Topics in Machine Learning Theory

Lecture 5: uniform convergence, tail inequalities, & VC-dimension

Avrim Blum

09/17/14

Today: back to distributional setting

- We are given sample $S = \{(x, l(x))\}$.
  - Assume x's come from some fixed probability distribution D over instance space.
  - View labels l as being produced by some target function. [Or can think of distrib over pairs $(x, l(x))$.]
- Alg does optimization over S to produce some hypothesis h. Want h to do well on new examples also from D.
- How big does S have to be to get this kind of guarantee?

Basic sample complexity bound recap

- If $|S| \geq (1/\varepsilon)[\ln(|C|) + \ln(1/\delta)]$, then with probability $\geq 1 - \delta$, all $h \in C$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

- Argument: fix bad h. Prob of fooling us on S is at most $(1-\varepsilon)^{|S|}$. Overall chance of being fooled at most $|C|(1-\varepsilon)^{|S|}$. Set to δ.
- So, if the target is in C, and we have an algo that can find consistent functions, then we only need this many examples to learn well.

Today: two issues

- If $|S| \geq (1/\varepsilon)[\ln(|C|) + \ln(1/\delta)]$, then with probability $\geq 1 - \delta$, all $h \in C$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

1. Look at more general notions of "uniform convergence".
2. Replace $\ln(|C|)$ with better measures of complexity.

Uniform Convergence

- Our basic result only bounds the chance that a bad hypothesis looks perfect on the data. What if there is no perfect $h \in C$?
- Without making any assumptions about the target function, can we say that whp all $h \in C$ satisfy $|err_D(h) - err_S(h)| \leq \varepsilon$?
  - Called "uniform convergence".
  - Motivates optimizing over S, even if we can't find a perfect function.
- To prove bounds like this, need some good tail inequalities.

Chernoff and Hoeffding bounds

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

Hoeffding bounds:

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

Chernoff bounds:

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

E.g,

- $\Pr[X > 2(\text{expectation})] \leq e^{-(\text{expectation})/3}$.
- $\Pr[X < (\text{expectation})/2] \leq e^{-(\text{expectation})/8}$.

Typical use of bounds

Thm: If $|S| \geq \frac{1}{2\epsilon^2}\left[\ln(|C|) + \ln\left(\frac{2}{\delta}\right)\right]$, then w.p. $\geq 1-\delta$, all $h \in C$ have $|err_D(h) - err_S(h)| \leq \epsilon$.

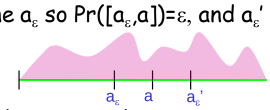
- Proof: Just apply Hoeffding.
  - $\Pr[X/m > p + \varepsilon] \leq e^{-2m\varepsilon^2}$, $\Pr[X/m < p - \varepsilon] \leq e^{-2m\varepsilon^2}$.
  - Chance of failure at most $2|C|e^{-2|S|\varepsilon^2}$.
  - Set to δ. Solve.
- So, whp, best on sample is ε-best over D.
  - Note: this is worse than previous bound ($1/\varepsilon$ has become $1/\varepsilon^2$), but conclusion is stronger.
  - Can also get bounds "between" these two.

Next topic: improving the |C|

- For convenience, let's go back to the question: how big does S have to be so that whp, $err_S(h) = 0 \Rightarrow err_D(h) \leq \varepsilon$.

VC-dimension and effective size of C

- If many hypotheses in C are very similar, we shouldn't have to pay so much
- E.g., consider the class $C = \{[0,a]: 0 \leq a \leq 1\}$.
  - Define a_ε so $\Pr([a_\varepsilon,a])=\varepsilon$, and a_ε' so $\Pr([a,a_\varepsilon'])=\varepsilon$.



  - Enough to get at least one example in each interval. Just need $(1-\varepsilon)^{|S|} \leq \delta/2$.
  - $(1/\varepsilon)\ln(2/\delta)$ examples.
- How can we generalize this notion?

Effective number of hypotheses

Define: C[m] = maximum number of ways to split m points using concepts in C. (Often called $\Pi_C(m)$.)
- What is C[m] for "initial intervals"?
- How about linear separators in R^2?

- **Thm**: For any class C, distribution D, if $|S| = m > (2/\varepsilon)[\log_2(2C[2m]) + \log_2(1/\delta)]$, then with prob. $1-\delta$, all $h \in C$ with error $> \varepsilon$ are inconsistent with data. [Will prove soon]
- I.e., can roughly replace "|C|" with "C[2m]".

Effective number of hypotheses

Define: C[m] = maximum number of ways to split m points using concepts in C. (Often called $\Pi_C(m)$.)
- What is C[m] for "initial intervals"?
- How about linear separators in R^2?

- 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 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 achievable using concepts in C.
- E.g., any 3 non-collinear points can be shattered by linear threshold functions in 2-D.
- But no set of 4 points in R^2 can be shattered by LTFs.

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.
- I.e., it's the largest m s.t. $C[m] = 2^m$.
- 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.

Upper and lower bound theorems

- Theorem 1: For any class C, distribution D, if $m=|S| > (2/\varepsilon)[\log_2(2C[2m]) + \log_2(1/\delta)]$, then with prob. $1-\delta$, all $h \in C$ with error $> \varepsilon$ are inconsistent with data.
- Theorem 2 (Sauer's lemma):
  $$C[m] \le \sum_{i=0}^{VCdim(C)} \binom{m}{i} = O(m^{VCdim(C)})$$
- Corollary 3: can replace bound in Thm 1 with
  $$O\left(\frac{1}{\epsilon}[VCdim(C)\log(1/\epsilon) + \log(1/\delta)]\right)$$
- Theorem 4: For any alg A, there exists a distrib D and target in C such that $|S| < (VCdim(C)-1)/(8\varepsilon) \Rightarrow E[err_D(A)] \ge \varepsilon$.