Learning Theory

Instructor: Jessica Wu -- Harvey Mudd College

The instructor gratefully acknowledges Andrew Ng (Stanford), Eric Eaton (UPenn), David Sontag (NYU), Carlos Guestrin (CMU), Piyush Rai (Utah), and the many others who made their course materials freely available online.

Learning Theory Motivation

Learning Goals

- Discuss the types of questions we can address using learning theory

Computational Learning Theory

(or why ML works)

We have seen a number of learning algorithms

How can we tell if a learning algorithm will do a good job?

- experimental results
- theoretical analysis

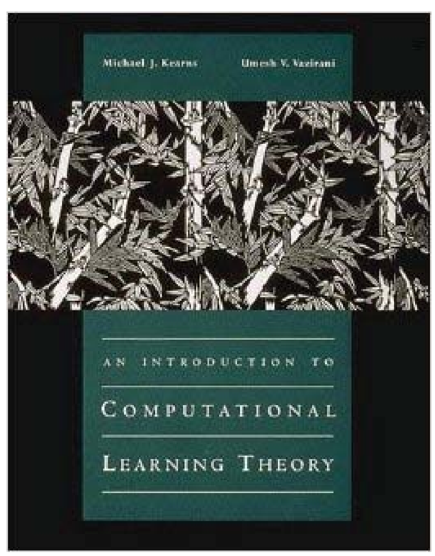
Why theory?

Based on slides by Piyush Rai

Computational Learning Theory

Subfield devoted to mathematical analysis of ML algos

- led to PAC learning and VC theory
  - PAC = probably and approximately correct
  - VC = Vapnik-Chervonenkis



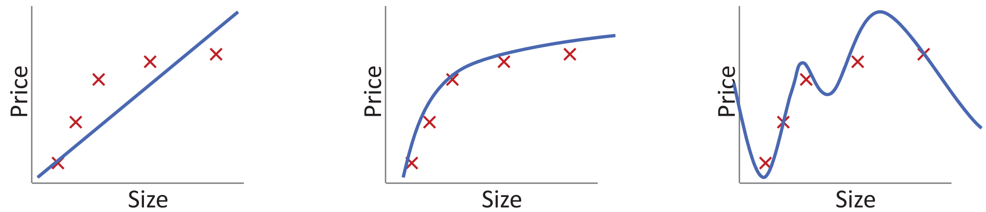
Relate theory to

- probability of successful learning
- number of training examples needed
- complexity of hypothesis space
- accuracy to which target function is approximated
- manner in which training examples should be presented

Annual conference:
Conference on Learning Theory (COLT)

Based on slides by Eric Eaton

Review: Bias-Variance Tradeoff



$\theta_0 + \theta_1 x$ — "simple" model

$\theta_0 + \theta_1 x + \theta_2 x^2$ — correct fit

$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$ — "complex" model

Both simple and complex models have large **generalization error**
BUT problems suffered by two models are very different

If relationship between x and y is not linear, then given large (infinite) training set, fitting linear model would still fail to capture data structure
$\Rightarrow$ model has high **bias**
$\Rightarrow$ bias = expected generalization error if fit to large (infinite) training set (aka structural error)

balance $\Leftrightarrow$

When fitting complex model, large risk of fitting patterns present in small, finite training set but do not reflect wider relationship between x and y
$\Rightarrow$ model has high **variance**
$\Rightarrow$ variance = expected spread in generalization error (aka estimation error)

Based on notes by Andrew Ng

Questions

Can we formalize bias-variance tradeoff?

- Can we automatically decide model complexity?

Why should doing well on training set tell us about generalization error?

- Can we relate error on training set to generalization error?

Are there conditions under which we can actually prove that learning algorithms work well?

Based on notes by Andrew Ng

Setup

- **Hypothesis class** $\mathcal{H}$ is a space of functions
- Learning algorithm learns function (hypothesis) $h \in \mathcal{H}$
- Assume h is learned using sample $\mathcal{D}$ of n iid training examples drawn from $P(\boldsymbol{x}, y)$

- 0/1 **training error** (aka empirical risk) of h

$$R_n(h) = \frac{1}{n} \sum_{i=1}^{N} \mathbb{I}\left[\left[h\left(\boldsymbol{x}^{(i)}\right) \neq y^{(i)}\right]\right]$$

- 0/1 **expected error** (aka risk) of h

$$R(h) = \mathbb{E}_{(\boldsymbol{x},y)\sim P}\left[\mathbb{I}\left[\left[h\left(\boldsymbol{x}\right) \neq y\right]\right]\right]$$

