10-701: Introduction to Machine Learning Lecture 20 – Learning Theory (Finite Case)

Henry Chai

4/1/24

Front Matter

- Announcements
  - HW5 released 3/22, due 4/1 (today!) at 11:59 PM
  - Project mentors have been assigned
    - **If you haven't already done so, please meet with your project mentors ASAP to discuss your proposals**
  - Project check-ins due on 4/8 at 11:59 PM
  - **Daniel is on leave and will be for an indeterminate amount of time, please direct all course requests/questions to Henry**

Recall: What is ~~Machine Learning~~ 10-701?

- Supervised Models
  - Decision Trees
  - KNN
  - Naïve Bayes
  - Perceptron
  - Logistic Regression
  - Linear Regression
  - Neural Networks
  - SVMs
- Unsupervised Learning
- Ensemble Methods
- Graphical Models
- Learning Theory
- Reinforcement Learning
- Deep Learning
- Generative AI
- Important Concepts
  - Feature Engineering
  - Regularization and Overfitting
  - Experimental Design
  - Societal Implications

Recall: What is ~~Machine Learning~~ 10-701?

- Supervised Models
  - Decision Trees
  - KNN
  - Naïve Bayes
  - Perceptron
  - Logistic Regression
  - Linear Regression
  - Neural Networks
  - SVMs
- Unsupervised Learning
- Ensemble Methods
- Graphical Models
- **Learning Theory**
- Reinforcement Learning
- Deep Learning
- Generative AI
- Important Concepts
  - Feature Engineering
  - Regularization and Overfitting
  - Experimental Design
  - Societal Implications

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

Recall: 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 the true error rate
- Validation error rate: used to set model hyperparameters
  - Slightly "optimistic" estimate of the true error rate
- Training error rate: used to set model parameters
  - Very "optimistic" estimate of the true error rate

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

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

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

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

$$\hat{R}(h) = P_{\boldsymbol{x} \sim \mathcal{D}}\left(c^*(\boldsymbol{x}) \neq h(\boldsymbol{x})\right)$$
$$= \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\left(c^*(\boldsymbol{x}^{(n)}) \neq h(\boldsymbol{x}^{(n)})\right)$$
$$= \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\left(y^{(n)} \neq h(\boldsymbol{x}^{(n)})\right)$$

where $\mathcal{D} = \left\{\left(\boldsymbol{x}^{(n)}, y^{(n)}\right)\right\}_{n=1}^{N}$ is the training data set and $\boldsymbol{x} \sim \mathcal{D}$ denotes a point sampled uniformly at random from $\mathcal{D}$

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,*

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

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 $\mathcal{H}$, i.e., $h_1, h_2, \ldots, h_K$ that all have $R(h_k) > \epsilon$
2. Pick one bad hypothesis, h_k
   A. Probability that h_k correctly classifies the first training data point $< 1 - \epsilon$
   B. Probability that h_k correctly classifies all M training data points $< (1 - \epsilon)^M$
3. Probability that at least one bad hypothesis correctly classifies all M training data points $=$

$$P(h_1 \text{ correctly classifies all } M \text{ training data points} \cup$$
$$h_2 \text{ correctly classifies all } M \text{ training data points} \cup$$
$$\vdots$$
$$\cup\, h_K \text{ correctly classifies all } M \text{ training data points})$$

Proof of Theorem 1: Finite, Realizable Case

$$P(h_1 \text{ correctly classifies all } M \text{ training data points} \cup$$
$$h_2 \text{ correctly classifies all } M \text{ training data points} \cup$$
$$\vdots$$
$$\cup\, h_K \text{ correctly classifies all } M \text{ training data points})$$

$$\leq \sum_{k=1}^{K} P(h_k \text{ correctly classifies all } M \text{ training data points})$$

by the union bound: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
$\leq P(A) + P(B)$

