Introduction to Machine Learning

Readings
Chapter-1: Nina Zumel & John Mount (2019). Practical Data Science with R, Second Edition. Manning Publications. https://multisearch.mq.edu.au/permalink/f/1od1ft6/TN_safari_s9781617295874
Chapter-1 and 2: Ozdemir, S. (2016). Principles of data science : Learn the techniques and math you need to start making sense of your data / Sinan Ozdemir. https://multisearch.mq.edu.au/permalink/f/i7uiug/MQ_ALMA51204622540002171
Chapter-1 and Chapter-2: Boehmke, Brad and Greenwell, Brandon M, Hands-on machine learning with R (CRC Press, 2019).https://bradleyboehmke.github.io/HOML/
Chapter-1: Sunila Gollapudi. (2016). Practical Machine Learning. Packt Publishing.https://multisearch.mq.edu.au/permalink/f/1lmkbbh/TN_pq_ebook_centralEBC4520739
Chapter-9 and Chapter 10: Statistics and Data Analysis for Financial Engineering with R examples Second Edition https://multisearch.mq.edu.au/permalink/f/i7uiug/MQ_ALMA51175555040002171

Contents

Part 1

Introduction to Machine Learning

Data science is a superset of Machine learning, data mining, and related subjects. It extensively covers the complete process starting from data loading until production.

1.1 What is Machine Learning?

Main reference: Chapter-1 of Gollapudi (2016). Whole of chapter-1 is relevant.

- Fig-1.1 presents an example concept map representing the key aspects of Machine learning (ML).

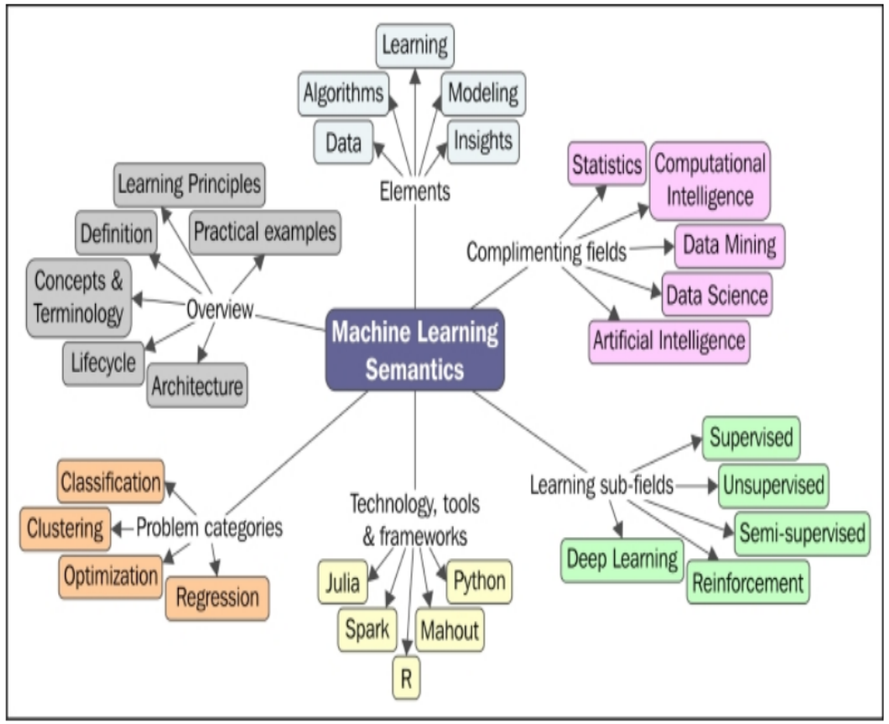

Figure 1.1: Concept Map Gollapudi (2016)

- There are various definitions for machine learning.

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." (Mitchell, 2017. Machine Learning, Mcgraw Hill)

- As per Wikipedia

"Machine learning is a scientific discipline that is concerned with the design and development of algorithms that allow computers to evolve behaviours based on empirical data, such as from sensor data or databases."

- Primary goal of a ML implementation is to develop a general purpose algorithm that solves a practical and focused problem.
- Important aspects in the process include data, time, and space requirements.
- The goal of a learning algorithm is to produce a result that is a rule and is as accurate as possible.

1.2 ML Process

- Types of datasets required: Training Set, Validation Set (may come from the initial data) and Testing Set
- Training set: data examples that are used to learn or build a classifier.
- Validation set: data examples that are verified against the built classifier and can help tune the accuracy of the output.
- Testing set: data examples that help assess the performance of the classifier.

Phase 1-Training Phase: Training data used to train the model by using expected output with the input. Output is the learning model.

Phase 2-Validation/Test Phase: Measuring the validity and fit of the model. How good is the model? Uses validation dataset, which can be a subset of the initial dataset.

