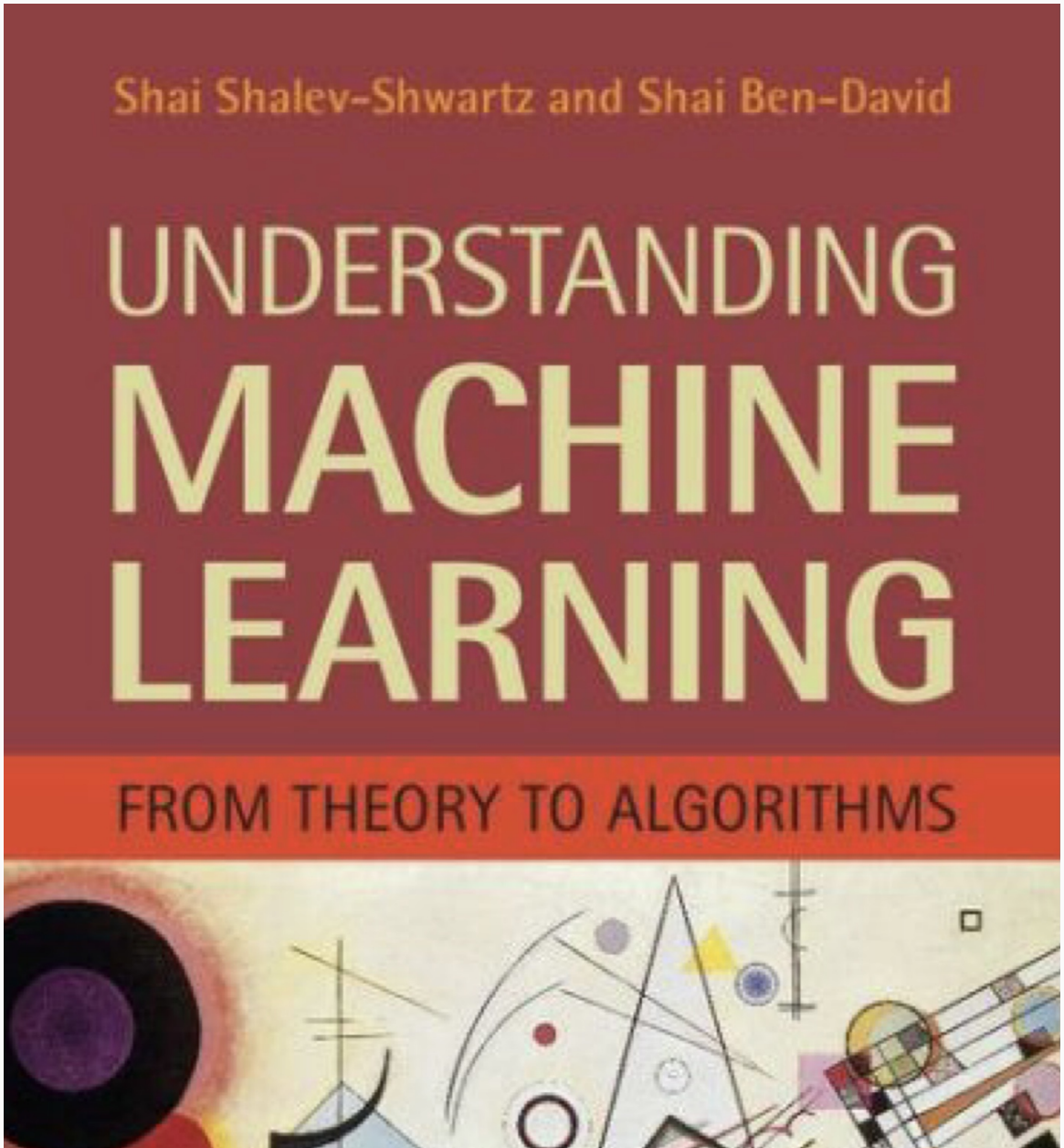
My thoughts on Shalev-Shwartz & Ben-David

