Foundation of Data analytics

UNIT -I

Understanding Buzzwords in Data Science

- Artificial Intelligence (AI)
- Learning System or Algorithm.
- How to decide which Learning Algorithm to use.
- **Machine Learning**.
- **Machine Learning** Algorithms.
- Deep Learning.
- Deep Learning.
- Artificial Neural Network (ANN)
    - Machine Learning Vs Deep Learning
    - AI Vs ML Vs DL
    - Deep Learning Vs Neural Network
    - Data Science
    - Data Science Flow Chart
    - Why Deep Learning and why not SVM?
    - What is deep learning? Why is this a growing trend in machine learning? Why not use SVMs?
    - Five main reasons why deep learning is so popular
    - Explainable AI (XAI)
    - References

Artificial Intelligence (AI)

According to John McCarthy, godfathers of AI, it is **"The science and engineering of making intelligent machines, especially intelligent computer programs".**

Artificial intelligence is the simulation of human intelligence processes by machines, especially computer systems.Somewhere I read an article (I missed the details) where D.J. Patil, from LinkedIn says:

"Artificial intelligence is a broad term that simply means: any intelligence run by a computer. Machine learning is a subpart of this, but you don't need machine learning to have an AI.

Andrew NG (One of my favorite Data Scientists) from Coursera tells us that AI is too big for any one person to understand—there are so many ways AI can be used and put into practice, it's nearly impossible to document and understand them all.

Some of the activities computers with artificial intelligence are designed for include:

- Speech recognition
- Learning
- Planning
- Problem-solving

Goals of AI

- To Create Expert Systems − The systems which show intelligent behavior, learn, demonstrate, explain, and advice its users.
- To Implement Human Intelligence in Machines − Building systems that understand, think, learn, and behave like people.

AI can be split between 2 branches as below

- Applied AI (Week AI) -->

  A machine which perform some specific tasks, such as Alexa; google assistant; email spam; housing prediction; stock prediction; weather predictions etc.

- Generalized AI (Strong AI) -->

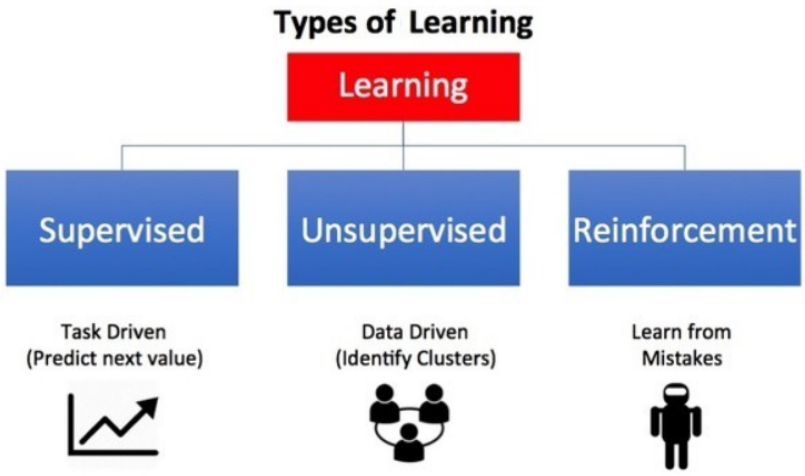
  A machine which acts like a human and perform task like a human

  Machines which can take decision, react instantly.

  Examples are Robots; Self-Driving cars etc.

Usually when anyone talks around us related to Artificial Intelligence, they are mainly referring to Week AI.

-



Supervised learning

- Meaning of supervised is "observe and direct the execution of (a task or activity)" or “supervisor” or “teacher”.
- Supervised learning is a type of machine learning algorithm that uses a known dataset (called the training dataset) to make predictions.
- o The training dataset includes input data (features / attributes / vector / independent variables etc.,.) and response values (dependent variable / class / target etc.,). From it, the supervised learning algorithm seeks to build a model that can make predictions of the response values for a new dataset. A test dataset is often used to validate the model. Using larger training datasets often yield models with higher predictive power that can generalize well for new datasets.
- Supervised learning classified into two categories of algorithms.
    - Classification: A classification problem is when the output variable is a category, such as “Red” or “blue” / “disease” and “no disease”.

      In-short, for categorical response values, where the data can be separated into specific “classes”.

    - Regression: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

      In-short, for continuous-response values.

