Introduction to Data Science and Machine Learning for Strategic Decision Making

Présenté par:
Mbaye Babacar GUEYE, Ph.D

THE TRAINER

Mbaye Babacar GUEYE, Ph.D

Ph.D in Computer Science: COmplex System Modeling of Université Cheikh Anta Diop de Dakar

Ph.D in Statistical Data Processing/AI of Sorbonne Paris VI

Principal (Manager) Modern Data Architecture (Data and Advanced Analytics) at Slalom Canada

Over 10 years of experience in the Cloud tech, Advanced Analytics , Big Data (NoSQL, Datalake, DWH, etc..)

Over 10 years of experience in the Teaching

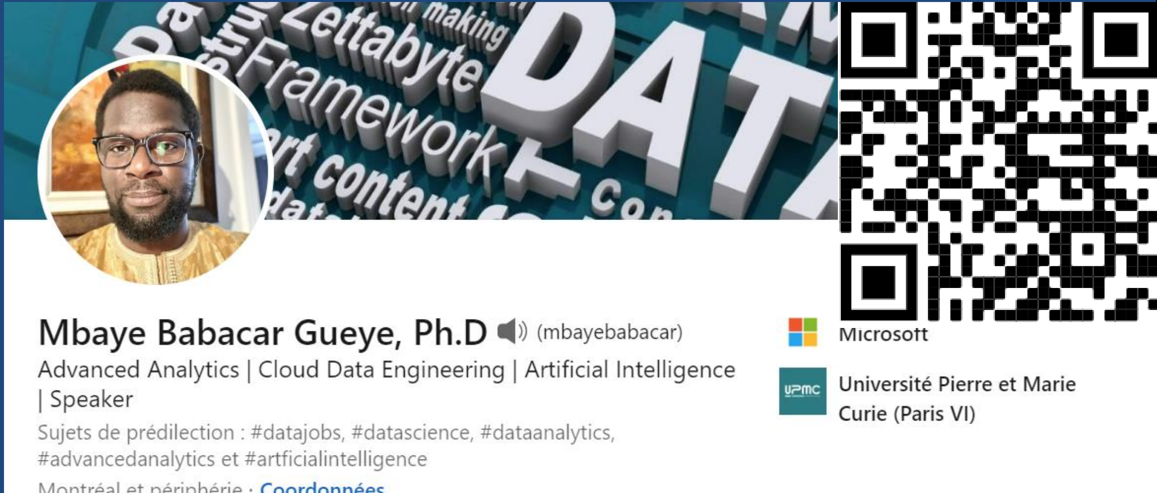
Over 15 years of experience in Sofware Engineering

Contact:

✉ gueye@mbayebabacar.me

@mbayebgueye

in www.linkedin.com/in/mbgueye/



Mbaye Babacar Gueye, Ph.D 🔊 (mbayebabacar)

Advanced Analytics | Cloud Data Engineering | Artificial Intelligence | Speaker

Sujets de prédilection : #datajobs, #datascience, #dataanalytics, #advancedanalytics et #artficialintelligence

Montréal et périphérie · Coordonnées

Microsoft

Université Pierre et Marie Curie (Paris VI)

OUTLINE

1. **Intro to Data Science**
   a. Definition of Data Science
   b. Basics in Data Science

2. **Basics in ML**
   a. Intro to AI
   b. ML concepts
   c. Data classification and Data clustering

3. **From ML to AI**
   a. Introduction to Deep Learning
   b. Scale AI with MLOps

4. **Advanced Analytics for Strategic Decision Making**
   a. Intro to Big Data
   b. Data and AI in the Cloud
   c. The notion of Advanced Analytics
   d. Strategic Decision-Making guided by Data

What's the Data Science.

Context: Computer Science, Data Science, and Real Science

Computer scientists, by nature, don't respect data. They have traditionally been taught that the algorithm was the thing, and that data was just meat to be passed through a sausage grinder.

To qualify as an effective data scientist, you must first learn to think like a real scientist. Real scientists strive to understand the natural world, which is a complicated and messy place.

THE Data Science Design MANUAL-Steven S. Skiena

What's the Data Science.

Context: Computer Science, Data Science, and Real Science

Data vs. method centrism: Scientists are data driven, while computer scientists are algorithm driven. Real scientists spend enormous amounts of effort collecting data to answer their question of interest.

By contrast, computer scientists obsess about methods: which algorithm is better than which other algorithm

Concern about results: Real scientists care about answers. They analyze data to discover something about how the world works.

THE Data Science Design MANUAL-Steven S. Skiena

What’s the Data Science.

Definition

Data Science is both

Computer Science (algorithm driven) + Real Science (Data driven)

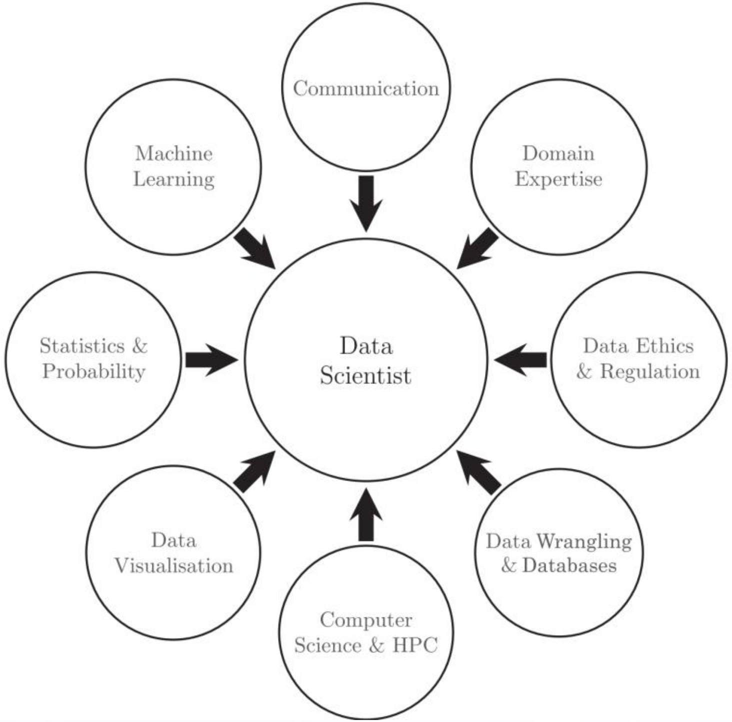
What's the Data Science.

Definition

Data science as lying at the intersection of **computer science, statistics, and substantive application domains.**

- **From computer science** comes machine learning and high-performance computing technologies for dealing with scale.
- **From statistics** comes a long tradition of exploratory data analysis, significance testing, and visualization.
- **From application** domains in business and the sciences comes challenges worthy of battle, and evaluation standards to assess when they have been adequately conquered.

What’s the Data Science.



What's the Data Science.

Definition

Data science is an amalgamation of various disciplines of research including the method ranging **from empirical methods** to current trends of data observation along with **the application of rigorous skepticism about observations in data,** resulting in interprets that involve formulating hypotheses and deducing inductions.

Data Science in Societal Applications

What's the Data Science.

Definition

Data science encompasses a set of principles, problem definitions, algorithms, and processes **for extracting nonobvious and useful patterns from data sets.**

Data Science
John D. Kelleher; Brendan Tierne

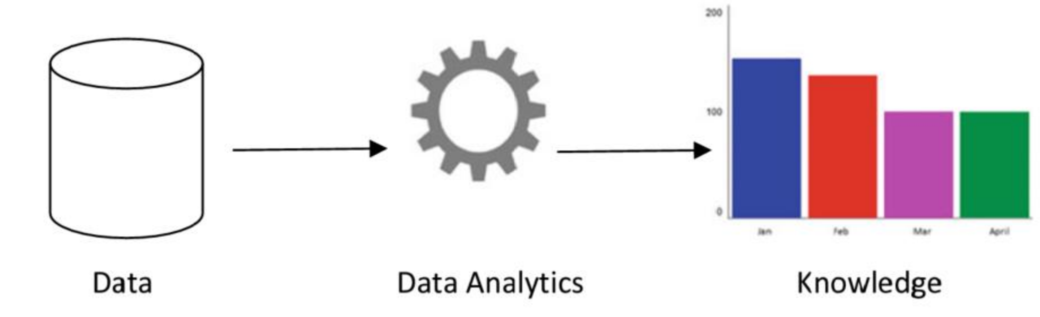
Data.

