Brought to you by:

IBM

Machine Learning

for dummies®

A Wiley Brand

Understand machine learning fundamentals

Make sense of machine learning algorithms

Build your data science team

Machine Learning

Machine Learning

IBM Limited Edition

by Judith Hurwitz and Daniel Kirsch

for dummies®
A Wiley Brand

Machine Learning For Dummies®, IBM Limited Edition

Published by
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

ISBN: 978-1-119-45495-3 (pbk); ISBN: 978-1-119-45494-6 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Publisher's Acknowledgments

Some of the people who helped bring this book to market include the following:

Project Editor: Carrie A. Burchfield

Editorial Manager: Rev Mengle

Acquisitions Editor: Steve Hayes

Business Development Representative: Sue Blessing

IBM Contributors:
Jean-Francois Puget, Nancy Hensley, Brad Murphy, Troy Hernandez

Table of Contents

Introduction

Machine learning is having a dramatic impact on the way software is designed so that it can keep pace with business change. Machine learning is so dramatic because it helps you use data to drive business rules and logic. How is this different? With traditional software development models, programmers wrote logic based on the current state of the business and then added relevant data. However, business change has become the norm. It is virtually impossible to anticipate what changes will transform a market.

The value of machine learning is that it allows you to continually learn from data and predict the future. This powerful set of algorithms and models are being used across industries to improve processes and gain insights into patterns and anomalies within data.

But machine learning isn't a solitary endeavor; it's a team process that requires data scientists, data engineers, business analysts, and business leaders to collaborate. The power of machine learning requires a collaboration so the focus is on solving business problems.

About This Book

Machine Learning For Dummies, IBM Limited Edition, gives you insights into what machine learning is all about and how it can impact the way you can weaponize data to gain unimaginable insights. Your data is only as good as what you do with it and how you manage it. In this book, you discover types of machine learning techniques, models, and algorithms that can help achieve results for your company. This information helps both business and technical leaders learn how to apply machine learning to anticipate and predict the future.

Foolish Assumptions

The information in this book is useful to many people, but we have to admit that we did make a few assumptions about who we think you are:

- You're already familiar with how machine learning algorithms are being used within your organization to create new software. You need to be prepared to lead your team in the right direction so that the company gains maximum value from the use of these powerful algorithms and models.
- You're planning a long-term strategy to create software that can stand the test of time. Management wants to be able to leverage all the important data about customers, employees, prospects, and business trends. Your goal is to be prepared for the future.
- You understand the huge potential value of the data that exists throughout your organization.
- You understand the benefits of machine learning and its impact on the company, and you want to make sure that your team is ready to apply this power to remain competitive as new business models emerge.
- You're a business leader who wants to apply the most important emerging technologies to be as creative and innovative as possible.

Icons Used in This Book

The following icons are used to point out important information throughout the book:

TIP

Tips help identify information that needs special attention.

WARNING

These icons point out content that you should pay attention to. We highlight common pitfalls in taking advantage of machine learning models and algorithms.

REMEMBER

This icon highlights important information that you should remember.

IN THIS CHAPTER

- Defining machine learning and big data
- Trusting your data
- Looking at why the hybrid cloud is important
- Using machine learning and artificial intelligence
- Understanding the approaches to machine learning

Chapter 1
Understanding Machine Learning

Machine learning, artificial intelligence (AI), and cognitive computing are dominating conversations about how emerging advanced analytics can provide businesses with a competitive advantage to the business. There is no debate that existing business leaders are facing new and unanticipated competitors. These businesses are looking at new strategies that can prepare them for the future. While a business can try different strategies, they all come back to a fundamental truth — you have to follow the data. In this chapter, we delve into what the value of machine learning can be to your business strategy. How should you think about machine learning? What can you offer the business based on advanced analytics technique that can be a game-changer?

What Is Machine Learning?

Machine learning has become one of the most important topics within development organizations that are looking for innovative ways to leverage data assets to help the business gain a new level of understanding. Why add machine learning into the mix? With the appropriate machine learning models, organizations have the ability to continually predict changes in the business so that they are best able to predict what's next. As data is constantly added, the machine learning models ensure that the solution is constantly updated. The value is straightforward: If you use the most appropriate and constantly changing data sources in the context of machine learning, you have the opportunity to predict the future.

Machine learning is a form of AI that enables a system to learn from data rather than through explicit programming. However, machine learning is not a simple process.

REMEMBER

Machine learning uses a variety of algorithms that iteratively learn from data to improve, describe data, and predict outcomes. As the algorithms ingest training data, it is then possible to produce more precise models based on that data. A machine learning model is the output generated when you train your machine learning algorithm with data. After training, when you provide a model with an input, you will be given an output. For example, a predictive algorithm will create a predictive model. Then, when you provide the predictive model with data, you will receive a prediction based on the data that trained the model. Machine learning is now essential for creating analytics models.

You likely interact with machine learning applications without realizing. For example, when you visit an e-commerce site and start viewing products and reading reviews, you're likely presented with other, similar products that you may find interesting. These recommendations aren't hard coded by an army of developers. The suggestions are served to the site via a machine learning model. The model ingests your browsing history along with other shoppers' browsing and purchasing data in order to present other similar products that you may want to purchase.

Iterative learning from data

Machine learning enables models to train on data sets before being deployed. Some machine learning models are *online* and continuously adapt as new data is ingested. On the other hand, other models, called *offline machine learning models*, are derived from machine learning algorithms but, once deployed, do not change. This iterative process of online models leads to an improvement in the types of associations that are made between data elements. Due to their complexity and size, these patterns and associations could have easily been overlooked by human observation. After a model has been trained, these models can be used in real time to learn from data.


TIP

In addition, complex algorithms can be automatically adjusted based on rapid changes in variables, such as sensor data, time, weather data, and customer sentiment metrics. For example, inferences can be made from a machine learning model — if the weather changes quickly, a weather predicting model can predict a tornado, and a warning siren can be triggered. The improvements in accuracy are a result of the training process and automation that is part of machine learning. Online machine learning algorithms continuously refine the models by continuously processing new data in near real time and training the system to adapt to changing patterns and associations in the data.

What's old is new again

AI and machine learning algorithms aren't new. The field of AI dates back to the 1950s. Arthur Lee Samuels, an IBM researcher, developed one of the earliest machine learning programs — a self-learning program for playing checkers. In fact, he coined the term *machine learning.* His approach to machine learning was explained in a paper published in the *IBM Journal of Research and Development* in 1959.

Over the decades, AI techniques have been widely used as a method of improving the performance of underlying code. In the last few years with the focus on distributed computing models and cheaper compute and storage, there has been a surge of interest in AI and machine learning that has lead to a huge amount of money being invested in startup software companies. Today, we

are seeing major advancements and commercial solutions. Why has the market become real? There are six key enablers:

- Modern processors have become increasingly powerful and increasingly dense. The density to performance ratio has improved dramatically.
- The cost of storing and managing large amounts of data has been dramatically lowered. In addition, new storage innovations have led to faster performance and the ability to analyze vastly larger data sets.
- The ability to distribute compute processing across clusters of computers has dramatically improved the ability to analyze complex data in record time.
- There are more commercial data sets available to support analytics, including weather data, social media data, and medical data sets. Many of these are available as cloud services and well-defined Application Programming Interfaces (APIs).
- Machine learning algorithms have been made available through open-source communities with large user bases. Therefore, there are more resources, frameworks, and libraries that have made development easier.
- Visualization has gotten more consumable. You don't need to be a data scientist to interpret results, making use of machine learning broader within many industries.

Defining Big Data

Big data is any kind of data source that has at least one of four shared characteristics, called the four Vs:

- Extremely large *Volumes* of data
- The ability to move that data at a high *Velocity* of speed
- An ever-expanding *Variety* of data sources
- *Veracity* so that data sources truly represent truth

The accuracy of a machine learning model can increase substantially if it's trained on big data. Without enough data, you are

trying to make decisions on small subsets of your data that might lead to misinterpreting a trend or missing a pattern that is just starting to emerge. While big data can be very useful for training machine learning models, organizations can use machine learning with just a few thousand data points.

WARNING

Don't underestimate the task at hand. Data must be able to be verified based on both accuracy and context. An innovative business in a fast-changing market will want to deploy a model that can make inferences in milliseconds to quickly assess the best offer for an at-risk customer to keep her happy. It is necessary to identify the right amount and types of data that can be analyzed to impact business outcomes. Big data incorporates all data, including structured, unstructured, and semi-structured data from email, social media, text streams, images, and machine sensors.

WARNING

Traditional Business Intelligence (BI) products weren't really designed to handle the complexities of constantly changing data sources. BI tools are typically designed to work with highly structured, well-understood data, often stored in a relational data repository. These traditional BI tools typically only analyze snapshots of data rather than the entire data set. Analytics on big data requires technology designed to gather, store, manage, and manipulate vast amounts data at the right speed and at the right time to gain the right insights. With the evolution of computing technology and the emergence of hybrid cloud architectures, it's now possible to manage immense volumes of data that previously could have only been handled by supercomputers at great expense.

Big Data in Context with Machine Learning

Machine learning requires the right set of data that can be applied to a learning process. An organization does not have to have big data in order to use machine learning techniques; however, big data can help improve the accuracy of machine learning models. With big data, it is now possible to virtualize data so it can be stored in the most efficient and cost-effective manner whether on-premises or in the cloud. In addition, improvements in network speed and reliability have removed other physical limitations of

being able to manage massive amounts of data at the acceptable speed. Add to this the impact of changes in the price and sophistication of computer memory, and with all these technology transitions, it's now possible to imagine how companies can leverage data in ways that would've been inconceivable only five years ago.


REMEMBER

No technology transition happens in isolation; change happens when there is an unsolved business problem combined with the maturation of technology. There are countless examples of important technologies that have matured enough to support the renaissance of machine learning. These maturing big data technologies include data virtualization, parallel processing, distributed file systems, in-memory databases, containerization, and micro-services. This combination of technology advances can help organizations address significant business problems. Businesses have never lacked large amounts of data. Leaders have been frustrated for decades about their inability to use the richness of data sources to gain actionable insights from their data.


REMEMBER

Armed with big data technologies and machine learning models, organizations are able to anticipate the future and be better prepared for disruption.

The Need to Understand and Trust your Data

It is not enough to simply ingest vast amounts of data. Providing accurate machine learning models requires that the source data be accurate and meaningful. In addition, these data sources are meaningful when combined with each other so that the model is accurate and trusted. You have to understand the origin of your data sources and whether they make sense when they're combined.

In addition to trusting your data, it also important to perform data cleansing or tidying. Cleaning data means that you transform your data into a form that can be understood by a machine learning algorithm. For example, algorithms use numbers, but data is often in the form of words. You have to turn those words into numbers. In addition, you have to make sure those numbers are

sensibly derived and internally consistent. You need to decide how you handle missing data and other data irregularities.


REMEMBER

Data refinement provides the foundation for building analytical models that deliver results you can trust. The process of data refinement will help to ensure that your data is timely, clean, and well understood.

