Learning Theory

Machine Learning – CSE546
Carlos Guestrin
University of Washington

November 25, 2013

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
  - N data points iid
  - **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?
  - $error_{true}(h) \geq \varepsilon$ for any $\varepsilon > 0$

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

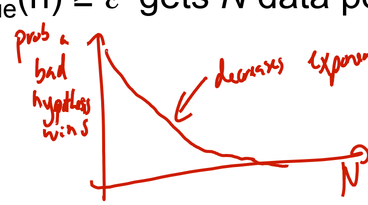
- Hypothesis *h* that is **consistent** with training data → got *N* i.i.d. points right
  - h "bad" if it gets all this data right, but has high true error $\varepsilon > 0$
- Prob. *h* with $error_{true}(h) \geq \varepsilon$ gets one data point right

less than $1-\varepsilon$

if error true $\varepsilon \geq 0.25$
75% prob h will get 1 point right

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

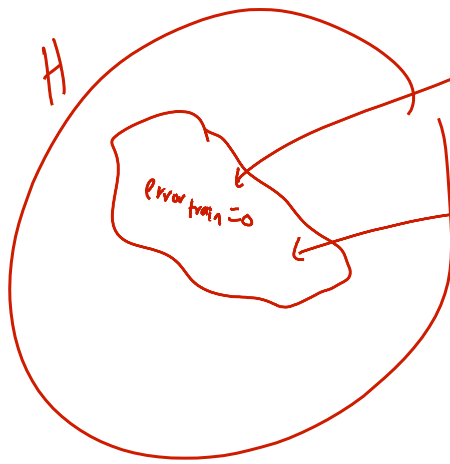
less $(1-\varepsilon)^N$



decreases exponentially in N

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

H



algorithm picked one of these

worst-case analysis

guarantee that there are no bad hypothesis consistent with the data

How likely is learner to pick a bad hypothesis

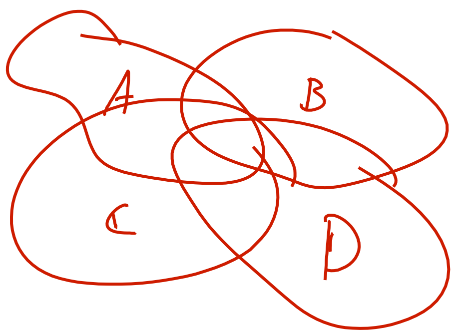
- Prob. a particular h with $\text{error}_{\text{true}}(h) \geq \varepsilon$ gets N data points right
  less $(1-\varepsilon)^N$
- There are k hypothesis consistent with data $h_1, \ldots, h_k$
  - How likely is learner to pick a bad one?

$$P(\exists h \ \text{error train} = 0 \ \& \ \text{error true} \geq \varepsilon)$$

$$= P(\text{error true}(h_1) \geq \varepsilon \text{ OR } \text{error true}(h_2) \geq \varepsilon \text{ OR} \ldots \text{ OR } \text{error true}(h_k) \geq \varepsilon)$$

Union bound

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



How likely is learner to pick a bad hypothesis

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

  less than $(1-\varepsilon)^N$

- There are *k* hypothesis consistent with data
  - How likely is it that learner will pick a bad one out of these *k* choices?

union bound

$\leq P(\exists\, h: \text{error}_{train}(h) = 0 \ \&\ \text{error}_{true}(h) \geq \varepsilon) \leq k(1-\varepsilon)^N$

$\leq |H|\,(1-\varepsilon)^N$

what's k?

$k \leq |H|$

total # hypotheses

crazy loose

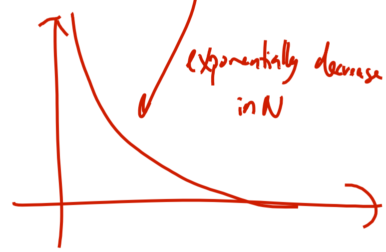
Generalization error in finite hypothesis spaces [Haussler '88]

- ***Theorem***: Hypothesis space H finite, dataset D with N i.i.d. samples, $0 < \varepsilon < 1$: for any learned hypothesis h that is consistent on the training data: $error_{train}(h)=0$

