PRINCETON UNIVERSITY F'02 COS 597D: A THEORIST'S TOOLKIT

Lecture 9: High Dimensional Geometry: VC-Dimension

Lecturer: *Sanjeev Arora* Scribe:*Renato F. Werneck*

We continue our study of high dimensional geometry with the concept of *VC-dimension*, named after its inventors Vapnik and Chervonenkis. This useful tool allows us to reason about situations in which the trivial probabilistic method (i.e., the union bound) cannot be applied. Typical situations are those in which we have to deal with uncountably many objects, as in the following example.

EXAMPLE 1 Suppose we pick set of m points randomly from the unit square in $\Re^2$. Let C be any triangle inside the unit square of area 1/4. The expected number of samplepoints that lie inside it is $m/4$, and Chernoff bounds show that the probability that the this number is not in $m/4[1 \pm \epsilon]$ is $\exp(-\epsilon^2 m)$.

Now suppose we have a collection of $2^{\sqrt{m}}$ triangles of area 1/4. The union bound shows that whp the sample is such that *all* triangles in our collection have $m/4[1 \pm \epsilon]$ samplepoints.

But what if we want the preceding statement to hold for all (*uncountably many*!) triangles of area 1/4 in the unit square? The union bound fails badly! Nevertheless the following statement is true.

THEOREM 1
Let $\epsilon > 0$ and m be sufficiently large. Let S be a random sample of m points in the unit sphere. Then the probability is at least $1 - o(1)$ that every *triangle of area 1/4 has between $m/4[1 \pm \epsilon]$ samplepoints.*

How can we prove such a result? The intuition is that even though the set of such triangles is uncountable, the number of "vastly different" triangles is not too large. Thus there is a small set of typical representatives, and we only need to argue about those.

1 VC-dimension

VC-dimension is used to formalize the above intuition. First, some preliminary definitions:

DEFINITION 1 *A* Range Space *is a pair (X, R), where X is a set and R is a family of subsets of X ($R \subseteq 2^X$).*[1]

DEFINITION 2 *For any $A \subseteq X$, we define the $P_R(A)$, the projection of R on A, to be $\{r \cap A : r \in R\}$.*

DEFINITION 3 *We say that A is* shattered *if $P_R(A) = 2^A$, i.e., if the projection of R on A includes all possible subsets of A.*

With these definitions, we can finally define the *VC-dimension* of a range space:

DEFINITION 4 *The* VC-dimension *of a range space (X, R) is the cardinality of the largest set A that it shatters: VC-dim $= \sup\{|A| : A \text{ is shattered}\}$. It may be infinite.*

[1] 2^X denotes the power set of X

Example 1. Consider the case where $X = \mathbf{R}^2$ (the plane) and R is the set of all axis-aligned squares. The sets A we consider in this case are points on the plane. Let's consider different values of $|A|$:

- $|A| = 1$: With a single point on the plane, there are only two subsets A, the empty set and A itself. Since there for every point there are axis-aligned squares that contain it and others that don't, A is shattered.
- $|A| = 2$: If we have two points, it is easy to find axis-aligned squares that cover both of them, none of them, and each of them separately.
- $|A| = 3$: It is possible to come up with a set of three points such that are shattered; the vertices of an equilateral triangle, for example: there are axis-aligned squares that contain none of the points, each point individually, each possible pair of points, and all 3 points.
- $|A| = 4$: In that case, it is impossible to come up with four points on the plane that are shattered by the set of axis-parallel squares. There will always be some pair of points that cannot be covered by a square.

Therefore, the VC-dimension of this range space if 3.

Example 2. Let $X = \mathcal{F}^n$ (where $\mathcal{F}$ is a finite field) and R be the set of all linear subspaces of dimension d. (Note that these cannot possibly shatter a set of $d+1$ linearly independent vectors.)

