Introduction to Machine Learning

Mahammad Valiyev

22.01.2022

Contents and timeline

1. Introduction to Machine Learning and use cases in O&G (Jan 2)
2. Overview of Machine Learning algorithms (Jan 8)
3. Machine Learning Life Cycle (Jan 15)
4. Overview of resources, skill sets, job types, general advice (Jan 22)

Part 4:

Overview of resources, skill sets, job types, general advice

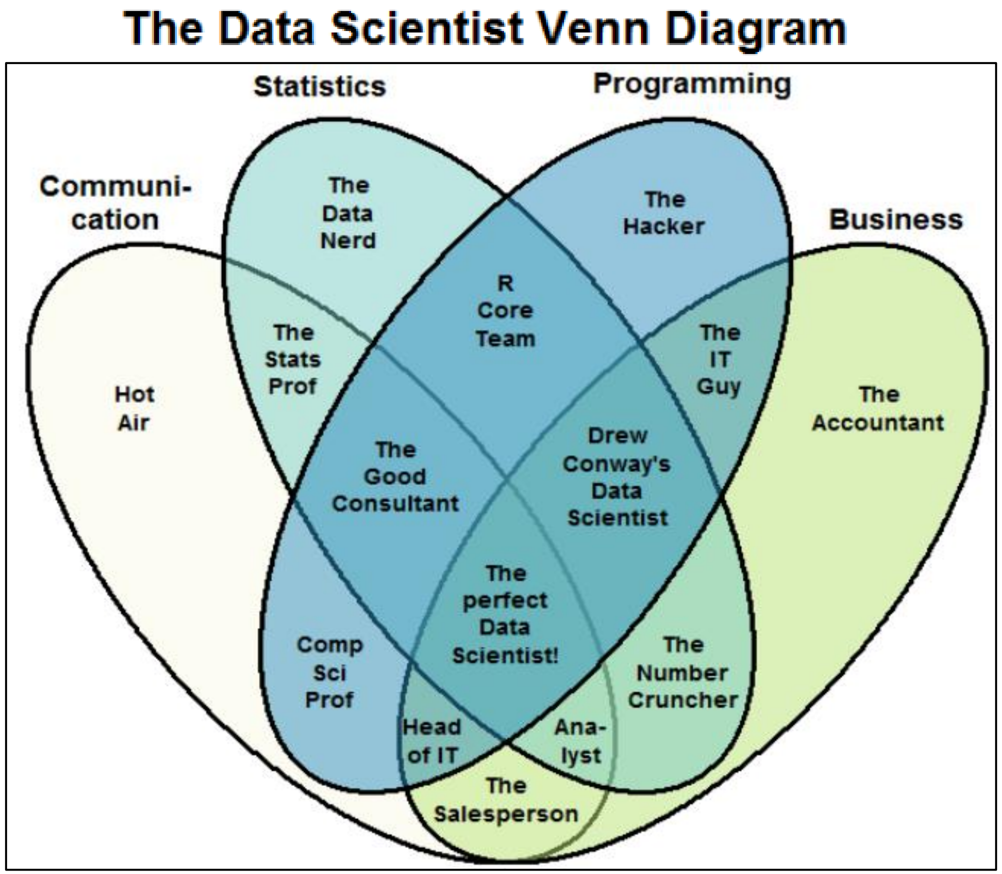
The big picture

Technical skills:

- Mathematics (including Statistics)
- Programming
- Machine/Deep Learning
- Domain knowledge

Non-technical skills:

- Communication (verbal & written)
- Curiosity & drive & passion



Credit: Wikimedia

Mathematics: Linear Algebra

- Branch of math, dealing with vectors, matrices
- Lots of applications in many engineering disciplines
- Why do you need linear algebra for ML?
  - Data for ML is represented with vectors, matrices, tensors
  - Theory for ML/DL is expressed with vectors, matrices
- Basics are enough to get started:
  - Notion of a scalar, vector, matrix, tensor
  - Basic arithmetic operations: e.g. addition, multiplication
  - Matrix multiplication properties and special matrices
  - Special operations: inverse, transpose
- Resources:
  - MIT OpenCourseWare, 18.06 SC
  - Khan Academy, Linear Algebra