The Importance of the Hybrid Cloud

When approaching machine learning and big data, many organizations have discovered that a combination of public and private cloud services is the most pragmatic way to ensure scalability, security, and compliance. To deepen learning, a company may, for example, want to leverage Graphics Processing Units (GPUs) on the cloud rather than building their own GPU-based environment. This is a hybrid approach.


REMEMBER

A hybrid cloud is a combination of on-premises and public cloud services intended to work in unison. The hybrid environment provides businesses with the flexibility to select the most appropriate service for specific workloads based on critical factors such as cost, security, and performance.

Cloud computing allows businesses to test new endeavors without the large upfront costs of on-premises hardware. Rather than going through procurement and integration, teams can immediately begin working with machine learning techniques. As the organization matures, it may choose to bring some of the hardware on-premises because of security and control or the cloud computing costs that can quickly escalate.

Leveraging the Power of Machine Learning

The role of analytics in an organization's operational processes has changed significantly over the past 30 years. Companies are experiencing a progression in analytics maturity levels ranging from descriptive analytics to predictive analytics to machine learning and cognitive computing. Companies have been successful at

using analytics to understand both where they've been and how they can learn from the past to anticipate the future. They are able to describe how various actions and events will impact outcomes. While the knowledge from this analysis can be used to make predictions, typically these predictions are made through a lens of preconceived expectations.


REMEMBER

Data scientists and business analysts have been constrained to make predictions based on analytical models that are based on historical data. However, there are always unknown factors that can have a significant impact on future outcomes. Companies need a way to build predictive models that can react and change when there are changes to the business environment.

In this section, we give you two types of approaches to advanced analytics.

Descriptive analytics

Descriptive analytics helps the analysts understand current reality in the business. You need to understand the context for historical data in order to understand the current reality of where the business is today. This approach helps an organization answer questions such as which product styles are selling better this quarter as compared to last quarter, and which regions are exhibiting the highest/lowest growth.

Predictive analytics

Predictive analytics helps anticipate changes based on understanding the patterns and anomalies within that data. With this model, the analyst assimilates a number of related data sources in order to predict outcomes. Predictive analytics leverages sophisticated machine learning algorithms to gain ongoing insights.


REMEMBER

A predictive analytics tool requires that the model is constantly provided with new data that reflects business change. This approach improves the ability of the business to anticipate subtle changes in customer preferences, price erosion, market changes, and other factors that will impact the future of business outcomes.

With a predictive model, you look into the future. For example, you can answer the following types of questions:

- How can the web experience be transformed to entice a customer to buy frequently?
- How do you predict how a stock or a portfolio will perform based on international news and internal financial factors?
- Which combination of drugs will provide the best outcome for this cancer patient based on the specific characteristics of the tumor and genetic sequencing?

The Roles of Statistics and Data Mining with Machine Learning

The disciplines of statistics, data mining, and machine learning all have a role in understanding data, describing the characteristics of a data set and finding relationships and patterns in that data to build a model. There is a great deal of overlap in how the techniques and tools of these disciplines are applied to solving business problems.

REMEMBER

Many of the widely used data mining and machine learning algorithms are rooted in classical statistical analysis. Data scientists combine technology backgrounds with expertise in statistics, data mining, and machine learning to use all disciplines in collaboration. Regardless of the combination of capabilities and technology used to predict outcomes, having an understanding of the business problem, business goals, and subject matter expertise is essential. You can't expect to get good results by focusing on the statistics alone without considering the business side.

The following points highlight how these capabilities relate to each other:

- **Statistics** is the science of analyzing the data. Classical or conventional statistics is inferential in nature, meaning it's used to reach conclusions about the data (various parameters). Statistical modeling is focused primarily on making inferences and understanding the characteristics of the variables. Machine learning models leverage statistical algorithms and apply them to predict analytics. In a statistical model, a hypothesis is a testable way to confirm the validity of the specific algorithm.

- **Data mining,** which is based on the principles of statistics, is the process of exploring and analyzing large amounts of data to discover patterns in that data. Algorithms are used to find relationships and patterns in the data, and then this information about the patterns is used to make forecasts and predictions. Data mining is used to solve a range of business problems, such as fraud detection, market basket analysis, and customer churn analysis. Traditionally, organizations use data mining tools on large volumes of structured data, such as customer relationship management databases or aircraft parts inventories. The goal of data mining is to explain and understand the data. Data mining is not intended to make predictions or back up hypotheses.

REMEMBER

Some analytics vendors provide software solutions that enable data mining of a combination of structured and unstructured data. Generally, the goal of the data mining is to extract data from a larger data set for the purposes of classification or prediction. In data mining, data is clustered into groups. For example, a marketer might be interested in the characteristics of people who responded to a promotional offer versus those who didn't respond to the promotion. In this example, data mining would be used to extract the data according to the two different classes and analyze the characteristics of each class. A marketer might be interested in predicting those who will respond to a promotion. Data mining tools are intended to support the human decision-making process. Therefore, data mining is intended to show patterns that can be used by humans. In contrast, machine learning automates the process of identifying patterns that are used to make predictions.

Machine learning algorithms are covered in the next section, "Putting Machine Learning in Context," in greater detail due to the importance of this discipline to advanced analytics.

Putting Machine Learning in Context

To understand the role of machine learning, we need to give you some context. AI, machine learning, and deep learning are all terms that are frequently mentioned when discussing big data, analytics, and advanced technology. AI can be understood as the

broadest way of describing systems that can "think." For example, thermostats that learn your preference or applications that can identify people and what they are doing in photographs can be thought of as AI systems.

As illustrated in Figure 1-1, there are four main subsets of AI. In this book, we focus on machine learning. However, in order to understand machine learning, it is important to put it in perspective.

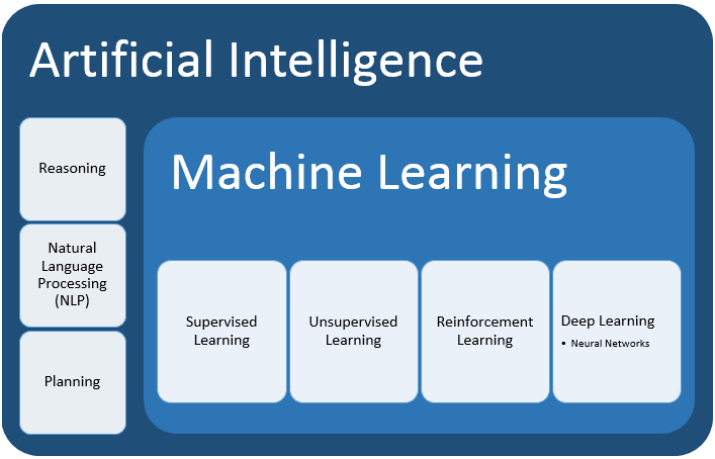

FIGURE 1-1: AI is the overall category that includes machine learning and natural language processing.

REMEMBER

When we explore machine learning, we focus on the ability to learn and adapt a model based on the data rather than explicit programming. In Chapter 6, we focus on applying machine learning to solving business problems.

Before we delve into the types of machine learning, it is important to understand the other subsets of AI:

- **Reasoning:** Machine reasoning allows a system to make inferences based on data. In essence, reasoning helps fill in the blanks when there is incomplete data. Machine reasoning helps make sense of connected data. For example, if a system has enough data and is asked "What is a safe internal temperature for eating a drumstick?" the system would be capable of telling you that the answer is 165 degrees. The

logic chain would be as follows: A drumstick that is eaten (as opposed to a part of a musical instrument) refers to a chicken leg, a chicken leg contains dark chicken meat, dark chicken meat needs to be cooked at 165 degrees, therefore the answer is 165 degrees. ***Note:*** In this example, the system was never explicitly trained on the safe internal temperature of chicken drumsticks. Instead the system used the knowledge it had to fill in the data gaps.

- **Natural Language Processing (NLP):** NLP is the ability to train computers to understand both written text and human speech. NLP techniques are needed to capture the meaning of unstructured text from documents or communication from the user. Therefore, NLP is the primary way that systems can interpret text and spoken language. NLP is also one of the fundamental technologies that allows non-technical people to interact with advanced technologies. For example, rather than needing to code, NLP can help users ask a system questions about complex data sets. Unlike structured database information that relies on schemas to add context and meaning to the data, unstructured information must be parsed and tagged to find the meaning of the text. Tools required for NLP include categorization, ontologies, tapping, catalogs, dictionaries, and language models.
- **Planning:** Automated planning is the ability for an intelligent system to act autonomously and flexibly to construct a sequence of actions to reach a final goal. Rather than a pre-programmed decision-making process that goes from A to B to C to reach a final output, automated planning is complex and requires a system to adapt based on the context surrounding the given challenge.

Approaches to Machine Learning

Machine learning techniques are required to improve the accuracy of predictive models. Depending on the nature of the business problem being addressed, there are different approaches based on the type and volume of the data. In this section, we discuss the categories of machine learning.

Supervised learning

Supervised learning typically begins with an established set of data and a certain understanding of how that data is classified. Supervised learning is intended to find patterns in data that can be applied to an analytics process. This data has labeled features that define the meaning of data. For example, there could be millions of images of animals and include an explanation of what each animal is and then you can create a machine learning application that distinguishes one animal from another. By labeling this data about types of animals, you may have hundreds of categories of different species. Because the attributes and the meaning of the data have been identified, it is well understood by the users that are training the modeled data so that it fits the details of the labels. When the label is continuous, it is a regression; when the data comes from a finite set of values, it known as *classification.* In essence, regression used for supervised learning helps you understand the correlation between variables. An example of supervised learning is weather forecasting. By using regression analysis, weather forecasting takes into account known historical weather patterns and the current conditions to provide a prediction on the weather.

TIP

The algorithms are trained using preprocessed examples, and at this point, the performance of the algorithms is evaluated with test data. Occasionally, patterns that are identified in a subset of the data can't be detected in the larger population of data. If the model is fit to only represent the patterns that exist in the training subset, you create a problem called *overfitting.* Overfitting means that your model is precisely tuned for your training data but may not be applicable for large sets of unknown data. To protect against overfitting, testing needs to be done against unforeseen or unknown labeled data. Using unforeseen data for the test set can help you evaluate the accuracy of the model in predicting outcomes and results. Supervised training models have broad applicability to a variety of business problems, including fraud detection, recommendation solutions, speech recognition, or risk analysis.

Unsupervised learning

Unsupervised learning is best suited when the problem requires a massive amount of data that is unlabeled. For example, social media applications, such as Twitter, Instagram, Snapchat, and

so on all have large amounts of unlabeled data. Understanding the meaning behind this data requires algorithms that can begin to understand the meaning based on being able to classify the data based on the patterns or clusters it finds. Therefore, the supervised learning conducts an iterative process of analyzing data without human intervention. Unsupervised learning is used with email spam-detecting technology. There are far too many variables in legitimate and spam emails for an analyst to flag unsolicited bulk email. Instead, machine learning classifiers based on clustering and association are applied in order to identify unwanted email.


REMEMBER

Unsupervised learning algorithms segment data into groups of examples (clusters) or groups of features. The unlabeled data creates the parameter values and classification of the data. In essence, this process adds labels to the data so that it becomes supervised. Unsupervised learning can determine the outcome when there is a massive amount of data. In this case, the developer doesn't know the context of the data being analyzed, so labeling isn't possible at this stage. Therefore, unsupervised learning can be used as the first step before passing the data to a supervised learning process.


