6.858/18.428 Machine Learning	**Lecture 7: 10/3/94**
Lecturer: Ron Rivest	*Scribe: Lewis Girod*

7.1 Outline

- The VC-dimension and PAC Learning infinite classes
- Definitions related to VC-dimension
- VC-dimension examples
- The Main property of $\Pi_c(m)$

7.2 The Vapnik-Chervonenkis Dimension

So far in our study of algorithms we have proven a number of finite concept classes to be PAC learnable. Recall from last lecture the Occams razor result: for any *finite* concept class $\mathcal{C}$ and any *finite* hypothesis class $\mathcal{H}$, if we draw $\frac{1}{\epsilon}(\ln(|\mathcal{H}|) + \ln(\frac{1}{\delta}))$ examples, and find a consistent hypothesis, then we PAC-learn. This a very strong result about finite classes – but none of the PAC learning results we have shown thus far apply to infinite concept classes.

Infinite-cardinality classes make up a large domain of important learning problems. As a simple example, the concept class formed by the half spaces of the plane is clearly infinite. Note that we have shown this concept class to be learnable with a finite mistake bound by the perceptron convergence algorithm, but we have not yet shown this class to be PAC learnable.

In this lecture we will begin what will be a major result of this course. The result will prove an analogous Occam's razor result for infinite concept classes. Essentially, $VC_{DIM}(\mathcal{H})$ replaces $\ln(|\mathcal{H}|)$ in the number of samples we must draw.

7.3 Definitions

First of all, our notation is changing slightly; $\mathcal{C}$ represents a concept class over an instance space X, both of which may now be infinite. We use S to represent a finite sample chosen from X. We will now define a function which characterizes the response of the concepts in $\mathcal{C}$ to a given sample S.

Definition 1 *For any concept class $\mathcal{C}$ over instance space X, for all finite samples $S \subseteq X$,*

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

It is important at this point to give some kind of explanation of what this means, especially since the set notation can be a little confusing. Essentially, $\Pi_{\mathcal{C}}(S)$ collects the set of all subsets of the sample set S which are made positive by some concept c in the concept class $\mathcal{C}$. Thus $c \cap S$ represents the elements of S that are labelled postive by a concept c, and $\Pi_{\mathcal{C}}(S)$ is all such subsets for all concepts.

Definition 2 *If $|\Pi_{\mathcal{C}}(S)| = 2^{|S|}$ then S is* **shattered** *by C. In other words, fix a sample and a concept class; if the following is true for every possible subset of the sample, then the concept class shatters the sample: the subset is made positive by some concept in the concept class, and that concept does not make any other element of S (not in the subset) positive. That is, S is shattered by C if C realizes all possible dichotimies of S.*

Finally, we define VC-dimension in these new terms. When reading this definition take careful note of the direction of the quantifiers.

Definition 3 *The* **Vapnik-Chervonenkis dimension** *of $\mathcal{C}$, denoted $VC_{DIM}(\mathcal{C})$, is the largest cardinality d such that there exists a sample set of that cardinality $|S| = d$ that is shattered by $\mathcal{C}$. If no largest cardinality exists then $VC_{DIM}(\mathcal{C}) = \infty$.*

7.4 Examples: Finding VC_{DIM}

1. Consider the concept class of intervals $[0, a]$ on the real number line, $0 < a < 1$. Clearly samples of size 1 can be shattered by this class, since if we pick some point x as our sample, the point a can be placed to the left or the right, thus

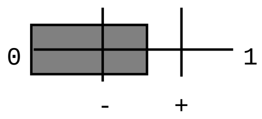

Figure 7.1: A left-bounded interval on the real axis: No set of two samples is shattered.

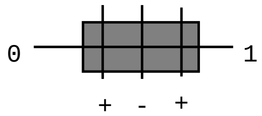

Figure 7.2: One interval on the real axis: No set of three points is shattered.

excluding or including x. However, as 7.1 shows, no sample of size 2 can be shattered since it is impossible to choose a such that x_2 is included and x_1 is excluded if $x_2 > x_1$. Hence the $VC_{DIM} = 1$.