Michael McMahon

January 25, 2022

Machine Learning Study Group

The Assignment



My Impression

Book Outline

Part I: Foundations

2. A Gentle Start
3. A Formal Learning Model
4. Learning via Uniform Convergence
5. The Bias-Complexity Tradeoff
6. The VC-Dimension
7. Nonuniform Learnability
8. The Runtime of Learning

Part II: From Theory to Algorithms

9. Linear Predictors
10. Boosting
11. Model Selection and Validation
12. Convex Learning Problems
13. Regularization and Stability
14. Stochastic Gradient Descent
15. Support Vector Machines
16. Kernel Methods
17. Multiclass, Ranking, and Complex Prediction Problems
18. Decision Trees
19. Nearest Neighbor
20. Neural Networks

Part III: Additional Learning Models

21. Online Learning
22. Clustering
23. Dimensionality Reduction
24. Generative Models
25. Feature Selection and Generation

Part IV: Advanced Theory

26. Rademacher Complexities
27. Covering Numbers
28. Proof of the Fundamental Theorem of Learning Theory
29. Multiclass Learnability
30. Compression Bounds
31. PAC-Bayes

- Appendix A Technical Lemmas
- Appendix B Measure Concentration
- Appendix C Linear Algebra

But before any of that...

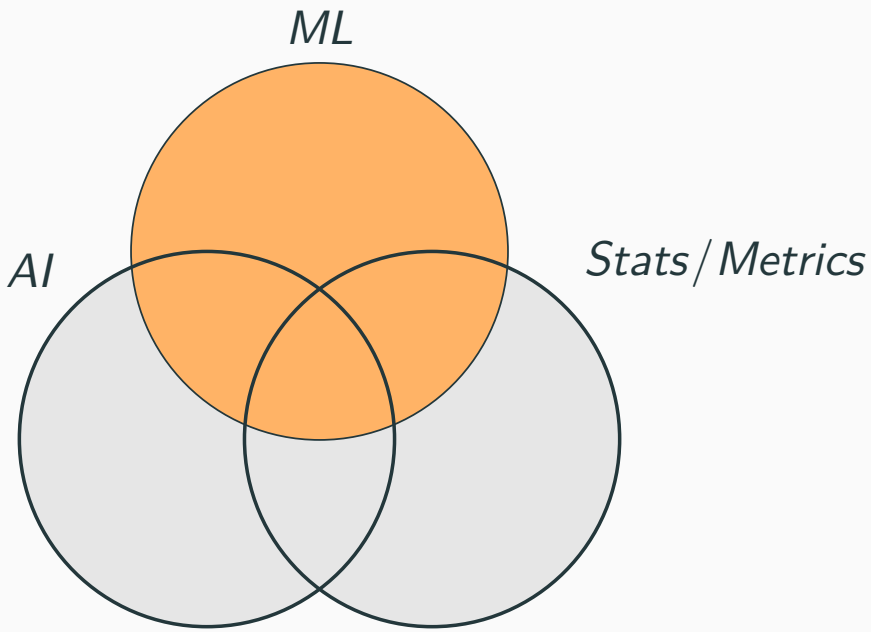
What I learned

What is Machine Learning

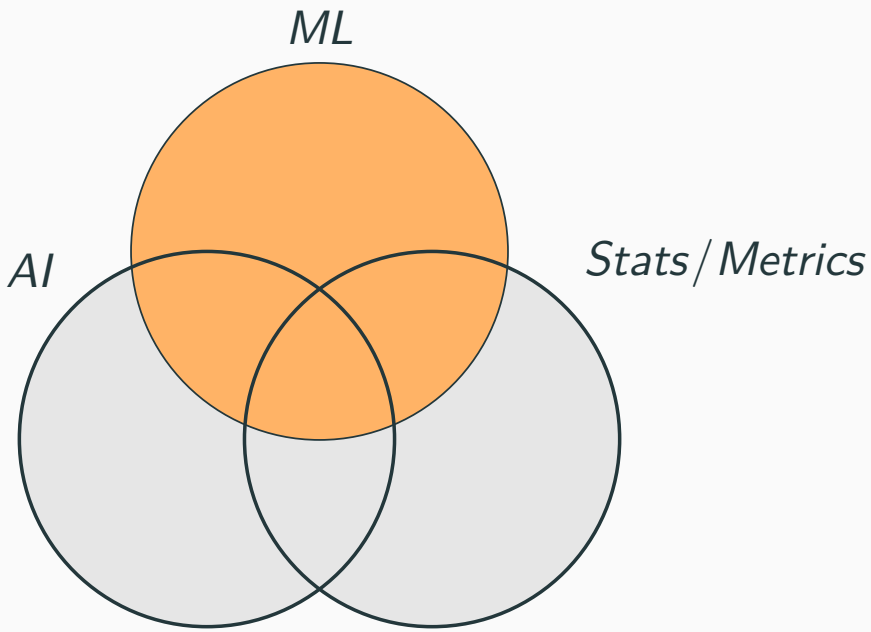
Some descriptions

- "we wish to program computers so that they can 'learn' from input available to them. Roughly speaking, learning is the process of converting experience into expertise or knowledge. The input to a learning algorithm is training data, representing experience, and the output is some expertise, which usually takes the form of another computer program that can perform some task."
- "Machine Learning is about the execution of learning by computers; hence algorithmic issues are pivotal. We develop algorithms to perform the learning tasks and are concerned with their computational efficiency."
- "given the size of available samples, machine learning theory aims to figure out the degree of accuracy that a learner can expect on the basis of such samples."

What ML is not...



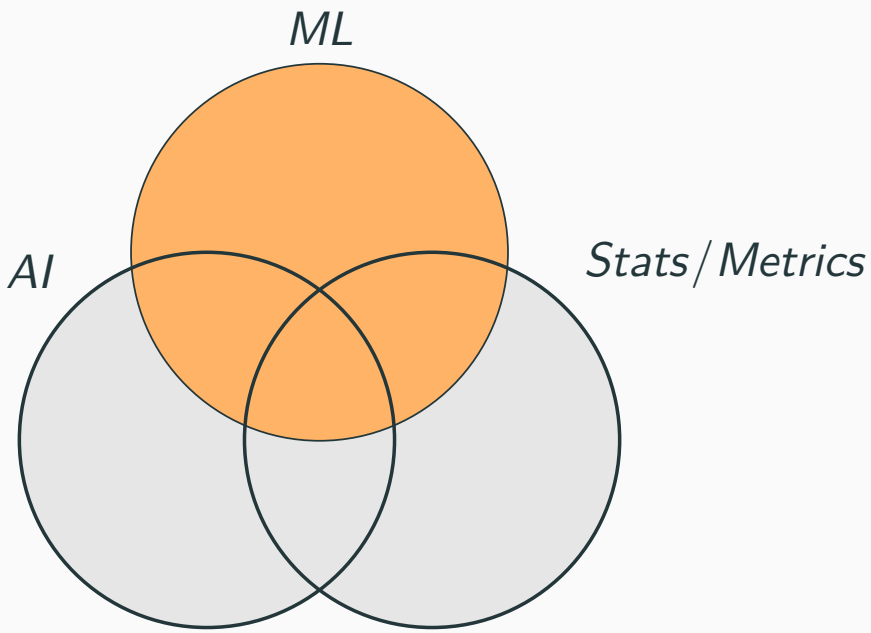
What ML is not...