Definition

A collection of discrete values that convey information, describing quantity, quality, fact, statistics, other basic units of meaning, or simply sequences of symbols that may be further interpreted (***Wikipedia)***.

Data <> information

Information = Data + interpretation model



Data properties.

Data properties.

3 data properties:

- Data category
- Data structure
- Data scale

Data category

Quantitative vs. Categorical Data

Data structure

Structured vs Semi-Structured vs Unstructured Data

Structured Data

Think of data that fits neatly within **fixed fields and columns in relational databases and spreadsheets**.

Data that resides in a fixed field within a file or record. **Structured data is typically stored in a relational database (RDBMS)**.

Examples of structured data include names, dates, addresses, credit card numbers, stock information, geolocation, and more.

Sname	Address	Course-id
Amadou	Amitie 3, Dakar	C01
Babacar	Escale Thies	C01
Christian	Park St. Dakar	C03
Mendy	Rue de Touba Diourbel	C04
Ousmane	Akbar Road Thies	C02
Malick	Rue Ababacar SY, Thies	C04

Structured vs Semi-Structured vs Unstructured Data

Unstructured Data

Unstructured data is most often categorized as qualitative data, and it cannot be processed and analyzed using conventional data tools and methods.

Examples of unstructured data include **text**, video files, audio files, mobile activity, social media posts, satellite imagery, surveillance imagery .

Example:

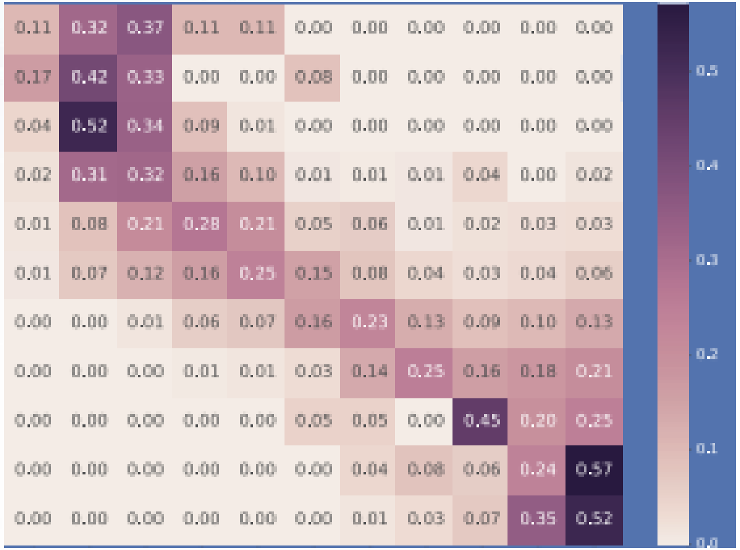
Advances in computing technologies have led to the advent of "Big Data". Big Data usually refers to very large quantities of data, usually at the petabyte scale. Using traditional data analysis methods and computing, working with such large (and growing) datasets is difficult, even impossible. (Theoretically speaking, infinite data would yield infinite information

Data structure

Structured vs Semi-Structured vs Unstructured Data

Dealing with unstructured Data

When confronted with an unstructured data source, such as a collection of tweets from Twitter or an image, our first step is generally to build a matrix to structure it.



0.11	0.32	0.37	0.11	0.11	0.00	0.00	0.00	0.00	0.00	0.00
0.17	0.42	0.33	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00
0.04	0.52	0.34	0.09	0.01	0.00	0.00	0.00	0.00	0.00	0.00
0.02	0.31	0.32	0.16	0.10	0.01	0.01	0.01	0.04	0.00	0.02
0.01	0.08	0.21	0.28	0.21	0.05	0.06	0.01	0.02	0.03	0.03
0.01	0.07	0.12	0.16	0.25	0.15	0.08	0.04	0.03	0.04	0.06
0.00	0.00	0.01	0.06	0.07	0.16	0.23	0.13	0.09	0.10	0.13
0.00	0.00	0.00	0.01	0.01	0.03	0.14	0.25	0.16	0.18	0.21
0.00	0.00	0.00	0.00	0.00	0.05	0.05	0.00	0.45	0.20	0.25
0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.08	0.06	0.24	0.57
0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.07	0.35	0.52

A bag of words model will construct a matrix with a row for each tweet, and a column for each frequently used vocabulary word. Matrix entry M[i, j] then denotes the number of times tweet i contains word j.

Data structure

Structured vs Semi-Structured vs Unstructured Data

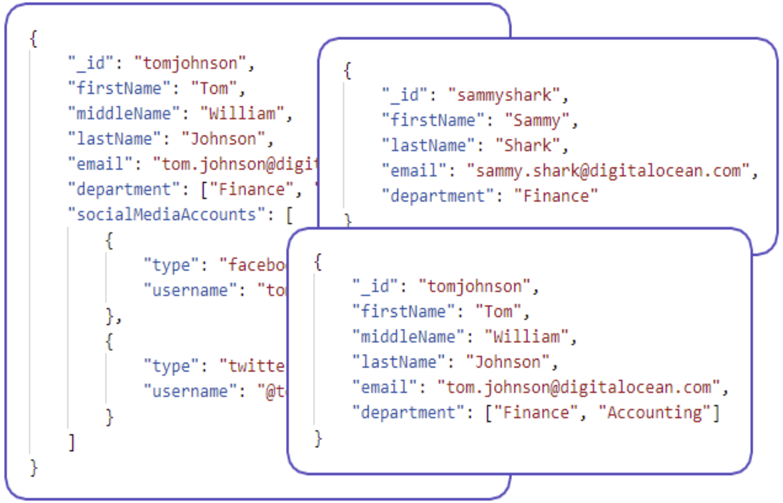
Semi-structured Data

Semi-structured **data maintains internal tags and markings that identify separate data elements**, which enables data analysts to determine **information grouping and hierarchies**. Both documents and databases can be semi-structured.

Semi-structured data lies midway between structured and unstructured data.

It doesn't have a specific relational or tabular data model but includes **tags and semantic markers that scale data into records and fields in a dataset**.

The most used format in the business world



```
{
    "_id": "tomjohnson",
    "firstName": "Tom",
    "middleName": "William",
    "lastName": "Johnson",
    "email": "tom.johnson@digit
    "department": ["Finance", '
    "socialMediaAccounts": [
        {
            "type": "facebo
            "username": "to
        },
        {
            "type": "twitte
            "username": "@t
        }
    ]
}
```

```
{
    "_id": "sammyshark",
    "firstName": "Sammy",
    "lastName": "Shark",
    "email": "sammy.shark@digitalocean.com",
    "department": "Finance"
}
```

```
{
    "_id": "tomjohnson",
    "firstName": "Tom",
    "middleName": "William",
    "lastName": "Johnson",
    "email": "tom.johnson@digitalocean.com",
    "department": ["Finance", "Accounting"]
}
```

Data scale

Big Data vs. Little Data

The analysis cycle time slows as data size grows: Computational operations on data sets take longer as their volume increases. Small spreadsheets provide instantaneous response, allowing you to experiment and play what if?.

Clever algorithms can permit amazing things to be done with big data but staying small generally leads to faster analysis and exploration.

