Olivier Maugain

Using Data

Course Notes

365√DataScience

Table of Contents

ABSTRACT

The introduction of Big Data and artificial intelligence (AI) has opened a whole new domain of opportunities for businesses to collect, manipulate and interpret data. Therefore, to become fully data literate means to understand the role of machine learning and AI in the modern data ecosystem.

The previous two data literacy course notes sections outlined the characteristics of a data literate person, as well as important data terminology and the different methods of data storage. The next logical territory to cover is the different methodologies for practical data usage.

In this section we are going to discuss topics such as

- The difference between Analysis and Analytics
- Business Intelligence
- The difference between Machine Learning and AI
- Real-life applications of machine learning and AI
- The types of machine learning

Keywords: data, data literacy, machine learning, data analysis, deep learning

1. Using Data

1.1. Analysis vs. Analytics

Definition:

"*Analysis is a detailed examination of anything complex in order to understand its nature or to determine its essential features.*"

- Data analysis is the in-depth study of all the components of a given data set
- Analysis is about looking backward to understand the reasons behind a phenomenon
- It involves the dissection of a data set and the examination of all parts individually and their relationship between one another
- The ultimate purpose of analysis is to extract useful information from the data (discovery of trends, patterns, or anomalies)
- It involves the detailed review of current or historical facts
- The data being analyzed describe things that already happened in the past

Examples:

- Comparison of the sales performances across regions or products
- Measurement of the effectiveness of marketing campaigns
- Assessment of risks (in finance, medicine, etc.)

Definition:

"Analytics is a broader term covering the complete management of data."

- It encompasses not only the examination of data, but also their collection, organization, and storage, as well as the methods and tools employed
- Data analytics implies a more systematic and scientific approach to working with data throughout their life cycle, which includes: data acquisition, data filtering, data extraction, data aggregation, data validation, data cleansing, data analysis, data visualization, etc.
- Analytics allows to make predictions about the future or about objects or events not covered in the data set

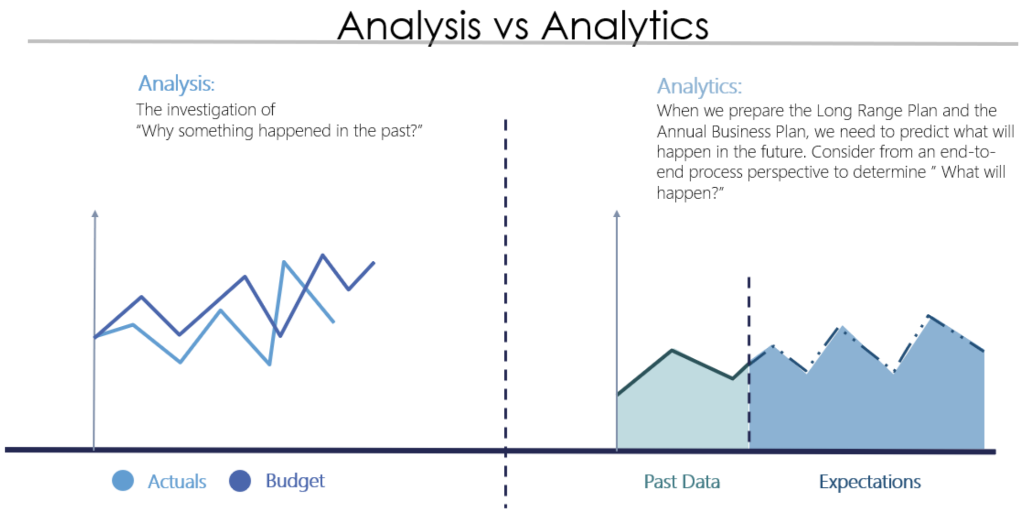
Examples:

- The semi-automatic or automatic analysis of data
- The extraction of previously unknown patterns
- The grouping of objects
- The detection of anomalies
- The discovery of dependence

Analysis vs. Analytics – differences:

- Analysis only looks at the past, whereas analytics also tries to predict the future
- Analysis can be performed with any spreadsheet software, in particular Microsoft Excel.

- For Analytics, you need an analytics tool, such as Python



Analysis vs Analytics

Analysis:
The investigation of
"Why something happened in the past?"

Analytics:
When we prepare the Long Range Plan and the Annual Business Plan, we need to predict what will happen in the future. Consider from an end-to-end process perspective to determine " What will happen?"

