# 15-859(B) Machine Learning Theory

Lecture 02/05/02, Avrim Blum

MB $\Rightarrow$ PAC, greedy set cover, VC-dim

# MB $\Rightarrow$ PAC (simpler version)

**Theorem 1** *If we can learn $C$ with mistake-bound $M$, then we can learn in the PAC model using a training set of size $O\left(\frac{M}{\epsilon}\log\left(\frac{M}{\delta}\right)\right)$.*

*Proof.*

- Assume MB alg is "conservative".
- Look at sequence of hypotheses produced: $h_1, h_2, \ldots$.
- For each one, if consistent with following $\frac{1}{\epsilon}\log\frac{M}{\delta}$ examples, then stop.
- If $h_i$ has error $> \epsilon$, the chance we stopped was at most $\delta/M$. So there's at most a $\delta$ chance we are fooled by any of the hypotheses.

# Chernoff/Hoeffding recap

Consider coin of bias $p$ flipped $m$ times. Let $S$ be the observed # heads. Let $\varepsilon \in [0,1]$.

Hoeffding bounds:

- $\Pr[\frac{S}{m} > p + \varepsilon] \leq e^{-2m\varepsilon^2}$, and
- $\Pr[\frac{S}{m} < p - \varepsilon] \leq e^{-2m\varepsilon^2}$.

Chernoff bounds:

- $\Pr[\frac{S}{m} > p(1+\varepsilon)] \leq e^{-mp\varepsilon^2/3}$, and
- $\Pr[\frac{S}{m} < p(1-\varepsilon)] \leq e^{-mp\varepsilon^2/2}$.

E.g., $\Pr[S < (expectation)/2] \leq e^{-(expectation)/8}$.

E.g., $\Pr[S > 2(expectation)] \leq e^{-(expectation)/3}$.

# MB $\Rightarrow$ PAC (better bound)

**Theorem 2** *We can actually get a better bound of* $O\left(\frac{1}{\epsilon}[M + \log(1/\delta)]\right)$.

To do this, we will split data into a "training set" $S_1$ of size $\max\left(\frac{4M}{\epsilon}, \frac{16}{\epsilon}\ln\frac{1}{\delta}\right)$ and a "test set" $S_2$ of size $\frac{32}{\epsilon}\ln\frac{M}{\delta}$. We will run alg on $S_1$ and test all hyps produced on $S_2$.

Claim 1: w.h.p., at least one hyp produced on $S_1$ has error $< \epsilon/2$. *Proof:*

- If all are $\geq \epsilon/2$ then expected number of mistakes is $\geq 2M$.
- By Chernoff, $\Pr[\leq M] \leq e^{(-expect)/8} \leq 1 - \delta$.

Claim 2: W.h.p., best one on $S_2$ has error $< \epsilon$.

*Proof.* Suffices to show that good one is likely to look better than $3\epsilon/4$ and all with true error $> \epsilon$ are likely to look worse than $3\epsilon/4$. Just apply Chernoff again....

# Learning an OR function revisited

Alternative greedy-set-cover approach to learning OR function:

- Pick literal that captures the most positive examples, without capturing any negatives.
- Cross of examples covered and repeat.

If there exists an OR function of size $r$, then:

- If continue until totally consistent, this will find one of size $O(r \log m)$, where $m =$ size of training set.
- If continue until training error $\leq \epsilon/2$ then find one of size $O(r \log \frac{1}{\epsilon})$.

Get sample-size bound $O\left(\frac{1}{\epsilon}\left[\left(r \log \frac{1}{\epsilon}\right) \log(n) + \ln \frac{1}{\delta}\right]\right)$.

This is slightly worse than Winnow.

# VC-dimension and "effective" hypothesis space size

If many hypotheses in $H$ are very similar, then we shouldn't have to pay so much for them.

E.g., we saw example of $C = \{[0, a] : 0 \leq a \leq 1\}$.

How can we make this formal?

# Effective number of Hypotheses

- Define: $C[m]$ = maximum number of ways to split $m$ points using concepts in $C$. Book calls this $\pi_C(m)$.

- **Theorem:** For any class $C$, distrib. $D$, if the number of examples seen $m$ satisfies:

$$m > \frac{2}{\epsilon}[\log_2(2C[2m]) + \log_2(1/\delta)]$$

then with prob. $(1-\delta)$, all bad (error $> \epsilon$) hypotheses in $C$ are inconsistent with data.

$C[m]$ is sometimes hard to calculate exactly, but can get a good bound using "VC-dimension". VC-dimension is roughly the point at which $C$ stops looking like it contains all functions.

# Shattering

Defn: A set of points $S$ is **shattered** by a concept class $C$ if there are concepts in $C$ that split $S$ in all of the $2^{|S|}$ possible ways.

In other words, all possible ways of classifying points in $S$ are acheivable using concepts in $C$.

E.g., any 3 non-collinear points can be shattered by linear threshold functions in 2-D.

# VC-dimension

The **VC-dimension** of a concept class $C$ is the size of the largest set of points that can be shattered by $C$.

So, if the VC-dimension is $d$, that means *there exists* a set of $d$ points that can be shattered, but there is *no* set of $d+1$ points that can be shattered.

E.g., VC-dim(linear threshold fns in 2-D) = 3.

What is the VC dim of intervals on the real line?

How about $C = \{$all boolean functions on $n$ features$\}$?

# Upper and lower bound theorems

- **Theorem 1:** $C[m] \leq \sum_{i=0}^{VCdim(C)} \binom{m}{i} = O(m^{VCdim(C)})$.
  "Sauer's lemma"

- **Theorem 2:** For any class $C$, distrib. $D$, if the number of examples seen $m$ satisfies:

  $$m > \frac{2}{\epsilon}[\log_2(2C[2m]) + \log_2(1/\delta)]$$

  then with prob. $(1-\delta)$, all bad (error $> \epsilon$) hypotheses in $C$ are inconsistent with data.

- **Theorem 3:** Can replace bound in Theorem 2 with:

  $$\frac{8}{\varepsilon}[VCdim(C)\log(1/\varepsilon) + \log(1/\delta)]$$

- **Theorem 4:** For any learning alg $A$, there exists a distribution $D$, and distribution on target concepts in $C$ such that expected error of $A$ is greater than $\epsilon$ if $A$ sees less than

  $$\frac{VCdim(C) - 1}{8\varepsilon}$$

  examples.