› Conclusions Paper

Statistics and Machine Learning at Scale

New Technologies Apply Machine Learning to Big Data

Insights From the Analytics 2014 Conference

Featuring

Herbert Bucheli, Head of Business Analytical Services, Senior Director, Aduno Group

Wayne Thompson, Manager of Data Sciences Technologies, SAS

sas
THE POWER TO KNOW.

Contents

Introduction

Imagine getting into your car and saying, "Take me to work," and then enjoying an automated drive as you read the morning paper. We're not there yet. But we're closer than you think. Google has already developed a prototype for a driverless car in the US.

Driverless cars are just one example of machine learning. It's used in countless applications, including those that predict fraud, identify terrorists, recommend the right products to customers at the right moment, and correctly identify a patient's symptoms in order to recommend the appropriate medications.

The concept of machine learning has been around for decades. What's new is that it can now be applied to huge quantities of data. Cheaper data storage, distributed processing, more powerful computers and the analytical opportunities available have dramatically increased interest in machine learning systems.

In this conclusions paper, based on a presentation given at the Analytics 2014 Conference, Wayne Thompson, Manager of Data Sciences Technologies at SAS, introduces key machine learning concepts and describes new SAS solutions – SAS In-Memory Statistics for Hadoop and SAS Visual Statistics – that enable machine learning at scale. In addition, Herbert Bucheli, Head of Business Analytical Services, Senior Director, Aduno Group, shares his experiences using a recommendation engine to differentiate his firm's new customer loyalty program.

Defining Data Science, Machine Learning and Statistics

As organizations gather big data, they're turning to data science to extract knowledge and meaning from it. Data science incorporates and builds on techniques and theories from many disciplines, including statistics, data mining, machine learning, artificial intelligence and more.

The interdisciplinary nature of data science means that it demands teams of practitioners with expertise in a variety of disciplines. For example, said Thompson, "In addition to knowing SAS, a lot of people on our team know programming languages such as Python and Java because you need to be proficient in computer science to do a good job with data science."

Within data science, machine learning is a branch of artificial intelligence that focuses on getting computers to act without being explicitly programmed. The idea is to automate the building of analytic models that use algorithms that learn from data interactively. By choosing better models, you can improve results over time with less human intervention. These models can then be used to produce reliable, repeatable decisions.

As Thompson explained, "Machine learning focuses on the construction and study of systems that can learn from data to optimize a performance function, such as optimizing the expected reward or minimizing loss functions. The goal is to develop deep insights from data assets faster, extract knowledge from data with greater precision, improve the bottom line and reduce risk."

Considerable overlap exists between statistics and machine learning. Both disciplines focus on studying generalizations (or predictions) from data. "The big difference between statistics and machine learning," Thompson explained, "is that statistics focuses more on inferential analysis or hypothesis testing to make predictions about a larger population than the sample represents. Statistics also looks at things like parameter estimates, error rates, distribution assumptions and so forth to understand empirical data with a random component.

"Machine learning," Thompson continued, "uses massive amounts of observational data and, as a branch of artificial intelligence, focuses on automation. [It focuses on] the algorithms, such as a random forest or gradient boosting, to automatically handle things like missing values, finding interactions and so forth."

Central to machine learning is the idea that with each iteration, the algorithm will learn from the data. Said Thompson, "To measure whether or not you're improving performance, you look at an objective function, such as minimizing a loss function. The algorithm iterates through the data until a convergence criterion is met. You typically use holdout data to see if you are overfitting."

Types of Machine Learning Algorithms

Four different types of machine learning algorithms are available that can be organized into a taxonomy based on the desired outcome of the algorithm or the type of input available for training the machine. Thompson noted, "The terminology used in machine learning is different than that used for statistics. For example, in machine learning, a target is called a label, while in statistics it's called a dependent variable."

The key types of machine learning include:

- Supervised learning.
- Unsupervised learning.
- Semisupervised learning.
- Reinforcement learning.

