Lecturer: Hung Q. Ngo
Last update: October 15, 2011

Vapnik-Chervonenkis Dimension

1 Vapnik-Chervonenkis Dimension

Valiant's theorem from the previous lecture is meaningless for infinite hypothesis classes, or even classes with more than exponential size. In 1968, Vladimir Vapnik and Alexey Chervonenkis wrote a very original and influential paper (in Russian) [5, 6] which allows us to estimate the sample complexity for infinite hypothesis classes too. The idea is that the size of the hypothesis class is a poor measure of how "complex" or how "expressive" the hypothesis class really is. A better measure is defined, called the *VC-dimension* (VCD) of a function class. Then, a version of Valiant's theorem is proved with respect to the VCD of $\mathcal{H}$, which can be finite for many commonly used infinite hypothesis class $\mathcal{H}$. (More technically, Vapnik-Chervonenkis used VCD to derive bounds for expected loss given empirical loss; more on this point later.)

Roughly speaking, the VC-dimension of a function (i.e. hypothesis) class is the maximum number of data points for which, no matter how we label them (with 0/1), there is always a hypothesis in the class which perfectly explains the labeling. This measure is a much better indicator of the model's capability than the number of parameters used to describe the models. Blumer et al. [1] first brought VCD to the attention of the COLT community. The following snippet from J. Hosking, E. Pednault, and M. Sudan (1997) describes the strength of VC theory well:

> Because VC dimension is defined in terms of model fitting and number of data points, it is equality applicable to linear, nonlinear and nonparametric models, and to combinations of dissimilar model families. This includes neural networks, classification and regression trees, classification and regression rules, radial basis functions, Bayesian networks, and virtually any other model family imaginable. In addition, VC dimension is a much better indicator of the ability of models to fit arbitrary data than is suggested by the number of parameters in the models. There are examples of models with only one parameter that have infinite VC dimension and, hence, are able to exactly fit *any* set of data. There are also models with billions of parameters that have small VC dimensions, which enables one to obtain reliable models even when the number of data samples is much less than the number of parameters. VC dimension coincides with the number of parameters only for certain model families, such as linear regression/discriminant models. VC dimension therefore offers a much more general notion of degrees of freedom than is found in classical statistics.

Let us now define VCD more formally. Each hypothesis $h : \Omega \to \{0, 1\}$ can naturally be viewed as a subset of Ω, where $h = \{\omega \mid h(\omega) = 1\}$. We will refer to h as a binary function and h as a subset interchangeably. For any finite subset $S \subseteq \Omega$, let $\Pi_{\mathcal{H}}(S) = \{h \cap S \mid h \in \mathcal{H}\}$. This is called the *projection* of h onto S, or the set of *dichotomies* or *behaviors* of h on S.

Definition 1.1. The set S is *shattered* by the function class $\mathcal{H}$ if $|\Pi_{\mathcal{H}}(S)| = 2^{|S|}$.