Proof of Theorem 1: Finite, Realizable Case

$$\sum_{k=1}^{K} P(h_k \text{ correctly classifies all } M \text{ training data points})$$

$$< k(1-\epsilon)^M \leq |\mathcal{H}|(1-\epsilon)^M$$

because $k \leq |\mathcal{H}|$

3. Probability that at least one bad hypothesis correctly classifies all M training data points $\leq |\mathcal{H}|(1-\epsilon)^M$
4. Using the fact that $1 - x \leq \exp(-x)\ \forall\, x$,
   $$|\mathcal{H}|(1-\epsilon)^M \leq |\mathcal{H}|\exp(-\epsilon)^M = |\mathcal{H}|\exp(-M\epsilon)$$
5. Probability that at least one bad hypothesis correctly classifies all M training data points $\leq |\mathcal{H}|\exp(-M\epsilon)$, which we want to be *low*, i.e., $|\mathcal{H}|\exp(-M\epsilon) \leq \delta$

Proof of Theorem 1: Finite, Realizable Case

$$|\mathcal{H}|\exp(-M\epsilon) \leq \delta \rightarrow \exp(-M\epsilon) \leq \frac{\delta}{|\mathcal{H}|}$$

$$\rightarrow -M\epsilon \leq \ln\left(\frac{\delta}{|\mathcal{H}|}\right)$$

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

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

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

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 ∃ a bad hypothesis $h_k \in \mathcal{H}$ with $R(h_k) > \epsilon$ and $\hat{R}(h_k) = 0$ is $\leq \delta$

$$\Updownarrow$$

7. 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 $\geq 1 - \delta$

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

7. 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 $\geq 1 - \delta$

$$\Updownarrow$$

8. 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 $\geq 1 - \delta$ ∎
(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

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

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 (setting the two sides equal to each other) 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 $|\mathcal{H}| = \infty$?

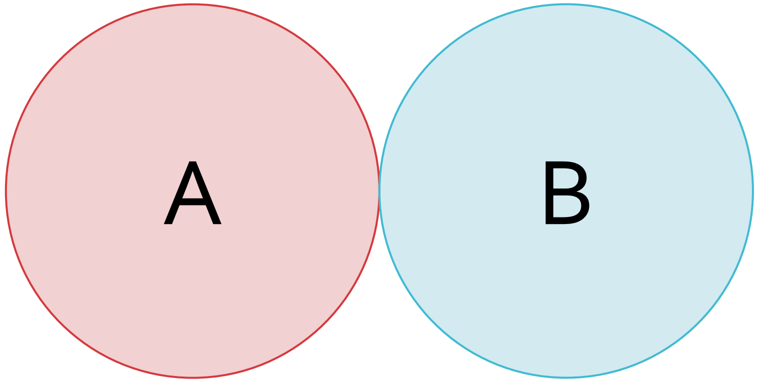
- For a finite hypothesis set $\mathcal{H}$ and arbitrary distribution p^*, given a training data set S s.t. $|S| = M$, all $h \in \mathcal{H}$ have

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

with probability at least $1 - \delta$.

The Union Bound…

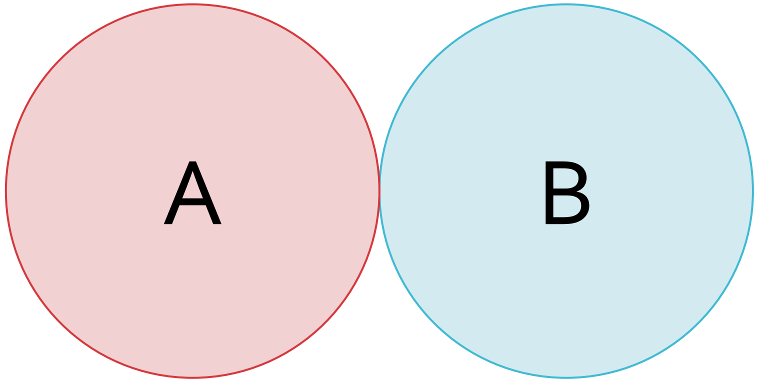
$$P\{A \cup B\} \leq P\{A\} + P\{B\}$$



The Union Bound is Bad!

