

## VC Clarification, Fall 2006

Several students have had questions about problem 1 on the homework. Here is an expanded explanation that is intended to help.

1. Consider the hypothesis class of homogeneous half-spaces in the  $\mathfrak{R}^d$  (i.e. each instance  $\mathbf{x} \in \mathfrak{R}^d$  and  $\mathcal{H}$  consists of all  $h_{\mathbf{w}}$  having the form  $h(\mathbf{x}) = +1$  if and only if  $\mathbf{w} \cdot \mathbf{x} > 0$ , so  $\mathbf{0}$  is always mapped  $-1$ ). Determine the VC-dimension of homogeneous half-spaces in  $\mathfrak{R}^d$ . In other words, for each  $d \geq 1$  find a set  $S \subset \mathfrak{R}^d$  is shattered by  $\mathcal{H}$  and show that no set  $S' \subset \mathfrak{R}^d$  with  $|S'| > |S|$  is shattered by  $\mathcal{H}$ . (Since if  $S'$  is shattered, then every subset of  $S'$  is also shattered, it suffices to show that no set  $S'$  with  $|S'| = |S| + 1$  is shattered). You may use the fact that in any set  $S$  of  $d + 1$  points in  $\mathfrak{R}^d$ , there is at least one point  $x \in S$  that can be expressed as a linear combination of the other points in  $S$ .

Recall that the VC-dimension is the size of the largest shattered set. To show the VC-dimension is some value  $k$ , you need to show that there is some set of size  $k$  that is shattered, and that no set of size  $k + 1$  is shattered.

For significant partial credit, find the VC-dimension of homogeneous half-spaces on the real line ( $d = 1$ ) and the real plane ( $d = 2$ ).

First, recall that a subset  $S$  of the domain (a set of *unlabeled* instances) is shattered by a hypothesis class  $\mathcal{H}$  if each of the  $2^{|S|}$  ways of labeling  $S$  is consistent with an  $h$  in  $\mathcal{H}$ .

The key question many students have is what is the hypothesis class for this problem. Each hypothesis is a partitioning of the domain and has (at least one) parameterization as a vector  $\mathbf{w}$  in  $\mathfrak{R}^d$ .

In one dimension, the  $h$ 's are parameterized by a scalar  $w \in \mathfrak{R}$ . When  $w = 0$  then  $wx = 0 \not> 0$  for all  $x$ , and the corresponding  $h$  maps all  $x$  to  $-1$ . When  $w > 0$  then  $wx > 0$  if and only if  $x > 0$ . So all  $w > 0$  are parameterizations of the same  $h$  that maps the positive reals to  $+1$  and the non-positive reals to  $-1$ . Finally, when  $w$  is negative, then  $wx > 0$  if and only if  $x < 0$ , so negative  $w$ 's are parameterizations of the  $h$  that maps the negative reals to  $+1$  and the non-negative reals to  $-1$ .

In two dimensions, things are a little more interesting. The  $h$  associated with  $(1, 1)$  maps instances  $\mathbf{x} = (x_1, x_2)$  to  $+1$  if and only if  $x_1 + x_2 > 0$ . So this  $h$  maps the positive quadrant to  $+1$  and the origin and the negative quadrant to  $-1$ . The other two quadrants are split by a diagonal line going through the origin. Similarly,  $(2, 2)$  or any parameterization  $(c, c)$  (where  $c$  is positive) represents the same hypothesis. On the other hand, the parameterization  $(-1, -1)$  has the same decision boundary, but maps the positive quadrant (and the origin) to  $-1$  while mapping the negative quadrant to  $+1$ .

Different parameterizations, like  $(1, 2)$  or  $(2, -1)$ , result in  $h$ 's with different decision boundaries, but multiplying a  $\mathbf{w}$  by the a positive scalar results in the same hypothesis  $h$ . For example,  $\mathbf{w} = (1, -1)$  represents the  $h$  that predicts  $+1$  if and only if  $x_1 > x_2$ , as does  $\mathbf{w} = (2, -2)$ .

Let me show what the VC-dimension is for a simple hypothesis class. Consider the hypothesis class  $\mathcal{H}'$  over domain  $\mathfrak{R}$  that consists of intervals of the real line. (Note that this

class has little if anything to do with the homework problem.) More precisely, for each pair  $a \leq b \in \mathfrak{R}$  there is an  $h$  in  $\mathcal{H}'$  such that  $h(x) = +1$  if and only if  $a \leq x \leq b$ , and  $h(x) = -1$  otherwise. No other  $h$ 's are in  $\mathcal{H}'$ . One consequence of this definition is that the empty hypothesis (that maps everything to  $-1$  is not in  $\mathcal{H}'$ ).

We claim that the VC dimension of  $\mathcal{H}'$  is 2. First, consider the set  $S = \{0, 2\}$ . This set is shattered by  $\mathcal{H}'$  since the interval  $(a = -2, b = -1)$  maps everything in  $S$  to  $-1$ , the interval  $(-3, 5)$  maps everything in  $S$  to  $+1$ , and the intervals  $(-1, 1)$  and  $(1, 3)$  give the other two possible labelings of  $S$ .

Next, we show that no 3-element set  $S = \{x_1, x_2, x_3\}$  is shattered by  $\mathcal{H}'$ . Assume to the contrary that some 3-element set  $S = \{x_1, x_2, x_3\}$  is shattered by  $\mathcal{H}'$ . Since  $S$  is a set,  $x_1, x_2$  and  $x_3$  are distinct (no two are equal). Without loss of generality assume that  $x_1 < x_2 < x_3$  (renaming the elements if needed). Now consider the  $h \in \mathcal{H}'$  such that  $h(x_1) = +1$ ,  $h(x_2) = -1$  and  $h(x_3) = +1$  and an  $a, b$  pair associated with that  $h$  (although there is only one parameterization for each  $h \in \mathcal{H}'$  there might be multiple parameterizations for different classes, like those in the homework). Such an  $h$  exists since, by assumption,  $S$  is shattered so there is some  $h$  consistent with any particular labeling of  $S$ .

Because of the way  $h$  labels  $S$ , we know that:

$$\begin{aligned} a &\leq x_1 \leq b \\ a &\leq x_3 \leq b \end{aligned}$$

Furthermore  $a < x_2$  since  $a \leq x_1 < x_2$  and  $x_2 < b$  since  $x_2 < x_3 \leq b$ . This implies that  $h(x_2) = +1$ , contradicting the fact that  $h(x_2) = -1$ . Therefore the assumption that  $\mathcal{H}$  shatters a 3-element set is false.

Since  $\mathcal{H}$  shatters a two-element set, but no three-element set, the VC dimension of  $\mathcal{H}'$  is two.