Large data sets are complex to visualize: Plots with millions of points on them are impossible to display on computer screens or printed images, let alone conceptually understand

Simple models do not require massive data to fit or evaluate: A typical data science task might be to make a decision on the basis of a small number of variables:

Basics in Data Science

Basic Algorithms and models

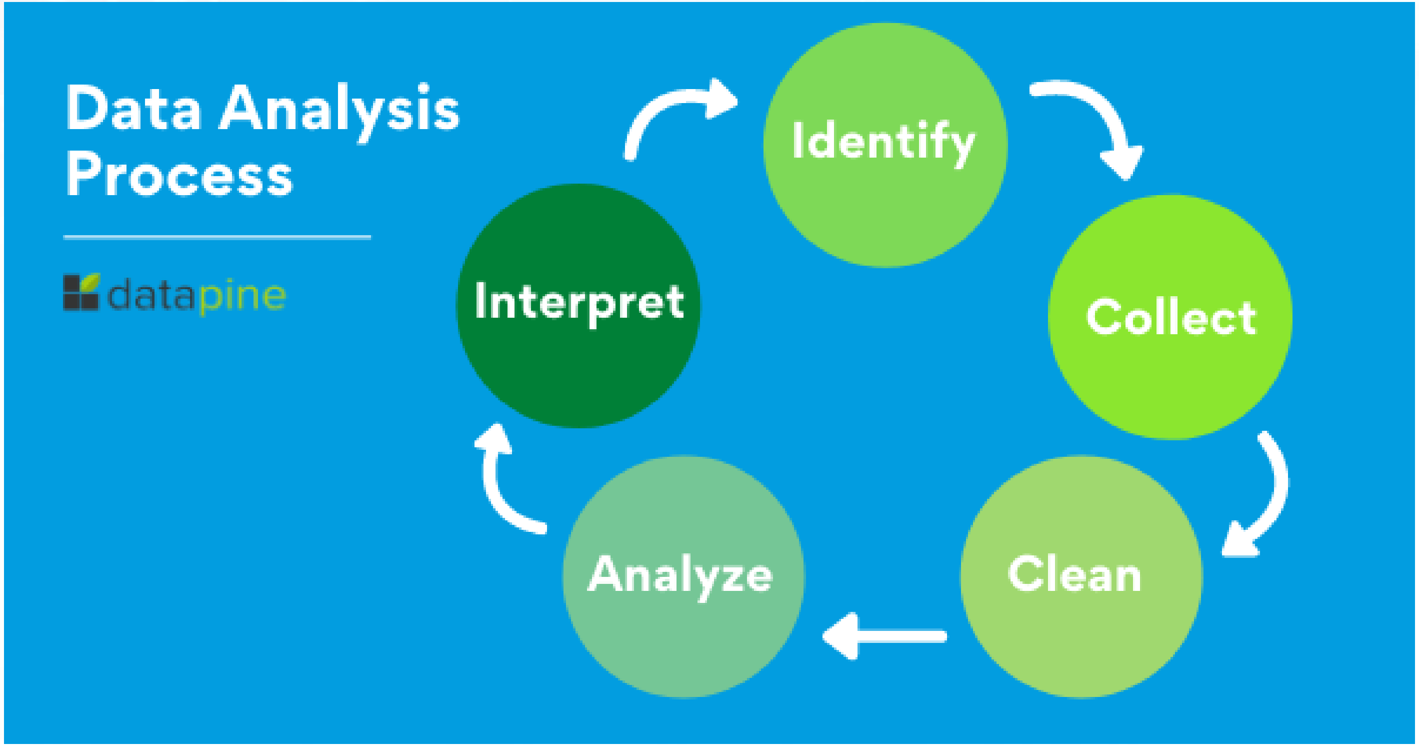
Data analysis

Data analysis is the process of collecting, modeling, and analyzing data to extract insights that support decision-making.

There are several methods and techniques to perform analysis depending on the industry and the aim of the investigation

Basic Algorithms and models

Data analysis process



Basic statistics

Mean and Weighted Average

The mean (also know as average), is obtained by dividing the sum of observed values by the number of observations,

$$\bar{X} = \frac{\sum_{i=1}^{i=n} X_i}{n}$$

the weighted average, which incorporates the standard deviation

$$X_{wav} = \frac{\sum w_i x_i}{\sum w_i}$$

Basic statistics

Median and Mode

Median: The median is the middle value of a set of data containing an odd number of values,

$$\bar{X} = \frac{\sum_{i=1}^{i=n} X_i}{n}$$

Mode: The mode of a set of data is the value which occurs most frequently. The excel syntax for the mode is MODE(starting cell: ending cell).

Basic statistics

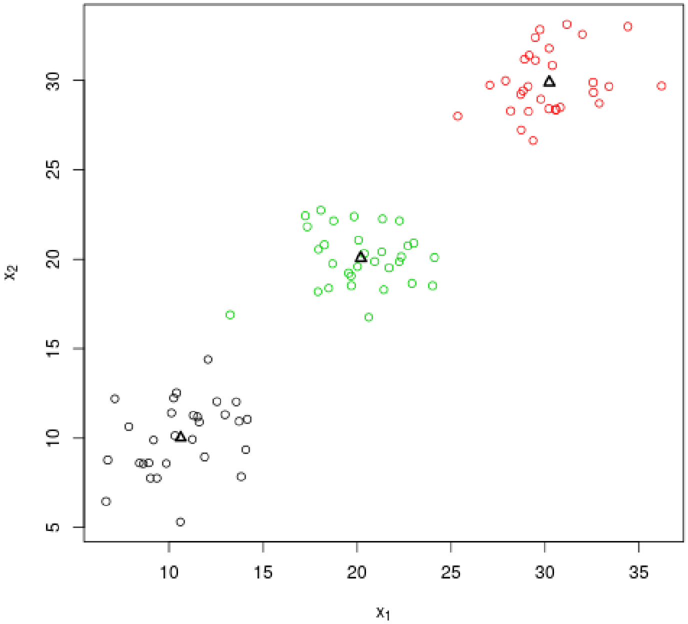
Standard Deviation and Weighted Standard Deviation

The most common measure of variability is the is the standard deviation gives an idea of how close the entire set of data is to the average value.

Datasets with a small standard deviation have tightly grouped, precise data. Data sets with large standard deviations have data spread out over a wide range of values

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{i=n}\left(X_i - \bar{X}\right)^2}$$

Interpretation: Mean without Standard deviation is not interpretable



Data exploration

Data exploration (Exploratory Data Analysis)

Exploratory data analysis (EDA) is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling and thereby contrasts traditional hypothesis testing

Data exploration

Data exploration (Exploratory Data Analysis)

Demo time

The code is available here: datasciencecourses/eda_ntr_data.ipynb at dev · bentechno/datasciencecourses (github.com)

The data I will be using in this article is from India. The data comes from NTR Vaidya Seva (or Arogya Seva) is the flagship health care program of the government of Andhra Pradesh, India, in which lower middle class and low-income citizens of the state of Andhra Pradesh can get free health care for many major illnesses and ailments. A similar program also exists in neighbouring Telangana state
The data is available here: https://1drv.ms/u/s!AidR7LjHHfHvhb0TQcCbd34gSIF59A?e=L5fDaQ

Introduction to Data Science and Machine Learning for Strategic Decision Making

Présenté par:
Mbaye Babacar GUEYE, Ph.D

Advanced Data exploration

Data exploration (Exploratory Data Analysis)

Correlation

PCA

Advanced Data exploration

Principal component analysis (PCA)

- The goal is to find a small set of artificial variables such that as much as possible of the variance of the data is retained.
- The first principal component is a vector in whose direction the variation of the data is the largest
- The next principal component is orthogonal to the first one, and it is in the direction of the second largest variation
- PCA is most suitable when the observations are Normally distributed or at least continuous. The output is hard to interpret if data are discrete.

Advanced Data exploration

Independent component analysis (ICA)

- A statistical technique that represents multidimensional data as a linear combination of nongaussian variables (‘independent components’)
- The variables are statistically as independent as possible
- Similar but stronger than PCA: suitable for non-Normally distributed data
- ICA has many applications in data analysis, source separation, and feature extraction

Advanced Data exploration

Kernel Principal component analysis (PCA)

- The goal is to find a small set of artificial variables such that as much as possible of the variance of the data is retained.
- The first principal component is a vector in whose direction the variation of the data is the largest
- The next principal component is orthogonal to the first one, and it is in the direction of the second largest variation
- PCA is most suitable when the observations are Normally distributed or at least continuous. The output is hard to interpret if data are discrete.