2. Consider the class of subintervals $[a, b]$, $0 < a, b < 1$ (see 7.2). Here a sample of size 2 is shattered, but no sample of size 3 is shattered, since no concept can satisfy a sample whose middle point is negative and outer points are positive. Hence, $VC_{DIM} = 2$.

3. What about the class of k non-intersecting subintervals (see 7.3)? A sample of size $2k$ shatters (just treat each pair of points as a separate case of example 2) but no sample of size $2k + 1$ shatters, since if the sample points are alternated positive/negative, starting with a positive point, the positive points can't be covered by only k intervals. Hence $VC_{DIM} = 2k$.

4. Consider the class of half-spaces of the plane. Three samples can be shattered, but four cannot; hence $VC_{DIM} = 3$. To see why four points can never be shattered, consider two cases. The trivial case is when one point can be placed

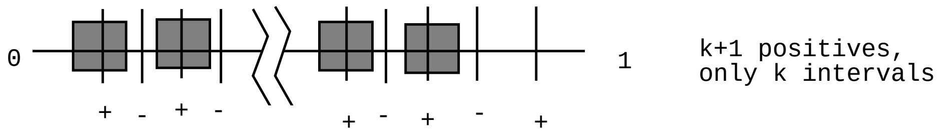

Figure 7.3: k subintervals on the real axis.

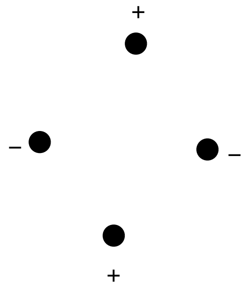

Figure 7.4: Half-spaces of the plane.

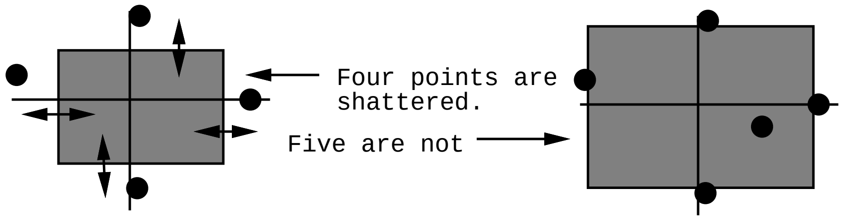

Figure 7.5: Axis-aligned rectangles in the plane.

within a triangle formed by the other three; then if the middle point is positive and the others are negative, no half space can contain only the positive points. If however the points cannot be arranged in that pattern, then label two points diagonally across from each other as positive, and the other two as negative (see 7.4). Similar results hold for higher dimensions: for half-spaces of $\mathbf{R}^\mathbf{n}$, $VC_{DIM} = n + 1$.

5. The class of circles in the plane has $VC_{DIM} = 3$, using the same argument as above.

6. The class of axis-aligned rectangles in the plane has $VC_{DIM} = 4$. The trick here is to note that for any collection of five points, at least one of them must be interior to any rectangle bounded by the other four; hence if the bounding points are positive, the interior point cannot be made negative (see 7.5).

7. The class of convex d-gons has $VC_{DIM} = 2d + 1$. To see that a set of $2d + 1$ points can be shattered, place the vertices on a circle; it is then possible to selectively include any of the vertices (see 7.6).

