13b6b78e64c21108a38a4346157385e7
ebrary


## 12.2 SHATTERING

If $C$ contains only a finite number of rules, then results on the estimation error as a function of the number of training examples can be obtained by some simple probabilistic arguments. However, if the number of rules in $C$ is large, then these results are not so useful. Moreover, if $C$ contains infinitely many rules, then the results are downright useless. Simply counting the number of rules in $C$ is not the right measure of complexity (or richness) of $C$. A good measure of richness should take into account the "expressive power" of the rules in $C$. The notion of "shattering" is one way to capture the expressive power. Before giving a precise definition, we first try to motivate this notion with an example.

Suppose someone tells us that before the start of each week they can predict whether the stock market (say the S&P 500 index) will end the week higher or lower. In order to make this prediction, they measure various features such as the price behavior of the index over the previous several weeks, the recent behavior of interest rates and other financial and economic indicators, perhaps some company specific features such as earnings, and possibly others (such as whether the AFC or NFC won the Super Bowl that January, which has been playfully suggested to predict stock market performance for the year).

13b6b78e64c21108a38a4346157385e7
ebrary

Kulkarni, Sanjeev, and Harman, Gilbert. Wiley Series in Probability and Statistics : Elementary Introduction to Statistical Learning Theory (1). Hoboken, US: Wiley, 2011. ProQuest ebrary. Web. 12 July 2016.
Copyright © 2011. Wiley. All rights reserved.

13b6b78e64c21108a38a4346157385e7
ebrary


If we know that they use a fixed decision rule and find that they make correct decisions for 10 weeks straight without yet making an error, we may be rather impressed and place some confidence in their decision rule. If the "winning streak" of correct redictions continues for 52 weeks, we would be extremely impressed (and also possibly wealthy if we had invested according to the predictions). The chance that a random decision rule could match the outcomes for 52 straight weeks is extremely small (one in $2^{52}$), so we would be quite confident that they are really on to something with the rule they have come up with. Even with just 10 straight correct predictions, the chance a random rule would achieve this performance is one in 1024.

On the other hand, suppose they tell us that instead of using a fixed decision rule, they have a collection of possible decision rules they may use. After the 10 weeks they search through their collection of decision rules and find one that agrees with the outcome for all the 10 weeks. Should we be impressed? Well, that should depend greatly on how many rules are in their class. If they have 1024 rules in their class and each of the possible 10-week outcomes is predicted by one of the rules, we should not be impressed at all. Of course one of the rules will agree, no matter what the results!

To be more precise, what really matters is not how many rules are in the class, but rather how many of the possible 10-week outcomes are represented by rules from the class. For example, even if there are thousands of rules in the class, but each one gives the same predictions over the 10 weeks, then we should still be impressed if the predictions are correct. Formalization of this leads to the notion of shattering.

**Definition (Shattering)** Given a set of feature vectors $\overline{x}_1, \ldots, \overline{x}_n$, we say that $\overline{x}_1, \ldots, \overline{x}_n$ are *shattered* by a class of decision rules $C$ if all the $2^n$ labelings of the feature vectors $\overline{x}_1, \ldots, \overline{x}_n$ can be generated using rules from $C$.

13b6b78e64c21108a38a4346157385e7
ebrary

Each rule from $C$ will classify each of the feature vectors $\overline{x}_1, \ldots, \overline{x}_n$ into either class 0 or class 1. Thus, each rule splits the set of feature vectors into a subset that gets labeled 1 and its complement that gets labeled 0. There are $2^n$ possible subsets (or possible labelings). $C$ shatters $\overline{x}_1, \ldots, \overline{x}_n$ if by using rules from $C$ we can carve the feature vectors in all possible ways. This means that using a suitable rule from $C$, we could generate any possible prediction for the given feature vectors.

## 12.3 VC DIMENSION

Suppose we see labeled examples $(\overline{x}_1, y_1), \ldots, (\overline{x}_n, y_n)$. If the set of feature vectors $\overline{x}_1, \ldots, \overline{x}_n$ is shattered by $C$, then certainly we can find a rule from $C$ that agrees completely with the training examples. However, from the discussion in the previous section, we expect that choosing such a rule will have little predictive power, even though it fits the data. Moreover, we will have little confidence that

13b6b78e64c21108a38a4346157385e7
ebrary

Kulkarni, Sanjeev, and Harman, Gilbert. Wiley Series in Probability and Statistics : Elementary Introduction to Statistical Learning Theory (1). Hoboken, US: Wiley, 2011. ProQuest ebrary. Web. 12 July 2016.
Copyright © 2011. Wiley. All rights reserved.

13b6b78e64c21108a38a4346157385e7
ebrar


the rule we pick is even close to the best rule from $C$. We will need much more data to be confident of this.

Hence, if $C$ shatters a large set of feature vectors, learning will be difficult if these feature vectors are observed as the training examples. The amount of data we need to learn will be large compared to the number of feature vectors shattered.

