CONCLUSIONS PAPER

Statistics and Machine Learning at Scale

New technologies apply machine learning to big data

sas
THE POWER TO KNOW.

Contents

Content Provider

This paper is based on several presentations given by Wayne Thompson, Manager of Data Science Technologies at SAS. Thompson is a globally known presenter, teacher, practitioner and innovator in the fields of data mining and machine learning. Over the course of his 25-year tenure at SAS, he has been credited with bringing to market landmark SAS Analytics solutions, and has worked to solve our customer's most challenging analytical problems.

Introduction

Imagine getting into your car and saying, "Take me to work," and then enjoying an automated drive as you read the morning news. We are getting very close to that kind of scenario, and companies like Ford expect to have production vehicles in the latter part of 2020.

Driverless cars are just one popular example of machine learning. It's also used in countless applications such as predicting fraud, identifying terrorists, recommending the right products to customers at the right time, and correctly identifying medical symptoms to prescribe appropriate treatments.

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 new analytical opportunities have dramatically increased interest in machine learning systems. Other reasons for the increased momentum include: maturing capabilities with methods and algorithms refactored to run in memory; the reduced cost of abundant computing power; and the simple fact that there is more data for computers to learn from.

This paper is based on presentations given over the last few years. Wayne Thompson, Manager of Data Science Technologies at SAS, introduces key machine learning concepts, explains the correlation between statistics and machine learning, and describes SAS® solutions that enable machine learning at scale.

Machine learning focuses on getting computers to act without being explicitly programmed. The idea is to automate the building of analytical models that use algorithms to learn from data interactively.

Defining Statistics and Machine Learning at Scale

As organizations gather big data, they're turning to a breadth of mathematical and computer science technologies to extract knowledge and meaning from it. This "data science," as it's been called the last few years, incorporates and builds on techniques and theories from many disciplines, including statistics, data mining, machine learning, artificial intelligence and more.

Within data science, machine learning focuses on getting computers to act without being explicitly programmed. The idea is to automate the building of analytical models that use algorithms to 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.

Thompson explains, "Machine learning focuses on the construction and study of systems that can learn from data to improve 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."

The concept of machine learning has been around for decades. A neural learning technique called the perceptron algorithm was developed as far back as 1958. The SAS DISCRIM procedure has been used for k-nearest-neighbor discriminant analysis

since 1979. (This early machine learning procedure was written by Jim Goodnight, SAS co-founder and CEO.) But neural network research and other automated machine learning techniques made slow progress until the early 1990s when the intersection of computer science and statistics reignited the popularity of these ideas.

What sets machine learning and statistics apart?

Considerable overlap exists between statistics and machine learning. Both disciplines focus on studying generalizations (or predictions) from data. But to understand machine learning, it's helpful to recognize the role that statistical analysis has played over the years.

Statistics has many goals. One of the most important is describing relationships between data attributes, which is realized through statistical modeling. The modeling phase creates a unique overlap between the fields of statistics and machine learning. In both areas, you are trying to understand the data that uses a model to best explain those relationships.

There are also differences. Inferential statistics makes assumptions that try to simplify the real world, and hypothesis testing makes predictions about a larger population than the sample represents. Statistics looks at things like parameter estimates, error rates, distribution assumptions and so forth to understand empirical data with a random component.

This is in contrast to machine learning, where all unique distributions do not have to be specified. Statistics also lends itself to smaller data environments where there may not be as many attributes or volumes of data. On the other hand, machine learning uses massive amounts of observational data with an emphasis on automation. It focuses on algorithms, such as a random forest or gradient boosting, to automatically handle things like missing values, find interactions, etc.

Central to machine learning is the idea that with each iteration, an algorithm learns from the data.

Central to machine learning is the idea that with each iteration, an algorithm learns from the data. Says 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."

What does scale have to do with it?

Let's first ask: What does "scale" mean?

Scale means a variety of things to different people within an organization. But when considering the world of statistics and machine learning, it can mean dealing with more data, more attributes and more variables. Scale can also mean the ability to process models faster. Or, using more techniques on that same set of large or high-velocity data. Lastly, it could entail creating many more models. Instead of creating 50 models per year, an organization could be asked to create 3,000 or many more. The desire to create better customer experiences or more targeted offers requires analyzing more granular segments of data. These issues all contribute to the importance of ability to scale. It means decisions that depend on huge quantities of data and numerous complex models can still be made quickly.

Types of Machine Learning Algorithms

Four different types of machine learning algorithms are available. They 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 notes, "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 are:

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

Supervised learning

Most experts estimate that approximately 70 percent of machine learning is supervised learning. These 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 know some transactions are fraudulent but are 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. The examples given to the learner have no explicit target outputs or reward signals associated with each input.