AI: trying to build automated imitation of intelligent behavior

ML: use the strengths and special abilities of computers to complement human intelligence, often performing tasks that fall way beyond human capabilities

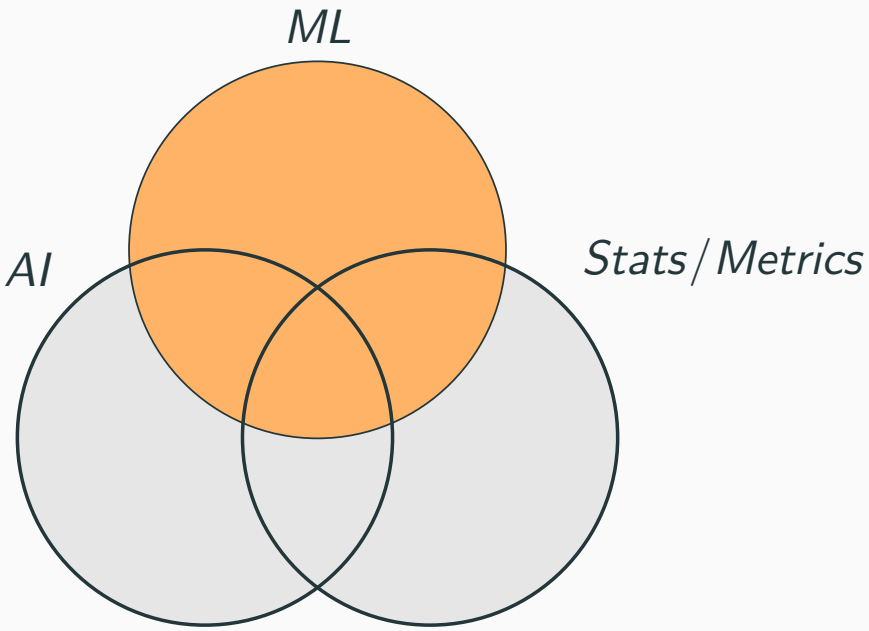
What ML is not...



Stats: test hypotheses

ML: use the data gathered from samples to come up with meaningful patterns (or hypotheses) that may have been missed by the human observer.

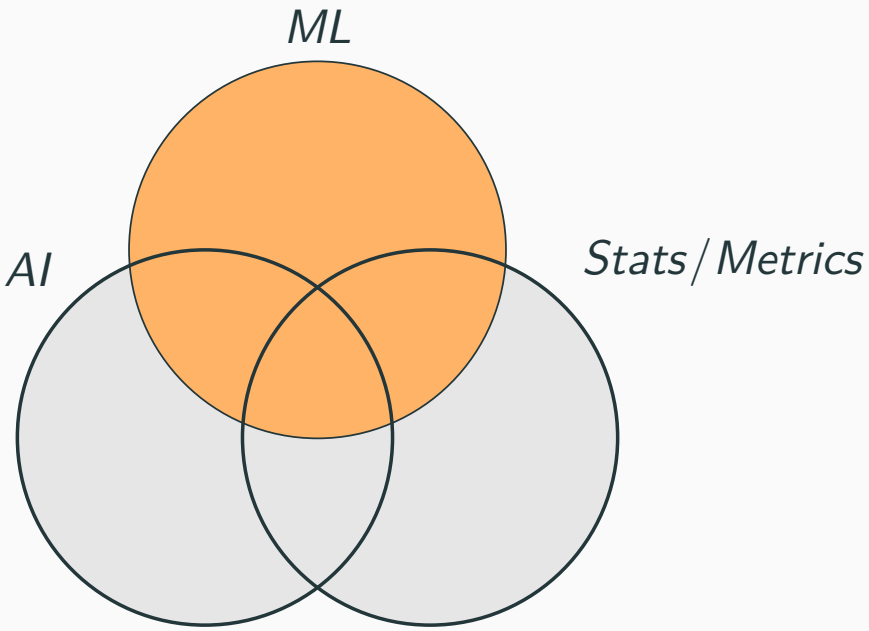
What ML is not...



Stats: basis on asymptotic properties

ML: the theory of machine learning focuses on finite sample bounds

What ML is not...



