Learning Theory

Aarti Singh and Eric Xing

Machine Learning 10-701/15-781
Nov 5, 2012

Slides courtesy: Carlos Guestrin

MACHINE LEARNING DEPARTMENT
Carnegie Mellon.
School of Computer Science

Learning Theory

- We have explored **many** ways of learning from data
- But…
  - How good is our classifier, really?
  - How much data do I need to make it "good enough"?

A simple setting

- Classification
  - m i.i.d. data points
  - **Finite** number of possible hypothesis (e.g., dec. trees of depth d)
- A learner finds a hypothesis *h* that is **consistent** with training data
  - Gets zero error in training, $\text{error}_{\text{train}}(h) = 0$
- What is the probability that *h* has more than ε true error?
  - $\text{error}_{\text{true}}(h) \geq \varepsilon$

Even if *h* makes zero errors in training data, may make errors in test

How likely is a bad hypothesis to get m data points right?

- Hypothesis *h* that is **consistent** with training data → got *m* i.i.d. points right
  - h "bad" if it gets all this data right, but has high true error

- Prob. *h* with $error_{true}(h) \geq \varepsilon$ gets one data point right
  $\leq 1 - \varepsilon$

- Prob. *h* with $error_{true}(h) \geq \varepsilon$ gets *m* data points right
  $\leq (1 - \varepsilon)^m$

How likely is a learner to pick a bad hypothesis?

- Usually there are many possible hypothesis that are consistent with training data.

- If there are k hypothesis consistent with data, how likely is learner to pick a bad one?

$$\begin{aligned}&\text{Prob}(\text{error}_{true}(h_1) \geq \varepsilon \text{ and } h_1 \text{ consistent OR}\\&\quad \text{error}_{true}(h_2) \geq \varepsilon \text{ and } h_2 \text{ consistent OR} \ldots \text{ OR}\\&\quad \text{error}_{true}(h_k) \geq \varepsilon \text{ and } h_k \text{ consistent})\end{aligned}$$

$$\begin{aligned}\leq\ &\text{Prob}(\text{error}_{true}(h_1) \geq \varepsilon \text{ and } h_1 \text{ consistent}) +\\&\quad \text{Prob}(\text{error}_{true}(h_2) \geq \varepsilon \text{ and } h_2 \text{ consistent}) + \ldots +\\&\quad \text{Prob}(\text{error}_{true}(h_k) \geq \varepsilon \text{ and } h_k \text{ consistent})\end{aligned}$$

Union bound

Loose but works

$$\leq \quad k\,(1-\varepsilon)^m$$

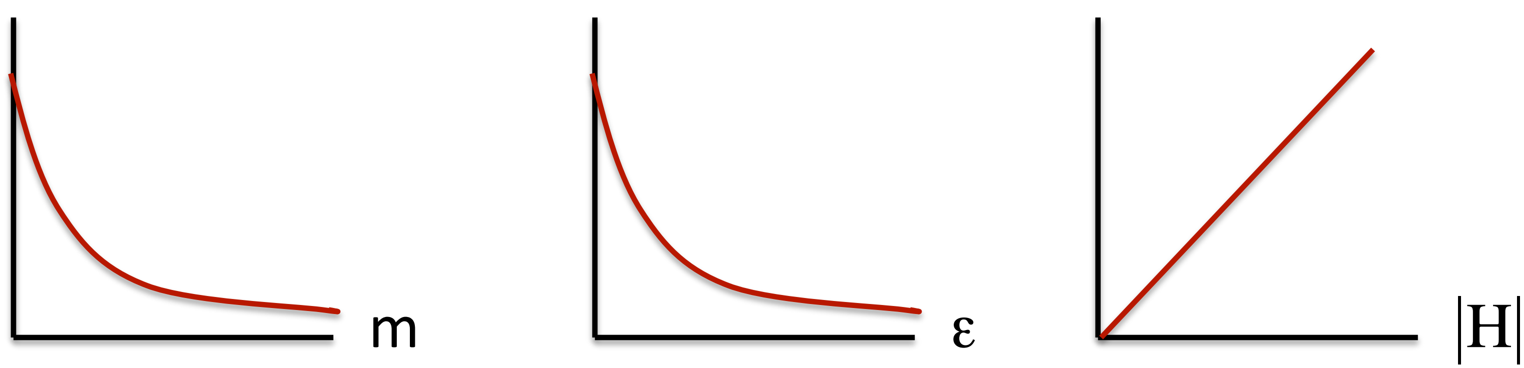
How likely is a learner to pick a bad hypothesis?