- Below are the few Algorithms used in Supervised Learning.

Unsupervised learning

- Meaning of unsupervised is “no supervisor” or “no teacher” or to act without anyone’s supervision or direction.
- Unsupervised learning means there is no training phase where we feed labelled data to the learning algorithm in order to train the model. Instead the algorithm must figure out things by itself.

- Unsupervised learning finds hidden patterns or intrinsic structures in input data. It is used to draw inferences from datasets consisting of input data without labeled responses
- In the unsupervised case, the goal is to discover patterns, deep insights, understand variation, find unknown subgroups (amongst the variables or observations), and so on in the data. Unsupervised learning can be quite subjective compared to supervised learning.
- The two most commonly used techniques in unsupervised learning are Association and Clustering.
- Association algorithms identity relationships between variables. A frequently quoted example is that if we feed sales data, it can identify patterns such as the people who bought item “A” has a probability of p% for buying item “B” too.
- Clustering algorithms groups data into clusters based on similar patterns. An example, if you feed many (say thousands; millions etc of) pictures of various animals, the clustering algorithm will group them into various clusters such as cats, dogs etc.
- Another example of Cluster algorithm, If a Telecom company wants optimize the locations where they build cell phone towers, they can use machine learning to estimate the number of clusters of people relying on their towers. A phone can only talk to one tower at a time, so the team uses clustering algorithms to design the best placement of cell towers to optimize signal reception for groups, or clusters, of their customers.
- Some techniques or algorithms that are used in Unsupervised learning are
    - PCA
    - SVM
    - k-Means,
    - Anomaly detection Algorithm
    - Neural Networks, and
    - Latent Variable Models

Reinforcement Learning

- Reinforcement learning is training by rewards and punishments.
- In reinforcement learning the system learns from the environment. When the system does something right, it is rewarded. When it does something wrong, it is not.
- The system learns in a very similar way to how a person would learn.
- In this type of machine learning, the machine itself learn how to behave in the environment by performing actions and comparing with the results.
- It is like machine performing trial and error method to determine the best action possible based on the experience.
- Reinforcement learning involves goal-oriented algorithms, which attain a complex goal with multiple steps which ultimately improves the performance of the machine to predict things.
- The aim of the game in reinforcement learning is to maximize the reward.
- There are many different types of algorithms for reinforcement learning in python.

Two of the most common for the multi-arm bandit problem are upper confidence bound and Thompson sampling.

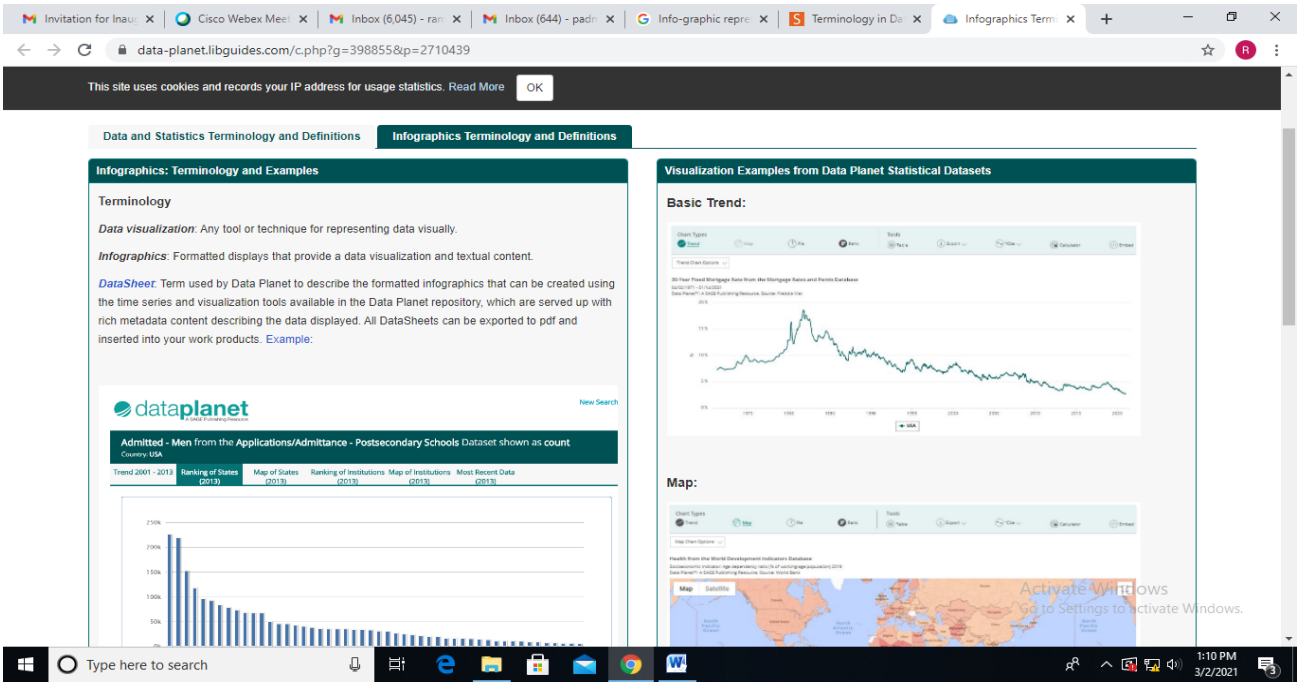
- Neural networks are the solution to most of the complex problems in Artificial intelligence like Computer vision, machine translation etc. If Neural networks combined with reinforcement learning, then it is very easy to solve even more complex problems. This way of integrating neural networks with reinforcement learning is known as Deep Reinforcement learning.

