MACHINE LEARNING

The Ultimate Guide to Understand Artificial Intelligence and Big Data Analytics. Learn the Building Block Algorithms and the Machine Learning's Application in the Modern Life.

MARK GRAPH

Machine Learning

The Ultimate Guide to Understand Artificial Intelligence and Big Data Analytics. Learn the Building Block Algorithms and the Machine Learning's Application in the Modern Life.

Mark Graph

By reading this document, the reader agrees that under no circumstances is the author responsible for any losses, direct or indirect, that are incurred as a result of the use of information contained within this document, including, but not limited to, errors, omissions, or inaccuracies.

Table Of Contents

References

Introduction

Machine learning is a field of artificial intelligence (AI) which gives us the ability to focus on the development of computer programs and understand the structure of various data models. As it is a major part of computer science, machine learning and artificial intelligence make use of specific algorithms, programs and instructions to solve specific problems. Furthermore, the main goal of understanding big data analytics is to search for patterns in data and allow the computers to learn automatically. Without the need for human intervention, machine learning methods have the capability to perform specific operations through the given algorithms and patterns.

There are different types, categories and implementation techniques of machine learning for which learning the basics of artificial intelligence are compulsory. The term “artificial” refers to something that is created by humans and “intelligence” is defined as the ability to think or understand a concept. Generally, AI is implemented into a system and is defined as the theory and development of computer systems that are able to perform specific tasks without human intervention, such as decision-making, speech recognition, translation between languages and visual perception. The aim of a machine learning algorithm is to go from data to insight in order to make relevant decisions, predictions and implementations.

“Machine Learning For Beginners: The Ultimate Guide to Understand Artificial Intelligence and Big Data Analytics. Learn the Building Block Algorithms and Machine Learning’s Application in the Modern Life” is the best source to read to understand machine learning, artificial intelligence and big data analytics for beginners. Along with detailed explanations and

examples for each topic, readers are given in-depth knowledge and understanding for each concept delivered in the book. Examples and applications are a great source for getting hands-on experience and skills related to machine learning and artificial intelligence implementation.

In traditional computing, algorithms used are comprised of predefined instructions and are used by computers to perform specific tasks or solve a certain problem. With the introduction of machine learning and artificial intelligence, computers can utilize data inputs and statistical analysis to automate the decision-making process. As it is a continuously developing field, developers should analyze the impact of machine learning processes and review the considerations before starting the implementation.

Machine learning was introduced by Arthur Samuel in 1959 and is based on statistical models that are used by computer systems to perform certain tasks without the need to use explicit instructions. As it is related to computational statistics, the study of machine learning is based on various methods, application domains and theory. However, the goal of researching artificial intelligence is to approach problems with different symbolic methods and other models. In artificial intelligence, a model is known as a mathematical expression that represents data in the context of a problem and provides an ongoing solution, such as prediction.

With the introduction of new computing technologies, machine learning has received several advancements. Although it started with pattern recognition and the approach that made computers able to learn without being explicitly programmed, the models in machine learning are now based upon new algorithms and produce better decisions and results.

Machine learning is categorized into different methods and sections which include regression, classification, clustering, dimensionality reduction,

neural networks and deep learning, ensemble methods, transfer learning, natural language processing, reinforcement learning and word embeddings. The major categories of machine learning are supervised and unsupervised approaches.

Chapter 1: What Is Machine Learning?

History

As previously mentioned, machine learning was discovered back in 1959 by Arthur Samuel who was an expert in the field of artificial intelligence (AI) and computer programming. To check whether a computer had real intelligence or not, Alan Turing created the "Turing Test" in 1950. As it all evoked from statistics and neural networks, the first program developed was the game of checkers in 1952 by Arthur Samuel which featured different winning strategies and self-playing techniques. Moreover, the "nearest neighbor" algorithm was written in 1967 which delivered basic pattern recognition abilities to the computer and was also used to map a certain route.

With further research and development, the work on machine learning was moved to a data driven approach from the knowledge driven approach in the 1990s. Machine learning and artificial intelligence have allowed researchers to solve specific problems by using symbolic methods and neural networks. However, the increasing focus on knowledge based and logical approach provided better implementation techniques for pattern recognition and information retrieval.

As a separate field, machine learning started to develop in the 1990s as it delivered new ways to solve problems of a different nature. Although the methods and models were inherited from probability theory and statistics, machine learning gained benefits from increasing the availability of digitized information.

Understanding And Development

Machine learning is a major field of computer science and there are several differences in the algorithms and problem-solving approaches as compared to traditional computation approaches. To understand machine learning algorithms in-depth, complete knowledge of statistical analysis and data inputs is needed, as machine learning algorithms facilitate computers in creating models from sample data to automated decision-making processes based on the given data inputs.

Take the example of students preparing for a specific exam, they try to learn the concepts, methods and theories step by step instead of going through the whole syllabus at once. As they have to get concepts and high-quality data into their brains, it has to deal with inputs, outputs and logic as well. With the passage of time, the processing power of their brains start to develop sharper and they get more of a grip on the subject. This is the same way machine learning models work since they allow computers to learn without further instructions.

In 1997, Tom Mitchell gave a definition that clearly defined machine learning: “A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.”

This means that if we want a program to predict something, such as traffic patterns at an intersection (T), we can run it through a machine learning algorithm featuring data about past traffic patterns (E) and if it has successfully learned, it will perform better and give more accurate

predictions for future traffic patterns (P). The selection for machine learning algorithms depends on the type of data and tasks that need to be automated.

Machine learning is a type of algorithm that enables software applications and computer systems to accurately predict outcomes for a specific scenario without the need to be programmed explicitly for that. These algorithms are built to use statistical analysis and receive input data to predict an output while performing relevant updates.

How Does Machine Learning Work?

The major aim of implementing machine learning is to go from data to insight for which different algorithms are used, also known as models. It is further defined as a category of an algorithm that enables software applications to deliver more accurate predictions or outcomes without being explicitly programmed. Basically, predictions were made by the computer through artificial intelligence which was then combined with machine learning algorithms.

Artificial intelligence was mainly based on statistical techniques through which researchers figured out new approaches for better prediction and data analysis. The statistical techniques include regression, classification,

clustering, probability theories and decision trees. In order to solve complex problems, more advanced method, such as deep learning, have proven to deliver accurate predictions.

Introduction To Data

Data that has been manipulated and interpreted is known as information, whereas the combination of experiences, inferred information and insights is called knowledge. In machine learning, any type of unprocessed value, text, fact or sound that is not interpreted or analyzed is known as data. It is the most important part of artificial intelligence, because models and algorithms cannot be implemented without data. Furthermore, data is essential for load and regression testing in machine learning models, which is an approach best suitable for big enterprises and businesses.

Data is classified into different categories in machine learning. These categories are:

- **Training data:** Training data is the part of data that is used to train a model. It is also considered as the input and output data that is utilized by the machine learning model.
- **Validation data:** Validation data is the part of data that is used to perform frequent evaluation of models and provides data sets as well. A data set is then used to improve hyperparameters that are set before the model starts learning, and also plays a vital role in the training of model.
- **Testing data:** In the testing data phase, the machine learning model is trained and tested to provide an unbiased evaluation. When inputs of testing data are entered into the model, it will predict some initial

values without considering the actual output. However, the testing data phase can also be used by the model to learn from experiences and make relevant predictions.

- **Data properties:** Generally, data is classified into various properties including volume, variety, velocity, value and veracity. Volume is the scale of data and it is increasing with each passing millisecond due to growing world population and exposure to technology. Variety includes different forms of data, such as text, images or videos. Rate of data streaming and generation is defined as velocity whereas the meaningfulness of data in terms of information is known as value. To check the correctness and certainty in the data that is being manipulated, veracity data property is used.

  Taking an example of a retail store, we can conduct a survey about sales, customer reviews and product quality. The answers received from customers can be referred to as the data and instead of reviewing each survey result manually, we can reduce time wastage and get instant results by manipulating data through calculations, graphs or software. Manipulated data is known as information and after the processing has been done, knowledge performs its role in differentiating between two data sets having the same information. It is more linked to human thought process and is not considered as technical content.

Get Started With Machine Learning

Machine learning makes use of artificial intelligence to allow machines to learn a specific task without human intervention or explicit programming.

The process begins with providing sample data to machine learning algorithms which is then processed to obtain relevant predictions and outcomes. Depending upon the process requirements, algorithms are selected and implemented on the type of data and system that needs to be automated.

In order to grasp the concepts of machine learning properly, it is advised that you are clear about linear algebra, statistics, python and multivariable calculus concepts as they will help in implementing algorithms and ML programming. You also need to focus on data science and artificial intelligence approaches as well.

Starting with linear algebra and multivariable calculus, mathematics plays a vital role in machine learning algorithms. Furthermore, data scientists who are focusing on research and development on machine learning have to create new algorithms which is made possible by working through making relevant calculations. Thus, to implement algorithms, expertise in linear algebra and multivariable calculus is mandatory.

Also, to handle collection, analysis and presentation of data, learning the key concepts of statistics is beneficial. As data plays a major role, we can perform multiple operations through probability distributions, statistical significance, regression and hypothesis testing. Developing complex machine learning algorithms is not possible if you are unable to learn linear algebra, statistics or multivariable calculus.

To get started with machine learning, considering the following approaches can be helpful:

- **Statistics:** Machine learning algorithms are designed to collect and manipulate data to perform specific tasks. To perform calculations, statistics concepts and methods are implemented over the algorithms,

such as probability distributions and regression. Furthermore, statistical approaches including priors, conditional probability and maximum likelihood are some of the key aspects of developing accurate machine learning algorithms.

- **Python:** For those who find it difficult to learn and implement multivariable calculus, linear algebra and statistics concepts, working on machine learning algorithms through Python programming language can be helpful. Python is the most suitable language that can be used for machine learning because it offers built in libraries that can be used for artificial intelligence as well. This means Python libraries, such as TensorFlow and Keras, are a great source to implement machine learning algorithms for beginners.

Machine Learning Terminologies

There are specific terminologies introduced in machine learning that can be used for working over different concepts and methodologies defined as follows:

- **Model:** A model is generated by applying machine learning algorithms and is also known as hypothesis. It is created by learning and manipulating the data provided in the algorithm.
- **Target:** The value predicted by a machine learning model is known as a target or label. Every input is labeled and classified for a specific category by implementing the target approach.

- **Feature:** Feature is an individual measurable property of specific data and is generally described by a feature vector. For example, in order to predict a car, the feature approach can give predictions for colors such as blue, silver, black or white.
- **Training:** In the training section, the algorithm is given a set of inputs and it is expected to deliver outputs or relevant predictions. After the training is complete, the model could then be mapped with new data in the categories.
- **Prediction:** Once the model is ready, data can be entered in a set of inputs so that it can be processed to deliver relevant predictions.

Handling Data

In machine learning, the data handling approach is split into three major categories that are known as training data, validation data and testing data. Training data is the process in which data is used to train the given machine learning model as it learns from both input and output to make relevant predictions. Validation data process is used to perform evaluations of a model and is used to improve other hyperparameters involved. Once the model is done with training and validation data processes, testing data delivers an unbiased evaluation.

Inputs of testing data are then fed into the machine learning model through which it predicts values without seeing the actual output. After the prediction is complete, the model is evaluated and compared with the actual output stored in the testing data phase. This process makes it simple to evaluate the prediction capabilities and performance of any machine learning model.

Types Of Machine Learning

Machine learning is categorized into three major categories called supervised, unsupervised and reinforcement. In general, supervised and unsupervised types of machine learning are commonly used because reinforcement learning is complex, difficult to learn and requires strong implementation skills.

- **Supervised learning:** Supervised learning includes learning from training data set along with a labeled data. The process is based on regression and classification models whereas the learning process continues until the required performance and functionality is achieved. This algorithm is comprised of a target variable which is supposed to be predicted from the provided set of independent variables or predictors. By utilizing a set of variables, a function is generated to map inputs to outputs.
- **Unsupervised learning:** Unsupervised learning involves the processing of unlabeled data. By using cluster analysis and factor models, we can learn more about data and utilize it for clustering population in different groups.

  Unsupervised learning algorithms, such as k-means and a priori algorithm, are widely used in machine learning models. The algorithm is given unlabeled data based on properties, unlike the supervised learning model, which makes it easier to make predictions on outcomes and results.

- **Reinforcement learning:** The reinforcement learning model involves learning of optimal actions and predictions through a trial and error approach. Depending upon the current state, the reinforcement learning model allows the agent to communicate with the environment and figure out the best possible outcome. For each correct or wrong answer, the agent is rewarded or penalized, and on the basis of positive rewards, the model trains itself to make future predictions.

  When compared with supervised and unsupervised learning models, the reinforcement learning model is complex and difficult to implement. At the start, the model makes several mistakes in predicting the outcomes or results, which means making appropriate adjustments in the machine learning model is necessary. Furthermore, the approach allows programmers to reinforce the algorithm to prefer good behaviors over negative ones because the reinforcement learning algorithm makes fewer mistakes after its learning is complete.

Chapter 2: Artificial Intelligence And Machine Learning

Introduction To Artificial Intelligence

Artificial intelligence is referred to as the science and engineering of intelligent machines. It is now considered as the most developed field of computer science as it allows machines to behave, think and make decisions like humans. Based on the principles of human based skills, such as reasoning, solving problems and learning, AI includes programmed algorithms which can work and deliver output using its own intelligence.

Machine learning and artificial intelligence methodologies are combined to develop strong models and algorithms that are able to make correct predictions. Without human intervention, machines and computer programs are able to learn and make accurate decisions in order to deliver positive outcomes.

The aim of AI is to enhance the performance of computer functions that are linked to human knowledge, such as learning, problem solving and reasoning. Currently, researchers are working on different objectives of AI including knowledge representation, natural language processing, learning, planning and realization.

Why Artificial Intelligence?

Based on the theories regarding computer science, psychology, biology, engineering and mathematics, artificial intelligence has made contributions in different fields including health care, medical, robotics and self-driving machines.

Artificial intelligence methods and techniques are used to create software and devices that can solve real world problems with a high level of accuracy. AI is intended to solve knowledge intensive tasks, replicate human intelligence and create an intelligent connection of action and perception. With the help of AI, we can create personal virtual assistants, build robots to work in remote environments and unleash new possibilities for the development of technologies.

Disadvantages Of AI

Although AI has provided long-term benefits and robust solutions, it has some disadvantages as well. No technology is perfectly developed and usually has initial problems that need to be addressed properly.

What's more, artificial intelligence is a costly solution because it requires lots of maintenance and routine upgrades. Moreover, the technology can never "think out of the box," as robots or machines will only perform the specific tasks that they are trained for. Humans are creative and can imagine new ideas, but with the rise of technology people are getting dependent on smart devices and are slowly losing their mental capabilities.

AI machines are great performers and provide accurate results, but they do not have any kind of emotional attachments or feelings for humans. This can sometimes be harmful for users as well.

Applications Of AI

Artificial intelligence is widely used in various development fields such as gaming, natural language processing, expert systems, vision systems, speech recognition, intelligent robots and handwriting recognition. From the past five to ten years, artificial intelligence has brought great advantages to the healthcare industry and has made it faster to diagnose diseases as compared to humans.

Moreover, AI is providing improved sales and profits to the e-commerce industry because it helps people discover new products. Last but not least, AI has provided outstanding improvements in education by bringing ease to both students and teachers through applications, such as AI chatbots and virtual tutors.

Differences Between Machine Learning And Artificial Intelligence

Machine learning is a subset of artificial intelligence and follows the symbolic rules, knowledge graphs, rules engines and models as defined in AI. When compared to AI, machine learning is dynamic and does not need human intervention to make the required changes as it is all about the acquisition of knowledge and skill. The aim of artificial intelligence is to increase the chance of success, whereas machine learning methods are more focused on achieving accuracy.

Artificial intelligence is also used to develop systems capable of mimicking humans in certain circumstances and will work to find the optimal solution. Models in AI are based on the decision-making approach whereas the machine learning systems adopt new things from data. Also, machine learning involves the development of self-learning algorithms and leads to knowledge, and devices designed by using AI techniques have the capability to handle any task and operation with accurate results.

Natural language processing, expert systems, neural networks, robotics and fuzzy logic are the major subfields of artificial intelligence. Machine learning has the capability to automate analytical model building as it utilizes methods from operations research, physics, statistics and neural networks to find hidden insights in data without being explicitly programmed.

Subfields Of Artificial Intelligence

Artificial intelligence is composed of multiple properties such as reasoning, learning, problem solving, perception and linguistic intelligence. There are several ways to simulate human intelligence because some methods or models are more intelligent than others and AI is completely based on *if-then* statements, such as knowledge graphs, expert systems and rules engines. On the other hand, machine learning models are said to learn by experience and improve their performance without being explicitly programmed.

Here are some of the major subfields of artificial intelligence:

- **Neural networks:** Neural network machine learning works similar to a human brain and is comprised of interconnected units. The information is then processed by responding to relaying information and external inputs between each unit. AI needs complete access to data for establishing connections to bring results from undefined data.
- **Computer vision:** Computer vision is focused on deep learning and pattern recognition approaches to recognize the information in a video or picture. With the help of AI models, machines are able to understand, analyze and capture images in real time. Interpretation and evaluation of visuals can also be performed through computer vision techniques.
- **Natural language processing:** Natural language processing is the ability of computers to understand, analyze, interpret and generate human language. As it is generally used in speech recognition models, humans can communicate with computers in their everyday language to perform specific tasks through natural language processing models. When an intelligence system performs any task as per the given

instructions, the process of natural language is used for both input and output tasks. For example, speech and written text.

The main components of natural understanding are mapping given input in natural language into useful representations and analyzing different aspects of language. To produce meaningful sentences and phrases in the form of natural language, the process used is known as natural language generation (NLG). This includes text planning, sentence planning and text realization.

- **Expert systems:** Expert systems are computer applications developed to solve complex problems at the level of human expertise and intelligence. Having high-performance, understandable, reliable, and highly responsive models, expert systems can easily instruct and assist humans in decision making. Moreover, the models are capable of advising, demonstrating and diagnosing different scenarios to interpret input and predict accurate results.

  Characteristics of expert systems provide researchers the ability to develop computer applications for solving complex problems in a particular domain. These systems deliver high performance and better responsiveness. Moreover, expert systems are reliable, understandable and capable of deriving a solution. Predicting results, explaining, demonstrating and suggesting alternative options to a problem are now possible with through expert system models. Major components of expert systems include knowledge base, interface engine and user interface.