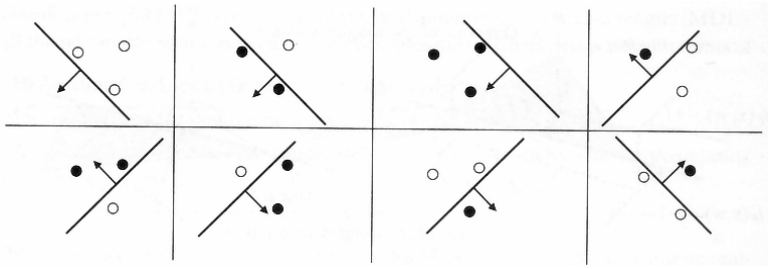

Figure 1: Three non-colinear points are shattered by the set of half-planes

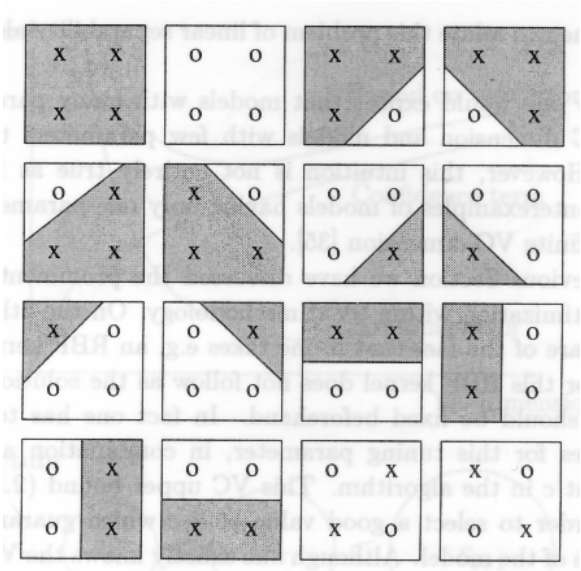

Figure 2: The class of half-planes cannot shatter 4 poitns in $\mathbb{R}^2$

In other words, the set S is shattered by $\mathcal{H}$ if, no matter how we assign $0/1$-labels to points in S, there's always a hypothesis in $\mathcal{H}$ which "explains" the labeling perfectly. For example, let $\mathcal{H}$ be the set of half-planes on $\mathbb{R}^2$. Let S be any three points on $\mathbb{R}^2$ which are not colinear. Then, S is shattered by $\mathcal{H}$ because all 8 ways of assigning labels to S are explainable by the half-planes. See Figure 1 for an illustration.

Definition 1.2. The *VC-dimension* (VCD) of a function class (or hypothesis class) $\mathcal{H}$ is the maximum size of a subset of Ω shattered by $\mathcal{H}$.

In other words, if $\text{VCD}(\mathcal{H}) = d$ then $\mathcal{H}$ cannot shatter any $d+1$ points *and* it can shatter some d points. Figure 1 explains intuitively why the class of half-planes cannot shatter four points in $\mathbb{R}^2$. We have thus shown that the set of all half-planes on $\mathbb{R}^2$ has VCD = 3. Here are some examples which should make things clearer:

Exercise 1. Show that

- If $\mathcal{H}$ is finite then $\text{VCD} \leq \log_2 |\mathcal{H}|$
- The set of all intervals $[x, \infty), x \in \mathbb{R}$ has VCD = 1.

- The set of all closed intervals on $\mathbb{R}$ has VCD = 2.
- The set of all axis-aligned rectangles on $\mathbb{R}^2$ has VCD = 4
- The set of all half-spaces on $\mathbb{R}^d$ has VCD = d+1

Exercise 2. Show that

- The set of all balls on $\mathbb{R}^d$ has VCD = d+1
- The set of all d-vertex convex polygons on $\mathbb{R}^2$ has VCD = 2d+1
- The set of all sets of intervals on $\mathbb{R}$ has VCD = ∞

We have claimed that VCD is a good measure of the "expressiveness" of a function class. Here is a result which roughly states that if the concept class is too expressive then we cannot PAC-learn it. We will not prove it formally because we shall prove a more general lower-bound in the next lecture.

Theorem 1.3. *If the concept class $\mathcal{C}$ has $\text{VCD}(\mathcal{C}) = \infty$, then there is no algorithm to PAC-learn $\mathcal{C}$ using $\mathcal{C}$.*

Proof. Fix $\epsilon = 1/8$ and $\delta = 1/10$. We can show that there exists a concept $c \in \mathcal{C}$ which no algorithm can PAC-learn with error ϵ and confidence $9/10$. The idea is to construct a particular probability distribution $\mathcal{D}$ on Ω. Then, pick a random target concept c. Finally, show that the algorithm produces a hypothesis whose expected error (over the random choices of c) is greater than ϵ. □

2 Sauer-Shelah Lemma

We need the following lemma to prove the VC-dimension-based sample complexity theorem. The following lemma was first proved by Vapnik-Chervonenkis [7], and rediscovered many times (Sauer [3], Shelah [4]), among others. It is often called the Sauer lemma or Sauer-Shelah lemma in the literature. (Sauer said that Paul Erdös posed the problem.)

Lemma 2.1 (Sauer-Shelah 1972)**.** *Suppose* VCD $(\mathcal{H}) = d < \infty$. *Define the* growth function

$$\Pi_{\mathcal{H}}(m) = \max\{|\Pi_{\mathcal{H}}(S)| \ : \ S \subseteq \Omega, |S| = m\}.$$

(Note that if VCD $(\mathcal{H}) = \infty$ *then* $\Pi_{\mathcal{H}}(m) = 2^m, \forall m$*.) Then,*

$$\Pi_{\mathcal{H}}(m) \leq \Phi_d(m) := \sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{em}{d}\right)^d = O(m^d).$$

Proof. Without loss of generality, we assume $m > d$, because if $m \leq d$ then $\Phi_d(m) = 2^m$ and the inequality is trivial.. Let S be an arbitrary set of m points from Ω. Let $\mathcal{F} = \Pi_{\mathcal{H}}(S)$, then $\mathcal{F}$ is a family of subsets of S. In fact, without loss of generality we set $S = [m]$ (just rename the members of S). We will show that $|\mathcal{F}| \leq \Phi_d(m)$.