1.2. Statistics

Definition:

"Statistics concerns the collection, organization, analysis, interpretation, and presentation of data."

" Population is the entire pool from which a statistical sample is drawn."

Well-known software tools for statistical analysis: IBM SPSS Statistics, SAS, STATA, and R.

Types of statistics:

- Descriptive statistics
- Inferential statistics

1.2.a Descriptive Statistics:

- Summarizes (describes) a collection of data

- Reveals useful or at least interesting things about the data, in particular "what is going in there".
- Presents and communicates the results of experiments and analyses.

Examples of questions to be answered through descriptive statistics:

- How many people in Georgia voted for Joe Biden during the US Presidential Election in 2020?
- What is the proportion of favorable opinions for Brand X among adults aged 18 to 34?
- What is the average salary of brand managers?

1.2.b Inferential Statistics:

- Helps us draw conclusions (inferences) about things that we do not fully grasp
- Helps us understand more about populations, phenomena, and systems from sample data (induction)
- Enables professionals to predict what is likely to happen if we make a specific decision or act on the entity in this or that manner

Examples of questions to be answered through inferential statistics:

- Is there a correlation between age and customer satisfaction?
- Does this new packaging significantly increase the sales volume of this product?
- In this country, is the brand preference of women the same as that of men?

Computational statistics (statistical computing):

- A combination between computer science and statistics
- The focus lies on computer-intensive statistical methods, e.g., a very large sample size

Practical applications of statistical computing:

- Econometrics
- Operations research
- Monte Carlo simulation

1.3. Business Intelligence (BI)

Definition:

"Business intelligence (BI) comprises the strategies and technologies used by enterprises for the data analysis of business information."

Practical Applications:

- Operating decisions, e.g., the allocation of marketing budget to different campaigns, brand positioning, pricing
- Strategic decisions, e.g., international expansion, opening or closing of plants, priority setting
- Enterprise reporting – the regular creation and distribution of reports describing business performance

- Dashboarding – the provision of displays that visually track key performance indicators (KPIs) relevant to a particular objective or business process
- Online analytical processing (OLAP) – querying information from multiple database systems at the same time, e.g., the multi-dimensional analysis of data
- Financial Planning and Analysis – a set of activities that support an organization's financial health, e.g., financial planning, forecasting, and budgeting

Vendors providing visualization or business intelligence software: MicroStrategy, Microsoft (with Power BI), Qlik, or Tableau.

1.4. Artificial Intelligence (AI)

Definition:

"*Artificial intelligence (AI) is the ability of a computer, or a robot controlled by a computer, to do tasks that are usually done by humans because they require human intelligence and discernment.*"

Related Fields:

Artificial intelligence research relies on approaches and methods from various fields: Mathematics, statistics, economics, probability, computer science, linguistics, psychology, philosophy, etc.

Practical Applications:

- Recognizing and classifying objects (on a photo or in reality)
- Playing a game of chess or poker
- Driving a car or plane
- Recognizing and understanding human speech
- Translating languages
- Moving and manipulating objects
- Solving complex business problems and making decisions like humans

Examples:

- In 2016, AlphaGo (developed by DeepMind Technologies, later acquired by Google) defeated Go champion LEE Sedol
- In 2017, Libratus (developed at Carnegie Mellon University) won a Texas hold 'em tournament involving four top-class human poker players
- In 2019, a report pooling the results of 14 separate studies revealed that AI systems correctly detected a disease state 87% of the time (compared with 86% for human healthcare professionals) and correctly gave the all-clear 93% of the time (compared with 91% for human experts)
- In 2020, researchers of Karlsruhe Institute of Technology developed a system that outperforms humans in recognizing spontaneously spoken language with minimum latency.

Narrow AI:

- AI programs that are able to solve one single kind of problem
- Narrow AI applications that work in different individual domains could be incorporated into a single machine
- Such a machine would have the capacity to learn any intellectual task that a human being can, a.k.a. "artificial general intelligence" (AGI)

Conclusions: AI is used to improve the effectiveness or efficiency of processes or decisions. It should be thought of as a facilitator of human productivity – not as a replacement for human intelligence.

1.5. Machine Learning (ML)

Definition:

"Machine learning (ML) is the study of computer algorithms that improve automatically through experience."