- **Deep Learning:** Deep learning is a subfield of machine learning and is based on the algorithms of artificial neural networks. This technique allows computers to perform activities in the same way humans do

because it is inspired by the function and structure of the human brain. Deep learning technology is mainly being used in driverless cars and voice control applications in smartphones or computers. Moreover, the computer model is intended to learn from classification tasks in the form of sound, text or image.

These models have a high level of accuracy and sometimes also exceed human level performance because they are trained by using a large set of neural network architectures and labeled data. A deep learning algorithm will perform tasks repeatedly and learn from experience to improve the outcome. Neural networks have various deep layers that enable learning and allow machines to solve complex problems even with complex data set input.

Applications of deep learning include virtual assistants, translations, chatbots and service bots, vision for driverless delivery trucks, autonomous cars and drones, image colorization, facial recognition, medicine, personalized shopping, and entertainment.

Impact Of AI In Everyday Life

Machine learning is a subfield of artificial intelligence and is based on the acquisition of knowledge. Through AI models and algorithms, we can make intelligent computer systems that can understand human language and respond accordingly. This technology is being implemented in hospitals to keep patients safer and make recovery faster. Moreover, AI can now be found in advanced security systems and cameras through which finding people and things in real time has now been made possible.

Artificial intelligence has impacted our lives on a daily basis at every level. Smartphones and digital devices are updated with latest features such as built in assistants Alexa and Siri. Also, the finance industry and other companies heavily rely on artificial intelligence models for customer service, investments, fraud protection and chat bots. Moreover, spam filters in Gmail inbox are powered with AI to avoid fake emails.

Chapter 3: Big Data Analytics And Tools

Introduction To Big Data

Big Data is known as a combination of structured, semi-structured and unstructured data which is utilized to perform different machine learning operations. It is generally comprised of characters, quantities and symbols from which a computer is intended to carry out multiple tasks. The processed data can be stored and transmitted through electrical signals as well. Big data is also known as a collection of data which is large in size and has the ability to grow exponentially with time.

Handling large volumes of heterogeneous data is not possible by using conventional methods. So, to deal with large data sets, big data systems are designed to handle processing, storage and operations at each stage.

These systems are designed with the following exceptional characteristics:

- **Volume:** Big data systems are subject to bulk data volumes and data sets having a magnitude larger than the conventional data sets. This requires special storage and processing models at each stage. Sometimes, the work requirements cannot be handled by a single computer, so the work is combined by allocating, coordinating and pooling resources from groups of computers.
- **Variety:** Big data operations are complex because data is ingested from internal systems, such as server logs, applications, social media feeds, external API's and other sources. Regardless of where the data is coming from, it is supposed to be consolidated into a single system and the transformation to the raw data is performed in memory at the

time of processing. Different types and formats of media such as images, audio recordings, text files, video files and structured logs are also a part of big data and need to be processed and stored in the system.

- **Velocity:** To differentiate big data systems from conventional data handling systems, the information or processed data is flowed into the system from multiple sources and processed in real time. As data is frequently flowing, it should be processed in real time to gain insights and develop a strong understanding of the system as well. Moreover, data is constantly being added, processed, analyzed and massaged in order to keep up with the manipulation of new information and to surface valuable information when required. These characteristics are essential for a robust big data system because the flow of data is massive and continuous.

  Velocity is known as the speed of generation of data and how fast the generated data or information is processed to meet demands. Big data velocity also refers to as how fast dataflow is generated and processed from sources like application logs, business processes, mobile devices, networks, sensors and servers. When inconsistency occurs in data, the process used to handle and manage the operations effectively is known as variability.

Big Data Analytics

To examining large and varied data sets, big data analytics are used in machine learning and artificial intelligence. This includes the operations to uncover information like market trends, unknown correlations and hidden

patterns. Through big data analytics, we can analyze data sets and draw conclusions to bring long-term benefits to business organizations.

Big data analytics is a category of advanced analytics which involves complex operations such as statistical algorithms and predictive models. To analyze large volumes of data, software tools and applications are specially designed to perform tasks such as data mining, predictive analysis, data optimization and forecasting.

Performing in-depth analysis for big data is challenging for organizations having huge volumes of data. Different formats of data, including structured and unstructured data, are collected through various channels which are then combined, contrasted and analyzed to find new patterns and useful business information.

Advantages

Businesses and industries are greatly benefiting from big data solutions. Because they are focusing on finding actionable insights into their data, most big data projects are designed to provide increased efficiency, improved operations and better risk management techniques. Data analytics involves the application of mechanical or algorithmic processes to derive insights and it is used in several industries for managing data sets. Moreover, data analytics also helps in searching for meaningful correlations between the data of different categories.

With the development of technology, organizations have implemented big data analytics to break down data silos and utilize analytics tools for

handling operations. Furthermore, industries use big data analytics for data procurement, data quality and integration, and data governance procedures.

In the data procurement process, there are different websites from where data is supposed to be collected along with the primary and secondary data collected. After the data has been collected, developers have to avoid data repetition and remove redundant or unauthentic information to save storage space and other expenses as well. This activity is performed through data quality and integration approaches in big data systems.

Data Segmentation And Modeling

Generally, data is distributed in different parameters and this process is called data segmentation. Different parameters, such as name, age, gender, city or country, can be a part of customer segmentation and are usually based on the decision tree technique and regression. In data modeling, the provided data is interlinked according to each category and a logical model is created to manage the big data process. Before implementing data modeling, it is compulsory that the provided data is of high quality and can be analyzed without any hassle.

Before the data can be analyzed, it needs to be well-governed in order to maintain standards for data quality. When the data is reliable, organizations are able to establish a master data management program that has the capability to ensure constant data flowing in and out.

In-memory And Predictive Analytics

In-memory analytics is the process of deriving immediate insights of data from system memory and performing the required actions. This technology can be used to remove analytical processing latencies and data prep in order to create models and test new scenarios. Although it is not simple for organizations to follow the agile development approach, it can deliver better interactive analytics scenarios and consistent data flow.

Predictive analysis technology is based on machine learning techniques and statistical algorithms that are used to identify future outcomes based on historical data. As it is all about the best assessment over future events, organizations can make the best suitable business decisions through predictive analysis and stay safe from any kind of adverse scenarios.

Moreover, text mining technology is used to analyze text data from the web, books and website comment fields to uncover data insights. Text mining makes use of natural language processing and machine learning technologies to go through documents such as emails, bolds, social media feeds and competitive intelligence to analyze large amounts of information.

Tools And Software

Big data analytics have now become an essential part of business workflow and to get the best outcome, selecting the most suitable big data solution or tool can deliver long-term benefits. For the sake of versatility and optimal performance, developers go for open source tools because they are rich in features, usefulness and provide better reliability.

Here are some of the most popular big data analytics tools and software:

- **Microsoft Power BI:** Microsoft Power BI is widely being used in the business intelligence space and provides ease of use and better accessibility. The platform is integrated with Azure Data Lake Storage Gen2, which also supports Hadoop Distributed File System to perform advanced analytics operations. With built-in dashboards and a report generating system for different types of data monitoring and analysis systems, Microsoft Power BI also provides alerting capabilities along with cloud-based business analytics solutions.
- **Apache Hadoop:** Apache Hadoop is one of the greatest big data analytics tools. It is known for its capabilities in the high scale data processing. Being an open source tool, Apache Hadoop framework can operate on-prem or in the cloud because it has low hardware requirements.

  One major advantage of using Apache Hadoop is the availability of built-in Hadoop libraries through which third party modules can be enabled to work with Hadoop. The MapReduce big data processing model is highly configurable, and the tool also includes a resource scheduler for Hadoop resource management known as YARN.

- **Zoho Analytics:** Zoho Analytics is a self-operating tool which is designed to gain insights from data and perform relevant operations without the need of professional data scientists or IT staff. Featuring a basic drag and drop interface, Zoho Analytics also provides spreadsheet style interface and can be easily connected with a wide range of data repositories. Furthermore, most of the databases, key business applications and custom-built big data applications can be operated through this tool without any issues.

  This platform is the best choice for businesses that need to ensure accessible and convenient data analytics insights. With the availability of the intuitive drag and drop interface, users can also take advantage of the built-in use cases which are quite simple to understand and implement.

- **Aqua Data Studio:** Aqua Data Studio is a universal SQL integrated development environment and productivity tool which is designed to analyze, manage and develop data for multiple platforms. From a single interface, the tool can be used to access information from big data sources such as MongoDB and Cassandra. Users can create graphs, visualizations and charts to display and analyze data through Aqua Data Studio.

  Being a feature rich application, users can connect to multiple data sources such as cloud platforms, NoSQL and relational databases. The tool also provides support to compare database objects and schema DDL of different platforms along with SQL queries operations to generate results for analysis.

- **Microsoft HDInsight:** Microsoft Azure HDInsight offers big data cloud solutions in two major categories known as premium and

standard. Being a Spark and Hadoop cloud service, the platform serves as an enterprise scale cluster for companies to manage their big data workloads. Furthermore, it offers enterprise grade security, monitoring and reliable analytics along with industry leading SLA support. Microsoft Azure is considered as a high productivity platform for scientists and developers as it provides integration with several high-end big data applications.

- **RapidMiner:** This tool is based on agile development and the decision-based approach. RapidMiner big data analytics tool has the capability to operate through visual programming and is capable of analyzing, manipulating and modeling data. RapidMiner can be used to perform all tasks such as model deployment, data preparation and machine learning, making it easier for researchers and big data analysts. Building analytical workflows for data science projects is now possible with RapidMiner because it offers improved efficiency and better value. For conducting big data analytics operations, industries including travel, hospitability, retail, government and healthcare can use RapidMiner.

- **SAS Visual Analytics:** Well known in the analytics market, SAS Visual Analytics tool has been used by companies for decades. The tool offers big data analytics solutions for SAS platforms and is built for delivering advanced visualizations. Being completely extendable, the tool has the capability to share information and comments on multiple devices including smartphones, Microsoft Office applications and web browsers. Furthermore, big data preparation capabilities to design and prepare models make this software popular within all kinds of industries. This tool also gives access to dashboards from which

users can observe data analytics and perform relevant operations without any hassle.

- **Google Fusion Tables:** When it comes to the selection of effective, high performing and reliable data analytics tools, Google Fusion Tables is certainly the best option to be considered. Featuring high-end support for data analysis, data set visualization and mapping, Google Fusion Tables can be used to filter and summarize thousands of rows, visualize bigger table data online and combine tables with other data over the internet.

Data Sciences And Big Data Analytics

Data obtained from different sources are manipulated and organized through big data analytics tools. The data sets that are received by companies have properties of high volume, variety and velocity as they belong from different web pages, online networks, social media feeds, blogs and other sources. Most of the big data analytics operations are performed by using popular techniques such as data mining, statistics, natural language processing and machine learning. The data is extracted and prepared to provide analysis for businesses and large enterprises through effective big data analytics and data science models.

Big data analytics tools help in reducing cost of storage, increase business efficiency and provide faster decision-making solutions for companies and businesses. There has been an enormous growth in the big data analytics field because of the long-term benefits it has to offer. Industries such as banking,

healthcare, manufacturing, technology, consumer and energy are the main users of big data analytics tools.

In the past, technology platforms were designed to process either structured or unstructured data for which integration of data types into machine learning models was compulsory. With the development of latest big data analytics tools and technologies such as Hadoop can be used to deal with structured information corporate systems including SAP, NetSuite and Oracle. Regardless of the size of business or type of industry, implementing effective big data analytics strategies and technologies can greatly help in achieving a competitive gain.

Chapter 4: Neural Networks

What Are Neural Networks?

Neural networks are based on the architecture and functionality of the human brain. Having the ability to recognize patterns, neural networks are basically a set of algorithms that have the capability to interpret sensory data through perception and labeling of raw input. Such systems generally learn to perform specific tasks without being explicitly programmed or provided with a set of instructions. They also have the characteristics to group unlabeled data by overviewing similarities from given inputs and also classifying data when provided with a labeled data set.

What's more, neural networks also have the capability to extract features that are modeled in other algorithms for classification and clustering. In this way, deep neural networks become a part of larger machine learning applications based on regression, classification and reinforcement learning properties. Neural networks and deep learning provide ultimate solutions to most of the problems in natural language processing, speech recognition and image recognition.

Classification And Clustering

- **Classification:** In neural networks, classification plays a major role in process in labeled data sets. In order to make a neural network learn the correlation between data and labels, it is mandatory that humans must transfer their knowledge to the data set. This process is also known as supervised learning in machine learning methodology.

Classification is used to perform multiple tasks such as identifying people in images, detecting faces and recognition of facial expressions. Moreover, the approach also has the capability to identify objects in images like lane markers or stop signs.

Classification is implemented in neural networks to classify text as spam in emails, recognize sentiment in text and flag fraudulent activities. Most of the labels generated by humans, or any other outcomes which correlate to data, are used to train a neural network through the classification approach.

- **Clustering:** In neural networks, clustering is the process of detecting similarities without labels in data sets. As it is based on unsupervised learning, algorithms are trained through unlabeled data. Therefore, the clustering approach is used to train machine learning algorithms through unsupervised learning and produce highly accurate models.

  Major functionalities of clustering are searching and comparing images, documents or sounds of similar data items. In most cases, the anomaly detection process of clustering helps in detecting unusual behavior and anomalies in the data such as fraud or spam.

  Regression is the ability to create correlations between past and future events through classification in deep learning. This process is also known as static prediction because it helps in evaluating future events and scenarios.

The Difference Between Neural Networks And Conventional Computing

In order to understand artificial neural computing, it is compulsory to know the functionality and software process information. A conventional computer has a central processor that addresses an array of memory locations to store data and instructions, whereas neural networks are not sequential and do not include any complex central processors. Neural networks also do not execute programmed instructions and respond in parallel to the pattern of inputs.

Computational steps in a serial system are logical, sequential and deterministic through which the status of a provided variable can be reviewed for each operation. Neural networks have a high tolerance to error and should not be used for financial tasks. However, they work best for capturing associations within a set of patterns where the number of variables or data diversity is high. Relationships between variables and data sets in neural networks are quite strong and difficult to describe with conventional approaches.

Limitations

Neural networks are limited to perform specific operations and work best if they are being used to model with a high tolerance to error. Therefore, we need to overview each individual type of network before implementing it into a system. In reference to back propagation networks, users should remain aware of the potential issues and limitations of a neural network. In back propagation neural networks, users are required to input random numbers or data into the system and just wait for the machine to train on its own.

A trained network that has no equations or coefficients defining a relationship is known as a self-trained neural network. Back propagation networks are slower to train as compared to other networks that are similar

because they are equipped with CPU machines that must compute the function of each connection separately.

How Do Neural Networks Learn?

Neural networks cannot be directly programmed to perform specific tasks because they are based on machine learning models that learn from input. The major learning strategies for neural networks are listed below:

- **Supervised learning:** Supervised learning is a simple approach in which the machine learning models are provided with a labeled data set. The data is then modified and processed to achieve desired results from the input.
- **Unsupervised learning:** The unsupervised learning strategy is considered when there is no labeled data set given to learn from. The neural network observes the data set in order to increase the accuracy of the algorithm without any provided instructions.
- **Reinforced learning:** The reinforced learning algorithm is rewarded for delivering positive results and is also punished for giving a negative output. In this way, the algorithm is able to learn on its own with the passage of time and events.

Deep Neural Networks

Deep learning networks are classified as per the number of node layers through which data is passed in a multistep process. This activity is also known as pattern recognition and deep neural networks are determined by their depth.

In deep learning networks, each layer of the node is based on previous layer's output and completes the training process by recognizing features from the previous layer. In the past, neural networks, like perceptrons, were composed of a single input and output layer, so more than three layers is known as deep learning.

In deep learning, each layer of the node trains on a distinct set of features which are based on the previous layer's output. Since deep neural networks recombine and aggregate features from the previous layer, they become more complex and nodes start to learn in more effective ways.

To achieve better abstraction and increase complexity, the feature hierarchy approach is implemented in deep neural networks, which has the capability to handle high dimensional data sets with different parameters passing through nonlinear functions. Neural nets are also capable of discovering latent structures from unlabeled and unstructured data.

Deep learning has the capability to process a million images and cluster each of them as per their similarities. Applying the same approach to data types, deep learning can also cluster raw text like articles or emails. Thus, deep learning networks are typically designed to perform automatic feature extraction tasks without the need for human intervention. Since extraction is a task which involves teams of data scientists and the process might take years to accomplish, to make things more manageable, deep learning is considered the best way to completely process vulnerable and structured data.

Moreover, training on unlabeled data allows each node layer to automatically learn new features by repeatedly trying to manage the input from where it draws its samples. This allows machines to minimize the difference between guesses and probability distribution within the data. Deep learning networks that are trained on labeled data can also be

implemented on unstructured data which in return gives access to more input as compared to machine learning nets.

Types of Neural Networks

Neural networks are differentiated on how many layers they have between input and output. Being associated with deep learning, neural networks can also be described by the number of hidden nodes in the model and the ratio of inputs and outputs. Although there might be variations in classic network design, deep learning networks are finished with an output layer which can be either a logistic or SoftMax to assign a particular outcome.

Major types of artificial neural networks are explained as follows:

- **Feed forward neural networks:** Feed forward neural networks have characteristics of transferring information through various input nodes in one direction to approach the final output node. This technology is mainly used in computer vision and facial recognition models because it is able to process large amounts of sound as well. Furthermore, the network can also have hidden node layers which increase the overall functionality of feed forward neural networks.
- **Convolutional neural networks:** Convolutional neural networks (CNN) are based on the variation of multilayer perceptrons containing one or more convolutional layers. As this approach is based on nonlinear processing, the convolutional layers can be connected or pooled to create feature maps as well. Furthermore, the CNN model is generally used to create image recognition models and has also been implemented in advanced artificial intelligence applications like natural language processing, text digitization and facial recognition.

- **Recurrent neural networks**: Recurrent neural networks (RNN) are complex and difficult to implement as compared to other types of neural networks. Based on the approach of saving output of processing nodes and bringing results back into the model, recurrent neural networks are designed to learn and predict outcome for each layer. Moreover, RNNs start with the same front propagation as feed forward networks. If the predictions from recurrent neural networks are incorrect, the system automatically learns from this and continues on working towards the right prediction through a back-propagation approach.
- **Modular neural networks:** Modular neural networks are comprised of multiple neural networks that work separately from one another. In order to perform high end computational processes, modular neural networks do not communicate with each other's activities during the work process.
- **Deconvolutional neural networks:** Deconvolutional neural networks are designed to locate lost signals or features that are not considered important by convolutional neural networks. Applications of deconvolutional neural networks include analysis and the synthesizing of images.