Introduction to Data Science and Machine Learning for Strategic Decision Making

Présenté par:
Mbaye Babacar GUEYE, Ph.D

Tools for Data Science using Python

1. Functions/models and algorithms
2. Package Scikit learn
3. Data visualization tools

Tools for Data Science using Python

Automate Exploratory Data Analysis With These 10 Libraries (analyticsvidhya.com)

Top 10 Exploratory Data Analysis (EDA) libraries you have to try in 2021. (malicksarr.com)

Basics in Machine Learning

Artificial Intelligence

Intelligence ?

According to you?

The ability to **restore / generalize** what we LEARNED.

LEARNING

Classic definition: Learning is the acquisition of **know-how**, through

observation, imitation, testing, repetition, presentation.

Artificial definition: process of adapting **the parameters of a system** to give a desired response to an external input or stimulation.

Artificial Intelligence

Artificial Intelligence ?

The ability of a machine to restore / generalize what it learned through a machine learning process

Machine Learning

Machine Learning ?

Considering a game, where the agent can choose between several actions.

The agent **chooses an action**, **applies** it, and **observes the result**.

It memorizes **the position, the action, the result**.

He *learns* to **predict the result according to the position and the action**: **build a model** (explicit or implicit).

Generalization:

In the presence of a **position**, the agent looks for **which action the prediction of the result is the most favorable.**

And decides to apply this action.

Analytics and ML

New dimension of ML

2000s,

Big data and new computing powers and infrastructures make it possible to explore unprecedented masses of data.

The more you learn, the more you are proficient, the more you know and the better you predict

So what does predict mean ?

IT’S ALL ABOUT MODELING

Modeling in Data Science/ML

Model Definition

ACCORDING TO YOU ?

AN **ABSTRACTION\REPRESENTATION** OF A **REAL WORLD** ASPECT

Modeling in Data Science/ML

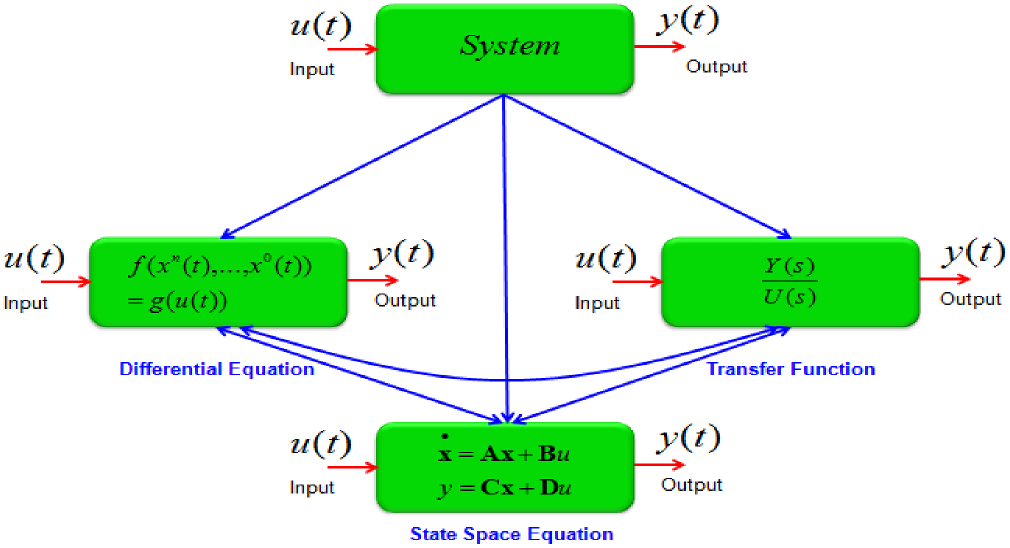
Definition

A mathematical model is a description of a system using **mathematical concepts and language.**

Wikipedia

The most important thing for modeling is to figure out the governing equation for the given situation (problem).

The governing equation varies depending the area/field, so you would need to have the basic understandings on theory for the given problem.



Modeling in Data Science/AI

MATHEMATICAL Modeling

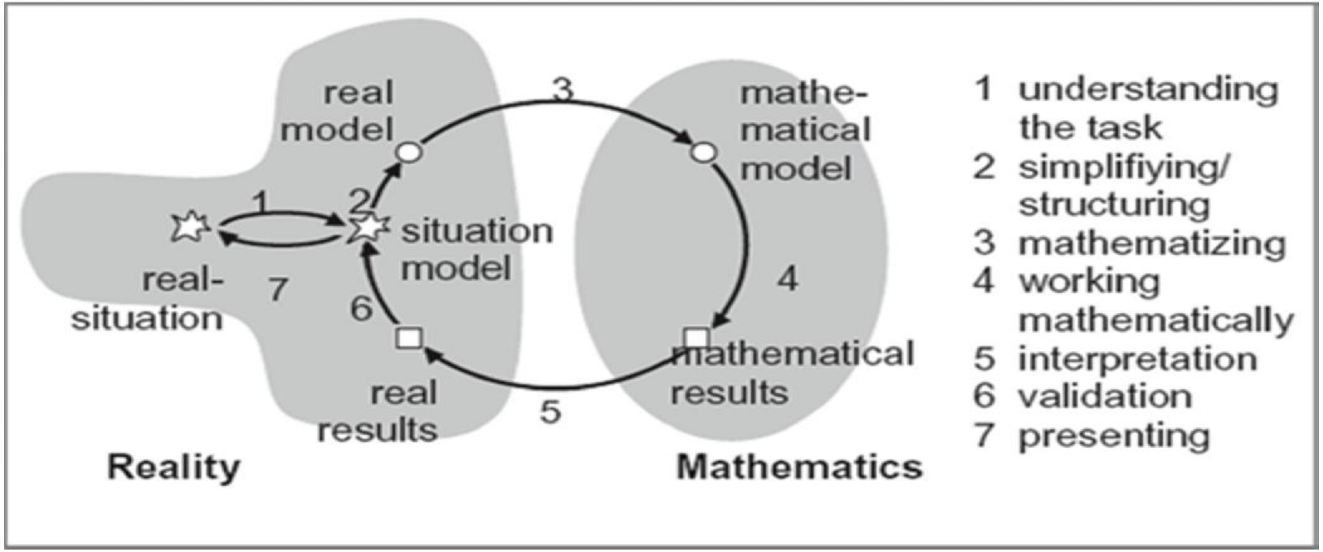


Figure 2. A mathematical modeling process (adapted from Blum and Liess, 2007)

Modeling in Data Science/AI

2 types of modeling:

1. Deterministic modeling ⇔ well-posed problem:
   a. $y = F(x)$, F is well known
   b. in physics $\mathbf{E = \frac{1}{2} mv^2}$, chemistry

2. Non-Deterministic modeling ⇔ ill-posed problem
   1. $y = F(x)$, F is not well known and his parameters may change
   2. data driven model
   3. **Data Science/AI approach**

Basics in Machine Learning

ML is not magic; it's just mathematics.

Learning to predict the result according to the position and the action: build a model

a Mathematical Model.

ML concepts

ML Models: Linear regression

Linear regression is the most representative "machine learning" method to build models for value prediction and classification from training data.

