CSE446: PAC-learning, VC Dimension

Winter 2015

Luke Zettlemoyer

Slides adapted from Carlos Guestrin

What now…

- 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 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 – $error_{train}(h) = 0$
- What is the probability that *h* has more than ε true error?
  - $error_{true}(h) \geq \varepsilon$

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
  - What is the probability of this happening?
- Prob. *h* with $\text{error}_{\text{true}}(h) \geq \varepsilon$ gets randomly drawn data point right

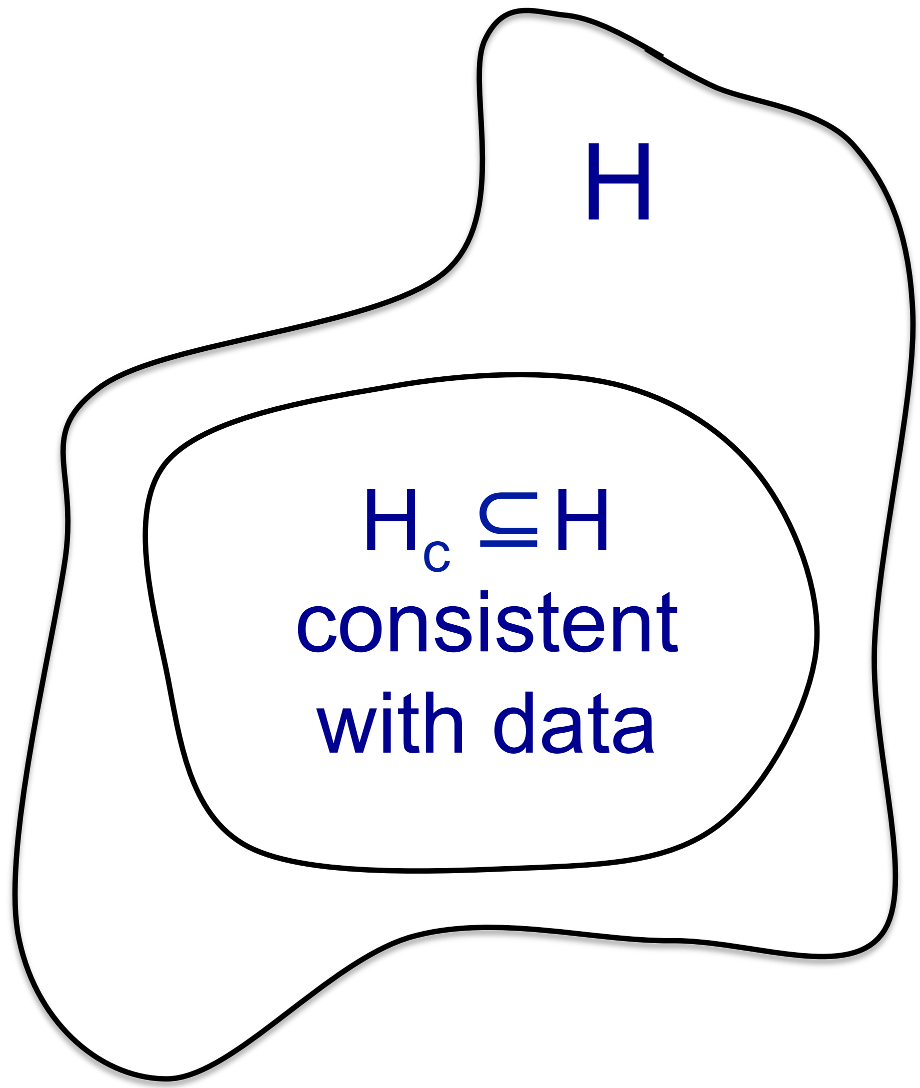
$$P(\text{error}_{\text{true}}(h) \geq \varepsilon, \text{gets one data point right}) \leq 1-\varepsilon$$

- Prob. *h* with $\text{error}_{\text{true}}(h) \geq \varepsilon$ gets *m* iid data points right

$$P(\text{error}_{\text{true}}(h) \geq \varepsilon, \text{gets m } \textit{iid} \text{ data point right}) \leq (1-\varepsilon)^m$$