- Expected error is generally worse than training error
  – We want to know how much worse it is
  – ... without doing experiments (e.g. cross-validation)

Based on slides by Piyush Rai

Roadmap

today

- We will start by analyzing finite hypothesis spaces ($|\mathcal{H}| < \infty$) with zero training error ($R_n(h) = 0$) $\Rightarrow$ **Haussler's Theorem**

- We will then generalize to finite hypothesis spaces ($|\mathcal{H}| < \infty$) with non-zero training error ($R_n(h) > 0$) $\Rightarrow$ **General PAC Bounds**

next time

- We will finally discuss infinite hypothesis spaces ($|\mathcal{H}| = \infty$) $\Rightarrow$ **VC-dimension**

Learning Theory for Finite Hypothesis Spaces

Learning Goals

- State PAC bounds
- Apply PAC bounds

Facebook Example (fictional)

- FB holds competition for best face recognition classifier (+1 if image contains face, -1 if not)
- FB receives 20k submissions
  - FB evaluates all 20k submissions on *n* labeled images (not previously shown to competitors) and chooses winner
  - Winner obtains 98% accuracy on *n* images
- FB already has algorithm known to be 95% accurate
  - Should FB deploy winner's algorithm?
  - FB cannot risk doing worse … would be PR disaster!

Based on slides by David Sontag

Generalization of Finite Hypothesis Spaces

Theorem [Haussler '88]

Given finite hypothesis space $\mathcal{H}$, dataset $\mathcal{D}$ with n iid samples, and probability of error on one sample > ϵ (where $0 \leq \epsilon \leq 1$), then for any learned hypothesis h that is consistent with the training data ($R_n(h) = 0$),

$$P(R(h) > \epsilon) \leq |\mathcal{H}| e^{-n\epsilon}$$

Observations

- Probability of h being "bad" (zero training error, positive generalization error) decreases exponentially with n
- While zero errors in training set does not imply zero errors in test set, it does bound expected error

Based on slides by Carlos Guestrin and David Sontag

Using a PAC Bound

(probably and approximately correct)

By Haussler's theorem, for all consistent h,

$$P(R(h) > \epsilon) \leq |\mathcal{H}| e^{-n\epsilon}$$

Suppose we are willing to tolerate at most a δ probability of having $> \epsilon$ error.

$$P(R(h) > \epsilon) \leq |\mathcal{H}| e^{-n\epsilon} \leq \delta$$
$$\ln(|\mathcal{H}| e^{-n\epsilon}) \leq \ln(\delta)$$
$$\ln(|\mathcal{H}|) - n\epsilon \leq \ln(\delta)$$

We have 2 typical use cases:

1) Pick ϵ and δ. Compute n.

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

larger hypothesis space $|\mathcal{H}|$ requires more examples

higher error threshold ϵ allows fewer examples

higher δ (larger tolerance) allows fewer examples

This gives the sufficient number of examples for which the learned hypothesis will be probably (with probability $1 - \delta$) and approximately (with error ϵ) correct. $\Rightarrow$ **PAC learning**

Based on slides by Carlos Guestrin and David Sontag

Using a PAC Bound

(probably and approximately correct)

We know that for all consistent h,

$$P(R(h) > \epsilon) \leq |\mathcal{H}| e^{-n\epsilon}$$

Suppose we are willing to tolerate at most a δ probability of having $>\epsilon$ error.

$$P(R(h) > \epsilon) \leq |\mathcal{H}| e^{-n\epsilon} \leq \delta$$
$$\ln(|\mathcal{H}| e^{-n\epsilon}) \leq \ln(\delta)$$
$$\ln(|\mathcal{H}|) - n\epsilon \leq \ln(\delta)$$

We have 2 typical use cases:

We supposed $P(R(h) > \epsilon) \leq \delta$.
Then $P(R(h) \leq \epsilon) > 1 - \delta$.
In other words, with probability at least $1 - \delta$, we can upper-bound generalization error $R(h) \leq \epsilon$.

2) Pick n and δ. Compute ϵ.

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

larger hypothesis space $|\mathcal{H}|$ raises error bound

more examples n lowers error bound

higher δ (larger tolerance) lowers error bound

Based on slides by Carlos Guestrin and David Sontag

Limitations of Haussler '88 Bound

There may be no consistent hypothesis h (where $R_n(h) = 0$)

$\Rightarrow$ extend to non-zero training error

The size of the hypothesis space $|\mathcal{H}|$ may be really big or continuous

$\Rightarrow$ extend to infinite sized hypothesis spaces

Based on slides by Carlos Guestrin and David Sontag

Extending to Non-Zero Training Error

So far…

- Learner with zero training errors ($R_n(h) = 0$) may make mistakes on test set ($R(h) > \epsilon$)

What if our classifier has $R_n(h) > 0$?

