Lesson 07 - The Vapnik-Chervonenkis Dimension

The VC Dimension

The Vapnik-Chervonenkis Dimension (VCD) is a measure of how ‘flexible’ or ‘expressive’ a concept class is. It can be used for both discrete spaces, like Σ^*, and for continuous spaces like $\mathbb{R}^n$.

In order to understand the VCD, one must first understand *shattering*.

Shattering

Consider a space X, a concept class $C \subseteq 2^X$, and a finite sample S drawn from X.

Definition. The concept class C **shatters** S if and only if for every subset $s \subseteq S$, there exists $c \in C$, such that $c \cap S = s$. In other words, for all the possible ways S can be divided into two disjoint groups, there is a concept in C realizing that division.

Exercise. Let Σ ={a,b,c} and consider the concept class 2-SL. Which of the following string sets does 2-SL shatter?

1. {aa, bb}
2. {aa, bb, cc}
3. {aa, a}

Exercise. Consider $\mathbb{R}^2$ and the concept class of axis-aligned rectangles.

1. Give four rectangles that shatter this set of points: { (2,2) (4,4) }.
2. Can this set of points be shattered: { (2,2) (4,4) (6,6)}?
3. Provide three points which can be shattered by the class of axis-aligned rectangles.

Shattering is a way measuring the limits to the kinds of distinctions a concept class can make. Making distinctions (and not making them) is what learning is all about!

Definition of the VC Dimension

Definition. The **VC dimension** of a concept class C is the size of the largest sample S that can be shattered by C. If there is no bound on the size of a sample that can be shattered by C then the VC dimension is said to be infinite.

Exercise. Consider $\mathbb{R}$ and the concept class of closed intervals over this space. For example, the interval [0,2] picks out all points between 0 and 2 inclusive, and the interval $[-\pi, \pi]$ picks out all the points between $-\pi$ and π inclusive.

1. Are there any samples of size one shatterable by this concept class? If so, then the VCD of this class is at least 1.
2. Are there any samples of size two shatterable by this concept class? If so, then the VCD of this class is at least 2.
3. Are there any samples of size three shatterable by this concept class? If so, then the VCD of this class is at least 3. If not, then the VCD of this class is at most 2.

Exercise. Consider $\mathbb{R}^2$ and the concept class of axis-aligned rectangles.

1. What is the VCD of this class?

Exercise. Let Σ ={a,b,c} and consider the concept class of finite languages.

1. What is the VCD of this class?

Exercise. Let Σ ={a,b,c} and consider the concept class of 2-SL languages.

1. What is the VCD of this class?

PAC learning and the VCD

Theorem. A concept class is PAC-learnable if and only if it has finite VC dimension.

- A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, *Learnability and the Vapnik-Chervonenkis dimension*, Journal of the ACM 36 (1989), 929-965.

Corollary. Every concept class of finite size is PAC-learnable. (Because they all have finite VC dimension.)

Corollary. The class of finite languages is not PAC learnable. (Because its VC dimension is infinite)

Corollary. None of the larger classes of the Chomsky Hierarchy (regular, context free, context sensitive classes are PAC learnable. (Because their VC dimension is infinite)

Exercise. How do these results compare to identification in the limit from positive data?