CS 683: Advanced Algorithms	VC-Dimension and Sampling
April 5, 2001	Scribes: Sharad Goel and David Kempe

1 The VC-Dimension

We have seen in previous lectures how to use sampling to estimate the number of perfect and near-perfect matchings on our way to computing the permanent of a matrix. It was an important issue to show that reasonably small sample sizes are sufficient to keep the error probability in our estimates small. We used specific properties of the sample space in the process.

We will now encounter a way of arguing about sufficient sample sizes from a purely combinatorial point of view, while treating the actual sampling method as a black box. We will exploit properties of the set of events that we are interested in, and obtain much stronger bounds than we could obtain by mere application of the Union Bound.

Throughout, U will be a (continuous or discrete) set of points, with an implicit collection of events and a probability function $\Pr[X]$ defined on sets $X \subseteq U$.

Consider the following most basic sampling problem: We know that the set X has probability mass $\Pr[X] \geq \varepsilon$, and want to sample enough points from U such that with probability at least $1-\delta$, our sample intersects the set X. The following Theorem gives an upper bound on the required sample size.

Theorem 1.1 *A set of $\frac{1}{\varepsilon}\ln(\frac{1}{\delta})$ points sampled independently and uniformly at random from U intersects X with probability at least $1-\delta$.*

Proof. Any single sample is in X with probability at least ε, and not in X with probability at most $1-\varepsilon$. Therefore, the probability that none of the samples is in X is at most

$$(1-\varepsilon)^{\frac{1}{\varepsilon}\ln(\frac{1}{\delta})} \quad \leq \quad e^{-\ln(\frac{1}{\delta})} \quad = \quad \delta.$$

■

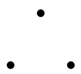
This Theorem states that for any given large set (large is used in the sense of having large probability mass), a random sample hits it with high probability. However, it does not guarantee that a single random sample has high probability of hitting *every* large set. Such a statement would correspond to exchanging the quantifiers in the Theorem — instead of $\forall\,\exists$, we would obtain $\exists\,\forall$. We are interested in characterizing when such an exchange of quantifiers is actually possible, i.e. when one random sample has high probability of hitting every large set.

Of course, this will depend on the sets of high probability that we consider. Let $(U,\mathcal{R})$ be a set system, where $\mathcal{R}$ is a collection of events $A \subseteq U$.

Definition 1.2 *We say that $X \subseteq U$ is* shattered *by $\mathcal{R}$ if for any $Y \subseteq X$, there exists a set $A \in \mathcal{R}$ with $X \cap A = Y$.*

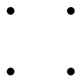
Example 1.1 Let $U = \mathbb{R}^2$ be the plane, and $\mathcal{R}$ the set of all closed half-spaces of the plane.

1. The following set

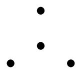
   is shattered by $\mathcal{R}$, since we can "cut off" any of its 8 subsets with a half-space.

2. The set

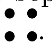
   

   is not shattered by $\mathcal{R}$. Two diagonally opposed points cannot be separated from the other two points by a half-space.

3. The set

   is not shattered by $\mathcal{R}$, because the point in the center cannot be separated from the other three points by a half-space.

We just saw that the set  is not shattered by $\mathcal{R}$. If instead of half-spaces, we had — say — ellipses available to separate out point sets, we could separate any subset of the four points, and hence shatter . In that sense, the set system of all ellipses in the plane seems to have "higher complexity" than the set system of all half-spaces. This notion is captured by the following definition.

Definition 1.3 *The* VC-dimension *of a set system $(U, \mathcal{R})$ is the maximum size of any finite set X shattered by $\mathcal{R}$. If arbitrarily large sets X are shattered by $\mathcal{R}$, then the VC-dimension is ∞.*

Example 1.2

1. For the set system $(U, \mathcal{R})$ from Example 1, we know that the VC-dimension is at least 3, because the set  is shattered. However, no set X of size 4 or more is shattered. If three points of X are collinear, then the middle one of those three points can not be singled out. Otherwise, there must be two points $x, y \in X$ such that the other two points lie on different sides of the line through x and y. (This can be seen by fixing the point x and looking at the lines to the three other points. Since no three points are collinear, one of the lines must lie between the other two.) The set $\{x, y\}$ cannot be separated by the set system $\mathcal{R}$.

2. If we consider half spaces on the line (i.e. intervals bounded on one side), then the VC-dimension is 2.

3. In general, the set $\mathbb{R}^d$ with $\mathcal{R}$ consisting of all half spaces in $\mathbb{R}^d$ has VC-dimension $d+1$. The VC-dimension is at least $d+1$, because a simplex is shattered by $\mathcal{R}$. To show that the VC-dimension is at most $d+1$, we can use Radon's Theorem (see Lecture 13, Theorem 2.4). It states that any set $X \subseteq \mathbb{R}^d$ can be partitioned into sets Y and Z whose convex hulls intersect. Any half-space containing Y must therefore meet the convex hull of Z, and thus contain a point of Z, i.e. the set $Y \subseteq X$ cannot be separated by a half-space.

