Machine Learning Basics Pt 1

Can Serif Mekik
PhD Candidate
Rensselaer Polytechnic Institute

April 14, 2022

CDSI
McGill Faculty of Science

McGill University • Co-Curricular Record
APPROVED

Are you interested in machine learning? If you are looking for a place to get started, then please join us for this workshop where we will cover basic concepts in machine learning theory and briefly survey some popular machine learning models.

At the end of this workshop (parts 1+2), you will be able to:
> Identify common machine learning problems and models.
> Understand basic concepts in machine learning theory and practice
> Create and run simple machine learning models using the scikit-learn Python library

Pre-requisites? Working knowledge of Python, linear algebra, calculus, and probability theory.

Today

1. Welcome
2. Software & Dataset
3. The Big Picture
4. The ML Cycle
5. An Initial Model
6. Feature Transformations
7. Overfitting
8. Topics Ahead

Software

scikit-learn – Python ML library

- Relatively Mature
- Versatile
- Available as part of the Anaconda stack

Anaconda: https://www.anaconda.com/

Dataset

Handwritten digit classification

- Classic task for computer vision & neural networks
- ‘digits’ dataset included in scikit-learn

Why do machine learning?

- Many interesting tasks present the following situation:
  - Manual coding is extremely difficult
  - Collecting data is easy
- Using ML we can
  - Take advantage of available data
  - Save work

Elements of an ML Project

- Want to compute some unknown function *f** that optimizes some performance metric *m* over samples taken from some domain *S*
- We know/choose *m* and we have can sample data from *S*; we use these resources to iterate through hypotheses for *f**

Learning Paradigms

- There are many paradigms for ML, each with its own set of metrics, domains, and even models.
- The three classic paradigms are
  - Supervised learning
  - Unsupervised learning
  - Reinforcement learning

Supervised Learning

- Training data contains explicit examples of correct output
- Work with explicit error measures
- Examples:
  - Classification
  - Regression

Reinforcement Learning

- Want to select an optimal sequence of actions
- So, agent interacts and affects domain
- Training data does not contain explicit information on correct outputs
- Work with rewards instead, which may be delayed and intermittent

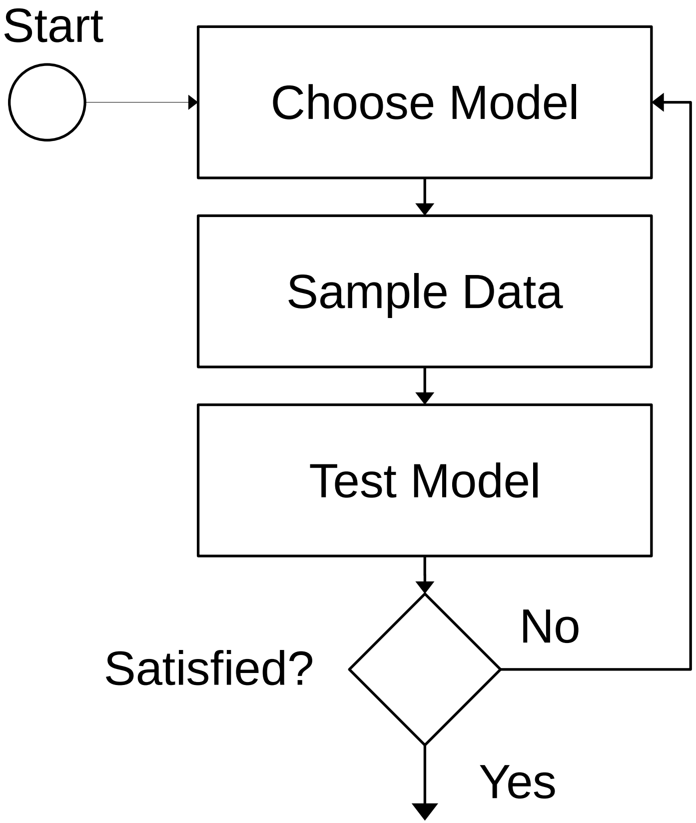
Unsupervised Learning

- Goal is to spontaneously discover patterns in data
- No output information available in data
- Examples:
  - Clustering
  - Dimensionality Reduction