REMEMBER

Unsupervised learning algorithms can help businesses understand large volumes of new, unlabeled data. Similarly to supervised learning (see the preceding section), these algorithms look for patterns in the data; however, the difference is that the data is not already understood. For example, in healthcare, collecting huge amounts of data about a specific disease can help practitioners gain insights into the patterns of symptoms and relate those to outcomes from patients. It would take too much time to label all the data sources associated with a disease such as diabetes. Therefore, an unsupervised learning approach can help determine outcomes more quickly than a supervised learning approach.

Reinforcement learning

Reinforcement learning is a behavioral learning model. The algorithm receives feedback from the analysis of the data so the user is guided to the best outcome. Reinforcement learning differs from other types of supervised learning because the system isn't trained with the sample data set. Rather, the system learns through trial and error. Therefore, a sequence of successful decisions will result in the process being "reinforced" because it best solves the problem at hand.

REMEMBER

One of the most common applications of reinforcement learning is in robotics or game playing. Take the example of the need to train a robot to navigate a set of stairs. The robot changes its approach to navigating the terrain based on the outcome of its actions. When the robot falls, the data is recalibrated so the steps are navigated differently until the robot is trained by trial and error to understand how to climb stairs. In other words, the robot learns based on a successful sequence of actions. The learning algorithm has to be able to discover an association between the goal of climbing stairs successfully without falling and the sequence of events that lead to the outcome.

Reinforcement learning is also the algorithm that is being used for self-driving cars. In many ways, training a self-driving car is incredibly complex because there are so many potential obstacles. If all the cars on the road were autonomous, trial and error would be easier to overcome. However, in the real world, human drivers can often be unpredictable. Even with this complex scenario, the algorithm can be optimized over time to find ways to adapt to the state where actions are rewarded. One of the easiest ways to think about reinforcement learning is the way an animal is trained to take actions based on rewards. If the dog gets a treat every time he sits on command, he will take this action each time.

Neural networks and deep learning

Deep learning is a specific method of machine learning that incorporates neural networks in successive layers in order to learn from data in an iterative manner. Deep learning is especially useful when you're trying to learn patterns from unstructured data.

Deep learning — complex neural networks — are designed to emulate how the human brain works so computers can be trained to deal with abstractions and problems that are poorly defined. The average five-year-old child can easily recognize the difference between his teacher's face and the face of the crossing guard. In contrast, the computer has to do a lot of work to figure out who is who. Neural networks and deep learning are often used in image recognition, speech, and computer vision applications.

A neural network consists of three or more layers: an input layer, one or many hidden layers, and an output layer. Data is ingested through the input layer. Then the data is modified in the hidden layer and the output layers based on the weights applied to

these nodes. The typical neural network may consist of thousands or even millions of simple processing nodes that are densely interconnected. The term deep learning is used when there are multiple hidden layers within a neural network. Using an iterative approach, a neural network continuously adjusts and makes inferences until a specific stopping point is reached. Neural networks are often used for image recognition and computer vision applications.

Deep learning is a machine learning technique that uses hierarchical neural networks to learn from a combination of unsupervised and supervised algorithms. Deep learning is often called a sub-discipline of machine learning. Typically, deep learning learns from unlabeled and unstructured data. While deep learning is very similar to a traditional neural network, it will have many more hidden layers. The more complex the problem, the more hidden layers there will be in the model.

REMEMBER

There are many areas where deep learning will have an impact on businesses. For example, voice recognition will have applications in everything from automobiles to customer management. In the Internet of Things (IoT) manufacturing applications, deep learning can be used to predict when a machine will malfunction. Deep learning algorithms can help law enforcement personnel keep track of the movements of a known suspect.

IN THIS CHAPTER

- Getting started with your strategy
- Looking at machine learning techniques in the business problem
- Tying machine learning to outcomes
- Understanding the business uses of machine learning

Chapter 2
Applying Machine Learning

With machine learning, you have the opportunity to use the data generated by your business to anticipate business change and plan for the future. While it is clear that machine learning is a sophisticated set of technologies, it is only valuable when you find ways to tie technology to outcomes. Your business is not static; therefore, as you learn more and more from your data, you can be prepared for business change.

Getting Started with a Strategy

Before you can define the strategy, you have to understand the problem that you're trying to solve. As businesses go through major strategy transitions, certain challenges present themselves. What is the status of existing business and existing customer engagement? What does the future hold for what customers will buy and expect from you in the future? The obvious answer is to ask customers if they are happy and what they will purchase in the future. While this is a sound starting point, it is not enough. Customers that are happy one minute become unhappy when something transformational comes along. If you do traditional

Business Intelligence (BI) analysis, you will have a good sense of where your business has been in the past but not where it is going in the future.

REMEMBER

Your business isn't static; much of the nuances and knowledge about your customers is hidden inside structured, unstructured, and semi-structured data. The value of machine learning techniques is to be able to uncover the patterns and anomalies in this massive amount of data. Selecting the right machine learning algorithms combined with the appropriate data sources helps you to determine what's next.

Using machine learning to remove biases from strategy

Typically, strategic planning and strategy exercises begin by gaining insights into customer satisfaction and future requirements. Where is the market headed? What are the competitive threats that could impact the company? But this is not enough. Even the best strategy consultants can't anticipate the sudden emergence of new discoveries or new trends.

WARNING

One of the traps that company leadership falls into is its assumptions and biases. Too often company management looks at the data presented and interprets the results through its own lens. Is the business sustainable in light of emerging competitors with unforeseen business models? While it is easy to be caught unaware of change, the seeds of change exist. However, those leading indicators are often buried inside huge amounts of unstructured or semi-structured data.

To gain benefit from a massive amount of unstructured data, it is important to truly understand these data sources. What is the source of the data? Who has manipulated that data? Are the data sources reliable? Early experiences in advanced analytics often resulted in disappointing results because analysts grabbed data sources without vetting them first. Before taking action, the data has to be verified as clean and accurate. After you are confident that you're using accurate data to address your business problem, machine learning approaches can provide significant insights. At the same time, you have to make sure that you have enough data to discover the patterns and anomalies within that data.

TIP

After the data quality is good, it is important to understand the context of the data being applied to the problem. For example, if a tree is losing its leaves in the middle of the summer, it is a sign that the tree is unhealthy. The same tree that has lost leaves in the middle of a cold winter day is a normal occurrence. Therefore, without understanding the context of data, you will likely misinterpret results. At the same time, there is considerable attention paid to correlation between data elements. What are the relationships between conditions? In the example of the health of trees, there is a direct correlation between the seasons and the color and amount of leaves on the trees. But you also have to be careful about correlations. You might find a correlation that makes no sense because the context is wrong. There may seem to be a correlation between leaves falling off the trees and the number of coats being purchased online. While both events are happening because the weather is colder, there is no relationship between trees and coats.

For the business to effectively use machine learning to support business strategy, you need these statistical methods to find patterns and anomalies in these data sets. With the best data available and in the right volume and the best level of cleanliness, it is possible to create a model by using the most appropriate machine learning algorithm based on the business problem being addressed. This model is only the beginning of the machine learning workflow.

By leveraging massive amounts of data, it is possible to model data, train the data, and then begin to learn from that data in order to improve the ability to make decisions. The value of learning from data means that the machine learning system is able to look at underlying patterns and anomalies that aren't necessarily obvious. Are there relationships between what customers buy with the time to repair? Are there impacts of weather on sales during a period of time? Are there indications in social media data that indicate subtle changes in customer perceptions or buying patterns? Being able to model massive amounts of data from different data sources can add insights that no single human could have understood by simply relying on data available in isolation.

WARNING

There has been much discussion about correlation of data as an analytic method. While data correlation is incredibly important, it can sometimes be misleading. There may seem to be a correlation between the consumption of orange juice in June and the rise in

traffic accidents in the same month, but there is no causal relationship. Therefore, while correlation might be useful in certain cases, it can also lead to inaccuracies. This is why context is even more important. If there were a useful context between orange juice and traffic accidents, then the correlation would be useful. Therefore, as you move to leverage machine learning as part of planning and strategy process, you need to make machine learning and advanced analytics indispensable tools.

More data makes planning more accurate

What difference could machine learning make in business strategy? Take the example of a business that executes a traditional data analysis of customer satisfaction. In analyzing the data, it becomes clear that some anomalies in the data exist. Because of the data set being used, the analyst throws out the data that doesn't conform, assuming that this data is not accurate. However, if more data did exist, it may become clear that those anomalies that were assumed to be errors are actually an indication of a change in customer buying patterns or customer satisfaction. As more data is added into a model, trained, and analyzed with the most appropriate machine learning algorithms, it becomes increasingly clear that there are changes that will directly impact the future of the business.

For example, data scientists seeing some subtle changes will begin to add new data sources that will strengthen or debunk a statistical analysis about business change or growth. Over time as more data is ingested into the model, the system learns and gains more insight and more sophistication in order to predict the future. Therefore, machine learning becomes an invaluable partner in strategic planning.

Understanding Machine Learning Techniques

In order to ensure that your data scientists are using the right machine learning techniques to achieve your business goals, it is important to understand how your organization can best apply these advanced techniques to manage your growth and keep focused on emerging opportunities.

Machine learning is a systematic approach to leveraging advanced algorithms and models to continually train data and test with additional data to begin to apply the most appropriate machine learning algorithms to a problem (we discuss this in more detail in Chapter 1). The advantage of machine learning is that it is possible to leverage algorithms and models to predict outcomes. The trick is to ensure that the data scientists doing the work are using the right algorithms, ingesting the most appropriate data (that is accurate and clean), and using the best performing models. If all these elements come together, it's possible to continuously train the model and learn from the outcomes by learning from the data. The automation of this process of modeling, training the model, and testing leads to accurate predictions to support business change.

Tying Machine Learning Methods to Outcomes

Machine learning techniques have the potential to reshape entire markets and business strategies. For example, machine learning techniques are being used to transform the automobile industry with self-driving cars. Machine learning algorithms and models are revolutionizing the way an x-ray image is analyzed. Machine learning can provide proactive ways of anticipated security vulnerabilities that can be repaired before damage is done. There are hundreds of different solutions that can be created that rely on machine learning techniques that can transform whole industries.

REMEMBER

Different approaches and algorithms exist for machine learning, depending on the problem being addressed. You need to understand the problem you're trying to solve. The model you design will represent an understanding of the data and your ability to predict outcomes based on that data.

Applying Machine Learning to Business Needs

Machine learning offers potential value to companies trying to leverage big data and helps them better understand the subtle changes in behavior, preferences, or customer satisfaction.

Business leaders are beginning to appreciate that many things happen within their organizations and with their industries that can't be understood through a query. It isn't the questions that you know; it's the hidden patterns and anomalies buried in the data that can help or hurt you. In this section, we provide some examples of how companies are beginning to use machine learning techniques to create business differentiation.

Understanding why customers are leaving

Have you ever heard, "It costs a lot less to keep an existing customer than to gain a new customer"? Customer churn is a constant problem in certain industries, such as telecommunications, retail, and financial services.