- Can we relate $R(h)$ to $R_n(h)$? That is, can we find bound on generalization error $R(h)$ for learner h with training error $R_n(h)$?

General PAC Bounds

Theorem [Generalization Bound for $|\mathcal{H}|$ Hypotheses]

Given finite hypothesis space $\mathcal{H}$, dataset $\mathcal{D}$ with n iid samples, and probability of error on one sample > ϵ (where $0 \leq \epsilon \leq 1$), then for any learned hypothesis h,

$$P(R(h) - R_n(h) > \epsilon) \leq |\mathcal{H}|e^{-2n\epsilon^2}$$

Compare to Haussler's Theorem

For any learned hypothesis h that is consistent with training data ($R_n(h) = 0$),

$$P(R(h) > \epsilon) \leq |\mathcal{H}|e^{-n\epsilon}$$

Using a PAC Bound

For all h,

$$P(R(h) - R_n(h) > \epsilon) \leq |\mathcal{H}|e^{-2n\epsilon^2}$$

As before, suppose we are willing to tolerate at most a δ probability of having $> \epsilon$ error.

$$P(R(h) - R_n(h) > \epsilon) \leq |\mathcal{H}|e^{-2n\epsilon^2} \leq \delta$$

$$n \geq \frac{1}{2\epsilon^2}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right)$$

n grows as *square* of $(1/\epsilon)$
for zero-error case, n grows *linearly* with $(1/\epsilon)$
$\Rightarrow$ since $\epsilon < 1$, then for given ϵ and δ, non-zero training error case requires more examples

$$\epsilon \geq \sqrt{\frac{1}{2n}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right)}$$

We supposed $P(R(h) - R_n(h) > \epsilon) \leq \delta$.
Then $P(R(h) - R_n(h) \leq \epsilon) > 1 - \delta$.
In other words, with probability at least $1 - \delta$, we have $R(h) - R_n(h) \leq \epsilon$.
That is, we can upper-bound generalization error $R(h) \leq R_n(h) + \epsilon$.

PAC Bound and Bias-Variance Tradeoff

With probability at least $1 - \delta$,

$$R(h) \leq \underbrace{R_n(h)}_{\text{bias}} + \underbrace{\sqrt{\frac{1}{2n}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right)}}_{\text{variance}}$$

For large $|\mathcal{H}|$

- low bias (assuming we can find good $h \in \mathcal{H}$)
- high variance (because bound is looser)

For small $|\mathcal{H}|$

- high bias (is there a good $h \in \mathcal{H}$?)
- low variance (because bound is tighter)

Important:

- PAC bound holds for all $h \in \mathcal{H}$.
- It does *not* guarantee that algorithm finds best h!

Based on slides by Carlos Guestrin and David Sontag

Facebook Example (fictional)

- FB holds competition for best face recognition classifier (+1 if image contains face, -1 if not)
- FB receives 20k submissions
  - FB evaluates all 20k submissions on *n* labeled images (not previously shown to competitors) and chooses winner
  - Winner obtains 98% accuracy on *n* images
- FB already has algorithm known to be 95% accurate
  - Should FB deploy winner's algorithm?
  - FB cannot risk doing worse … would be PR disaster!

Applying PAC Bounds to Facebook

$R(\mathrm{FB}) = 0.05$ (existing system)

new system

- suppose we want at least 99% confidence

$$R(h) \leq R_n(h) + \sqrt{\frac{1}{2n}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right)} \quad \delta = 0.01$$

$R_n(h) = 0.02$ $|\mathcal{H}|$ = 20k models

- what if $n = 100$?

$$R(h) \leq 0.02 + \sqrt{\frac{1}{200}(\ln 20k + \ln 100)} \approx 0.29$$

$\Rightarrow$

- what if $n = 10\mathrm{k}$?

$$R(h) \leq 0.02 + \sqrt{\frac{1}{20k}(\ln 20k + \ln 100)} \approx 0.047$$

$\Rightarrow$

(extra slides)

Learning Theory Proofs

Learning Goals

- Glimpse into the black-box
  - Formally prove Haussler's Theorem
  - Gain intuition towards proving general PAC bounds

How Likely will a Bad Hypothesis be Consistent with the Training Set?