Supervised Learning

"Most machine learning – about 70 percent – is supervised learning," said Thompson. Supervised learning algorithms are "trained" using labeled examples where the desired output is known. Supervised learning is commonly used in applications that use historical data to predict likely future events.

For example, it can anticipate which credit card transactions are likely to be fraudulent or which insurance customer is likely to make a claim. In the case of fraud, you have known customers were fraudulent and not in your training data. The learning algorithm receives a set of inputs along with the corresponding correct outputs, and the algorithm learns by comparing its actual output with the correct outputs so it can find errors and modify the model accordingly. The inputs are called features in machine learning. In the case of fraud, example features may be account balances, number of daily transactions, and so on. Through methods like classification, regression, prediction and gradient boosting, supervised learning uses the inputs to predict the values of the labels. Applying the model to new cases to classify the transactions as either fraudulent or not is called scoring.

Unsupervised Learning

About 10 to 20 percent of machine learning is unsupervised learning, although this area is growing rapidly. Unsupervised learning is a type of machine learning where the system operates on unlabeled examples. In this case, the system is not told the "right answer." The algorithm tries to find a hidden structure or manifold in unlabeled data. By contrast with supervised learning and reinforcement learning, the examples given to the learner have no explicit target outputs or reward signals associated with each input.

"The goal of unsupervised learning," Thompson said, "is to explore the data to find intrinsic structures within it using methods like clustering or dimension reduction. [Unsupervised learning] works very well on transactional data."

The intrinsic structure and associated unsupervised learning methods vary depending on the nature of the data. For example, the data in a Euclidean space can be structurally modeled by a probability density, and its dimensionality can be reduced using methods such as k-means clustering, Gaussian mixtures and principal component analysis (PCA); while the data in a general topological space is only locally Euclidean, and its structure is better modeled as a nonlinear manifold whose dimension reduction can be achieved by ISOMAP, local linear embedding (LLE), Laplacian eigenmaps, kernel PCA, and other methods. In addition, matrix factorization, topic models and graphs are popular structural models for unsupervised learning of text, imagery and social media data.

Semisupervised Learning

Semisupervised learning is used for the same applications as supervised learning. But this technique uses both labeled and unlabeled data for training – typically, a small amount of labeled data with a large amount of unlabeled data.

This type of learning can be used with methods such as classification, regression and prediction. Semisupervised learning is useful when the cost associated with labeling data is too high to allow for a fully labeled training process, but acquiring unlabeled data is relatively inexpensive.

Semisupervised learning may be interpreted in at least two different ways. In the first interpretation, one uses unlabeled data to inform a computer algorithm of the structural information of the data that is relevant to supervised learning, which is considered the primary goal. In this view, unlabeled data provides side information to help enhance supervised learning when labels are insufficient. In the second interpretation, the primary goal is unsupervised learning (clustering, for example), and labels are viewed as side information (cluster indicators in the case of clustering) to help the algorithm find the right intrinsic data structure. In this case, the labels are particularly helpful when the intrinsic data structure is not very clear and poses challenges to regular unsupervised learning methods.

Early examples of this include image analysis – for example, identifying a person's face on a webcam – textual analysis, and disease detection.

Reinforcement Learning

With reinforcement learning, the algorithm discovers for itself which actions yield the greatest rewards through trial and error. Reinforcement learning has three primary components:

1. The agent – the learner or decision maker.
2. The environment – everything the agent interacts with.
3. Actions – what the agent can do.

Said Thompson, "The objective is for the agent to choose actions that maximize the expected reward over a given period of time. The agent will reach the goal much quicker by following a good policy, so the goal in reinforcement learning is to learn the best policy." Reinforcement learning is often used for robotics and navigation.

Reinforcement learning has strong connections with optimal control, statistics and operational research. Markov decision processes (MDPs) are popular models used in reinforcement learning. MDPs assume the state of the environment is perfectly observed by the agent. When this is not the case, one can use a more general model called partially observable MDPs (or POMDPs) to find the policy that resolves the state uncertainty while maximizing the long-term reward.

