

10-301/601: Introduction to Machine Learning Lecture 24 – Learning Theory (Finite Case)

Henry Chai

7/15/24

Front Matter

- Announcements
 - HW8 released 7/11, due 7/18 at 11:59 PM
- Recommended Readings
 - Mitchell, Chapters 7.1-7.3

What is Machine Learning 10-301/601?

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

What is Machine Learning 10-301/601?

- Supervised Models
 - Decision Trees
 - KNN
 - Naïve Bayes
 - Perceptron
 - Logistic Regression
 - Linear Regression
 - Neural Networks
- Unsupervised Learning
- Ensemble Methods
- Deep Learning & Generative AI
- **Learning Theory**
- Reinforcement Learning
- 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

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

2. Labels are generated from some *unknown* function

$$y^{(n)} = c^*(\mathbf{x}^{(n)})$$

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)

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

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

$$\begin{aligned}\hat{R}(h) &= P_{\mathbf{x} \sim \mathcal{D}} (c^*(\mathbf{x}) \neq h(\mathbf{x})) \\ &= \frac{1}{N} \sum_{n=1}^N \mathbb{1} (c^*(\mathbf{x}^{(n)}) \neq h(\mathbf{x}^{(n)})) \\ &= \frac{1}{N} \sum_{n=1}^N \mathbb{1} (y^{(n)} \neq h(\mathbf{x}^{(n)}))\end{aligned}$$

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

Three Hypotheses of Interest

1. The *true function*, c^*

2. The *expected risk minimizer*,

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

3. The *empirical risk minimizer*,

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

Which of the following statements must be true?

$$c^* = h^*$$

0%

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

0%

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

0%

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

0%

None of the above

0%

Key Question

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

PAC Learning

- PAC = Probably Approximately Correct
- PAC Criterion:

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

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 $\rightarrow |\mathcal{H}| = \infty$

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, \dots, 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 =
 $P(h_1$ correctly classifies all M training data points \cup
 h_2 correctly classifies all M training data points \cup
 \vdots
 $\cup h_K$ correctly classifies all M training data points)

Proof of Theorem 1: Finite, Realizable Case

$P(h_1$ correctly classifies all M training data points \cup
 h_2 correctly classifies all M training data points \cup
 \vdots
 $\cup h_K$ correctly classifies all M 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 \exists a bad hypothesis $h_k \in \mathcal{H}$ with $R(h_k) > \epsilon$ and $\hat{R}(h_k) = 0$ is $\leq \delta$

\Leftrightarrow

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 an 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 $\leq \delta$



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 $\geq 1 - \delta$

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 $\geq 1 - \delta$

\Leftrightarrow

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 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

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

- 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 $|\mathcal{H}| = \infty$?

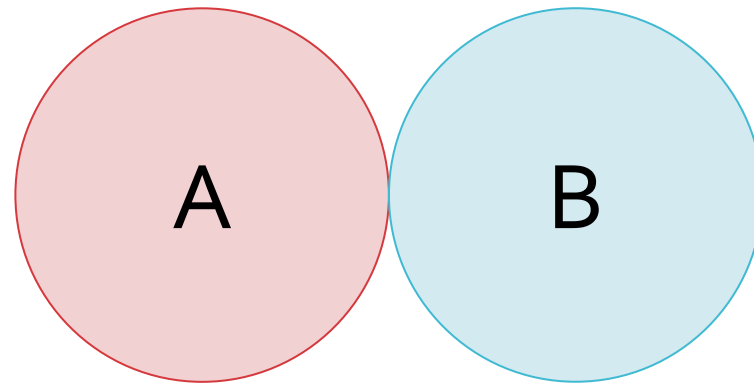
- For a finite hypothesis set \mathcal{H} and arbitrary distribution p^* , given a training data set S s.t. $|S| = M$, all $h \in \mathcal{H}$ have

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

with probability at least $1 - \delta$.

The Union Bound...