Assume **finite hypothesis space** ($|\mathcal{H}| < \infty$) with some $h \in \mathcal{H}$ with **zero training error** ($R_n(h) = 0$)

we will generalize later

Hypothesis h is "bad" if $R_n(h) = 0$ and $R(h) > \epsilon$

- h gets all training points right despite true error $> \epsilon$

How likely is a bad hypothesis to get n data points correct?

Interpretation

$$P(h \text{ gets } n \text{ iid data points } \textit{right} \mid R(h) > \epsilon) \leq e^{-n\epsilon}$$

What This Says

- If true error $> \epsilon$, then h gets n data points right with very low probability ($P \leq e^{-n\epsilon}$)

Equivalent Statement

- If h gets n data points right with very high probability ($P > 1 - e^{-n\epsilon}$), then it is close to perfect ($R(h) \leq \epsilon$)

Are We Done?

No! This only considers **one** hypothesis!

We need to account for **multiple hypotheses**

- Suppose 1 billion people entered competition, and each submitted a *random* function
- For small enough n, one submission could be consistent *by chance* despite all submissions having very large true error

How Likely will At Least One Bad Hypothesis be Consistent with the Training Set?

Let $\mathcal{H}_\epsilon \subseteq \mathcal{H}$ be set of hypotheses with $R(h) > \epsilon$

- How likely will any $h \in \mathcal{H}_\epsilon$ be consistent with training data?
- We need a bound that holds for all $h \in \mathcal{H}_\epsilon$!

Lemma [Union Bound]
Let $A_1, A_2, \ldots, A_k$ be k different events (not necessarily independent). Then
$P(A_1 \cup \ldots \cup A_k)$
$\leq P(A_1) + \ldots + P(A_k)$.

Intuitively, the probability of any one of k events happening is at most the sums of the probabilities of the k different events. (The bound is tight for disjoint events.)

What is the probability that at least one $h \in \mathcal{H}$ is "bad"?

Based on slides by Carlos Guestrin and David Sontag

(This slide intentionally left blank.)

Generalization Error of Finite Hypothesis Spaces

Theorem [Haussler '88]

Given finite hypothesis space $\mathcal{H}$, dataset $\mathcal{D}$ with n iid samples, and probability of error on one sample > ϵ (where $0 \leq \epsilon \leq 1$), then for any learned hypothesis h that is consistent with the training data ($R_n(h) = 0$),

$$P(R(h) > \epsilon) \leq |\mathcal{H}| e^{-n\epsilon}$$

Next

Extending to Non-Zero Training Error

(This slide intentionally left blank.)

Simpler Question: What is the Expected Error of a Hypothesis?

Lemma [Chernoff Bound] (aka Hoeffding inequality)

Let $Z_1, \ldots, Z_m$ be m iid random variables drawn from a $\mathrm{Bernouilli}(\phi)$ distribution, i.e. $P(Z_i = 1) = \phi$ and $P(Z_i = 0) = 1 - \phi$. Let $\hat{\phi} = \frac{1}{m}\sum_{i=1}^{m} Z_i$ be the mean of these r.v.s, and let any $\gamma > 0$ be fixed. Then

$$P(|\phi - \hat{\phi}| > \gamma) \leq 2e^{-2\gamma^2 m}$$

Idea: If we take $\hat{\phi}$ (the average of m $\mathrm{Bernouilli}(\phi)$ r.v.s) to be our estimate of ϕ, then the probability of our being far away from the true value is small so long as m is large.

- Example: Suppose you have a coin whose chance of landing on heads is ϕ. If you toss it m times and calculate the fraction of times that it came up heads, that will be a good estimate of ϕ with high probability (if m is large).

Based on notes by Andrew Ng

Generalization Error for $|\mathcal{H}|$ Hypotheses

Applying similar reasoning as before

- For a single hypothesis $h \in \mathcal{H}$, apply Chernoff bound

$$P(R(h) - R_n(h) > \epsilon) \leq e^{-2n\epsilon^2}$$

- For at least one hypothesis $h \in \mathcal{H}$, apply Union bound

$$P(R(h) - R_n(h) > \epsilon) \leq |\mathcal{H}|e^{-2n\epsilon^2}$$

Theorem [Generalization Bound for $|\mathcal{H}|$ Hypotheses]

Given finite hypothesis space $\mathcal{H}$, dataset $\mathcal{D}$ with n iid samples, and probability of error on one sample > ϵ ($0 \leq \epsilon \leq 1$), then for any learned hypothesis h,

$$P(R(h) - R_n(h) > \epsilon) \leq |\mathcal{H}|e^{-2n\epsilon^2}$$

Based on slides by Carlos Guestrin and David Sontag