First assume $d \leq n/2$. We claim that any set $A = \{v_1, v_2, \ldots, v_d\}$ of d linearly independent vectors is shattered by X. Take any subset I of the vectors. We can build a basis of a d-dimension subspace as follows: first, take the $|I|$ vectors of this set (which are all linearly independent); then, complete the basis with $d - |I|$ linearly independent vectors z_j in $\mathcal{F}^n$ that do not belong to the (d-dimensional) subspace determined by $v_1, v_2, \ldots, v_d$. This basis determines the subspace

$$\text{span}\{\{v_i\}_{i \in I}, z_{d+1}, z_{d+2}, \ldots, z_{d+d-|I|}\},$$

which is d-dimensional, as desired.

But note that this construction works as long as there are $d - |I|$ linearly independent vectors "available" to complete the basis. This happens when $d \leq n/2$. If $d > n/2$, we can pick at most $n - d$ vectors outside the set. Therefore, the VC-dimension of the range space analized is $\min\{d, n - d\}$.

We now study some properties of range spaces related to VC-dimension. The next theorem gives a bound; it is tight (exercise: prove this using the ideas in the previous example).

THEOREM 2
If (X, R) has VC-dimension d and $A \subseteq x$ has n elements then $|P_R(A)| \leq g(d, n)$ where $g(d, n) = \sum_{i \leq d} \binom{n}{i}$.

To prove this theorem, all we need is the following result:

LEMMA 3
If (X, R) has VC-dimension d and $|X| = n$ then $|R| \leq g(d, n)$.

This lemma clearly implies the theorem above: just apply the lemma with A instead of X, and look at the range space $(A, P_R(A))$. We now proceed to the proof of the lemma.

PROOF: We prove Lemma 3 by induction on $d + n$. Let $S = (X, R)$, and consider some element $x \in X$. By definition, we have that

$$S - x = (X - \{x\}, \overbrace{\{r - \{x\} : r \in R\}}^{R_1})$$

and

$$S \setminus x = (X - \{x\}, \overbrace{\{x \in R : x \notin r \text{ but } r \cup \{x\} \in R\}}^{R_2})$$

Note that every element of R that does not contain x is contained in R_1; and every element of R that does contain x is in R_2 (with x removed). Therefore,

$$|R| = |R_1| + |R_2|.$$

A well-known property of binomial coefficients states that

$$\binom{n}{i} = \binom{n-i}{i} + \binom{n-i}{i-1},$$

which implies that

$$g(d, n) = g(d, n-1) + g(d-1, n-1),$$

according to our definition of $g(d, n)$.

In the subproblems defined above, $|X - \{x\}| = n - 1$. We claim that there exists an x such that $S \setminus x$ has VC-dimension at most $d - 1$. That being the case, the inductive hypothesis immediately applies:

$$|R| = |R_1| + |R_2| \leq g(d, n-1) + g(d-1, n-1) = g(d, n).$$

So once we prove the claim, we will be done. Let $A \subseteq S$ be shattered in S with $|A| = d$. We want to prove that in $S \setminus x$ no set $B \subseteq X - \{x\}$ can be shattered if $|B| = d$. But this is straighforward: if B is shattered in $S \setminus x$, then $B \cup \{x\}$ is shattered in S. This completes the proof of Lemma 3. (Incidentally, the bound given by the lemma is tight.) □

The following result is due to Haussler and Welzl (1987), and also appears in Vapnik and Chervonenkis's paper.

THEOREM 4
Let (X, R) be a range space with VC-dimension d, and let $A \subseteq X$ have size n. Suppose S is a random sample of size m drawn (with replacement) from A, and let m be such that

$$m \geq \max\left\{\frac{4}{\epsilon}\log\frac{2}{\delta}, \frac{8d}{\epsilon}\log\frac{8d}{\epsilon}\right\}.$$

Then, with probability at least $1 - \delta$, S is such that for all $r \in R$ such that $|r \cap A| \geq \epsilon n$, we have $|r \cap S| \neq \emptyset$.

