1. Intro, Ch. 1 Shalev-Shwartz & Ben-David

Friday, January 7, 2022 5:19 PM

aka "SS" for Shai, Shai or Shalev-Shwartz

- Go over syllabus, websites (Canvas, github)
- Recommended reading: preface to § 2.3.1 (p.17 in print version) of SS
- Intro: what is machine learning? (ML = machine learning)
  ... and what kind of ML class is this?

Some authors distinguish ML and AI

Not too relevant for us: we study the subset of ML known as statistical learning*, a special case of the branch of math/statistics known as empirical process theory.

* Not directly related to the series of books "The Elements of Statistical Learning" by Friedman, Tibshirani, & Hastie. Some overlap.

Applications of ML/AI

Traditionally: OCR (optical character recognition)
Email Spam filtering
Machine translation
Speech recognition
Vision
Face recognition / detection
Control systems, e.g. self-driving cars
Search, ads
Recommendation Systems
Fraud detection

Recently, more applications than we can keep track of!

In academics, major examples:
protein folding, bioinformatics, astronomy, medicine, solving PDE

How does ML compare to similar fields?

descriptive
- knowledge discovery, data mining
  Ex. tools: tableau
  Ex: smoking is related to cancer

predictive
- statistics
  Ex. tools: R
  smoking causes lung cancer w/ p-value 10^{-20}

prescriptive
- machine learning
  Ex. tools: PyTorch, Scikit-learn
  for all hospital patients, which treatments should we suggest?

... of course it's not really that simple, the lines are very blurry.

Why do ML? (instead of using human labor*?)

- to mimick human intuition/expertise that is hard to duplicate and/or codify (since experts are hard to come by)
- superhuman tasks (faster or with more data)
- easily adapt to variations

*this includes direct programming

Types of learning problems

We'll focus on topics in red

(1) supervised (i.e. labelled) vs unsupervised (vs semi-supervised) (vs Reinforcement Learning)

(Ex: Spam filtering. It's easy for most of us to give a label to each email (spam or not)

(Ex: clustering
"There are two types of people in this world..."
No right answer!

② active vs passive ___ training data is fixed.

(for semi-supervised learning,
you can request an oracle (expert) to label a few data
(Related to experimental design)

③ data assumptions: statistical learning vs adversarial
(i.i.d) (non-stationary)

④ online vs. batch

(you need to start making decisions before you have all the data! Every time step, make a decision!

Ex: restaurant problem: try a new dish or order your favorite?
exploration exploitation

We focus on supervised, batch, passive, i.i.d
since it's simplest but still contains essential ideas

(⚠ ML is not a closed field --- many theoretical aspects are not understood. More on this later)

Types of tasks

Our focus is on binary classification

Classification (binary or multiclass)

Regression (predict a $\mathbb{R}$ value; unlike classification, now you get credit for being close)

Clustering (unsupervised task)

Ranking (e.g. search: is a good page returned in top 10 results?)

Dimensionality Reduction, manifold learning and misc.
(e.g. learning reduced order models for PDE)