10-301/601: Introduction to Machine Learning Lecture 15 – Learning Theory (Finite Case)

Matt Gormley & Henry Chai

3/10/25

Front Matter

- Announcements
  - HW5 released 2/27, due 3/16 at 11:59 PM
  - Exam 1 Exit Poll due 3/10 (today!) at 11:59 PM
  - Peer tutoring information will be posted to Piazza some time this week

Statistical Learning Theory Model

independent and identically distributed

1. Data points are generated i.i.d. from some *unknown* distribution

$$\boldsymbol{x}^{(n)} \sim p^*(\boldsymbol{x})$$

2. Labels are generated from some *unknown* function

$$y^{(n)} = c^*\left(\boldsymbol{x}^{(n)}\right)$$

3. The learning algorithm chooses the hypothesis (or classifier) with lowest *training* error rate from a specified hypothesis set, $\mathcal{H}$

4. Goal: return a hypothesis (or classifier) with low *true* error rate

Statistical Learning Theory Model

1. Data points are generated i.i.d. from some *unknown* distribution
$$\boldsymbol{x}^{(n)} \sim p^*(\boldsymbol{x})$$
2. Labels are generated from some *unknown* function
$$y^{(n)} = c^*\left(\boldsymbol{x}^{(n)}\right) \in \{-1, +1\}$$
3. The learning algorithm chooses the hypothesis (or classifier) with lowest *training* error rate from a specified hypothesis set, $\mathcal{H}$
4. Goal: return a hypothesis (or classifier) with low *true* error rate

Types of Error

- True error rate
  - Actual quantity of interest in machine learning
  - How well your hypothesis will perform on average across all possible data points
- Test error rate
  - Used to evaluate hypothesis performance
  - Good estimate of your hypothesis's true error
- Validation error rate
  - Used to set hypothesis hyperparameters
  - Slightly "optimistic" estimate of your hypothesis's true error
- Training error rate
  - Used to set model parameters
  - Very "optimistic" estimate of your hypothesis's true error

Types of Risk (a.k.a. Error)

- Expected risk of a hypothesis h (a.k.a. true error)

$$R(h) = P_{\vec{x} \sim p^*}\left(h(\vec{x}) \neq c^*(\vec{x})\right)$$

- Empirical risk of a hypothesis h (a.k.a. training error)

$$\hat{R}(h) = P_{\vec{x} \sim D}\left(h(\vec{x}) \neq c^*(\vec{x})\right)$$

training dataset

$\vec{x} \sim D \Rightarrow \vec{x}$ is uniformly at random chosen from $\{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$

$$\hat{R}(h) = \underbrace{\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left(h(\vec{x}^{(i)}) \neq c^*(\vec{x}^{(i)})\right)}_{\text{training error rate}}$$

indicator function

Three Hypotheses of Interest

1. The *true function,* c^*

2. The *expected risk minimizer,*

$$h^* = \operatorname*{argmin}_{h \in \mathcal{H}} R(h)$$

3. The *empirical risk minimizer,*

$$\hat{h} = \operatorname*{argmin}_{h \in \mathcal{H}} \hat{R}(h)$$

$\mathcal{H} = \{$ all linear decision boundaries in 2D $\}$

Poll Question 1: Which of the following are *always* true?

