.
- Interactive graphical tasks in SAS Studio support point-and-click machine learning.
- SAS Studio tasks generate SAS code for getting started quickly and automating machine-learning tasks.
- Collaborative environment enables easy sharing of data, code snippets and best practices.

Highly scalable, in-memory analytical processing

- Distributed in-memory processing of complex analytical calculations on large data sets provides low-latency answers.
- Analytical tasks are chained together as a single in-memory job without having to reload the data or write out intermediate results to disks.
- Concurrent access to the same data source in memory by many users for efficiency.
- Data and intermediate results are held in memory as long as required, reducing processing latency.
- Built-in workload management ensures efficient use of compute resources.
- Built-in failover management guarantees submitted jobs always finish.

Analytical data preparation

- Distributed SAS DATA step language:
  - Run SAS DATA step code in parallel in a distributed, in-memory computing environment.
  - Control the level of parallelism per execution node and how many nodes to engage.
- Data summarization and cardinality profiling:



METHODS

Creating Trees

*Number of iterations: 200

*Proportion of training observations sampled for each iteration: 0.5

*Learning rate: 0.01

Splitting Nodes

*Maximum depth of a tree: 2

*Minimum number of observations for a leaf: 5

*Maximum number of branches for a node: 2

Sampling

PLOTS

□ Misclassifications by number of iterations

```
/*
 *
 * Task code generated by SAS Studio 4.0
 *
 * Generated on '4/19/16, 2:33 PM'
 * Generated by 'laryan'
 * Generated on server 'RDCESX14194.RACE.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 2.6.32-573
 * Generated on SAS version 'V.03.00M0P04132016'
 * Generated on browser 'Mozilla/5.0 (Windows NT 6.1;
 * Generated on web client 'http://rdcesx14194.race.sa
 *
 */

ods noproctitle;

proc gradboost data=MYCAS.DONOR ntrees=200 maxdepth=2
        learningrate=0.01;
    partition fraction(validate=0.3);
    target TARGET_B / level=nominal;
    input MONTHS_SINCE_ORIGIN DONOR_AGE IN_HOUSE INCOM
        MOR_HIT_RATE WEALTH_RATING MEDIAN_HOME_VALUE M
        PCT_OWNER_OCCUPIED PER_CAPITA_INCOME PCT_ATTRI
        PCT_ATTRIBUTE3 PCT_ATTRIBUTE4 PEP_STAR RECENT_
        FREQUENCY_STATUS_97NK RECENT_RESPONSE_PROP REC
        RECENT_CARD_RESPONSE_PROP RECENT_AVG_CARD_GIFT
        RECENT_CARD_RESPONSE_COUNT MONTHS_SINCE_LAST_F
        LIFETIME_PROM LIFETIME_GIFT_AMOUNT LIFETIME_GI
        LIFETIME_GIFT_RANGE LIFETIME_MAX_GIFT_AMT LIFE
        CARD_PROM_12 NUMBER_PROM_12 MONTHS_SINCE_LAST_
        FILE_AVG_GIFT FILE_CARD_GIFT / level=interval;
    input URBANICITY SES CLUSTER_CODE HOME_OWNER DONOR
        RECENCY_STATUS_96NK / level=nominal;
    ods output VariableImportance=Work._Gradboost_VarI
```

TRAINING OPTIONS

Number of factors:

○ Use 1 (Default)

● Use custom value

*Number of factors: 20

Learning step size:

○ Use 0.01 (Default)

● Use custom value

*Step size: 0.15

Maximum number of iterations:

○ Use 1 (Default)

● Use custom value

*Number of iterations: 20

It is very easy to create complex data mining and machine-learning models. (Above) Selecting model options in the middle panel generates code that can be edited and shared, in this case for a gradient boosting model. (Left) Running automatically generated code produces this factorization machine model.

secure, multiuser environment for concurrent access to data in memory. Many users can collaborate to explore the same raw data and build models simultaneously.

Data and analytical workload operations are automatically distributed across the cores of a single server or the nodes of a massive compute cluster, taking advantage of parallel architectures for blazingly fast speed. All data, tables and objects are held in memory as long as required, allowing for efficient in-memory processing.

Elastic scalability lets you run more experiments with more complex approaches on larger amounts of data. It increases confidence in the results, vastly improves productivity and enables more efficient model building.

Also, with built-in fault tolerance and memory management, advanced workflows can be applied to data, ensuring that processes always finish.

Powerful data manipulation and management

Take advantage of powerful data manipulation and management capabilities within the same distributed, in-memory environment to prepare data for analytics. Access data, join tables, subset and filter data, and create the final table for your machine-learning projects.

Data exploration, feature engineering and dimension reduction

Explore your data with descriptive statistics and powerful graphical programming. Discover data issues and fix them with advanced analytical techniques. Identify potential predictors quickly, reduce the dimensions of large data sets and easily create new features from your original data.

Modern statistical, data mining and machine-learning techniques

Apply powerful unsupervised and supervised learning algorithms, such as clustering, principal component analysis, linear and nonlinear regression, logistic regression, decision trees, random forests, gradient

Key Features (continued)

  - Large-scale data exploration and summarization through parallelized processing.
  - Ability to generate comprehensive descriptive statistics for your data quickly and easily.
  - Intelligent recommendation for variable measurement and role (categorical, numeric, interval and ID).
- Sampling:
  - Random and stratified; oversampling for rare events.
  - Creates indicator variable for sampled records for easy handling and tracking.