Scalar

1

Vector

$\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$ or $\begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$

Matrix

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}$$

Tensor

$$\begin{bmatrix} [1 \;\; 2] & [3 \;\; 4] \\ [5 \;\; 6] & [7 \;\; 8] \\ [9 \;\; 0] & [1 \;\; 2] \end{bmatrix}$$

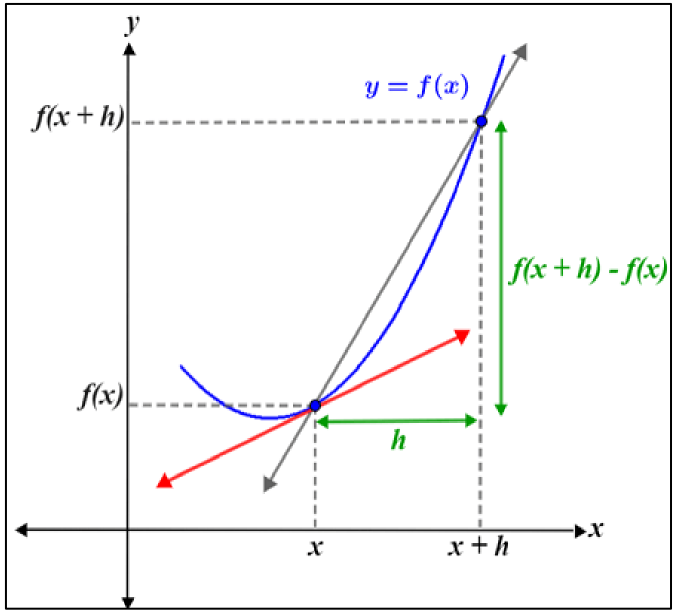
$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} + \begin{bmatrix} 9 & 8 & 7 \\ 6 & 5 & 4 \\ 3 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 1+9 & 2+8 & 3+7 \\ 4+6 & 5+5 & 6+4 \\ 7+3 & 8+2 & 9+1 \end{bmatrix}$$

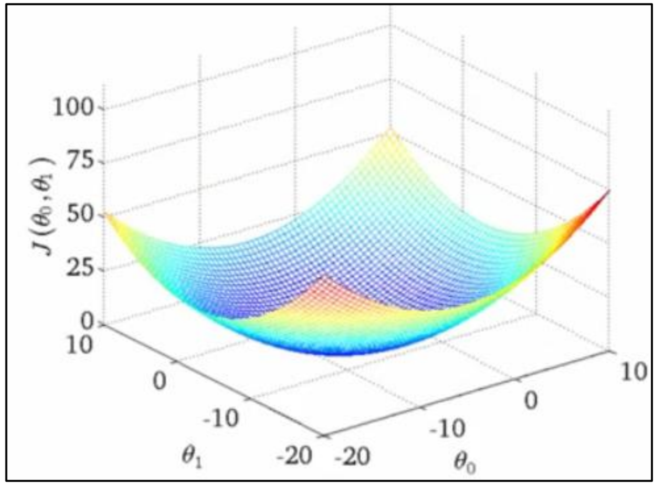
$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \times \begin{bmatrix} w & x \\ y & z \end{bmatrix} = \begin{bmatrix} aw+by & ax+bz \\ cw+dy & cx+dz \end{bmatrix}$$

Mathematics: Multivariable Calculus

$$\lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

- Branch of math involving study of continuous change
- Two major branches: differential and integral
- Multivariable calculus is extension of single variable calculus to multiple variables
- ML theory needs mostly differential calculus
- Why do you need calculus for ML?
  - Internal workings of algorithms (backpropagation for DL)
  - Optimization of objective functions
- Basics are enough to get started:
  - Notion of a derivative, partial derivative
  - Differentiation rules
  - Calculus on vectors, e.g. gradient
- Resources:
  - MIT OpenCourseWare, 18.02 SC
  - Khan Academy, Multivariable Calculus





