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

Matt Gormley & Henry Chai

10/21/24

# Front Matter

- Announcements
  - HW5 released 10/9, due 10/27 at 11:59 PM

# 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

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

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)

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

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

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

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,*

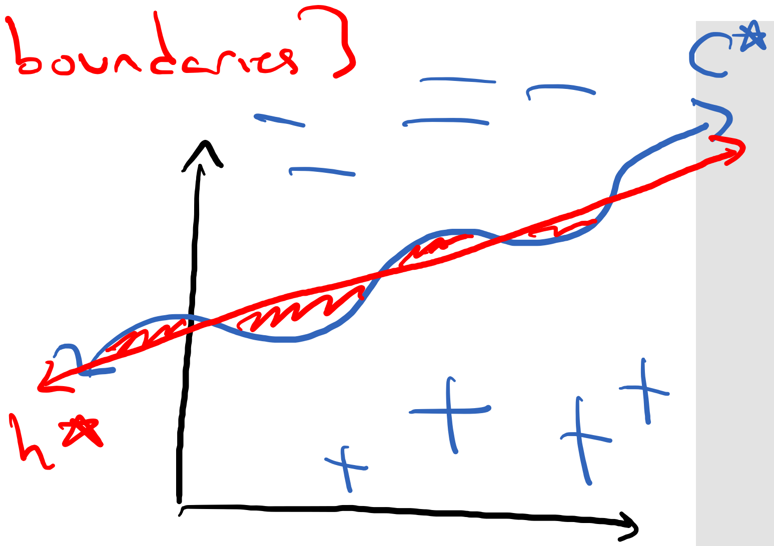
$$\hat{h} = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, \hat{R}(h)$$

$\mathcal{H} = \{$ all linear decision boundaries $\}$

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

A. $c^* = h^*$
B. $c^* = \hat{h}$
C. $h^* = \hat{h}$
D. $c^* = h^* = \hat{h}$
E. None of the above
F. **TOXIC**

- The *true function,* $c^*$




- The *expected risk minimizer,*

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

- 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 $\epsilon$ (difference between expected and empirical risk) and $\delta$ (probability of "failure")

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

# 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 $\delta$ and $\epsilon$
- 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 $R(h) \leq \epsilon$

$P(A \cup B) =$

$P(A) + P(B) - P(A \cap B)$

$\leq P(A) + P(B)$

# Proof of Theorem 1: Finite, Realizable Case

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

2. Pick 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 points) $< (1-\epsilon)^M$

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

$= P(h_1$ correctly classifies all $M$ training data points

$\cup\ h_2$ "

$\vdots$

$\cup\ h_K$ " $)$

# Proof of Theorem 1: Finite, Realizable Case

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

$$\le \sum_{i=1}^{K} P(\hat{R}(h_i)=0) < \sum_{i=1}^{K} (1-\epsilon)^M$$

$$= K(1-\epsilon)^M$$

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

5. P(at least one bad hypothesis correctly classifies all M training data points)

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

6. We want $|H|(1-\epsilon)^M \le \delta$

# Proof of Theorem 1: Finite, Realizable Case

7. Using the fact that $1 - x \le \exp(-x)\ \forall x$

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

$$= |H| \exp(-\epsilon M) \le \delta$$

$$\Rightarrow \exp(-\epsilon M) \le \frac{\delta}{|H|}$$

$$\Rightarrow -\epsilon M \le \log \frac{\delta}{|H|}$$

$$\Rightarrow \epsilon M \ge \log \frac{|H|}{\delta}$$

$$\Rightarrow \epsilon M \ge \log |H| + \log \frac{1}{\delta}$$

$$\Rightarrow M \ge \frac{1}{\epsilon}\left(\log(|H|) + \log \frac{1}{\delta}\right)$$

# Proof of Theorem 1: Finite, Realizable Case

8. Given $M \geq \frac{1}{\epsilon}\left(\log|H| + \log\frac{1}{\delta}\right)$ labelled training data points, the probability $\exists$ a bad hypothesis $h_i \in \mathcal{H}$ where $R(h_i) > \epsilon$ and $\hat{R}(h_i) = 0 \leq \delta$

$$\Updownarrow$$

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

# Proof of Theorem 1: Finite, Realizable Case

# 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 ⇒ Henry brings am umbrella”

  is the same as saying

  “Henry didn’t bring an umbrella ⇒ 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 $\geq 1 - \delta$

$$A \Rightarrow B$$

$$\Updownarrow$$

$$\neg B \Rightarrow \neg A$$

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

# 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

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

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 $\epsilon$ 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 $\epsilon$, e.g., halving $\epsilon$ means we need four times as many labelled training data points
- Again, making the bound tight and solving for $\epsilon$ 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 $1 - \delta$.