But there are many possible hypothesis that are consistent with training data

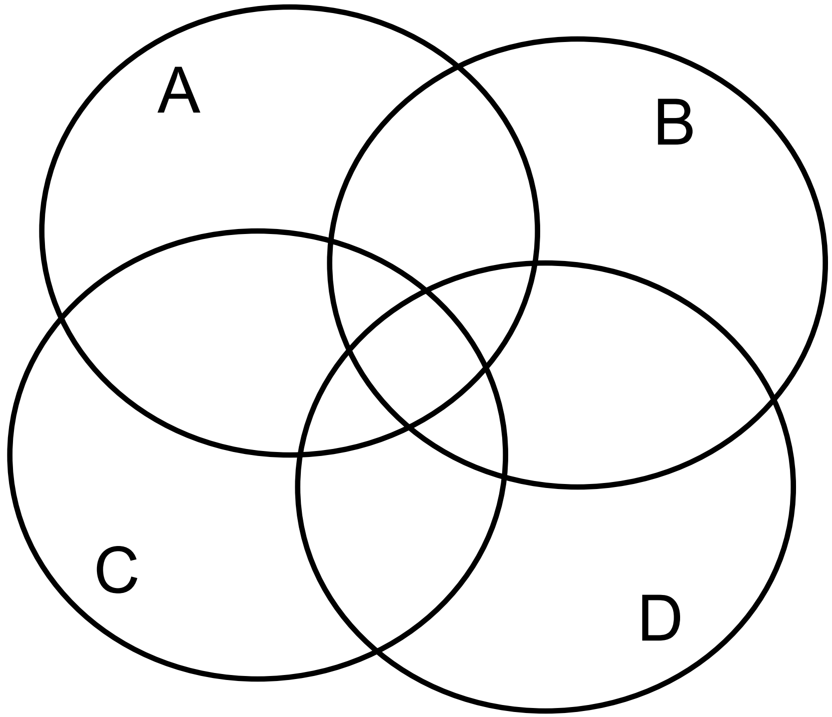
- Which classifier should be learn?
  - and how to we generalize the bounds?
- We want to make as few assumptions as possible!
- So, pick any $h \in H_c$
- But wait, we had a bound on a single h, now we need to bound the worst $h \in H_c$



Union bound

- P(A or B or C or D or …)

  $\leq$ P(A) + P(B) + P(C) + P(D) + …



Q: Is this a tight bound? Will it be useful?

How likely is learner to pick a bad hypothesis

$P(\text{error}_{true}(h) \geq \varepsilon$, gets m *iid* data point right$) \leq (1-\varepsilon)^m$

There are *k* hypothesis consistent with data

- How likely is learner to pick a bad one?
- We need to a bound that holds for all of them!

$$P(\text{error}_{true}(h_1) \geq \varepsilon \text{ OR } \text{error}_{true}(h_1) \geq \varepsilon \text{ OR } \ldots \text{ OR } \text{error}_{true}(h_k) \geq \varepsilon)$$

$$\leq \sum_k P(\text{error}_{true}(h_k) \geq \varepsilon) \quad \leftarrow \text{Union bound}$$

$$\leq \sum_k (1-\varepsilon)^m \quad \leftarrow \text{bound on individual } h_j\text{s}$$

$$\leq |H|(1-\varepsilon)^m \quad \leftarrow k \leq |H|$$

$$\leq |H|\, e^{-m\varepsilon} \quad \leftarrow (1-\varepsilon) \leq e^{-\varepsilon} \text{ for } 0 \leq \varepsilon \leq 1$$

Generalization error in finite hypothesis spaces [Haussler '88]

- ***Theorem***: Hypothesis space *H* finite, dataset *D* with *m* i.i.d. samples, 0 < ε < 1 : for any learned hypothesis *h* that is consistent on the training data:

$$P(\text{error}_{true}(h) > \epsilon) \leq |H|e^{-m\epsilon}$$

Using a PAC bound

Argument: For all h we know that

$$P(\text{error}_{true}(h) > \epsilon) \leq |H|e^{-m\epsilon}$$

so, with probability 1-δ the following holds…

- Typically, 2 use cases:
  - 1: Pick ε and δ, compute *m*
  - 2: Pick m and δ, compute ε

$$P(error_{true}(h) \leq \epsilon) \leq |H|e^{-m\epsilon} \leq \delta$$

$$\ln\left(|H|e^{-m\epsilon}\right) \leq \ln\delta$$

$$\ln|H| - m\epsilon \leq \ln\delta$$

Case 1

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

Case 2

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

Log dependence on |H|, ok if exponential size (but not doubly)

ε has stronger influence than δ

ε shrinks at rate O(1/m)

Limitations of Haussler '88 bound

$$P(\text{error}_{true}(h) > \epsilon) \leq |H|e^{-m\epsilon}$$

- Do we really want to pick a consistent hypothesis h? (where $error_{train}(h)=0$)
- Size of hypothesis space
  - What if $|H|$ is really big?
  - What if it is continuous?
- First Goal: Can we get a bound for a learner with $error_{train}(h)$ in training set?

Question: What's the expected error of a hypothesis?

- The error of a hypothesis is like estimating the parameter of a coin!

- Chernoff bound: for *m* i.i.d. coin flips, $x_1,\ldots,x_m$, where $x_i \in \{0,1\}$. For $0<\varepsilon<1$:

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

Generalization bound for |H| hypothesis

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

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

Why? Same reasoning as before. Use the Union bound over individual Chernoff bounds

PAC bound and Bias-Variance tradeoff

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

or, after moving some terms around,

with probability at least 1-δ:

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

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

PAC bound and Bias-Variance tradeoff

for all h, with probability at least 1-δ:

$$\text{error}_{true}(h) \leq \underbrace{\text{error}_{train}(h)}_{\text{“bias”}} + \underbrace{\sqrt{\frac{\ln|H| + \ln\frac{1}{\delta}}{2m}}}_{\text{“variance”}}$$

- For large |H|
  - low bias (assuming we can find a good h)
  - high variance (because bound is looser)
- For small |H|
  - high bias (is there a good h?)
  - low variance (tighter bound)

PAC bound: How much data?