Understanding how to prevent customers from leaving is more important than ever. We are in an era where emerging companies are offering new innovative business models. For example, mobile phone service providers used to demand a two-year contract, which was extended each time the service changed. As the competitive landscape shifted, companies found that they had to get rid of the contracts. This change was beneficial to customers but resulted in a huge spike in customer churn. Without the protection of customer contracts, mobile companies are turning to new approaches to keep customers.


REMEMBER

In order to prevent customer churn, it is critical that you have enough data about the customer's history, his preferences, the services he has purchased in the past, and his complaints. In a highly stable market, this approach to analytics might have been a predictor of the future. But in volatile markets, this approach will not work. You have to be able to anticipate market changes and changes in customer buying patterns. Using machine learning models can help you predict changes that will impact revenue. In essence, the mobile provider needs to be able to look at patterns from data as well as anomalies. The mobile provider has the benefit of having access to huge volumes of data across many different customers. By using the right algorithm, the vendor can create a model that maps the types of offerings and promotions that will retain customers and add new ones. How much will it cost to retain and add new customers? Will new plans reduce revenue significantly? Will the spending justify the efforts? These

are the types of predictions that a machine learning technique can provide.

What is the difference between a traditional BI approach and a machine learning approach to customer churn? With traditional BI, the organization is able to understand what has happened in the past and can evaluate trends of customer loyalty. In contrast, the machine learning algorithm creates a model that brings in massive amounts of both internal and external data. After the data is trained and tested, analysts can begin to anticipate changes in customer preferences. The model may be able to anticipate how customers' buying patterns will change in the future.


REMEMBER

Machine learning uses statistical algorithms as the foundation to creating a model that can learn and predict. The most common models used for predictive models for churn analysis are classification statistical algorithms, such as logistic regression and neural networks.

Recognizing who has committed a crime

Police departments have a difficult task when tracking criminals. Increasingly, there are more and more cameras in neighborhoods that help identify unlawful activity. But who has committed the act? While a picture may be worth a thousand words, without someone to identify the bad actor, it isn't easy to solve crimes. One of the ways law enforcement is trying to leverage image data is through the use of machine learning.


REMEMBER

Specifically, deep learning algorithms and neural network–based algorithms are best suited to deal with facial recognition. In essence, neural networks are intended to emulate the human brain. By using a neural network algorithm, people can identify clusters and patterns in images. Image analytics can index and search video events by classifying objects into different categories, such as people, cars, roads, or streetlights. Further, facial recognition algorithms can be used to digitize sections of a photograph of a person in a way that eliminates extraneous data that isn't useful. The most important elements needed to identify a person include the eyes, nose, mouth, and things like scars. By collecting massive amounts of data of facial images, the algorithm can identify patterns in faces. Testing becomes a core technique that helps the model discriminate between two different

faces. Some of the emerging neural network techniques enable this type of training to be done with sparse data, which makes these systems more practical for a police force.

How would a police force take advantage of this type of neural network? The solution incorporates image data of known criminals. It includes data collected by surveillance cameras as well as images of suspicious individuals who might be involved in crimes locally. When a crime happens, such as a robbery at a local store, the images from the cameras can identify the faces of the individuals involved. These images can be matched against the quantity of data. Basically, the model is looking to match the pattern of a specific face against the collection of images to see if there is a match. If police can find the match, they will be able to quickly make an arrest without first taking the time to interview witnesses and spending hours reviewing store videos.

Preventing accidents from happening

Many industries rely on sophisticated preventive maintenance approaches to ensure that processes and systems are safe and operate as expected. Industries such as manufacturing, oil and gas, and utilities succeed or fail based on their ability to prevent accidents. While it is common to have a maintenance schedule, that is often not enough. For example, there may be environmental conditions that impact the operations of a machine or system. For example, there may be a failure of a heating or air conditioning system. There could be a dramatic shift in weather conditions that could impact machinery.


TIP

Machine learning algorithms can be applied to preventive maintenance in a number of ways. For example, a regression algorithm can be used as the foundation for a model that can predict time to failure of a machine. Various classification algorithms can be used to model the patterns associated with machine failures. Data generated by sensors provides a huge volume of semi-structured data that can model and compare patterns of performance so that an anomaly from normal performance can be detected.

IN THIS CHAPTER

- Transforming applications through machine learning
- Understanding your data
- Looking at the machine learning cycle

Chapter 3
Looking Inside Machine Learning

Machine learning is a powerful set of technologies that can help organizations transform their understanding of data. This technology approach is dramatically different from the ways companies have traditionally leveraged data. Rather than beginning with business logic and then applying data, machine learning techniques enable the data to create the logic. One of the greatest benefits of this approach is to remove business assumptions and biases that can cause leaders to adapt a strategy that might not be the best.

Machine learning requires a focus on managing the right data that is well prepared. Organizations also must be able to select the right algorithms that can provide well-designed models. The work does not end there. Machine learning requires a cycle of data management, modeling, training, and testing. In this chapter, we focus on the technology underpinning that supports machine learning solutions.

The Impact of Machine Learning on Applications

We made a bold statement that with machine learning you begin with the data and let that data lead you to logic. How does a business execute on the goal? As with everything in complex application development and deployment, it requires a planning process for understanding the business problem that needs to be solved and collecting the right data sources.

How does this approach to creating applications have an impact on the business? When building applications from logic, you assume that business processes will remain constant. However, the reality is that processes change. If you can begin by modeling data, it will lead you to changes in process and logic. Therefore, machine learning can make the creation of applications much more dynamic and effective.

The role of algorithms

No discussion about machine learning would be complete without a section devoted to algorithms.

REMEMBER

Algorithms are a set of instructions for a computer on how to interact with, manipulate, and transform data. An algorithm can be as simple as a technique to add a column of numbers or as complex as identifying someone's face in a picture.

To make an algorithm operational, it must be composed as a program that computers can understand. Machine learning algorithms are most often written in one of several languages: Java, Python, or R. Each of these languages include machine learning libraries that support a variety of machine learning algorithms. In addition, these languages have active user communities that regularly contribute code and discuss ideas, challenges, and approaches to business problems.

WARNING

Machine learning algorithms are different from other algorithms. With most algorithms, a programmer starts by inputting the algorithm. However, with machine learning the process is flipped. With machine learning, the data itself creates the model. The more data that is added to the algorithm, the more sophisticated the algorithm becomes. As the machine learning algorithm

is exposed to more and more data, it is able to create increasingly accurate algorithm.

Types of machine learning algorithms

Selecting the right algorithm is part science and part art. Two data scientists tasked with solving the same business challenge may choose different algorithms to approach the same problem. However, understanding different classes of machine learning algorithms helps data scientists identify the best types of algorithms. This section gives you a brief overview of the main types of machine learning algorithms.

Bayesian

Bayesian algorithms allow data scientists to encode prior beliefs about what models should look like, independent of what the data states. With so much focus on data defining the model, you might wonder why people would be interested in Bayesian algorithms. These algorithms are especially useful when you don't have massive amounts of data to confidently train a model.

REMEMBER

A Bayesian algorithm would make sense, for example, if you have prior knowledge to some part of the model and can therefore code that directly. Let's take the case of a medical imaging diagnosis system that looks for lung disorders. If a published journal study estimates the probability of different lung disorders based on lifestyle, those probabilities can be encoded into the model.

Clustering

Clustering is a fairly straightforward technique to understand — objects with similar parameters are grouped together (in a cluster). All objects in a cluster are more similar to each other than objects in other clusters. Clustering is a type of unsupervised learning because the data is not labeled. The algorithm interprets the parameters that make up each item and then groups them accordingly.

Decision tree

Decision tree algorithms use a branching structure to illustrate the results of a decision. Decision trees can be used to map the possible outcomes of a decision. Each node of a decision tree represents a possible outcome. Percentages are assigned to nodes based on the likelihood of the outcome occurring.

TIP

Decision trees are sometimes used for marketing campaigns. You may want to predict the outcome of sending customers and prospects a 20 percent coupon. You can break customers into four segments:

- Persuadables who will likely shop if they receive an outreach
- Sure things that will buy no matter what
- Lost causes that will never buy
- Fragile customers who may react negatively to an outreach attempt

If you send out a marketing campaign, you clearly want to avoid sending items to three of the groups because they will either not respond, shop anyway, or actually negatively respond. Targeting the *persuadables* will give you the best return on investment (ROI). A decision tree will help you map out these four customer groups and organize prospects and customers based on who will react best to the marketing campaign.

Dimensionality reduction

Dimensionality reduction helps systems remove data that's not useful for analysis. This group of algorithms is used to remove redundant data, outliers, and other non-useful data. Dimensionality reduction can be helpful when analyzing data from sensors and other Internet of Things (IoT) use cases. In IoT systems, there might be thousands of data points simply telling you that a sensor is turned on. Storing and analyzing that "on" data is not helpful and will occupy important storage space. In addition, by removing this redundant data, the performance of a machine learning system will improve. Finally, dimensionality reduction will also help analysts visualize the data.

Instance based

Instance-based algorithms are used when you want to categorize new data points based on similarities to training data. This set of algorithms are sometimes referred to as *lazy learners* because there is no training phase. Instead, instance-based algorithms simply match new data with training data and categorize the new data points based on similarity to the training data.

REMEMBER

Instance-based learning is not well-suited for data sets that have random variation, irrelevant data, or data with missing values. Instance-based algorithms can be very useful in pattern recognition. For example, instance learning is used in chemical and biological structure analysis and spatial analysis. Analysis in the biological, pharmaceutical, chemistry, and engineering fields often uses various instance-based algorithms.

Neural networks and deep learning

A neural network attempts to mimic the way a human brain approaches problems and uses layers of interconnected units to learn and infer relationships based on observed data. A neural network can have several connected layers. When there is more than one hidden layer in a neural network, it is sometimes called *deep learning.* Neural network models are able to adjust and learn as data changes. Neural networks are often used when data is unlabeled or unstructured. One of the key use cases for neural networks is computer vision. (For more details on neural networks, refer to Chapter 1).

Deep learning is being leveraged today in a variety of applications. Self-driving cars use deep learning to help the vehicle understand the environment around the car. As the cameras capture images of the surrounding environment, deep learning algorithms interpret the unstructured data to help the system make near real-time decisions. Likewise, deep learning is embedded in applications that radiologists use to help interpret medical images.

Figure 3-1 depicts the architecture of a neural network. Each layer of the neural network filters and transforms the data before passing it to the next layer.

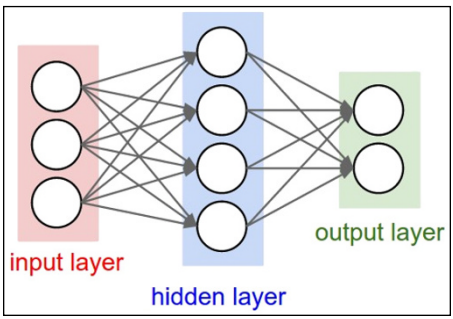

FIGURE 3-1: The architecture of a neural network.

Linear regression

Regression algorithms are commonly used for statistical analysis and are key algorithms for use in machine learning. Regression algorithms help analysts model relationships between data points.