It offers a study in contrasts:

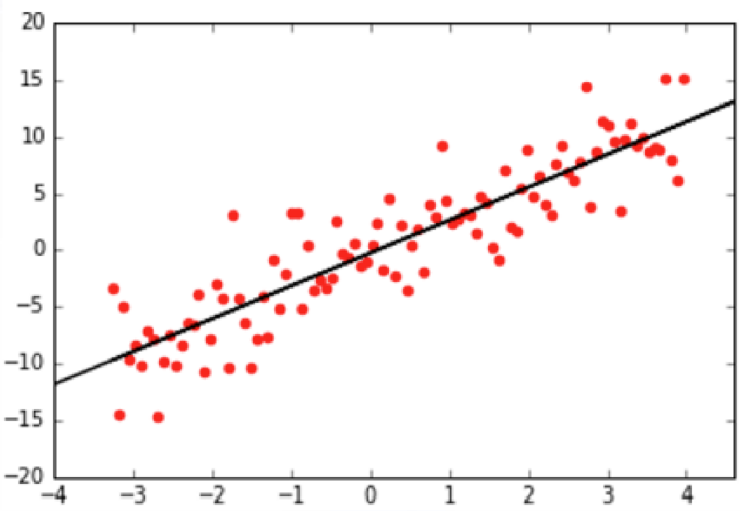
- Linear regression has a **beautiful theoretical foundation** yet, in practice, this algebraic formulation is generally discarded in favor of faster, more heuristic optimization.
- Linear regression models are, by definition, **linear**. This provides an opportunity **to witness the limitations of such models**, as well as develop clever techniques to generalize to other forms.
- Linear regression simultaneously encourages model building with **hundreds of variables, and regularization techniques to ensure** that most of them will get ignored.

ML Models: Linear regression

Linear regression is a bread-and-butter modeling technique that **should serve as your baseline approach to building data-driven models**

Linear regression seeks the line y = f(x) which minimizes the sum of the squared errors over all the training points, i.e. the coefficient vector w that minimizes

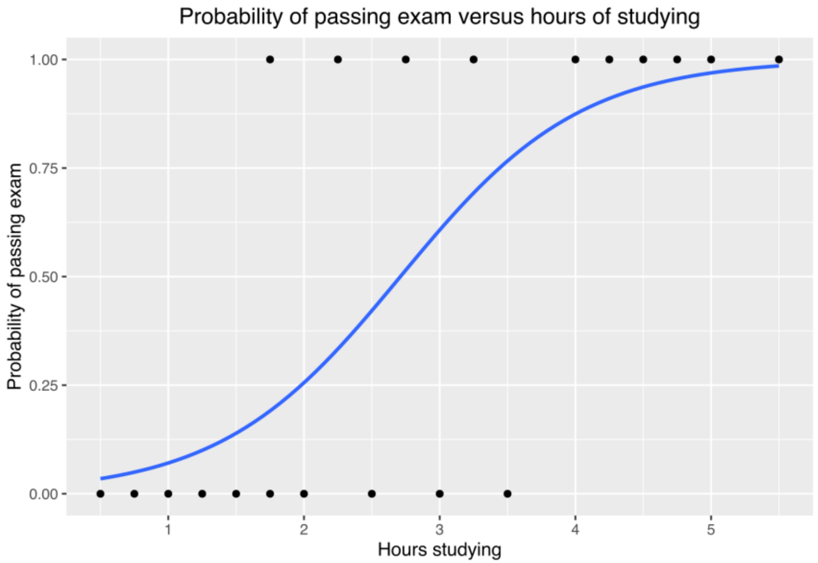
$$\sum_{i=1}^{n}(y_i - f(x_i))^2, \text{ where } f(x) = w_0 + \sum_{i=1}^{m-1} w_i x_i$$



ML Models: logistic regression

Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable(or output), y, can take only discrete values for a given set of features(or inputs), X

The Logistic regression equation can be obtained from the Linear Regression equation.



ML Models: logistic regression

The Logistic regression equation can be obtained from the Linear Regression equation.

We know the equation of the straight line can be written as:

$$y= b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$$\frac{y}{1-y}\text{; 0 for y= 0, and infinity for y=1}$$

But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

ML Models: logistic regression

Type of Logistic Regression: On the basis of the categories, Logistic Regression can be classified into three types:

Binomial: In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
Multinomial: In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"
Ordinal: In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

Data classification and Data clustering

Machine learning types

3 types

1. Supervised learning
2. unsupervised learning
3. reinforcement learning.

Supervised learning,

Supervised learning,

In supervised machine learning, **the program already knows the output**. Note that this is opposite to conventional programming where we feed input to the program and program gives output.

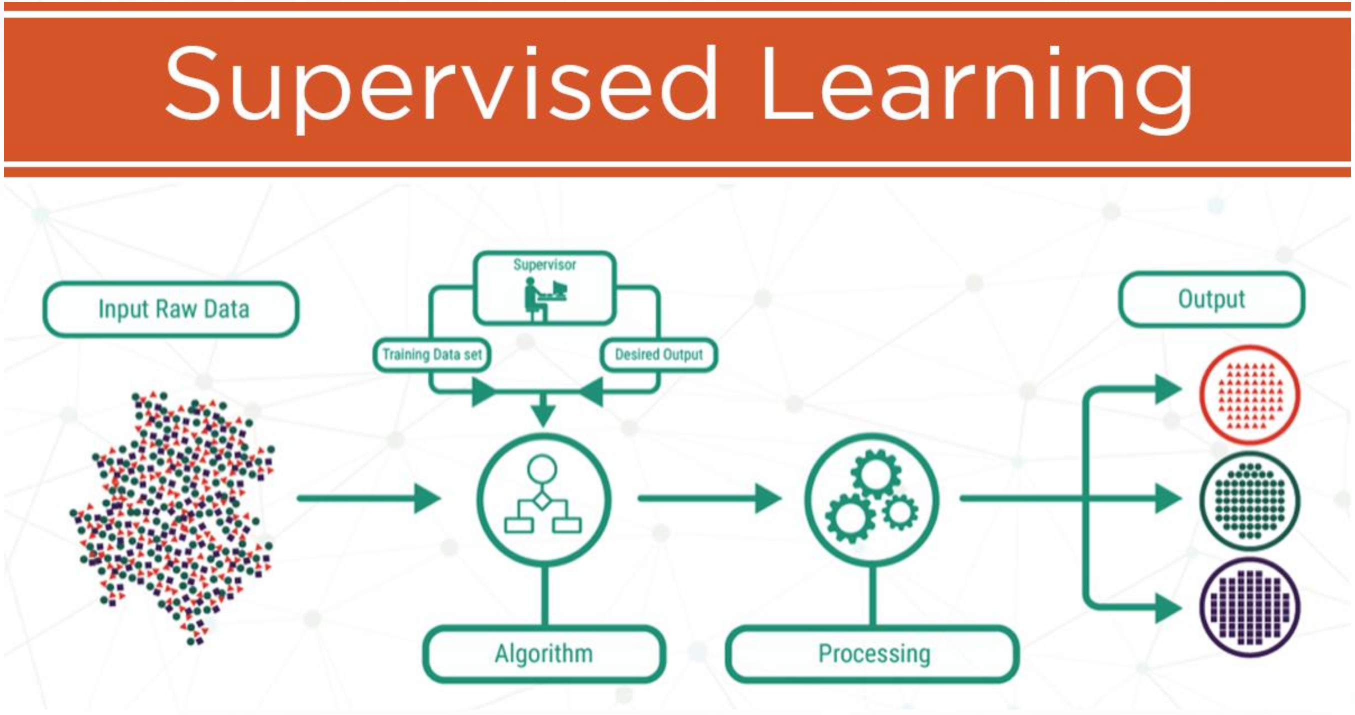
Here in this case, we give input and output at the same time, in order to make program learn that in case of any of this or related input what program has to output.

This learning process is called model building. It means that through provided input and output, the system will have to build a model that maps the input to output.