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

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

- Given δ,ε how big should m be?

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

Decision Trees

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

- Bound number of decision trees with depth k with data that has *n* features:

$$2 * (2n)^{2^k - 1}$$

- Bad!!! Need exponentially many data points (in k)!!!

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

- But, for *m* data points, tree can't get too big…
  - **Number of leaves never more than number data points**
  - **Instead, lets bound number of decision trees with k leaves**

$$H_k = n^{k-1}(k+1)^{2k-1}$$

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

$$H_k = n^{k-1}(k+1)^{2k-1}$$

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

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

Bias / variance again

- k << m: high bias, low variance
- k=m: no bias, high variance
- k>m: we would never do this!!!

What did we learn from decision trees?

- Bias-Variance tradeoff formalized

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

- 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

What about continuous hypothesis spaces?

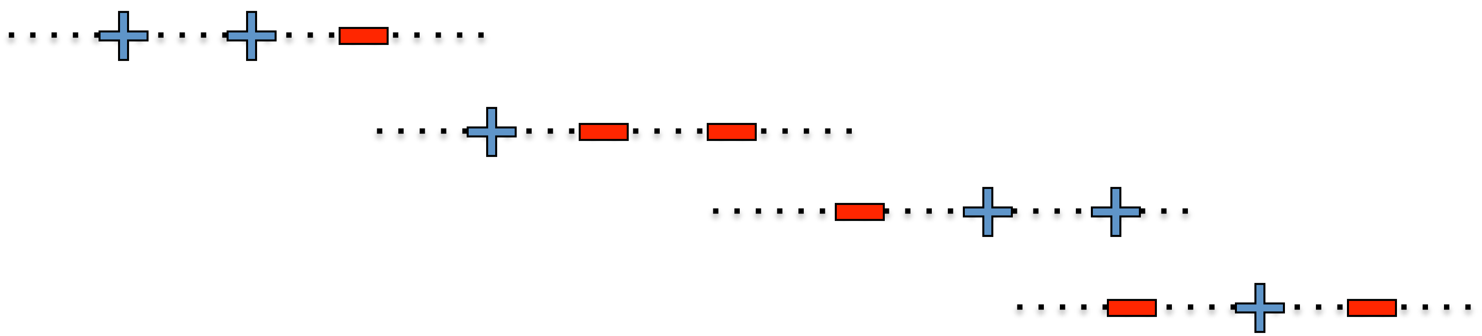
$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{\ln|H| + \ln\frac{1}{\delta}}{2m}}$$

- Continuous hypothesis space:
  - $|H| = \infty$
  - Infinite variance???
- **As with decision trees, only care about the maximum number of points that can be classified exactly!**

How many points can a linear boundary classify exactly? (1-D)

2 Points: Yes!!

3 Points: No…



etc (8 total)

Shattering and VC Dimension

A set of points is *shattered* by a hypothesis space H iff:

- For all ways of *splitting* the examples into positive and negative subsets
- There exists some *consistent* hypothesis h

The *VC Dimension* of H over input space X

- The size of the *largest* finite subset of X shattered by H

How many points can a linear boundary classify exactly? (2-D)

3 Points: Yes!!

etc.

4 Points: No…

etc.

How many points can a linear boundary classify exactly? (d-D)

- A linear classifier $w_0+\sum_{j=1..d} w_j x_j$ can represent all assignments of possible labels to d+1 points
  - But not d+2!!
  - Bias term w_0 required!
  - Rule of Thumb: number of parameters in model often matches max number of points
- Question: Can we get a bound for error in as a function of the number of points that can be completely labeled?

PAC bound using VC dimension

- VC dimension: number of training points that can be classified exactly (shattered) by hypothesis space H!!!
  - Measures relevant size of hypothesis space, as with decision trees with k leaves

$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{VC(H)\left(\ln \frac{2m}{VC(H)} + 1\right) + \ln \frac{4}{\delta}}{m}}$$

- Same bias / variance tradeoff as always
  - Now, just a function of VC(H)

Examples of VC dimension

$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{VC(H)\left(\ln \frac{2m}{VC(H)} + 1\right) + \ln \frac{4}{\delta}}{m}}$$

- Linear classifiers:
  - VC(H) = d+1, for *d* features plus constant term *b*
- Neural networks (we will see this next)
  - VC(H) = #parameters
  - Local minima means NNs will probably not find best parameters
- 1-Nearest neighbor
  - VC(H) = ∞
- SVM with Gaussian Kernel
  - VC(H) = ∞

What you need to know

- Finite hypothesis space
  - Derive results
  - Counting number of hypothesis
  - Mistakes on Training data
- Complexity of the classifier depends on number of points that can be classified exactly
  - Finite case – decision trees
  - Infinite case – VC dimension
- Bias-Variance tradeoff in learning theory
- Remember: will your algorithm find best classifier?