TIP

Regression algorithms can quantify the strength of correlation between variables in a data set. In addition, regression analysis can be useful for predicting the future values of data based on historical values. However, it is important to remember regression analysis assumes that correlation relates to causation. Without understanding the context around data, regression analysis may lead you to inaccurate predictions.

Regularization to avoid overfitting

Regularization is a technique to modify models to avoid the problem of overfitting. You can apply regularization to any machine learning model. For example, you can regularize a decision tree model. Regularization simplifies overly complex models that are prone to be overfit. If a model is overfit, it will give inaccurate predictions when it is exposed to new data sets.


REMEMBER

Overfitting occurs when a model is created for a specific data set but will have poor predictive capabilities for a generalized data set.

Rule-based machine learning

Rule-based machine learning algorithms use relational rules to describe data. A rule-based system can be contrasted from machine learning systems that create a model that can be generally applied to all the incoming data. In the abstract, rule-based systems are very easy to understand: If X data is inputted, do Y. However, as systems become operationalized, a rule-based approach to machine learning can become very complex.

For example, a system may include 100 predefined rules. As the system encounters more and more data and is trained, it is likely that hundreds of exemptions to the rules might emerge. It is important to be careful when creating a rule-based approach that it doesn't become so complicated that it loses its transparency. Think about how complicated it would be to create a rule-based algorithm to apply the tax code.

Training machine learning systems

Through an iterative process of developing and refining a model, selecting the correct algorithm, training, and testing a system can begin. Training is a critical step in the machine learning process.

TIP

When you're training a machine learning system, you know the inputs (for example customer income, buying history, location, and so on), and you know your desired goal (predicting a customer's propensity to churn). However, the great unknown is the mathematical functions to transform that raw data into a customer churn prediction. As the learning algorithm is exposed to more and more customer data, the system will become more accurate at predicting the likelihood of customer churn.

Training a machine learning algorithm to create an accurate model can be broken down into three steps:

1. **Representation.**

   The algorithm creates a model to transform the inputted data into the desired results. As the learning algorithm is exposed to more data, it will begin to learn the relationship between the raw data and which data points are strong predictors for the desired outcome.

2. **Evaluation.**

   As the algorithm creates multiple models, either a human or the algorithm will need to evaluate and score the models based on which model produces the most accurate predictions. It is important to remember that after the model is operationalized, it will be exposed to unknown data. As a result, make sure the model is generalized and not overfit to your training data.

3. **Optimization.**

   After the algorithm creates and scores multiple models, select the best performing algorithm. As you expose the algorithm to more diverse sets of input data, select the most generalized model.

REMEMBER

The most important part of the training process is to have enough data so you're in a position to test your model. Often the first pass at training provides mixed results. This means that you either might need to refine your model or provide more data.

This process is not unlike learning any new discipline where you start with your assumptions based on incomplete knowledge. As you learn more, you can decide if you need more data from more sources. As you gain more insights from the data, your assumptions will probably change. One of the values of machine learning is that you don't start the learning process by deciding in advance what the answer to the problem will be.

When you have completed the training process, you're ready to test your understanding of the domain to see whether you have the right amount of knowledge or whether you are still required to collect more data and learn more. This is precisely what happens in an automated fashion when you design a machine learning system.

Data Preparation

Machine learning algorithms often get the majority of the attention when people discuss machine learning; however, success depends on good data.


REMEMBER

Understanding your data is critical to your success. If you create a model based on faulty data, your predictions will obviously be inaccurate. In addition, you need to think about what data should be included in your machine learning application.

Identify relevant data

Business decisions need to be made based on constantly changing data from a variety of sources. Your data sources may include both traditional systems of record data (such as customer, product, transactional, and financial data) and external data (for example, social media, news stories, weather data, image data, or geospatial data). In addition, many data structures are critical to analyzing information, including structured and unstructured data.

Structured data sources

Structured data is typically stored in traditional relational databases and refers to data that has a defined length and format. Most organizations have a large amount of structured data in their on-premises data centers. Examples of structured data include the following:

- **Sensor data:** Examples include radio frequency ID (RFID) tags, smart meters, medical devices, and Global Positioning System (GPS) data.
- **Weblog data:** When servers, applications, networks, and so on operate, they capture all kinds of data about their activity.
- **Point-of-sale data:** When the cashier swipes the bar code of any product that you purchase, all that data associated with the product is generated.
- **Financial data:** Many financial systems are now programmatic; they operate based on predefined rules that automate processes.
- **Weather data:** Sensors to collect weather data are being deployed across towns, cities, and regions to collect data on things like temperature, wind, barometric pressure, and precipitation. This data can help meteorologists create hyperlocal forecasts.
- **Click-stream data:** Data is generated every time you click a link on a website. This data can be analyzed to determine customer behavior and buying patterns.

Unstructured data sources

Although unstructured data has some implicit structure, it doesn't follow a specified format. Unstructured data is still widely underutilized by businesses and provides a great opportunity for monetization. Cloud, mobile, and social have contributed to a huge increase of unstructured data. Examples of unstructured data include the following:

- **Text internal to your company:** Think of all of the text within documents, logs, survey results, and emails. Enterprise information actually represents a large percent of the text information in the world today.
- **Social media data:** This data is generated from the social media platforms, such as YouTube, Facebook, Twitter, LinkedIn, and Flickr.
- **Mobile data:** This includes text messages, notes, calendar inputs, pictures, videos, and data entered into third-party mobile applications.
- **Satellite images:** This includes weather data or the data that the government captures in its satellite surveillance imagery.

- **Photographs and video:** This includes security, surveillance, and traffic data.
- **Radar or sonar data:** This includes vehicular, meteorological, and oceanographic data.

Governing data

Understanding and governing your data are prerequisites for an effective use of machine learning to solve real business problems. There will be a different level of governance when you are training data than when you use that data in a production environment. In the traditional world of data warehouses or relational database management, it is likely that your company has well-understood rules about how data needs to be handled and protected. For example, in the retail industry, it is critical that certain security provisions are placed around customers' personably identifiable information. You have to make sure that unauthorized individuals can't access private or restricted data. You also have control over who is allowed to both view data and change that data.

WARNING

As your organization begins to use machine learning–based solutions to predict outcomes, you must consider the data governance implications. When building machine learning applications, think about the following three data governance considerations:

- **Ensure that private data isn't compromised.** At the onset of a project, understand what types of data will be touched by a machine learning application. For example, will the applications process customer or employee data that is covered by industry rules or governmental regulations? If the results of a machine learning algorithm produce additional customer data, those results may need to be secured.
- **Data placement must be driven by governance rules.** Understand where data will be physically located and where the machine learning will take place. Some countries require that citizen data be kept within the country. Other rules and regulations may prohibit certain data from moving to a public cloud. These data locality requirements are important to consider if applications move data to different locations to perform machine learning tasks.
- **Maintain the privacy of sensitive data.** Understand who is allowed to see data being ingested into a machine learning application.

The Machine Learning Cycle

Creating a machine learning application or operationalizing a machine learning algorithm is an iterative process. You can't simply train a model once and leave it alone — data changes, preferences evolve, and competitors will emerge. Therefore, you need to keep your model fresh when it goes into production. While you will not have to do the same level of training that was needed when you built the model, you can't assume that it will be self-sufficient.

REMEMBER

The machine learning cycle is continuous, and choosing the correct machine learning algorithm is just one of the steps. The steps in the machine learning cycle are as follows:

- **Identify the data:** Identifying the relevant data sources is the first step in the cycle. In addition, as you develop your machine learning algorithm, think about expanding the target data to improve the system.
- **Prepare data:** Make sure your data is clean, secured, and governed. If you create a machine learning application based on inaccurate data, the application will fail.
- **Select the machine learning algorithm:** You may have several machine learning algorithms applicable to your data and business challenge.
- **Train:** You need to train the algorithm to create the model. Depending on the type of data and algorithm, the training process may be supervised, unsupervised, or reinforcement learning.
- **Evaluate:** Evaluate your models to find the best performing algorithm.
- **Deploy:** Machine learning algorithms create models that can be deployed to both cloud and on-premises applications.
- **Predict:** After deployment, start making predictions based on new, incoming data.
- **Assess predictions:** Assess the validity of your predictions. The information you gather from analyzing the validity of predictions is then fed back into the machine learning cycle to help improve accuracy.

TIP

After your model begins to make predictions, start the process over again by assessing the data you're evaluating. Is all of the data relevant? Are there new data sets that could help improve the accuracy of predictions? By continually improving models and evaluating new approaches you will be able to keep your machine learning-based applications relevant.

IN THIS CHAPTER

- Understanding how machine learning supports your goals
- Focusing on the business problem
- Requiring collaboration
- Selecting a pilot project
- Determining the best learning model

Chapter 4
Getting Started with Machine Learning

Using machine learning techniques to help your business achieve a level of sophistication of advanced analytics requires a plan and roadmap. You can't simply hire a group of data scientists and hope that they are able to produce results for the business.

In this chapter, we focus on the best approach to begin the process of enabling machine learning to support your business goals. Think about how you can get started so you can gain insights from the data generated by your company. If you approach the adoption of machine learning techniques in a systematic way, you'll be in a good position to anticipate changes in your market and changes in the way customers expect to do business with you.

Understanding How Machine Learning Can Help

Before you pick a target project, begin by helping business management understand what machine learning is all about. It isn't a cure-all. Machine learning is an approach that allows you to use

algorithms to create models based on data. Therefore, it is important to set expectations. In Chapter 5, we discuss the types of skills your team needs. While you'll certainly have experts, such as data scientists, it is important that business analysts and business strategists understand how machine learning can be applied to the business to solve some very complex problems. The abundance and variety of data can provide the business with a valuable weapon to help your business grow and change.

Focus on the Business Problem

When you begin to apply machine learning techniques to support your business strategy, you have to understand three fundamentals:

- **What is the business problem you are trying to solve?**

  Make sure that you have a good understanding of the nature of the problem you're trying to address. You may see changes in your revenue or perhaps in the types of products your customers are buying. Do you understand why your customers are buying? Do you understand how changes in the market are affecting your ability to satisfy customers? You have a lot of information about customers, product mix, and your market in general, but you need to conduct a deeper analysis so you're prepared for the future. Perhaps you're considering offering a new product to your traditional customer base. You need to understand how the new products will impact your revenue over the coming year.

- **Where are the hidden data resources that you can take advantage of to better understand your opportunities and threats?**

  Your organization probably has much more information about your business than you realize. There are customer support logs that can give you insights into the issues that confront your customers. Data will give you insights into the amount of time it takes to repair a problem. Some data is also stored in text that indicates what customers are looking for in the future. While this data exists, it may never have been used to make sense of your business. Ironically, you may already have all the data that you need to begin

assessing your future. All of this data has the potential to help you look beyond the obvious and anticipate the future.

- **How can you prepare to get your data in order?**

  The challenge is to make sure that you have your data ready to perform the type of analytical analysis so that you can learn from the data that you have. Are you using the right data sources that are the most up to date? Have you put the data into a form that is usable? Are you protecting the identity of your customers' private data? Are you selecting the best third-party data sources that will put your own data in context with your industry?

While machine learning has captured the attention of the technology and business market, you want to make sure that you select the approach and tools that best match the problem you need to solve. There will be different approaches depending on your industry and the type of data you're dealing with and the type of results you're looking to achieve.