Machine learning is about teaching computers to:

- Find hidden insights without being explicitly programmed where to look
- Make predictions based on data and finding
- Produce a "model", e.g., a predictive model

A model can be described as:

- A formula that takes some values as input and delivers one or more values as output
- The computer can use this description to learn and then make predictions

- It is only a mathematical description of phenomena or events, and can never fully represent the reality

The creation of a model requires two elements: 1) an algorithm and 2) data

Definition:

"An algorithm is a procedure for solving a mathematical problem in a finite number of steps that frequently involves repetition of an operation."

An algorithm:

- Provides step-by-step guidance on what to do to solve the problem
- Can be run again with a few parameters variations, to see if the new configuration leads to a better result
- It can take several iterations for the algorithm to produce a good enough solution to the problem
- Selects the model that yields the best solution
- Machine learning algorithms find their way to better solutions without being explicitly programmed where to look
- Users must determine how to define when the problem is solved or what kind of changes need to be made for each iteration

Training data:

- These are past observations, used to "feed" the algorithm so that it can gain initial experience
- These observations represent units of information that teach the machine trends and similarities derived from the data

- The machine gets better with every iteration
- Once the algorithm is able to distinguish patterns, you can make predictions on new data
- The process an algorithm goes through training data again and again is called "training the model".

1.6. Supervised Learning

Definition:

"Mapping is the association of all elements of a given data set with the elements of a second set."

Supervised learning splits the data into:

- Training data: an algorithm analyzes the training data (related to existing observations)
- Validation data: an algorithm produces a function for the mapping of validation data

During the training process:

- The algorithm "learns" the mapping function from the input to the output
- The algorithm learns by comparing its own computed output with the correct outputs (provided in the training data) to find errors
- The goal is to plot a function that best approximates the relationship between the input and output observable in the training data
- Once the patterns and relationships have been computed, these can then be applied to new data, e.g., to predict the output for new observations

- The process is called “supervised” learning, because it is as if the learning took place in the presence of a supervisor checking the outcome
- In every iteration, the model provides a result
- That result is then compared with the “correct” answer (provided by the labels in the training data)
- Direct feedback is given to the algorithm, which takes it into consideration for its next iteration
- It requires technical proficiency to develop, calibrate, and validate supervised learning models
- The models produced through supervised learning are such that they are fed an input and they deliver an output
- Each observation in the training data set is tagged with the outcome the model is supposed to predict
- Outcome can also be continuous or categorical with two or more classes

Types of supervised machine learning techniques:

- Regression (when the output is continuous data)
- Time series forecasting (when input and output are arranged as a sequence of time data)
- Classification (when the output is categorical data)

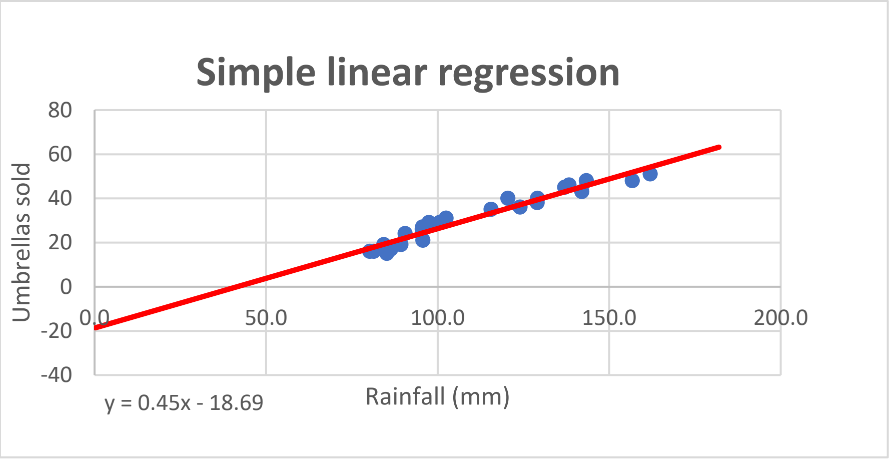
1.6.1. Regression

Regression is a method that is used to determine the strength and character between a dependent variable and one or more independent variables.