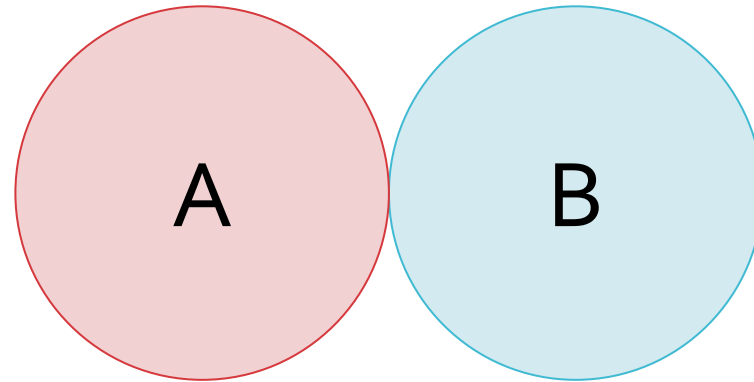
$$P\{A \cup B\} \leq P\{A\} + P\{B\}$$



The Union Bound is Bad!

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

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

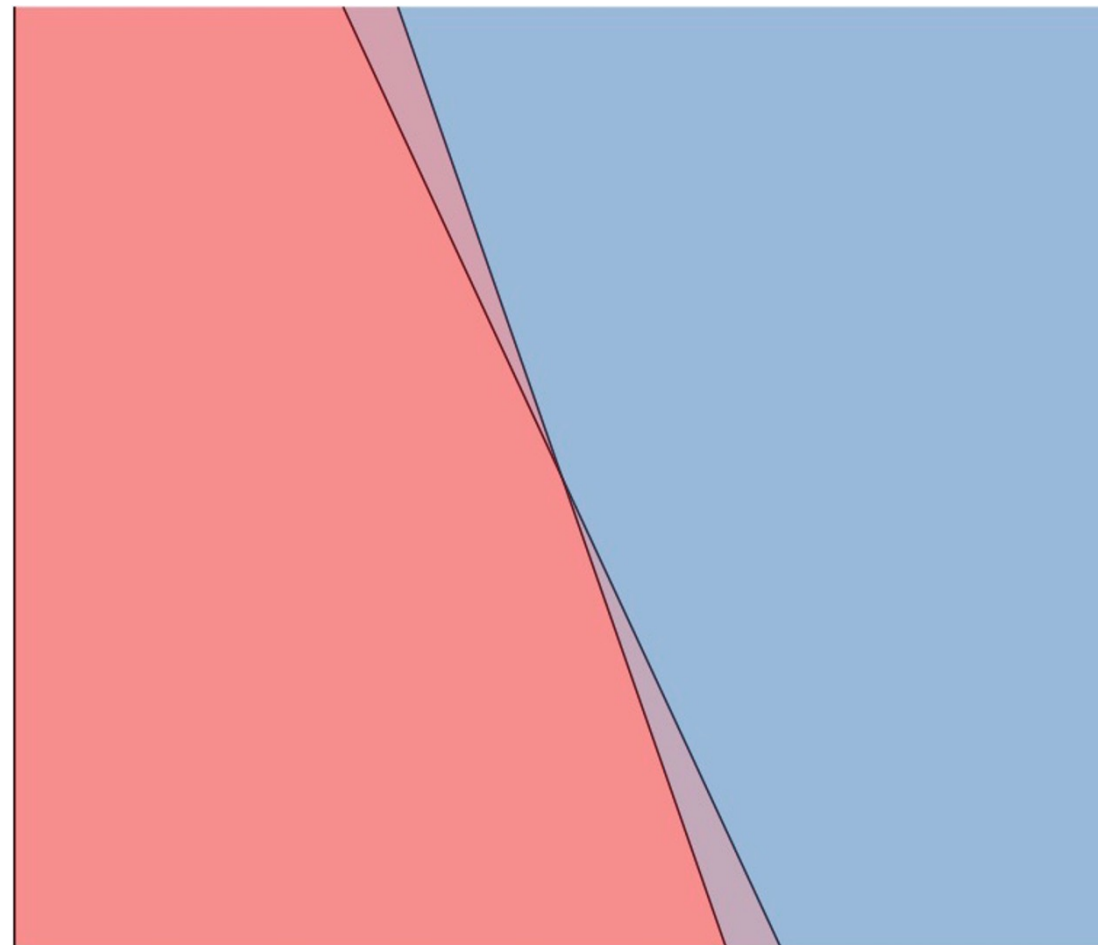


Intuition

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!