"The goal of unsupervised learning," Thompson says, "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). 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 incorporates 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 of 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, unlabeled data is used to inform an algorithm of the data's structural information that is relevant to supervised learning, which is considered the primary goal. In this view, unlabeled data provides side information to enhance supervised learning when labels are insufficient. In the second interpretation, the primary goal is unsupervised learning (clustering, for example). 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 include image analysis (e.g., 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.

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, a more general model called partially observable MDPs (or POMDPs) can be used to find the policy that resolves the state of uncertainty while maximizing the long-term reward.

Creating, Evaluating and Selecting Models for Machine Learning Applications

Regardless of the learning method used, the goal is to enable your models to perform accurately on new, unseen examples or tasks. And then, the machine improves the models by learning over time.

Thompson says, "Developing the right model to fit the data is like Goldilocks. 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. "It's nice to have parsimonious models with very few terms," says Thompson, "but this model doesn't do a good job of fitting."

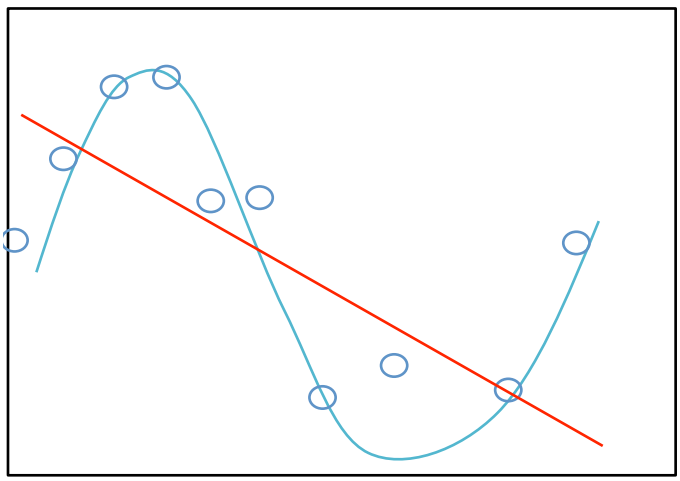

Figure 1: Underfitting.

Figure 2 shows overfitting, where the predictor is too complex. Thompson explains, "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 notes, "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."

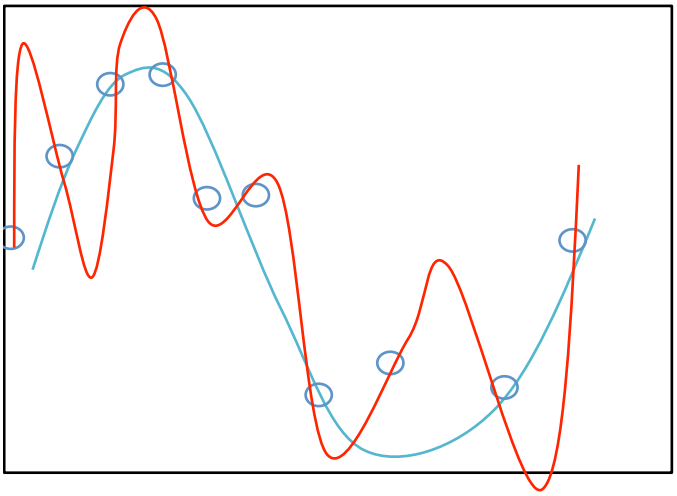

Figure 2: Overfitting.

Model building

Data scientists need to be able to look at data of any complexity and build a model that sizes well to that data. They may need to look at all the data, or perhaps only 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 explains, "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 permutating not only 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 and process random forests faster than ever.

In using machine learning models for business applications of data mining, Thompson says, "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. If we're making a decision about what to do with a customer, it is 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, a typical approach is to perform 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 notes, "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 produce 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. Thompson says, "If I'm evaluating models that I use in database marketing, 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. 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."

Machine Learning in Action

Machine learning has long been used in predictive analytics. Typical applications include preventing churn, identifying fraud and reducing risk. The technologies are also being used to predict consumer demand, recommend next-best offers and detect defects as early in the manufacturing process as possible. There is increasing interest in using machine learning for cybersecurity and bio-imaging analysis to improve medical outcomes.

Manufacturing optimization

High-tech manufacturing already uses robotics extensively, and streaming sensor and image data play a key role in finding where and when defects are looming. New semiconductor manufacturing processes include 3-D printing to generate chips. In the digital printing of semiconductor components, a one in a billion failure rate for droplets doesn't sound like a bad statistic. But when you consider that up to 50 million droplets can be pushed per second, that defect rate translates to one defect every 20 seconds.

Let's look at how a major global semiconductor manufacturer is using SAS machine learning techniques (specifically pattern detection) to improve wafer production. They start by training the machine to learn what a good wafer looks like. A wafer is a thin slice of semiconductor material used in electronics for the fabrication of integrated circuits. Each wafer is defined by more than 90,000 measurements. Data comes from images taken of the wafers. An even, smooth surface (good quality) has the same pixel values over the entire surface. Deviations from that base value are troughs and peaks, which result in an uneven surface (poor quality). Using image processing, patterns of defects are identified.