REMEMBER

For many organizations, being able to understand the hidden patterns within their data offers a huge potential advantage. Most companies have important data that is stored in silos across different business units. Some of the important data may be found in social media sources. Data may also be found in unstructured data sources such as documents related to new research findings. Data is also found in semi-structured sources such as sensor and IoT-based systems.

Your first task is to determine what data sources and types of data are best suited to solving your problem. After you understand this, you will be in a good position to determine which algorithms will be used to create the most appropriate models. While there are hundreds of use cases to illustrate how to use machine learning algorithms to solve specific problems, we give you three examples in this section.

Bringing data silos together

You are in a competitive market with a lot of emerging companies determined to disrupt the market. Therefore, you have to figure out a way to understand the subtle changes to customer preferences and requirements. While you are diligent about conducting customer surveys and responding to customer complaints, this

information tends to be siloed across business units. Each organization that engages with customers understands a different view of the customer. What if you could get a broader perspective of all those touch points and interactions with clients? What could it tell you about your client's preferences that you didn't know? Many of these business units will deal with different product lines with different buyers.


TIP

With machine learning, you can bring together a variety of internal and external data sources and create a model that helps uncover patterns and anomalies that impact what you offer to customers and how you offer products and services.

Imagine a clothing chain that had the data and applied the most appropriate algorithm to gain an understanding of the changing customer expectations — what they are happy with and where they are dissatisfied. This data provides insights into the changing buying patterns. Is the customer base growing? Are existing customers beginning to leave? What are the demographics of new customers? Are new customers buying the same products in the same way as existing customers? Successful companies have the ability to truly leverage their data by breaking down silos of data across organizational boundaries. Disruptive businesses are agile and understand the value of their data in growing their customer base and revenue. Gaining early insights and indicators from data can turn a problem into an opportunity.

Avoiding trouble before it happens

Large cities often have limited resources to cope with complexity. Some issues can undermine their ability to react to problems that have the potential to overwhelm governments. Traffic problems can cause gridlock, lead to accidents, cause pollution, and make cities unlivable. When there are incidents such as a flood or a bridge collapse, city support services need to be prepared to act before populations are impacted dramatically. An unlivable city has a tough time attracting new companies to move in.


REMEMBER

Modeling traffic patterns by ingesting weather data, data about alternative traffic routes, and social media, for example, can help alert city management so they can alert citizens and reroute traffic away from danger zones. Being able to anticipate problems before they happen can improve conditions that make a city vibrant and avoid the loss of lives and property. How do you do it?

Machine learning can learn the patterns and conditions that can change the traffic patterns at a pace that the human mind can't.

Getting customer focused

Innovations often happen when a business begins to understand that there is a better way to create business opportunities. The only way to be prepared for change is to have the data and analytics that help you determine the next best action to get the results you hope for. Searching for the answer to a problem only works when you have an idea of what the answer might be. With machine learning, it is possible to find solutions when you can't anticipate what the answers or results will be.


TIP

Based on understanding changing expectations, it is possible to help customers know what they want before they can articulate the need. Understanding the subtle changes in customer buying patterns can help streamline the business to constantly change packaging and offers. Ironically, companies can often grab this data from a variety of public data sources. Matching this data to information about your own customers can lead to some potentially winning approaches.

Machine Learning Requires Collaboration

Much of the focus on machine learning is the viability of the models that are created by data scientists. But for those models to be able to predict business outcomes, there has to be a rigorous collaboration with the business. Line of Business (LoB) leaders are best able to understand the important data that is used to analyze the business. However, they often have a bias about what is most important to customers and the data that is most important. It is critical that the data science and data analytics teams discover new data sources that can improve the ability of businesses to uncover the hidden patterns and trends. The appropriate level of collaboration between business units, corporate leadership, and data scientists can create value that leads to true differentiation and meaningful change.

Executing a Pilot Project

After you have an idea of what types of problems you can begin to solve with machine learning, you may be ready to begin experimenting. Don't expect that you can do it all by hiring a few data scientists and having them experiment in isolation. You need to create collaboration between those business analysts, executives, strategists, data scientists, and analysts.

In this section, we give you the steps to help you successfully execute a pilot project.

Step 1: Define an opportunity for growth

Start with a problem that can be tied to a business outcome. But don't get ahead of yourself. Make sure that you select a small problem where you can readily identify the data that you already have and the data sources that you can obtain.


REMEMBER

What is the opportunity that your business can take advantage of to grow the business? Perhaps you have identified a traditionally popular product that is no longer selling well. For example, what is the optimal packaging of your products that will increase sales in the future? By understanding and modeling the data, you can understand how the data you have helps you predict the best set of options that meets customers' changing needs. Your selected pilot is also a marketing tool to demonstrate to the business that you can anticipate the future requirements of customers.

Step 2: Conducting a pilot project

Begin conducting your pilot with your concrete idea from Step 1. Make sure that you explain the purpose of the project and the type of data you are using. Good pilot projects are a subset of a larger problem that you're trying to solve for the business. If the pilot is successful, you'll already have put your goals in context. You will have gained insights to define your next steps.


REMEMBER

You will learn a tremendous amount from this pilot project. While you certainly will learn a lot from a successful pilot, you may learn even more from a failed pilot. What can you learn about customer buying patterns? Can you determine the ways that customers

purchase your products today and how that has begun to change? These new patterns that emerge from the data and being able to predict what this could mean for your business could help your transformation strategy.

Step 3: Evaluation

Let's assume that you learn some interesting things from conducting your pilot (see Step 2). You are seeing some interesting patterns emerge about customers and their future requirements. How do your results differ from the way you conduct business today? Your management has made certain assumptions about your customers and what they want. Did the results of your pilot indicate that you were making assumptions that were not in line with the pilot results?


TIP

You may be surprised that when you remove biases, your results differ significantly from what you thought you would find. This is one of the great benefits from applying machine learning to a business problem — you will be able to understand your business in a very different way. You may find that the pilot indicates that your customer expectations are much different than your assumptions and biases. As you add new data sources, the changes in customer requirements become more defined. These answers will then feed into your business planning and can enable your company to move more quickly to try new approaches that can positively impact revenue.

Step 4: Next actions

The benefit of the pilot project is that it begins to give you an understanding of how you can use machine learning and predictive analytics to better understand your business. If you planned your pilot as the first step of a succession of projects, you will be well on your way to learning from your data. At this stage, you want to expand to incorporate more data and bring more business leaders into the process. Select more data sources from many different areas of the business that help your analytics process. With machine learning, the more data that you can apply to a project the better your chance of gaining insights that you can apply to your business strategy.

Determining the Best Learning Model

One of the most complex tasks for applying machine learning to a business problem is selecting the most appropriate model. Selecting the most appropriate model is the best starting point in the journey to making machine learning an indispensable tool for predicting business outcomes. One of the most complex issues with selecting a model is to make sure that the model will perform well in the future when new data is introduced.

WARNING

The selected algorithm has to be generalized enough that it can be accurate with new data. If the algorithm is too tightly tied to an existing set of data, this type of overfitting will cause problems in the future. Therefore, when you select an algorithm, begin by making sure that the data set being used is a representative sample of your information. Your pilot will be much more successful if your data set is a representative sample of the aspect of the business that you are focused on. For example, you might begin selecting an algorithm by selecting a sample data set that is well known in your organization. As a next step, you can add a data set from a totally different source that could be relevant to your hypothesis. How does the algorithm you've selected predict outcomes from both the well-understood data set and the new data set?

Tools to determine algorithm selection

It is definitely not easy to select the algorithm that is best suited for your data and your challenge. Luckily, the market is beginning to recognize that in order to move forward, tools need to exist to help with algorithm selection. How do you choose the right model? It is a difficult problem.

WARNING

While overfitting may be one problem, a more serious problem is that models lose accuracy over time. Therefore, you have to continuously retrain the model as the data changes. Selecting the right algorithm can be best accomplished by automating the selection of an algorithm. Take the example of a classification algorithm. There are as many as 40 different classifier algorithms. These different algorithms can be combined depending on the approach the data scientist is using. Therefore, you can have hundreds of combinations to choose from. If your data scientists need to test for potentially valid algorithms, it could take a long time to pick

the best ones. Using an automation tool enables your scientists to more quickly determine the best combination of algorithms that will provide the highest score and the best fit for your data.

Automation tools are important not just because of the complexity of the algorithms but also because you have to make sure that the algorithms you select to build your models will not impact data latency and data consistency.

Approaching tool selection

A variety of open-source tools are intended to help data scientists select the right algorithm. These tools are often tied directly to the language (Python, R, Java, and so on) being used. Why should data scientists use tools for algorithm selection? Many different machine learning models may all be useful in solving problems. If a data scientist can experiment with different algorithms, he will be able to improve the ability of models to predict outcomes and create models that will scale.

IN THIS CHAPTER

- Identifying the skills for your team
- Finding resources to help learn more about machine learning

Chapter 5
Learning Machine Skills

If you have been reading the book up to this point, you have a good sense of the complexities and benefits of leveraging machine learning to solve business problems. You know that you need to arm your team with the right skills, including languages and tools. In this chapter, we provide you with an understanding of the technologies that help your organization successfully leverage the benefits of machine learning to support your organization's business goals.

Defining the Skills That You Need

Your team needs a variety of tools to successfully apply machine learning to solve some of your most complex business problems. At first glance, you might expect that you can employ a large team of data scientists. However, the reality is that it is difficult to find the data scientists that you need to move your company forward quickly. There are simply not enough skilled data scientists. And, because this talent is difficult to find, you will have to pay high salaries to those data scientists that you do discover. The answer is that you need to think differently about how you staff a department focused on innovating with machine learning. You can focus the data scientists on building models that can be used

by experienced data analysts. At the same time, you can begin selecting next-generation tools that can help a smart analyst be trained to achieve many of the machine learning techniques. There are many online training courses that can help educate your team.

In this section, we give you ten areas of skill that we recommend be your focus. Each one of these areas has many elements. Therefore, ensure that your team dives deeply into the areas that impact your organization's ability to support the business.

Understand what tools are available

What are the characteristics of leadership to support your company's goals with machine learning? There isn't *one* single tool or technique that you can use for machine learning; you can use a variety of tools. You should spend some time experimenting with different approaches that best match the problem you're trying to solve. There are best practices that can help in this process of tool selection.

Learn languages

A number of popular languages can be useful in moving forward with machine learning. The popularity of languages changes over time so it is often useful to learn more than one language. Languages such as Python, R, Java, and C++ are fundamental for moving forward with machine learning. Tools such as Linux, Hadoop, Spark, and cloud services are required to operate in an environment where you're investing in machine learning.

Explore algorithms

You need to understand the countless algorithms that will be useful in machine learning. A good data scientist will have deep understanding of probability and statistical methods because these are often used in creating effective machine learning models. Key algorithms that come in to play for machine learning include model creation to determine patterns and correlations and clusters from the data. For more details on machine learning algorithms, see Chapter 3.