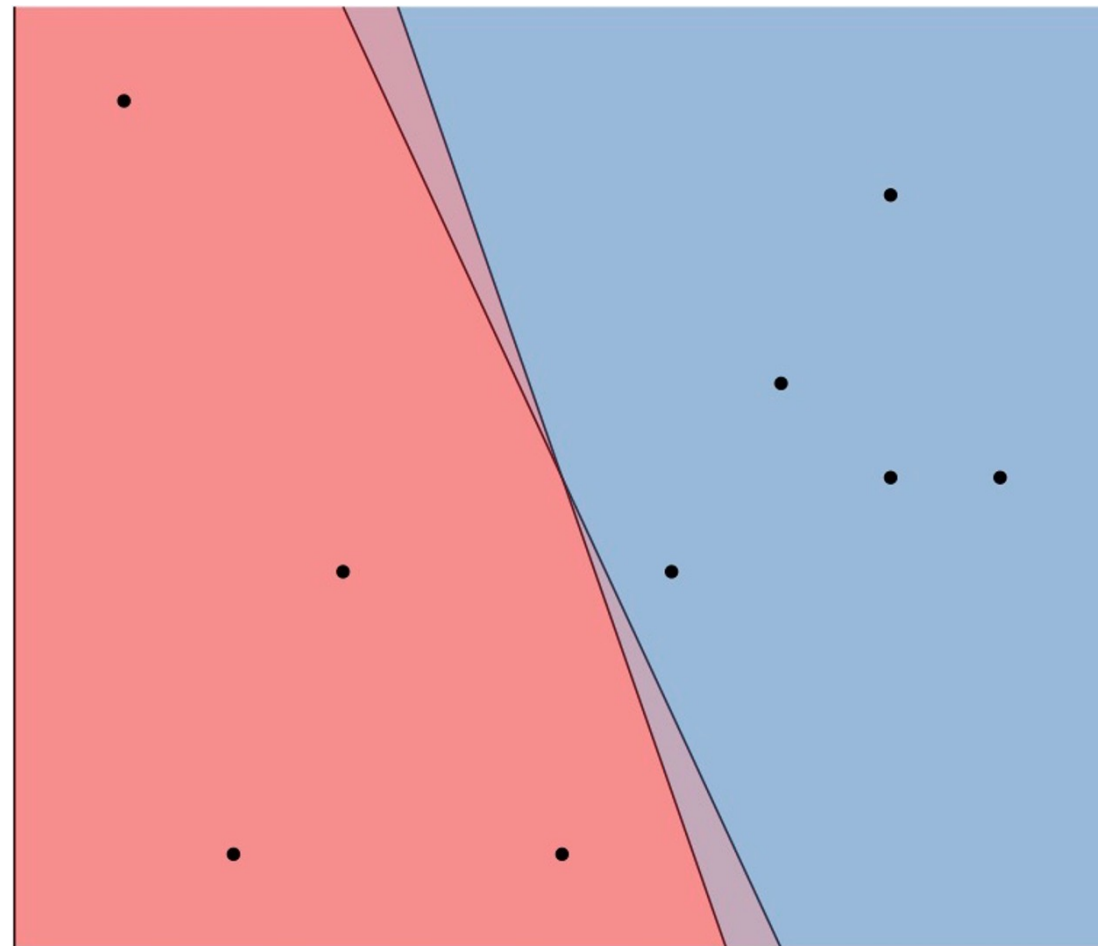


Intuition

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