

# Machine Learning

Nicholas Lambert

6/25/2021

## Definition

“The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience.”<sup>1</sup>

## Origin

Computer Science: \* how to manually program computers to solve tasks

Statistics: \* what conclusions can be inferred from data

Machine Learning: \* intersection of computer science and statistics \* how to get computers to program themselves from experience plus some initial structure \* effective data capture, store, index, retrieve and merge \* computational tractability<sup>2</sup>

## Types of machine learning

Machine learning approaches are divided into two main types

- **Supervised**
  - training of a “predictive” model from data
  - one (or more) attribute of the dataset is used to “predict” another attribute e.g., classification
- **Unsupervised**
  - discovery of descriptive patterns in data
  - commonly used in data mining e.g., clustering

## Supervised

- Training dataset
  - input attribute(s)
  - attribute to predict
- Testing dataset
  - input attribute(s)
  - attribute to predict
- Type of learning model
- Evaluation function
  - evaluates difference between prediction and output in testing data

---

<sup>1</sup>Mitchell, T. (1997). Machine Learning. McGraw Hill.

<sup>2</sup>Mitchell, T.M., 2006. The discipline of machine learning (Vol. 9). Pittsburgh, PA: Carnegie Mellon University, School of Computer Science, Machine Learning Department.

## Unsupervised

- Dataset
  - input attribute(s) to explore
- Type of model for the learning process
  - most approaches are iterative
  - e.g., hierarchical clustering
- Evaluation function
  - evaluates the quality of the pattern under consideration during one iteration

## Semi-supervised learning

Supervised learning requires “labelled data”

- which can be expensive to acquire

Semi-supervised learning

- combines a small amount of labelled data with a larger un-labelled dataset
  - train on small labelled dataset
  - apply model to larger unlabelled dataset generating “pseudo-labels”
  - re-train the model with all data (including “pseudo-labels”)
  - assumptions: continuity, cluster, and manifold (lower dimensionality)

## Reinforcement learning

Based on the idea of training agents to learn how to

- take actions
  - which affect
  - agent state
  - environment
- to maximize reward
- balancing
  - exploration (new paths/options)
  - exploitation (of current knowledge)