Generalization, Evaluation and Model Selection

Regardless of the method, all types of machine learning develop models that enable the learning machine to perform accurately on new, unseen examples or tasks. Then the machine can improve these models by learning over time.

"Developing the right model to fit the data is like Goldilocks," said Thompson. "We want the fit to be not too much, not too little, but just right." Figure 1 is an example of "too little" fit, or underfitting, where the predictor is too simplistic to capture salient patterns in the data. It will not do a good job of resolving future examples. Said Thompson, "It's nice to have parsimonious models with very few terms, but this model doesn't do a good job of fitting."

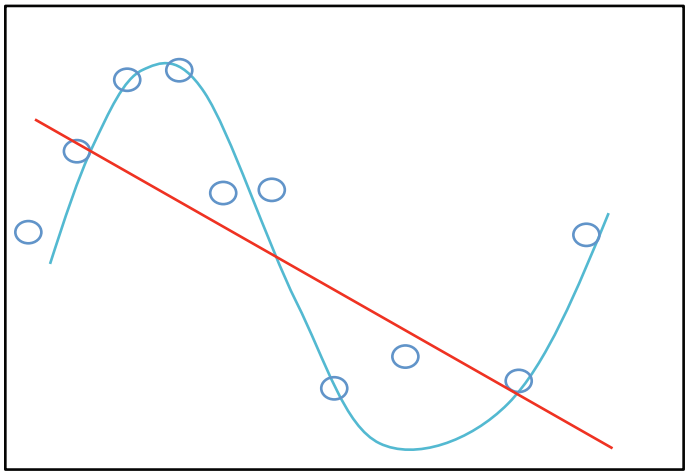

Figure 1: Underfitting.

Figure 2 shows overfitting, where the predictor is too complex. Thompson explained, "This model won't generalize well when I try to score the new population. I want something with fewer parameters – perhaps using penalty functions or holdout functions – to find models that fit the data better."

Data scientists often use average squared error or the misclassified rate of holdout data to measure if the model is overfitting or not. But Thompson noted that, "Some machine learning algorithms can look at your model and see whether you're using too many variables and can automatically adjust the model to use fewer variables."

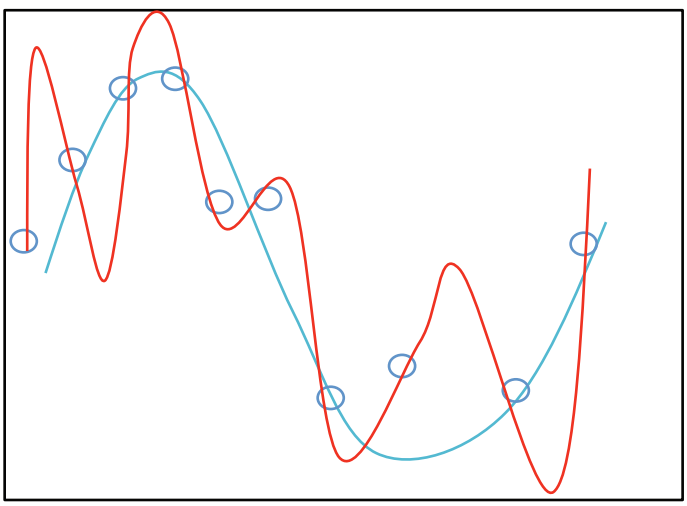

Figure 2: Overfitting.

Choosing a Model

Data scientists need to be able to look at data of any complexity and size and build a model that sizes well to that data. They may need to look at all the data or a subset to create an accurate model.

One of the more powerful machine learning algorithms is random forest, which has become a powerful tool for data mining. A random forest takes individual decision trees and combines them. When a new input is entered into the system, it runs down all of the trees. The result is either an average or a weighted average of all the terminal nodes that are reached.