Regression is a supervised learning method

Supervised learning,

Supervised learning,



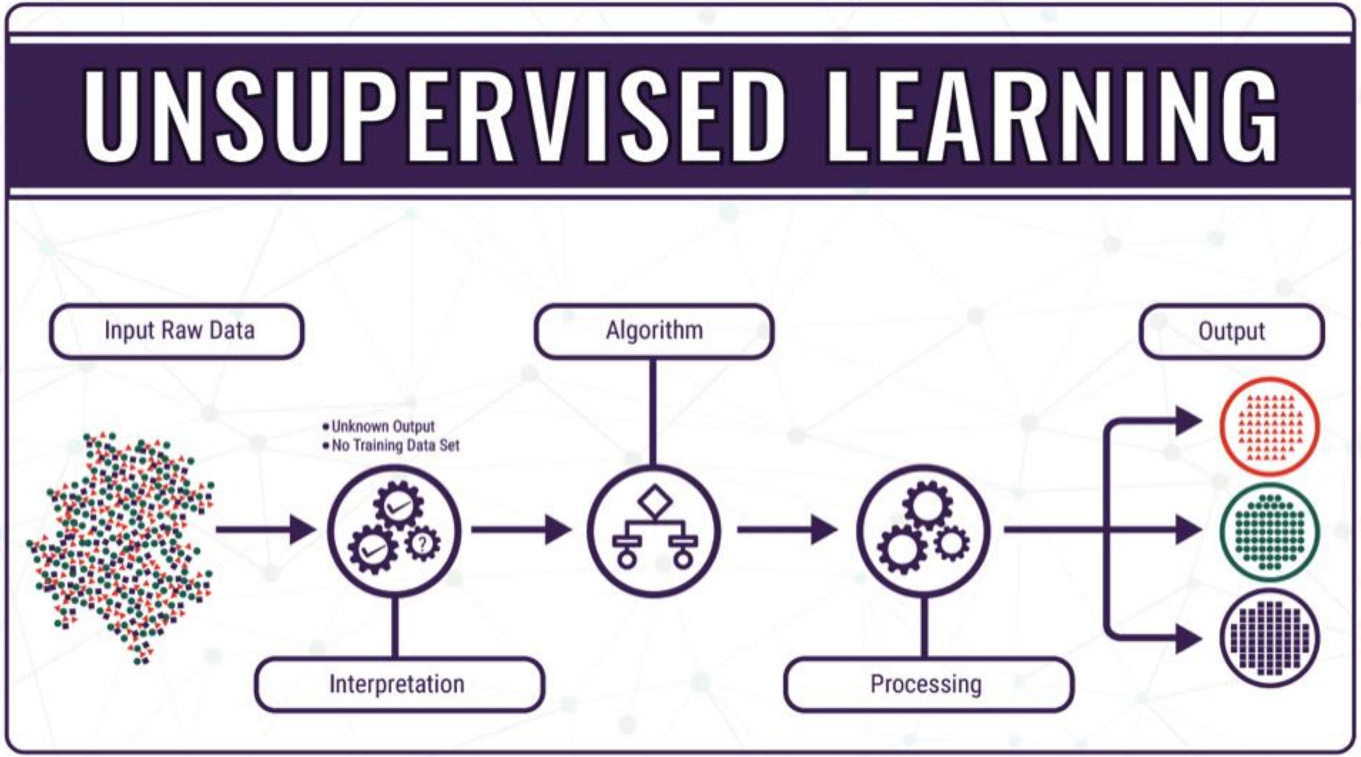
Supervised learning,

Unsupervised learning,

We may have applications where the class labels are not known. The scenario is called unsupervised learning. n unsupervised learning, the machine learns from the training dataset and groups the objects on the basis of similar features, e.g. we may group the fruits on the basis of colors, weight, size, etc. This makes it challenging as we do not have any heuristics to guide the algorithm. However, it also opens the new opportunities to work with the scenarios where the outcomes are not known beforehand. The only thing available is the set of operations available to predict the group of the unknown data.

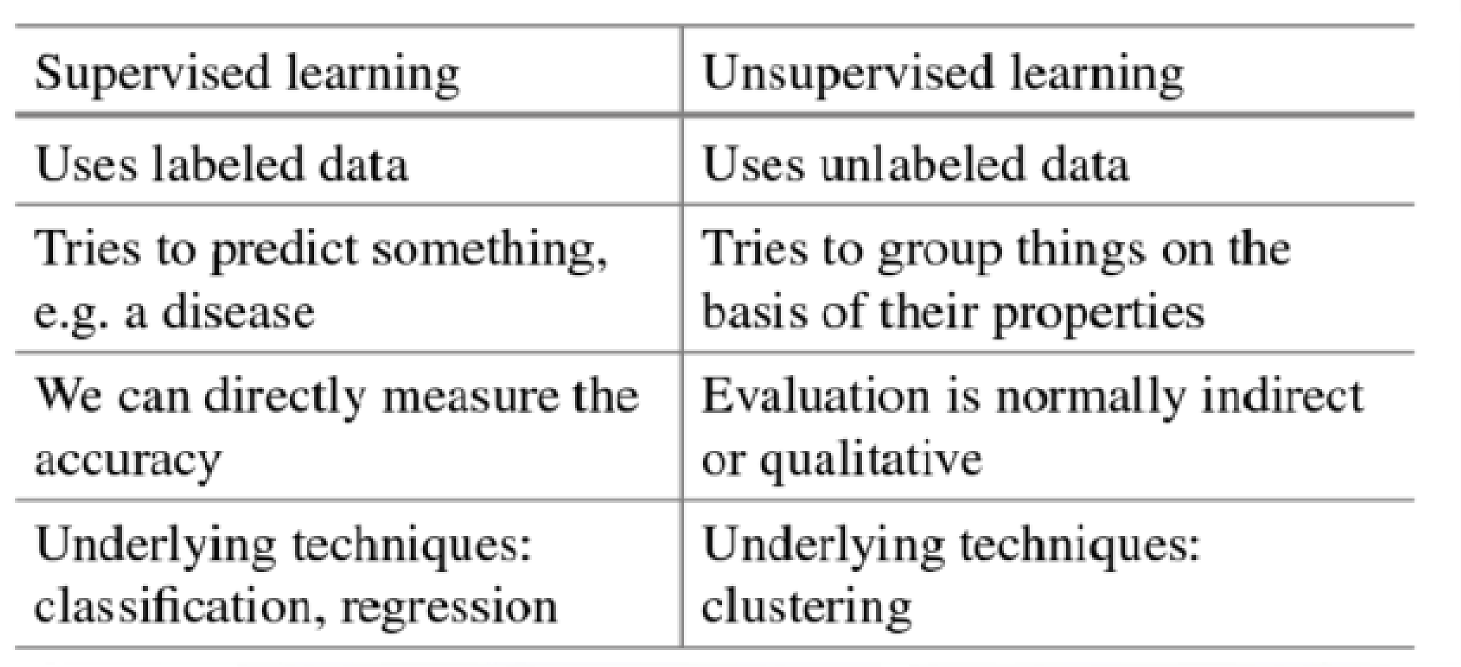
Supervised learning,

Unsupervised learning,



Supervised learning,

Unsupervised learning vs supervised learning,



Supervised learning	Unsupervised learning
Uses labeled data	Uses unlabeled data
Tries to predict something, e.g. a disease	Tries to group things on the basis of their properties
We can directly measure the accuracy	Evaluation is normally indirect or qualitative
Underlying techniques: classification, regression	Underlying techniques: clustering

