Machine Learning and Scalable Analytics in SAS®

What Is Machine Learning?

Machine learning is a branch of artificial intelligence that is based on two things: mathematical algorithms and automation. Building analytic models can be automated using algorithms to "learn" from data in an iterative fashion. The "machine" (it's really an algorithm) learns from its mistakes in previous steps to derive the best results without human intervention.

This iterative aspect is important, because models don't get smarter by themselves. They learn from previous computations to produce the best results.

Statistics and data mining overlap with machine learning, but the latter is geared toward accuracy of predictions as opposed to inference and insights into causal effects. Machine learning models are sometimes called "black boxes" because they have low interpretability, but they can be very powerful. Machine learning is also more heavily focused on automated model building. Furthermore, a machine learning toolset can be augmented by practically effective heuristic methods like genetic and pattern search algorithms that enable meta-optimization of tasks like parameter selection and tuning.

Machine learning algorithms can be used for many day-to-day activities, such as fraud detection, real-time ads for web and mobile, text-based theme identification, next-best offers, equipment failure prediction, telematics, graph-based entity analysis, network intrusion detection, and email spam filtering.

How SAS Can Help

SAS includes many machine learning algorithms:

- neural networks
- decision trees and random forests
- gradient boosting and bagging
- multivariate adaptive regression splines
- support vector machines
- recommendation engines
- associations and sequence discovery
- Bayesian networks
- *k*-nearest-neighbor
- clustering and self-organizing maps
- pattern search and genetic algorithms
- kernel density estimation
- Gaussian mixture models
- singular value decomposition

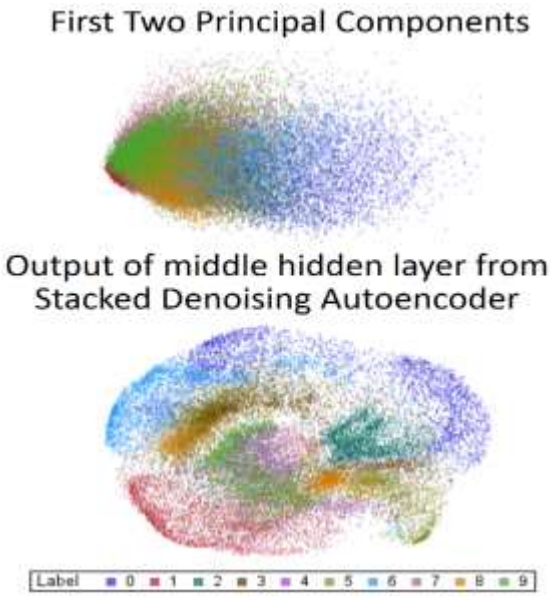
Example: Deep Learning in SAS® Enterprise Miner™

Deep learning is a machine learning field that is based on endowing large neural networks with multiple hidden layers in order to learn features that have strong predictive abilities.

"Autoencoders" (based on neural network techniques) are trained by using the same unlabeled inputs as both training examples and target labels. Because autoencoders use the training examples as targets instead of using the training labels themselves, they are a semi-supervised learning technique. When a large number of inputs is used in conjunction with a much smaller number of hidden units, the features that are extracted as outputs of the hidden units are a nonlinear projection of the training examples onto a lower-dimensional space.

Such features can be highly predictive of a training example's class label. These deep learning techniques are used to create predictive models for problems such as handwriting and facial recognition. Figure 1 shows a deep learning algorithm that differentiates nine digits that are automatically recognized from handwriting better than a traditional method such as principal component analysis.

Figure 1. Principal components and the nonlinear features extracted from a stacked denoising autoencoder.



Example: Product Recommendations with SAS® In-Memory Statistics for Hadoop

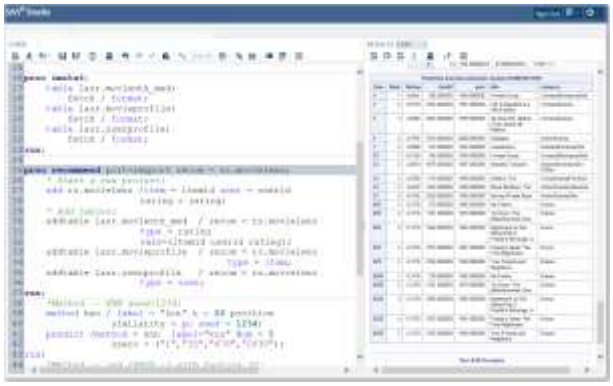
With a wealth of new data sources on customer behavior and preferences, many organizations are exploring recommender system techniques. These systems can provide personalized, meaningful recommendations in real time.

For building accurate implicit and explicit recommendation systems within Hadoop, SAS® In-Memory Statistics for Hadoop provides multiple machine learning and statistical algorithms:

- regression-based recommender
- associative rule mining recommender
- collaborative filtering recommender
- cluster-based recommender
- SVD-based recommender
- ensemble recommender
- average-based recommender

Statisticians and data scientists can prepare and explore data, build and test different models, and deploy the model for production scoring—all within Hadoop in one single, interactive programming environment. Training data for recommendation systems (often stored in Hadoop) are usually high-dimensional and very sparse. Users can apply different learning algorithms directly to the data stored in Hadoop, avoiding time-consuming data transfers and taking advantage of the parallel processing capabilities. SAS in-memory capabilities provide interactive and extremely fast computations on even very large data sets and regardless of the analytical algorithms that are applied to the data.

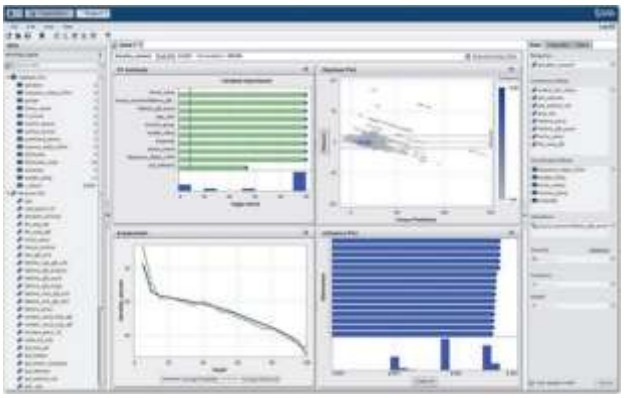
Figure 2: Generate personalized, meaningful recommendations in real time with a high level of customization in the new web-based programming interface—SAS® Studio.



Example: Interactive Data Exploration at Scale with SAS® Visual Statistics

Exploring big data presents new demands and complications that are made easier with large-scale visualization and interactivity. SAS® Visual Statistics enables statisticians and data scientists to quickly apply advanced algorithms to big data in an intuitive interface to discover insights and explore predictive models. Multiple users can customize models—by adding or changing variables, removing outliers, and so on—and quickly see how those changes affect model outcomes (see Figure 3).

Figure 3: Interactive exploration of multiple linear regression models with integrated model diagnostics.



For More Information

For more information about machine learning in SAS, visit the web page: www.sas.com/machine-learning.