Learning Theory

Machine Learning – CSE446
Carlos Guestrin
University of Washington

May 10, 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 – $\text{error}_{train}(h) = 0$
- What is the probability that *h* has more than ε true error?
  - $\text{error}_{true}(h) \geq \varepsilon$ ← large for given $\varepsilon > 0$

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

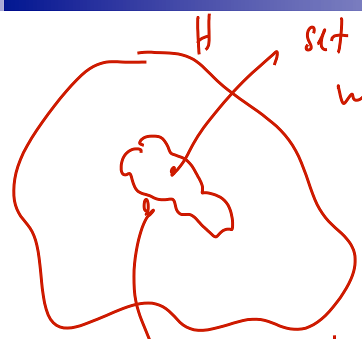
- Hypothesis *h* that is **consistent** with training data → got *N* i.i.d. points right
  - h "bad" if it gets all this data right, but has high true error $\geq \varepsilon$
- Prob. *h* with $\text{error}_{true}(h) \geq \varepsilon$ gets one data point right

  less than $1-\varepsilon$

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

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

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



set of hypotheses that are consistent with data set, $error_{train}(h) = 0$

Some h in set are great but some are bad

I don't know which h learner will pick from here.

Choose worst-case

How likely is learner to pick a bad hypothesis

hypothesis good if $error_{true}(h) \le \epsilon$

bad if $error_{true}(h) > \epsilon$

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

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

- There are *k* hypothesis consistent with data $h_1, \dots, h_k$

  some good, some bad... worst case analysis over

  - How likely is learner to pick a bad one?

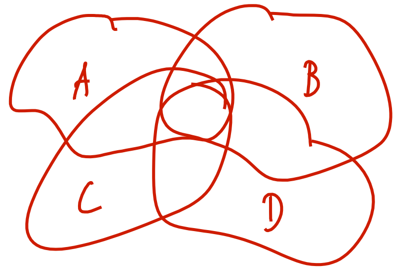
worst case

$$P(\exists h \text{ consistent with data}, \; error_{true}(h) \geq \epsilon)$$

$$= P(error_{true}(h_1) \geq \epsilon \text{ OR } error_{true}(h_2) \geq \epsilon \dots \text{ OR } error_{true}(h_k) \geq \epsilon)$$

Union bound

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



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 $(1-\varepsilon)^N$
- There are *k* hypothesis consistent with data
  - How likely is it that learner will pick a bad one out of these *k* choices?

$P(\exists h \text{ consistent with data}, \text{error}_{\text{true}}(h) \geq \varepsilon) \leq k(1-\varepsilon)^N$

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

what's k??

$k \leq |H|$

total number of hypotheses

crazy loose

Generalization error in finite hypothesis spaces [Haussler '88]

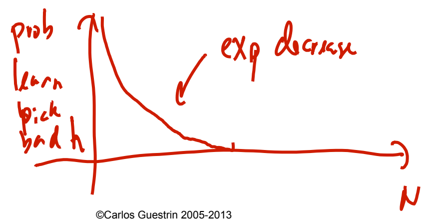
- ***Theorem***: Hypothesis space *H* finite, dataset *D* with *N* i.i.d. samples, 0 < ε < 1 : for any learned hypothesis *h* that is consistent on the training data:

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

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

for $0 \le \varepsilon \le 1$

$1-\varepsilon \le e^{-\varepsilon}$



Using a PAC bound

$-\ln\delta = \ln\delta^{-1} = \ln\frac{1}{\delta}$

$\le \delta$

- Typically, 2 use cases: $P(error_{true}(h) > \epsilon) \le |H|e^{-N\epsilon}$
  - 1: Pick ε and δ, give you *N*
  - 2: Pick N and δ, give you ε → $\varepsilon \ge \dfrac{\ln|H| + \ln\frac{1}{\delta}}{N}$

$P(error_{true}(h) \ge \varepsilon) \le |H|e^{-N\varepsilon} \le \delta$

upper bound; desired level

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

how much data you "need"

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

tolerance ε; failed prob

commit to a tolerance at least this high

N ← decreases linearly in N, very nice rate!! :)

only depend on ln|H| !! :)

|H| can be very large, but not very very large

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:

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

h

our bound

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

  $error_{train}(h) = 0$

  highly unrealist, and bad WRT overfitting

- Size of hypothesis space

  $\ln|H|$, bad is H is continuous (infinite)

  or H is very very large

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?

is $error_{train}(h) > 0$

what about $error_{true}(h)$?

in Logistic Regression, there are infinitely many h, parameterized by w

©Carlos Guestrin 2005-2013 13

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

- The error of a hypothesis is like estimating the parameter of a coin! $\theta \simeq \hat{\theta} = \frac{3}{5}$

data, estimate true θ

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

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

true param.

©Carlos Guestrin 2005-2013 14

Using Chernoff bound to estimate error of a single hypothesis

For some h:

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

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

$= \int_x \mathbb{1}(h(x) \neq y)\, p(x)\, dx$

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

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

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 h_1 better than h_2?

Danger: $error_{train}(h_1) < error_{train}(h_2)$, but $error_{true}(h_1) > error_{true}(h_2)$

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

$\leq p(error_{true}(h_1) - error_{train}(h_1) > \varepsilon) + p(error_{true}(h_2) - error_{train}(h_2) > \varepsilon)$

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

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*:

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

hold $\forall h_i$

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

$$\varepsilon \geq \sqrt{\frac{\ln|H| + \ln\frac{1}{\delta}}{2N}}$$