ML concepts

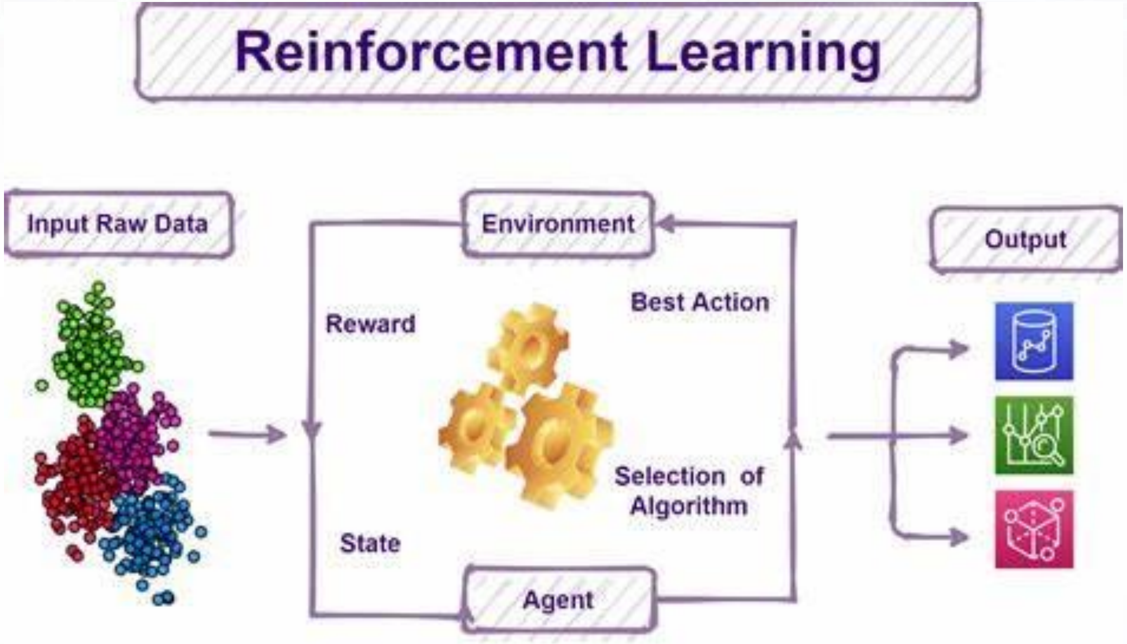
Reinforcement Learning

It is the process of training a model so that it could make series of decisions. In reinforcement learning, a machine agent interacts with its environment in uncertain conditions in order to perform some actions.

The agent is guided in order to achieve the intended output with the help of rewards and penalties.

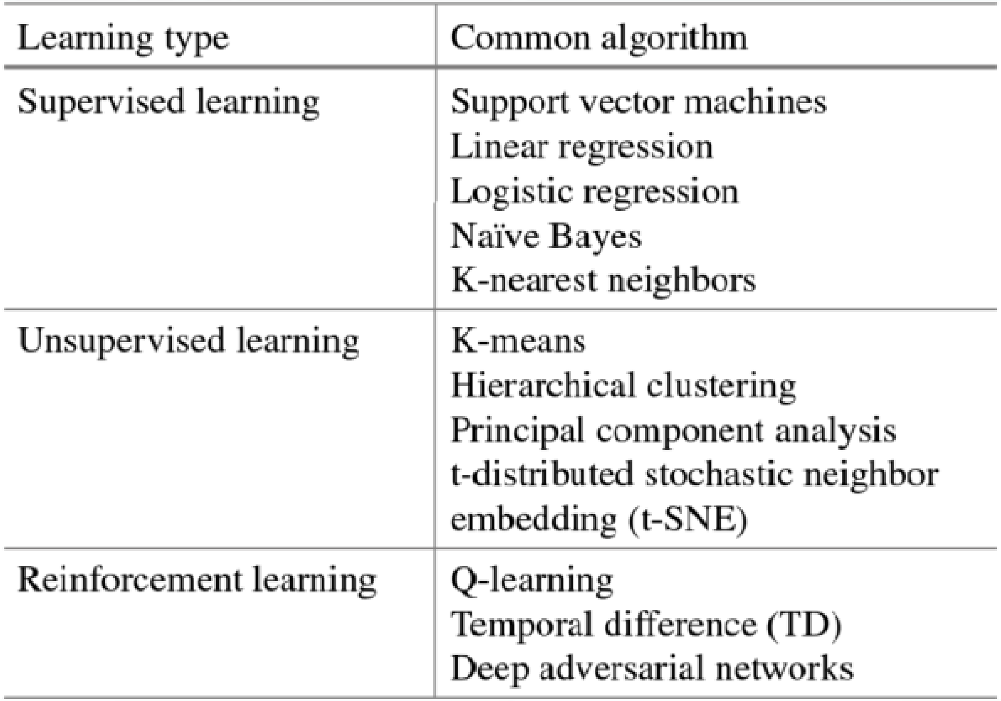
The overall goal is to increase the total number of rewards. The designer sets the rewards policy. Now it is up to the model to perform the actions in order to maximize the rewards.

Reinforcement Learning



ML concepts

Common algorithms in each type of learning



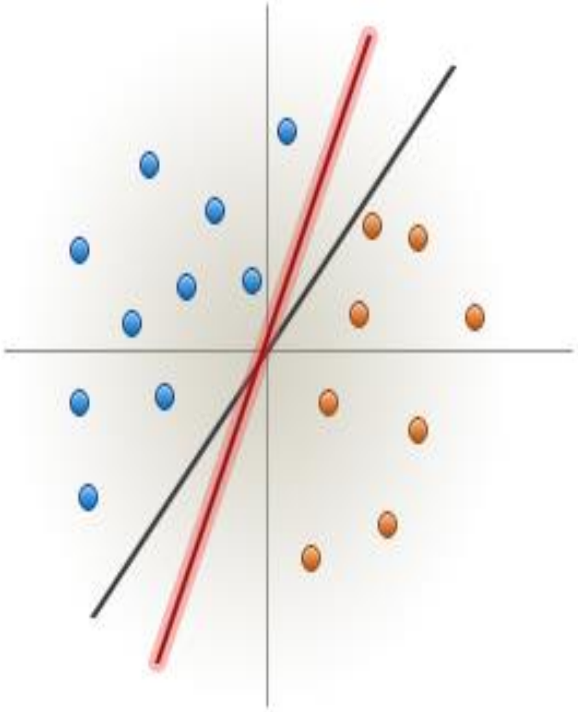
Learning type	Common algorithm
Supervised learning	Support vector machines Linear regression Logistic regression Naïve Bayes K-nearest neighbors
Unsupervised learning	K-means Hierarchical clustering Principal component analysis t-distributed stochastic neighbor embedding (t-SNE)
Reinforcement learning	Q-learning Temporal difference (TD) Deep adversarial networks

Classification and clustering

CLASSIFICATION

Constituer des groupes d'objets homogènes et différenciés, des groupes d'objets tels que :

- les objets soient les plus similaires possibles au sein d'un groupe (critère de compacité),
- les groupes soient aussi dissemblables que possible (critère de séparabilité),



La ressemblance ou la dissemblance étant mesurée sur l'ensemble des variables descriptives.

Classification : Calcul de la dissimilarité

Dissimilarité par Calcul des distances:

elle est définie de I x I → R+ par

1) $\forall\ i \in I,\ d(i,i)$=0

2) $\forall\ i,\ \ i' \in I$ $\qquad d(i,i')$=$d(i',i)$

Les algorithmes de classification ont pour point de départ une mesure des distances entre les objets. La plus utilisée:

la distance euclidienne usuelle:

$$d^2(i,i') = \sum_{j=0}^{n}(x_{ij} - x_{i'j})^2$$

$d^2(i,i')$=distance entre i et i', j est la variable considérée

Classical classification algorithm

k-nearest-neighbors classification

- This is supervised learning
- In your training data, you are given observations and their class labels
- For a new observation, find the k nearest neighbors among the training observations using a suitable distance function
- Classify the new observation by the major vote of those *k* labeled observations

Classical classification algorithm

k-nearest-neighbors algorithm

Il affecte un élément à la classe la plus représentée parmi les K plus proches éléments de la base d'apprentissage. L'algorithme des K-Plus Proches Voisins est donc le suivant :

```
KNN(trainX,trainY,testX,k)
for each observation $\vec{x}$ in testX:
    compute the Euclidean distance between x and all observations $\overrightarrow{x_i}$ in trainX: $d(\vec{x},\overrightarrow{x_i})$
    find the k observations having the smallest distance to x
    find the most frequent class label in these k observations
    output the class label for observation x
```

Classical classification algorithm

Maximum likeliwood

Il se base sur l'analyse statistique de la distribution des éléments de la base d'exemples pour définir des probabilités d'appartenance à chaque classe. Le nouvel objet est assigné à la classe pour laquelle la probabilité d'appartenance est la plus élevée.

