CSE 513T Theory of Artificial Intelligence & Machine Learning February 12, 2015

Lecture 10

Lecturer: Brendan Juba *Scribe: Weimin Zhou*

1 Today's topics

1. Finish No Free Lunch
2. ϵ–nets and Inductive Generalization
3. The "Growth Function"
4. Sauer's Lemma

2 No Free Lunch Theorem

2.1 WARM–UP PROBLEM

Suppose we have a highly biased coin that only comes up "heads" with probability ϵ. Asymptotically (as a function of k and ϵ), how many times must we toss the coin before it comes up heads k times with constant probability $p \in (0, 1)$?
If n numbers toss...
$\sum_{i=k}^{n} \binom{n}{i}$ (1-ϵ)$^{n-i}\epsilon^i \geq p$

2.2 Definitions from last time

The set of dichotomies of a set $S \subseteq X$ realized by the class $\mathcal{C}$, $\Pi_{\mathcal{C}}(S)$ is $\{C \cap S : C \in \mathcal{C}\}$ (Using $x \in C \leftrightarrow C(x) = 1$). If every subset $S' \subseteq S$ is in $\Pi_{\mathcal{C}}(S)$, we say S is "shattered" by $\mathcal{C}$.
The VC-dimension of $\mathcal{C}$ is the size of the largest set shattered by $\mathcal{C}$.

Note: Chernoff bound
$\sum_{i=1}^{n} X_i$ *(X_i=1 $\leftrightarrow$ heads)*
$E[\sum X_i]=\epsilon n$
$Pr[\sum X_i \geq (1+\gamma)\epsilon n] \leq e^{\frac{-\epsilon n \gamma^2}{3}} < e^{-\theta(k)}$. *((1+$\gamma$)$\epsilon n$ is k, ϵn is θ(k), γ is a constant)*
Consider γ=2
$Pr[\sum X_i <$*(1-γ)ϵn*]$\leq e^{\frac{-\epsilon n \gamma^2}{2}} < e^{-\theta(k)}$.
$\epsilon n = k$
$n = \theta(\frac{k}{\epsilon})$

2.3 No Free Lunch Theorem

Any PAC-learning algorithm for a class $\mathcal{C}$ of VC-dimension d that produces h. $\Pr_{x \in D}$[h(x)=c(x)]$\geq 1 - \epsilon$ for the target c w.p. $1/7 + \delta$ must use $\Omega(\frac{d}{\epsilon})$ examples from D.

Note: Summery of last time
On a set $\{X^{(1)}, ..., X^{(d)}\}$=S shattered by $\mathcal{C}$. If I choose c$\in \mathcal{C}$ s.t. a uniform random labeling of S is realized, then if my alg. only sees $\frac{d}{2}$ of the ex's, then it incorrectly labels $\geq \frac{d}{8}$ wtih probability $\geq 1/7$ (for some c$\in \mathcal{C}$)

To obtain the final bound.
We choose D to force the alg. to take many ex's just to see a new example. Simultaneously, we need to ensure that the unseen examples can cause error$\geq \epsilon$.

Let's choose D to give $\mathrm{X}^{(1)}$ prob. $1-8\epsilon$.
Let's give the other $d-1$ examples in S probability $\frac{8\epsilon}{d-1}$ each.
If the algorithm misses $\geq \frac{d-1}{2}$ of these, then as before, if we choose $c \in \mathcal{C}$ to induce a uniform, random labeling on S, then it must be wrong on $\geq \frac{1}{2}\frac{d-1}{2}$ of these unseen points in expectation. By the same calculation as before, the probability that it is wrong on $< \frac{1}{8}$ of these overall satisfies $p\frac{d-1}{8}+(1-p)(d-1) \geq \frac{d-1}{4}$.
So $p \leq \frac{6}{7}$
So therefore, again, w.p.$> 1/7$, it is wrong on $\frac{d-1}{8}$, then its error is $> \frac{d-1}{8}\frac{8\epsilon}{d-1} = \epsilon$
We now show that $\Omega(\frac{d}{\epsilon})$ examples from d are necessary to obtain $\geq \frac{d-1}{2}$ out of $\{X^{(2)}, ..., X^{(d)}\}$.
Define $R_i = \begin{cases} 1 & \text{if } \mathrm{i^{th}} \text{ example drawn from D is among } \{\mathrm{X}^{(2)}, ..., \mathrm{X}^{(\mathrm{d})}\} \\ 0 & \text{if } \mathrm{i^{th}} \text{ example is } \mathrm{X}^{(\mathrm{i})} \end{cases}$
If my alg. uses m examples, and it is obtaining $\geq \frac{d-1}{2}$ rare examples w.p. $\delta > 0 (\delta \in (0,1))$.
We then have that $m = \Omega(\frac{d}{\epsilon})$.
$\Pr[R_i = 1] = \sum_{k=2}^{d} \Pr_D[\mathrm{X}^{(K)} \in S] = (d-1)\frac{8\epsilon}{d-1} = 8\epsilon$ by the earlier calculation.