Advantages Of Neural Networks

Information stored through neural networks is different from conventional databases because they can be managed throughout the network. With the ability to learn and model complex and nonlinear relationships, neural networks can be implemented to solve real life relationships and operations between input and output data. Moreover, there are no restrictions for input

variables, and they can be distributed through parallel processing as well. Neural networks have a high level of fault tolerance, and corruption of more than one cell of an artificial neural network will not affect output generation.

With the need of complete knowledge, neural networks have the power to produce output and there are no restrictions placed on the input variables. Moreover, machine learning allows artificial neural networks to learn from events and improve their decision-making abilities.

Disadvantages Of Neural Networks

Although neural networks have some amazing benefits to offer to the technological world, there are also some disadvantages. The requirement of processors that have parallel processing power makes neural networks absolutely hardware dependent which can increase the cost of setting up the system in an industry. Furthermore, there are no set rules to determine network structure because artificial neural network architecture is based only on a trial and error approach. Lack of explanation behind the provided solutions is yet another great drawback for neural networks because the inability to explain working procedure of neural networks results in trust issues.

Feed Forward Networks

A neural network is implemented to arrive at a point of least error in a minimal time frame. The network starts working from the point where the weights are initialized and end at the state of those parameters when they have the capability of producing accurate predictions and classifications.

This process involves several steps and scenarios as each step in a neural network involves error management, a guess and a small update in its weights. Because each step taken by a neural network is based on a guess, it is mandatory that the system is updated with incremental adjustments and error management facilities.

The collection of weights in both start and end states is known as a model. In order to adapt the data's structure, models usually change with time as the neural network is updated with new parameters. In order to perform operations with the model's data relationship with labels, the neural network must always be updated with the latest parameters. This process is done because of the ignorance characteristic within the neural networks. They do not know which biases and weights will translate best for the input to make appropriate guesses.

The maps, coefficients and weights that are a part of a neural network can be defined through the equation below:

input * weight = guess.

For computing errors, a neural network utilizes weighted input results and then takes a guess to compare it with the data through the equation mentioned as follows:

truth – guess = error.

To calculate adjusting weights that occur during calculating the difference between the network's guess and the truth, the equation is:

error * weight's contribution to error = adjustment.

These mathematical formulas are responsible for performing major machine learning and neural network operations such as application of an update to

the model, calculating loss and scoring input. A neural network is based on a corrective feedback loop which is responsible for rewarding weights or correct guesses and punishing weights that result in an error.

Gradient Descent

The gradient descent method is used to optimize a function that adjusts weights according to the error that has occurred in the neural network. Generally, the gradient is represented in a typical form on an x-y graph to show how two variables are related to each other. Further, the slope describes the relationship between a single weight and the network's error.

With the passage of time and events, a neural network learns slowly and adjusts many weights so that they can map signal onto meaning in a better way. Moreover, the relationship between a network error and weights is a derivative which is used to analyze how a change in weight causes a change in the error.

Afterward, the signal of weight passes through sums and activations over other layers and involves other types of transformations as well. This process is known as the chain rule of calculus and is considered to move back through the network's outputs and activations. Once the process is complete, the network arrives at the weight and its relationship to the error. The mathematical equation for chain rule in calculus is defined as follows:

$$\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx}.$$

Or

$$\frac{dError}{dweight} = \frac{dError}{dactivation} * \frac{dactivation}{dweight}$$

Through this equation, we can calculate how a change in weight directly affects a change in error. To perform calculations through the above-cited mathematical equation, we need to first calculate how a change in activation affects a change in error. This is also linked with how a change in weight affects a change in activation. In the deep learning approach, we are required to adjust a model's weight with respect to the error it is producing until the minimum reduction in error is achieved.

Neural networks and artificial intelligence both start and end the models in the same way. Because they are somehow inefficient in their approach to modeling, we cannot make appropriate assumptions related to their input and output functional dependencies. The number of updates is calculated through computation and through the orders of magnitude in the algorithms.

Loss Functions And Stochastic Gradient Descent (SGD)

Loss function in neural networks is used to evaluate how far the predicted output is from the actual output. The task is performed during training and can be done through a simple mathematical calculation based on certain functions and factors. Other aspects, such as regression and classification, might also be considered for calculating loss functions in neural networks.

Moreover, loss functions also have the capability to evaluate a network's performance during training. As a result, the loss in backpropagation is calculated as the gradient of change so that the weights can be updated in

order to minimize the loss. In the end, output values of each node are cached and calculated in the forward pass within the network.

In the network's computational graph, the partial derivative of error from loss function, with respect to other parameters, is calculated in a backward pass. To update the parameters for reducing loss, the derivative of error is used to examine whether loss is increasing or decreasing and by what rate. This procedure helps researchers to optimize the amount of weights and take further steps to make sure that the training is not taking extra steps in the loss space. This whole procedure is known as the stochastic gradient descent (SDC) in the field of neural networks.

Regression And Classification

Networks that are used for regression generally do not have any sort of training loss function. Root mean square and mean square errors are usually used to calculate the loss functions within a neural network. When training the networks for classification, entropy loss function is considered because it provides a loss that is based on the prediction of the correct class and models to judge whether the prediction is right or wrong.

This formula is implemented by applying two rules that are defined as follows:

1. If the training example is of a class, it has a value of 1.
2. If the training example is not of a class, it has a value of 0.

If the prediction is between 0 and 1, then the training example is 1, so the loss is calculated through the log (prediction) formula. In case the training example is actually 0, the loss is calculated through the log (1-prediction) formula.

Chapter 5: Probability And Statistics In Machine Learning

Machine Learning And Statistics

Machine learning is based on algorithms and models that use statistics and probability methods to provide insights for data. This also helps in building accurate and intelligent business applications which can bring long-term benefits to the industry. Probability deals with the likelihood of future events for prediction whereas statistics involves the analysis and evaluation of the frequency of past events. Both of these mathematical approaches are implemented in machine learning models to perform operations such as regression, probabilistic inference, density estimation, regularization, and model selection.

Before starting with machine learning, it is advised that you are aware about the basic probability and statistics concepts. This will help in understanding different machine learning models and approaches which are essential to achieve accurate predictions and outcomes.

Probability

Probability is a specific field of mathematics which is focused on quantifying uncertainty. An event is a set of outcomes which a probability is assigned in an experiment. E represents an event and P(E) is known for probability that will occur for E. In the event of a success or failure where the event might happen or not, the process is known as a trail. In machine learning, probability is recommended as a prerequisite subject to study before getting started because it makes it easier to understand and evaluate machine learning models, algorithms and processes.

Most of the algorithms are designed by using tools and techniques from probability such as probabilistic graphical and naïve Bayes models. Depending upon the approach for machine learning, developers can learn the concepts of probability and statistics to design specific algorithms. The event can be anything, including rolling a die or tossing a coin, because in these examples the event is random and the outcome of is known as a random variable.

Types And Categories

- **Theoretical probability:** Theoretical probability is described by the number of ways the specific events can occur divided by the total number of possible outcomes. Taking the example of a coin, there are

two outcomes, heads or tails, for which we can expect about half of tosses to be heads and the other half to be tails. So, the probability for the event of getting heads is 50/50. In this way, we can implement theoretical probability into machine learning algorithms and allow them to make better predictions in the future.

- **Empirical probability:** The empirical probability of an event is known by the number of times the event occurs divided by the total number of incidents. In the coin example, the tosses might have more or less than 50% heads and 50% tails and the success for this event is observed by S/N calculation.

- **Class membership:** When an example is assigned and given a label, it is considered as a classification predictive modeling problem and needs to be solved through probability methods. To frame the problem as a probabilistic class membership, the probability of an observation, such as input and output to each class, is predicted. Framing the problem as a prediction of class membership makes it convenient to model the problem. This also makes it easier for models to learn and capture any ambiguity in data to ensure a downstream process.

  Remember that probabilities can be transformed into a crisp class label as well. To perform this operation, we are required to choose the class that has the largest probability and can also be scaled through the probability calibration process. Most of the machine learning models are trained by using an iterative algorithm that is created through a probabilistic framework. The expectation-maximization algorithm, which is also known as the EM algorithm, is the best approach for unsupervised data clustering as it provides a framework for minimizing

the difference between a predicted and observed probability distribution.

Probability And Machine Learning

Machine learning algorithms are designed to harness methods from probability and are constructed to deliver accurate data insights. Algorithms such as naïve Bayes, probabilistic graphical models and Bayesian belief networks are a major part of machine learning on which various systems are designed and operated.

It is a wise approach to tune the hyperparameters of machine learning models or learning rate in a neural network. Bayesian optimization is an efficient approach to hyperparameter optimization. It involves a directed search of the space of possible combinations, and the other configurations based on those combinations are likely to deliver better results and performance in the future.

Probabilistic measures are used to evaluate model skill and evaluation of algorithms. There are several measures to conclude the performance of a model based on its predicted probabilities, evaluation measures are skills required to summarize performance of a specific model. Furthermore, receiver operating characteristic (ROC) curves can be constructed to examine different trade-offs and aggregate measures.

Statistics

Statistics is a major field of mathematics that is based on analysis, collection, presentation, interpretation and organization of data. In order to

adapt and implement machine learning concepts in a better way, it is important to learn key concepts such as conditional probabilities, posterior probabilities, maximum likelihood and data statistics. In fact, computer programming can also help in the process of learning the key concepts of statistics because it will combine your theory with practice to overcome many programming challenges.

For learning statistics in the field of data science, understanding core statistical concepts such as regression, descriptive statistics, hypothesis testing and distribution is mandatory. Afterward, learn the basic machine learning concepts and the best methods to use statistics in machine learning.

Statistics In Machine Learning

Learning statistics in data science in usually based on Bayesian probability theory. In this theory, the level of uncertainty before collecting data is known as “prior probability” and after the data has been collected, it is referred to as the “posterior probability.” Bayes’ theorem is a formula which describes how to update the probabilities of hypothesis in the presence of given evidence.

Bayes’ theorem states that the relationship between probability of hypothesis before getting evidence P (H) and after getting the evidence P (H | E) can be implemented through the following formula:

$$\Pr(\mathrm{H}|\mathrm{E}) = \frac{\Pr(\mathrm{E}|\mathrm{H})\,\Pr(\mathrm{H})}{\Pr(\mathrm{E})}$$

Statistical modeling techniques are usually applied to low dimensional data sets for which understanding how the data was collected and other statistical properties is required. Machine learning has no prior assumptions for the

underlying relationships between variables and the data must be given to the algorithm for processing and discovering patterns. The discovered patterns and results are then processed by models to make relevant predictions.

Descriptive Statistics

Descriptive statistics is a branch of statistics that calculates useful properties of a sample. This includes mean, mode and median calculations. Mean is referred to as the average value of the sample whereas Median is the middle value that separates the higher half from the lower half of the provided sample. Mode is the most frequent value occurring in the sample. Variance and standard deviation techniques are used to evaluate the spread of a data set from their mean.

Statistical analysis and machine learning work in collaboration to apply data science to the problem or to get insights from the data that is making a higher impact on business. Descriptive statistic covers calculation of statistical values on samples of data. This also helps in summarizing properties of sample data such as mean, median, mode and spread of data known as standard deviation. Descriptive statistics also covers graphical methods that are used to visualize samples of data.

Inferential Statistics

Inferential statistics is implemented in methods which support in quantifying properties of a particular domain or population from a set of obtained observations. Statistical inference tools can be used to quantify the likelihood of observing data samples that are given an assumption.

There are several examples of inferential statistical methods with specific ranges of hypotheses and constraints of data. For increasing the likelihood that the findings of the test are correct, we can impose constraints on the data as well.

Probability Distributions For Machine Learning

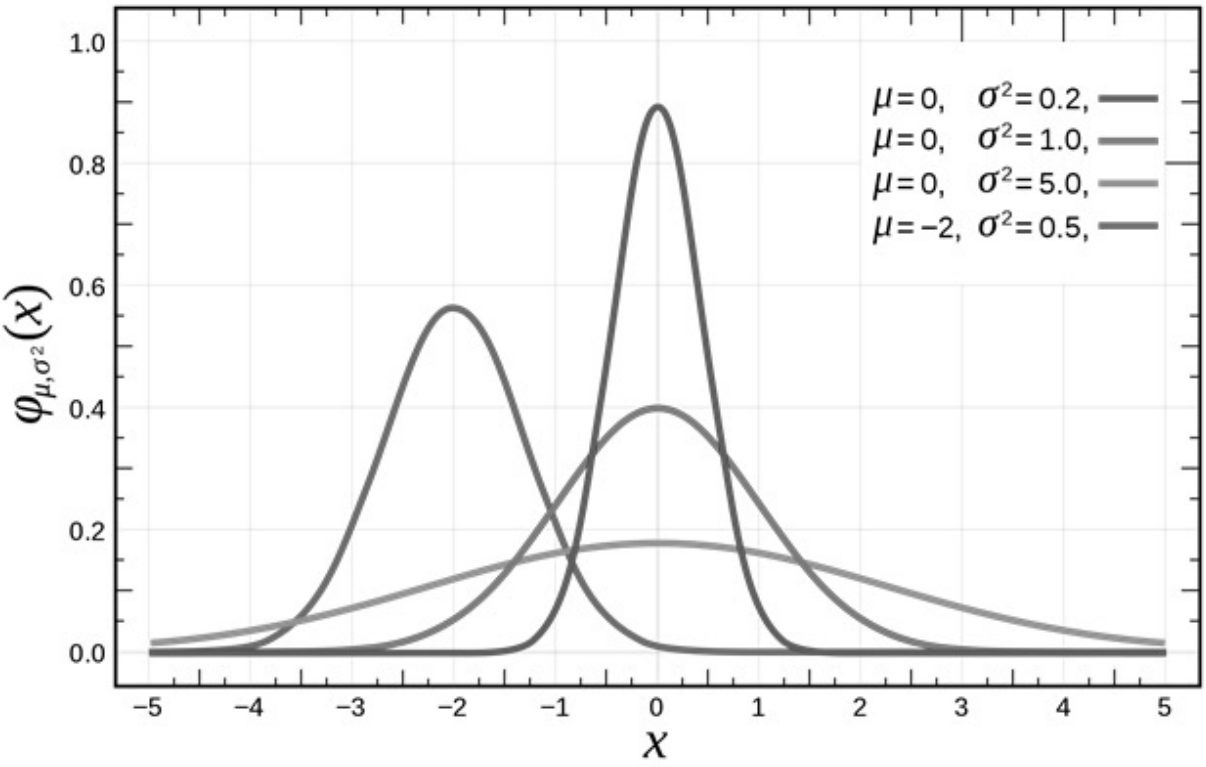
The relationship between events of a continuous random variable and its probabilities is known as the continuous probability distribution. In machine learning models, probability and statistics play a major role in determining expected outcomes and predicting results for a model. Continuous probability distributions are a major part of operations performed by the models over numerical input and output variables. Also, in-depth knowledge and understanding of continuous probability distribution is essential to program machine learning models for performing complex operations.

Probability of an event equal to or less than a defined value is known as the cumulative distribution function (CDF) whereas the inverse of this function is called the percentage point function. Probability density function returns the probability of a specific continuous outcome and cumulative distribution function delivers the probability of a value less than or equal to a given outcome.

Normal Distribution

Normal distribution is also known as Gaussian distribution. This distribution is based on the probability of real valued events that occur from different problem domains. A continuous random variable that has a normal distribution is known as normal or normally distributed. For example, the heights of people, test scores and marks of students.

Normal distribution is defined by using two main parameters that are mean (mu) and variance (sigma^2). Standard deviation is the average spread from the mean and is denoted by sigma as well. Visual representation of normal distribution can be learned from the graph below:



In this graph, we can notice that Gaussian distribution or normal distribution model can be identified with an iconic bell-shaped curve. This can also be represented in the form of mathematical formula described as follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\mu - x)^2}{2\sigma^2}}$$

Normal distribution is currently the most used probability distribution. It is used in various fields including marketing, psychology, finance and economics. Because of its fine symmetry properties, normal distribution provides the probability of a variable being a given distance below the mean of distribution equal to the probability of it being the same distance above the mean. Due to its reach of implementation and interpretation, normal distribution is most used in the finance industry.

Some of the machine learning algorithms are based on distance and gradient descent measures such as k-means and k-nearest neighbors. They are quite sensitive to the scale of the provided numeric values, and for an algorithm to provide a more exact solution, it is mandatory that rescaling of distribution is confirmed. Rescaling usually mutates range of values and can also affect variance. To perform rescaling, we can use statistical standardization (z-score normalization) and min-max transformation (normalization) techniques as well.

Python and R programming languages provide built in functions for distribution transformation and we can also achieve standardization by using the scale function (try help scale) in R programming language. To perform the same operation in Python programming language, we can consider the scikit-learn module preprocessing approach as well.

In machine learning, independent variables are also known as features. These are the input for a process which is being analyzed whereas the dependent variables are output of the process.

Chapter 6: Types Of Machine Learning

Machine learning is the study of teaching an algorithm the different techniques of how to progressively improve on a given task. There are several techniques for building machine learning and artificial intelligence applications for which it is mandatory that the developers are aware about the types, categories and functionality of algorithms and models. As we have learned, machine learning is classified into two major categories: supervised learning and unsupervised learning. It is interesting to learn and implement new techniques and strategies into machine learning models to avail better data insights and prediction results.

Supervised Learning

Generally, machine learning is based on the supervised learning models and algorithms. Because they are easy to design, learn and implement, achieving data insights and accurate predictions have become a lot easier with supervised learning models. The data that machine learning takes as input is also considered as training data. This includes both input and labels. Take the example of subtracting two numbers, a = 10, b = 6. So, the inputs are 10 and 6 and the target is 4.

At first, the model is trained with providing lots of data and inputs so that it is able to predict the output with any kind of new data. This training process is known as supervised learning, which is highly accurate and yields fast results as well. Once the model gets trained, it can start making a decision or prediction when new data is provided without running into any problems.

So, how does it work? Well, a supervised learning algorithm is comprised of a target or dependent variable which is supposed to be predicted from a given set of independent variables or predictors. By utilizing these sets of variables, a function can be generated that maps inputs to desired outputs. This training process continues until the model achieves the best level of accuracy on training data. Major examples of supervised learning are the decision tree, regression, k-nearest neighbor (KNN), logistic regression and random forest.

Regression

- **Linear regression:** This approach is best suitable for estimating real values based on continuous variables. A relationship is established between both dependent and independent variables by fitting a best line, which is also known as a regression line. Linear equation to represent regression is defined as follows:

  $$Y = a*X + b.$$

  In linear regression, relationships are modeled through linear predictor functions. Parameters of linear predictor functions are estimated from data which are then used to create linear models. In machine learning, linear regression is considered for many practical uses and most of the applications are designed for prediction and error reduction. Moreover, the analysis can also be applied to quantify the strength of the relationship between explanatory variables and response.