Info-graphic Representation of Terminologies

(a clipped compound of "information" and "graphics") are graphic visual representations of information, data, or knowledge intended to present information quickly and clearly. They can improve cognition by utilizing graphics to enhance the human visual system's ability to see patterns and trends.[3][4] Similar pursuits are information visualization, data visualization, statistical graphics, information design, or information architecture. Info-graphic evolved in recent years to be for mass communication, and thus are designed with fewer assumptions about the readers' knowledge base than other types of visualizations. so types are an early example of info-graphic conveying information quickly and easily to the masses.

Data and Statistics: Terminology and Definitions: Info-graphic Terminology and Definitions

Defines terms relevant to using data and statistics.



Terminology in Data Analytics

As data continue to grow at a faster rate than either population or economic activity, so do organizations' efforts to deal with the data deluge, and use it to capture value. And so do the methods used to analyze data, which creates an expanding set of terms (including some buzzwords) used to describe these methods.

This is a field in flux, and different people may have different conceptions of what terms mean. Comments on this page and its "definitions" are welcome. Since many of these terms are subsets of others, or overlapping, the clearest approach is to start with the more specific terms and move to the more general.

Predictive modeling:

Used when you seek to predict a target (outcome) variable (feature) using records (cases) where the target is known. Statistical or machine learning models are "trained" using the known data, then applied to data where the outcome variable is unknown. Includes both classification (where the outcome is categorical, often binary) and prediction (where the outcome is continuous).

Predictive analytics:

Basically the same thing as predictive modeling, but less specific and technical. Often used to describe the field more generally.

Supervised Learning:

Another synonym for predictive modeling.

Unsupervised Learning:

Data mining methods not involving the prediction of an outcome based on training models on data where the outcome is known. Unsupervised methods include cluster analysis, association rules, outlier detection, dimension reduction and more.

Business intelligence:

An older term that has come to mean the extraction of useful information from business data without benefit of statistical or machine learning models (e.g. dashboards to visualize key indicators, queries to databases).

Data mining:

This term means different things in different contexts. To a lay person, it might mean the automated searching of large databases. To an analyst. it may refer to the collection of statistical and machine learning methods used with those databases (predictive modeling, clustering, recommendation systems, …)

Text mining:

The application of data mining methods to text.
Text analytics:

A broader term that includes the preparation of text for mining, the mining itself, and specialized applications such as sentiment analysis. Preparing text for analysis involves automated parsing and interpretation (natural language processing), then quantification (e.g. identifying the presence or absence of key terms).
Data science, data analytics, analytics:

Cover all of the concepts described on this page. "Data science" is often used to define a (new) profession whose practitioners are capable in many or all the above areas; one often sees the term "data scientist" in job postings. While "statistician" typically implies familiarity with research methods and the collection of data for studies, "data scientist" implies the ability to work with large volumes of data generated not by studies, but by ongoing organizational processes. Due to the complexity of dealing with large datasets and data flows, most of the day-to-day work of a data scientist lies in data pipeline challenges – storing relevant data, getting it into appropriate form for analysis, and managing the real-time implementation of models. "Data analytics" and "analytics," by contrast, are general terms used to describe the field and a comprehensive collection of associated methods. Wikipedia references here and here. All these terms tend to be used for the application of analytic methods to data that large organziations generate or have available ("big data").
Statistics:

Covers nearly all of the above methods, and also carries the mantle of a well-established profession dating back to the mid 1800's. Although statisticians work on "big data" problems, the field of statistics has traditionally been focused on focused research studies (e.g. drug trials).

