IX The VC Dimension

1. Introduction
2. Review: Sample Complexity and $|H|$
3. The Vapnik-Chervonenkis (VC) Dimension
4. Sample Complexity and $VC(H)$
5. Some Theorems on $VC(H)$ and Learnability
6. VC Dimension and Bias
7. Summary of Lecture
8. References

1 Introduction

We have spent the last 3 lectures on efficiency.

Now we change subject and look at a new and better measure of hypothesis complexity.

This will improve our sample complexity bounds.

Notes:
This lecture covers chapter 7 of Anthony and Biggs (but not 7.4 or 7.5). We see a few results from chapter 8. We will not cover chapters 9 or 10 in this course.

See also Mitchell, chapter 7.

2 Review: Sample Complexity and $|H|$

In lecture 5 we proved that if H is finite it is potentially learnable.

We also saw how many examples m are needed (the *sample complexity*) for a consistent algorithm to PAC learn H:

$$m \geq \left\lceil \frac{1}{\epsilon} \ln \frac{|H|}{\delta} \right\rceil \tag{1}$$

We are using $|H|$ to measure the complexity of H.

This is useful, but:

- We cannot use (1) when $|H|$ is infinite.
- $|H|$ is not the best estimate of the complexity of H.

3 The Vapnik-Chervonenkis (VC) Dimension

$VC(H)$ is a measure of the expressive power of H.

If we measure the complexity of H using $VC(H)$ instead of $|H|$:

- We can sometimes estimate the sample complexity of infinite H.
- We often obtain better upper bounds on sample complexity than when using $|H|$.
- We can obtain a *lower* bound on sample complexity.
- It can be shown that these bounds are fairly tight.

Shattering a Sample

Suppose we have a sample of length m:

$$(x_1, x_2, \ldots x_m)$$

How many possible hypotheses are there for the sample?

- Each example can be labelled as either 0 or 1, so there are 2^m possible hypotheses.

How many of these hypotheses does H actually contain?

- The more hypotheses H contains, the greater its expressive power is.

Shattering a Sample

We say H *shatters* a sample if H contains all 2^m hypotheses for it.

- In other words, if H does *not* shatter a sample there is at least one hypothesis it cannot represent.

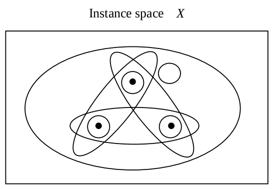


Figure from "Machine Learning" by Tom Mitchell.

The VC Dimension

The VC dimension of H – written $VC(H)$ – is the maximum length of sample shattered by H.

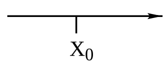
Example Suppose:

- $X = \Re$, the set of real numbers.
- H = the set of intervals on the real number line (i.e., hypotheses have the form $a < x < b$, where a and b are any real constants).

What is the VC dimension of H?

Shattering 1 Example

Suppose the sample contains 1 example x_0:



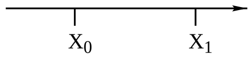
Can H shatter this sample? There are $2^{m=1}$ ways to classify $m = 1$ examples. We can express these cases as follows:

case 1: $a < b < x_0$ or $x_0 < a < b$

case 2: $a < x_0 < b$

Shattering 2 Examples

Suppose the sample contains 2 examples $x_0 < x_1$:



The following $2^{m=2}$ cases shatter any two examples:

case 1: $a < b < x_0 < x_1$ or $x_0 < x_1 < a < b$

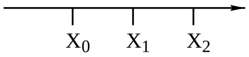
case 2: $a < x_0 < b < x_1$

case 3: $x_0 < a < x_1 < b$

case 4: $a < x_0 < x_1 < b$

(Not) Shattering 3 Examples

Suppose the sample contains 3 examples $x_0 < x_1 < x_2$:



Notice that no interval can contain both x_0 and x_2, but not x_1.

Therefore, no set of 3 examples can be shattered, and $VC(H) = 2$.

Comparing $|H|$ and $VC(H)$

In the last example $|H|$ is infinite but $VC(H)$ is finite.

For this H the two measures of complexity are very different:

- $|H|$ suggests this H will be difficult to learn.
- $VC(H)$ suggests this H will not be so difficult to learn.

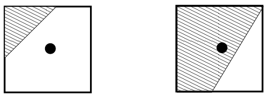
Another Example

Let:

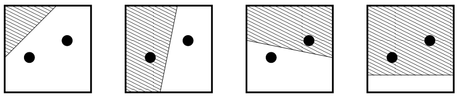
- X be the set of points in the x, y plane.
- H be the set of all linear decision surfaces in the plane.

What is the VC dimension of H?