- The dependent variable is what is to be explained, or what is to be predicted
- The independent variables (a.k.a predictors) are the explanatory variables, i.e., the variables that are used to make the prediction
- Regression is intended for the prediction of continuous variables, i.e. numeric data that have an infinite number of values in a range between any two values
- The goal of the regression analysis is to draw a red line on a plot, which is the visualization of an equation connecting both variables
- Regression is particularly useful when the independent variables represent factors that can be controlled (which is not the case for rainfall).

Simple linear regression - we can make predictions as follows "given a particular x value, what is the expected value of the y variable".

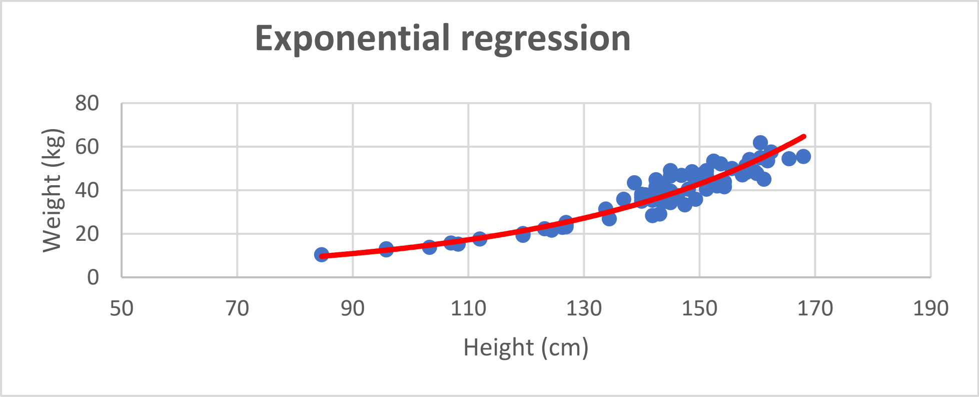
Example:



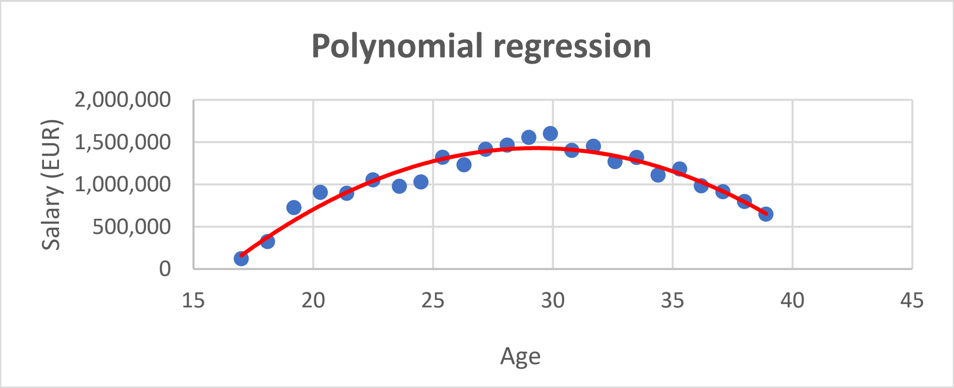
- The dots on the scatter plot correspond to the observed data, each of them representing one month (24 months in total)
- For each month, we can see the rainfall in mm (the independent variable, on the horizontal axis) and the number of umbrellas sold (the dependent variable, on the vertical axis)
- There's a positive relationship between both variables: the larger the rainfall, the higher the number of umbrellas sold – and vice versa

Polynomial regression - when the relationship between the dependent and independent variable is non-linear.

Example 1:



Example 2 :



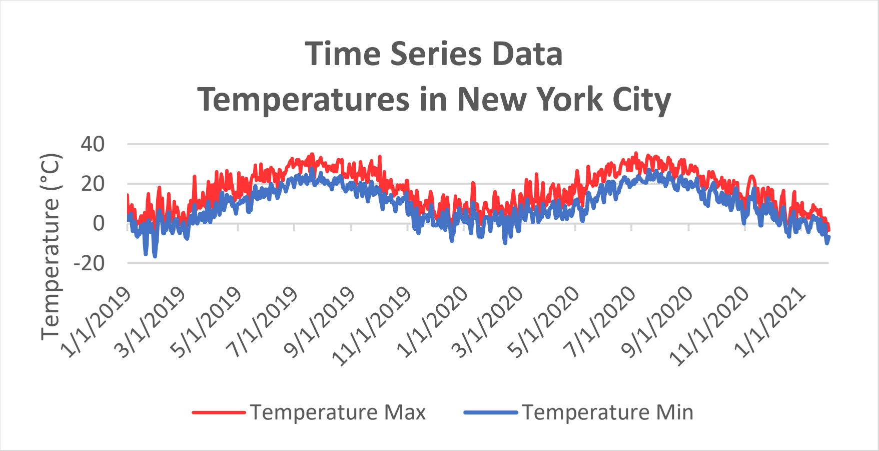
Multivariate regression - when we use two or more independent variables to predict a single outcome. The number of independent variables is virtually unlimitedMultivariate regression can determine how the variables relate to one another and what variables matter the most when explaining a phenomenon:

1.6.2. Time Series Forecasting

Definition:

"Time series forecasting is defined as the use of models to predict future values based on previously observed values."

- It is a supervised machine learning, where past values can be the training data used for the mapping of future values
- A time series is a sequence of discrete-time data, taken at equally spaced time intervals, e.g., days, months, or years
- We employ previous time periods as input variables, while the next time period is the output variable
- Time series analysis has the objective to forecast the future values of a series based on the history of that series
- Time series are used in various fields, e.g., economics, finance, weather forecasting, astronomy, supply chain management, etc.
- Time series are typically plotted via temporal line charts (run charts):



Examples:

- Daily closing prices of stocks
- Weekly sales volume of laundry detergents
- Monthly count of a website's active users
- Annual birth or death rate for a given country

1.6.3. Classification

Definition:

"*Classification is a supervised machine learning approach where a class (also called category, target, or label) is predicted for a new observation. The main goal is to identify which class the new observation will fall into.* "

- The learning is based on a training set of data with observations for which category membership is already known
- Classification can be performed on both structured and unstructured data
- The output and input data must be categorical
- Classification is best use when the output has finite and discrete values

- When there are only two categories, we have "binary" classification, e.g. Yes vs. No, Churn vs. Non-churn, etc.
- For more categories, the term used is "multi-class" classification, e.g., Positive/Neutral/Negative, Win/Draw/Loss, 0 to 9

Email spam detection - example:

Spam detection can be considered as a binary classification problem, since there two classes – "Spam" or "Not spam", respectively Spam "yes" or "no". The input data consist of variables, such as:

- Title includes keywords such as "winner", "free", "dollar", etc.
- Email body contains special formatting (such as bold, entire words in capital letters, etc.)
- Number of addressees in the "To" field of the email
- Email includes one or more attachments

The model (here - a mapping function) quantifies the extent to which the input variables affect the output (or "predicted") variable. The objective is to approximate the mapping function so that when we can predict the nature of an email (spam or not) as soon as we get an email in the mailbox.

Classifier:

An algorithm that implements classification. Classifiers are used for all kinds of applications:

- Image classification
- Fraud detection -> Is this transaction fraudulent?

- Direct marketing -> Will this customer accept this offer?
- Churn prediction -> What is the chance that this subscriber switch to another provider (high, medium, low)?
- Credit scoring -> Will this prospective borrower default?

1.7. Unsupervised Learning

Definition:

"An unsupervised learning is based on an algorithm that analyzes the data and automatically tries to find hidden patterns."

- No guidance or supervision required
- No specific desired outcome or correct answer is provided
- The only objective is for the algorithm to arrange the data as best as it can
- It identifies structure in the data
- The training data are unlabeled; more cost-effective than supervised techniques
- Here, analysts do not worry about quality and accurate labeling
- In unsupervised learning, only input data are provided, but no corresponding output
- No explicit instructions on what to do with the data is given
- The algorithm does not know what is right or wrong
- It is difficult to assess or compare the truthfulness of results obtained

- The two most useful unsupervised learning techniques are clustering and association

Unsupervised learning techniques are used to:

- Detect anomalies
- Fraud detection (via flagging of outliers)
- Predictive maintenance (via discovery of defective parts in a machine or system)
- Network intrusion detection
- Reduce features in a data set
- Describing customers with 5 attributes almost as precisely as with 10 attributes

1.7.1. Clustering Analysis

Definition:

"Clustering is the splitting of a data set into a number of categories (or classes, labels), which are initially unknown."

