

Recall that a *hypotheses* is a mapping from the domain  $\mathcal{X}$  to two values (like  $+, -$  or  $0, 1$ ). Each hypothesis in the class can be viewed as a subset of  $\mathcal{X}$  (the points that it maps to  $+$  or  $1$ ). A *hypothesis class* is a set of hypotheses.

A subset  $S$  of the domain  $\mathcal{X}$  (a set of *unlabeled* instances) is *shattered* by a hypothesis class  $\mathcal{H}$  if:

for each of the  $2^{|S|}$  ways of labeling  $S$  there is a hypothesis  $h$  in  $\mathcal{H}$  consistent with that labeling.

Note that if  $S$  is shattered by  $\mathcal{H}$ , then every subset of  $S$  is also shattered by  $\mathcal{H}$ . Also, if  $S$  contains  $k$  points and  $S$  is shattered by  $\mathcal{H}$  then  $\mathcal{H}$  must contain at least  $2^k$  hypotheses.

The VC-dimension of a hypothesis class is the size (cardinality) of the largest set shattered by the hypothesis class. In other words, hypothesis class  $\mathcal{H}$  has VC-dimension  $k$  if there exists some subset  $S \subseteq X$  of the domain with  $|S| = k$  that is shattered by  $\mathcal{H}$ , and every subset  $S' \subseteq X$  with  $|S'| > k$  is *not* shattered by  $\mathcal{H}$ .

To show the VC-dimension of a hypothesis class is some value  $k$ , you need to show both:

1. there is some set of size  $k$  that is shattered, and
2. no set of size  $k + 1$  is shattered (i.e. all sets of size  $k + 1$  are *not* shattered)

If no set of size  $k + 1$  is shattered by  $\mathcal{H}$  then no set of size  $k' > k$  can be shattered by  $\mathcal{H}$ . This can be shown by contradiction: assume to the contrary that some set  $S'$  of size  $k' > k$  is shattered by  $\mathcal{H}$ . By the observation in the 2nd paragraph, every subset of  $S'$  is shattered by  $\mathcal{H}$ . Since  $S'$  has more than  $k$  elements,  $S'$  has a subset of size exactly  $k + 1$  that is shattered, giving the contradiction.

The VC dimension is used in generalization bounds: based on the VC dimension of the hypothesis class and the error rate on the training set, the error rate on new unseen points can be bounded. Assume that:

1. training examples and (labeled) test examples are drawn i.i.d. from the same (possibly unknown) distribution on  $\mathcal{X} \times \{0, 1\}$ ,
2. the algorithm always outputs a hypothesis in a class of VC-dimension  $d$ , and
3. the training set has  $N > d$  examples.

Pick any  $\eta > 0$ . Then, with probably at least  $1 - \eta$ , the hypothesis output by the learner will have an error probability on new examples that is bounded by:

$$\text{error rate on training set} + \sqrt{\frac{d(\log(2N/d) + 1) + \log(\frac{4}{\eta})}{N}}.$$

Note that this result doesn't require that the "target" is in the hypothesis class or that the distribution is noise free.

This bound can be used to choose between hypothesis classes (as in "structural risk minimization"), or to estimate the number of samples needed. However, the bound is often pessimistic, and in practice error probabilities are often much smaller than the bound.

For more information, see Section 2 of the Burges SVM tutorial (on the web page), the wikipedia page on VC dimension, and/or Andrew Moore's VC dimension tutorial at <http://www-2.cs.cmu.edu/~awm/tutorials/vcdim.html>.