PROOF: Define r to be "heavy" if $|r \cap A| \geq \epsilon n$. We will now analyze a single experiment from two different points of view.

1. From the first point of view, we pick *two random samples* N and T, each of size exactly m (note that they both have the same characteristics as the set S mentioned in the statement of the theorem).

   Consider the following events:

   - E_1: there exists a heavy r such that $r \cap N = \emptyset$ (this is a "bad event"; we must prove that it happens with probability at most δ);
   - E_2: there exists a heavy r such that $r \cap N = \emptyset$, but $|r \cap T| \geq \epsilon m/2$.

   We claim that, if E_1 happens with high probability, then so does E_2. Specifically, we prove that $Pr(E_2|E_1) \geq 1/2$. Suppose E_1 has happened and that some specific r is the culprit. Now, pick T. Consider the probability that this r satisfies $|r \cap T| \geq \frac{\epsilon m}{2}$ (this is a lower bound for $Pr(E_2|E_1)$). We know that $|r \cap T|$ is a binomial random variable; its expectation is ϵm and its variance is $\epsilon(1-\epsilon)m$ (and therefore strictly smaller than ϵm. Using Chebyschev's inequality, we get:

   $$\Pr\left(|r \cap T| < \frac{\epsilon m}{2}\right) \leq \frac{\epsilon m}{(\epsilon m/2)^2} = \frac{4}{\epsilon m} \tag{1}$$

   Because the statement of the theorem assumes that $m > 8/\epsilon$, we conclude that indeed $Pr(E_2|E_1) \geq 1/2$.

2. Now consider the same experiment from a different point of view: we pick a *single sample* of size $2m$, which is then partitioned randomly between N and T. Consider the following event:

   E_r: $r \cap N = \emptyset$, but $|r \cap T| \geq \epsilon m/2$.

   It is clear that

   $$E_2 = \bigcup_{r:\text{heavy}} E_r.$$

   Given $N \cup T$, the number of distinct events E_r is no greater than the number of different sets in the projection $P_{N \cup T}(R)$. From Lemma 3, the fact that the VC-dimension of X is d implies that $P_{N \cup T} \leq g(d, 2m)$. (Note: This is the crucial step that uses the VC dimension.)

   Given a choice of $N \cup T$, we now estimate the probability of E_r happening when we select N and T:

   $$P' = \Pr\left(r \cap N = \emptyset |\ |r \cap (N \cup T)| > \frac{\epsilon m}{2}\right).$$

   Let $p = |r \cap (N \cup T)|$. Then, the conditional probability above is

   $$\begin{aligned} P' &= \frac{(2m-p)(2m-p-1)\cdots(m-p+1)}{2m(2m-1)\cdots(m+1)} \\ &= \frac{m(m-1)\cdots(m-p+1)}{2m(2m-1)\ldots(2m-p+1)} \leq 2^{-p} \leq 2^{-\frac{\epsilon m}{2}} \end{aligned}$$

There are no more than $g(d, 2m)$ distinct events E_r. Applying the union bound, we see that the probability that at least one of the occurs for a fixed $N \cup T$ is at most $g(d, 2m) \cdot 2^{-\epsilon m/2}$. But this holds for any choice of $N \cup T$, so

$$\Pr(E_2) \geq g(d, 2m) \cdot 2^{-\epsilon m/2}. \tag{2}$$

Together, Facts 1 and 2 imply that

$$\Pr(E_1) \leq 2g(d, 2m) \cdot 2^{-\epsilon m/2}.$$

It remains to be proven that the RHS of the above expression is a lower bound on δ. The hard case is $d \geq 2$. The definition of $g(d, 2m)$ ensures that $g(d, 2m) \leq (2m)^d$, so it's enough to prove that

$$2 \cdot (2m)^d 2^{-\epsilon m/2} < \delta. \tag{3}$$

Rearranging this expression and taking the (base-two) logarithm of both sides,

$$\frac{\epsilon m}{2} \geq d \log(2m) + \log \frac{2}{\delta} \Longrightarrow \frac{\epsilon m}{4} + \frac{\epsilon m}{4} \geq d \log(2m) + \log \frac{2}{\delta}$$