Big Data:

Refers to the huge amounts of data that large businesses and other organizations collect and store. It might be unstructured text (streams of tweets) or structured quantitative data (transaction databases). In the 1990's organizations began making efforts to extract useful information from this data. The challenges of big data lie mainly

. DATA ANALYSIS VS. . DATA ANALYTICS

. DATA ANALYSIS

- According to Merriam Webster, analysis is the division of a whole into small components, and analytics is the science of logical analysis. While analysis looks backward over time and works on the facts and figures of what has happened, analytics work towards modeling the future or predicting a result. In

other words, the analysis restructures existing available information or data. And, the analytics uses this analyzed information to predict what may happen.

- Data analysis is the process of studying a given data set (in close detail), dividing them into small components, and studying the subcomponents individually and their relationship with each another.
- Data analysis is a process involving the collection, manipulation, and examination of data for getting a deep insight.
- Data analysis helps design a strong business plan for businesses, using its historical data that tell about what worked, what did not, and what was expected from a product or service.
- In data analysis, experts explore past data, break down the macro elements into the micros with the help of statistical analysis, and draft a conclusion with deeper and significant insights.

- Tools used for data analysis are Open Refine, Rapid Miner, KNIME, Google Fusion Tables, Node XL, Wolfram Alpha, Tableau Public, etc. Tools used in Data analytics are Python, Tableau Public, SAS, Apache Spark, Excel, etc.
- Data analysis is a process of studying, refining, transforming, and training of the past data to gain useful information, suggest conclusions and make decisions.

DATA ANALYTICS

- Data analytics, on the other hand, is a more comprehensive term referring to a discipline that comprises the complete management of data, including collection, cleaning, organizing, storing, administering, and analysis of data with the help of specialized tools and techniques. In other words, data analysis is a process or method, whereas data analytics is an overarching discipline (science).
- Data analytics is taking the analyzed data and working on it in a meaningful and useful way to make well-versed business decisions.
- Data analytics helps businesses in utilizing the potential of the past data and in turn identifying new opportunities that would help them plan future strategies. It helps in business growth by reducing risks, costs, and making the right decisions.
- Data analytics utilizes different variables and creates predictive and productive models to challenge in a competitive marketplace.

- Data analytics is more extensive in its scope and encompasses data analysis as a sub-component. The life cycle of data analytics also comprises data analysis as one of the significant steps.
- Data analytics is using data, machine learning tools, statistical analysis, and computer-based patterns to gain better insight and design better strategies.

14 Fascinating Data Analytics Real Life Applications in 2021

Table of Contents

1. Security

Data analytics applications or, more specifically, predictive analysis has also helped in dropping crime rates in certain areas. In a few major cities like Los Angeles and Chicago, historical and geographical data has been used to isolate specific areas where crime rates could surge. On that basis, while arrests could not be made on a whim, police patrols could be increased. Thus, using applications of data analytics, crime rates dropped in these areas.

2. Transportation

Data analytics can be used to revolutionize transportation. It can be used especially in areas where you need to transport a large number of people to a specific area and require seamless transportation. This data analytical technique was applied in the London Olympics a few years ago.

For this event, around 18 million journeys had to be made. So, the train operators and TFL were able to use data from similar events, predict the number of people who would travel, and then ensure that the transportation was kept smooth.

3. Risk detection

One of the first data analytics applications may have been in the discovery of fraud. Many organizations were struggling under debt, and they wanted a solution to this problem. They already had enough customer data in their hands, and so, they applied data analytics. They used ‘divide and conquer’ policy with the data, analyzing recent expenditure, profiles, and any other important information to understand any probability of a customer defaulting. Eventually, it led to lower risks and fraud.