A. $c^* = h^*$ 38%

B. $c^* = \hat{h}$

C. $h^* = \hat{h}$

D. $c^* = h^* = \hat{h}$

E. None of the above 42%

F. **TOXIC**

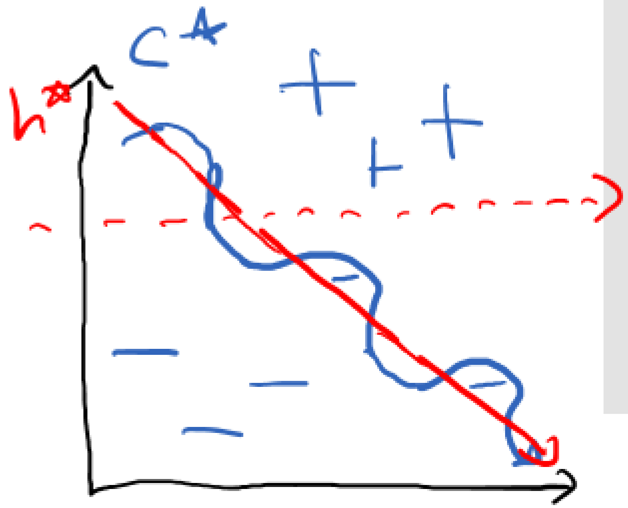
- The *true function*, c^* $R(c^*) = 0$

- The *expected risk minimizer*,

$$h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} R(h) \overset{?}{=} 0$$

- The *empirical risk minimizer*,

$$\hat{h} = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \hat{R}(h)$$



Key Question

- Given a hypothesis with zero/low training error, what can we say about its true error?

PAC Learning

- PAC = **<u>P</u>**robably **<u>A</u>**pproximately **<u>C</u>**orrect

- PAC Criterion:

$$P\left(\left|R(h) - \hat{R}(h)\right| \leq \epsilon\right) \geq 1 - \delta \ \forall \ h \in \mathcal{H}$$

for some ϵ (difference between expected and empirical risk) and δ (probability of “failure”)

  - We want the PAC criterion to be satisfied for $\mathcal{H}$ with small values of ϵ and δ

Sample Complexity

- The sample complexity of an algorithm/hypothesis set, $\mathcal{H}$, is the number of labelled training data points needed to satisfy the PAC criterion for some δ and ϵ
- Four cases
  - Realizable vs. Agnostic
    - Realizable $\rightarrow c^* \in \mathcal{H}$
    - Agnostic $\rightarrow c^*$ might or might not be in $\mathcal{H}$
  - Finite vs. Infinite
    - Finite $\rightarrow |\mathcal{H}| < \infty$
    - Infinite $\rightarrow |\mathcal{H}| = \infty$

most realistic

Theorem 1: Finite, Realizable Case

- For a finite hypothesis set $\mathcal{H}$ s.t. $c^* \in \mathcal{H}$ and arbitrary distribution p^*, if the number of labelled training data points satisfies

$$M \geq \frac{1}{\epsilon}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right)\right)$$

then with probability at least $1 - \delta$, all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$

Proof of Theorem 1: Finite, Realizable Case

(sketch)

1. Assume the worst!!
   Assume that every hypothesis in H is bad! ($R(h) > \epsilon$)
2. The probability that a bad hypothesis "tricks" me ($\hat{R}(h) = 0$) is tiny! And shrinks as M grows
3. The probability that any of the bad hypotheses "tricks" me is pretty small and also shrinks as $M \uparrow$

Proof of Theorem 1: Finite, Realizable Case

do some math

⋮

$P($at least one "bad" hypothesis correctly classifies M training data points$) \leq |H|(1-\epsilon)^M \leq \delta$

do some more math

⋮

$$M \geq \frac{1}{\epsilon}\left(\ln|H| + \ln\left(\frac{1}{\delta}\right)\right)$$

Proof of Theorem 1: Finite, Realizable Case

Given $M \geq \frac{1}{\epsilon}\left(\ln |H| + \ln\left(\frac{1}{\delta}\right)\right)$ labelled training data points sampled from p^*, the probability $\exists$ some hypothesis $h \in H$ with $R(h) > \epsilon$ and $\hat{R}(h) = 0$ is $\leq \delta$

$\Updownarrow$

"Given $M \geq$ sampled from p^*",

the probability that all bad hypotheses $h \in H$ with $R(h) > \epsilon$ have $\hat{R}(h) > 0$ is $\geq 1 - \delta$

Aside: Proof by Contrapositive

- The contrapositive of a statement $A \Rightarrow B$ is $\neg B \Rightarrow \neg A$
- A statement and its contrapositive are logically equivalent, i.e., $A \Rightarrow B$ means that $\neg B \Rightarrow \neg A$
- Example: “it’s raining $\Rightarrow$ Henry brings am umbrella”

  is the same as saying

  “Henry didn’t bring an umbrella $\Rightarrow$ it’s not raining ”