- Usually there are many possible hypothesis that are consistent with training data.

- If there are k hypothesis consistent with data, how likely is learner to pick a bad one?

$$\leq \; k\,(1-\varepsilon)^m \quad \leq \; |H|\,(1-\varepsilon)^m \leq \; |H|\,e^{-\varepsilon m}$$

$|H|$ → Size of hypothesis class



PAC (Probably Approximately Correct) bound

- ***Theorem [Haussler'88]***: Hypothesis space *H* finite, dataset *D* with *m* i.i.d. samples, $0 < \varepsilon < 1$: for any learned hypothesis *h* that is consistent on the training data:

$$P(\mathrm{error}_{true}(h) \geq \epsilon) \leq |H|e^{-m\epsilon} \leq \delta$$

- Equivalently, with probability $\geq 1 - \delta$

$$\mathrm{error}_{true}(h) \leq \epsilon$$

Important: PAC bound holds for all *h*, but doesn't guarantee that algorithm finds best *h*!!!

Using a PAC bound

$$|H|e^{-m\epsilon} \leq \delta$$

- Given ε and δ, yields sample complexity

#training data, $m \geq \dfrac{\ln |H| + \ln \frac{1}{\delta}}{\epsilon}$

- Given m and δ, yields error bound

error, $\epsilon \geq \dfrac{\ln |H| + \ln \frac{1}{\delta}}{m}$

Limitations of Haussler'88 bound

- Consistent classifier

  h such that zero error in training, $\text{error}_{\text{train}}(h) = 0$

- Dependence on Size of hypothesis space

$$m \geq \frac{\ln |H| + \ln \frac{1}{\delta}}{\epsilon}$$

what if |H| too big or H is continuous?

What if our classifier does not have zero error on the training data?

- A learner with zero training errors may make mistakes in test set
- What about a learner with $error_{train}(h) \neq 0$ in training set?
- The error of a hypothesis is like estimating the parameter of a coin!

$$error_{true}(h) := P(h(X) \neq Y) \equiv P(H=1) =: \theta$$

$$error_{train}(h) := \frac{1}{m}\sum_i \mathbf{1}_{h(X_i)\neq Y_i} \equiv \frac{1}{m}\sum_i Z_i =: \widehat{\theta}$$

Hoeffding's Bound for a single hypothesis

- Consider *m* i.i.d. flips $x_1,\ldots,x_m$, where $x_i \in \{0,1\}$ of a coin with parameter θ. For $0<\varepsilon<1$:

$$P\left(\left|\theta - \frac{1}{m}\sum_i x_i\right| \geq \epsilon\right) \leq 2e^{-2m\epsilon^2}$$

- For a single hypothesis h

$$P\left(|\text{error}_{true}(h) - \text{error}_{train}(h)| \geq \epsilon\right) \leq 2e^{-2m\epsilon^2}$$

PAC bound for |H| hypotheses

- For each hypothesis h_i:

$$P\left(|\text{error}_{true}(h_i) - \text{error}_{train}(h_i)| \geq \epsilon\right) \leq 2e^{-2m\epsilon^2}$$

- What if we are comparing |H| hypotheses?

  Union bound

- ***Theorem***: Hypothesis space *H* finite, dataset *D* with *m* i.i.d. samples, $0 < \varepsilon < 1$: for any learned hypothesis $h \in H$:

$$P\left(|\text{error}_{true}(h) - \text{error}_{train}(h)| \geq \epsilon\right) \leq 2|H|e^{-2m\epsilon^2} \leq \delta$$

Important: PAC bound holds for all *h*, but doesn't guarantee that algorithm finds best *h*!!!

