Machine Learning Theory: Overview

COMS 4773 Spring 2024 (Daniel Hsu)

Agenda

- Today:
  - About machine learning theory
  - About the course
  - Example of learning problem

What is machine learning (ML)?

Examples:

- Spam filtering (from email text)
- Ad click prediction (from user profile and context)
- Gene expression level prediction (from upstream DNA)
- Best-next-move prediction (from state of chess board)
- ...
- Programming-by-demonstration

What is machine learning (ML)?

Examples:

- Spam filtering (from email text)
- Ad click prediction (from user profile and context)
- Gene expression level prediction (from upstream DNA)
- Best-next-move prediction (from state of chess board)
- ...
- Programming-by-demonstration

Note:

- This is *not* an introductory course in ML
- Also, won't discuss application-oriented aspects of ML
- No programming or "data" in this class
- See COMS 4771 and elsewhere instead

What is learning theory?

- Design/analysis of machine learning algorithms/problems
  - Computational resources: running time, memory, ...
  - Data resources: sample size, rounds of interaction, ...
- Many different models for theoretical analysis
  - Online learning
  - Statistical learning
  - Learning with queries
  - Finding planted structures
  - ...
- Each one tries to capture some essential aspect of "learning"
- A cross between theoretical computer science and statistics

Why study learning theory? (1)

Relevance to machine learning practice

- Breiman (1995) "Reflections After Refereeing Papers for NIPS"

2. USES OF THEORY

- **Comfort:** We knew it worked, but it's nice to have a proof.
- **Insight:** Aha! So that's why it works.
- **Innovation:** At last, a mathematically proven idea that applies to data.
- **Suggestion:** Something like this might work with data.

Why study learning theory? (1)

Relevance to machine learning practice

- Breiman (1995) "Reflections After Refereeing Papers for NIPS"

2. USES OF THEORY

- **Comfort:** We knew it worked, but it's nice to have a proof.
- **Insight:** Aha! So that's why it works.
- **Innovation:** At last, a mathematically proven idea that applies to data.
- **Suggestion:** Something like this might work with data.

Breiman's "Post World War II" Examples (in 1995):

- Asymptotic analyses of decision trees, nearest neighbor, universal approximation
- Nonparametric regression, sparsity in inverse problems
- Spectral analysis in time series, information theory, bootstrap
- Theory-inspired heuristics for function fitting

Why study learning theory? (2)

Relevance to machine learning practice

- Breiman (1995) "Reflections After Refereeing Papers for NIPS"

> **Mathematical theory is not critical to the development of machine learning.**
>
> ***But scientific inquiry is.***

3.5 INQUIRY

INQUIRY = sensible and intelligent efforts to understand what is going on. For example:

- mathematical heuristics
- simplified analogies (like the Ising Model)
- simulations
- comparisons of methodologies
- devising new tools
- theorems where useful (rare!)
- shunning panaceas

Why study learning theory? (3)

Insights into general phenomenon of learning

- Valiant (1984) "A Theory of the Learnable"

ABSTRACT: *Humans appear to be able to learn new concepts without needing to be programmed explicitly in any conventional sense. In this paper we regard learning as the phenomenon of knowledge acquisition in the absence of explicit programming. We give a precise methodology for studying this phenomenon from a computational viewpoint.*

Why study learning theory? (3)

Insights into general phenomenon of learning

- Valiant (1984) "A Theory of the Learnable"

ABSTRACT: *Humans appear to be able to learn new concepts without needing to be programmed explicitly in any conventional sense. In this paper we regard learning as the phenomenon of knowledge acquisition in the absence of explicit programming. We give a precise methodology for studying this phenomenon from a computational viewpoint.*

My suggestions

- Study learning theory for its breadth of topics and the wide applicability of its methods
- Theorem + proof = demonstration of understanding

About this course

- COMS 4773
  - Was COMS 4995-1 in Spring 2020
  - Large overlap with COMS 4252 and COMS 4774
- Website (with syllabus, schedule, etc):
  http://www.cs.columbia.edu/~djhsu/LT/
- Topics:
  - Online learning (e.g., learning with experts, multi-arm bandits)
  - Statistical learning (e.g., generalization theory)
  - Unsupervised learning (e.g., clustering models)
- Learning goals:
  - Rigorously analyze ML problems/algorithms
  - Read/understand research papers in ML theory

Course requirements

Prerequisites

- Mathematical maturity; reading and writing proofs
- Probability, linear algebra, a bit of convex analysis and algorithm design/analysis
- Prior exposure to machine learning is helpful for motivation

