VC Dimension: Examples and Tools

Carlos C. Rodríguez
http://omega.albany.edu:8008/

October 5, 2004

Recall the Definitions

Let $\mathcal{A}$ be a class of subsets of R^d. The shatter coefficients of $\mathcal{A}$ are denoted by $\mathcal{S}\ (n, \mathcal{A}\ )$ and defined as the maximum number of subsets of a set of n elements that appear in the elemts of $\mathcal{A}$. We say that the class $\mathcal{A}$ shatters a given set of n elements if each of its 2^n subsets are picked by the members of $\mathcal{A}$. The Vapnik-Chervonenkis (VC) dimension of $\mathcal{A}$ is the size of the largest set that can be shattered by $\mathcal{A}$. It is denoted by $V = V(\mathcal{A}\ )$. Thus, if $V < \infty$ we must have $\mathcal{S}\ (n, \mathcal{A}\ )\ = 2^n$ for each $n \leq V$ and $\mathcal{S}\ (V+1, \mathcal{A}\ ) < 2^{V+1}$. In words: If the VC dimension is V the class $\mathcal{A}$ shatters some sets with V elements but it shatters no set with $V+1$ (or more) elements. Here are the simplests examples.

Simplest Examples

Half lines in R $\mathcal{A}\ = \{(-\infty, x] : x \in R\}$. Clearly, $\mathcal{S}\ (n, \mathcal{A}\ )\ = n+1$ and so $V = 1$. Notice, that,

$$\mathcal{S}\ (n, \mathcal{A}\ )\ = n + 1 = \binom{n}{0} + \binom{n}{V}$$

Intervals in R $\mathcal{A}\ = \{(a, b) : a, b \in \mathrm{R}\}$. To compute $\mathcal{S}\ (n, \mathcal{A}\ )$ just think of the $(n+1)$ spaces defined by n different real numbers. Any choice of two of these spaces corresponds to an (a, b) picking up the points inside. Thus,

$$\mathcal{S}\ (n, \mathcal{A}\ )\ = \binom{n+1}{2} + 1$$

where the extra interval is the one picking none of the n points. Therefore, $V = 2$ and again,

$$\mathcal{S}\ (n, \mathcal{A}\ )\ = \frac{n(n+1)}{2} + 1 = \binom{n}{0} + \binom{n}{1} + \binom{n}{V}$$

More Examples

It is not difficult to generalize the previous examples to R^d.

South-West Intervals in R^d $\mathcal{A} = \{(-\infty, a_1] \times (-\infty, a_2] \ldots \times (-\infty, a_d] : a_j \in R\}$. By what we found for the case $d = 1$ and property 5 above, we can deduce that, $\mathcal{S}(n, \mathcal{A}) \leq (n+1)^d$ and looking ahead we guess $V = d$. This can be shown by noticing that $\mathcal{A}$ shatters the d cannonical vectors $E = \{e_1, \ldots, e_d\}$ with $e_j = (0, \ldots, 0, 1, 0, \ldots, 0)$ (i.e. all zeroes except a 1 in the jth entry). Any $B \subset E$ is picked by a member of $\mathcal{A}$, e.g., by the one with $a_j = 1 + \epsilon_j/2$ where $\epsilon_j = +1$ if $e_j \in B$ and $\epsilon_j = -1$ otherwise. Furthermore, no set with $d+1$ points can be shatter by $\mathcal{A}$ since it is imposible to pick only the d points where the first has the largest first coordinate, the second has the largest second coordinate, ..., the dth has the largest dth coordinate, and not the other. Thus, $V = d$. •

All the rectangles in R^d For this class $V = 2d$. Again the case $d = 1$ and property 5 above shows that,

$$\mathcal{S}(n, \mathcal{A}) \leq \left(\frac{n(n+1)}{2} + 1 \right)^d < (n+1)^{2d}$$

and looking ahead we'd guess $V = 2d$. A proof along the lines of the previous example is straight forward.

Power Tools

The following simple result is known as Sauer's Lemma. It provides a way to bound shatter coefficients in terms of the VC dimension.

Sauer's Lemma If $V(\mathcal{A}) < \infty$, for all n

$$\mathcal{S}(n, \mathcal{A}) \leq \sum_{i=0}^{V} \binom{n}{i}$$

Proof: By induction on n. It is true for $n \leq V$ since for these n,

$$\mathcal{S}(n, \mathcal{A}) = 2^n = \sum_{i=0}^{n} \binom{n}{i} = \sum_{i=0}^{V} \binom{n}{i}.$$

Now let us assume the result true for n and deduce it for $n+1$. Clearly $\mathcal{S}(n+1, \mathcal{A})$ is at most $2\mathcal{S}(n, \mathcal{A})$ (see property 4 above) but this bound can be reduced by using the fact that the vc dimension is V so we know that $\mathcal{A}$ shatters no set with $V+1$ points. On the other hand, a set with $n+1$ points has n choose V more subsets with $V+1$ elements than a set with n points has, and each of these must hide at least one new subset that cannot be picked out by $\mathcal{A}$.