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

Henry Chai
10/24/22
Front Matter

- Announcements
  - HW5 released 10/13, due 10/27 at 11:59 PM
  - Exam 3 scheduled
    - Thursday, December 15th from 9:30 AM to 11:30 AM
  - Sign up for peer tutoring! See Piazza for more details
Q & A:
Where have you been???

- Sorry, I’ve been training a fresh neural network...
Q & A:

My HW5 code isn’t working, what should I do???

• Review the recitation material!
  • Specifically, test your implementation against the numerical examples our TAs worked through and make sure you’re getting the same values
### Learning Paradigms:
What data is available and when? What form of prediction?
- supervised learning
- unsupervised learning
- semi-supervised learning
- reinforcement learning
- active learning
- imitation learning
- domain adaptation
- online learning
- density estimation
- recommender systems
- feature learning
- manifold learning
- dimensionality reduction
- ensemble learning
- distant supervision
- hyperparameter optimization

### Problem Formulation:
What is the structure of our output prediction?
- boolean
- categorical
- ordinal
- real
- ordering
- multiple discrete
- multiple continuous
- both discrete & cont.

### Theoretical Foundations:
What principles guide learning?
- probabilistic
- information theoretic
- evolutionary search
- ML as optimization

### Facets of Building ML Systems:
How to build systems that are robust, efficient, adaptive, effective?
1. Data prep
2. Model selection
3. Training (optimization / search)
4. Hyperparameter tuning on validation data
5. (Blind) Assessment on test data

### Big Ideas in ML:
Which are the ideas driving development of the field?
- inductive bias
- generalization / overfitting
- bias-variance decomposition
- generative vs. discriminative
- deep nets, graphical models
- PAC learning
- distant rewards

### Application Areas:
Key challenges?
NLP, Speech, Computer Vision, Robotics, Medicine, Search
Learning Paradigms: What data is available and when? What form of prediction?
- supervised learning
- unsupervised learning
- semi-supervised learning
- reinforcement learning
- active learning
- imitation learning
- domain adaptation
- online learning
- density estimation
- recommender systems
- feature learning
- manifold learning
- dimensionality reduction
- ensemble learning
- distant supervision
- hyperparameter optimization

Problem Formulation: What is the structure of our output prediction?
- boolean: Binary Classification
- categorical: Multiclass Classification
- ordinal: Ordinal Classification
- real: Regression
- ordering: Ranking
- multiple discrete: Structured Prediction (e.g. dynamical systems)
- multiple continuous: (e.g. mixed graphical models)
- both discrete & cont.

Theoretical Foundations: What principles guide learning?
- probabilistic
- information theoretic
- evolutionary search
- ML as optimization

Facets of Building ML Systems: How to build systems that are robust, efficient, adaptive, effective?
1. Data prep
2. Model selection
3. Training (optimization / search)
4. Hyperparameter tuning on validation data
5. (Blind) Assessment on test data

Big Ideas in ML: Which are the ideas driving development of the field?
- inductive bias
- generalization / overfitting
- bias-variance decomposition
- generative vs. discriminative
- deep nets, graphical models
- PAC learning
- distant rewards

Application Areas Key challenges?
NLP, Speech, Computer Vision, Robotics, Medicine, Search

10/24/22
1. Data points are generated iid from some *unknown* distribution

\[ x^{(i)} \sim p^*(x) \]

2. Labels are generated from some *unknown* function

\[ y^{(i)} = c^*(x^{(i)}) \]

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

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

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

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

$$= \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(c^*(x^{(i)}) \neq h(x^{(i)}))$$

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

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

- The true function, $c^*$
- The expected risk minimizer,
  $$h^* = \underset{h \in \mathcal{H}}{\text{argmin}} R(h)$$
- The empirical risk minimizer,
  $$\hat{h} = \underset{h \in \mathcal{H}}{\text{argmin}} \hat{R}(h)$$
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,
  \[ h^* = \arg\min_{h \in \mathcal{H}} R(h) \]

- The empirical risk minimizer,
  \[ \hat{h} = \arg\min_{h \in \mathcal{H}} \hat{R}(h) \]
Key Question

• Given a hypothesis with zero/low training error, what can we say about its true error?
The sample complexity of an algorithm/hypothesis set is the number of labelled training data points needed to satisfy the PAC criterion for some $\delta$ and $\epsilon$.