$$P\{A \cup B\} \leq P\{A\} + P\{B\}$$

$$P\{A \cup B\} = P\{A\} + P\{B\} - P\{A \cap B\}$$

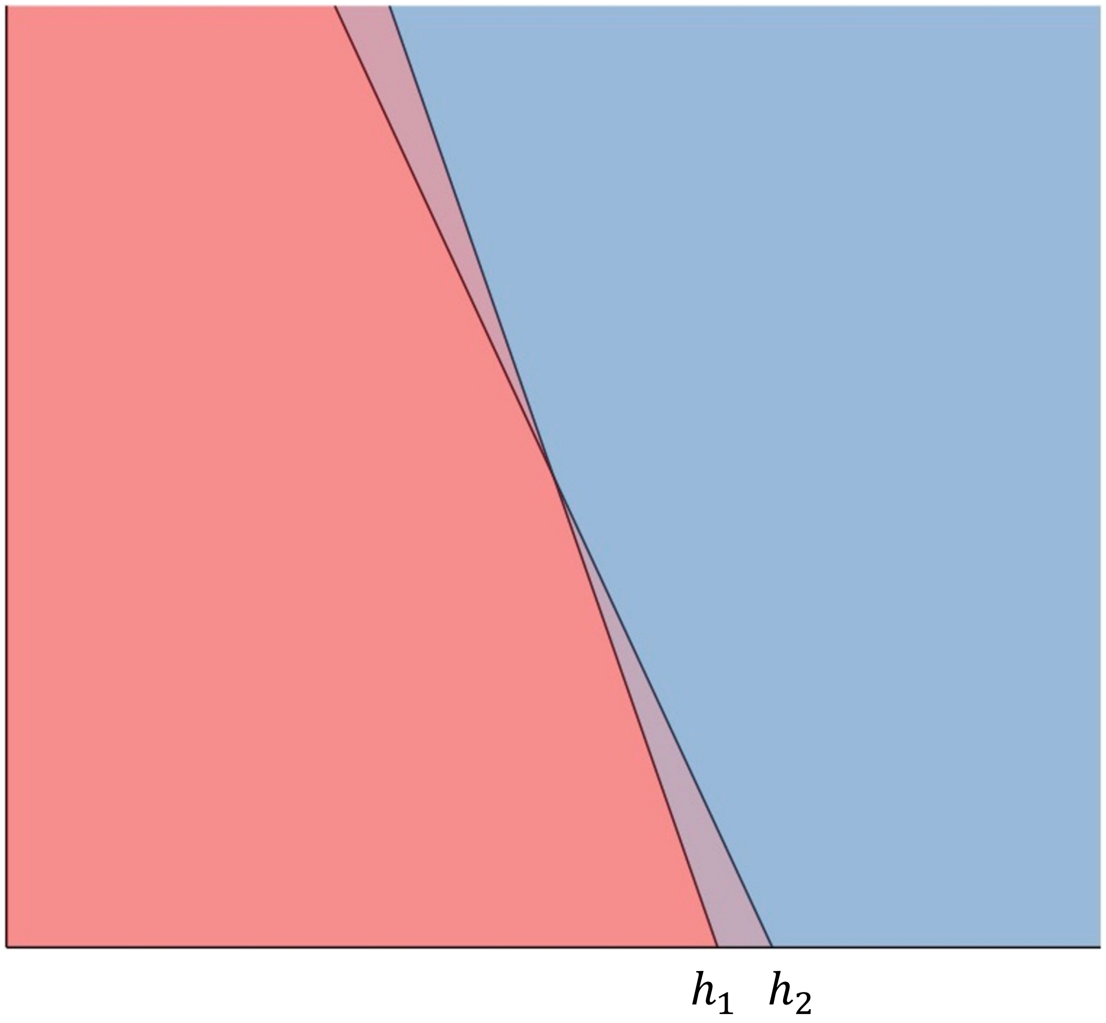


Intuition

If two hypotheses $h_1, h_2 \in \mathcal{H}$ are very similar, then the events

- "h_1 is consistent with the first m training data points"
- "h_2 is consistent with the first m training data points"

will overlap a lot!