Mathematics: Probability

- Branch of math involving study and quantification of uncertainty
- Lots of applications in science, engineering, industry for modeling and risk assessment
- Why do you need probability for ML?
  - Some algorithms are directly designed based on probabilistic laws
  - Models are trained with probabilistic frameworks
  - Models are tuned with a probabilistic framework
  - Models are evaluated with probabilistic measures
- Basics are enough to get started:
  - Notion of a probability and probability axioms
  - Conditioning and Bayes theorem
  - Idea of random variable, PDF, CDF
- Resources:
  - MIT OpenCourseWare, 6.041 SC
  - Khan Academy, Probability and Random variables

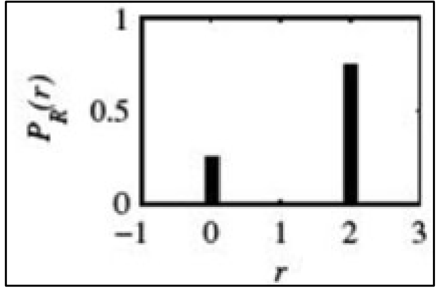
$$P\,(\text{heads}) = \frac{1}{2} = 0.5$$

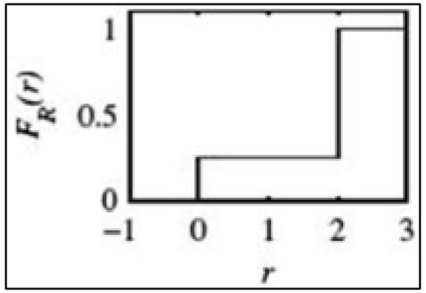
Axiom 1 : $P(A) \leq 1$

Axiom 2 : $P(S) = 1$

Axiom 3 : $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \Phi$

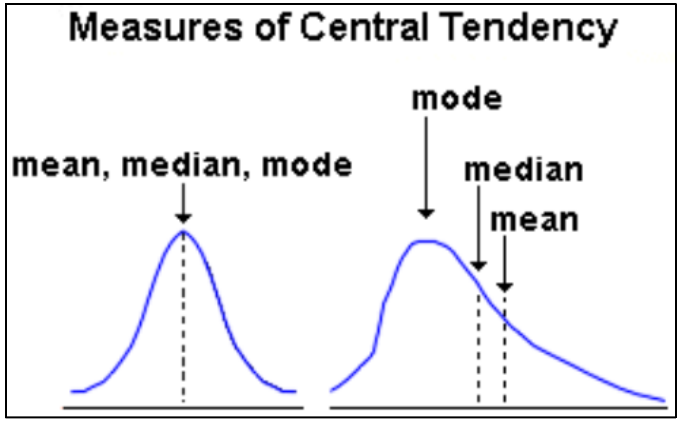
$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

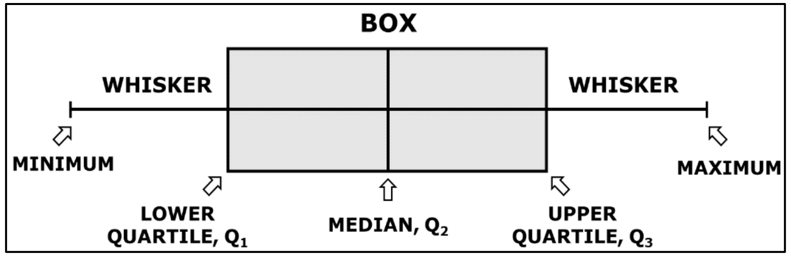


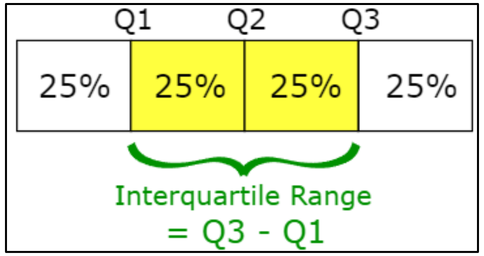


Mathematics: Statistics

- Branch of math concerning collection, analysis, interpretation, visualization of data
- 2 major subfields:
    - descriptive (summarize feature of data: e.g. mean)
    - inferential (infer properties of distribution using data: e.g. mean from sample data)
- Why do you need statistics for ML?
    - Statistics offers a collection of tools to deal with all aspects of data
        - Collection, cleaning, visualization, modeling
    - Foundations of many models are based on statistics
    - Statistics concepts and terminologies are used in ML
- Basics are enough to get started:
    - Descriptive statistics (summarizing data), e.g. mean, variance
    - Data visualization techniques
- Resources:
    - Khan Academy, Statistics
    - MIT OpenCourseWare, 18.650