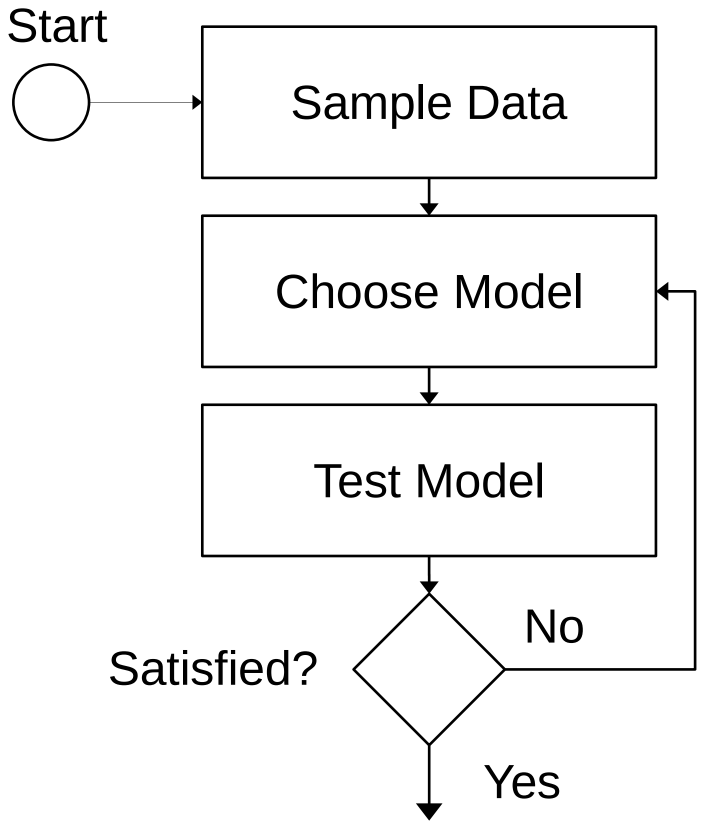
The ML Cycle

- ML is very diverse in its techniques and approaches
- Nevertheless ML projects tend to follow a common cycle

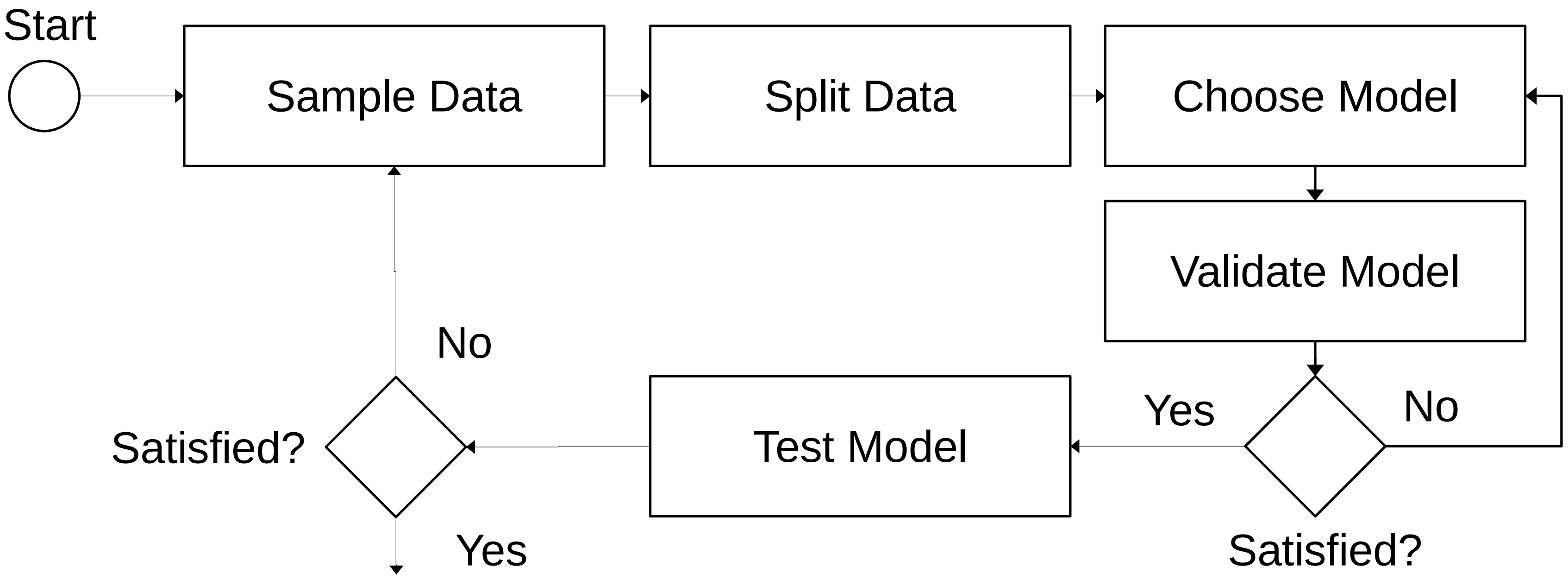
ML Utopia



ML Dystopia



ML Reality



Start
Sample Data
Split Data
Choose Model
Validate Model
No
Yes
No
Satisfied?
Test Model
Yes
Satisfied?

Features

- Features are variables constructed from raw data that are used by a model in decision making
- Features can be manually desinged or learned

The Perceptron Model

input

$$h(\boldsymbol{x_i}|\boldsymbol{w}) = sign\left(\sum_{j=1}^{N} w_{ji}\, x_i + b\right)$$

parameter intercept

Accuracy

$$\frac{1}{N}\sum_{i=1}^{N} [\![h(\boldsymbol{x_i}|\boldsymbol{w}) \neq y_i]\!]$$

Looking Ahead

- Theory of generalization
- Quantifying model complexity
- Gradient descent
- Various more sophisticated models