10-301/601: Introduction to Machine Learning Lecture 17 – Learning Theory (Finite Case)

Henry Chai

5/28/25

Front Matter

- Announcements:
  - HW4 released on 5/23, due 5/28 (today!) at 11:59 PM
  - Midterm on 5/30 at 9:30 AM in BH A36
    - Lectures 1 – 14 are in-scope; **this week's lectures will not be tested on the midterm**
    - Recitation on 5/29 will be a review of the practice problems

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

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

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)

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

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

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

$$= \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\left(c^*(\vec{x}^{(n)}) \neq h(\vec{x}^{(n)})\right)$$

↑ indicator

where $D = \{(\vec{x}^{(n)}, y^{(n)})\}_{n=1}^{N}$ is a training dataset & $\vec{x} \sim D$ means uniform sampling

Three Hypotheses of Interest

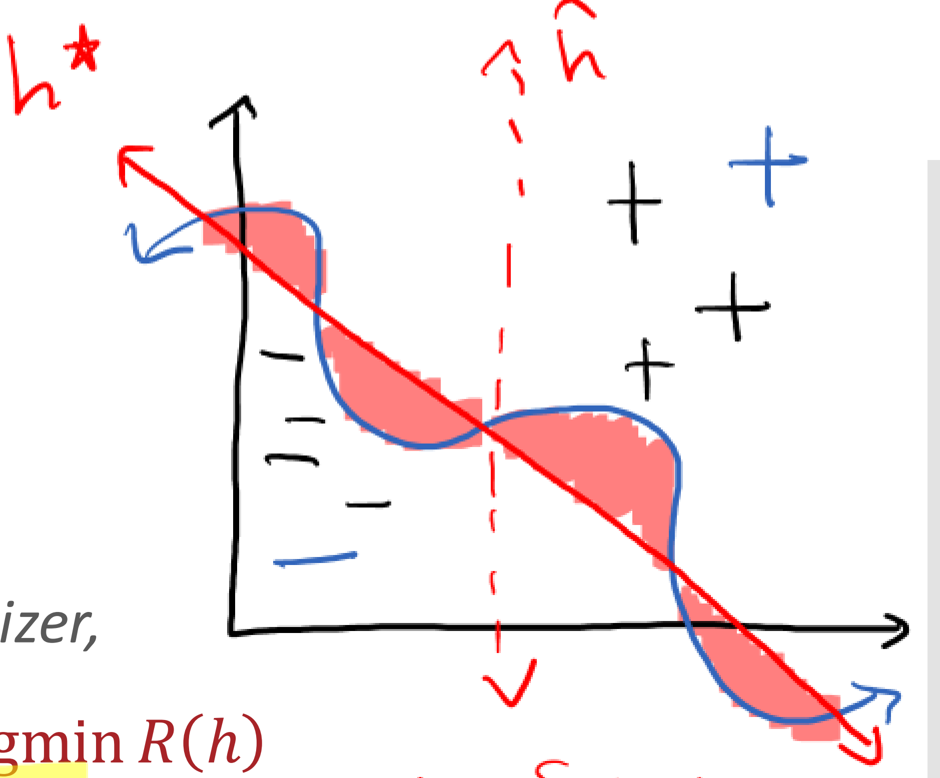
1. The *true function*, c^*

2. The *expected risk minimizer*,

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

3. The *empirical risk minimizer*,

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



$\mathcal{H}$ = {all linear decision boundaries} c^*

Lecture 17 + 18 Polls

0 surveys completed

0 surveys underway

Select all that apply: Which of the following statements is *always* true?

$c^* = h^*$

$c^* = \hat{h}$

$h^* = \hat{h}$

All of the above ($c^* = h^* = \hat{h}$)

None of the above

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 → $c^* \in \mathcal{H}$
    - Agnostic → c^* might or might not be in $\mathcal{H}$
  - Finite vs. Infinite
    - Finite → $|\mathcal{H}| < \infty$
    - Infinite → $|\mathcal{H}| = \infty$