Avantages: une classe, un degré de confiance lié à ce choix.

Classical classification algorithm

Maximum likeliwood

Type de distribution des éléments de la base des exemples. Dans le cas d’une distribution gaussienne, on cherche à maximiser pour chaque nouvel objet $\vec{x} \in$ R^{m} la probabilité d’appartenance à la classe y_i :

$$arg\max_{y_i} P(\vec{x}/y_i) = arg\max_{y_i} \frac{1}{\sqrt{2\pi^m|Q_i|}} \exp\left(-\frac{1}{2}\left(\vec{x}-\overrightarrow{\mu_i}\right)^T . Q_i^{-1} . \left(\vec{x}-\overrightarrow{\mu_i}\right)\right)$$

Où $\overrightarrow{\mu_i}$ et Qi désigne respectivement la moyenne et la matrice de covariance associées à la classe y_i.

Classical classification algorithm

Maximum likeliwood

Il se base sur l'algorithme du Maximum de Vraisemblance est donc le suivant :

– On calcule des statistiques pour chaque classe de la base d'apprentissage : moyennes $\overrightarrow{\mu_i}$ *et matrices de covariance* Q_i,

– Pour chaque point :

✓ *on calcule les probabilités d'appartenance à chaque classe :* $P(\vec{x}/y_i)$

✓ *on assigne le point à la classe ayant la plus grande probabilité.*

Classical classification algorithm

Decision trees

- Supervised learning: prediction and classification
- Suitable especially when the data contain both numerical and categorical variables
- Then it is difficult to construct a distance function for regression or k-nearest-neighbors etc.
- A decision tree can help

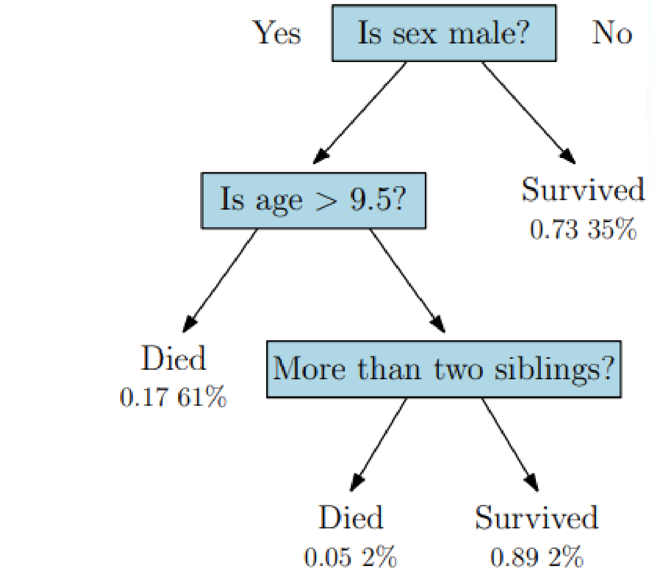
Classical classification algorithm

Decision trees

A binary branching structure used to classify an arbitrary input vector X.

Each node in the tree contains a simple feature comparison against some field xi ∈ X, like "is xi ≥ 23.7?"

The result of each such comparison is either true or false, determining whether we should proceed along to the left or right child of the given node.



Clustering

Objectives

- This is unsupervised learning
- Task: Divide the observations into groups (clusters) so that “similar”

observations go together

- Must choose:
  - – Which method to use?
  - – How to measure the goodness of clustering?
  - – Number of clusters K

Classical Clustering algorithm

2 common types of clustering methods

- Partition the data directly into K clusters (example: K-means)

- Hierarchical methods: iteratively merge the observations into groups, and merge similar groups together, until you have K group

Classical Clustering algorithm

K-means clustering

Start from a random clustering, and iteratively change the clustering so that the goodness of clustering (=similarity of observations in a cluster) increases at each step. Continue until the clustering does not change.

Classical Clustering algorithm

K-means clustering

Méthode simple de classification automatique qui sépare les données $\overrightarrow{x_j}$ en K classes en minimisant : $U = \sum_{i=1}^{k} \sum_{\overrightarrow{x_j} \in C_i} \left\| \overrightarrow{x_j} - \overrightarrow{\mu_i} \right\|^2$ (1) où $\overrightarrow{\mu_i}$ désigne la moyenne des éléments de la classe Ci

Problème (a): K à fixer, convergence vers un minimum local et n'est pas adapté au cas de classes de structures non convexes et de tailles différentes.

Solution pour (a): ajout d'un terme entropique défini par

$= -\sum_{i=1}^{k} pi.\log(pi)$ où $pi = \frac{Card(C_i)}{N}$ représente la probabilité qu'une donnée de l'ensemble de taille N appartienne à la classe Ci

Donc (1) devient $U = \sum_{i=1}^{k} \sum_{\overrightarrow{x_j} \in C_i} (\left\| \overrightarrow{x_j} - \overrightarrow{\mu_i} \right\|^2 - \alpha.\log(pi))$ (2)

Classical Clustering algorithm

K-means clustering algorithm

– Initialisation : un grand nombre de classe K et positionnement aléatoire des K centroïdes sur des points de l'ensemble,

– Itérations : tant que les centroïdes changent de position :

- *on assigne chaque point à la classe du centroïde le plus proche au sens de la distance définie par :* $d(\vec{x},\vec{\mu_i}) = (\|\vec{x}-\vec{\mu_i}\|^2 - \alpha . log(pi))$
- *si une classe ne comporte plus d'élément, elle est éliminée*
- *on recalcule le centre de gravité* $\vec{\mu_i}$ *des éléments de chaque classe*

La partition obtenue par l'algorithme des k-moyennes dépend des représentants initialement choisis.

Exécuter l'algorithme des k-moyennes (k et d étant fixés) avec des initialisations différentes, et on retient la meilleure partition.

La qualité d'une partition est mesurée par la quantité : $D = \sum_{i=1}^{k} \sum_{\vec{x_j} \in C_i} d(\vec{x},\vec{\mu_i})$ qui mesure la cohésion des classes obtenues.

Classical Clustering algorithm

2 common types of clustering methods

- Partition the data directly into K clusters (example: K-means)
- Hierarchical methods: iteratively merge the observations into groups, and merge similar groups together, until you have K group

DEMO

DEMO TIME

Demo

datasciencecourses/1BlogPostMonthKmeans.ipynb at dev · bentechno/datasciencecourses (github.com)

More context here: Unsupervised Learning with k-means part 1 | by Mbaye Babacar GUEYE, Ph.D | Analytics Vidhya | Medium

The data is available here: https://1drv.ms/u/s!AidR7LjHHfHvhb0j-Se-t8ujxdyxyw?e=AWAh4j

Design Principles and Best Practices

Sources:

THE Data Science Design MANUAL, Steven S. Skiena

Logistic Regression in Machine Learning - Javatpoint

https://dev.to/luminousmen/what-are-the-best-software-engineering-principles--3p8n

https://medium.com/agileactors/4-basic-principles-of-software-engineering-787b495c2870

https://luminousmen.com/post/what-are-the-best-engineering-principles

The Data Engineering Cookbook, Mastering The Plumbing Of Data Science, Andreas Kretz, May 18, 2019

TDWI CHECKLIST REPORT: FIVE DM AND ANALYTICS BEST PRACTICES FOR BECOMING DATA-DRIVEN

http://softwaretestingfundamentals.com

https://www.callicoder.com/software-development-principles/

THANKS AND QUESTIONS

Mbaye Babacar Gueye, Ph.D

Advanced Analytics | Big Data | Artificial Intelligence

Sujets de prédilection : #datajobs, #datascience, #dataanalytics, #advancedanalytics et #artficialintelligence

Montréal et périphérie · Coordonnées

Slalom

Université Pierre et Marie Curie (Paris VI)

Contact:

gueye@mbayebabacar.me,

@mbayebgueye

www.linkedin.com/in/mbgueye/