Proof of Theorem 1: Finite, Realizable Case

Given $m \geq \frac{1}{\epsilon}\left(\ln|H| + \ln\left(\frac{1}{\delta}\right)\right)$ labelled training data points sampled from P^*, the probability that all $h \in H$ with $\underbrace{R(h) > \epsilon}_{A} \Rightarrow \underbrace{\hat{R}(h) > 0}_{B}$ is $\geq 1-\delta$

$$\Updownarrow$$

$$\overbrace{\hat{R}(h) = 0}^{\neg B} \Rightarrow \overbrace{R(h) \leq \epsilon}^{\neg A} \quad \blacksquare$$

Poll Question 2:

- Let $\mathcal{H}$ be the set of all *conjunctions* over M Boolean variables, $\boldsymbol{x} \in \{0,1\}^M$; examples of conjunctions are
  - $h(\boldsymbol{x}) = x_1(1 - x_2)x_4x_{10}$ ✱
  - $h(\boldsymbol{x}) = (1 - x_3)(1 - x_4)x_8$
- Assuming $c^* \in \mathcal{H}$, if $M = 10$, $\epsilon = 0.1$, and $\delta = 0.01$, at least how many labelled examples do we need to satisfy the PAC criterion using Theorem 1?

$\frac{1}{0.01} = 100$

$\ln 3^{10} = 10 \ln 3$

A. 1 **(TOXIC)**

10% B. $10(2\ln 10 + \ln 100) \approx 92$

13% C. $10(3\ln 10 + \ln 100) \approx 116$

30% D. $10(10\ln 2 + \ln 100) \approx 116$

38% E. $10(10\ln 3 + \ln 100) \approx \underline{156}$

F. ~~$100(2\ln 10 + \ln 10) \approx 691$~~

G. ~~$100(3\ln 10 + \ln 10) \approx 922$~~

H. ~~$100(10\ln 2 + \ln 10) \approx 924$~~

I. ~~$100(10\ln 3 + \ln 10) \approx 1329$~~

Theorem 1: Finite, Realizable Case

- For a finite hypothesis set $\mathcal{H}$ s.t. $c^* \in \mathcal{H}$ and arbitrary distribution p^*, if the number of labelled training data points satisfies

$$M \geq \frac{1}{\epsilon}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right)\right)$$

then with probability at least $1 - \delta$, all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$

- Making the bound tight and solving for ϵ gives...

Statistical Learning Theory Corollary

- For a finite hypothesis set $\mathcal{H}$ s.t. $c^* \in \mathcal{H}$ and arbitrary distribution p^*, given a training data set S s.t. $|S| = M$, all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have

$$R(h) \leq \frac{1}{M}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right)\right)$$

with probability at least $1 - \delta$.

Theorem 2: Finite, Agnostic Case

- For a finite hypothesis set $\mathcal{H}$ and arbitrary distribution p^*, if the number of labelled training data points satisfies

$$M \geq \frac{1}{2\epsilon^2}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{2}{\delta}\right)\right)$$

then with probability at least $1 - \delta$, all $h \in \mathcal{H}$ satisfy

$$\underbrace{\left|R(h) - \hat{R}(h)\right| \leq \epsilon}$$

- Bound is inversely quadratic in ϵ, e.g., halving ϵ means we need four times as many labelled training data points
- Again, making the bound tight and solving for ϵ gives…

Statistical Learning Theory Corollary

- For a finite hypothesis set $\mathcal{H}$ and arbitrary distribution p^*, given a training data set S s.t. $|S| = M$, all $h \in \mathcal{H}$ have

$$\hat{R}(h) - \sqrt{\cdots} \lesssim R(h) \leq \hat{R}(h) + \sqrt{\frac{1}{2M}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{2}{\delta}\right)\right)}$$

with probability at least $1 - \delta$.

What happens when $|\mathcal{H}| = \infty$?