3 ϵ–nets and Inductive Generalization

"ϵ–*net*": Intuitively, what we need to know for generalization is that if some $h \in H$ is wrong about the target c on any set r that has probability $\geq \epsilon$ under the unknown D, then we have some example (x, c(x)) in our training set that catches this error (Then, as before, solving $Consis_H$ solves the learing problem).
Definition For any target concept c, the class of error regions of $\mathcal{H}$(w.r.t. c), denote $\Delta_{\mathcal{H}}(c)$, is $\Delta_{\mathcal{H}}(c) = \{[h \neq c](\mathrm{x}) : h \in \mathcal{H}\}$
The set of error regions of probability $\geq \epsilon$ under D is $\Delta_{\mathcal{H},\epsilon}(c) = \{r \in \Delta_{\mathcal{H}}(c) : \Pr_{x \in D}[x \in r] \geq \epsilon\}$
Definition For any $\epsilon > 0$, we say that a set $S \subseteq X$ is an ϵ–net for $\Delta_{\mathcal{H}}(c)$ w.r.t. D if for every $r \in \Delta_{\mathcal{H},\epsilon}(c)$ we have some $x \in r \cap s$.
Example: Suppose $C(X) \equiv 0, \mathcal{H}$ is the class of intervals of $\Re$, D is the uniform distribution on $[0,1]$.
Then $\Delta_{\mathcal{H}}(c) = \{[h \neq 0](\mathrm{x}) : h \in \mathcal{H}\} = \mathcal{H}$. ($[h \neq 0](\mathrm{x}) \equiv h(x)$).
Since for each $[a,b] \in \Delta_{\mathcal{H}}(c)$, $\Pr_{x \in D}[X \in [a,b]] = min(1,b) - max(0,a)$.
$\Delta_{\mathcal{H},\epsilon}(c) = \{[a,b] : min(1,b) - max(0,a) \geq \epsilon\}$.
The set $S = \{k\epsilon : k = 1, 2, ..., [\frac{1}{\epsilon}]\}$ is an ϵ–net for $\Delta_{\mathcal{H}}(c)$.
Since every $r \in \Delta_{\mathcal{H},\epsilon}(c)$ has $length \geq \epsilon$ in $[0,1]$, it must hit this grid of ϵ–spaced points.
Revisiting the standard calculation:
If we fix any $r \in \Delta_{\mathcal{H},\epsilon}(c)$, the probability that we fail to hit r after m examples is $\Pr_{x \in D}[x \notin r]^m = (1 - Pr_{x \in D}[x \in r])^m \leq (1-\epsilon)^m$ since $r \in \Delta_{\mathcal{H},\epsilon}(c)$.
So the probability that m examples don't form an ϵ–net for $\Delta_{\mathcal{H}}(c)$ is at most $|\Delta_{\mathcal{H}}(c)|(1-\epsilon)^m$ by a union bound.
$\Delta_{\mathcal{H}}(c)$: Trouble is usually infinitly for numerical data.
Want: analysis in terms of how $\mathcal{H}$ behaves on finite sample.

4 The Growth Function

Definition: The growth function of a representation class $\mathcal{H}$, $\Pi_{\mathcal{H}} : N \to N$ is given by $\Pi_{\mathcal{H}}(m) = max_{s \subseteq x : |s| = m} |\Pi_{\mathcal{H}}(s)|$

So if $\mathcal{H}$ shatters a set S of size m, $\Pi_{\mathcal{H}}(m) = 2^m$.
(note: there are at most 2^m possible labelings of any set of m points, so, $max_{s \subseteq x, |s| = m} |\Pi_{\mathcal{H}}(S)| \leq 2^m$).
Notice: if $\mathcal{H}$ shatters S, $\mathcal{H}$ also shatters all subsets $\mathrm{S}' \subseteq S$

So...
proposition: If the VC-dimension of $\mathcal{H}$ is d, for every $m \leq d$, $\Pi_{\mathcal{H}}(m) = 2^m$.

5 Sauer's Lemma

If $\mathcal{H}$ has VC-dimension d, then for every m, $\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} = \mathcal{O}(m^d)$ for $m > d$.
Lemma 1 For any $m, d \in N$, define $\Phi_d(m)$ inductively, by $\Phi_0(m) = \Phi_d(0) = 1$ and otherwise $\Phi_d(m) = \Phi_d(m-1) + \Phi_{d-1}(m-1)$. Then, if VCD($\mathcal{H}$)=$d$, for every m $\Pi_{\mathcal{H}}(m) \leq \Phi_d(m)$.
Lemma 2 $\Phi_d(m) = \sum_{i=0}^{d} \binom{m}{i} = \mathcal{O}(m^d)$ for $m > d$