Requirements

- Reading assignments – schedule on website
  - Primarily from a few textbooks, available on the website
- Homework assignments – will be posted on website
  - 60% of overall grade
- Midterm exam (week of March 4?) – probably take-home, but do this by yourself!
  - 20% of overall grade
- Reading Project – instructions forthcoming
  - 20% of overall grade

Example: Linear regression

- Hypothesis class: linear functions in $\mathbb{R}^d$

$$\boldsymbol{x} \mapsto \boldsymbol{x}^\mathsf{T}\boldsymbol{w}$$

- Data model: iid random examples $((\boldsymbol{X}_i, Y_i))_{i=1}^n$ from $\mathbb{R}^d \times \mathbb{R}$ (at least satisfying moment conditions s.t. stuff below is finite)
- Success criterion:
  - Let $(\boldsymbol{X}, Y)$ be an independent copy of $(\boldsymbol{X}_1, Y_1)$
  - Learner returns a linear function $\hat{\boldsymbol{w}} = \hat{\boldsymbol{w}}((\boldsymbol{X}_1, Y_1), \dots, (\boldsymbol{X}_n, Y_n))$
  - Learner is successful if the mean squared error of $\hat{\boldsymbol{w}}$, defined by

$$\mathcal{R}(\boldsymbol{w}) := \mathbb{E}[(\boldsymbol{X}^\mathsf{T}\boldsymbol{w} - Y)^2] \quad (\forall \boldsymbol{w} \in \mathbb{R}^d),$$

  is not much larger than $\min_{\boldsymbol{w} \in \mathbb{R}^d} \mathcal{R}(\boldsymbol{w})$.

Example: Linear regression

- Hypothesis class: linear functions in $\mathbb{R}^d$

$$\boldsymbol{x} \mapsto \boldsymbol{x}^\mathsf{T}\boldsymbol{w}$$

- Data model: iid random examples $((\boldsymbol{X}_i, Y_i))_{i=1}^n$ from $\mathbb{R}^d \times \mathbb{R}$ (at least satisfying moment conditions s.t. stuff below is finite)
- Success criterion:
  - Let $(\boldsymbol{X}, Y)$ be an independent copy of $(\boldsymbol{X}_1, Y_1)$
  - Learner returns a linear function $\hat{\boldsymbol{w}} = \hat{\boldsymbol{w}}((\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n))$
  - Learner is successful if the mean squared error of $\hat{\boldsymbol{w}}$, defined by

$$\mathcal{R}(\boldsymbol{w}) := \mathbb{E}[(\boldsymbol{X}^\mathsf{T}\boldsymbol{w} - Y)^2] \quad (\forall \boldsymbol{w} \in \mathbb{R}^d),$$

  is not much larger than $\min_{\boldsymbol{w} \in \mathbb{R}^d} \mathcal{R}(\boldsymbol{w})$.

What is a good strategy here?

Empirical risk minimization

- Let P be the probability distribution of $(\boldsymbol{X}, Y)$.
  - Risk $\mathcal{R}(\boldsymbol{w})$ is mean squared prediction error of $\boldsymbol{w}$ w.r.t. P.
  - If we know P, then in principle we can just minimize $\mathcal{R}$.
  - What if we just have the random examples $((\boldsymbol{X}_i, Y_i))_{i=1}^n$?

Empirical risk minimization

- Let P be the probability distribution of $(\boldsymbol{X}, Y)$.
  - Risk $\mathcal{R}(\boldsymbol{w})$ is mean squared prediction error of $\boldsymbol{w}$ w.r.t. P.
  - If we know P, then in principle we can just minimize $\mathcal{R}$.
  - What if we just have the random examples $((\boldsymbol{X}_i, Y_i))_{i=1}^n$?
- **Plug-in principle**: pretend P_n is P, where

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{(\boldsymbol{X}_i, Y_i)}$$

  is the "empirical distribution", and proceed as above.

Empirical risk minimization

- Let P be the probability distribution of $(\boldsymbol{X}, Y)$.
  - Risk $\mathcal{R}(\boldsymbol{w})$ is mean squared prediction error of $\boldsymbol{w}$ w.r.t. P.
  - If we know P, then in principle we can just minimize $\mathcal{R}$.
  - What if we just have the random examples $((\boldsymbol{X}_i, Y_i))_{i=1}^n$?
- **Plug-in principle**: pretend P_n is P, where

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{(\boldsymbol{X}_i, Y_i)}$$

  is the "empirical distribution", and proceed as above.

This general approach—not just for linear functions—is called "Empirical Risk Minimization (ERM)".

Recap

- Recap:
  - About machine learning theory
  - About the course
  - Examples of learning problem
- Homework:
  - Please do HW0 (self-assessment)
  - Read (Blum, 1998) Section 2