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

$0 \leq \varepsilon \leq 1 \qquad (1-\varepsilon) \leq e^{-\varepsilon}$

$|H|(1-\varepsilon)^N \leq |H| e^{-N\varepsilon}$



exponentially decrease in N

Using a PAC bound

- Typically, 2 use cases:
  - 1: Pick ε (error) and δ (← prob lose your job), give you N
  - 2: Pick N and δ, give you ε

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

Prob(lose job) = prob(error true(h) ≥ ε)

$\leq |H| e^{-N\varepsilon} \leq \delta$

$\ln|H| - N\varepsilon \leq \ln \delta$

$\Rightarrow N \geq \dfrac{\ln|H| + \ln\frac{1}{\delta}}{\varepsilon}$

log dependence on |H|

log dependence on tolerance

collect

linear dependence on accuracy

$\varepsilon \geq \dfrac{\ln|H| + \ln\frac{1}{\delta}}{N}$

decrease at rate $O(\frac{1}{N})$

great rate

more general setting, decrease as $O\left(\frac{1}{\sqrt{N}}\right)$

Summary: Generalization error in finite hypothesis spaces [Haussler '88]

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

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

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

Limitations of Haussler '88 bound

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

- Consistent classifier

$error_{train}(h) = 0 \Rightarrow$ highly unrealistic

→ label noise ⇒ two data points with same x have different y

→ overfit

- Size of hypothesis space

ln |H| can be bad

→ finite |H| if |H| is very very large

→ |H| is infinite, e.g. in SVMs or LR

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 $error_{train}(h)$ in training set?

what happens error train (h) > 0

⇒ error true (h)?

Simpler question: What's the expected error of a hypothesis?

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

$\theta \approx \hat{\theta} = \frac{3}{5}$

- Chernoff bound: for N i.i.d. coin flips, $x^1,\ldots,x^N$, where $x^j \in \{0,1\}$. For $0<\varepsilon<1$:

$$P\left(\theta - \frac{1}{N}\sum_{j=1}^{N} x^j > \epsilon\right) \leq e^{-2N\epsilon^2}$$

truth

3/5 estimate of mean

ε

Something ← decrease exponentially in N

Using Chernoff bound to estimate error of a single hypothesis

$$P\left(\theta - \frac{1}{N}\sum_{j=1}^{N} x^j > \epsilon\right) \leq e^{-2N\epsilon^2}$$

$\hat{\theta} = error_{train}(h) \leftarrow \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}(h(x^j) \neq y^j)$

$\theta = error_{true}(h)$

$= \int_X p(x)\, \mathbb{1}(h(x) \neq f(x))\, dx$

in reality bad h

h looks good

$P(error_{true}(h) - error_{train}(h) > \varepsilon) \leq e^{-2N\varepsilon^2}$

But we are comparing many hypothesis: **Union bound**

For each hypothesis h_i:

$$P\left(error_{true}(h_i) - error_{train}(h_i) > \epsilon\right) \leq e^{-2N\epsilon^2}$$

What if I am comparing two hypothesis, h_1 and h_2?

is there an h_2 that is truly better than my h_1

$P(error_{train}(h_1) < error_{train}(h_2)$ but $error_{true}(h_1) > error_{true}(h_2))$

instead worst case analysis

$P([error_{true}(h_1) - error_{train}(h_1) > \varepsilon]$ OR $[error_{true}(h_2) - error_{train}(h_2)])$

union bound

$\leq 2e^{-2N\varepsilon^2}$

Generalization bound for |H| hypothesis

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

$$P\left(error_{true}(h_i) - error_{train}(h_i) > \epsilon\right) \leq e^{-2N\epsilon^2}$$

hold $\forall h$: $P(\text{error true}(h) - \text{error train}(h) > \varepsilon) \leq |H| e^{-2N\varepsilon^2}$

for $\delta > 0$ $\quad \varepsilon \geq \sqrt{\frac{\ln|H| + \ln \frac{1}{\delta}}{2N}}$ $\Rightarrow$ rate is only $O\left(\frac{1}{\sqrt{N}}\right)$

much worse than $O\left(\frac{1}{N}\right)$

but still works!! :)

©Carlos Guestrin 2005-2013 17

PAC bound and Bias-Variance tradeoff

$$P\left(error_{true}(h) - error_{train}(h) > \epsilon\right) \leq |H| e^{-2N\epsilon^2}$$

or, after moving some terms around, ε from previous slide
with probability at least 1-δ:

$$error_{true}(h) \leq error_{train}(h) + \sqrt{\frac{\ln|H| + \ln\frac{1}{\delta}}{2N}}$$