  There are three major uses of linear regression. The first is determining the strength of predictors. Regression is the best approach to be implemented in machine learning to identify the effect that an independent variable might have on a dependent variable. The second

major use is trend forecasting as the model is also implemented to estimate or forecast the impact of changes. Finally, it's used for forecasting an effect. That is, linear regression is used to predict trends and future values for getting point estimates.

In the case that linear regression occurs with multiple independent variables, forward selection, step wise and backward elimination approaches can be implemented.

- **Logistic regression:** This approach is used for finding the probability of event success and event failure. In logistic regression, the dependent variable must be in binary 0/1, true/false or yes/no format. The equation to represent logistic regression is defined as follows:

  $$\text{Odds} = p/(1-P)$$

  $$\text{Ln(odds)} = \ln(p/(1-p))$$

  $$\text{Logit}(p) = \ln(p/(1-p))$$

  Because we are working with a binomial distribution, we have to select a link function that is best suited for the specific distribution. Logical regression does not require a linear relationship between the dependent and independent variables.

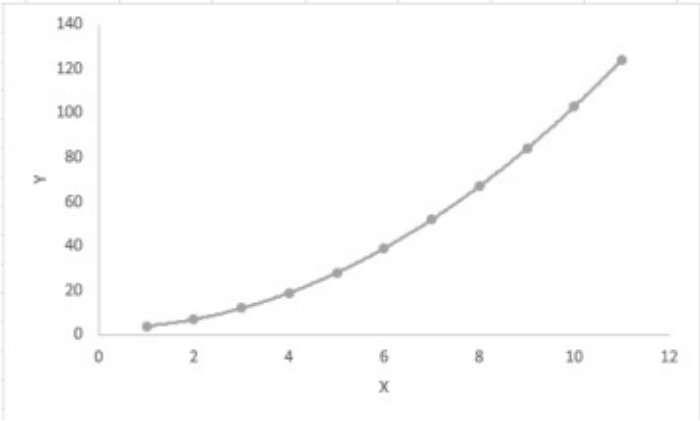
  Since it is widely implemented for classification problems, linear regression generally requires large sample sizes. It can also handle different types of relationships because it is based on a non-linear log transformation approach for predicting odds ratios.

- **Stepwise regression:** Stepwise regression is used to perform operations with multiple independent variables. By observing statistical values such as AIC metric, t-stats and R-square, the selection

of independent variables is performed with the support of an automatic process. Stepwise regression brings benefit to the regression model by adding covariates one at a time. The procedure is based on a predefined criterion.

Standard stepwise regression adds and removes predictors whereas the forward selection approach starts with a significant predictor and adds variables for every step in the machine learning model. Furthermore, backward elimination begins with all predictors in the model and also removes the least significant variable for very step.

- **Polynomial regression:** If the power of an independent variable is more than 1, the regression equation is referred to as a polynomial regression equation. The best line for this approach is a curve that fits into data points defined in the graph below:



Mathematical representation:

$$Y= a+b*x^2$$

- **Ridge regression:** Ridge Regression technique is implemented when data undergoes multicollinearity. This is a situation in which independent variables are highly correlated and the least square estimates are totally unbiased. A degree of bias is added to regression

estimates in order to reduce standard errors and deviations. There are high chances of the occurrence of prediction errors because of the two subcomponents known as biased and variance. For example:

Y=a+y a+ b1x1+ b2x2+…. +e; for multiple independent variables.

We must remember that the assumptions of regression are the same as the least squared regression except for the fact that it cannot be assumed normally. Moreover, the value of coefficients will never reach a zero point which is also known as a regularization method.

- **Least absolute shrinkage and selector operator regression (lasso regression)**

  Lasso Regression has the capability of reducing the variability and works similar to ridge Regression. Along with improving the accuracy of linear regression models, lasso regression also leads to penalizing the sum of absolute values of the estimates that eventually decrease the value some of the parameters estimates to zero. Depending upon the penalty applied, the estimates get shrunk towards absolute zero which in return helps in feature selection.

  Being a regularization method, lasso picks only a single option from a group of predictors and shrinks the others to zero. Lasso regression is mainly used in machine learning algorithms to gain better data insights and making predictions more accurate.

Selecting A Regression Model

There is a simple rule for selecting a regression model. If the outcome is continuous, linear regression can be applied and if the outcome is binary,

logistic regression should be implemented. With different types and categories of regression models, selecting the best suitable approach is based on other factors such as the type of dependent and independent variables, characteristics of data and dimensionality within the data.

For finding the right regression model, first step to consider is the relationship and impact of variables. On the other hand, we can analyze different metrics such as R-square, statistical significance of parameters, AIC, adjusted R-square, error term and BIC. Furthermore, selection also depends upon your objective and if your data set is having multiple confounding variables, the automatic selection method should not be implemented.

Decision Trees

Machine learning makes use of decision trees as a practical method for supervised learning. The goal is to create a model which has the capability to predict the value of a target variable by learning decision rules from data. Decision trees (DTs) are also considered as a supervised learning technique that has the capability to predict values of responses by learning decision rules. As these rules are derived from features, they can be easily implemented in regression and classification methods.

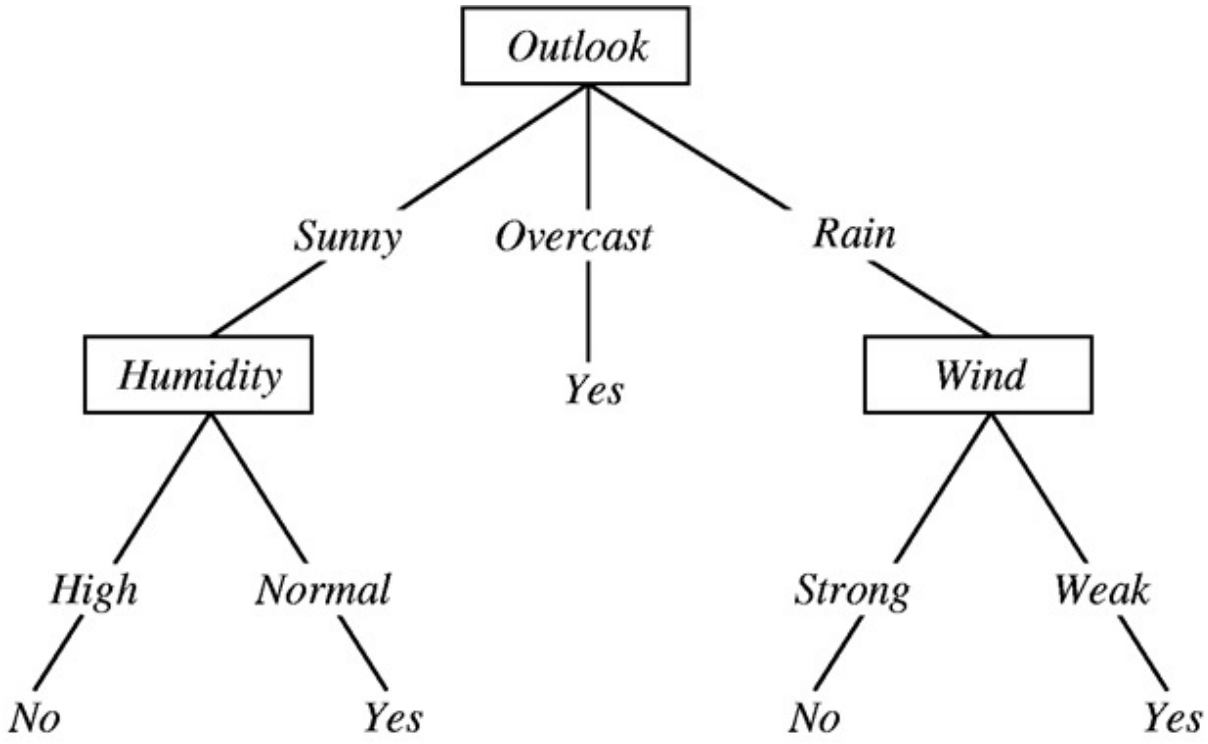
Decision trees are also known as classification and regression trees (CART) which are based on adaptive function models. The goal is to create a model that has the power to predict the value of a target variable for which decision rules are written in *if-then-else* statements only. Here are a few of the main terms to be learned before starting with decision trees:

- **Attribute:** A quantity that describes an instance.

- **Instance:** Attributes that define input space and usually refer to the vector of features.
- **Sample:** A set of inputs combined with a label that is also known as the training set.
- **Concept:** A function that maps an input to an output.
- **Target concept**: The actual answer or correct function.
- **Hypothesis class:** A set of possible functions.
- **Testing set:** An approach used to check the candidate concept and oversee its performance.
- **Candidate concept:** Target concept.

A decision tree is the same as a tree graph, with nodes representing the place of attribute as a question, edges representing answers to the questions and leaves showing the actual class label or output. They are also used in non-linear decision-making approaches along with simple linear decision surface models. Each node in the tree serves as a test case for a sample attribute and each edge descending from that specific node corresponds to one of the possible answers for the test case. This process is repeated for every subtree rooted at the new nodes and is recursive in nature.

Here is an example of a decision tree representing weather:



When looking at the algorithm of decision trees, we first need to select the best attribute or feature that can separate the data in the best possible manner. Next, we need to ask the relevant question and follow the answer path. The expressiveness of decision trees can be represented through any Boolean function of the input attributes. Boolean gates to be used are AND, OR and XOR.

Steps For Making A Decision Tree

1. Get the data set or list of rows that are required for making the decision tree and calculate the uncertainty for the data set.
2. Create a list of all relevant questions that need to be asked at that specific node.

3. Depending upon the question, partition the rows into “True” rows and “False” rows.
4. Calculate the information gain through the partition of data from the previous step.
5. Update the question based on the information gain.
6. To finish, divide the node on the question and start again from step 1 until the leaf node is achieved.

Below is the sample code to implement a decision tree in Python programming language:

```
import numpy as np

import pandas as pd

from sklearn.metrics import confusion_matrix

from sklearn.cross_validation import train_test_split

from sklearn.tree import DecisionTreeClassifier

from sklearn.metrics import accuracy_score

from sklearn.metrics import classification_report


# Function importing data set

def importdata():

    balance_data = pd.read_csv(

'https://archive.ics.uci.edu/ml/machine-learning-'+

'databases/balance-scale/balance-scale.data',
```

```
    sep= ',', header = None)

    # Printing the dataswet shape
    print ("data set Length: ", len(balance_data))
    print ("data set Shape: ", balance_data.shape)

    # Printing the data set obseravtions
    print ("data set: ",balance_data.head())
    return balance_data

# Function to split the data set
def splitdata set(balance_data):

    # Seperating the target variable
    X = balance_data.values[:, 1:5]
    Y = balance_data.values[:, 0]

    # Spliting the data set into train and test
    X_train, X_test, y_train, y_test = train_test_split(
    X, Y, test_size = 0.3, random_state = 100)
```

```
    return X, Y, X_train, X_test, y_train, y_test

# Function to perform training with giniIndex.
def train_using_gini(X_train, X_test, y_train):

    # Creating the classifier object
    clf_gini = DecisionTreeClassifier(criterion = "gini",
            random_state = 100,max_depth=3, min_samples_leaf=5)

    # Performing training
    clf_gini.fit(X_train, y_train)
    return clf_gini

# Function to perform training with entropy.
def tarin_using_entropy(X_train, X_test, y_train):

    # Decision tree with entropy
    clf_entropy = DecisionTreeClassifier(
            criterion = "entropy", random_state = 100,
            max_depth = 3, min_samples_leaf = 5)
```

```
    # Performing training
    clf_entropy.fit(X_train, y_train)
    return clf_entropy


# Function to make predictions
def prediction(X_test, clf_object):

    # Predicton on test with giniIndex
    y_pred = clf_object.predict(X_test)
    print("Predicted values:")
    print(y_pred)
    return y_pred

# Function to calculate accuracy
def cal_accuracy(y_test, y_pred):

    print("Confusion Matrix: ",
        confusion_matrix(y_test, y_pred))

    print ("Accuracy : ",
```

```
    accuracy_score(y_test,y_pred)*100)

    print("Report : ",
    classification_report(y_test, y_pred))

# Driver code
def main():

    # Building Phase
    data = importdata()
    X, Y, X_train, X_test, y_train, y_test = splitdata set(data)
    clf_gini = train_using_gini(X_train, X_test, y_train)
    clf_entropy = tarin_using_entropy(X_train, X_test, y_train)

    # Operational Phase
    print("Results Using Gini Index:")

    # Prediction using gini
    y_pred_gini = prediction(X_test, clf_gini)
    cal_accuracy(y_test, y_pred_gini)
```

```
    print("Results Using Entropy:")
    # Prediction using entropy
    y_pred_entropy = prediction(X_test, clf_entropy)
    cal_accuracy(y_test, y_pred_entropy)


# Calling main function
if __name__=="__main__":
    main()
```

(Sharma, n.d.)

How Does A Decision Tree Work?

At first, a decision tree selects the best attribute. To perform this task, specific attribute selection measures (ASM) are implemented to split records. For making the attribute a decision node, the data set is broken into smaller subsets. Next, the decision tree algorithm starts tree building by repeating this process recursively for every child until all the tuples belonging to the same attribute value or there are no more remaining attributes left in the tree. In the case that there are no more instances left, the decision tree algorithm will repeat the process to meet the given conditions.

For choosing the splitting criterion, attribute selection measure is used because it is a heuristic approach to split the partitioned data in a suitable manner. This technique is also known as splitting rules because it allows

researchers to determine tuple breakpoints on a node. The best score attribute is selected as a splitting attribute because the attribute selection measure gives a rank to each feature or attribute from the given data set. Remember that split points for branches need to be defined for a continuous valued attribute. Popular selection measures for this are gain Ration, Gini Index and Information Gain.

Impurity concept is used for determining which of the splits is better and it also concludes that none of the leaves deliver a 100 percent result for a single node. However, for calculating Gini impurity of the left leaf, the fraction of left node is subtracted by 1 and the fraction of people from the right node.

How To Avoid Overfitting

In a decision tree model, overfitting is a major problem occurring in every machine learning model. Models that are overfitted are more likely to poorly generalize new sample.

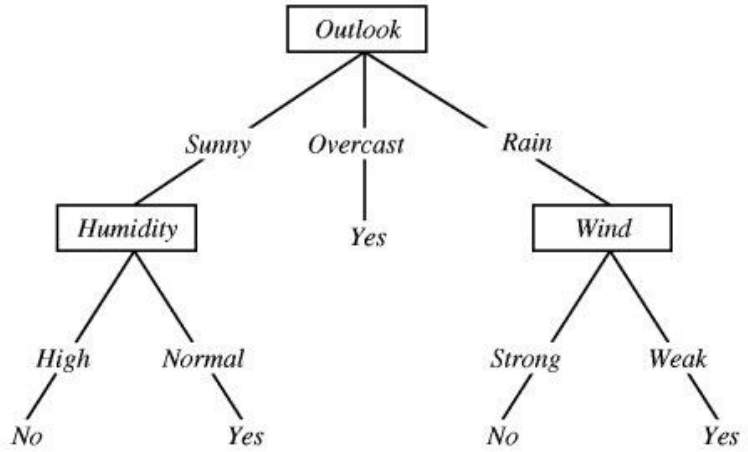
To stop decision trees from overfitting, the branches that make use of features that have low importance are removed. This approach is also known as post-running or pruning. Without effective predictive accuracy, pruning has the capability to reduce the size of a learning tree through a cross validation set.

Major pruning techniques are minimum error and smallest tree. In the minimum error technique, a tree is pruned back to the point where the cross validated error is a minimum whereas the smallest tree approach occurs when the tree is pruned back slightly further than the minimum error. In this way, pruning creates a decision tree with cross validation error within 1 standard error of the minimum error.

To prevent overfitting, we can try and stop the tree building process before it starts to produce leaves with small samples. At each stage of splitting a tree, we can use the early stopping heuristic or pre-pruning decision trees to check the cross-validation error. In case the error does not decrease, early stopping and pruning can be used in combination to save time.

Random Forests

Random forests are a combination of hundreds of thousands of decision trees. Each single tree is trained through a slightly different set of observations to split nodes in every tree with a limited number of features. The final predictions of random forests are made by averaging predictions of each individual tree. Random forests are bagged decision tree models that split data into smaller groups that are based on the features of data to create a smaller set of data which rests under one single label only.



In the above example of a decision tree, we can notice multiple features until we derive to a conclusion of “Yes” to playing tennis and “No” to not playing

tennis. To determine the decision for the lines along the tree, the conclusion for the overcast outlook is “Yes” to playing tennis and is the outlook is sunny along with high humidity, then the outcome is “No” resulting in the conclusion that we should not play tennis. Random decision tree models have splits that are chosen according to a purity measure and at each node, the information gain must be maximized.

Bagging

Bootstrap aggregation is a simple yet powerful ensemble method which is also known as bagging. An ensemble method is an approach that combines predictions from various machine learning algorithms together for achieving mode accurate predictions as compared to any individual model. Decision trees are sensitive to the specific data they are trained for and in case the training data is changed, the resulting decision tree can be different and will yield predictions in other ways.

Bagging can also be considered an application of the bootstrap procedure which is implemented to high variance machine learning algorithms and decision trees. Random forests are an enhancement of bagging which allows algorithms and decision trees to improve variable selection processes. Bootstrap aggregation decision trees are very sensitive to the data on which they are trained because small changes in a training set can result in significantly different tree structures.

In bagging, the training data is not divided into subsets, and each training tree is given a specific sample size. Instead of the original training data, we can utilize a random sample of size N as a replacement. In a normal decision tree, feature randomness occurs when a node is split, and we have to select the one that produces the most separation between left and right nodes. For

this activity, different features and methods can be implemented because each tree in a random forest can select only a random subset of features.

How Are Random Forests Better Than A Decision Tree?

Random forests are better than a single decision tree because they are used for pooling predictions and can be incorporated with other machine learning models, which a single decision tree cannot do. As each individual brings their own information and background information sources to any problem, a limited scope of data are utilized to perform specific tasks in machine learning and artificial intelligence.

There are several sources on which we can rely to make a prediction, and in a decision tree, a random subset of training data points is utilized to increase diversity in the forest. When it comes to making a prediction, random forests usually take advantage of all individual decision tree estimates.

Features

Random forest models are designed to provide high accuracy predictions because the biological interpretability of generated models is complex. Furthermore, an ensemble of regression trees is difficult to analyze from

direct connectivity or biological topology perspective. When compared to linear regression models, we can determine that they have the capability to deliver direct interpretability of individual features in terms of regression coefficients. To overcome this scenario, extra analysis of random forest models can be performed to evaluate different features.