Select appropriate models

It is important that you apply the right machine learning algorithms to solve the problem at hand. An increased number of packaging of machine learning algorithms exist through APIs, including Spark MLlib, H2O, and TensorFlow. One of the most important skills for developers is to understand which algorithm is the best fit for the problem. For example, a linear regression model fits the problem when you're trying to understand how two points are related. On the other hand, if you are dealing with understanding the content of images, you may want to explore TensorFlow. Many machine learning techniques match a variety of learning problems. The data scientist needs to be able to determine which algorithm and libraries make the most sense.

Understand the value of probability and statistics

A large number of learning algorithms are based on probability and statistics. Naive Bayes, Gaussian Mixture Models, and Hidden Markov Models are some of the methods that are important to understand.

Understand data management

Data scientists also have to understand the data that is being used. What is the source of the data? Is that source reliable and traceable? Do the sources you bring together to solve a problem make sense? In this case, the programmer or data scientist needs to work collaboratively with the business to vet the data sources.

Evaluate the cleanliness of your data sources

How good your data sources are will make the difference between success and failure of your machine learning projects. You need to understand the origins of your data and make sure that they're reputable. You also need to determine if you are selecting a combination of data sources that make sense when brought together.

A CHECKLIST: BUILDING YOUR TEAM

How should you plan your data science team? No matter what size your company is, there are some common characteristics that will make you successful. Remember, you are building a team to solve a business problem. The team might be one or two people that do everything, or in a larger company, you might have a person for each skillset. Chances are you won't find one person with all of these skills (what we refer to as a *unicorn*), but here is a checklist that helps you get started:

- Build a team with a mix of skills. You want to make sure that you are balancing technical team members with business members.
- Pick a lead data scientist who is well versed in both programming and architectural principles. In addition, the individual must have proven leadership skills in order to direct the team to execute on business goals.
- Bring in a business analyst who knows your industry as well as your company.
- Make sure a member of the team can tell a story from the data. This skill is different than interpreting the data or understanding the data; it's using the data to frame a discussion or provoke an action.
- Select representative business leaders who understand what they need to gain from the project.
- Add subject matter experts to the team who really understand the details of how processes work and the nature of the data. These experts should collaborate with a data engineer who understands how to capture and process the data.
- Find consultants when needed who can help train the team on new languages or new tools that support the project goals.
- Bring in specialists for specific technical areas where you don't have in-house talent.

If you work in a large company, you may have a variety of people inside your organization that are right for the tasks. In this case, you want to make sure that you have good leaders who can create a collaborative environment. If you're operating in a small company, select team members who really understand the fundamentals of your organization and your goals. Use technical team members you already have and supplement with industry specialists who will mentor your staff.

Understand how to piece work together

The bottom line is that with machine learning you are building an application based on a business outcome. Therefore, you need to understand how all the elements of the software and infrastructure support those outcomes. How do the elements fit together and communicate with each other to form a system? How do you create an environment that scales as more data and more logic are added? You need to understand that you are building a system that requires testing, management, documentation, and so on.

Understand the life cycle of data

One of the great benefits of machine learning is the fact that it requires a constant ingestion of new data in order to be able to make accurate predictions. Therefore, you need to understand that machine learning isn't a one-time task. Rather, machine learning is a continuum. The more accurate and plentiful your data is, the better your results.

Identify new use cases

Machine learning can be helpful across many different industries and many different functions. Exploring machine learning from pilots to production will help you gain insights into new uses. There may be many other areas within your business that can benefit from the type of predictive analytics that machine learning can provide.

Getting Educated

Because machine learning is an emerging market, there is a great demand for skilled personnel to help support organizations' efforts. It is becoming clear that companies can't wait to find all the skilled professionals they need. This means there is a great opportunity for IT professionals to up their game and become experts in data science and machine learning techniques. Luckily, there are a lot of resources out there that can help you learn. In this section, we provide a list of resources that are available to give you a great start.

Medium: Inside Machine Learning

This site gives you deep-dive articles on a wide range of machine learning topics. From weather predictions to robots, you can explore the top machine learning case studies and get insights from industry experts. Visit `medium.com/inside-machine-learning` for more information.

CognitiveClass.ai

Visit `https://cognitiveclass.ai` to build data science and cognitive computing skills for free today. Classes are based on an IBM community initiative. Courses include "Machine Learning with Apache SystemML."

Coursera online learning

Coursera is an online learning platform that offers courses and degrees in a variety of areas, including machine learning. It works with universities to offer more than 2,000 courses. Sign up today at `www.coursera.org/learn/machine-learning`.

Udacity courses on machine learning

Udacity is a for-profit educational organization that offers Massive Open Online Courses online (MOOCs). You can find it at `www.udacity.com/course/intro-to-machine-learning--ud120`.

Galvanize

Immersive data science curriculum includes a dive into machine learning and working on real problems in classification, regression, and clustering by utilizing structured and unstructured data sets. Students discover libraries like scikit-learn, NumPy, and SciPy, and use real-world case studies to root understanding of these libraries to real world applications. Learn more at `www.galvanize.com/san-francisco/data-science`.

edX courses

edX is an MOOC provider. It hosts online university-level courses. Some of the courses are even offered at no charge. Visit `www.edx.org/course/machine-learning-data-science-analytics-columbiax-ds102x-1` to find out more about the online "Machine Learning for Data Science and Analytics" course.

MITOpenCourseware

MIT has set up a site that includes all of its courses. It is offered at no cost to participants. You can learn more about machine learning at `http://bit.ly/1tP7pPU`.

Google Research Blog

Google researchers publish a variety of papers on topics related to machine learning and deep learning. You can learn more about deep learning here: `research.googleblog.com/2016/01/teach-yourself-deep-learning-with.html`.

Kaggle Wiki

The Kaggle Public Wiki is a resource for learning statistics, machine learning, and other data science concepts. It offers tutorials as well as a platform for data science competitions. Visit `www.kaggle.com/wiki/Home` today.

KDnuggets

KDnuggets is a popular site that provides a vast amount of information on analytics and a variety of information on data science. Check out the content at `www.kdnuggets.com/about/index.html`.

Data Science Central

Data Science Central is an online site for big data practitioners. It includes a community platform with technical forums for information exchange and technical support. Head to `www.datasciencecentral.com` for more information.

IBM-Recommended Resources

The IBM machine learning community can provide you with sources to add to your machine learning knowledge. For more information, visit these sites:

- **ibm.com/machinelearning:** See how companies are using machine learning to address challenges and pursue new opportunities.
- **ibm-ml-hub.com:** Get practical know-how to quickly and powerfully apply machine learning to start transforming your business.
- **ibm.com/datascience:** Research the capabilities that best meet your needs and learn how collaboration is enabling data science teams to innovate with quick time to value.
- **datascienceforall.com:** Whether you're a coder interested in the latest open-source capabilities or an analyst looking for drag-and-drop tools to collaborate on data science projects and move quickly, visit the data science community to find the latest best practices and resources to help you succeed.
- **datasciencemeetups.com:** Keep up to date on the latest meetups in your area, or join a virtual meetup featuring data science experts and sharing.

You can also use social media to stay connected to the data science world. Visit these two communities:

- **Facebook:** www.facebook.com/IBMDataScience
- **Twitter:** twitter.com/IBMDataScience or @IBMDataScience

IN THIS CHAPTER

- Seeing how machine learning works with patient health
- Using the Internet of Things to make predictions
- Responding to potential IT issues
- Preventing fraud

Chapter 6
Using Machine Learning to Provide Solutions to Business Problems

Machine learning is finding its way into every aspect of computing from social media to complex financial applications. Machine learning can be used to enhance the customer experience, better handle and predict results from complex data, and even transform the way different businesses can operate. Being able to correlate data to detect patterns and anomalies can help an organization predict outcomes and improve operations. There are numerous examples in almost every industry. In this chapter, we give you a few examples of how machine learning can be applied to solving complex business problems.

Applying Machine Learning to Patient Health

One of the biggest problems in treating patients is that drugs often affect individuals differently. Some medications may cause terrible side effects for one patient while being an effective

treatment for a different patient. A patient may have additional medical conditions that may cause a reaction to a treatment. Age and gender may also impact the effectiveness of a drug. Too often physicians have to resort to trial and error to find the right treatment.


TIP

One solution to selecting the most effective treatment is to build a machine learning model based on classification and regression algorithms. The classification model is needed to predict the impact of the drug based on known results from patient tests and conditions. The regression model is then used to predict the changes in the patient's condition when she takes a certain drug. Creating this model by using data helps provide researchers with an understanding of how a population of patients historically reacts to various drugs. As the model is built and trained, it will be able to determine the probability that a certain drug will be most effective for a patient.

If the model is online, it will continue to evolve as more patient data is added. A solution can be built to include a conversational interface using cognitive Application Programming Interfaces (APIs). In this way, a physician can interact with the model and ask a variety of questions to ensure that the right treatment is provided with fewer side effects.

Leveraging IoT to Create More Predictable Outcomes


TIP

Machine learning models are an ideal application for the Internet of Things (IoT). The first thing to understand about analytics on the IoT data is that it involves data sets generated by sensors. These sensors are now both cheap and sophisticated enough to support a seemingly endless variety of applications. The data generated by sensors contains a specific structure and is therefore ideal for applying machine learning techniques. While the data itself is not complex, there is often an enormous amount of data produced. By using this sensor data, along with known outages, machine learning algorithms can build models to predict future mechanical problems. The model would include data about the optimal indicators of a baseline of a well-run machine as well as data points the preceded a failure. As the model is trained, it will be able to determine anomalies that will predict the potential for failure.

HOW IT USED TO BE DONE

Machinery needs to be managed, maintained, and monitored regularly to ensure quality control and effective performance. Taking equipment offline for unneeded maintenance means downtime. Likewise, running equipment until it fails will result in unscheduled outages and potentially catastrophic results. Therefore, organizations want the ability to spot potential problems and fix them before they can cause downtime.

Reaching this level of preventative maintenance has not been easy. With traditional diagnosis methods, you can understand what has happened in the past month or even the past day. Manufacturing companies were early adopters of sensor technology in order to monitor how well equipment was operating. The typical way companies would monitor the output of sensors was to determine if they were matching the anticipated output. However, in order to prevent failure, it is important to anticipate and predict failures before they can cause damage.

While equipment has been outfitted with sensors for decades, there was no easy way to aggregate the data created by sensors. With advances in networking and the advent of inexpensive cloud compute and storage, it is now possible to aggregate this sensor data. With the advent of advanced analytics techniques, it is possible to capture the information generated by sensors and apply machine learning techniques to predict when a machine is likely to fail.

Proactively Responding to IT Issues

IT operations have always been complicated because of the array of different network devices, servers, applications, storage systems, endpoints, and so on. Each system has its unique ways of managing its components. As new versions of software are implemented, configuration updates may be necessary to keep the system running as expected. This is the normal way that systems need to interact in order to maintain a steady state. Often a single mistake in one area can lead to a massive outage, which can be difficult to determine the original cause of a problem — despite the fact that there is significant instrumentation within the data center.

A typical organization might deploy a dozen different monitoring tools to try to keep track of the health if its systems. These monitoring tools capture a huge amount of data about the systems they are monitoring. However, a key challenge is interpreting the large volume of system data and the fact that the data is contained in logs. To understand the data, the logs must be understood. In addition to this log and system data, valuable data can also be found in trouble tickets that include text describing a problem or data from application performance management systems.