Phase 3-Application Phase: Run the model with real world data to generate results.

- Fig-1.2 example flowchart on how learning can be applied to predict

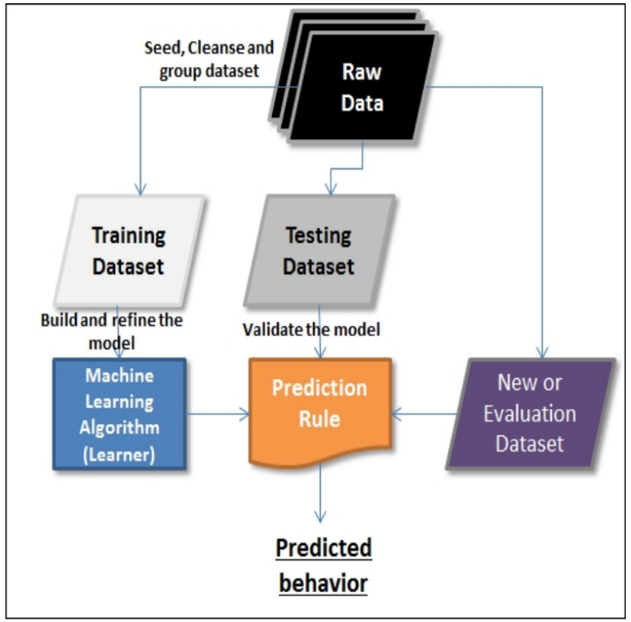

Figure 1.2: Example Flowchart for predictive ML workflow

1.3 Models

- Central to any ML implementation
- At a high level
  - Logical : Rule based (if else...), for example, decision trees.
  - Geometric: Use geometric concepts like lines, planes etc. Linear transformations are often used.
  - Probabilistic: Statistical models. Defines relationship between two variables.

1.4 Types of Learning Problems

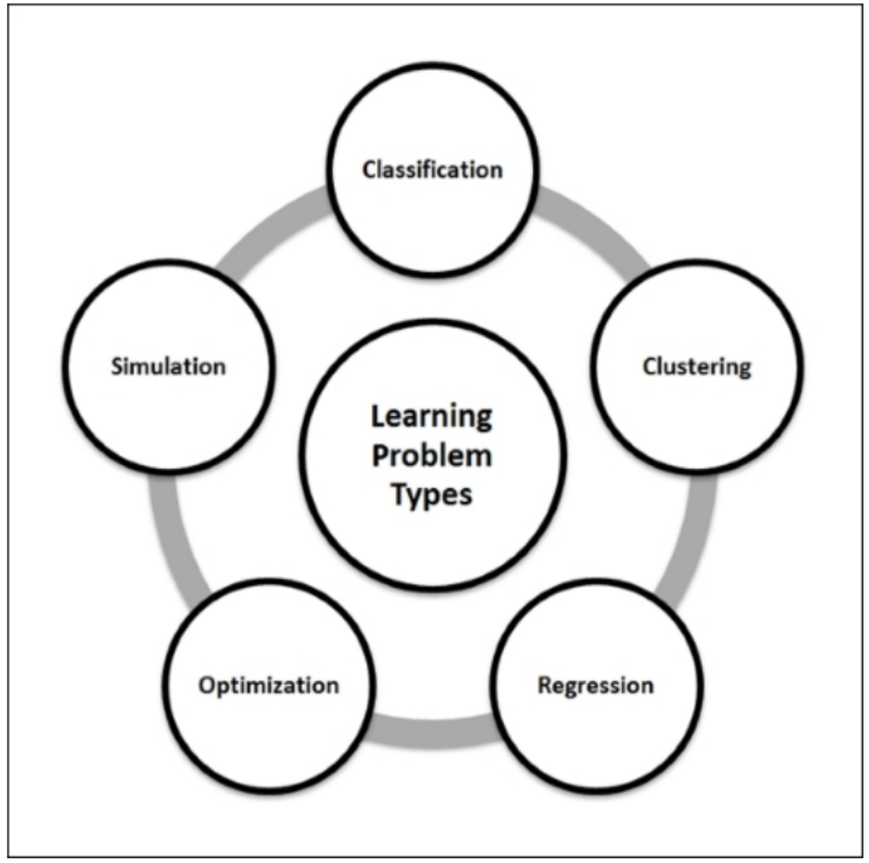

Figure 1.3: Learning Problems Categories

1.5 Machine Learning Algorithms

- Decision tree based algorithms
- Bayesian method based algorithms
- Kernel method based algorithms
- Clustering methods
- Artificial neural networks
- Dimensionality reduction
- Ensemble methods (combining multiple methods)
- Instance based learning algorithms
- Regression Analysis based algorithms
- Association rule based learning algorithms