The Naïve Bayes Method For Machine Learning

The naïve Bayes algorithm is a powerful approach for predictive modeling. In machine learning, operations are performed on given data for which we are required to select the best hypothesis and utilize our prior knowledge about the problem to achieve accurate predictions. The Bayes theorem provides a great way to calculate the probability of a hypothesis, as is defined below:

$$P(h|d) = (P(d|h) * P(h))/P(d)$$

For this equation, P(h|d) is the probability of hypothesis h from provided data d. This approach is also known as posterior probability. P(d|h) is known as the probability of data d provided that the hypothesis h is true. Moreover, P(h) is considered as hypothesis h being true and it is also known as the prior probability of h. At last, P(d) is the probability of data without consideration of hypotheses and data.

This theorem is also considered to calculate the posterior probability for different types of hypotheses, and we can select the hypothesis that has the highest probability. If we are given an even number of instances for the training data, then the probability of every class will be considered as equal. For this situation, the equation can also be expressed as follows:

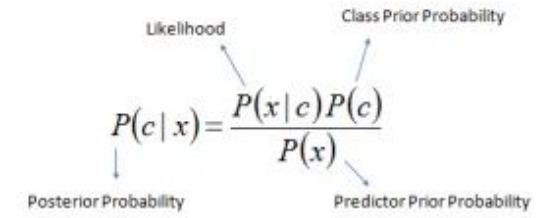
MAP(h) = max(P(d|h))

The Naïve Bayes Classifier

The naïve Bayes classification algorithm is used for multi-class and binary classification problems so that the probabilities for each hypothesis are simplified. This technique is simple to understand when described through categorical or binary input values. Class probabilities and conditional probabilities are known as the representation of the naïve Bayes model. In class probabilities are the probabilities of each class in the training data set whereas conditional probabilities are determined from each input value given each class value from the probability.

How Does The Naïve Bayes Algorithm Work?

To better understand the functionality of naïve Bayes algorithm, we must understand the method of how to calculate the posterior probability P(c|x) from P(c), P(x) and P(x|c). The equation for naïve Bayes algorithm is defined as follows:



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

To understand the algorithm in a better way, we can convert the given data into a frequency table and use the naïve Bayesian equation for calculating posterior probability for each class. The class having the highest posterior probability is likely to be the outcome of prediction. Naïve Bayes makes use

of the same method for predicting the probability of the next class depending upon some attributes and the algorithm is widely used for text classification as well.

Advantages Of Using The Naïve Bayes Algorithm

The naïve Bayes algorithm is widely used in machine learning models because it delivers outstanding results for multi-class prediction. It is generally used for categorical input values and performs better than other models, such as logistic regression, because it requires a lot less training data.

The algorithm is an eager learning classifier and is best suitable for real time prediction. Prediction for the probability of multiple classes of target variables through the multi-class prediction feature is also possible through this algorithm.

Naïve Bayes classifiers are generally used in text classification and have a higher success rate as compared to other machine learning algorithms. Moreover, the naïve Bayes classifier and collaborative filtering approach together create a recommendation system which uses data mining and other machine learning techniques to filter unseen information.

Improving The Naïve Bayes Model

To improve the naïve Bayes model, we can consider the simple tips and techniques listed below:

- Apply the Laplace correction technique if the data set has a zero-frequency issue for predicting the class of a test data set.

- Use the transformation of any other method to convert data into normal distribution if continuous features are not achieving normal distribution.
- Delete correlated features because highly correlated features are considered twice on the model and often lead to overinflating.
- The naïve Bayes algorithm has no variance to minimize, and techniques such as bagging, boosting or assembling must not be used while applying a classifier combination technique.
- Focusing on the preprocessing of data and feature selection is important because naïve Bayes classifiers have limited options for parameter tuning.
- The naïve Bayes classifier assumes that each feature in the data is unrelated to one another and the presence or absence of any feature will never affect the presence or absence of any other feature.

  In real data sets, a test hypothesis is given with multiple evidence by which calculations become complicated. To solve this problem, we use the feature independence approach to uncouple multiple evidence and take each evidence as an independent one.

The KNN Algorithm

The K-nearest neighbors' algorithm is also known as the KNN algorithm. It is a supervised machine learning algorithm and is implemented to solve classification and regression predictive problems. KNN is well-defined, with two major properties: it is a lazy learning and non-parametric learning algorithm. It is considered as a lazy algorithm because the model does not

have a specialized training phase and utilizes all data for training while in the classification phase. Furthermore, the algorithm is a non-parametric learning algorithm because it never assumes anything about the underlying data.

Steps For Implementing The KNN Algorithm

1. Load the data and initialize K to your selected number of neighbors. Training data and test data must be uploaded as well.
2. Determine the value of K and calculate the distance between the current example and the query example from the given data.
3. For every point in the test data, calculate the distance between test data and each row of training data for which methods, such as Manhattan, Hamming distance or Euclidean, can be considered.
4. Sort the values in ascending order based on the distance value and the algorithm will choose top k rows and assign a class to the test point which is based on the most frequent class of the rows.
5. Get the labels of selected k entries. For regression, return the mean of k labels, and for classification, return the mode of k labels.
6. End.

The model representation for the KNN algorithm is the entire training data set and it has no model except of storing the entire data set. Because there is no learning required, the algorithm is easy to update as new data becomes available and old data can be removed without any hassle. Efficient implementations allow the KNN algorithm to store data by taking help from complex data structures such as k-d trees for matching new patterns during

prediction efficient. Predictions for a new instance (x) are made by undergoing the entire training set for k similar instances.

The popular distance measure approach for the KNN algorithm is Euclidean distance which is calculated from the equation below:

EuclideanDistance(x, xi) = sqrt(sum((xj – xij)^2))

Other distance measures, such as cosine distance, Jaccard, Tanimoto and Mahalanobis, can be considered too by overviewing the properties of your data. K-value is found through algorithm tuning because the computational complexity of KNN increases depending upon the size of the training data set.

Applications Of The KNN Algorithm

The KNN algorithm is widely used in banking systems to predict whether the applicant is fit for loan approval or not because it compares the characteristics of the applicant to the default ones. The model is also implemented to find credit ratings for individuals by comparing them with people who have similar traits. Other applications of KNN algorithm include handwriting detection, speech recognition, video recognition and image recognition.

Below is the function code to predict KNN in R:

```
knn_predict <- function(test, train, k_value){
  pred <- c()
  #LOOP-1
  for(i in c(1:nrow(test))){
```

```
        dist = c()
        char = c()
        setosa =0
        versicolor = 0
        virginica = 0
}
        #LOOP-2-looping over train data
        for(j in c(1:nrow(train))){}

      dist <- c(dist, ED(test[i,], train[j,]))
      char <- c(char, as.character(train[j,][[5]]))

        df <- data.frame(char, dist$SepalLength)
        df <- df[order(df$dist.SepalLength),]       #sorting dataframe
        df <- df[1:k_value,]
        #Loop 3: loops over df and counts classes of neighbors.
        for(k in c(1:nrow(df))){
    if(as.character(df[k,"char"]) == "setosa"){
    setosa = setosa + 1
     }else if(as.character(df[k,"char"]) == "versicolor"){
    versicolor = versicolor + 1
```

```
      }else
    virginica = virginica + 1
        }
        n<-table(df$char)
   pred=names(n)[which(n==max(n))]
  return(pred) #return prediction vector
}

#Predicting the value for K=1
K=1
predictions <- knn_predict(test, train, K)
```

Output:

For k=1

[1] “Iris-virginica”

(Srivastava, 2018)

KNN algorithm is a simple supervised machine learning algorithm which is used to solve regression and classification problems. It is very easy to implement. However, the algorithm is reported to become very slow when provided with heavy data inputs.

Chapter 7: Types Of Machine Learning Continued

Unsupervised Learning

As we now know, machine learning is divided into two major types called supervised learning and unsupervised learning. In this chapter, we will be discussing unsupervised learning in detail and how it is implemented in machine learning models.

In machine learning, unsupervised algorithms are used to create inferences from data sets that are comprised of input data without labeled responses. Unsupervised learning models are provided with input data (x) and have no correspondence with output variables. This approach is known as unsupervised learning because the algorithms are supposed to discover and present the structure in data without any instruction.

Unsupervised learning models are further classified into association and clustering problems. Traditional data sets in machine learning worked through labels and followed the logic of "x leads to y." In the case of unsupervised learning models, there are no outcomes and the model just analyzes through the inputs. No training is given to the machines and therefore they are programmed to find the unknown structure in unlabeled data on their own. Unsupervised learning is also known as the training of an artificial intelligence algorithm which is neither classified nor labeled. This allows the algorithm to act on the information without guidance.

Artificial intelligence systems are capable of handling unsupervised learning models for which they generally use a retrieval-based approach. They also have the capability to perform more complex processing tasks when

compared to the supervised learning systems. Major applications of unsupervised learning models are self-driving cars, chatbots, expert systems and facial recognition programs.

Clustering And Association

Clustering

Clustering problems occur when we must discover the inherent groupings in data. In the association rule learning problem, we are required to discover rules that describe large proportions of data. Clustering is one of the major types of unsupervised machine learning algorithms as it runs through the given data and finds any natural clusters if they exist. For example, there could be different groups of users that are differentiated across a few criteria. For example, the criteria can be age, gender or height.

Major types of clustering are defined as follows:

- **Probabilistic clustering:** This involves clustering data points into clusters on a specific probabilistic scale.
- **K-means clustering:** K-means clustering involves the clustering of data points into a number (k) of mutually exclusive clusters whereas the method is more focused on selecting the right number for K.
- **Hierarchical clustering:** Hierarchical clustering classifies data points into parent and child clusters.
- **Gaussian mixture models:** These are the models featuring a mixture of multivariate normal density components.
- **Hidden Markov models:** Used to observe data and recover the sequence of states in the model.

- **Self-organizing maps:** Self-organizing maps use neural networks that learn the distribution and topology of the data.

Generative Models

Generative models are a major part of unsupervised learning models. They create new samples from the same distribution of the given training data. Generally, these models are designed to learn and discover the essence of the given data for generating similar data. Image data set is an example of generative models and has the capability to create a set of images like the given set. Generative models are purely based on the characteristics of unsupervised machine learning models.

Data Compression

Data compression is the process for keeping data sets as small and efficient as possible. Unsupervised learning methods support data compression and perform the task through the dimensionality reduction process. The dimensionality reduction approach assumes that the given data is redundant and utilizes the same concepts as the information theory. Also known as the number of columns in a data set, dimensionality reduction can be used to represent information in a data set with only a specific size of the actual content.

Singular value decomposition (SVD) and principal component analysis (PCA) are the main approaches for data compression. PCA is used to find linear combinations that communicate variance in data whereas the SVD model factorizes data into the product of three other, smaller matrices.

For generating new samples from the same distribution, generative models are used, and they are a major part of the unsupervised machine learning

approach as well. The model also has the power to automatically learn the features of given data and implement them into the produced data.

Association

Association is a rule-based machine learning method that is used to discover relations between databases. In large sets of data items, we can find interesting associations such as dependencies and relationships through association rules. In data sets, items are generally stored in the form of transactions which can be generated or extracted through external processes and relational databases. Unsupervised machine learning delivers full scalability to association rules algorithms which in return help in accumulating the ever-growing size of data. Furthermore, the evaluation and discovery of interesting associations make is easier for companies and industries to handle big data.

What Is An Association Rule?

An association rule is a crucial part of unsupervised machine learning concepts. This rule has the capability to find relationships between sets of elements of every distinct transaction and does not extract an individual's preference. An association rule is comprised of a consequent and an antecedent as both of them are list items. The item set is the list for all items in the consequent and antecedent. A typical example of an association rule is market-based analysis, explained as follows:

TID	ITEMS
1	Bread, Milk

2	Bread, Biscuits, Drink, Eggs
3	Milk, Biscuits, Drink, Coke
4	Bread, Milk, Biscuits, Drink
5	Bread, Milk, Drink, Coke

Basic definitions:

Support count for frequency of occurrence of an item set:

Here ({Milk, Bread, Biscuit}) =2.

Association rule for implication expression of the form X->Y, where X and Y are any of the two item sets. For example:

{Milk, Biscuits}->{Drinks}

A frequent item set is an item set having support greater than or equal to the minimum threshold of the data. To better understand the strength of association between the item set s, we can make use of the following metrics:

- **Support:** The support measure gives an idea of how frequently an item set occurs in all transactions. Taking the example of itemset1 = {Milk} and itemset2 = {Drinks}, there will be more transactions containing milk as compared to those containing drinks. This means that itemset1 has a higher support than itemset2.

  In order to calculate the total number of transactions in which the item set occurs, we can make our calculation through the following mathematical equation:

$$Support(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Total\ number\ of\ transactions}$$

- **Confidence:** In unsupervised machine learning, a confidence measure is used to define the likeliness of occurrence of consequent. Given that the cart has antecedents already, a confidence measure could be implemented on all transactions of an item set. For example, of all the transactions containing {James}, how many had {Drink} on them? We can determine by common knowledge that {James} -> {Drink} should be a high confidence rule. Equation to calculate confidence mathematically is defined as follows:

$$Confidence(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Transactions\ containing\ X}$$

  It does not matter what type of item set is present in the antecedent in the case of a frequent consequent because the confidence for an association rule that has a very frequent consequent will be high.

- **Lift:** The lift approach is implemented to support the frequency of a consequent while calculating the conditional probability of the occurrence of {Y} given {X}. It is also defined as the rise in probability of having {Y} on the cart along with the knowledge of {X} being present over the probability having {Y} in the cart. This operation is performed without the presence of {X}. In cases where {X} leads to {Y} in the cart, remember that the value of lift will always be greater than 1. The mathematical equation to represent Lift approach is defined as follows:

$$Lift(\{X\} \rightarrow \{Y\}) = \frac{(Transactions\ containing\ both\ X\ and\ Y)/(Transactions\ containing\ X)}{Fraction\ of\ transactions\ containing\ Y}$$

How Does An Association Rule Work?

In machine learning and data mining, association rules are used to represent the probability of relationships for data items in large data sets. Actually, they are *if-then* statements that are implemented in various models to discover sales correlations in transactional data.

At the basic level, association rule mining involves the use of data patterns and co-occurrence. Machine learning models are used to identify frequent *if-then* associations which are also known as association rules.

Popular machine learning algorithms that are associated with association rules are SETM, AIS and a priori. With the help of the AIS algorithm, the model can generate and count item set s as it scans data. This also helps the algorithm in determining which large item set s contained a transaction and how new candidate items are created by entering large item set s along with other items in the transaction data. Moreover, SETM algorithm is yet another way to generate candidate item set s while the model scans a database.

SETM algorithm is accountable to item set s at the end of scan because new candidate item set s is generated in the same way as the AIS algorithm. Along with this operation, transaction ID of the generated transaction is saved along with the candidate item set for a sequential structure. An a priori algorithm allows candidate item set s to generate only with the support of large item set s from the previous pass. Large item set s of the previous pass can be joined to generate specific item set s of a larger size.

Association rules are a powerful source in data mining because they are used for predicting customer behaviors and analyzing customer analytics, product clustering, store layout and market-based analysis. Generally, programmers make use of association rules to build programs that are capable of handling

machine learning. Association rule mining actually searches for frequent items in the data set and is used to generate interesting associations and correlations between item set s on both relational and transactional databases.

Density Estimation

Density is the relationship between observations and their probability. Some outcomes of a random variable have low probability density while other outcomes have a higher probability density. For sample data, probability density is considered as probability distribution which is a major part of machine learning models.

In order to determine the probability density function for sample data, we must determine whether the given observation is unlikely or not. Selecting appropriate learning methods that require input data is best for obtaining a specific probability distribution.

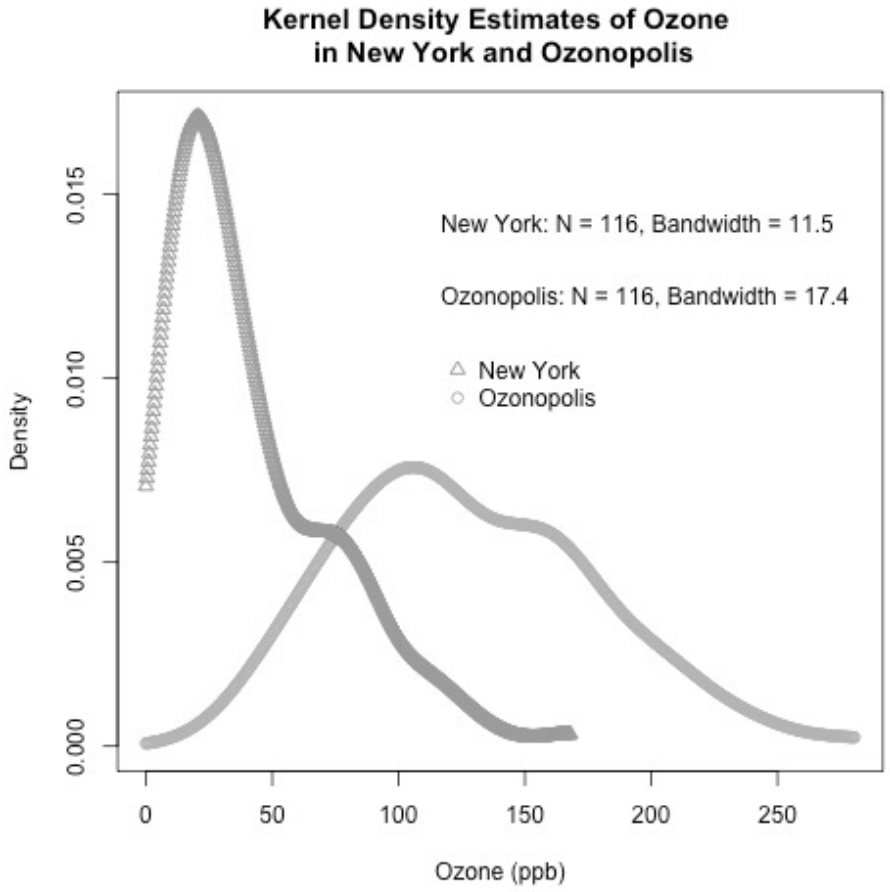
The probability density function for a random sample of data is calculated because the probability density is supposed to be approximated through the probability density estimation process. Density estimation uses statistical models to find underlying probability distribution which in return gives rise to the observed variables.

Kernel Density Estimation

Kernel density estimation is a method of estimating the probability density function of a continuous variable. Being a non-parametric approach, kernel density estimation does not assume any underlying distribution for the variable and is exclusively created with datum.

The probability density function is then estimated by adding all of the given kernel functions and dividing the total number to ensure that every possible value of the probability density function is non-negative and is a definite integral of probability density function over its support sets equal to 1.

