Learning Theory

Machine Learning – CSE546
Carlos Guestrin
University of Washington

October 27, 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
  - **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?
  - $\text{error}_{\text{true}}(h) \geq \varepsilon$

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

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

How likely is learner to pick a bad hypothesis

- Prob. *h* with $error_{true}(h) \geq \varepsilon$ gets *N* data points right
- There are *k* hypothesis consistent with data
  - How likely is learner to pick a bad one?

Union bound

- P(A or B or C or D or …)

©Carlos Guestrin 2005-2013

How likely is learner to pick a bad hypothesis

- Prob. a particular *h* with $\mathrm{error}_{true}(h) \geq \varepsilon$ gets *N* data points right
- There are *k* hypothesis consistent with data
  - How likely is it that learner will pick a bad one out of these *k* choices?

©Carlos Guestrin 2005-2013

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:

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

©Carlos Guestrin 2005-2013

Using a PAC bound

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

- Typically, 2 use cases:
  - 1: Pick ε and δ, give you N
  - 2: Pick N and δ, give you ε

©Carlos Guestrin 2005-2013

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

- ***Theorem***: Hypothesis space *H* finite, dataset *D* with *N* i.i.d. samples, 0 < ε < 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

©Carlos Guestrin 2005-2013

Limitations of Haussler '88 bound

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

- Consistent classifier

- Size of hypothesis space

©Carlos Guestrin 2005-2013

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?

©Carlos Guestrin 2005-2013

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

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

©Carlos Guestrin 2005-2013

Using Chernoff bound to estimate error of a single hypothesis

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

©Carlos Guestrin 2005-2013

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?

©Carlos Guestrin 2005-2013

Generalization bound for |H| hypothesis

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

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

PAC bound and Bias-Variance tradeoff

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

or, after moving some terms around, with probability at least 1-δ:

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

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

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?

©Carlos Guestrin 2005-2013

Boolean formulas with *m* binary features

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

©Carlos Guestrin 2005-2013

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

PAC bound for decision trees of depth k

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

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

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

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 $m^k k^{2k}$
    - A very loose bound

©Carlos Guestrin 2005-2013

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

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

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

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

©Carlos Guestrin 2005-2013

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

©Carlos Guestrin 2005-2013

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

©Carlos Guestrin 2005-2013

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?

©Carlos Guestrin 2005-2013