Machine Learning Theory (CS 6783)

Tu-Th 1:25 to 2:40 PM
Kimball, B-11

Instructor : Karthik Sridharan
No exams!

5 assignments that count towards your grades (55%)

One term project (40%)

5% for class participation
Pre-requisites

- Basic probability theory
- Basics of algorithms and analysis
- Introductory level machine learning course
- Mathematical maturity, comfortable reading/writing formal mathematical proofs.
One of the following three options:

1. Pick your research problem, get it approved by me, write a report on your work

2. Pick two papers on learning theory, get it approved by me, write a report with your own views/opinions

3. I will provide a list of problems, workout problems worth a total of 10 stars out of this list

Oct 16th submit proposal/get your project approved by me
Finals week projects are due
Lets get started …
Use past observations to automatically learn to make better predictions/decisions in the future.
WHERE IS IT USED?

Recommendation Systems

House of Cards
2013-2014 TV-MA 2 Seasons

Bad, for a greater good.
Season 2 of this acclaimed original thriller series earned a total of 13 Emmy Award nominations including Outstanding Drama Series. Outstanding Lead Actor nominee Kevin Spacey stars as ruthless, cunning Congressman Francis Underwood, who will stop at nothing to conquer the halls of power in Washington D.C. His secret weapon: his gorgeous, ambitious, and equally conniving wife Claire (Outstanding Lead Actress nominee Robin Wright).

Directors’ Commentary Available
Watch Season 1 of this Emmy-winning series with exclusive scene-by-scene audio commentary from directors including David Fincher and Joel Schumacher.

Genres: TV Shows, TV Dramas
This show is: Witty, Cerebral, Dark
Pedestrian Detection
Where is it used?

Market Predictions
WHERE IS IT USED?

Spam Classification
Online advertising (improving click through rates)

Climate/weather prediction

Text categorization

Unsupervised clustering (of articles …)

…
Oops . . .
Cognitive theories look beyond behavior to explain brain-based learning. And constructivism views learning as a process in which the learner actively constructs or builds new ideas or concepts. Behaviorism. Behaviorism as a theory was primarily developed by B. F. Skinner.

Learning theory (education) - Princeton University
www.princeton.edu/.../Learning_theory_(education)... ▼ Princeton University ▼
What is Machine Learning Theory

- How do formalize machine learning problems
- Right framework for right problems (Eg. online, statistical)
- What does it mean for a problem to be “learnable”
- How many instances do we need to see to learn to given accuracy
- How do we build sound learning algorithms based on theory
- Computational learning theory: which problems are efficiently learnable
Outline of Topics

• Learning problem and frameworks, settings, minimax rates

• Statistical learning theory
  • Probably Approximately Correct (PAC) and Agnostic PAC frameworks
  • Empirical Risk Minimization, Uniform convergence, Empirical process theory
  • Finite model classes, MDL bounds, PAC Bayes theorem
  • Infinite model classes, Rademacher complexity
  • Binary Classification: growth function, VC dimension
  • Real-valued function classes, covering numbers, chaining, fat-shattering dimension
  • Supervised learning: necessary and sufficient conditions for learnability

• Online learning theory
  • Sequential minimax and value of online learning game
  • Martingale Uniform convergence, sequential empirical process theory
  • Sequential Rademacher complexity
  • Binary Classification: Littlestone dimension
  • Real-valued function classes, sequential covering numbers, chaining bounds, sequential fat-shattering dimension
  • Online supervised learning: necessary & sufficient conditions for learnability

• Designing learning algorithms: relaxations, random play-outs

• Computational Learning theory and more if time permits . . .
Learning Problem: Basic Notation

- **Input space/feature space**: \( \mathcal{X} \)
  
  (Eg. bag-of-words, n-grams, vector of grey-scale values, user-movie pair to rate)

  Feature extraction is an art, … an art we won’t cover in this course

- **Output space/label space**: \( \mathcal{Y} \)
  
  (Eg. \( \{\pm 1\}, [K], \mathbb{R}\)-valued output, structured output)

- **Loss function**: \( \ell: \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R} \)
  
  (Eg. 0–1 loss \( \ell(y', y) = \mathbf{1}\{y' \neq y\} \), sq-loss \( \ell(y', y) = (y - y')^2 \), absolute loss \( \ell(y', y) = |y - y'| \)

  Measures performance/cost per instance (inaccuracy of prediction/cost of decision).

- **Model class/Hypothesis class**: \( \mathcal{F} \subset \mathcal{Y}^{\mathcal{X}} \)
  
  (Eg. \( \mathcal{F} = \{x \mapsto f^T x : \|f\|_2 \leq 1\} \), \( \mathcal{F} = \{x \mapsto \text{sign}(f^T x)\} \))
Formalizing Learning Problems

- How is data generated?
- How do we measure performance or success?
- Where do we place our prior assumption or model assumptions?
Formalizing Learning Problems

- How is data generated?
- How do we measure performance or success?
- Where do we place our prior assumption or model assumptions?
- What we observe?
\[ \mathcal{Y} = \{\pm 1\} , \quad \ell(y', y) = 1 \{y' \neq y\} , \quad \mathcal{F} \subset \mathcal{Y}^\mathcal{X} \]