PAC = **Probably Approximately Correct**

PAC Criterion:

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

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 $\delta$
The sample complexity of an algorithm/hypothesis set 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 $R(h) \leq \epsilon$
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 $h_k$ is consistent with (correctly classifies) the first training data point $\leq 1 - \epsilon$
   
   B. Probability that $h_k$ is consistent with (correctly classifies) all $M$ training data points $\leq (1 - \epsilon)^M$

3. Probability that at least one bad hypothesis is consistent with all $M$ training data points $= P(h_1$ is consistent with all $M$ training data points $\cup h_2$ is consistent with all $M$ training data points $\cup \vdots \cup h_K$ is consistent with all $M$ training data points)
Proof of Theorem 1: Finite, Realizable Case

\[ P(h_1 \text{ is consistent with all } M \text{ training data points } \cup \]
\[ h_2 \text{ is consistent with all } M \text{ training data points } \cup \]
\[ \vdots \]
\[ \cup h_K \text{ is consistent with all } M \text{ training data points}) \]
\[ \leq \sum_{k=1}^{K} P(h_k \text{ is consistent with all } M \text{ training data points}) \]

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

\[ \sum_{k=1}^{K} P(h_k \text{ is consistent with all } M \text{ training data points}) \]

\[ \leq k (1 - \epsilon)^M \leq |\mathcal{H}|(1 - \epsilon)^M \]

because \( k \leq |\mathcal{H}| \)

3. Probability that at least one bad hypothesis is consistent with 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 is consistent with 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

\[ |\mathcal{H}| \exp(-M\epsilon) \leq \delta \Rightarrow \exp(-M\epsilon) \leq \frac{\delta}{|\mathcal{H}|} \]

\[ \Rightarrow -M\epsilon \leq \log\left(\frac{\delta}{|\mathcal{H}|}\right) \]

\[ \Rightarrow M \geq \frac{1}{\epsilon}\left(-\log\left(\frac{\delta}{|\mathcal{H}|}\right)\right) \]

\[ \Rightarrow M \geq \frac{1}{\epsilon}\left(\log\left(\frac{|\mathcal{H}|}{\delta}\right)\right) \]

\[ \Rightarrow M \geq \frac{1}{\epsilon}\left(\log(|\mathcal{H}|) + \log\left(\frac{1}{\delta}\right)\right) \]
Proof of Theorem 1: Finite, Realizable Case

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 $\leq \delta$

$\uparrow$

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 $\geq 1 - \delta$
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 by contrapositive)
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”.
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 $R(h) \leq \epsilon$
Let $\mathcal{H}$ be the set of all conjunctions over $M$ Boolean variables, $\mathbf{x} \in \{0,1\}^M$; examples of conjunctions are
\[
\cdot h(\mathbf{x}) = x_1(1 - x_2)x_4x_{10}
\]
\[
\cdot h(\mathbf{x}) = (1 - x_3)(1 - x_4)x_8
\]
Assuming $c^* \in \mathcal{H}$, if $M = 10$, $\epsilon = 0.1$, and $\delta = 0.01$, at least how many labelled examples do we need to satisfy the PAC criterion using Theorem 1?

A. 1 (TOXIC)

B. $10(2 \ln 10 + 100) \approx 92$

C. $10(3 \ln 10 + 100) \approx 116$

D. $10(10 \ln 2 + 100) \approx 116$

E. $10(10 \ln 3 + 100) \approx 156$

F. $100(2 \ln 10 + 10) \approx 691$

G. $100(3 \ln 10 + 10) \approx 922$

H. $100(10 \ln 2 + 10) \approx 924$

I. $100(10 \ln 3 + 10) \approx 1329$
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 $\epsilon$ gives...
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 $\epsilon$, e.g., halving $\epsilon$ means we need four times as many labelled training data points

• Solving for $\epsilon$ gives...
• 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 $1 - \delta$. 
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 $|\mathcal{H}| = \infty$?
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

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

with probability at least $1 - \delta$. 
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

$$R(h) \leq \hat{R}(h) + \infty$$

(not a very meaningful result...) with probability at least $1 - \delta$.

- Insight: $|\mathcal{H}|$ measures how complex our hypothesis set is
- Idea: define a different measure of hypothesis set complexity

What happens when $|\mathcal{H}| = \infty$?
\[ P\{A \cup B\} \leq P\{A\} + P\{B\} \]
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 all $M$ training data points”
- “$h_2$ is consistent with all $M$ 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 all $M$ training data points”
- “$h_2$ is consistent with all $M$ training data points”

will overlap a lot!
Theorem 3: Vapnik-Chervonenkis (VC)-Bound

- Infinite, realizable case: for any hypothesis set $\mathcal{H}$ and distribution $p^*$, if the number of labelled training data points satisfies

$$M = O\left(\frac{1}{\epsilon} \left( d_{VC}(\mathcal{H}) \log \left( \frac{1}{\epsilon} \right) + \log \left( \frac{1}{\delta} \right) \right) \right)$$

then with probability at least $1 - \delta$, all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$

- $d_{VC}(\mathcal{H})$ is the VC-dimension of $\mathcal{H}$, a measure of how complex our hypothesis set is, suitable when $|\mathcal{H}| = \infty$