The constraints on m defined in the statement of the theorem take care of the second "half" of this inequality:

$$m \geq \frac{4}{\epsilon} \log \frac{2}{\delta} \Longrightarrow \frac{\epsilon m}{4} \geq \frac{2}{\delta}.$$

It remains to be proven that

$$\frac{\epsilon m}{4} \geq d \log(2m).$$

This follows from a simple calculation (omitted).

□

1.1 VC dimension of intersections of range spaces

Let (X, R) be a range space of VC dimension d. Consider the range space $(X, R_{\cap h})$ where $R_{\cap h}$ consists of h-wise intersections of sets in R. We claim that the VC-dimension is at most $2dh \log(dh)$.

For, suppose A is some subset of R that is shattered in $(X, R_{\cap h})$, and let $k = |A|$. Then $|P_{R_{\cap h}}(A)| = 2^k$, but we know that it is at most $\binom{m}{h}$ where $m = |P_R(A)| \leq g(k, d)$. Hence

$$2^k \leq \binom{g(k,d)}{h} \leq k^{(d+1)h},$$

whence it follows that $k \leq 2dh \log(dh)$.

Exercise: Generalize to range spaces in which the subsets are boolean combinations of the sets of R.

2 Applications to Learning Theory

VC-dimension has a close connection to Valiant's PAC (Probably Approximately Correct) model, introduced in 1984. Specifically, VC-dimension gives conditions under which one can learn a concept in *principle* (using possibly a very tedious nonpolynomial time computation). Valiant's notion requires learning algorithms be polynomial. Nevertheless, in many cases one can obtain a polynomial-time algorithm using VC-dimension arguments, as in the following simple example.

Suppose we are trying to learn an unknown rectangle C contained in $[0;1] \times [0;1]$. Let μ be an unknown measure on $[0;1] \times [0;1]$ — one can think of if as concetrated on grid points of a fine grid. Somebody picks in samples according to μ and labels them $+$ or $-$ depending on whether C contains them or not. Our task is to guess what C is.

Our algorithm is the obvious one: find some rectangle C' that separates $+$ from $-$ (note: one exists, namely, C) and output that as your guess. Intuitively this makes sense, but how well does it approximate the actual rectangle C we are looking for?

Let's look at the symmetric difference between C and C', defined as follows.

$$C \Delta C' = (C \setminus C') \cup (C' \setminus C).$$

(Observation: the range space defined by the difference of rectangles has constant VC-dimension, as shown by Section 1.1.) This is the set of all points that are wrongly classified (points that belong to C and not to C' or vice-versa). Suppose that the probability of error is high, i.e., that $Pr_\mu[C \Delta C']$ is large. With high probability the sample is such that for every C_1, C_2 such that $\mu(C_1 \Delta C_2)$ is large, $C_1 \Delta C_2$ contains a sample point. Therefore, with high probability (over the choice of the sample), C' is such that $\mu(C' \Delta C)$ is small.

3 Geometric Algorithms

VC dimension arguments are invaluable for designing (randomized) geometric algorithms. We now consider a problem that has attracted great interest recently: nearest neighbor searching in $\mathbf{R}^d$, where d is large. Given n points in $\mathbf{R}^d$, our goal is to build a data structure using which one can, given any $y \in \mathbf{R}^d$, return the closest neighbor in the set. (A number of norms can be used to define "closeness": ℓ_1, ℓ_2, and ℓ_∞ are among those commonly used.)

A typical application is as follows. Assume you have a class of objects (images, for example, each of which is a vector of pixel values). Each object can be modeled as a point in $\mathbf{R}^d$. This would allow you to make similarity queries: given a new point (image), find the existing point that is closest (the image in the database that is similar to the new one, for example).Of course, this assumes that the modeling of objects as vectors is such that similar objects map to nearby points, which happens to be true in many applications.