Stats: common to work under the assumption of certain presubscribed data models

ML: emphasis is on 'distribution-free' setting

Key Dimensions of the ML Problem

1. Supervised versus Unsupervised
    - Classifier versus plain-vanilla LDA

Key Dimensions of the ML Problem

1. Supervised versus Unsupervised
    - Classifier versus plain-vanilla LDA
2. Active versus Passive Learners
    - Pose questions vs take what you're given

Key Dimensions of the ML Problem

1. Supervised versus Unsupervised
    - Classifier versus plain-vanilla LDA
2. Active versus Passive Learners
    - Pose questions vs take what you're given
3. Helpfulness of the Teacher
    - ML assumes statistical learning

Key Dimensions of the ML Problem

1. Supervised versus Unsupervised
    - Classifier versus plain-vanilla LDA
2. Active versus Passive Learners
    - Pose questions vs take what you're given
3. Helpfulness of the Teacher
    - ML assumes statistical learning
4. Online versus Batch Learning Protocol
    - Real-time vs ex-post learning

Basic Statistical Learning: Elements

- Domain set - $\mathcal{X}$
  - Domain points are *instances* and a vector of *features*
- Label set - $\mathcal{Y}$
- Training data - $\mathcal{S} = ((x_1, y_1), (x_2, y_2), ...(x_m, y_m))$
- Learner's Output - $h : \mathcal{X} \to \mathcal{Y}$
- Model of Data Generation
  - Assume: probability distribution over $\mathcal{X}$ is $\mathcal{D}$
    - No knowledge of $\mathcal{D}$ necessarily assumed
  - Correct labelling function $\Rightarrow f : \mathcal{X} \to \mathcal{Y}$
    - $y_i = f(x_i)\ \forall i$

Basic Statistical Learning: An Approach I

Objective

Based on $\mathcal{S}$, can we figure out $h : \mathcal{X} \to \mathcal{Y}$ that matches (approximates) $f : \mathcal{X} \to \mathcal{Y}$

- Measure of Success: Low error of the classifier
  - Probability, according to $\mathcal{D}$, a random instance x has $h(x) \neq f(x)$

$$L_{\mathcal{D},f} \equiv \mathbb{P}_{x \sim \mathcal{D}} \left[h(x) \neq f(x) \right]$$

  - True error of h / generalisation error / the risk

Basic Statistical Learning: An Approach I

Objective

Based on $\mathcal{S}$, can we figure out $h : \mathcal{X} \to \mathcal{Y}$ that matches (approximates) $f : \mathcal{X} \to \mathcal{Y}$

- Measure of Success: Low error of the classifier
  - Probability, according to $\mathcal{D}$, a random instance x has $h(x) \neq f(x)$

$$L_{\mathcal{D},f} \equiv \mathbb{P}_{x \sim \mathcal{D}} \left[h(x) \neq f(x) \right]$$

  - True error of h / generalisation error / the risk

Problem

The learner is blind to the underlying distribution $\mathcal{D}$ over the world and to the labeling function f. The only way the learner can interact with the environment is through observing the training set.

Basic Statistical Learning: An Approach II

Solution

A learning algorithm receives as input a training set $\mathcal{S}$, sampled from an unknown distribution $\mathcal{D}$ and labeled by some target function f, and should output a predictor $h_{\mathcal{S}} : \mathcal{X} \to \mathcal{Y}$. The goal of the algorithm is to find $h_{\mathcal{S}}$ that minimizes the error with respect to the unknown $\mathcal{D}$ and f.

Basic Statistical Learning: An Approach II

Solution

A learning algorithm receives as input a training set $\mathcal{S}$, sampled from an unknown distribution $\mathcal{D}$ and labeled by some target function f, and should output a predictor $h_{\mathcal{S}} : \mathcal{X} \to \mathcal{Y}$. The goal of the algorithm is to find $h_{\mathcal{S}}$ that minimizes the error with respect to the unknown $\mathcal{D}$ and f.

- Use the training error / empirical error / empirical risk:

$$L_{\mathcal{S}}(h) \equiv \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}$$

- This learning paradigm: Empirical Risk Minimization (ERM)

$$ERM_{\mathcal{H}}(\mathcal{S}) \in \operatorname*{argmin}_{h \in \mathcal{H}} L_{\mathcal{S}}(h)$$