	"bias"	"variance"
"complex hypothesis space"	low	large ⇐ |H| is large
"simple hypothesis space"	high	low ⇐ |H| is small

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

©Carlos Guestrin 2005-2013 18

What about the size of the hypothesis space?

$$N \geq \frac{\ln |H| + \ln \frac{1}{\delta}}{2\epsilon^2}$$

- How large is the hypothesis space?

|H|?

|H| is really large

⇒ ln|H| = only large

⇒ ok

|H| = really really large

⇒ ln|H| = really large

⇒ lots of data needed

Boolean formulas with *m* binary features

$X_1, \ldots, X_m$ $X_1 \wedge \neg X_2 \vee X_3 \wedge \neg X_2 \ldots$

$$N \geq \frac{\ln |H| + \ln \frac{1}{\delta}}{2\epsilon^2}$$

H: any boolean formula |H|?

what does one h represent:

$X_1 \ldots X_m$ | Y

0 0 0 0 0 0 0 | 1

0 0 0 0 0 0 1 | 0

⋮

1 1 1 1 1 1 1

2^m rows

2 possibilities for Y in each row

$|H| = 2^{2^m}$

$\ln|H| = 2^m \ln 2$ ← exponentially large

H: conjunctions only:

$X_1 \wedge \neg X_2 \wedge X_7$

each feature → positive, → negated, → absent } 3 possibilities

$|H| = 3^m$ ← only really large

$\ln|H| = m \ln 3$

↑ linear in # of features

Number of decision trees of depth k

$$N \geq \frac{\ln|H| + \ln\frac{1}{\delta}}{2\epsilon^2}$$

Recursive solution
Given *m* attributes
H_k = Number of decision trees of depth k
H_0 =2
H_{k+1} = (#choices of root attribute) *
(# possible left subtrees) *
(# possible right subtrees)
= m * H_k * H_k

Write $L_k = \log_2 H_k$
$L_0 = 1$
$L_{k+1} = \log_2 m + 2L_k$
So $L_k = (2^k-1)(1+\log_2 m) +1$

Simplify to

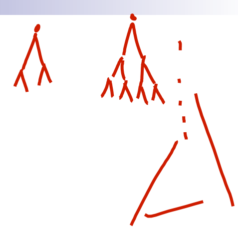
$\ln|H| \leq 2^k \log m$

really really large in depth

really nice in terms of # of features

PAC bound for decision trees of depth k

$$N \geq \frac{\overbrace{2^k \log m}^{\ln|H|} + \ln\frac{1}{\delta}}{\epsilon^2}$$



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

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

no reason to have more than N leaves

Number of leaves never more than number data points

Number of Decision Trees with k Leaves

- Number of decision trees of depth k is really really big:
  - ln |H| is about $2^k \log m$
- Decision trees with up to k leaves:
  - |H| is about $m^k k^{2k}$ ← only really large
    - A very loose bound

$\ln|H| \leq k \ln m + 2k \ln k$

much better

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

$$\ln|H_{\text{DTs k leaves}}| \leq 2k(\ln m + \ln k)$$

$$error_{true}(h) \leq error_{train}(h) + \sqrt{\frac{\overbrace{\ln|H|}^{\ln|H|} + \ln\frac{1}{\delta}}{2N}}$$

$$error_{true}(h) \leq error_{train}(h) + \sqrt{\frac{2k(\ln m + \ln k) + \ln\frac{1}{\delta}}{2N}}$$

	"bias"	"variance"
$k \approx N$	goes to 0	LARGE bound > 1
$k << N$	potentially larger	much smaller

What did we learn from decision trees?

- Bias-Variance tradeoff formalized

$$error_{true}(h) \leq error_{train}(h) + \sqrt{\frac{2k(\ln m + \ln k) + \ln \frac{1}{\delta}}{2N}}$$

- 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 N – no bias, lots of variance
  - Lower than N – some bias, less variance

What about continuous hypothesis spaces?

$$error_{true}(h) \leq error_{train}(h) + \sqrt{\frac{\ln |H| + \ln \frac{1}{\delta}}{2N}}$$

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

∞ many planes

- **As with decision trees, only care about the maximum number of points that can be classified exactly!**
  - **Called VC dimension… see readings for details**

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?