4. Risk Management

Risk management is an essential aspect in the world of insurance. While a person is being insured, there is a lot of data analytics that goes on during the process. The risk involved while insuring the person is based on several data like actuarial data and claims data, and the analysis of them helps insurance companies to realize the risk.

Underwriters generally do this evaluation, but with the advent of data analysis, analytical software can be used to detect risky claims and push such claims before the authorities for further analysis.

5. Delivery

Several top logistic companies like DHL and FedEx are using data analysis to examine collected data and improve their overall efficiency. Using data analytics applications, the companies were able to find the best shipping routes, delivery time, as well as the most cost-efficient transport means. Using GPS and accumulating data from the GPS gives them a huge advantage in data analytics.

6. Fast internet allocation

While it might seem that allocating fast internet in every area makes a city 'Smart', in reality, it is more important to engage in smart allocation. This smart allocation would mean understanding how bandwidth is being used in specific areas and for the right cause.

It is also important to shift the data allocation based on timing and priority. It is assumed that financial and commercial areas require the most bandwidth during weekdays, while residential areas require it during the weekends. But the situation is much more complex. Data analytics can solve it.

For example, using applications of data analysis, a community can draw the attention of high-tech industries and in such cases, higher bandwidth will be required in such areas.

7. Reasonable Expenditure

When one is building Smart cities, it becomes difficult to plan it out in the right way. Remodeling of the landmark or making any change would incur large amounts of expenditure, which might eventually turn out to be a waste. Data analytics can be used in such cases. With data analytics, it will become easier to direct the tax money in a cost-efficient way to build the right infrastructure and reduce expenditure.

8. Interaction with customers

In insurance, there should be a healthy relationship between the claims handlers and customers. Hence, to improve their services, many insurance companies often use customer surveys to collect data. Since insurance companies target a diverse group of people, each demographic has their own preference when it comes to communication. Data analysis can help in zeroing in on specific preferences. For example, a study showed that modern customers prefer communication through social media or online channels, while the older demographic prefers telephonic communication.

9. Planning of cities

One of the untapped disciplines where data analysis can really grow is city planning. While many city planners might be hesitant towards using data analysis in their favour, it only results in faulty cities riddled congestion. Using data analysis would help in bettering accessibility and minimizing overloading in the city.

Overall, it will generate more efficiency in the planning process. Just erecting a building in a suitable spot will not create an overall benefit for a city since it can harm the neighbors or the traffic in the area. Using data analytics and modeling, it will be easy to predict the outcome of placing a building in a specific situation and therefore, plan accordingly.

10. Healthcare

While medicine has come a long way since ancient times and is ever-improving, it remains a costly affair. Many hospitals are struggling with the cost pressures that modern healthcare has come with, which includes the use of sophisticated machinery, medicines, etc.

But now, with the help of data analytics applications, healthcare facilities can track the treatment of patients and patient flow as well as how equipment are being used in hospitals. It has been estimated that there can be a 1% efficiency gain achieved if data analytics became an integral part of healthcare, which will translate to more than $63 billion in healthcare services.

11. For Travelling

If you ever thought travelling is a hassle, then data analytics is here to save you. Data analysis can use data that shows the desires and preferences of different customers from social media and helps in optimizing the buying experience of travellers. It will also help companies customize their own packages and offer and hence boost more personalized travel recommendations with the help data collected from social media.

12. Managing Energy

Many firms engaging with energy management are making use of applications of data analytics to help them in areas like smart-grid management, optimization of energy, energy distribution, and automation building for other utility-based companies. How does data analytics help here?

Well, it helps by focusing on controlling and monitoring of a dispatch crew, network devices, and management of service outages. Since utilities integrate about millions of data points within the network performance, engineers can use data analytics to help them monitor the entire network.

13. Internet searching

When you use Google, you are using one of their many data analytics applications employed by the company. Most search engines like Google, Bing, Yahoo, AOL, Duckduckgo, etc. use data analytics. These search engines use different algorithms to deliver the best result for a search query, and they do so within a few milliseconds. Google is said to process about 20 petabytes of data every day.

14. Digital advertisement

Data analytics has revolutionized digital advertising, as well. Digital billboards in cities as well as banners on websites, that is, most of the advertisement sources nowadays use data analytics using data algorithms. It is one of the reasons why digital advertisements are getting more CTRs than traditional advertising techniques. The target of digital advertising nowadays is focused on the analysis of the past behaviour of the user.