Thompson explained, "If I'm fitting around a random forest, I'll build decision trees on many random subsets of the data and then average them to build the final model. I also split on different variables at each split point in the construction of the decision tree. If I have 100 variables, I might look at only 10 variables at random at each split point; so I'm not only permutating the observations, but also the data." While single decision trees can suffer from high variance or high bias, this averaging balances the two extremes.

New technologies, such as in-memory analytics, allow queries of data residing in a computer's random access memory (RAM) and across a distributed computing environment to divide processing across multiple computers. This allows data scientists to build random forests faster than ever.

In using machine learning models for business applications of data mining, Thompson observed that, "Customers often don't know the expected profit or cost from working with their customers. When I use SAS® Enterprise Miner™ for predictive modeling, I try to select models that maximize profit or revenue. For example, if we're making a decision about what to do with a customer, it's not a yes or no decision. Rather, I want to determine the expected outcome in revenue associated with that decision. That's really important to add to your models."

Model Evaluation and Selection

Once you've built a model, you need to validate it to determine whether it can make effective predictions. Typically data scientists use a training data set to develop the model, and then use known out-of-sample data to test the model.

If not enough data is available to allow some of it to be kept back for testing, Thompson said, "People typically do random subsampling or random stratified subsampling of the data. You can also use techniques such as k-fold cross-validation or leave one out (LOO) cross-validation."

But Thompson noted, "If I have a million observations and an event rate of 1 percent, I find it useful to evaluate all the data to understand whether I can classify or predict the event. In certain cases, such as fraud, where the event rate is small, I find that using oversampling to correct for a bias in the original data set and developing bio samples that put more weight on looking at the rare event derive better models."

Some models are developed for use in database marketing to score customers. For example, a marketer needs to know which customers are most likely to purchase a product to target special offers at those customers. Marketing efforts can also have a rather small event rate, commonly called response rate – often in the range of 1 percent.

"If I'm evaluating models that I use in database marketing," said Thompson, "I'd use statistics that look at lift or how well the model performs at a particular depth of file. I may not be interested in the overall misclassification rate for my model. I only have 1 percent responders, so the null model is 99 percent accurate. So here I like to first develop predictions, generate a prediction profile with regard to a lift, and select models that maximize lift at depth of file."

Viseca: Using a Recommender Engine to Drive Competitive Advantage

One customer taking advantage of machine learning and recommendation engines, in particular, is Viseca Card Services SA, a Swiss credit card company with 1.3 million cards. Viseca is part of the Aduno Group,[1] which also offers point-of-sale (POS) payment solutions, cloud-based POS terminals, as well as consumer credit and leasing services.

Within Switzerland, Aduno Group is unique in delivering services that give it direct relationships with both credit cardholders and merchants. The company is attempting to use these relationships to differentiate itself through a new loyalty program called "surprize." This isn't the first loyalty program in Switzerland, but the company views the program as a necessity to engage customers.

[1] The Aduno Group offers the entire range of products and services for cashless payment from a single source. By integrating card-issuing activities (Viseca Card Services SA), card acceptance agreements and payment terminals (Aduno Payment Services), the Aduno Group is the only company in Switzerland to bring customers, merchants, partners and banks together in this unique manner. The personal loan and leasing business (cashgate AG and Revi-Leasing und Finanz AG) rounds out the group's product and service offerings.

"This loyalty program offers all the standard features," said Bucheli." Cardholders collect points by making purchases anywhere in the world and redeem them on the redemption platform or redemption shop for items from us and from our partners. They also get special offers, including coupons, rebates or a percentage off tailored to them based on behavioral insights that increase the value of the program."

Merchants that participate in the program benefit from Viseca's ability to create marketing campaigns that target Viseca's customer base to develop existing customer relationships and create relationships with new customers. Through the analysis of these marketing campaigns, Viseca grants merchants access to customer intelligence they can use to optimize targeted campaigns and gain valuable insights about their customers.