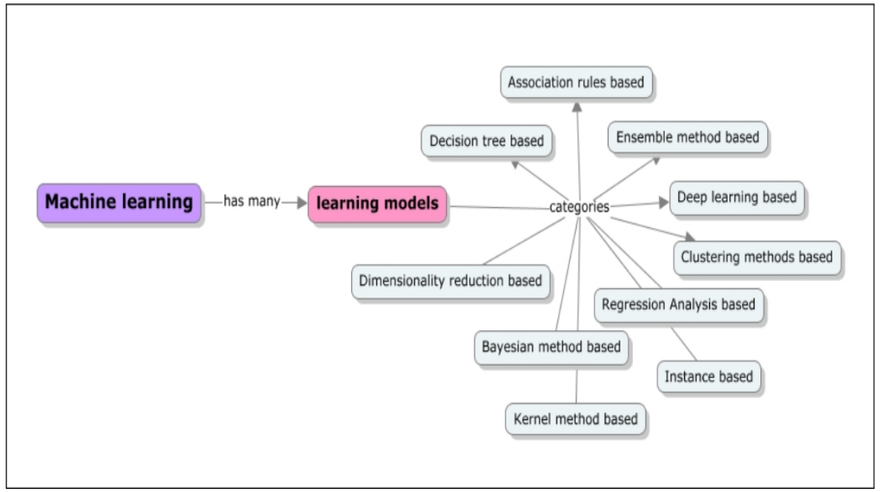

Figure 1.4: Machine learning algorithms/methods Gollapudi (2016)

1.6 Subfields of Machine Learning

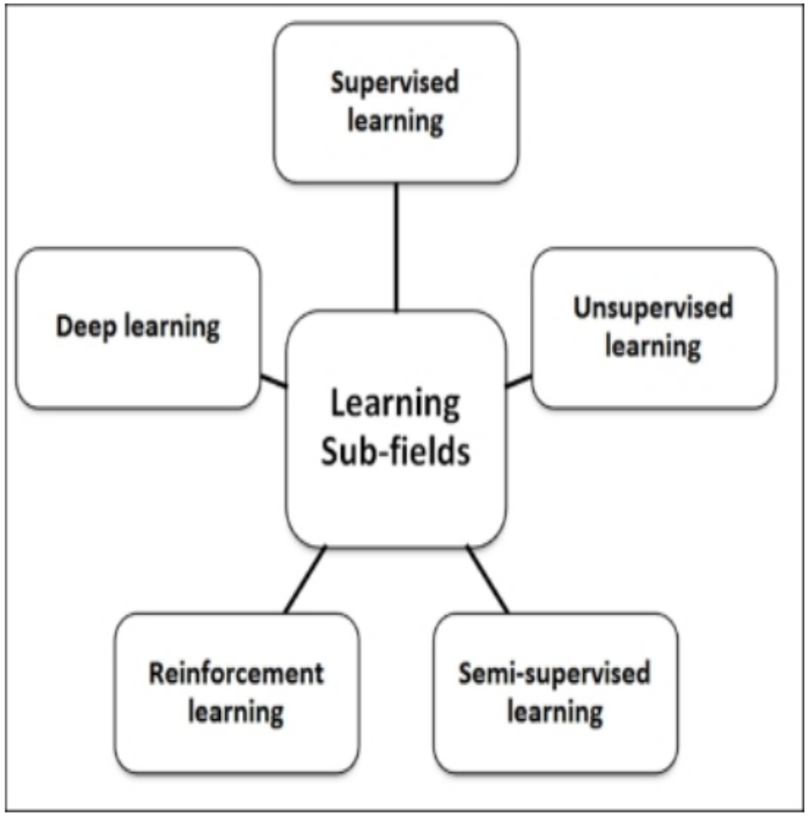

Figure 1.5: Subfields of ML

1.6.1 Supervised Learning

- Construct predictive models
- Prediction of a given output (or target) using other variables (or features) in the data set.
- Supervision refers to the fact that the target values provide a supervisory roles. Indicates to the learner the task it needs to learn.
- Uses labelled data.
- Most supervised learning problems are either regression or classification.

Also review Chapter-1 and 2 from Boehmke & Greenwell (2019)
`https://bradleyboehmke.github.io/HOML/`

1.6.2 Unsupervised Learning

- Statistical tools to conduct descriptive analysis; for better understanding of the data.
- No specific target to solve, for example, clustering to identify groups.
- Unsupervised learning is often performed as part of an exploratory data analysis (EDA).
- Unlabelled dataset

References

Boehmke, Brad, & Greenwell, Brandon M. 2019. *Hands-on machine learning with R.* CRC Press.

Gollapudi, Sunila. 2016. *Practical Machine Learning.* Packt Publishing.