TIP

Applying machine learning algorithms to this complex IT operations data allows organizations to proactively respond to potential IT issues. Traditionally, event correlation has been used to look for patterns in performance data. There are times, however, when correlation alone might be misleading. Therefore, to gain better accuracy, data scientists are beginning to cluster machine learning algorithms to identify event anomalies. The value of applying machine learning is that it can create a model based on a complex set of data created within the data center including alerts, logs, and instrumentation or sensors. The machine learning algorithm creates a model based on all the relevant data. The model can understand the dependencies between the various elements that comprise the environment. The model can also help identify patterns for ideal performance metrics and compare that to the current state of the environment. As more data is added, the model can be continuously updated.

Protecting Against Fraud


WARNING

Detecting fraud is a cat and mouse game. Bad actors are becoming increasingly sophisticated in perpetrating fraud. As more and more customers use online services, the potential for fraud has increased dramatically. In addition, payment processors want to make sure that customers have a friction-free transaction and do not want to block legitimate payments. Many companies are finding that the only approach that can help stop fraud is to use software, based on machine learning algorithms. A trained model can identify an anomaly before a fraud event is perpetrated. In essence, the model can identify an action that's associated with an intrusion or an unauthorized action and block the intruder before damage can occur.

REMEMBER

Combatting fraud has become a complex challenge and takes the combination of a variety of techniques. Linear techniques, neural networks, and deep learning are used together in order to spot fraudulent behavior (for more details, check out Chapters 1 and 3). Linear algorithms have been used for a long time to separate valid activities from fraudulent ones. However, a simple algorithm can't anticipate that the criminal will constantly change his techniques. It is difficult to stay one step ahead of the criminal activity.

Because linear algorithms on their own can't spot advanced fraudulent techniques, more advanced machine learning algorithms are used. For example, neural networks and deep learning are being used by payment processors. The deep learning models take into account thousands of data points in order to understand the context around a transaction.

TIP

An organization won't use neural networks or deep learning in isolation. Instead, it will use all three techniques together in order to perform *ensemble modeling,* which has its advantages. For example, while the linear algorithm might miss some fraudulent activity, it may be very good at catching the most common and straightforward schemes. The final model will take votes from each machine learning model and either approve or block a transaction. This sort of assessment is very similar to a medical patient getting multiple doctors' opinions. In the end, the goal is that the multiple opinions will yield more accurate results.

IN THIS CHAPTER

- Embedding machine learning in applications
- Making trained data as a service a prerequisite
- Investing in machine learning as a service
- Streamlining the machine learning pipeline
- Automating algorithm selection
- Requiring transparency and trust
- Making machine learning an end-to-end process

Chapter 7
Ten Predictions on the Future of Machine Learning

Machine learning is emerging as one of the most important developments in the software industry. While this advanced technology has been around for decades, it is now becoming commercially viable. We're moving into an era where machine learning techniques are essential tools to create value for businesses that want to understand the hidden value of their data. What does the future hold for machine learning? In this chapter, you explore our top ten predictions.

Machine Learning Will Be Embedded in Most Applications

Today, machine learning techniques are beginning to become popular in a variety of specialized environments. Businesses are looking to machine learning techniques to help them anticipate the future and create competitive differentiation.

In the next several years, you'll begin to see machine learning models embedded in nearly every application and on a variety of devices, including mobile devices and IoT hubs. In many cases, users will not know that they're interacting with machine learning models. Two examples where machine learning models are already embedded into everyday applications are retail websites and online advertisements. In both cases, machine learning models are often used to provide a more customized experience for users.

The impact of machine learning on a variety of industries will be dramatic and disruptive. Therefore, machine learning will significantly change how you do things. For example, hospitals can use machine learning models to anticipate the rate of admission based on conditions within their communities. Admissions can be related to weather conditions, the outbreak of a communicable illness, and other situations such as large events taking place in the city.

We are just beginning to see more and more machine learning models embedded into packaged solutions, such as customer management solutions and factory management systems. With the addition of machine learning models, these same systems become smarter and are able to provide predictive capability to enhance the value for the organization.

Trained Data as a Service Will Become a Prerequisite

One of the major obstacles in developing cognitive and machine learning models is training the data. Traditionally, data scientists have had to assume the jobs of gathering, labeling, and training the data. Another approach is to use publicly available data sets or crowdsourcing tools to collect and label data. While both of these approaches work, they are time consuming and complicated to execute.

TIP

To overcome these difficulties, a number of vendors offer pre-trained data models. For example, a company may provide hundreds of thousands of pre-labeled medical images to help customers create an application that can help screen medical images and spot potential health issues.

Continuous Retraining of Models

Currently, the majority of machine learning models are offline. These offline models are trained using trained data and then deployed. After an offline model is deployed, the underlying model doesn't change as it is exposed to more data. The problem with offline models is that they presume the incoming data will remain fairly consistent.

Over the next few years, you will see more machine learning models available for use. As these models are constantly updated with new data, the better the models will be at predictive analytics. However, preferences and trends change, and offline models can't adapt as the incoming data changes. For example, take the situation where a machine learning model makes predictions on the likelihood that a customer will churn. The model could have been very accurate when it was deployed, but as new, more flexible competitors emerge, and once customers have more options, their likelihood to churn will increase. Because the original model was trained on older data before new market entrants emerged, it will no longer give the organization accurate predictions. On the other hand, if the model is online and continuously adapting based on incoming data, the predictions on churn will be relevant even as preferences evolve and the market landscape changes.

Machine Learning as a Service Will Grow

As the models and algorithms that support machine learning mature, you'll see the growing popularity of Machine Learning as a Service (MLaaS). MLaaS describes a variety of machine learning capabilities that are delivered via the cloud. Vendors in the MLaaS market offer tools like image recognition, voice recognition, data visualization, and deep learning. A user typically uploads data to a vendor's cloud, and then the machine learning computation is processed on the cloud.

WARNING

Some of the challenges of moving large data sets to the cloud include networking costs, compliance and governance risks, and performance. However, by using a cloud service, organizations can use machine learning without the upfront time and costs associated with procuring hardware.

In addition, MLaaS abstracts much of the complexity involved with machine learning. For example, a team can use Natural Language Processing (NLP) — a tool used to interpret text or image recognition — to create a dialog between humans and machines. Both NLP and image recognition are well suited for the application of cloud services that has been designed to process specific compute intensive tasks. The performance differences are especially important when training and iterating many models. Large Graphic Processing Units (GPUs) are designed to speed the rendering of images so that they can significantly reduce the cycle time.

The Maturation of NLP

We expect that in the coming decade, NLP will mature enough to be the norm for users to communicate with systems via a written or spoken interface. NLP is the technology that allows machines to understand the structure and meaning of the spoken and written languages of humans. In addition, NLP technology allows machines to output information in spoken language understood by humans. Researchers have been working on NLP technology for decades, and machine learning is helping to accelerate the implementation of NLP systems. Currently, it is very difficult for machines to understand the context of words and sentences. By applying machine learning to NLP, systems are able to learn the context and meaning of words and sentences. Take for example the sentence "A bat flew toward the crowd." The sentence could be referring to a baseball bat that a hitter inadvertently let go of or a flying mammal that was heading toward a crowd of people. To understand the meaning of the sentence, a system would need to ingest the context around that phrase.

More Automation Will Streamline Machine Learning Pipelines

Automating the machine learning process will give less-technical employees access to machine learning capabilities. Additionally, by

adding automation, technical users will be able to focus on more challenging work rather than simply automating repetitive tasks. There are many tedious details involved with machine learning that are important but ripe for automation (for example, data cleaning). Data visualization is another area where automation is helping to streamline the machine learning process. Systems can be designed to select the most appropriate visualization for a given data set, making it easy to understand the relationship between data points.

Specialized Hardware Will Improve the Performance of Machine Learning

We are approaching an era where sophisticated hardware is now affordable. Therefore, many organizations can procure hardware that is powerful enough to quickly process machine learning algorithms. In addition, this powerful hardware removes the processing bottleneck of machine learning, thus allowing machine learning to be embedded in more applications.

Traditionally, CPUs have been used to support the deep learning training process with mixed results. These CPUs are problematic because of the cumbersome way that they process steps in a neural network. In contrast, GPUs have hundreds of simpler cores that allow thousands of concurrent hardware threads. Because of the importance of GPUs in deep learning applications, there has been considerable research going into the technology in order to offer more powerful chips. Cloud computing vendors also recognize the value of GPUs, and more of them are offering GPU environments on the cloud.

In addition to GPUs, researchers are using Field-Programmable Gate Arrays (FPGAs) to successfully run machine learning workloads. Sometimes FPGAs outperform GPUs when running neural network and deep learning operations.

Automate Algorithm Selection and Testing Algorithms

Data scientists typically need to understand how to use dozens of specific machine learning algorithms. In Chapter 3, we discuss

the main types of machine learning algorithms. A variety of algorithms are used for different types of data or different types of questions you're trying to answer.

Choosing the right algorithm to create a machine learning model is not always easy. A data scientist may try several different algorithms until he finds the one that creates the best model. This process takes time and requires a high degree of expertise. Automation is being applied to help speed the task of algorithm selection. By using automation, data scientists are able to quickly focus on just one or two algorithms rather than manually testing many more. In addition, this automation helps developers and analysts with less machine learning experience work with machine learning algorithms.

Transparency and Trust Become a Requirement

Understanding not just how but why a machine learning model recommends a specific outcome will be essential in order to trust the results. A deep learning model used for medical image scanning may flag an image for a potential cancerous growth. However, simply identifying the image isn't enough. The physician will need to understand why the machine model thought the growth was cancerous. What information was analyzed to lead the model to conclude the diagnosis? The physician must be convinced that the results are confirmed by the data.

Machine Learning as an End-to-End Process

Now that we are moving into an era of commercialization of machine learning, we will begin to see machine learning as an end-to-end process from a development and operations perspective. This means that the process includes identifying the right data to solve a complex problem, ensuring that the data is properly trained, modeled, and managed on an ongoing basis. This life cycle of machine learning is critical because there is so much at stake. Machine learning models can be a powerful tool for predicting the future.

The future with machine learning

New competitors are emerging and customer expectations have never been higher. Your business and reputation requires you to adapt to new trends and anticipate customer demands. Machine-learning technology is embedded in applications throughout enterprises in order to improve performance, increase customer satisfaction, reduce customer churn, and boost revenue. In this book, you discover what machine learning is, how to adopt machine learning in your company, and how machine learning can help your company.

Inside...

- What is machine learning?
- Explaining the business imperative
- The key machine learning algorithms
- Skills for your data science team
- How businesses use machine learning
- The future of machine learning

Learn more at:
IBM.com/machinelearning
IBM.com/datascience

Judith Hurwitz, President, Hurwitz & Associates, is a consultant and thought leader. **Daniel Kirsch,** principle analyst, Hurwitz & Associates, is a researcher and consultant in machine learning, cloud, and security.

Go to Dummies.com®
for videos, step-by-step photos, how-to articles, or to shop!

ISBN: 978-1-119-45495-3
Part #: IMM14209USEN-00

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.