Mathematics: Optimization

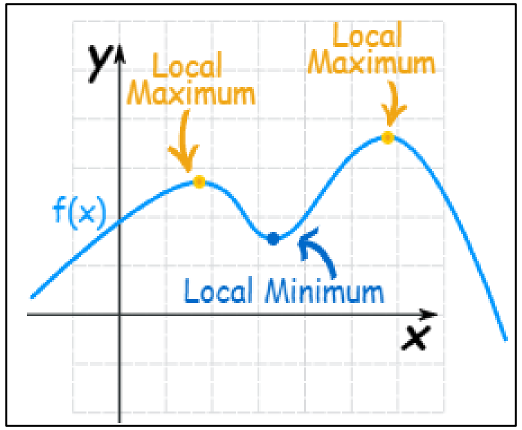
- Branch of math involving study of algorithms to determine maxima or minima of functions under or without constraints

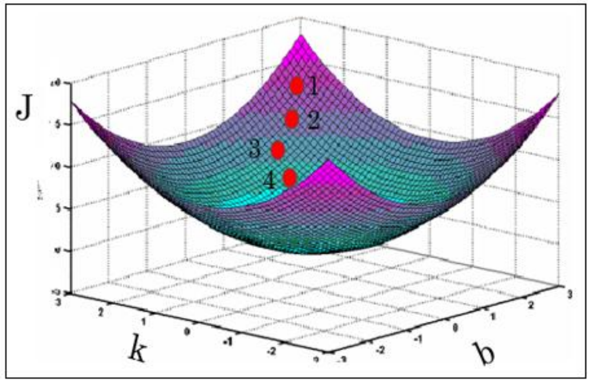
- Being very applied field, there are lots of applications in science, engineering, economics

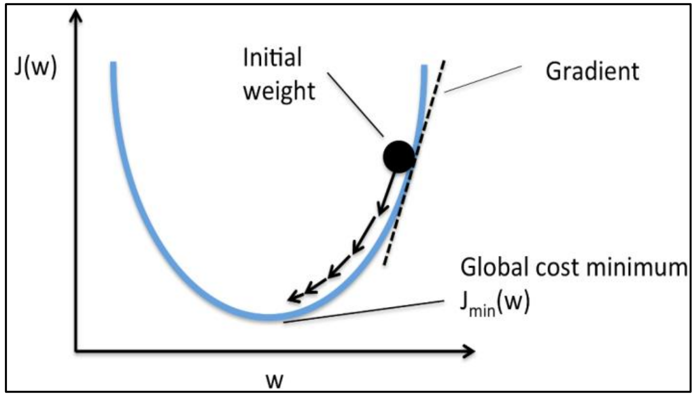
- Why do you need statistics for ML?
  - All Machine Learning problems are optimization problems
  - Each ML algorithm has its own objective/cost function

- Basics are enough to get started:
  - Notions of objective function, maxima, minima
  - Gradient descent

- Resources
  - No open-source resource to suggest







$$\boxed{\textbf{Model: } \mathrm{y}(\mathrm{k},\mathrm{b}) = kx + b}$$

$$\boxed{\textbf{Cost function: } J(k,b) = \frac{1}{2N}\sum_{i=1}^{N} \left((kx_i + b) - y_i\right)^2}$$