For example:



The above graph is based on the "Ozone" data from the built-in "air quality" data set and the previously simulated Ozone data for Ozonopolis City. It also shows rug plots and how they affect kernel density plots. It can be determined that Ozonopolis has more ozone pollution when compared to New York City, and this is the main reason why the density plots have higher bumps at specific places.

Parameters such as bandwidth and smoothing parameter have the capability to control the scope from data samples which contribute to estimating the probability of a given sample. Moreover, kernel density is also referred to as the Parzen-Rosenblatt window.

Pseudocode for kernel estimation:

- **Input**: Kernel function K(x), kernel width w, training data instances x1, xn.
- **Output:** An estimate of the probability density function underlying the training data.
- **Process:** initialize dens(x) = 0 at all points x

```
for i = 1 to n:
    for all x:
        dens(x) += (1/n)*(1/w)*K((x-xi)/w)
```

Latent Variables

What Are Latent Variables?

Latent variables are the variables that are inferred from other variables and are not directly observed. Variables that are observed are also known as directly measured variables and are implemented mathematically to design latent variable models in economics, machine learning, artificial intelligence, psychology and natural language processing. A latent variable cannot be observed neither in training nor in the testing phase and these variables can also not be measured on a quantitative scale.

Taking the example of a probabilistic language model of new articles, we can notice that each article x focuses on a specific topic y. For example, sports, economics, politics or finance articles each focus on their respective topics. By using this knowledge, we can build a more accurate model that can also be expressed as p (x | t)p(t) for which additionally added variable t is unobserved. However, we cannot use learning methods because it is unobserved, and the unobserved variables often make learning a difficult task.

Gaussian Mixture Models

Gaussian Mixture Models are also considered as latent variable models. These models are widely being used in machine learning because they postulate that our observed data is comprised of k clusters. It is assumed that each cluster is specified by $\pi 1,\ldots,\pi K$; whereas the distribution within each cluster is known as Gaussian.

Chapter 8: Machine Learning And Business

Machine learning extracts meaningful insights from raw data to solve complex and data rich business problems. By focusing on the development of computer programs and applications that can access data for learning autonomously, different activities such as data preparation, deploying, tuning, sharing of machine learning models and collaboration can be performed without any issues.

Machine learning in business also helps in enhancing business scalability along with improved operations for different companies from all over the world. Artificial intelligence tools and machine learning algorithms are designed to perform operations over big data and learn without human intervention. Organizations and companies can benefit by understanding the need to implement machine learning algorithms into their systems.

Benefits Of Machine Learning To The Business World

Machine learning allows businesses to grow and flourish by analyzing big data, user trends and market scenarios in an effective way. As it extracts meaningful insights in order to solve complex and data rich business problems from raw data, algorithms have the capability to learn from data iteratively and allow computers to figure out different categories of hidden insights without being explicitly programmed. Machine learning models and algorithms have gained immense popularity in the business analytics community to cater factors such as growing volumes of data.

With the increasing level of data, companies are now taking advantage of machine learning and artificial intelligence models to manage operations and

everyday tasks. Without the need to regularly update the systems, big data can be effectively handled by machine learning algorithms as they are designed to learn through experiences.

Industries related to finance, economics, education, transport, healthcare and medicine have now updated their systems with the latest machine learning models to automate processes for accurate results.

Applications Of Machine Learning

- **Financial analysis:** Machine learning is used to perform financial analysis for companies that have large volumes of accurate and quantitative historical data. As it is already being used for algorithmic training, portfolio management, fraud detection and loan underwriting, future applications of machine learning might include chatbots for customer service, sentiment analysis and security purposes. Machine learning models learn with experience and utilize data to make future financial predictions in a better way.
- **Predictive maintenance:** Corrective and preventive maintenance are a major part of manufacturing industries. Although this process is complex and expensive when conducted with conventional approaches, machine learning has now made it easier to discover meaningful insights and hidden patterns in factory data. Because this process helps in reducing risks associated with unexpected failures, companies can also reduce unnecessary expenses by implementing machine learning models. What's more, artificial intelligence and machine learning algorithms work in collaboration to analyze historical data and ensure workflow visualization.

- **Fraud detection:** Machine learning models are best for detecting spam and fraud in systems. Spam filters are now designed by using the latest artificial intelligence and machine learning algorithms which include neural, network rules for detecting phishing messages and spam.

  Major search engines such as Google and Bing, work perfectly because the system is programmed to learn the ranking of pages and eliminate spam in real time. They are powered by effective machine learning algorithms. Also, there is a specific number of spam filtering approaches that are used by email clients to make sure that these spam filters are regularly updated.

  To perform such activities, a rule-based spam filtering approach is considered along with multi-layer perceptron and decision tree induction. Furthermore, system security programs that are designed by artificial intelligence and machine learning algorithms understand the coding pattern and remove fraudulent activities before they cause any problems. It is predicted that online credit card fraud will reach an expected worth of $32 billion in 2020. So, this is a serious issue that must be solved quickly.

  Fraud detection is one of the major applications of machine learning that helps make online transactions safer. The number of transactions from credit cards, debit cards, UPI, numerous wallets and smartphones have been increasing with time and have also been a major target for online criminals. To make online transactions smooth and safe, machine learning model thoroughly investigates the processes and searches for suspicious patterns. This approach in machine learning is also known as the classification problem and is handled in real time.

- **Image recognition:** Image recognition is also known as computer vision which involves machine learning, data mining and database knowledge for discovery. Having the capability to produce symbolic and numeric information from high dimensional data and images, the approach is widely used in various industries such as automobiles and healthcare. Moreover, image recognition also helps in detecting criminals or suspected persons in real time.
- **Dynamic pricing:** Setting a fixed price for a service or a product is considered as a traditional means of trading. Nowadays, pricing strategies are based on objectives and targets for which different kinds of promotions and discounts are given. Services, such as air tickets or cab fares, are dynamically priced depending upon various external scenarios such as traffic or number of passengers.

  Artificial intelligence is now able to track buying trends and provides pricing solutions for determining competitive product prices. Further, machine learning algorithms are implemented to predict accurate service rates depending upon load and user demands.
- **Medical diagnosis:** Machine learning in the medical industry has supported healthcare organizations and companies in great ways. By reducing medical costs, the technology has provided effective treatment plans and better diagnostic tools in order to improve patients' health. Also, machine learning is now being used in healthcare systems for identifying high-risk patients, recommending medicines and predicting readmissions as well.

  Prediction and analysis in the medical industry is generally performed through machine learning algorithms because they work with patient records and data sets to yield relevant outcomes. The healthcare

industry has greatly benefited from artificial intelligence and machine learning because these technologies have reduced overall operational costs and bring highly accurate prediction results.

- **Google Translate:** Google Translate is yet another amazing application of machine learning that has helped people in translating and understanding the text of any language. Google has introduced the Google Neural Machine Translation (GNMT) model that is based on neural machine learning.

  Having the capability to work on thousands of languages and dictionaries, Google Translate uses the natural language processing approach for finding the accurate translation of words and phrases. Other techniques, such as named entity recognition, POS tagging and chunking, are also modeled in Google Translate systems to yield accurate translation results.

- **Customer lifetime value prediction:** Machine learning is an approach that can be effectively used to derive meaningful business insights. Customer lifetime value prediction and customer segmentation are two of the major challenges that are being faced by marketers. To handle the growing number of customers, companies have to update their systems with artificial intelligence and machine learning algorithms for better forecasting results.

  Thus, machine learning and artificial intelligence allows businesses to predict customer behavior and purchasing patterns. Furthermore, companies can greatly improve their sales by analyzing customer purchase trends over the internet.

- **Financial evaluation:** The finance industry has taken advantage of the benefits associated with artificial intelligence and machine learning

models. Machine learning algorithms have the capability to perform in-depth financial analysis and is widely implemented for different procedures such as algorithmic trading, fraud detection and portfolio management.

The models have the capability to give an overview of an individual's past record and credit history in loan approval cases. Financial institutions and companies can also stop fraudulent activities, such as illegal transactions or money laundering, by using machine learning algorithms, as artificial intelligence and machine learning models have the capability to stop such activities and report them in real time.

- **Social media:** Social media platforms have millions of active users with tons of data stored in their storage systems. In order to handle big data analytics and prediction requirements at a larger scale, companies, such as Facebook and Twitter, have state of the art data centers that are designed with powerful machine learning and artificial intelligence algorithms. Taking the example of Facebook, we can notice that whenever a picture of you with a friend is uploaded, Facebook instantly recognizes the person and suggests a tag. This activity is made possible through face recognition machine learning models.

  Moreover, Facebook's "people you may know" feature is based on machine learning models that are designed to understand and make decisions based on experience. To make this happen, Facebook consistently notices the friends you are in contact with and make suggestions accordingly.

- **Online customer support:** To cater customer requirements, websites are now taking the help of chatbots to instantly answer their queries

and present the most suitable solutions. Not all websites have deployed a customer support representative at the backend. However, in most cases, all you need is to talk with a chatbot. These chatbots are designed with effective machine learning models that learn with time and present the required information to customers instantly. The ability to understand customer queries and problems makes chatbots one of the best options to be considered for online businesses.

- **Product recommendations and guidance:** Online shopping has become the latest trend and it is expanding globally. We can see that online shopping stores or websites often recommend suitable items through email or on the website. This activity is performed through intelligent machine learning algorithms that understand customer requirements and suggest options accordingly. Depending upon the market trends and behavior of the customers, the system is designed to learn and adapt to new changes for making accurate predictions. Factors, such as brand preferences, website purchases or liked items, allow machine learning models to send relevant predictions and suggestions to the customers regularly.

- **Search engines:** Search engines, such as Google and Bing, are powered by machine learning algorithms to improve search results and rankings. Each time a search is executed by the user, algorithms at the backend deliver accurate findings and also learn from search results.

  Pages having regular and high amounts of traffic are displayed by the search engines at the top in accordance to the query. In this way, algorithms make use of search engine estimates and improve search results in great ways. The algorithms also match user requirements and show pages that contain the most relevant information.

- **Self-driving cars:** Self-driving or AI powered cars are currently the most impressive technological invention and application of machine learning. Based on the unsupervised learning algorithm, self-driving cars work according to the rules of deep learning and crowdsource data from all similar vehicles and drivers.

  This application has the capability to guide a vehicle in case of emergency situations or drive on its own by receiving and utilizing information from sensor data fusion systems. External and internal information sources such as lidar, cameras, IoT or radars feed the system with relevant data and allow the vehicle to operate on its own.

  In an autonomous car, one major task of the machine learning algorithm is to continuously observe the surroundings and environment to forecast changes. Generally, these sub-tasks include detection of an object, identification of an object and prediction of movement.

- **Traffic alerts:** Google Maps is the most popular application when it comes to location searching and directions. The application is featured with machine learning and artificial intelligence algorithms to find both the shortest and fastest paths. Google Maps is being used widely, by people from all over the world, as it provides accurate locations, best possible routes, average speeds and predictions about traffic rates as well.

- **Government:** Government agencies are regularly in need of machine learning models because they have various data sources and need to oversee information regularly. By identifying useful patterns and insights, data can be analyzed in different ways to minimize cost and increase the efficiency of projects.

- **Cyber security:** Cyber security is one of the major factors that makes the internet safe for online transactions, data transfer and information handling. Take the example of the finance industry, cyber security systems are implemented to stop money laundering and distinguish between legitimate or illegitimate transactions that take place between buyers and sellers. This is all possible due to the implementation of machine learning algorithms that compare millions of transactions taking place all over the world in real time.
- **Email spam:** Spam filtering methods are used by email clients to avoid fake emails or malware. As it can greatly damage a system if opened in a computer, malware software is used for phishing purposes by spammers to obtain highly sensitive information such as bank card details or passwords. Machine learning and artificial intelligence algorithms, such as decision tree induction and multi-layer perceptron, help in avoiding cybercrime and deliver real time protection as well.
- **Video surveillance:** Video surveillance systems are the best way to monitor a specific location. While managing multiple security cameras by a single person is not possible, video surveillance systems are designed with machine learning and artificial intelligence technologies that train computers to do this well.

  Video surveillance systems that are powered by artificial intelligence have the capability to detect and report a crime even before it happens because the system has the capability to track unusual behavior such as someone standing motionless for a long period of time or napping regularly on a specific location.

  Each reported event is handled by a machine learning model at the backend so that immediate responses can be generated to stop any kind

of criminal activity or mishap.

- **Virtual personal assistants:** With the latest advancements in technology, smartphones and digital devices now feature virtual personal assistants that help users find useful information when requested via text or voice. Major applications of virtual personal assistants designed with machine learning models are speech recognition, speech to text conversion, text to speech conversion and natural language processing. Siri, Alexa, Google Assistant and Cortana are few of the most popular virtual personal assistants that help users in performing routine activities every day.

The Impacts Of Machine Learning

Machine learning has revolutionized the working and performance of systems that are being used in our daily lives. From automation to accurate prediction, artificial intelligence and machine learning models have the capability to complete specific tasks without human intervention. The workings and functionality of systems in different industries have now changed totally since machines have been programmed to operate on their own.

We can notice that the need for human labor is decreasing by each coming day which is a major issue to be considered. Although there are amazing advantages of machine learning in our daily lives, it has affected thousands of people's jobs. The need for a human workforce has been eliminated to a high extent with the implementation of automated machines that are programmed to work and learn on their own. On the other hand, while the cost of setting

up machine learning and artificial intelligence models on new systems is higher, it definitely yields more benefits to industries in the long run.

Every technology has some cons along with the benefits and this is no different with machine learning and artificial intelligence technologies. The need for machine learning engineers is in high demand because of the ever-increasing data in systems. Artificial intelligence is everywhere, and it is possible that we are using it one way or the other without even knowing. Machine learning is modeled in computers and software as it allows devices to perform via cognition and work similarly to the human brain.

Today, companies are focusing on machine learning to improve business decisions and increase productivity as well. With the exponential growth of technology, we need better tools to analyze and understand the data we are receiving and also to prepare our systems for the increasing data or big data in upcoming years.

Considerations And Implications

The level of intelligence that a machine exerts is known as the moral component and it is a direct result of data that is being received. Based on the input data, machines can train themselves to work in a specific direction and even work against the interest of some humans. Scientists and researchers are still unable to remove bias from a machine learning algorithm which may produce results that are not suitable for a politically correct society. Moreover, researchers, experts and scientists believe that artificial intelligence will be harmful to society as it can be developed to mirror the human brain and all its prejudices.

The accuracy of risk assessment is another considerable aspect of machine learning algorithms. Risk assessments are used to measure and evaluate potential risks that might be involved in specific scenarios and allowing artificial intelligence to make important decisions on behalf of humans can result in more trust between humans and machines.

However, the supporters of creating transparency in artificial intelligence request for the creation of a regulated and shared database which is not in possession of a single authority that has the power to manipulate data. Transparency of algorithms will help in creating a high level of trust between users and machines as well.

Risks Associated With Machine Learning

Although machine learning models are highly accurate and make correct predictions, there are some risks associated with intelligent machines that can bring negative results to humans. Machines can surpass the highest level of intelligence to a point where they can psychologically manipulate people and become destructive to humans. Usually, people are unable to align artificial intelligence's goals with the values of humanity which allow machines to make decisions that are not feasible for human beings.

Machines that are performing multiple operations on large amounts of unrelated big data find correlations and patterns that are meaningless at best and wrong at the worst. In the long run, machines train themselves even on faulty inputs which cause them to learn to make future decisions that are not in favor of humans. Machine learning risks are real and can prove to be very dangerous if not managed properly.

Data, design and output are the three major categories of machine learning that are directly associated with risks. Bias, data, lack of model variability and output interpretation are the top risks that can result during machine learning operations.

When learning and making predictions, machine learning models can be affected with bias and make inaccurate predictions. As it mainly depends on data input, not having enough data or having faulty data can also bring enormous risks to the modeling process. Moreover, lack of model variability and output interpretation can also be a huge risk for humans if the models are not trained properly.

Avoiding Bias In Machine Learning

Self-correction is the main focus of artificial intelligence industry because researchers are finding new ways to strengthen ethics and reduce bias in rule based artificial systems. By taking human biases into consideration, there are several important aspects about ethics in artificial intelligence to work on.

Here are some of the best ways to avoid bias when designing machine learning and artificial intelligence models:

- **Selecting the right learning model:** Not all artificial intelligence models are suitable for every task. As each problem requires a specific solution and requires varying data sources, there is no proven way to avoid bias. To solve this problem, machine learning models should be designed with parameters that are most suitable to the industry where the system will be implemented. Furthermore, informing the development team in advance can also help companies and industries to avail the most suitable machine learning models.

Supervised and unsupervised machine learning models have their own advantages and disadvantages. The supervised models provide more control over bias in data selection whereas the unsupervised models that do dimensional reduction have the capability to lean bias from their data set. Data scientists need to identify and present the best model for a specific situation so that effective strategies can be implemented for long-term benefit.

- **Choosing the training data set:** Artificial intelligence and machine learning models learn from the given data inputs. So, to obtain accurate and secure predictions, data scientists are required to feed the AI models with data of different types and rule out any vulnerabilities. Training algorithms is done to avoid bias in data selection for which we have to make sure that the training data is diverse. When there is insufficient data for a group, the weighting approach can be used to increase its effectiveness on training, but this needs to be done with extreme care.

- **Monitoring performance through real data:** Discriminatory models usually work in controlled environments and no company is knowingly creating biased artificial intelligence models. Unfortunately, test groups are used to check algorithms that are already in production which result in future problems. Using statistical methods instead of real data is a better approach for monitoring performance through real data because when searching for equality, we need to examine data for equality of outcome and equality of opportunity.

  Although it is hard to prove opportunity equality, real-world testing and oversight of machine learning models can surely help researchers in developing highly accurate models and algorithms. For self-learning

systems, using biased data to feed the models can sometimes result in unintended and dangerous outcomes.

Chapter 9: Integration Of Machine Learning In Business

Businesses can benefit from any application that has the capability to organize, interpret and manipulate large amounts of data. Working on business intelligence will help companies in finding correct patterns and automate value extraction with minimum resources.

Nowadays, the production of goods or services is dependent on demand, and with lower failure rates and better predictability, businesses can operate efficiently. The amount of data in businesses is enormous and machines that have a large capacity will perform better than any human analyst.

How Should Businesses Use Machine Learning?

Before selecting a machine learning model, businesses and companies need to take an overview of some of the key factors that will provide great benefit in the future. Choosing an inappropriate artificial intelligence or machine learning model can yield negative outcomes and false predictions for your business. To avail accurate predictions and data analytics, focusing on the factors explained below will deliver great outcomes.

