CS 395T Computational Complexity of Machine Learning

Lecture 5: September 24, 2005 — Lecturer: Adam Klivans

Scribe: Warren Hunt

5.1 Relationship of $\mathsf{VC-D}$ to Learning and Sample Complexity

First we show that the $\mathsf{VC-D}$ provides a lower bound on the sample complexity requred to PAC learn a concept class C.

Theorem 1 *Assume concept class C is PAC learnable, then any learner L must draw $\Omega(\frac{d}{\epsilon})$ examples in order to achieve accuracy ϵ.*

Before proving the general case, we will provide the proof for the specific case where accuracy $\epsilon = \frac{1}{8}$.

Proof: Let S be a set C shatters of size d. Let D be the uniform distribution over S. Choose a concept $c \in C$ at random. This is equivelant to independantly randomly choosing lables for d points because D is uniform. Assume for contradiction that L takes $m < O(d)$ examples and outputs $h \in \mathcal{C}$ with accuracy better than $\frac{1}{8}$.

$error := Pr[h(s) \neq c(x)]$
$E[error] = \frac{d-m}{2d}$
By Markov's inequality, with $probably \geq \frac{1}{2}$, $E[error] \geq \frac{d-m}{4d}$
$m = \frac{d}{2} \Rightarrow E[error] \geq \frac{1}{8} = \epsilon$
Therefore, we must take at least $\Omega(d)$ examples to achieve accuracy $\frac{1}{8}$

Now for the proof of the general case.

Proof: Let S be a set C shatters of size d Define the weight $x_1 \in S = 1 - 8\epsilon$ and weight of $x_{\{i|i\neq 1\}} \in S = \frac{8\epsilon}{d-1}$ Take m examples with ($\frac{d}{2}$ distinct).
$m < d \Rightarrow E[Pr[h(s) \neq c(x)]] = \frac{\delta}{8} \geq \epsilon$ by Markov's inequality again.

However we must take $\Omega(\frac{\delta}{\epsilon})$ random draws to see some $x_j \neq x_1$ Therefore we also need $\Omega(\frac{\delta}{\epsilon})$ random draws to see $\frac{d}{2}$ distinct examples, which were required for accuracy better than ϵ. Therefore we must draw at least $\Omega(\frac{d}{\epsilon})$ examples in order to achieve accuracy ϵ.

5.2 The VC-Theorem

Theorem 2 VC-Theorem: *It suffices to choose $O(\frac{\mathsf{VC-D}}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon})$ examples and output a consistent h to have, with high probability, error $\leq \epsilon$. Defining the error of a hypothesis h with respect to a concept c by the weight of $c \triangle h$ with respect to distribution D.*

Define the concept class $\Delta c = \{c' \triangle c | c' \in C\}$.

Fact 1 $\mathsf{VC-D}(\Delta c) = \mathsf{VC-D}(C) \Rightarrow |\Pi_{\Delta c}| = |\Pi_C|$

For a distribution D, consider $\Delta_\epsilon c$, all regions Δc such that the weight of $\Delta c \geq \epsilon$ with respect to D.

Consider the following "bad event", call it A, for the learner: Choose S_1 a set of m examples. There exists some "bad region" $r \in \Delta_\epsilon(c)$ such that $S_1 \cap r \neq \emptyset$. We would like to bound $Pr(A)$ to be $= (1-\epsilon)^m$ per (fixed) error region.

In order to do this, we will consider another even B and bound $Pr[A]$ by $Pr[B]$. Choose S_1, S_2 both sets of m examples. B occurs if A occurs and $|r \cap S_2| \geq \frac{\epsilon m}{2}$ (this equation by Markov's inequality)

$Pr[B] = Pr[B|A]Pr[A]$; we may assume $Pr[B|A] \geq \frac{1}{2} \Rightarrow Pr[A] \leq 2Pr[B]$

We will now bound $Pr[B]$: $B \Leftrightarrow \exists r \in \Pi_{\Delta_\epsilon c}(S_1 \cup S_2)$ s.t. $r \cap S_1 = \emptyset |r| = \frac{\epsilon m}{2}$ Replace S_1 and S_2, with S, a set of $2m$ random examples and redefine S_1 and S_2 by randomly partitioning S into two groups of size m. For any S and any fixed region r, The probability $|S_1 \cap r| \geq \frac{\epsilon m}{2}$ is $\leq 2^{\frac{\epsilon m}{2}}$ for each fixed error region r. Since we have $|\Pi_{\Delta_\epsilon c}(S_1 \cup S_2)| \leq O((2m)^d)$ error regions.

$Pr[B] \leq (2m)^d 2^{-\frac{\epsilon m}{2}}$
$Pr[A] \leq 2(2m)^d 2^{-\frac{\epsilon m}{2}}$
$m = O(\frac{d}{\epsilon} \log \frac{1}{\epsilon} + \frac{\log \frac{1}{\delta}}{\epsilon})$