Programming

- Basics of programming (a course in any programming language is okay):
  - Conditionals (if/else), loops (for, while)
  - Defining functions
  - Basic data structures: e.g., strings, lists, arrays
- Python for data analysis
  - Jupyter notebook
  - Operations with vectors and matrices: numpy
  - Data manipulation (clean, merge, reshape etc) and exploration: pandas
  - Data visualization: matplotlib
- Python for Machine Learning:
  - scikit-learn: for Machine Learning
  - Keras, Pytorch for Deep Learning
- Resources:
  - Intro to CS and Programming using Python, edx.org
  - Python for data analysis (book), Wes Mckinney
  - Keras/Pytorch documentation, https://keras.io/, https://pytorch.org/

Machine Learning algorithms

- **Supervised**
  - Regression
    - Linear regression and extensions (ridge, lasso)
    - K-nearest neighbors
    - Support vector machine and its extensions (kernels)
    - Decision trees and its extensions (ensemble methods)
  - Classification
    - Logistic regression
    - K-nearest neighbors
    - Support vector machine and its extensions (kernels)
    - Naïve Bayes
    - Decision trees and its extensions (ensemble methods)

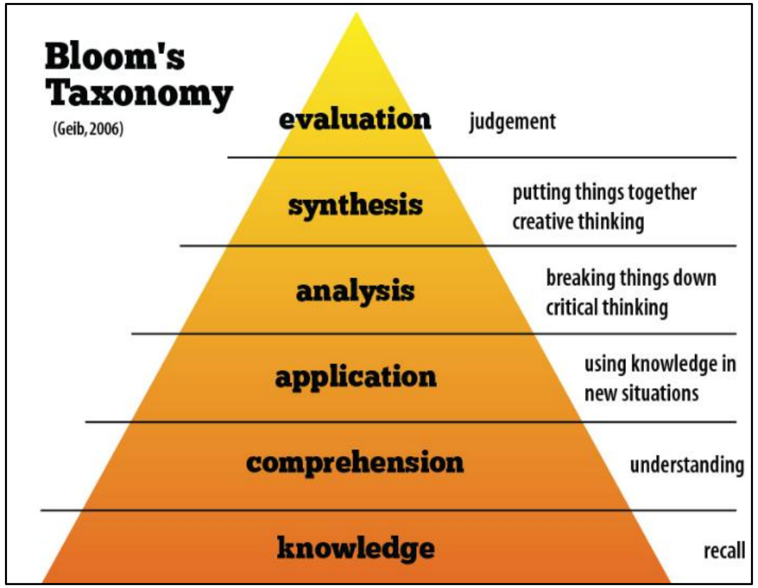
- **Unsupervised**
  - Clustering
    - K-means
  - Dimensionality reduction
    - Principal component analysis

- **Reinforcement Learning**

- **Deep Learning**: can be used for supervised, unsupervised and reinforcement learning
  - Multilayer perceptron (regression and classification)
  - Autoencoders (dimensionality reduction)
  - Convolutional neural networks (regression and classification)
  - Recurrent neural networks (regression and classification)
  - Generative adversarial neural networks (unsupervised: new data generation)
  - Transformers (regression and classification)

Multiple levels of understanding of an algorithm

- Basic
  - Type of ML task used to solve
  - Rough idea/intuition about how algorithm works
- Intermediate
  - The gist of mathematics underlying the algorithm
  - Intuition behind pros and cons of algorithm
  - Application of algorithm to solve a problem using a library
- Advanced
  - The details of mathematics underlying the algorithm
  - Detailed knowledge of advantages and limitations
  - Implementation from scratch



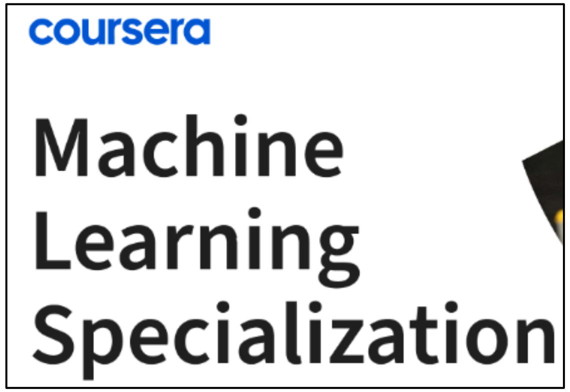
Bloom's taxonomy, Credits: psia-nw.org

Resources for learning Machine/Deep Learning

- **Online courses** (Intermediate):
  - Machine Learning Specialization, Coursera
  - Deep Learning Specialization, Coursera