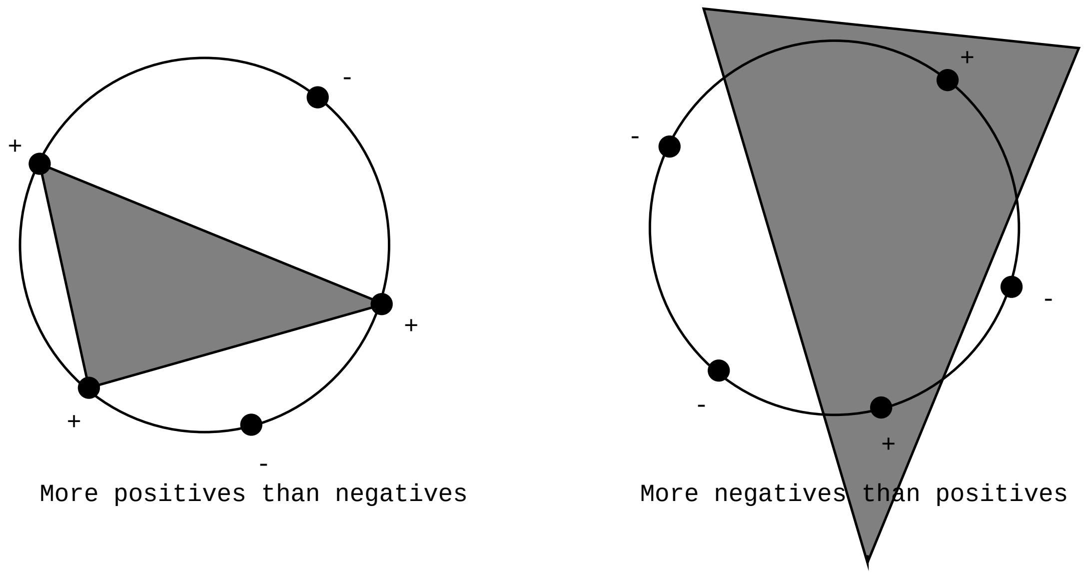

Figure 7.6: Convex d-gons.

8. The class of arbitrary convex polygons has infinite VC_{DIM}.

7.5 The Main Property of $\Pi_{\mathcal{C}}(m)$

We begin our investigation of infinite concept classes by deriving a polynomial bound on the magnitude of $\Pi_{\mathcal{C}}(S)$. In order to show this we need to introduce a new function which in the interests of greater confusion is called $\Pi_{\mathcal{C}}(m)$:

Definition 4 *For any natural number m,*

$$\Pi_{\mathcal{C}}(m) = \max\{|\Pi_{\mathcal{C}}(S)| : |S| = m\}.$$

In other words, $\Pi_{\mathcal{C}}(m)$ takes as its argument a sample size m and returns the maximum number of ways a sample of that size can be labeled in the concept class $\mathcal{C}$. This function behaves as a measure of concept class complexity. Suppose $d = VC_{DIM}(\mathcal{C})$; then $m \leq d$ implies $\Pi_{\mathcal{C}}(m) = 2^m$, and $m > d$ implies $\Pi_{\mathcal{C}}(m) < 2^m$. We will further demonstrate that $\Pi_{\mathcal{C}}(m)$ grows polynomially, within a constant factor of m^d, whenever $m > d$. In order to show this bound we need to define a new function.

Definition 5 *For any natural numbers m and d,*

$$\Phi_d(m) = \begin{cases} 1 \text{ if } d = 0 \text{ or } m = 0 \\ \Phi_d(m-1) + \Phi_{d-1}(m-1) \end{cases}.$$

Our plan is to first show that $\Phi_d(m)$ grows polynomially, and then show that $\Pi_{\mathcal{C}}(m) \leq \Phi_d(m)$, proving the main property of $\Pi_{\mathcal{C}}(m)$: that it is polynomially bounded for $m > d$.

Lemma 1 $\Phi_d(m) = \sum_{i=0}^{d} \binom{m}{i}$

Base cases:

- if $d = 0$, $\binom{m}{0} = 1$
- if $m = 0$, $\sum_{i=0}^{d} \binom{0}{d} = \binom{0}{0} = 1$.