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

1. Assume there are K "bad" hypotheses in H: $\{h_1, h_2, \ldots, h_K\}$ with $R(h_i) > \epsilon$

2. Think about one bad hypothesis, h_i

   a. $P(h_i$ correctly classifies the first training data point$) < 1 - \epsilon$

   b. $P(h_i$ correctly classifies all M training data point$) < (1-\epsilon)^M$

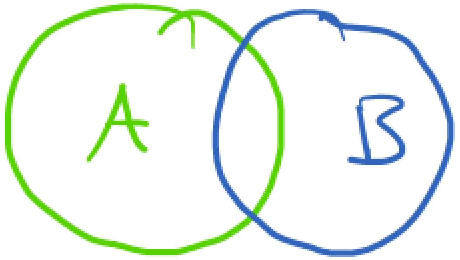
3. $P($that at least one bad hypothesis correctly classifies all M training data points$) =$

$$P(\hat{R}(h_1) = 0 \cup \hat{R}(h_2) = 0 \cup \ldots \cup \hat{R}(h_K) = 0)$$

Proof of Theorem 1: Finite, Realizable Case

4. Use the union bound

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$\leq P(A) + P(B)$$



$$P(\hat{R}(h_1) = 0 \cup \hat{R}(h_2) = 0 \cup \ldots \cup \hat{R}(h_K) = 0)$$
$$\leq \sum_{k=1}^{K} P(\hat{R}(h_k) = 0) < \sum_{k=1}^{K} (1-\epsilon)^M$$
$$= K(1-\epsilon)^M$$

P(at least one bad hypothesis tricks us) $< K(1-\epsilon)^M$

Proof of Theorem 1: Finite, Realizable Case

5. $$K(1-\epsilon)^M < |H|(1-\epsilon)^M$$

6. Use the fact that $1 - x \leq \exp(-x) \quad \forall x$

$$\Rightarrow (1-\epsilon)^M \leq \exp(-\epsilon)^M$$

$$|H|(1-\epsilon)^M \leq |H| \exp(-\epsilon M) \leq \delta$$

7. $$\Rightarrow \exp(-\epsilon M) \leq \frac{\delta}{|H|}$$

$$\Rightarrow -\epsilon M \leq \ln(\delta / |H|)$$

$$\Rightarrow M \geq \frac{1}{-\epsilon} \ln(\delta / |H|)$$

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

Proof of Theorem 1: Finite, Realizable Case

6. Given $M \geq \frac{1}{\epsilon}(\ln(|H|) + \ln(1/\delta))$ training data points, the probability $\exists$ a bad hypothesis $h_i \in \mathcal{H}$ with $R(h_i) > \epsilon$ and $\hat{R}(h_i) = 0$ is $\leq \delta$

$$\Updownarrow$$

Given $M \geq \frac{1}{\epsilon}(\ln(|H|) + \ln(1/\delta))$ t.d.p., the probability that all bad hypotheses with $R(h_i) > \epsilon$ have $\hat{R}(h_i) > 0$ 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

6. Given $M \geq \frac{1}{\epsilon}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right)\right)$ labelled training data points, the probability that $\exists$ a bad hypothesis $h_k \in \mathcal{H}$ with $R(h_k) > \epsilon$ and $\hat{R}(h_k) = 0$ is $\leq \delta$

$$\Updownarrow$$

Given $M \geq \frac{1}{\epsilon}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right)\right)$ labelled training data points, the probability that all hypotheses $h_k \in \mathcal{H}$ with $R(h_k) > \epsilon$ have $\hat{R}(h_k) > 0$ is $\geq 1 - \delta$

$A \Rightarrow B$

Proof of Theorem 1: Finite, Realizable Case

6. Given $M \geq \frac{1}{\epsilon}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right)\right)$ labelled training data points, the probability that all hypotheses $h_k \in \mathcal{H}$ with $R(h_k) > \epsilon$ have $\hat{R}(h_k) > 0$ is $\geq 1 - \delta$

$$\Updownarrow$$

Given $M \geq \frac{1}{\epsilon}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right)\right)$ labelled training data points, the probability that all hypotheses $h_k \in \mathcal{H}$ with $\hat{R}(h_k) = 0$ have $R(h_k) \leq \epsilon$ is $\geq 1 - \delta$

$\neg B \Rightarrow \neg A$

(proof by contrapositive)

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 $\epsilon = \frac{1}{M} \cdots$

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

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

- 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

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

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 $1 - \delta$.

Key Takeaways

- Statistical learning theory model
- Expected vs. empirical risk of a hypothesis
- Four possible cases of interest
  - realizable vs. agnostic
  - finite vs. infinite
- Sample complexity bounds and statistical learning theory corollaries for finite hypothesis sets