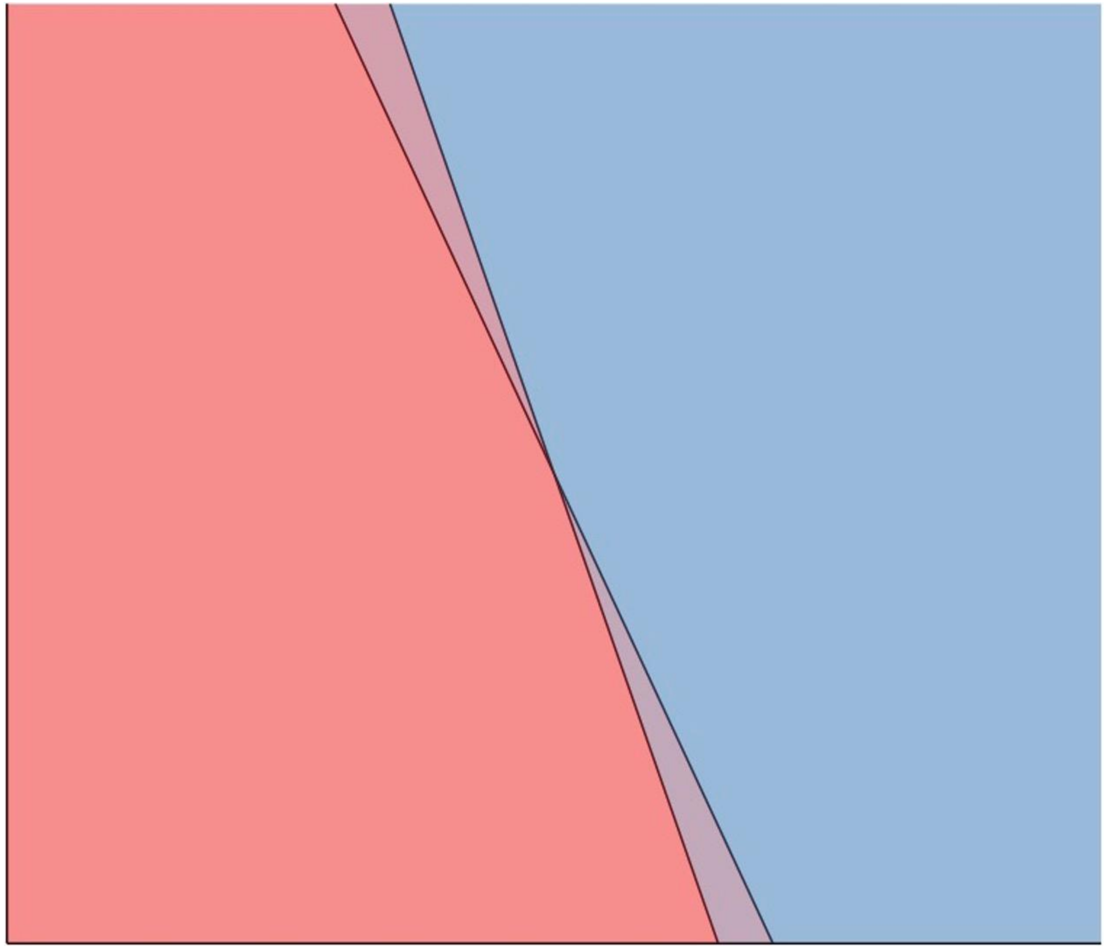
- For a finite hypothesis set $\mathcal{H}$ and arbitrary distribution p^*, given a training data set S s.t. $|S| = M$, all $h \in \mathcal{H}$ have

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{1}{2M}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{2}{\delta}\right)\right)}$$

with probability at least $1 - \delta$.

Intuition

For most infinite hypothesis sets $\mathcal{H}$, many hypotheses in $\mathcal{H}$ will behave very similarly



Intuition

For most infinite hypothesis sets $\mathcal{H}$, many hypotheses in $\mathcal{H}$ will behave very similarly

Relative to a given dataset, these two hypotheses are *identical*!