To compete with other loyalty programs, Viseca enrolls cardholders in its program and encourages them to use it. Said Bucheli, "Engaging customers requires good offers, which means we need an attractive partner network. To get good partners, we need to engage customers. It's a chicken and egg problem."

For Viseca, engaging customers is a three-step process. First, Viseca must raise awareness by communicating the availability of the program to cardholders. Next, there's a growth phase, where it can build the perceived value of the program so cardholders use it. Finally, a consolidation phase occurs that encourages ongoing use of the program so it becomes a habit.

"We believe our recommender system can help us make suitable recommendations to customers to build the perceived value of the program," said Bucheli.

Viseca's recommender system relies on an event processing engine that tracks events, processes the events based on business rules and then takes action. Roughly 70 applications run on this engine, which sits directly on the data warehouse.

What is an event? Said Bucheli, "A customer clicking on an item, searching for a term, entering the site and so on – all are events.

"The recommender system observes cardholder behavior on the redemption platform by event tracking," Bucheli continued. "And the recommender system offers a web service that the redemption platform calls for item recommendations. Because all of these components sit on the same DWH platform, we can also use and include data from our data warehouse.

"What's important," said Bucheli, "is that we've implemented a closed loop so we can learn from each customer interaction. We track items the cardholders bought, as well as whether they click on detailed information for the items we present to them. We believe there's a strong correlation between item clicks and item orders.

"We also include the number of impressions of items without clicks as small signals of negative user feedback. In the future, we'll be able to look at items we've presented, including context information such as page and position, as well as search terms, filter and sorting criteria."

The next phase will be to pair this information with other data, including demographic data, transaction history, customer preferences, CRM (call center and campaign information) and other data. They will then be able to test prebuilt models on the new data included in the analysis.

SAS® Analytics Solutions

SAS solutions that support machine learning at scale and are also interactive include SAS In-Memory Statistics for Hadoop and SAS Visual Statistics.

SAS® LASR™ Analytic Server

The SAS LASR Analytic Server is the back end for solutions such as SAS In-Memory Statistics for Hadoop and SAS Visual Statistics. This server is an analytics platform that provides a secure, multiuser environment for concurrent access to data loaded into memory. Said Thompson, "This server loads data into memory once, and users can then repetitively analyze the data without dropping it down the disk.

"The advantage of the SAS LASR Analytic Server's in-memory processing is speed," Thompson explained. "More importantly, it allows data scientists to work with the data interactively. I can try many different permutations, create new columns, summarize data, look at outlier detection and description, build models in a champion/challenger type tournament and evaluate them. The ability to work with the data in memory is not just about speed, but about allowing data scientists to be flexible and learn from the data."

The SAS LASR Analytic Server can run in a distributed computing environment or on a single machine. For distributed deployments, the SAS LASR Analytic Server supports the Hadoop Distributed File System as a collocated data provider. Hadoop spreads data over large clusters of commodity servers

and performs processes in parallel. It also detects and handles failures. In addition to low distributed hardware cost and data redundancy, Hadoop delivers parallel processing to process huge amounts of data, scalability, and flexible storage for structured and unstructured data.

SAS® In-Memory Statistics for Hadoop

"Many people are aware of SAS Visual Analytics solutions, which provide a complete platform for analytics visualization, enabling users to identify patterns and relationships in data that weren't evident before," said Thompson. "SAS Visual Analytics is point and click. But data scientists like to code. So SAS developed SAS In-Memory Statistics for Hadoop."

SAS In-Memory Statistics for Hadoop enables multiple users to concurrently manage and prepare data stored in Hadoop. They can explore and visualize this data, develop accurate statistical and machine learning models quickly, and deploy and execute these models in their Hadoop ecosystem.

Said Thompson, "This product focuses on Hadoop not because it can't run elsewhere, but because our customers are moving heavily to Hadoop as an open source transactional system. SAS wants to be No. 1 in analytics workload on Hadoop."