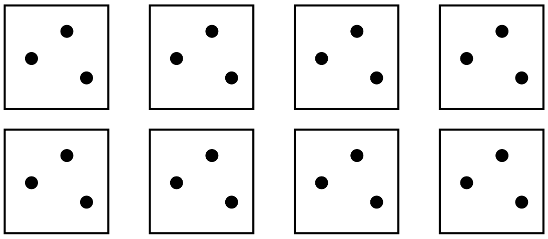
Shattering 1 Example



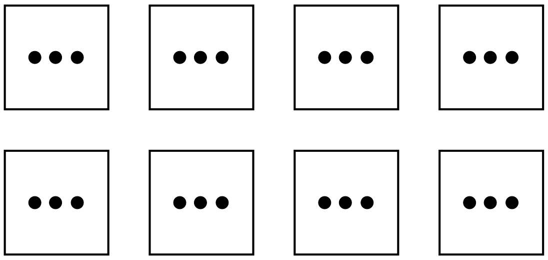
Shattering 2 Examples



Shattering 3 Examples



Shattering 3 Collinear Examples



VC Dimension of the Example

Since there exists a 3-example sample which can be shattered, $VC(H)$ is at least 3.

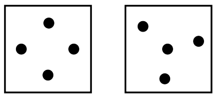
To show $VC(H) = 3$ we must show that no 4-example sample can be shattered.

What did we learn about this H from shattering 3 examples?

- A set of examples $\mathbf{x}$ is shattered by H only if any subset of $\mathbf{x}$ and its difference are linearly separable.
- No 3 collinear points are linearly separable.

Shattering 4 Examples

Is it possible for a set of 4 points not to contain 3 non-collinear points? There are 2 ways:



On the left the 4 examples are boundaries for a polygon, while on the right one example lies within the polygon.

In neither case is every subset of examples linearly separable. Therefore, no sample of length 4 can be shattered, and $VC(H) = 3$.

4 Sample Complexity and $VC(H)$

$VC(H)$ is a better measure of the complexity of H than $|H|$.

Using $VC(H)$ we can obtain a new upper bound on sample complexity:

$$m \geq \frac{1}{\epsilon}\left(4 \log_2(2/\delta) + 8VC(H) \log_2(13/\epsilon)\right) \qquad (2)$$

This bound is often tighter than the old bound (1).

See Mitchell p. 217, or Anthony and Biggs p. 94.

Lower Bounds on Sample Complexity

It is also possible to obtain *lower* bounds on complexity:

- See Anthony and Biggs p. 96 for a simple bound on the worst possible consistent learning algorithm.
- See Anthony and Biggs p. 97 or Mitchell p. 217 for the following more powerful result due to Ehrenfeucht et al. (1989), which applies to any consistent algorithm:

$$m \geq \max\left[\frac{1}{\epsilon}\ln(1/\delta), \quad \frac{VC(C)-1}{32\epsilon}\right] \qquad (3)$$

for $0 < \epsilon < 1/8$ and $0 < \delta < 1/100$.

The lower bounds tell us that if there are not enough examples, no learner can PAC-learn every target concept in C.

Comparison of Upper and Lower Bounds

Note that the upper (2) and lower (3) bounds based on VC dimension are similar:

Both are logarithmic in $1/\delta$ and linear in $VC(H)$.

This shows the bounds are fairly tight.

5 Some Theorems on $VC(H)$ and Learnability

Theorem If H is finite then $VC(H) \leq \log_2|H|$.

Proof See Anthony and Biggs page 76, or Mitchell p. 215.

Theorem If $VC(H)$ is infinite then H is not potentially learnable.

Proof See Anthony and Biggs page 87.

Theorem If $VC(H)$ is finite then H is potentially learnable.

Proof See Anthony and Biggs page 90.

6 VC Dimension and Bias

The larger $VC(H)$ is, the less bias H has.

Recall from lecture 3 that an *Unbiased Generalisation Language* (UGL) can express all hypotheses.

If H shatters a sample, H is a UGL *for that sample.*

In lecture 3 we saw that we needed bias for inductive learning.

In this lecture we saw that infinite $VC(H)$ means H is not potentially learnable.

7 Summary of Lecture

We noted that we cannot use (1) when $|H|$ is infinite.

We saw the VC dimension is a better measure of the complexity of H than $|H|$:

- We saw a bound on sample complexity (2) using $VC(H)$ which is often better than the old bound (1) which uses $|H|$.
- We can sometimes use VC(H) to bound sample complexity of infinite H.

We noted that there exist lower bounds on sample complexity, and that the upper and lower bounds based on VC dimension are close.

Summary continued

We noted that:

- Infinite $VC(H)$ implies H is not PAC learnable.
- Finite $VC(H)$ implies H is PAC learnable.

We related the VC dimension and bias.

8 References

Ehrenfeucht, A., Haussler, D., Kearns, M. and Valiant, L. (1989). A general lower bound on the number of examples needed for learning. *Information and Computation*, 82, 247–261.

Mitchell, T. (1997). Machine Learning. McGraw Hill.