As a warmup, let's consider the 2-dimensional case. This is a well-solved problem. First, we build the *Voronoi diagram* of the original points: this diagram partitions the space into separate regions, one for each point in the original set. The region associated with point p represents the set of points that are closer to q than to any other point in the space (the way ties are handled depends on the application). This diagram can be built in $O(n \log n)$ time.

The original nearest neighbor searching problem then reduces to point-location in the Voronoi diagram. One way to do that is to divide the space into *vertical slabs* separated by vertical lines — one passing through each vertex of the Voronoi diagram. There are at most $O(n)$ slabs and each is divided into up to $O(n)$ subregions (from the Voronoi diagram). These slabs can be maintained in a straightforward way in a structure of size $O(n^2)$ (more sophisticated strucutures can reduce this significantly). Queries can be thought of as two binary searches (one to find the correct slab and another within the slab), and require $O(\log n)$ time.

The notion of Voronoi diagrams can be generalized in a straighforward way to arbitrary dimensions. For every two points p and q, the points that are equidistant from both define a set H_{pq} given by:

$$H_{pq} = \{x : d(x,p) = d(x,q)\}.$$

Considering a d-dimensional space with the ℓ_2 norm, we get that

$$\begin{aligned} d(x,p) &= d(x,q) \\ \sum_{i=1}^{d}(x_i - p_i)^2 &= \sum_{i=1}^{d}(x_i - q_i)^2 \\ 2\sum_{i=1}^{d} x_i(p_i - q_i) &= d(q_i^2 - p_i^2). \end{aligned}$$

Since p_i, q_i, and d are constants, H_{pq} must be a hyperplane.

The number of hyperplanes is $m = \binom{n}{2}$, quadratic in the number of points. However, the number of cells in the Voronoi diagram could be $n^{O(d)}$. How can we store this diagram so as to quickly answer nearest neighbor queries? Here we describe Meiser's algorithm (1986), which does so in $O(f(d)\log n)$ time, where $f(d)$ is some exponential function of d.

First, take a random sample S cointaining $O(d^2 \log^2 d)$ points. Then, construct a data structure for doing point location within S. Triangulate each cell of the arrangement of S. (In a plane, triangulation means dividing the cell into triangles; in higher dimensions one gets simplices. Recall that a simplex in $\Re^d$ is an intersection of $d+1$ halfspaces.) For each cell C in this triangulated arrangement, recursively construct a data structure for point location with respect to the hyperplanes that intersect C.

CLAIM: With high probability, S is such that for every cell C, the number of hyperplanes that intersect C is at most $m/2$.

Assuming the claim is true, it is easy to see that $Q(m)$, the query time for a structure of size m, is

$$Q(m) \leq f(d^2 \log^2 d) + Q(m/2) \approx f(d^2 \log^2 d) \log m,$$

where f is the time required to perform point location within S. Moreover, the total space $P(m)$ necessary to process instances with m hyperplanes is

$$P(m) \leq g(d^2 \log^d) \cdot S(m/2) \approx [g(d^2 \log^2 d)]^{\log m},$$

where g determines the size of the structure for point location within the sample. We don't compute f, g because that is not the point of this simplified description. (Using some care, the running time is $d^{O(d)} \log n$ and storage is $n^{O(d)}$.) We now proceed to prove the claim:

PROOF: Consider the range space (X, R), where X is the set of all m hyperplanes of the Voronoi diagram and R contains for each simplex D a subset $R_D \subseteq X$ containing all hyperplanes that intersect the interior of D. From Section 1.1 we know that the VC-dimension of this range space is $O(d^2 \log d)$. Then with high probability the sample must be such that for every simplex D that is intersected by at least $m/2$ hyperplanes is intersected by at least one hyperplane in the sample. This means that D cannot be a cell in the arrangement of the sample. $\square$