Rather than requiring organizations to extract data from Hadoop to a SAS environment, SAS In-Memory Statistics for Hadoop brings the analytics to the Hadoop environment. This minimizes the need to move big data and allows SAS to take advantage of the computing power offered by the distributed, in-memory processing environment.

Figure 3 shows how SAS solutions interact with Hadoop. Said Thompson, "The Hadoop cluster on the left can run on the Cloudera, Hortonworks, Apache, IBM BigInsights and Pivotal platforms. All the data is in Hadoop, but SAS resides inside the Hadoop cluster. SAS is installed on data nodes and a head node. When data is loaded into memory, it's automatically distributed across a cluster, so the solution can use all the individual nodes to do the analysis very quickly."

Thompson continued, "One of the big advantages is that the data nodes talk to each other using something called message processing interface (MPI). Our competitors don't do this, so they can't parcel out the processing to divide and conquer. This means SAS is much faster."

Another unique advantage of SAS In-Memory Statistics for Hadoop is that statisticians and data scientists do not need to piece together different programming languages or products to manage the variety of analytical lifecycle tasks in Hadoop.

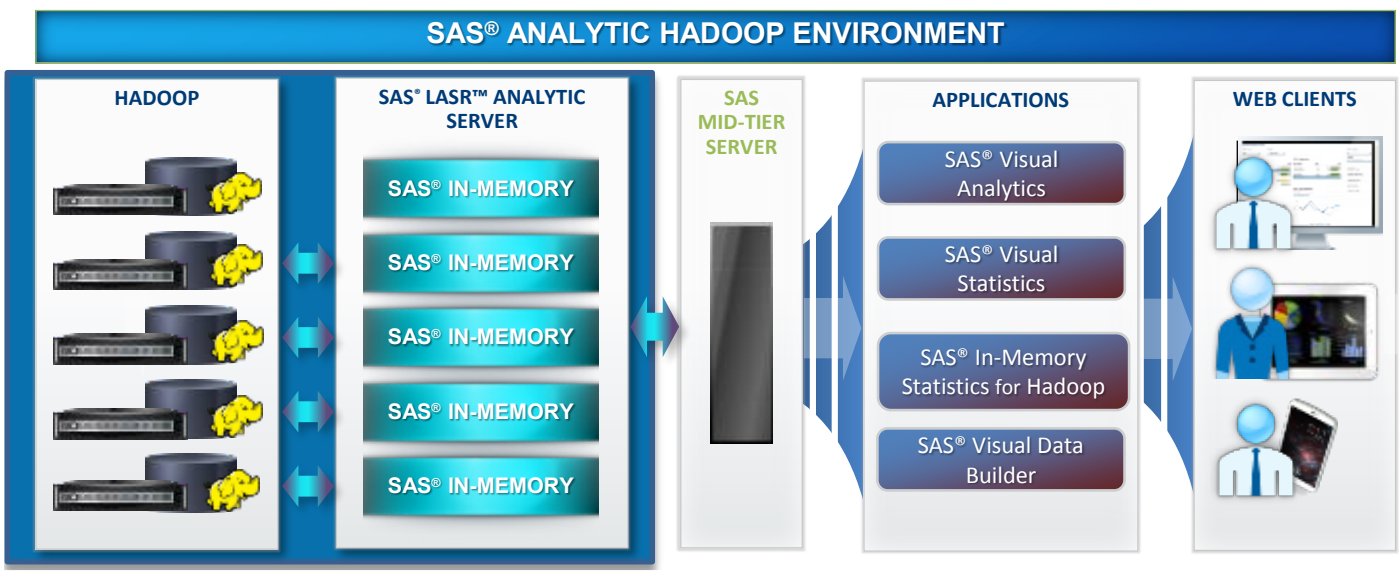

Figure 3: How SAS In-Memory Statistics for Hadoop and SAS LASR Analytic Server work with Hadoop.