Now remember that for PAC learning, we wish to learn for *any* distributions, that is any prior probabilities $P(0)$, $P(1)$ and any conditional distributions $P(\overline{x}|0)$, $P(\overline{x}|1)$. Recall from Chapter 11 that this is equivalent to learning for any distribution $P(\overline{x})$ for the feature vector $\overline{x}$, and any conditional probabilities $P(0|\overline{x})$, $P(1|\overline{x})$. The amount of data needed for $\epsilon, \delta$ learning of a class $C$ is governed by the "bad" distributions that make learning difficult.

Thus, if $C$ shatters *some* set of feature vectors $\overline{x}_1, \ldots, \overline{x}_n$, then for *some* distributions learning will be difficult. In particular, the distribution could be concentrated on these feature vectors, so as we see labeled examples, there are always rules agreeing with the data, but providing no information about unseen feature vectors. This discussion leads to the following definition.

**Definition (VC Dimension)** The *Vapnik-Chervonenkis dimension* (or *VC dimension*) of a class of decision rules $C$, denoted by $\text{VCdim}(C)$, is the largest integer $V$ such that *some* set of $V$ feature vectors is shattered by $C$. If arbitrarily large sets can be shattered, then $\text{VCdim}(C) = \infty$.

Remember that an important point here is that we only need *some* set of $V$ points (as opposed to *all* sets of $V$ points) to be shattered to make the VC dimension equal to $V$. The VC dimension of a class is the measure of richness that we are after. As we have argued intuitively and state precisely in the next section, this measure characterizes the PAC learnability of a class $C$.

Kulkarni, Sanjeev, and Harman, Gilbert. Wiley Series in Probability and Statistics : Elementary Introduction to Statistical Learning Theory (1). Hoboken, US: Wiley, 2011. ProQuest ebrary. Web. 12 July 2016.
Copyright © 2011. Wiley. All rights reserved.

13b6b78e64c21108a38a4346157385e7
ebrary


## 12.5 SOME EXAMPLES

In this section we consider several examples for which computation of the VC dimension is relatively straightforward. As we see in the examples, in order to find the VC dimension of a class, the usual approach is to obtain both upper and lower bounds. If we are lucky (as we will be in the examples below), the bounds will match and we will determine the VC dimension exactly. In more complicated situations, we often have to satisfy ourselves with bounds that do not match, but give some idea of the dimension.

In order to get lower bounds on $\mathrm{VCdim}(C)$ (that is, to show that the VC dimension is larger than some quantity), it is enough to find *some* set of feature vectors

13b6b78e64c21108a38a4346157385e7
ebrary

Kulkarni, Sanjeev, and Harman, Gilbert. Wiley Series in Probability and Statistics : Elementary Introduction to Statistical Learning Theory (1). Hoboken, US: Wiley, 2011. ProQuest ebrary. Web. 12 July 2016.
Copyright © 2011. Wiley. All rights reserved.

13b6b78e64c21108a38a4346157385e7
ebrary


that are shattered. This is usually done by selecting some specific points and explicitly showing that all labelings (subsets) of these points can be generated by rules from $C$. If we find $k$ feature vectors that are shattered by $C$, then we know that $\text{VCdim}(C) \geq k$.

Obtaining upper bounds (that is, showing that $\text{VCdim}(C)$ is less than some quantity) is usually more difficult. To show that $\text{VCdim}(C) < k$, we need to argue that *no* set of $k$ feature vectors can be shattered by $C$. For the upper bound, it is not enough to exhibit some set of points that cannot be shattered. This usually requires a more careful argument, sometimes considering several cases for different arrangements of the feature vectors.

**Example 12.1 (Intervals in 1-Dimension)** Let each feature vector consist of a single real value. Let the decision rules $C$ be the set of all closed intervals of the form $[a,b]$ for real numbers $a,b$. That is, we decide 1 if $a \leq x \leq b$, and we decide 0 otherwise. What is $\text{VCdim}(C)$?

**Example 12.2 (Unions of Intervals)** As in the previous Example, let each feature vector consist of such a single real value. But now, let the decision rules $C$ be the set of all finite unions of intervals. What is $\text{VCdim}(C)$?

Kulkarni, Sanjeev, and Harman, Gilbert. Wiley Series in Probability and Statistics : Elementary Introduction to Statistical Learning Theory (1). Hoboken, US: Wiley, 2011. ProQuest ebrary. Web. 12 July 2016.
Copyright © 2011. Wiley. All rights reserved.

13b6b78e64c21108a38a4346157385e7
ebrary


**Example 12.3 (Half-spaces in 2-D)** Now, let each feature vector be a point in the plane, so that the feature space is two-dimensional. Let $C$ be the set of all half-spaces in the plane. This is precisely the set of all decision regions with a linear decision boundary, which are the decision rules that can be represented by perceptrons. What is VCdim($C$)?

Kulkarni, Sanjeev, and Harman, Gilbert. Wiley Series in Probability and Statistics : Elementary Introduction to Statistical Learning Theory (1). Hoboken, US: Wiley, 2011. ProQuest ebrary. Web. 12 July 2016.
Copyright © 2011. Wiley. All rights reserved.