Machine Learning Theory (CS 6783)

Tu-Th 1:25 to 2:40 PM
Olin Hall, 255

Instructor : Karthik Sridharan
TA: Ayush Sekhari

About the course

- 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.*

Term project

One of the following three options :

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

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

October 5th submit proposal/get your project approved by me
Finals week projects are due

Assignments

1. 2.5 before fall break, 2.5 after fall break
2. You are allowed at most 2 late submissions (up to 3 days on each) without penalty, but do notify me
3. Beyond this late submissions will be penalized for each day its late by
4. Assignment submission via CMS, submit as PDF.

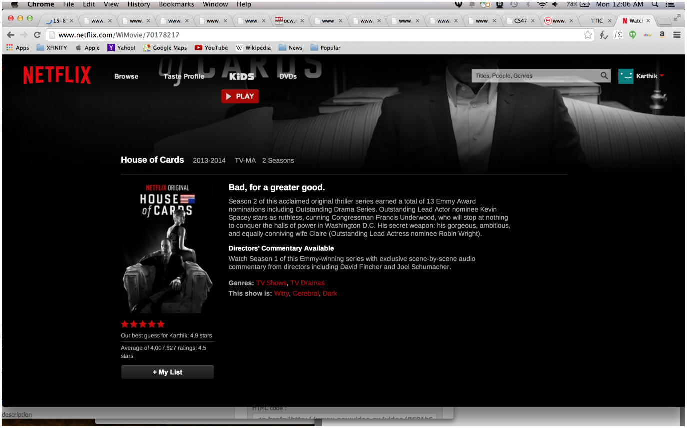
Lets get started …

What is Machine Learning

Use past observations to automatically learn to make better predictions/decisions in the future.

Where is it used ?

Recommendation Systems



Where is it used ?

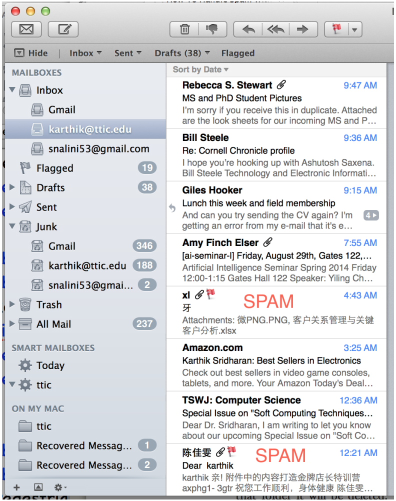
Pedestrian Detection

Where is it used ?

Market Predictions

Where is it used ?

Spam Classification



Where is it used ?

- Online advertising (improving click through rates)
- Climate/weather prediction
- Text categorization
- Unsupervised clustering (of articles ...)
- ...

What is Learning Theory

What is Learning Theory

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

Feedback

What is Machine Learning Theory

- How do we formalize machine learning problems
- Right framework for right problems (Eg. online , statistical)
- How do we pick the right model to use and what are the tradeoffs between various models
- How many instances do we need to see to learn to given accuracy
- How do we design learning algorithms with provable guarantees on performance
- *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
  - Bound on learning rates: MDL bounds, PAC Bayes theorem, Rademacher complexity, VC dimension, covering numbers, fat-shattering dimension
  - Supervised learning : necessary and sufficient conditions for learnability
- Online learning theory
  - Sequential minimax and value of online learning game
  - Regret bounds: Sequential Rademacher complexity, Littlestone dimension, sequential covering numbers, sequential fat-shattering dimension
  - Online supervised learning : necessary & sufficient conditions for learnability
- Algorithms for online convex optimization: Exponential weights algorithm, strong convexity, exp-concavity and rates, Online mirror descent
- Deriving generic learning algorithms : relaxations, random play-outs
- If time permits, uses of learning theory results in optimization, approximation algorithms, perhaps a bit of bandits, . . .

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. $\mathcal{F} = \{x \mapsto f^\top x : \|f\|_2 \leq 1\}$, $\mathcal{F} = \{x \mapsto \mathrm{sign}(f^\top x)\}$)

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 ?*

Probably Approximately Correct Learning

$$\mathcal{Y} = \{\pm 1\}\ ,\ \ \ell(y', y) = \mathbf{1}\{y' \neq y\}\ ,\ \ \mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$$

- Learner only observes training sample $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
  - $x_1, \ldots, x_n \sim \mathbf{D}_X$
  - $\forall t \in [n], y_t = f^*(x_t)$ 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)

Probably Approximately Correct Learning

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

Non-parametric Regression

$$\mathcal{Y} \subset \mathbb{R}, \quad \ell(y', y) = (y - y')^2, \quad \mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$$

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

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

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

(Either in expectation or in high probability)

Eg. : clinical trials (inference problems) model class known.

Non-parametric Regression

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

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

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

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

(Either in expectation or in high probability)

Eg. : clinical trials (inference problems) model class known.

Statistical Learning (Agnostic PAC)

- Learner only observes training sample $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn iid from joint distribution $\mathbf{D}$ on $\mathcal{X} \times \mathcal{Y}$

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

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

$$L_{\mathbf{D}}(\hat{y}) - \inf_{f \in \mathcal{F}} L_{\mathbf{D}}(f) \leq \epsilon$$

Well suited for *Prediction* problems.

Statistical Learning (Agnostic PAC)

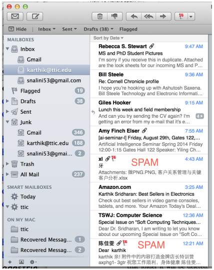
Definition

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 $1 - \delta$,

$$L_{\mathbf{D}}(\hat{y}) - \inf_{f \in \mathcal{F}} L_{\mathbf{D}}(f) \leq \epsilon$$

Learning Problems

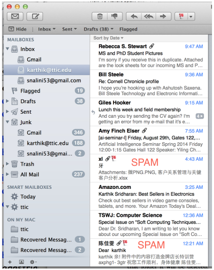
Pedestrian Detection



Spam Classification

Learning Problems

Pedestrian Detection
(Batch/Statistical setting)



Spam Classification
(Online/adversarial setting)

Online Learning (Sequential Prediction)

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

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

Other Problems/Frameworks

- Unsupervised learning, clustering
- Semi-supervised learning
- Active learning and selective sampling
- Online convex optimization
- Bandit problems, partial monitoring, . . .

Sneek Peek

- No Free Lunch Theorems
- Minimax rates for various setting/problems
- Comparing the various settings