PAC bound and Bias-Variance tradeoff

$$P\left(|\text{error}_{true}(h) - \text{error}_{train}(h)| \geq \epsilon\right) \leq 2|H|e^{-2m\epsilon^2} \leq \delta$$

- Equivalently, with probability $\geq 1 - \delta$

$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{\ln|H| + \ln\frac{2}{\delta}}{2m}}$$

- Fixed m

hypothesis space	$\text{error}_{train}(h)$ ↓	$\sqrt{\frac{\ln\lvert H\rvert + \ln\frac{2}{\delta}}{2m}}$ ↓
complex	small	large
simple	large	small

What about the size of the hypothesis space?

$$2|H|e^{-2m\epsilon^2} \leq \delta$$

- Sample complexity

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

- How large is the hypothesis space?

Number of decision trees of depth k

Recursive solution:

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

Given *n* attributes

H_k = Number of decision trees of depth k

$H_0 = 2$

H_k = (#choices of root attribute)

*(# possible left subtrees)

*(# possible right subtrees) $= n * H_{k-1} * H_{k-1}$

Write $L_k = \log_2 H_k$

$L_0 = 1$

$L_k = \log_2 n + 2L_{k-1} = \log_2 n + 2(\log_2 n + 2L_{k-2})$

$= \log_2 n + 2\log_2 n + 2^2\log_2 n + \ldots + 2^{k-1}(\log_2 n + 2L_0)$

So $L_k = (2^k-1)(1+\log_2 n) + 1$

PAC bound for decision trees of depth k

$$m \geq \frac{\ln 2}{2\epsilon^2}\left((2^k - 1)(1 + \log_2 n) + 1 + \log_2 \frac{2}{\delta}\right)$$

- Bad!!!
  - Number of points is exponential in depth k!

- But, for *m* data points, decision tree can't get too big…

Number of leaves never more than number data points

Number of decision trees with k leaves

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

H_k = Number of decision trees with k leaves

H_1 =2

H_k = (#choices of root attribute) *

[(# left subtrees wth 1 leaf)*(# right subtrees wth k-1 leaves)

+ (# left subtrees wth 2 leaves)*(# right subtrees wth k-2 leaves)

+ …

+ (# left subtrees wth k-1 leaves)*(# right subtrees wth 1 leaf)]

$$H_k = n\sum_{i=1}^{k-1} H_i H_{k-i} = n^{k-1} C_{k-1}$$

(C_{k-1} : Catalan Number)

Loose bound (using Sterling's approximation):

$$H_k \leq n^{k-1} 2^{2k-1}$$

Number of decision trees

- With k leaves

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

$$\log_2 H_k \leq (k-1)\log_2 n + 2k - 1$$

linear in k

number of points m is linear in #leaves

- With depth k

$$\log_2 H_k = (2^k-1)(1+\log_2 n) +1$$

exponential in k

number of points m is exponential in depth

PAC bound for decision trees with k leaves – Bias-Variance revisited

With prob ≥ 1-δ $\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{\ln|H| + \ln\frac{2}{\delta}}{2m}}$

With $H_k \leq n^{k-1}2^{2k-1}$, we get

$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{(k-1)\ln n + (2k-1)\ln 2 + \ln\frac{2}{\delta}}{2m}}$$

	↓	↓
k = m	0	large (~ > ½)
k < m	>0	small (~ <½)

What did we learn from decision trees?

- Bias-Variance tradeoff formalized

$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{(k-1)\ln n + (2k-1)\ln 2 + \ln \frac{2}{\delta}}{2m}}$$

- Moral of the story:

  Complexity of learning not measured in terms of size hypothesis space, but in maximum *number of points* that allows consistent classification

  – Complexity *m* – no bias, lots of variance

  – Lower than *m* – some bias, less variance

What about continuous hypothesis spaces?

$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{\ln |H| + \ln \frac{2}{\delta}}{2m}}$$

- Continuous hypothesis space:
  - $|H| = \infty$
  - Infinite variance???

- **As with decision trees, only care about the maximum number of points that can be classified exactly!**