Inductive step:

$$\begin{aligned} \Phi_d(m) &= \Phi_d(m-1) + \Phi_{d-1}(m-1) \\ &= \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \\ &= \sum_{i=0}^{d} \left[\binom{m-1}{i} + \binom{m-1}{i-1} \right] \\ &= \sum_{i=0}^{d} \binom{m}{i} \end{aligned}$$

Lemma 2 $\Phi_d(m)$ *grows polynomially when $m > d$.*

Since we are concerned with times when $m > d$, we have $0 \leq \frac{d}{m} < 1$. We can now write:

$$\left(\frac{d}{m}\right)^d \sum_{i=0}^{d} \binom{m}{i} \leq \sum_{i=0}^{d} \left(\frac{d}{m}\right)^i \binom{m}{i} \leq \sum_{i=0}^{m} \left(\frac{d}{m}\right)^i \binom{m}{i} = \left(1 + \frac{d}{m}\right)^m \leq e^d.$$

Dividing both sides by $\left(\frac{d}{m}\right)^d$, we get

$$\Phi_d(m) \leq e^d \left(\frac{m}{d}\right)^d \leq \left(\frac{me}{d}\right)^d.$$

We now wish to show the main property of $\Pi_{\mathcal{C}}(m)$, that it is less than $\Phi_d(m)$ and is therefore polynomially bounded.

Lemma 3 *If $d = VC_{DIM}(\mathcal{C})$, then for all m, $\Pi_{\mathcal{C}}(m) \leq \Phi_d(m)$*

The proof proceeds by double induction on m and d. We must show the base cases whenever $m = 0$ or $d = 0$. When $m = 0$, there can only be one subset, hence $\Pi_{\mathcal{C}}(0) \leq 1 = \Phi_d(0)$. When $d = VC_{DIM}(\mathcal{C}) = 0$, no set of points can be shattered, hence all points can be labelled only one way. From this we conclude that $\Pi_{\mathcal{C}}(m) = 1 \leq \Phi_0(m)$. So the lemma holds for the base case.

We choose as the induction hypothesis that for all m', d' such that $m' \leq m$ and $d' \leq d$ and at least one of these inequalities is strict, we assume $\Pi_{\mathcal{C}}(m') \leq \Phi_{d'}(m')$.

Now suppose we have a set S of cardinality m. In order to use the induction hypothesis, we consider the number of ways of labelling S after we remove some element x from it:

$$|\Pi_{\mathcal{C}}(S - \{x\})| \leq \Pi_{\mathcal{C}}(m-1) \leq \Phi_d(m-1).$$

Our goal is to count the possible labellings of S; certainly there are at least as many as there are ways of labelling $S - \{x\}$, but this may not account for all of them. So we need to count the number of labellings in S that correspond to a single labelling in $S - \{x\}$ (i.e. this is the amount by which we're undercounting); if we add in this correction term we will get our desired count. In order to do this, we consider the set of labellings of S which may be extended to include x:

$$\mathcal{C}' = \{c \in \Pi_{\mathcal{C}}(S) : x \notin c, c \cup \{x\} \in \Pi_{\mathcal{C}}(S)\}.$$

By definition, $|\mathcal{C}'|$ is the number of labellings in S that map to a single labelling in $S - \{x\}$. We also note that $\mathcal{C}' = \Pi_{\mathcal{C}'}(S - \{x\})$ since $\mathcal{C}'$ consists only of possible labellings of $S - \{x\}$. Furthermore, we know that the $VC_{DIM}(\mathcal{C}') \leq d - 1$ because if $VC_{DIM}(\mathcal{C}') = d$ then $VC_{DIM}(\mathcal{C}) = d + 1$ since by definition of $\mathcal{C}'$ there exists some element x that can be arbitrarily included or excluded in a labelling of S, thus shattering S if $S - \{x\}$ can be shattered. From this we conclude:

$$\mathcal{C}' = \Pi_{\mathcal{C}'}(S - \{x\}) \leq \Phi_{d-1}(m-1).$$

Putting this together with our previous result we have:

$$|\Pi_{\mathcal{C}}(S)| \leq \Phi_d(m-1) + \Phi_{d-1}(m-1) = \Phi_d(m).$$

This concludes the proof of the Lemma, demonstrating the main property of $\Pi_{\mathcal{C}}(m)$.

7.6 Next Lecture

How do we PAC learn over an infinite concept class? We need to show that with high probability no bad concepts will survive. It turns out that a finite set of samples *can* kill off all the bad concepts!