Said Thompson, "We can read the data just once into memory and go through that full analytical life cycle. We're the first to invent this analytical life cycle – and no other vendor is doing that. We've done a great job of going from data management to exploration, to model building, to model governance, to model monitoring."

Machine Learning and SAS® In-Memory Statistics for Hadoop

SAS In-Memory Statistics for Hadoop comes with a number of state-of-the art machine learning techniques. Here are just a few:

- **Unsupervised learning.** For unsupervised learning, the solution has density-based spatial clustering with noise (DBSCAN) and k-means clustering. It also uses singular value decomposition (SVD) in the recommendation engine and to do text parsing to develop topics.
- **Supervised learning.** For supervised learning, the solution offers random woods, the SAS version of random forest. It also offers generalized linear models with many different link functions and logistics, as well as decision trees that use the C4.5 methodology.
- **Recommendation engine.** The SAS recommendation system generates recommendations in real time. Collaborative filters, matrix factorization, hybrid models and affinity analysis can be used to build customized recommendation systems.

With these state-of-the-art machine-learning techniques, multiple users can explore and use various analytic approaches to build models, quickly determine the best ones, and then evaluate them.

SAS® Visual Statistics

SAS Visual Statistics is aimed at business analysts, statisticians and data scientists who want to quickly test variables and build and refine models to optimize performance. Said Thompson, "SAS Visual Statistics extends SAS Visual Analytics to build descriptive and predictive models using techniques like regression, clustering, interactive decision trees and so forth. The presentation layer is fun to use, very interactive and fast."

With SAS Visual Statistics running on SAS LASR Analytic Server, multiple users can collaborate and interact with the same data sets. They can develop, test and refine their models quickly because the in-memory engine provides a highly scalable processing environment and the visualization techniques let them see how their models are being fitted and evaluated on the fly.

Conclusion

With more and more data available, machine learning techniques are becoming increasingly popular as they get better at looking at massive amounts of data. SAS In-Memory Statistics for Hadoop and SAS Visual Statistics running on the SAS LASR Analytic Server allow data scientists to use machine learning techniques to take advantage of the most advanced in-memory distributed computing platforms to uncover insights within big data. Rather than having to work in batch mode, these solutions give statisticians and data scientists the tools to learn about data interactively. Said Thompson, "Because they can submit data on the fly and make adjustments on the fly, SAS solutions allow machine learning techniques to mimic human thought."

Speakers

Herbert Bucheli, Head of Business Analytics Services, Senior Director, Aduno Group

Herbert Bucheli is Senior Director at the Aduno Group, a leading Swiss credit card issuing and acquiring company, and heads the Business Analytics department. Bucheli is responsible for all activities for data analytics within the Aduno Group, including fraud and risk scoring, marketing optimization, and the development of analytics-based customer services. Currently he is engaged in the development of a recommender system for a recently launched loyalty program. Bucheli holds an MSc in physics from the Swiss Institute of Technology in Zurich and now has more than 16 years of analytics and data warehousing experience.

Wayne Thompson, Manager of Data Science Technologies, SAS

Wayne Thompson, Manager of Data Science Technologies at SAS, is a globally renowned presenter, teacher, practitioner and innovator in the fields of data mining and machine learning. He has worked alongside the world's biggest and most challenging organizations to help them harness analytics to build high-performing organizations. Over the course of his 20-year tenure at SAS, he has been credited with bringing to market landmark SAS Analytics technologies (SAS Text Miner, Credit Scoring for SAS Enterprise Miner, SAS Model Manager, SAS Rapid Predictive Modeler, SAS Scoring Accelerator for Teradata, SAS High-Performance Data Mining and SAS Analytics Accelerator for Teradata). Current focus initiatives include easy-to-use, self-service data mining tools for business analysts, outlier detection and description, entity analytics, and recommendation engines with a heavy focus on highly interactive, in-memory analytics optimized for Hadoop.

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

107284_S127971.1014

THE POWER TO KNOW®