- **Analyze sales level and data:** Sales function has gained a lot of advantages through the development of in sales focused data. This approach allows sales teams to predict customer turnover for a new product or service and the data can also be fed into machine learning algorithms for accurate predictions. Due to the huge inflow of data, sales teams are often distracted by the time and analysis that is

required to make accurate insights. Not only does machine learning save time and effort, but the algorithms, when implemented correctly, can deliver accurate predictions and analysis by overviewing sales data.

- **Identify the ideal outcomes and define your success metrics:** Machine learning models are supposed to produce desirable outcomes for which relevant data is fed into the systems for obtaining accurate insights. Although the outcome can be different from how you assess the model and its functionality, defining your ideal outcome at first will surely help in selecting the best suitable machine learning algorithms.

  Moreover, list the metrics for success and failure for the machine learning system because failure metrics are immensely important when it comes to automating processes in a business. Most of the time, metrics are linked with ideal outcomes as specified by the customers because it is totally dependent upon predictions and input data sets. The metrics you select must be measurable because some of them cannot be measured offline. Thinking about the ideal outcome that you noticed in the previous step can also help in measuring the success or failure of a machine learning model.

  The performance of a machine learning model is not only limited to binary success or failure. Other factors such as catastrophic, about the same as before, but not as good as we expected, improvement or everything is perfect can also be considered as metrics to measure effectiveness of a system. Furthermore, keep in mind the maintenance and engineering costs for long term deployment of artificial intelligence and machine learning models. Although failure is

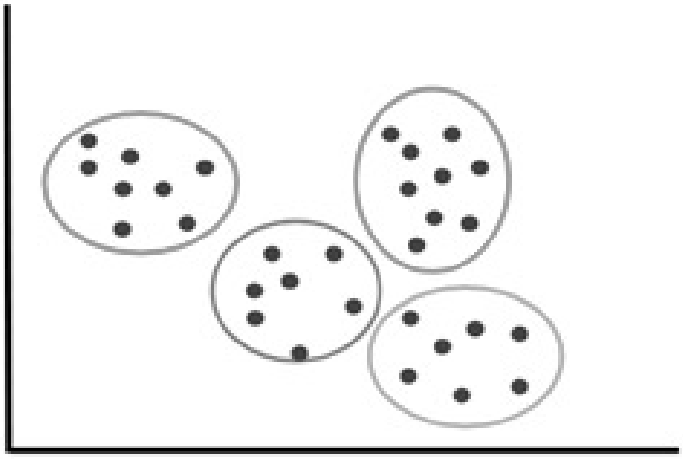
inevitable and can occur despite a successful metric, you must always remain prepared for the worst-case scenario.

- **Predictions vs. decisions:** Machine learning is best at making decisions and providing meaningful insights. Decisions are the actions which a product must take in the output of a model, and machine learning is best at developing insights and making predictions. For businesses, predictions that allow you to take useful actions can only prove to be beneficial because sometimes the prediction and decision are closely aligned.

  Taking the example of YouTube, predictions are used to generate an estimate of what video the learner wants to watch next, a fraction of a video ad the user will likely be interested in and the probability of users clicking on a search result. In the case of decisions, YouTube algorithms are designed to show videos in the recommendation bar and do not show users can add if the video is of a short length.

Machine Learning Problems

- **Clustering:** Clusters are a major part of every unsupervised learning problem. Actions for machine learning models are generally based on clustering and it can often become challenging to determine what action is needed for a specific cluster. Taking the example of Google Photos, clustering is used to group pictures of the same person together as shown in the image below:



- **Causation:** Machine learning models have the ability to identify correlations or connections between two or more things. The algorithms find it difficult to determine causation in a system and cannot interpret why a specific event happened. Moreover, we cannot determine causation with the help of observational data without making an experiment. Without the availability of relevant data, we can never train a model to perfection. It should be remembered that rule-based, simple and heuristic systems only have the capability to train without data.

  Once you have selected the training data, the next step is to find patterns. In the case that there are no patterns found or only trivial patterns are available, machine learning will not be able to deliver accurate predictions or values. Generally, training data sets have a high number of patterns which allow machine learning models to follow the right approach for making accurate predictions.

- **Anomaly detection:** Machine learning models are best suitable for identifying anomalies. This approach is dependent on how you decide what constitutes an anomaly. To achieve labeled data to solve the

issue, defining a heuristic to label anomalies is recommended. However, once you are done defining the heuristic, you can also use the heuristic for the production system because machine learning models can never defeat the heuristic on which they are trained. For complicated heuristics, replacing the system with machine learning models can deliver many benefits.

Data Preparation In Machine Learning

Machine learning provides different models and algorithms that help us in finding data patterns. To make the predictions correct and feasible, we must design the data set and transform data in the best possible manner. At first, find the best quality and suitable size of data that is supposed to be used for updating the algorithm. The process for data preparation in machine learning is explained as follows:

1. **Define a problem and propose a relevant solution:** For your business requirements, take an overview of the requirements of the functional process and then list out the models that can make accurate predictions for the data set. Selecting the wrong machine learning model will not only yield inaccurate predictions, but it can also disturb the operational system in the near future.
2. **Construct data set:** After you are done with selecting the machine learning model, the next step is to construct your data set and transform

the data. In this phase, you will have to collect raw data and identify the feature and label sources as well. Moreover, choose a sampling strategy and split the data into different categories so that high quality training data set can be prepared. After you are done with constructing the data set, the next step is to transform data to bring feature engineering into your machine learning model. This approach also helps in exploring and cleaning the training data.

Simple models on large data sets can beat complex models on small data sets for which you need to determine the quality of training data. A quality data set lets you succeed in solving the business problem for which the machine learning model is being designed because focusing on reliability, feature representation and minimizing skew for training data is mandatory. Reliability refers to the extent your data is trustworthy and a model that is trained on reliable data is more likely to deliver positive predictions as compared to a model that is trained on unreliable data.

Furthermore, business owners must also understand the importance of feature representation and how it directly influences business processes. Feature representation includes various factors such as how data is shown to the model, methods to handle outliers and normalizing numeric values in training data.

3. **Train a model and use it for predictions:** In this step of model development, you can train the machine learning model according to your business needs. Repeatedly insert the training data you have collected through the previous steps and analyze the performance of the model by creating a chart. This will ultimately help you in getting accurate predictions and positive outcomes in the future. Upon the

completion of training, your model must be powerful enough to make accurate predictions for every scenario in your business.

Remember that the training process is not always supposed to be sequential, because in some cases, you will have to split your data after transformation. Furthermore, extra data also needs to be collected if your model is not making accurate predictions. Even after the training is started, you can modify the feature set and apply other rules or linear regression and neural nets to achieve the perfect outcome.

Major Challenges In Machine Learning

Engineers and data scientists have greatly revolutionized business processes with the help of machine learning and artificial intelligence models. Deep learning algorithms have greatly helped in the development of major business projects such as driverless cars, automatic disease detection systems, image recognition, natural language processing, independent human robots and predictive maintenance in industries. Most of the companies face the challenge of educating customers regarding the applications of their innovative technology which requires a lot of hard work and is also time consuming.

Commercial implementation of artificial intelligence and deep learning is still in the development phase and there are a lot of challenges that need to be handled in future. These systems require huge sets of properly organized and prepared data sets to yield accurate answers or predictions. Generally, an artificial neural network is based on millions of parameters and is capable of remembering training sets to provide answers with 100 percent accuracy.

Other challenges are included in the list below.

- **Availability of data and experts:** Although artificial intelligence models and deep learning networks are great performers, the approach is still in the development phase with a limited number of researchers working with this technology.

  In order to train a machine learning model for greatest accuracy, big sets of data are required. Finding and storing suitable data can result in increased expenses for the company because ready sets of training data are costly and hard to find. Companies who have to prepare data for algorithm training need to focus on major factors such as classification, ranking, regression and clustering for which machine learning experts are required.

  There are different types of data collection mechanisms which help to reduce data with record sampling, attribute sampling and aggregating. Data scientists are required to decompose the data and resell it to meet the business requirements of a specific company. Even if you have unlimited storage space in your system, the process is expensive and requires a lot of hard work as well.

  Nowadays, people are more concerned about privacy and hesitate to share their personal information and data online. This is another major challenge that needs to be handled when creating data sets. Big data and personal data activities have now become more costly, risky and difficult due to the implementation of rules for the protection of personal data. Even if you are able to develop a proper training data set, hiring the best suitable data scientist to train your business model will remain expensive in each case.

- **Memory networks and natural language processing:** In machine learning models, memory networks require large working memories to

store data which is a type of neural network that needs to be combined with a memory block to read and write the network.

For attaining fully effective and efficient artificial intelligence, we need to overcome this challenge and find a better approach for networks to find relevant data and store it for access when needed. Moreover, natural language processing is yet another major challenge for machine learning algorithms that takes a lot of effort and money to be implemented in the long run.

An understanding of language is a major factor of deep networks for which computers and machine learning models are supposed to be designed for delivering accurate predictions. Machine learning development has several layers and engineers need to write an algorithm that will write more programs on its own.

- **Development of machine learning models:** As compared to traditional development techniques, machine learning and artificial intelligence model programming is quite different and complex. There are more uncertainties because machine learning engineers are not aware of how deep learning networks will respond to different types of data sets. Moreover, the behavior and performance of a model is absolutely dependent on how well it is trained. This makes it difficult for machine learning engineers and data scientists to guarantee replication of the training process in a model.

  A wrong problem formulation can lead to unexpected outcomes, and before attempting any approach, the engineers have to think carefully about whether the new change will yield positive results or not. Performing data engineering and feature extraction are recommended

approaches when designing a machine learning model because only proper parameters can yield accurate predictions in the future.

- **Model retraining:** Model retraining is one of the biggest challenges that is being faced by industries and businesses that have implemented machine learning and artificial intelligence models in their systems. This process is challenging because training is done offline whereas the models are loaded at prediction or inference time in a serving system. In the event of the occurrence of new events for which the models are not trained, the process of model retraining needs to be repeated. Deciding when to train a model can be challenging because detecting change in data is critical and will result in future problems if not done properly.

  Standard cross validation techniques help in estimating the generalization error and tune precision thresholds for missing value imputation models. In the presence of training data, it is mandatory that some predictions are audited by humans which can make it easier to tackle future model management challenges. Furthermore, to achieve any sort of large-scale data processing, we need powerful CPUs that are capable of handling tons of data transactions per second without interruption.

- **Data selection:** Machine learning models training is highly dependent upon the quality and reliability of data that is being fed into the system. So, generally, we use supervised learning to train models. Along with data, the supervised learning also requires labels and information about the data which can sometimes become challenging to find. The availability of labeled data is difficult and also restricts the performance of machine learning models in some cases.

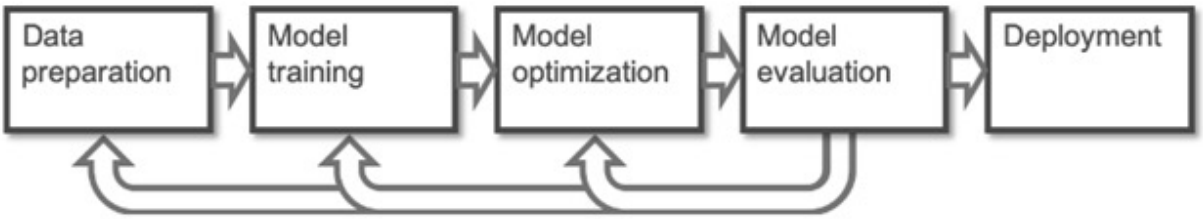
Furthermore, labeled data does not naturally occur, and it can be created with quantitative data as well. In the case of an unsupervised learning algorithm, there are no suitable answers given in the training data set and algorithms are made to find relevant solutions on their own for making accurate predictions. As compared to supervised learning, unsupervised learning problems are difficult to solve because they require more data to achieve the desired solution.

Overcoming The Challenges Of Machine Learning Implementation

The complexity of machine learning models and deep learning algorithms can be solved by considering hyperparameters and automated machine learning theories. Traditional methods of developing a machine learning model is time consuming and requires a lot of trial and error. In order to overcome the circumstances, auto-machine learning algorithms are implemented in systems which provide better data analytics and accurate predictions.

Automated machine learning is the process of iterating through an exploration of space which also helps in selecting the best model. The results are dependent on training scores and the exploration of space is generally comprised of hyperparameters selection, feature extraction and other major AI algorithms. A robust implementation of auto-machine learning systems is difficult and requires a lot of effort.

Here are the main parts of machine learning model designing:



Data management plays a vital role in the development of machine learning algorithms. As it is capable of data preparation and ingestion, it is considered as the most demanding relation for an infrastructure. In particular, machine learning can also help increase efficiencies by optimizing the path the query takes to approach data. Moreover, it also supports confidence-based querying that provides answers with the most predicted accuracy.

The challenges that are related to data access and preparation are directly affected by the selection of a right data management environment. As the connectivity is created by data virtualization, it also enables access from multiple repositories with different data types. Although data management solutions exceed budget limitations for small scale businesses, they surely deliver better returns on investment and improved efficiencies.

Machine learning models are built using major programming languages like Python, Java, R, C and SQL. Depending upon the prediction and data analytics requirements, artificial intelligence algorithms and machine learning models are combined to improve business processes and bring long-term benefits as well. Moreover, machine learning algorithms also provide a platform to predict proactive measures and help in avoiding possible failures in the production department.

Ethical Issues With Artificial Intelligence

Artificial intelligence has somehow affected the hierarchy of labor because it is primarily concerned with automation. With more automated processes and self-learning systems, the need for a human workforce is reducing with time. This results in the downsizing of companies because automation has the capability to control processes and yield better outcomes in most of the cases. Although it is not possible to completely hand over the processes and routine tasks to machine learning models, there is a higher percentage of tasks that are now automated which leads to a lesser need of humans in every industry.

Taking the example of self-driving vehicles, millions of individuals in the United States are working in truck driving industry and the introduction of self-driving cars would affect their jobs. Although self-driving vehicles are safe and have a lower risk of accidents, this type of artificial intelligence would greatly affect the workforce in underdeveloped countries.

Further, artificially intelligent bots are now becoming better at modeling human relationships and conversions as they work through natural language processing algorithms. This will eventually lead to the start of an age where we will be frequently interacting with machines instead of humans to avail different types of services.

Humans have the power to express emotions and gratitude for another person, and in case of an AI powered robot, this might not be possible, leading to major problems such as a communication gap and/or racism. With accurate training data, machine learning algorithms can bring positive

predictions and work in ways that will not affect humans or their employment.

Chapter 10: Artificial Intelligence And Real World

Artificial intelligence systems and cognitive technologies have greatly transformed the way we used to perform routine tasks such as online shopping, financial transactions or transportation. Companies who are using cognitive technologies are likely to achieve a major transform within three years because artificial intelligence systems not only enhance business processes, they also provide better data analytics and predictions for the future.

Types Of Artificial Intelligence

Process Automation

It is beneficial for companies to give an overview of artificial intelligence and their business capabilities instead of technologies. Artificial intelligence has the capability to support businesses by introducing automated business processes, engaging with customers and employees and providing insights through data analysis. Automation of physical and digital tasks is the most used type of artificial intelligence because it helps companies to improve both administrative and financial activities.

RPA is an advanced business process automation tool which has the capability to perform multiple IT system tasks without interruption. Mainly, these tasks include transferring data from call center systems and email into a system of records. Other tasks, such as updating customer files, service additions and reaching multiple systems to update records, can be performed through process automation. Moreover, reading contractual and legal

documents to extract provisions is now possible with natural language processing methods. Developers are working to bring more intelligence and learning capabilities in cognitive technologies for backend systems as well. RPA is easy to implement, and costs less compared to other artificial intelligence models as it brings the highest return on investment.

Cognitive Insight

Cognitive insight is yet another popular artificial intelligence approach for detecting patterns in huge volumes of data and interpreting their meaning to avail accurate insights. Machine learning applications with cognitive insight are used to identify credit fraud in real time and detect insurance claims fraud as well. Moreover, the algorithm has the capability to predict customer buying behaviors and analyze warranty data to identify quality problems or safety concerns in manufactured products.

In machine learning, cognitive insights are different from traditional analytics. They are data sensitive and more detailed because they are trained on a specific training data set. Furthermore, cognitive insights have the ability to use new data to make accurate predictions and manage categories as well.

Cognitive insight applications are generally used to improve performance in jobs that are only suitable for machines. For example, programmatic ad buying, which is based on high speed data crunching and automation, done by a human being is not possible.

Different versions of machine learning and deep learning have the capability to recognize patterns and perform tasks such as recognizing speech and images. Although the activity of data curation is quite labor intensive,

machine learning has now introduced state of the art algorithms and models that are capable of learning and identifying probabilistic matches as well.

Cognitive Engagement

The cognitive engagement approach is used by companies to communicate with employees for getting updates regarding day to day activities through artificial intelligence models. By using natural language processing, intelligent agents and bots communicate with both employees and customers to address issues and also provide technical support. All of these processes are done in understandable language, including product and service recommendations. Cognitive engagement also covers internal sites for answering employee questions, implementation of HR policies and is responsible for managing employee benefits as well.

The growing number of employee and customer interactions without requiring staff can be managed through artificial intelligence cognitive engagement models. Nowadays, companies follow a conservative approach while developing customer facing cognitive engagement technologies to achieve the best outcomes. Although cognitive engagement is a beneficial approach for businesses to communicate with their employees and customers, they are not able to answer each query, so human intervention is necessary. As a result, bot-based interfaces are now limited to specific topics or certain domains.

Understanding The Technologies

Each business and company does not need to understand every artificial intelligence model. The importance of research and evaluation before selecting an artificial intelligence model depends on the specific business.

Take the example of robotic process automation and rule based expert systems. They are transparent in how they work but both of them do not have the capability to improve and learn. On the other hand, deep learning is best at learning from large volumes of labeled data but finding out how it actually creates the models is difficult. These problems are also known as black box issues and might not be suitable for highly regulated industries like banks and financial institutions.

Long-term artificial intelligence projects require in-house data scientists and machine learning experts so that more progress can be achieved. At first, companies need to understand the performance and functionality of different technologies and choose the one that best addresses their issues. Acquiring knowledge for machine learning and artificial intelligence systems need ongoing research and interest as well. Moreover, businesses need to leverage the capabilities of employees, such as data scientists, who are well aware of big data and the statistical needs of the company.

Creating A Portfolio And Determining Use Cases Of Projects

Artificial intelligence programs are designed after evaluating the needs and requirements of a business. At first, companies need to identify opportunities and perform assessments regarding which areas of the business can achieve the most gain out of the cognitive AI applications. Insights derived from knowledge and data analysis can greatly help in the development of a portfolio needed to design a machine learning model.