We will use the "shifting technique" [2] to construct a family $\mathcal{G}$ of subsets of $[m]$ satisfying the following three conditions:

1. $|\mathcal{G}| = |\mathcal{F}|$

2. $\mathcal{G}$ is closed under containment, i.e. if $A \in \mathcal{G}$, then every subset of A is in $\mathcal{G}$.

3. Every $A \subset [m]$ shattered by $\mathcal{G}$ is also shattered by $\mathcal{F}$

So, instead of upperbounding $|\mathcal{F}|$ we can just upperbound $\mathcal{G}$. Because $\mathcal{G}$ closed under containment, every member of $\mathcal{G}$ is shattered by $\mathcal{G}$, and thus every member of $\mathcal{G}$ is shattered by $\mathcal{F}$. Thus, every member of $\mathcal{G}$ has size at most d, implying $|\mathcal{G}| \leq \Phi_d(m)$ as desired.

We next describe the *shifting* operation which achieves properties 1, 2, 3 by an algorithm.

1: **for** $i = 1$ to m **do**
2: **for** $F \in \mathcal{F}$ **do**
3: **if** $F - \{i\} \notin \mathcal{F}$ **then**
4: Replace F by $F - \{i\}$
5: **end if**
6: **end for**
7: **end for**
8: Repeat steps 1–7 until no further changes is possible.

The algorithm terminates because some set gets smaller at each step. Properties 1 and 2 are easy to verify.

We next verify that property 3 holds. Let A be shattered by $\mathcal{F}$ *after* executing lines 2–6 at any point in the execution. We will show that A must have been shattered by $\mathcal{F}$ *before* the execution. Let i be the element examined in that iteration. To avoid confusion, let $\mathcal{F}'$ be the set family after the iteration. We can assume $i \in A$, otherwise the iteration does not affect the "shatteredness" of A.

Let R be an arbitrary subset of A. As A is shattered by $\mathcal{F}'$, we know there is a set $F' \in \mathcal{F}'$ such that $F' \cap A = R$. If $i \in R$, then $F' \in \mathcal{F}$. Suppose $i \notin R$. There is $T \in \mathcal{F}'$ such that $T \cap A = R \cup \{i\}$. This means $T - \{i\} \in F$, or else T would have been replaced in step 4. But, $T - \{i\} \cap A = R$ as desired. □

Exercise 3 (Inductive proof of Sauer-Shelah lemma)**.** We induct on $m + d$. The $m = 0$ and $d = 0$ cases are trivial. Now consider $m > 0, d > 0$. Let S be an arbitrary set of m points from Ω. Rename members of S so that $S = [m]$. Let $\mathcal{F} = \Pi_{\mathcal{H}}([m])$. Define

$$\mathcal{F}' = \{F \subseteq [m-1] \mid F \in \mathcal{F} \text{ and } F \cup \{m\} \in \mathcal{F}\}.$$

Prove the followings:

(a) $|\mathcal{F}| = |\mathcal{F}'| + |\Pi_{\mathcal{F}}([m-1])|$.

(b) $\Pi_{\mathcal{F}}([m-1]) = \Pi_{\mathcal{H}}([m-1])$

(c) $\text{VCD}(\mathcal{F}') \leq d - 1$.

(d) Finally, by applying the induction hypothesis, show that

$$|\mathcal{F}| \leq \Phi_{d-1}(m-1) + \Phi_d(m-1) = \Phi_d(m).$$

References

[1] A. BLUMER, A. EHRENFEUCHT, D. HAUSSLER, AND M. WARMUTH, *Classifying learnable geometric concepts with the vapnik-chervonenkis dimension*, in Proceedings of the eighteenth annual ACM symposium on Theory of computing, STOC '86, New York, NY, USA, 1986, ACM, pp. 273–282.

[2] P. FRANKL, *Shadows and shifting*, Graphs Combin., 7 (1991), pp. 23–29.

[3] N. SAUER, *On the density of families of sets*, J. Combinatorial Theory Ser. A, 13 (1972), pp. 145–147.

[4] S. SHELAH, *A combinatorial problem; stability and order for models and theories in infinitary languages*, Pacific J. Math., 41 (1972), pp. 247–261.

[5] V. N. VAPNIK AND A. Y. CHERVONENKIS, *On the uniform convergence of relative frequencies of events to their probabilities*, Doklady Akademii Nauk USSR, 181 (1968).

[6] ———, *On the uniform convergence of relative frequencies of events to their probabilities*, Theory of Probability and its Applications, 16 (1971), pp. 264–280.

[7] ———, *Theory of uniform convergence of frequencies of events to their probabilities and problems of search for an optimal solution from empirical data*, Avtomat. i Telemeh., (1971), pp. 42–53.