- You do not know in advance what we are looking for
- The data are unlabeled
- The objective is to discover possible labels automatically
- These categories produced are only interpreted after their creation
- A clustering algorithm automatically identifies categories to arrange the data in a meaningful way

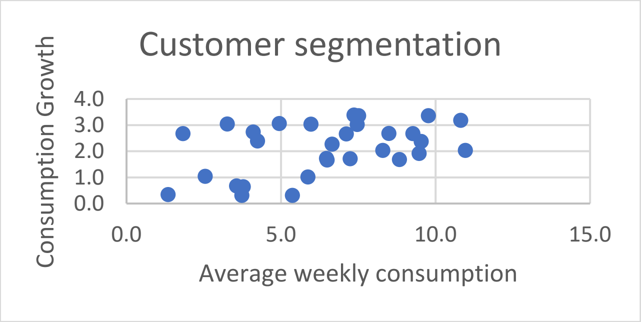
- The grouping of the data into categories is based on some measure of inherent similarity or distance
- The algorithm must group the data in a way that maximizes the homogeneity within and the heterogeneity between the clusters
- The interpretation of the clusters constitutes a test for the quality of the clustering

Example:

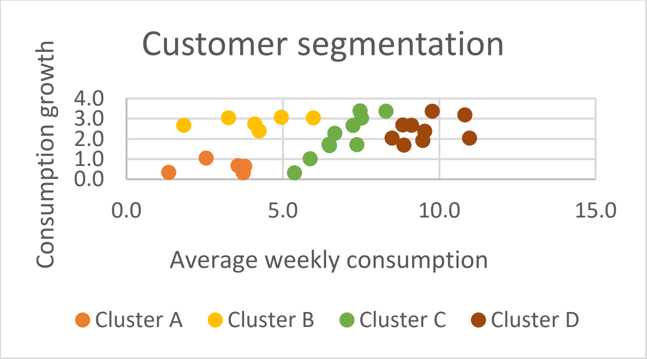
30 data points (each data point corresponding to one consumer of chocolate bars) and two variables:

X-axis - the average weekly consumption of chocolate bars in the last 12 months

Y-axis - the growth in consumption of chocolate bars in the 12 months, compared to the previous 12 months



By applying the clustering technique, 4 distinct categories emerge:



These four customer segments could be specifically targeted with promotions or adverts. Cluster B could be lured with discounts; Cluster A might prefer healthier products.

How to interpret these groups is a question to be answered by the business:

- What do you know about chocolate bar consumption habits and their effects on buying patterns?
- Why did these patterns emerge and how are they relevant to our business?
- What could we offer to consumers in Cluster C (average weekly consumption) so that they buy more of our products?

Practical applications:

- Customer segmentation
- Sorting of e-mails or documents into themes
- Grouping of retail products based on customer behavior or other characteristics
- Image recognition, e.g., grouping of similar images

1.7.2. Association Rules

Definition:

"Association rule learning is *an unsupervised machine learning method to discover interesting associations (relationships, dependencies) hidden in large data sets."*

The associations are usually represented in the form of rules or frequent item sets.

Practical applications:

- Market basket analysis: The exploration of customer buying patterns by finding associations between items that shoppers put into their baskets, e.g., “tent and sleeping bags” or “beer and diapers”; it produces the probability of a joint occurrence
- Product recommendation: When a customer puts a tent in his online shopping cart, the website displays an offer for a sleeping bag
- Design of promotions: When the retailer offers a special price for just one of the products (crackers) but not for the other one at the same time (dip)
- Customization of store layout: When you increase customer traffic so that customers have more opportunities to buy products; strawberries and whipped cream can be placed at the opposite ends of the store

1.8. Reinforcement Learning

Definition:

“Reinforcement learning is a technique where *a machine (“agent”) learns through trial and error, using feedback from its own actions.”*

- The learning is based on the rewarding of desired behaviors and/or the punishing undesired ones (“the carrot and stick for machines”)

- The agent is given an overarching goal to reach. It performs tasks randomly, and “sees what happens” in every iteration. Over time, it learns what brings a reward or a punishment
- The agent is programmed to seek the maximum overall reward when interacting within a given situation (“environment”)
- Used for the development of robots, autonomous vehicles, and game-playing programs, and in various business areas, e.g., marketing

Reinforcement learning vs. supervised learning:

- In both approaches, algorithms are given specified goals
- They use mapping between input and output; outcomes can be positive or negative
- Reinforcement learning does not require labelled data sets
- The agent is not given any directions on how to complete the task
- The agent can uncover entirely new solutions

Example:

In 2016, AlphaGo defeated LEE Sedol, the world champion in the game of Go. Within a period of just a few days, AlphaGo had accumulated thousands of years of human knowledge, enough to beat a champion with 21 years of professional experience.