A major global semiconductor manufacturer is using SAS® machine learning techniques (specifically pattern detection) to improve wafer production.

Once the defects are known to the computer, pattern detection techniques match new wafers to the dictions of defects, and bad wafers are automatically detected. Today's new technologies and machine learning will continue to enhance quality control in manufacturing.

SAS® Analytics Solutions for Machine Learning

SAS has a long history in generalized statistics. The company was founded in 1976, based on statistical software that was developed and leased to agricultural departments to analyze the effect of soil, weather and seed varieties on crop yields. Over the decades, SAS has been defined as a leader in the world of descriptive, predictive and prescriptive analytics, and holds the world's largest market share of the analytics market.

SAS offers a variety of analytical solutions, from desktop to enterprise. There are many options for specific types of analysis, including forecasting and operations research, and solutions for specific industry sectors. However, two solutions were created specifically for data mining and machine learning.

SAS® Visual Data Mining and Machine Learning

SAS Visual Data Mining and Machine Learning provides unified, in-memory analytics processing that is cutting-edge in terms of multi-tenant performance, elasticity and resilience. It includes modern machine learning algorithms, such as extreme gradient boosting and factorization machines. Feature engineering and data reduction techniques identify potential predictors, reduce the dimensions of large data sets and create new features from your original data to produce more reliable outcomes.

Additionally, automated intelligent autotuning for selected machine learning models provides fast and easy identification of optimal parameter settings for maximized model accuracy and improved efficiency. You can code in Java, R, Python or Lua, and use REST APIs to call algorithms and capabilities in SAS Visual Data Mining and Machine Learning from third-party applications.

SAS Visual Data Mining and Machine Learning provides:

- An interactive, web-based programming environment.
- Highly scalable, in-memory analytical processing.
- Analytical data preparation.
- Data exploration, feature engineering and dimension reduction.
- Model development with modern statistical, data mining and machine learning algorithms.
- Integrated text analytics.
- Model assessment and scoring.

The solution takes advantage of the SAS® Viya™ engine – a modernization of the SAS Platform. SAS Viya is optimized for multipass analytical computations and provides a secure, multiuser environment for concurent 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

processing for extremely fast speeds. All data, tables and objects are held in memory as long as required, allowing for efficient in-memory processing. With built-in fault tolerance and memory management, advanced workflows can be applied to data, ensuring that processes always finish.

SAS® Enterprise Miner™

SAS Enterprise Miner has a process-flow GUI based on SAS®9, with drag-and-drop task-oriented icons and prompting wizards for data mining and machine learning algorithms (both supervised and unsupervised).

Decision trees, bagging and boosting, time series data mining, neural networks, memory-based reasoning, hierarchical clustering, linear and logistic regression, associations, sequence and web path analysis are all included. The breadth of analytical algorithms also extends to industry-specific algorithms (such as credit scoring) and state-of-the-art methods (such as gradient boosting and least angular regression splines).

The software also includes data preparation techniques, variable selection methods, text mining approaches, model assessment and numerous other tasks. All of these SAS Enterprise Miner capabilities make it very convenient to take an iterative approach to data mining and machine learning. After viewing the results of any task, simply make changes to the task property settings or parameter values and rerun the relevant task(s).

To take advantage of the enhancements SAS Viya brings to the SAS Platform, including high availability, faster in-memory processing, image data types and native cloud support, you can also submit and execute SAS Viya code directly in a SAS Enterprise Miner process flow. And, you can easily integrate R and Python code inside of a SAS Enterprise Miner process flow diagram. This enables you to perform data transformation and exploration, as well as train and score supervised and unsupervised models, in other programming languages. You can then integrate the results, assess your R or Python models and compare them to models generated by SAS Enterprise Miner to find the best performer.

With the ability to submit data modifications on the fly and adjust analytical models while they're still running, SAS® solutions enable machine learning techniques to closely mimic human thought.

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 solutions allow data scientists to use machine learning techniques to take advantage of the most advanced, in-memory distributed computing platforms to quickly uncover insights within big data. These solutions give statisticians and data scientists the tools to learn about data interactively and create advanced analytical models quickly. They can work in the environment they feel most comfortable with and still take advantage of trusted SAS algorithms and modern machine learning techniques.

With the ability to submit data modifications on the fly and adjust analytical models while they're still running, SAS solutions enable machine learning techniques to closely mimic human thought. At the same time, they take advantage of extreme processing power to produce faster insights than ever before.

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

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 © 2017, SAS Institute Inc. All rights reserved. 107284_G52461.0917