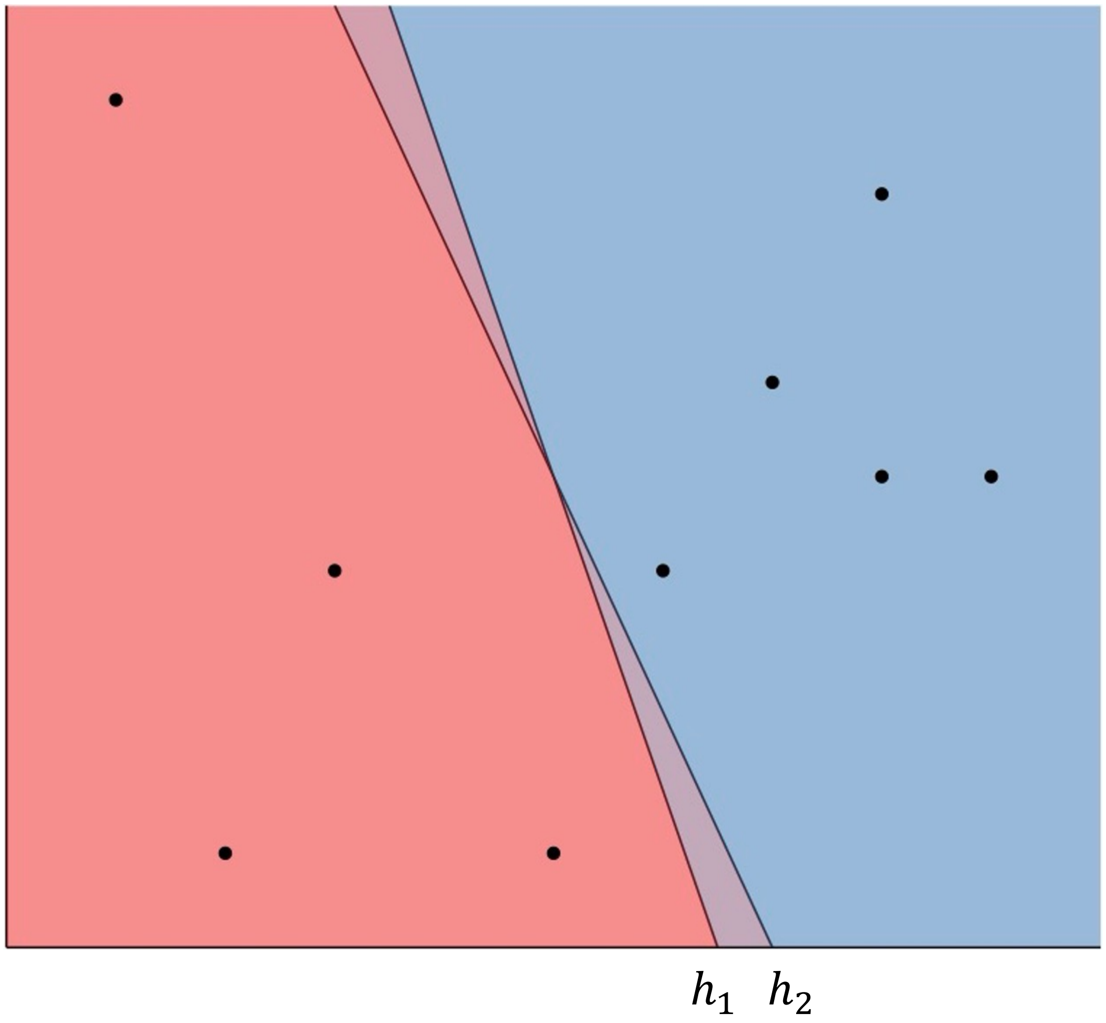


Intuition

If two hypotheses $h_1, h_2 \in \mathcal{H}$ are very similar, then the events

- "h_1 is consistent with the first m training data points"
- "h_2 is consistent with the first m training data points"

will overlap a lot!



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