1.9. Deep Learning

Definition:

"Deep learning is a subfield of machine learning, applying methods inspired by the structure and function of the human brain."

- Deep learning can be supervised, unsupervised or take place through reinforcement learning
- With deep learning, computers learn how to solve particularly complex tasks – close to what we describe as "artificial intelligence"
- Deep learning teaches machines to imitate the way humans gain knowledge
- The algorithm performs a task repeatedly, adjusting it a bit to improve the outcome
- Neural networks can be taught to classify data and identify patterns like humans
- When the machine receives new information, the algorithm compares it with known objects and tries to draw similar conclusions
- The more it knows about the objects' distinguishing characteristics, the more likely the algorithm is to solve the problem, e.g., recognize a certain animal on an image
- The system improves its ability as it gets exposed to more examples

Artificial neural networks (ANN) - The algorithms used to achieve deep learning.

- ANN were inspired by the way the brain processes information and how communication nodes are distributed
- A neural network is made of a collection of connected nodes (called neurons)
- By connecting the nodes to other another, information can "flow" through the network
- Neurons are organized in layers
- The more layers involved, the "deeper" the neural network

Practical applications:

- Image recognition
- Autonomous driving: The computer can differentiate between pedestrians of different sizes, e.g., children vs. adults
- Speech recognition, e.g., to detect "Hey Siri" spoken by iPhone users, language translation, e.g., Google Translate
- Product recommendations, e.g., Netflix or YouTube

1.10. Natural Language Processing (NLP)

Definition:

"Natural Language Processing (NLP) is a field of artificial intelligence concerned with the interactions between computers and human (natural) language."

- It enables computers to process and analyze large amounts of natural language data

- The goal of NLP is to decipher and to derive meaning from human languages
- It taps into the treasure troves of unstructured data that are held by companies (emails, chats, blogs, images, etc.)
- There are two sub-topics of NLP

1. Natural Language Understanding (NLU) – It deals with the reading comprehension of machines, i.e., the ability to process documents, understand its meaning, and to integrate with what the reader already knows.

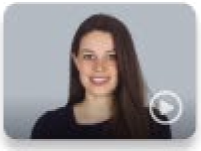
Sentiment Analysis – the interpretation and classification of emotions in text.

- A machine tries to identify and categorize terms and opinions expressed in a written or spoken text
- The objective is to determine the author's attitude towards a subject
- The output is a sentiment score, like positive, neutral, or negative
- Used in marketing, customer relationship management or customer service, e.g., to determine how they feel about their brands, products, services, etc.
- Brands can apply it to emails, social media posts, phone calls, reviews on eCommerce platforms, etc.
- It can be a valuable source of insights, as the feedback is left in an unprompted fashion

- It delivers information about customers' preferences and choices, and decisions
- The sentiment score is numerical
- It is used in subsequent analyses, e.g., to establish a quantitative relationship between customer satisfaction and revenue

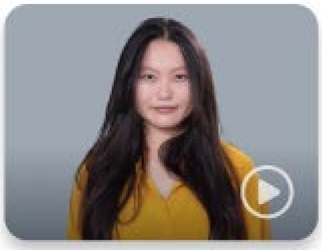
2.Natural Language Generation (NLG) – It deals with the creation of meaningful sentences in the form of natural language.

- It is about transforming structured data into natural language
- An NLG system makes decisions about how to turn a concept into words and sentences
- Less complex than NLU since the concepts to articulate are usually known
- An NLG system must simply choose the right expressions several potential ones
- Classic examples: the production of (written out) weather forecasts from weather data, automated journalism, generate product descriptions for eCommerce sites, interact with customers via chatbots, etc.



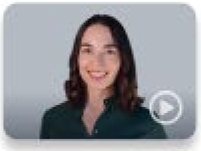
Danielle Thé
Esade Ramon Llull University

Bernard Marr
Cambridge University



Tina Huang
University of Pennsylvania

Ken Jee
DePaul University



Anastasia Kuznetsova
Université Côte d'Azur

Olivier Maugain

Email: team@365datascience.com

365√DataScience