CS 395T Computational Learning Theory

Lecture 4: September 19, 2005 Lecturer: Adam Klivans

Scribe: Justin Brickell

4.1 Using Set-Cover Approximation to learn Conjunctions

This material is covered in depth on pages 38-42 of Kearns and Vazirani, so we will only give a brief outline here. Recall from the previous lecture that there is a greedy algorithm to find a set over of size at most $O(opt(S)\log m)$ where $opt(S)$ denotes the number of sets in a minimum cardinality cover and m is the size of the universe from which the set elements are drawn.

We relate the Set Cover Problem to the Conjunction Problem by noting that each literal z appearing in the hypothesis h "covers" a subset $N_z \in S$ of the negative examples in S. Specifically, a z covers those examples that are made negative by its inclusion. The Conjunction Problem is equivalent to finding the minimum number of sets N_z that cover the entire set of negative examples.

Applying the cardinality version of Occam's Razor (Theorem 2.2) to the Occam algorithm based on the greedy set cover algorithm, we arrive at the conclusion that the Conjunction Problem is PAC learnable provided that the sample size is

$$m \geq c_1 \left(\frac{1}{\epsilon} \log \frac{1}{\delta} + \frac{size(c) \log n (\log size(c) + \log \log n)}{\epsilon} \right).$$

By modifying the algorithm to stop the greedy algorithm before its completion, we are able to improve the sample size bound to

$$m \geq c_1 \left(\frac{1}{\epsilon} \log \frac{1}{\delta} + \frac{size(c) \log(2/\epsilon) \log n}{\epsilon} \right).$$

4.2 The Vapnik-Chervonenkis Dimension

The Vapnik-Chervonenkis (VC) Dimension is a measure that assigns to each concept class C a number that captures the complexity of C. It is given a thorough treatment in chapter 3 of the text; rather than repeat the proofs given in the text we will just highlight the important definitions and theorems presented in lecture.

Definition 1 *For any concept class C over X, and any $S \subseteq X$,*

$$\Pi_C(S) = \{c \cap S : c \in C\}.$$

Definition 2 *If $\Pi_C(S) = \{0,1\}^m$ (where $m = |S|$), then we say that S is* **shattered** *by C. Thus, S is shattered by C if C realizes all possible dichotomies of S.*

Definition 3 *The* **Vapnick-Chervonekis (VC) dimension** *of C, denoted as $VCD(C)$, is the cardinality d of the largest set S shattered by C. If arbitrarily large finite sets can be shattered by C, then $VCD(C) = \infty$*

4.2.1 Examples of VC Dimension

We derived the VC dimensions of several well-known concept classes.

- *Intervals on the real line.* $VCD = 2$
- *Halfspaces on the plane.* $VCD = 3$
- *Halfspaces in d dimensions.* $VCD = d + 1$
- *Axis aligned rectangles in the plane.* $VCD = 4$
- *Convex polygons in the plane with d sides.* $VCD = 2d + 1$
- *Monotone disjunctions on $k << n$ variables.* $VCD = k \log n$

4.2.2 A Polynomial Bound on $|\Pi_C(S)|$

Definition 4 *for any natural number m we define*

$$\Pi_C(m) = \max\{|\Pi_C(S)| : |S| = m\}.$$

Note that if the VC dimension of C is infinite, then $\Pi_C(m) = 2^m$. However, in the following lemma we show that if the VC dimension is finite, then $\Pi_C(m)$ grows only polynomially with m.

Lemma 1 *If $VCD(C) = d$, then for any m, $\Pi_C(m) \leq \Phi_d(m) = \sum_{i=0}^{d} \binom{m}{i} = O(m^d)$ The proof of this lemma is by induction on both d and m. It is found on page 55 and is omitted here to avoid redundancy.*