In most of the cases, knowledge exists in the organization, but it cannot be distributed evenly due to the lack of cognitive insights. This situation is also

considered as a bottleneck point which hurdles in distribution of data and stops knowledge from reaching AI and machine learning models.

On the other hand, knowledge exists within organizations but the processes that are responsible for data transfer and management take more time than expected. This issue is also known as a scaling challenge for businesses and organizations.

Use cases are another essential area of assessment for cognitive applications. Through use cases, the substantial value and contribution to a business can be measured. By determining the overall strategy and approach to address a target problem, we can launch artificial intelligence models by prioritizing use cases that yield both short-term and long-term benefits to the company. Furthermore, cognitive capabilities are also used to create a competitive advantage to suit the business perspectives of a company that is purely based on artificial intelligence and machine learning systems.

Moreover, we need to examine whether the artificial intelligence tools for each use case are meeting business requirements or not. Taking the example of intelligent agents and chatbots, we can notice that they are still not able to meet human problem-solving capabilities, even with the implementation of scripted cases. Although they are improving with time, the process is slow and still requires a lot of development. But, in time, cognitive technologies will be able to transform the process of how companies do business.

Redesigning The Business Process

With the development of cognitive technologies, business workflows and processes are meant to be changed over time. The main focus of cognitive technologies is increasing the percentage of decisions made by machines as compared to humans in a predefined business model. To make this happen, systematic redesign of workflows is necessary to make sure that humans and machines both utilize each other's strengths to bring long-term benefits to the organization.

Nowadays, most companies have successfully deployed cognitive pilots in their system architecture to achieve their goals and annual targets. To make this happen, organizations require a complete plan and research to scale up for which deep understanding between owners of business process and technology experts is mandatory. Cognitive technologies support individual tasks rather than entire processes for which the existing systems need to be scaled up with latest machine learning processes.

In the case that the application is dependent on a special technology that is difficult to implement, discussing scaling considerations during the pilot phase with the IT team will prove to be of great benefit. Furthermore, simple technologies, such as RPA, are the best types of implementation for artificial intelligence and machine learning models that work in collaboration to improve business processes. If the scale up is performed to achieve the

desired results, companies should also focus on improving productivity and delivering their services in a better way.

The Future Of Cognitive Technologies

Machine learning and artificial intelligence have now transformed the way companies work and deliver their services. Companies that are adopting artificial intelligence in moderation surely have extensive plans for the future and will be able to make a strong market position in the upcoming years.

Information sensitive domains, like financial services, marketing, professional services and education, can obtain long-term benefits from cognitive technologies. Not only does it automate routine tasks in a system, but it also reduces the need for a human workforce while delivering accurate predictions and data analytics.

Cognitive technologies also serve as a catalyst for the success of other data intensive technologies such as the Internet of Things, autonomous vehicles, multi-channel consumer technologies and smart phones.

It is feared that mass implementation of cognitive technologies will lead to the unemployment of millions of humans. Although the system is capable of performing smart tasks that were traditionally done by humans, it is still not able to accomplish entire jobs. Most of the cognitive tasks nowadays are managed by humans and technology is only able to work for basic artificial intelligence approaches such as big data analytics.

The Future Of Artificial Intelligence

The future of artificial intelligence and machine learning is about less data, and the models will be designed to become less artificial and more intelligent. This ability will cause artificial intelligence to be broadly applied and will also create new opportunities for adopters in different ways.

Nowadays, deep learning and machine learning systems are built by training them on tons of data so that they are able to make actions in any possible situation. By relying less on big data, machine learning models will be designed to work through logical reasoning so that the cost and effort to obtain training data can be reduced.

Again, taking the example of a driverless car, we can notice that the vehicles are trained to drive perfectly in almost any traffic situation for which the neural networks are fed with intensive data. A driverless car can handle any situation or hurdle while considering the safety of pedestrians and other vehicles in its immediate surroundings.

Systems that are dependent on data also have to face ethical and business constraints. But not all businesses and companies have huge data sets to train models which is necessary for neural networks to perform with best accuracy.

In the future, we will have top down systems that are flexible and do not require a huge amount of training data. To make this happen, we must focus on efficient robot reasoning, which is a concept that allows machines to understand every situation in the way humans do. Consider the example of the CAPTCHA code. They are quite easy for humans to solve but cannot be evaluated by computers or robots. Furthermore, we need to create models with ready expertise that will allow systems to do exactly what humans would in cases of high uncertainty and complex situations.

Currently, organizations are struggling to teach machines to navigate the world by utilizing basic principles of human common sense. The introduction of common sense in the machine learning approach will allow models to understand everyday actions and objects, handle unforeseen situations, learn from experiences and communicate naturally.

Although machines can never reach the intelligence and intellect of a human brain, they can definitely be trained with explicit training data to make vital decisions as humans would. By helping machines to understand the importance of common sense, researchers are working to develop a system that is capable of untangling ambiguities in natural language and mimicking core domains of human cognition.

How Can Artificial Intelligence Become An Everyday Technology?

The evolution of artificial intelligence and machine learning started in the early 1950's and is still receiving major breakthroughs and improvements to help in automating services and tasks. In recent years, artificial intelligence has undergone a remarkable refinement and has brought a wide range of resources, services, practices and products for the benefit of human beings. We are now entering an age of deployed artificial intelligence as most of the business processes and systems are already working on artificial intelligence and machine learning models.

It is hard to imagine how accessible and powerful machine learning tools are because they are trained to solve real life problems without bringing damage to the environment or to humans. To make artificial intelligence an everyday technology, we need to work on producing training data that delivers more consistency and yields better results when fed into machine learning models.

Among many other challenges, misleading input is one major problem that leads to AI systems misbehaving and yielding negative outcomes.

Technology transitions form an elite niche to a mainstream tool are affected by AI models and training data. Intelligent algorithms are designed after years of research on neural networks and have the capability to solve complex everyday problems without any hassle. Although machine learning models are suitable for each of the everyday tasks, it still has some accessibility limitations that are yet to be accomplished.

At a time when machines will become as intelligent as humans, there will be several factors for humans to worry about. Although artificial intelligence algorithms receive constant development and improvements so that they never bring harm to humans, an excess of automation can result in harmful consequences for humans as well.

It is thought that if artificial intelligence is able to reach human intelligence level, it might become difficult for humans to interpret AI models and gain control over the insights and actions they are performing. As a result, we will never be able to comprehend things that a super intelligent AI model does even if the machine is consistently trying to explain the situation.

Artificial Intelligence Transforming The World

The world as we know it is quickly evolving with artificial intelligence and is changing the way we live and perform everyday tasks. For most people, the definition of artificial intelligence still remains unclear and they find the models to be unsafe and unsuitable for human beings.

Over the years, researchers and scientists have been performing extensive research and evaluation to develop safe machine learning models for human

beings. The implementation of artificial intelligence in security frameworks, organizations, government departments, energy industries and natural resource management departments is increasing rapidly because of the accurate predictions and data analytics machine learning models provide.

Most of artificial intelligence developers are ultimately directed towards achieving a basic goal as they are supposed to develop AI models that are capable of reducing human effort to complete a specific process. In some cases, human labor is held responsible for inefficiency, inaccuracy and other failures which result in great losses for a company. When trained properly, machine learning models will always work in the best interest of the organization and dutifully perform each assigned task.

Artificial intelligence has the power to change the perspectives in our lives and how we accomplish everyday tasks. With the help of artificial intelligence, we are able to remove mundane tasks from our lives as AI assistance has the capability to schedule appointments, achieve a deadline and perform routine tasks without any problems. However, artificial intelligence will only function as programmed and work on multiple assignments depending upon the training data it is trained on.

Machine learning and artificial intelligence are known for helping businesses in overcoming challenges in a more productive way. The technology is able to make use of innovative diffusions and adopt proactive decision implementations to improve the overall work procedure for an organization. Generally, most of the artificial intelligence modules are connected with other frameworks such as big data, cloud database, block chain, Internet of Things and cryptography. Furthermore, the technologies also allow companies to benefit from the effectiveness of information transmission and reproducibility from each source.

Artificial intelligence is transforming tasks and operations that are being performed all over the world. However, the possibility of an inability of humans to control accompanying negative effects of artificial intelligence is an alarming situation for humans. Researchers and developers who are working on AI technologies focus on the limitations of machine learning and artificial intelligence models so that they do not bring any type of harm to humans in the near future. Irrespective of the negatives of AI, the technology is widely accepted because of its amazing services and benefits.

Chapter 11: Managing Big Data, Artificial Intelligence And Machine Learning

In the present era, data science has allowed artificial intelligence to find appropriate and meaningful information in an efficient way. Artificial intelligence, big data and machine learning work in collaboration to design state of the art models and algorithms for predictions and automated processes. Although these processes are complex, the effectiveness and results they offer have great benefits for today's world.

Big Data Services

Companies that are planning to implement big data services need robust IT computer systems and expert artificial intelligence programmers to train algorithms. Generally, big data engineers can help in setting up the machine learning and big data models on large mainframe systems. Dedicated or colocation servers are required to support the processes of analysis of big data and this solution is generally managed off-site. Data mining of big data is usually achieved through artificial intelligence programming that works along with other machine learning algorithms to find suitable patterns in big data.

Furthermore, it also makes it easier to make informed decisions and it is a cost-effective approach for companies to outsource IT consultant tasks for big data analytics development. Moreover, big data specialists can also be arranged on an "as needed" basis for designing, developing and updating the artificial intelligence systems. Support services for big data operations can also be availed through IP transit and managed hosting.

To manage big data effectively, large data centers are required that are powered with high-end processing systems. Business critical hardware is required to have at least triple redundancy for achieving 99.9 percent uptime performance. This is because IT structures are designed by overviewing the risk of mishaps and load issues. To balance load in real time, network servers have to reroute the processing to servers that remain operating in the network so that none of the processes are affected.

Big Data Storage Handling

Data storage requirements for big data are increasing with time and the approach to handle operations in the system is meant to be effective and reliable. Generally, localized data is captured and processed in big data storage centers or an extensive storage system that is maintained in the cloud.

The operations can also be managed through a virtualized data system which creates a virtual layer of data. This virtual layer is responsible for determining the location where data is stored on the network and whenever any changes or calculations are performed on an AI system, virtual layer only extracts the required data and the original storage data is never affected in any way. To complete big data management tasks, network-wide data management protocols also need to be implemented as they deliver better computational processing speeds and reduce the need for data storage memory as well.

Artificial Intelligence And Big Data

Artificial intelligence and big data are widely being used to make predictions and perform data insights to complete certain tasks. In big data

systems, artificial intelligence is used to detect anomalies and unusual occurrences in the data set through different parameters and sensors. Any node that is not meeting the rules defined by AI models is immediately removed so that it does not cause a potential problem in the future. Moreover, probabilities of future outcomes can be analyzed by artificial intelligence through the Bayes theorem as briefed in the earlier parts of this book.

Recognizing patterns in big data is done through artificial intelligence models because AI can analyze big data by searching through patterns in graphs and bars that are made from underlying data sets. Machine learning builds the connection between artificial intelligence and data science because machine learning itself is based on the process of learning from data over time.

Big Data's Role In Artificial Intelligence

Artificial intelligence models are absolutely dependent on training and we can increase the performance and effectiveness of an AI system by using appropriate training data sets. Generally, big data is considered to feed processors of artificial intelligence systems through which machine learning algorithms can learn how to make certain decisions and reproduce a certain behavior. Although artificial intelligence does not deduce conclusions the way humans do, they have the capability of learning through trial and error for which massive amount of training data is needed.

The more data that is given to AI systems, the better it will perform and yield more accurate outcomes. Usually, data science gathers data from different sources and applies machine learning, artificial intelligence, sentiment analysis and predictive analysis rules to extract critical information from collected data sets. This process is complex and needs a lot of focus in terms

of business point of view because only accurate predictions and insights can help in making better business decisions.

Machine Learning And Data Science

Data science is based on the basic principles of data mining, data analytics and artificial intelligence. A data scientist is supposed to forecast the future based on the past patterns and have the model extract meaningful insights from various sources. For becoming a data scientist, having knowledge of mathematical statistics, a fluent understanding of R and Python, learning data wrangling methods and a deep understanding of PIG/HIV approaches can prove to be of great help.

Data science involves the processes of data extraction, data cleansing, analysis, visualization and actionable insights generations. Machine learning also has a different perspective on statistics and you can gain expertise in the domain by focusing on the basics of probability and statistics, data modeling, computer programming fundamentals and essential evaluation skills. Furthermore, a data scientist is supposed to gather, process and derive valuable insights from data for which different experiments and machine learning methods can be implemented to make better business decisions.

We must remember that machine learning is a part of data science and it draws aspects from algorithms and statistics to work on the generated data. Data in machine learning models are present in enormous amounts and data scientists have to work on making a system capable of learning and predicting. Machine learning is also an all-encompassing term that includes aspects of artificial intelligence as well.

Handling Machine Learning Projects

A typical machine learning workflow is based on simple steps comprised of managing data, training models, evaluating models, deploying models, making predictions and monitoring those predictions. With the help of online inferencing and prediction, applications can access centralized data centers and use standard algorithms for training machine learning models. Once the development of model is complete, it is evaluated for accuracy which is dependent on different parameters and metrics. Moreover, each model can be deployed for online and offline predictions depending upon the requirements.

Failure Of Machine Learning Models

Machine learning models will only deliver positive predictions and evaluations if they are trained on a suitable data. As long as you have loads of training data, you can make the model perform as per your requirements. This is one major issue that often leads to the failure of machine learning models when a company needs to retrain a model for new functionalities. To manage things in a better way, computer vision approach can be used to perform multiple operations from a machine learning model.

Best Practices To Deal With Data Training Problems

Data set training is an essential part of machine learning algorithms and model development. Start with the simplest task you have and test the limitations of the model before adding further complexity. This will eliminate the chances of false predictions in the future and will make your machine learning model more reliable.

Nearly all machine learning and artificial intelligence models fail and need to be handled with determination. In batch processes, you can focus on building human in loop systems that deliver low confidence predictions to an operator for making the system work reliably.

This process will also help in collecting high quality training data and, with other use cases, you can present low confident predictions to flag potential errors for end users. As we are increasingly relying on machine learning systems, the field has now become a complete engineering discipline and has provided unlimited opportunities for companies to manage their routine system operations in a better way.

Testing And Evaluation

Machine learning models can be tested and evaluated to measure performance, reliability and effectiveness. For checking a model on test distribution, it is mandatory that you understand the differences between set distributions and training. Moreover, revisit model evaluation metrics to ensure that the model is behaving properly. Writing tests for input data pipeline, model inference functionality, model inference performance on validation data and explicit scenarios expected in production can also help in examining the overall performance and reliability of a machine learning model.

For model deployment, we can use REST API and deploy the new model to a small subset of users to make sure that each process is done accurately. Furthermore, machine learning engineers must maintain the ability to roll back to previous versions of model development and monitor live model prediction and data analytics.

Last, but not least, periodic maintenance of machine learning and artificial intelligence models helps in retaining their performance, effectiveness and reliability. Changes can affect AI models in unexpected ways and the models must be maintained regularly to obtain accurate predictions and data insights.

Advanced Artificial Intelligence

The next generation of artificial intelligence is said to have advanced computational powers and might also involve human intelligence. The artificial general intelligence (AGI) systems will have the capability to learn, adapt, self-improve and solve problems in the best interest of human beings. Furthermore, they will even perform tasks for which they are not programmed.

The introduction of AGI systems will also bring artificial super intelligence (ASI) in the near future. Fully functional AGI systems are expected to be launched after 2029 for which research and development are being performed at this moment.

However, there are certain risk factors involved with AGI systems that could transform humanity in different ways. To handle problem solving at a global level, companies and businesses need to deploy intelligent systems that can effectively learn from training data and deliver predictions that do not harm human beings in any way. Human factors and ergonomics are also associated with AGI systems for which new strategies for achieving goals need to be developed.

Controlling AGI Systems is tricky and we need to implement high end techniques and approaches to control the performance and prediction powers of AGI systems. Without proper testing and development, introducing AGI

systems at a global level could lead to negative circumstances as well. That being said, when designed, developed and tested with 100 percent accuracy, AGI systems will surely bring positive changes for humans and organizations.

Conclusion

“Machine Learning For Beginners: The Ultimate Guide to Understand Artificial Intelligence and Big Data Analytics. Learn the Building Block Algorithms and the Machine Learning’s Application in the Modern Life” is a comprehensive and detailed book covering all of the basic and major findings of artificial intelligence. With the help of comprehensive explanations and briefings on each aspect of the machine learning approach, the concepts are delivered in an easy to understand manner so that beginners can learn the basics of artificial intelligence and its impacts on our daily lives.

Artificial intelligence has been transforming the technology industry from 1950’s and is still receiving major upgrades and developments for the benefit of humans. Machine learning models and algorithms have the capability to bring positive changes in our lives as they make accurate predictions and automate everyday processes without yielding many negative consequences.

By learning the concepts as explained in this book, beginners can develop a strong understanding and interest in the field of machine learning, big data analytics and artificial intelligence.

Over the years, artificial intelligence and machine learning have defined new pathways for researchers and engineers for designing technologies that are in the best interest of human beings. Furthermore, the scope, deployment and scale of artificial intelligence and machine learning will help us in bringing positive changes in our daily lives and making organizations more profitable.

References

Big data analytics. Retrieved from https://searchbusinessanalytics.techtarget.com/definition/big-data-analytics

Brownlee, J. (2018). What is Machine Learning? Retrieved from https://machinelearningmastery.com/what-is-machine-learning

Carrigan, A. (2019). Machine Learning for Business. Retrieved from https://medium.com/mindsdb/machine-learning-for-business-c24cb198dbb5

Marr, B. (2016). What is The Difference Between Artificial Intelligence And Machine Learning? Retrieved from https://www.forbes.com/sites/bernardmarr/2016/12/06/what-is-the-difference-between-artificial-intelligence-and-machine-learning/#5e296ac32742

Nicholson, C. (n.d.). A Beginner's Guide to Neural Networks and Deep Learning. Retrieved from https://skymind.ai/wiki/neural-network

Ray, S. (2017.) Commonly used Machine Learning Algorithms. Retrieved from https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

Vigo, J. (2019). Why Machine Learning Is The Future Of Business Culture. Retrieved from https://www.forbes.com/sites/julianvigo/2019/06/06/why-machine-learning-is-the-future-of-business-culture/#2c3697b2bab7

Yiu, T. (2019). Understanding Neural Networks. Retrieved from https://towardsdatascience.com/understanding-neural-networks-19020b758230