The VC Dimension

These notes are slightly edited from scribe notes in previous years.

Let $X$ be an instance space, $S \subseteq X$, $C$ is a concept class. Define:

$$\Pi_C(S) = \{c \cap S | c \in C\} \quad (1)$$

i.e. the set of subsets of $S$ that can be obtained using $c \in C$

$$\Pi_C(m) = \max_{S:|S|=m} |\Pi_C(S)| \quad (2)$$

We say that $C$ shatters $S$ if $|\Pi_C(S)| = 2^{|S|}$, i.e., we get all possible subsets of $S$. The Vapnik-Chervonenkis Dimension of $C$ is the side of the largest shattered set:

$$VCD(C) = \max_{S:S \text{ shatters } S} |S|.$$ 

If $C$ can shatter arbitrarily large sets, we say that $VCD(C) = \infty$.

To prove $VCD(C) = d$, we need to show:

1. There is at least one set of size $d$ that is shattered.
2. No set of $d+1$ points is shattered.

Thus showing $VCD \geq k$ requires just one example of a shattered set, whereas showing $VCD < k+1$ require a proof that no such set can be shattered (which might be harder to do).

As will be shown in the next two lectures the significance of these definitions is that if $VCD(C) = d$ then $\Omega(\frac{d}{\varepsilon})$ examples are required for PAC learning $C$ and $O(\frac{1}{\varepsilon} \ln \frac{1}{\delta} + \frac{d}{\varepsilon} \ln \frac{1}{\varepsilon})$ suffice for PAC learning $C$ by an algorithm finding a consistent hypothesis. Therefore the VCD provides a good measure for the sample complexity of PAC learning.

It follows immediately from the definitions that $VCD(C) \leq \log_2(|C|)$ (because to shatter a set of size $k$ we need at least $2^k$ hypotheses). Thus we see with respect to the convergence comment above that VCD replaces $\ln |H|$ in the Occam results, and might yield a tighter bound in some cases.

Example 1: $VCD$(Intervals) = 2: we can see that a labeling ”$+ - +$” of 3 points can not be obtained by any single interval. On the other hand, any two points can be shattered since all labelings: $- -, - +, + -, ++$ can be easily obtained.
Example 2: \(VCD(\text{Union of two Intervals})\). Here concepts have the form \([a, b] \cup [c, d]\):

1. Pick any 4 points, we can simply shatter the “left” 2 points with one interval, and the “right” 2 points with one interval to get any labeling of the 4 points.
2. For 5 points, we can see that “+ − − +” can not be obtained by any two intervals.

Example 3: \(VCD(\text{Union of } k \text{ Intervals})\). Here concept have the form \([a_i, b_i] \cup \cdots [a_k, b_k]\). We can see that similar to the argument for \(k = 2\), the pattern ”+ − − − +” cannot be realized and it is easy to shatter \(2k\) points in parts.

Example 4: finite union of intervals. Here the VCD is infinite because we can shatter sets of any size by using a sufficiently large \(k\) in example 3. Given the above comment it follows that there is no uniform sample bound \(m\) that holds for PAC learning this class. Redefining to allow the learner’s complexity to depend on target size (as we have already done) the class is learnable with the obvious algorithm.

Example 5: \(VCD(\text{axis−parallel rectangle})\). In class we showed that any labeling of some 4 point configuration (for example the set \((0,2), (1,0), (2,3), (3,1)\)) can be obtained. For any five points, we can pick 4 of the points that together include the minimum and maximum x and y coordinates (pick arbitrarily if there is more than one choice). Assign the 4 points with these extremal values the label + and the fifth one the label −. Then there is no concept that can obtain the labeling.

Example 6: \(VCD(\text{Conjunctions}) = n\) where \(n\) is the number of features in the domain, and we assume that the empty conjunction is true on every example). Here we only prove \(VCD(\text{Conjunctions}) \geq n\). The other direction appears in the homework.

Let \(O_i = \text{string of all 1’s except in } i\text{'th position}\). We define \(S = \{O_i\}_{i=1}^n\). For example if \(n = 5\) we have \(01111, 10111, 11011, 11101, 11110\). We next show that \(S\) is shattered.

Pick any arbitrary labeling for \(S\). Let \(R = \{i|O_i \text{ is a negative example}\}\) and \(c = \wedge_{i \in R} x_i\).

Recall that the only zero bit in \(O_i\) is \(x_i\). It is easy to see that if \(i \in R, O_i \text{ is a negative example for } c\) since \(x_i\) is in the conjunction; otherwise, \(O_i \text{ is positive since } x_i\) is not in conjunction. Since the choice of labeling was arbitrary we have shown that all labelings can be obtained and \(S\) is shattered.