10-301/601: Introduction to Machine Learning Lecture 16 – Learning Theory (Finite Case)

Henry Chai

7/3/23

Front Matter

- Announcements
  - **No class or quiz tomorrow** for July 4th
  - PA4 released 6/15, due 7/13 at 11:59 PM
    - You still have one week from this Thursday!
- 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
  - SVMs
  - Linear Regression
  - Neural Networks
- Unsupervised Models
  - K-means
  - GMMs
  - PCA
- Graphical Models
  - Bayesian Networks
  - HMMs
- Learning Theory
- Reinforcement Learning
- Important Concepts
  - Feature Engineering and Kernels
  - Regularization and Overfitting
  - Experimental Design
  - Ensemble Methods

What is ~~Machine Learning~~ 10-301/601?

- Supervised Models
  - Decision Trees
  - KNN
  - Naïve Bayes
  - Perceptron
  - Logistic Regression
  - SVMs
  - Linear Regression
  - Neural Networks
- Unsupervised Models
  - K-means
  - GMMs
  - PCA
- Graphical Models
  - Bayesian Networks
  - HMMs
- **Learning Theory**
- Reinforcement Learning
- Important Concepts
  - Feature Engineering and Kernels
  - Regularization and Overfitting
  - Experimental Design
  - Ensemble Methods

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)

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

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

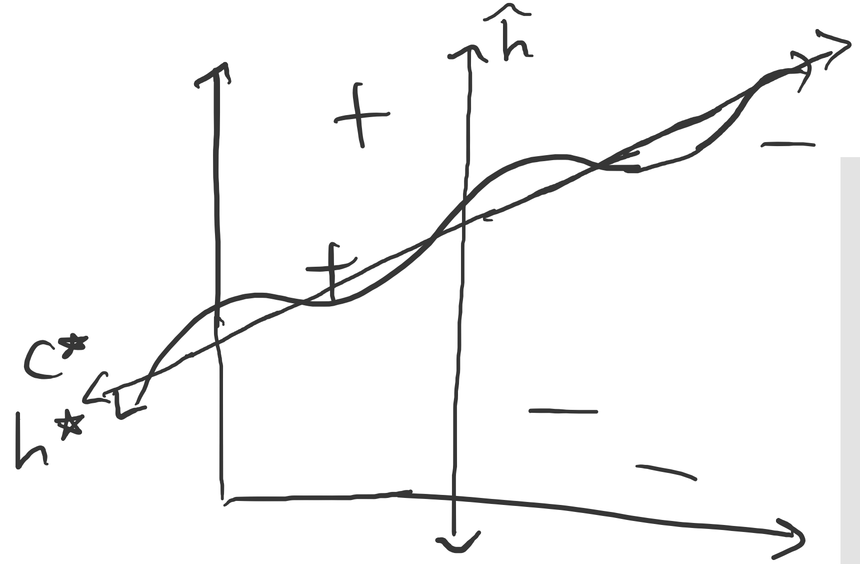
$$\hat{R}(h) = P_{x \sim \mathcal{D}}\left(c^*(x) \neq h(x)\right)$$

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

where $\mathcal{D} = \{(x^{(n)}, y^{(n)})\}_{n=1}^{N}$ and $x \sim \mathcal{D}$ denotes a point uniformly sampled from $\mathcal{D}$

Three Hypotheses of Interest

1. The *true function*, c^*



2. The *expected risk minimizer*,

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

$\mathcal{H}$ = all linear classifiers

3. The *empirical risk minimizer*,

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

Which of the following statements must be true?

- $c^* = h^*$
- $c^* = \hat{h}$
- $h^* = \hat{h}$
- $c^* = h^* = \hat{h}$
- 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 = **P**robably **A**pproximately **C**orrect

- PAC Criterion:

$$P\left(\left|R(h) - \hat{R}(h)\right| \le \epsilon\right) \ge 1 - \delta \; \forall \, h \in \mathcal{H}$$

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

What is the probability that a "bad" hypothesis exists in H where bad means that it has low training error and high test error

1. Assume there are K bad hypotheses in H $h_1, h_2, \ldots, h_K \in H$ and these all have $R(h) > \epsilon$
2. Pick one bad hypothesis, h_i
   - → $P(h_i$ correctly classifies the first training data point$) < 1 - \epsilon$
   - → $P(h_i$ correctly classifies all N training data points$) < (1-\epsilon)^N$

Proof of Theorem 1: Finite, Realizable Case (Continued)

3. P(at least one bad hypothesis $h_1, \ldots, h_K$ correctly classifies all N training data points)

= P(h_1 correctly classifies all N training data points

$\cup$ h_2 " "

$\cup$ h_3 " "

⋮

$\cup$ h_K " ")

$\leq \sum_{k=1}^{K}$ P(h_k correctly classifies all N training data points) by the union bound

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

Proof of Theorem 1: Finite, Realizable Case (Continued)

4. $\sum_{k=1}^{K} P(h_k \text{ correctly classifies all } N \text{ of my training data points}) < \sum_{k=1}^{K} (1-\epsilon)^N$

$= K(1-\epsilon)^N \leq |H|(1-\epsilon)^N$

$(K \leq |H|)$

Use the fact that $1-x \leq \exp(-x) \; \forall x$

$|H|(1-\epsilon)^N \leq |H| \exp(-\epsilon)^N = |H| \exp(-\epsilon N)$

5. P(at least one bad hypothesis achieves 0 empirical risk or training error) $\leq$

$|H| \exp(-\epsilon N)$

Proof of Theorem 1: Finite, Realizable Case (Continued)

6. We want $|H| \exp(-\epsilon N) \leq \delta$

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

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

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

$$\rightarrow \epsilon N \geq -\left(\ln(\delta) - \ln(|H|)\right)$$

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

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

Proof of Theorem 1: Finite, Realizable Case (Continued)

6. Given $M \geq \frac{1}{\epsilon}\left(\log(|\mathcal{H}|) + \log\left(\frac{1}{\delta}\right)\right)$ 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 $M \geq \frac{1}{\epsilon}\left(\log(|\mathcal{H}|) + \log\left(\frac{1}{\delta}\right)\right)$ 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 (Continued)

6. Given $M \geq \frac{1}{\epsilon}\left(\log(|\mathcal{H}|) + \log\left(\frac{1}{\delta}\right)\right)$ 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 $M \geq \frac{1}{\epsilon}\left(\log(|\mathcal{H}|) + \log\left(\frac{1}{\delta}\right)\right)$ 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 (Continued)

6. Given $M \geq \frac{1}{\epsilon}\left(\log(|\mathcal{H}|) + \log\left(\frac{1}{\delta}\right)\right)$ 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 $M \geq \frac{1}{\epsilon}\left(\log(|\mathcal{H}|) + \log\left(\frac{1}{\delta}\right)\right)$ 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 $R(h) \leq \epsilon$

- Solving for ϵ gives...