15-859(A) Machine Learning Theory

Lecture 02/03/04, Avrim Blum

- Chernoff/Hoeffding and Uniform Convergence recap
- VC-dimension

Uniform convergence

Theorem 1 *After m examples, with probability $\geq 1-\delta$, all $h \in C$ with $err(h) > \varepsilon$ have $\widehat{err}(h) > 0$, for*

$$m \geq \frac{1}{\varepsilon}\left[\ln(|C|) + \ln\left(\frac{1}{\delta}\right)\right].$$

Proof: Chance of failure at most $|C|(1-\varepsilon)^m$. Set to δ and solve.

Theorem 2 *After m examples, with probability $\geq 1-\delta$, all $h \in C$ have $|err(h) - \widehat{err}(h)| < \varepsilon$, for*

$$m \geq \frac{1}{2\varepsilon^2}\left[\ln(|C|) + \ln\left(\frac{2}{\delta}\right)\right].$$

Proof: Apply Hoeffding.

- Chance of failure at most $2|C|e^{-2m\varepsilon^2}$.
- Set to δ and solve.

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

VC-dimension and "effective" hypothesis space size

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

E.g., how many examples do we need to PAC-learn the class $C = \{[0,a] : 0 \leq a \leq 1\}$?

[e.g., define a_ϵ so $\Pr([a_\epsilon, a]) = \epsilon$, define a'_ϵ so $\Pr([a, a'_\epsilon]) = \epsilon$.]

How can we make this general notion formal?

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.

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:

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

- **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 ϵ if A sees less than

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

  examples.

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.

  [Will prove this in a bit]
- In other words, can roughly replace "$|C|$" with "$C[2m]$".

$C[m]$ 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.

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$\}$?