- Learner only observes training sample \( S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \)
  - \( x_1, \ldots, x_n \sim D_X \)
  - \( \forall t \in [n], y_t = f^*(x_t) \) where \( f^* \in \mathcal{F} \)
- Goal: find \( \hat{y} \in \mathcal{Y}^\mathcal{X} \) to minimize

\[ \mathbb{P}_{x \sim D_X} (\hat{y}(x) \neq f^*(x)) \]

(Either in expectation or with high probability)
**Definition**

Given $\delta > 0$, $\epsilon > 0$, sample complexity $n(\epsilon, \delta)$ is the smallest $n$ such that we can always find forecaster $\hat{y}$ s.t. with probability at least $1 - \delta$,

$$\mathbb{P}_{x \sim D_X} (\hat{y}(x) \neq f^*(x)) \leq \epsilon$$

(efficiently PAC learnable if we can learn efficiently in $1/\delta$ and $1/\epsilon$)

Eg. : learning output for deterministic systems
\( Y \subset \mathbb{R} \), \( \ell(y', y) = (y - y')^2 \), \( \mathcal{F} \subset \mathcal{Y}^\mathcal{X} \)

- Learner only observes training sample \( S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \)
  - \( x_1, \ldots, x_n \sim \mathcal{D}_X \)
  - \( \forall t \in [n], y_t = f^*(x_t) + \epsilon_t \) where \( f^* \in \mathcal{F} \) and \( \epsilon_t \sim N(0, \sigma) \)

- Goal: find \( \hat{y} \in \mathbb{R}^\mathcal{X} \) to minimize

\[
\|\hat{y} - f^*\|_{L_2(D_X)}^2 = \mathbb{E}_{x \sim D_X} \left[ (\hat{y}(x) - f^*(x))^2 \right]
\]

(Either in expectation or in high probability)

Eg.: clinical trials (inference problems) model class known.
NON-PARAMETRIC REGRESSION

\[ Y \subset \mathbb{R}, \quad \ell(\hat{y}, y) = (y - \hat{y})^2, \quad F \subset Y^X \]

- Learner only observes training sample \( S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \)
  - \( x_1, \ldots, x_n \sim D_X \)
  - \( \forall t \in [n], y_t = f^*(x_t) + \varepsilon_t \) where \( f^* \in F \) and \( \varepsilon_t \sim N(0, \sigma) \)

- Goal: find \( \hat{y} \in \mathbb{R}^X \) to minimize

\[
\|\hat{y} - f^*\|^2_{L_2(D_X)} = \mathbb{E}_{x \sim D_X} \left[ (\hat{y}(x) - f^*(x))^2 \right] = \mathbb{E}_{x \sim D_X} \left[ (\hat{y}(x) - y)^2 \right] - \inf_{f \in F} \mathbb{E}_{x \sim D_X} \left[ (f(x) - y)^2 \right]
\]

(Either in expectation or in high probability)

Eg. : clinical trials (inference problems) model class known.
Learner only observes training sample $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn iid from joint distribution $D$ on $\mathcal{X} \times \mathcal{Y}$

Goal: find $\hat{y} \in \mathbb{R}^\mathcal{X}$ to minimize expected loss over future instances

$$\mathbb{E}_{(x,y) \sim D} [\ell(\hat{y}(x), y)] - \inf_{f \in \mathcal{F}} \mathbb{E}_{(x,y) \sim D} [\ell(f(x), y)] \leq \epsilon$$

$$L_D(\hat{y}) - \inf_{f \in \mathcal{F}} L_D(f) \leq \epsilon$$
Definition

Given \( \delta > 0 \), \( \epsilon > 0 \), sample complexity \( n(\epsilon, \delta) \) is the smallest \( n \) such that we can always find forecaster \( \hat{y} \) s.t. with probability at least \( 1 - \delta \),

\[
L_D(\hat{y}) - \inf_{f \in F} L_D(f) \leq \epsilon
\]
Learning Problems

Pedestrian Detection

Spam Classification
Learning Problems

Pedestrian Detection
(Batch/Statistical setting)

Spam Classification
(Online/adversarial setting)
For $t = 1$ to $n$

Learner receives $x_t \in \mathcal{X}$
Learner predicts output $\hat{y}_t \in \mathcal{Y}$
True output $y_t \in \mathcal{Y}$ is revealed

End for

Goal: minimize regret

$$\text{Reg}_n(\mathcal{F}) := \frac{1}{n} \sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \ell(f(x_t), y_t)$$
Other Problems/Frameworks

- Unsupervised learning, clustering
- Semi-supervised learning
- Active learning and selective sampling
- Online convex optimization
- Bandit problems, partial monitoring, …
No Free Lunch Theorems

Statistical learning theory

- Empirical risk minimization
- Uniform convergence and learning
- Finite model classes, MDL, PAC Bayes theorem, …
Brush up Markov inequality, Chebychev inequality, central limit theorem

Read up or brush up, concentration inequalities
(specifically Hoeffding bound, Bernstein bound, Hoeffding-Azuma inequality, McDiarmid’s inequality also referred to as bounded difference inequality)

Brush up union bound

Watch out for homework 0, no need to submit, just a warmup