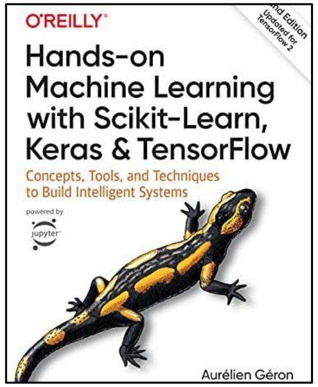
- **Books** (Intermediate):
  - Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, Aurélien Géron
  - Python Machine Learning, Sebastian Raschka
  - Introduction to Statistical Learning, Gareth James et al.,

- **Youtube channel** :
  - GeostatsGuyLectures, Michael Pyrcz (for ML for Petroleum Engineering)



coursera

Machine Learning Specialization



O'REILLY

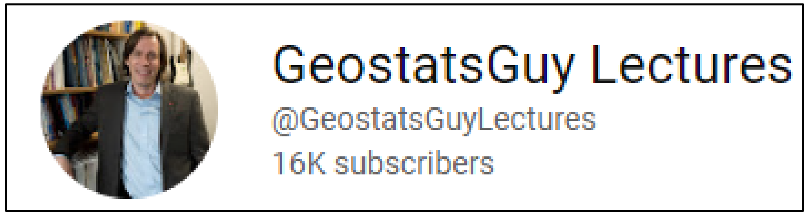
2nd Edition Updated for TensorFlow 2

Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow

Concepts, Tools, and Techniques to Build Intelligent Systems

powered by jupyter

Aurélien Géron



GeostatsGuy Lectures

@GeostatsGuyLectures

16K subscribers

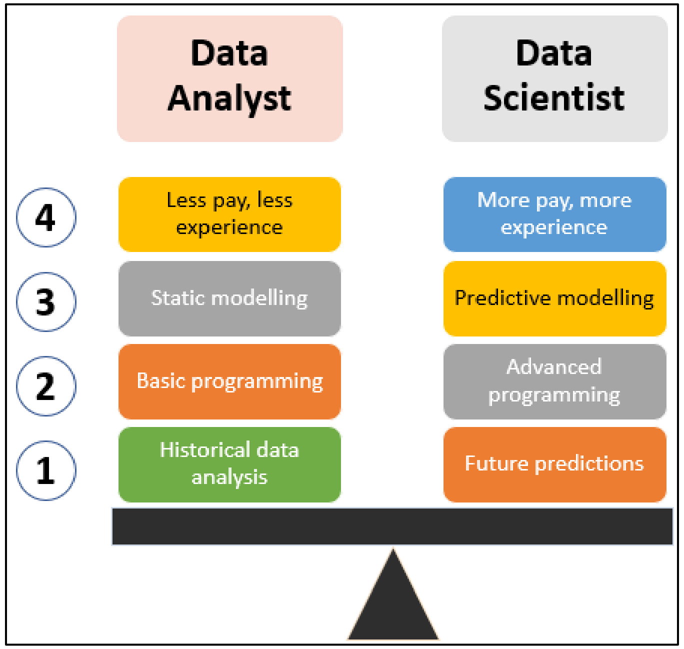
Some types of Data Science Jobs

Data Analyst:

- Focus is mostly on data analysis & visualization and some basic predictive modeling
- Skills: basic mathematics and programming skills are needed
- Deliverable: mostly reports, presentations, dashboards

Data Scientist:

- Focus is on predictive modeling
- Skills: intermediate math and programming skills are needed
- Deliverable: mostly predictive models



Credits: towardsdatascience.com

Recap

1. Introduction to Machine Learning and use cases in O&G
2. Overview of Machine Learning algorithms
3. Machine Learning Life Cycle
4. Overview of resources, skill sets, job types, general advice

Thank you