Data exploration, feature engineering and dimension reduction

- Large-scale binning of continuous features.
- High-performance imputation of missing values in features with user-specified values, mean, pseudo-median and random value of nonmissing values.
- Large-scale dimensions reduction for continuous and categorical features:
  - Reduces dimensionality for structured inputs and selects a subset of the original features to maximize predictive power of the supervised models.
  - Performs unsupervised variable selection by identifying a set of variables that jointly explain the maximum amount of data variance (covariance analysis).
- Large-scale principal components analysis (PCA):
  - Provides the eigenvalue decomposition, NIPALS and ITERGS algorithms.
  - Outputs principal component scores across observations.
  - Creates scree plots and pattern profile plots.
- Unsupervised learning with cluster analysis:
  - K-means clustering for continuous and nominal variables.
  - Various distance measures for similarity.
  - Automated estimation of optimal number of clusters.
  - Outputs cluster membership and distance measures across observations.

Model development with modern statistical, data mining and machine-learning algorithms

- Linear and logistic regression models with continuous and classification variables:
  - Supports any degree of interaction and nested effects, polynomial and spline effects.
  - Automated model selection based on forward, backward, stepwise, least angle regression and lasso selection methods.
  - Provides a variety of diagnostic statistics and automated model assessment.
- Generalized linear models:
  - Supports responses with variety of distributions, including binary, normal, Poisson and gamma.
  - Supports any degree of interaction as well as nested, polynomial and spline effects.
  - Automated model selection based on forward, backward, stepwise, least angle regression and lasso selection methods.
  - Provides a variety of diagnostic statistics and automated model assessment.
- Nonlinear regression:
  - Fits nonlinear regression models using least squares or maximum likelihood estimation.
  - Supports standard distributions for responses, such as binary, Poisson and normal.
  - Supports programming syntax for specifying customized distributions for responses.
  - Supports programming syntax for specifying models and expressions of parameters.
  - Provides a variety of optimization methods for parameter estimation.
- Decision trees:
  - Supports classification and regression trees with categorical and numerical features.
  - Provides the cost-complexity, C4.5 and reduced-error methods of pruning trees.
  - Automated pruning and final tree selection based on holdout data.
  - Automated handling of missing values, including surrogate rules.
  - Automated model fit assessment, including model-based (resubstitution) statistics.

boosting, neural networks and support vector machines to your structured and unstructured data, and quickly identify the champion model. With matrix factorization, you can build customized recommendation systems.

Get faster results with higher confidence by running automated tuning scenarios for your complex machine-learning algorithms. Using advanced optimization techniques, an integrated autotuning process searches through the possible combinations of parameter settings and returns the optimal set. (See an example on Page 1.)

Integrated text analytics

Designed with big data in mind, you can examine large collections of text documents in 13 languages. You can explore textual data to gain new insights about unknown themes and connections using powerful text preprocessing, natural language processing, topic detection and more. The integrated text analytics capabilities also enable data scientists to easily use the insights hidden in unstructured data for improved supervised learning.

Model assessment and scoring

Test different modeling approaches in a single run and compare results of multiple supervised learning algorithms with standardized tests to quickly identify champion models. Classification models can be evaluated using lift charts, ROC charts, concordance statistics and misclassification tables. When the champion model has been identified, operationalize analytics in distributed and traditional environments with automatically generated SAS score code.

Key Features (continued)

- Random forests for binary, nominal and interval labels:
  - Automated ensemble of decision trees predict a single target.
  - Automated distribution of independent training runs.
  - Automated intelligent tuning of parameter set to identify optimal model.
- Gradient boosting for binary, nominal and interval labels:
  - Automated iterative search for optimal partitioning of data in relation to selected label variable.
  - Automated generation of weighted averages for final supervised model.
  - Automated stopping criteria based on validated data scoring to avoid overfitting.
- Neural networks for binary, nominal and interval labels:
  - Intelligent defaults for most neural network parameters such as activation and error functions.
  - Customizable neural networks architecture and weights.
  - Ability to use an arbitrary number of hidden layers to support deep learning.
  - Automatic out-of-bag validation for early stopping to avoid overfitting.
  - Automated intelligent tuning of parameter set to identify optimal model.
- Support vector machines for binary labels:
  - Model training with linear and polynomial kernels.
  - Ability to apply the interior-point method and the active-set method.
  - Support for data partitioning for model validation.
  - Support for cross-validation for penalty selection.
- Factorization machines:
  - Develop recommender systems based on sparse matrices of user IDs and item ratings.
  - Apply full pairwise-interaction tensor factorization.
  - Supercharge models with time stamps, demographic data and context information.
  - Supports warm restart so you can update models with new transactions without having to fully retrain.

Integrated text analytics

- Supports 13 native languages out of the box (English, German, French, Italian, Spanish, Portuguese, Dutch, Russian, Finnish, Turkish, Japanese, Chinese and Korean).
- Automatically identifies term part of speech (more than 15 are system defined).
- Extracts standard entities such as location, time, date and address from predefined options.
- Detects noun groups and multiterm lists, and creates single terms for processing.
- Finds term variants automatically with synonym detection.
- Uses default start and stop lists to manage specific terms for parsing and downstream processing.
- Machine-learned topics represent the term-by-document, matrix-generated text processing as a structured numeric representation of the document collection.

Model assessment and scoring

- Supervised learning model performance statistics are automatically calculated for selected model with binary, nominal or interval label.
- Creates lift table for interval and categorical target.
- Creates ROC table for categorical target.
- Automated generation of SAS DATA step code for model scoring.
- Score statement for applying scoring logic to training, holdout data and new data.

To contact your local SAS office, please visit: sas.com/offices

sas
THE POWER TO KNOW.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. Copyright © 2016, SAS Institute Inc. All rights reserved. 108275_G31354.0616