Given $(U, \mathcal{R})$, we can build a "smaller" set system from any $X \subseteq U$, namely $(X, \mathcal{R}|X)$, where $\mathcal{R}|X = \{A \cap X \mid A \in \mathcal{R}\}$. Then, the VC-dimension of $(X, \mathcal{R}|X)$ is at most that of $(U, \mathcal{R})$, and equality holds if X has a shattered subset of size VC-dimension$(U, \mathcal{R})$.

2 Consequences of VC-dimension for Sampling

Definition 2.1 *$N \subseteq U$ is an ε-net for $(U, \mathcal{R})$ if for all $A \in \mathcal{R}$ with $Pr[A] \geq \varepsilon$, we have $N \cap A \neq \emptyset$. That is, N meets* every *large set in $\mathcal{R}$.*

We have the following theorem (which will be proved later in class):

Theorem 2.2 *Let d be the VC-dimension of $(U, \mathcal{R})$, and $\delta > 0$. If we draw $O(\frac{d}{\varepsilon}\log(\frac{d}{\varepsilon}) + \frac{1}{\varepsilon}\log(\frac{1}{\delta}))$ points at random from U, the resulting set N is an ε-net with probability at least $1 - \delta$.*

Example 2.1 Suppose U is a set of n points in the plane, and $\mathcal{R}$ is the set of all intersections of U with some half-space. We have already seen that the VC-dim of $(U, \mathcal{R})$ is 3, so the above theorem tells us that we can with high probability generate an ε-net for $(U, \mathcal{R})$ by randomly selecting a relatively small number of points.

Example 2.2 Suppose U is again a set of n points in the plane, but $\mathcal{R}$ is now the set of all subsets of U. The VC-dimension of $(U, \mathcal{R})$ is n, and the smallest ε-net has size $(1-\varepsilon)n + 1$.

Definition 2.3 *A measurable set $S \subseteq U$ is an ε-sample for $(U, \mathcal{R})$ if $|Pr[A] - \frac{Pr[S \cap A]}{Pr[S]}| \leq \varepsilon$, for all $A \in \mathcal{R}$.*

Any ε-sample is also an ε-net, but the converse does not necessarily hold. Next time, we will prove the following theorem:

Theorem 2.4 *A random set of $O(\frac{d}{\varepsilon^2}\log(\frac{d}{\varepsilon}) + \frac{1}{\varepsilon^2}\log(\frac{1}{\delta}))$ points from U is an ε-sample with probability at least $1 - \delta$.*

3 Applications

Example 3.1 Suppose we want to pre-process a polytope $P \subset \mathbb{R}^d$ so that for any $c \in \mathbb{R}^d$ we can efficiently approximate the validity of statements of the form: Is $c \cdot x \leq \gamma$, for all $x \in P$? By approximate validity, we mean that our algorithm makes the following guarantees:

- If the answer is "No", then the algorithm provides a witness $x \in P$, satisfying $c \cdot x > \gamma$.
- If the algorithm answers "Yes", then with high probability, $c \cdot x \leq \gamma$ for at least a $1 - \varepsilon$ fraction of the points.

Note that we are trying to pre-process P so that we can determine if it is on a given side of a half-space.

Consider the set-system $\mathcal{A} = (P, 1/2 - \text{spaces})$. Then the VC-dimension of $\mathcal{A}$ is $d+1$, so by our sampling theorem, we can generate an ε-net N with high probability consisting of $O(\frac{d}{\varepsilon}\log(\frac{d}{\varepsilon}) + \frac{1}{\varepsilon}\log(\frac{1}{\delta}))$ points. Now, we check the condition $c \cdot x \leq \gamma$ for each $x \in N$. If it fails for some x, then we have our witness. Otherwise, let $\mathcal{F} = \{x \in P \mid c \cdot x > \gamma\}$ i.e. the set where our condition fails to hold. Since N is with high probability an ε-net, we can conclude that with high probability $\Pr[\mathcal{F}] < \varepsilon$. That is, with high probability the statement is true for at least a $1 - \varepsilon$ fraction of P.

The important observation here is that we don't have to take a new sample every time, which is particularly important if sampling is difficult and expensive.

Definition 3.1 *If P is a set of n points in $\mathbb{R}^d$, then an ε-approximate centerpoint is a point $x \in \mathbb{R}^d$ so that any hyperplane through x divides P into two pieces which both have size at least $(\frac{1}{d+1} - \varepsilon)n$.*

Example 3.2 Consider the problem of finding an ε-approximate centerpoint for a set P of n points. By our sampling theorem, we can with high probability generate an ε-sample S by randomly drawing $O(\frac{d^2}{\varepsilon^2}\log(\frac{d}{\varepsilon}) + \frac{1}{\varepsilon^2}\log(\frac{1}{\delta}))$ from P. So suppose S is in fact an ε-sample. We can find a center point x of S in constant time (since $|S|$ is independent of $|P|$). Suppose x is not an ε-approximate center point for P. Then there is some half space H through x with

$$\frac{|P \cap H|}{|P|} \leq \frac{1}{d+1} - \varepsilon.$$

But by the centerpoint property,

$$\frac{|S \cap H|}{|S|} \geq \frac{1}{d+1},$$

and by the ε-sample property,

$$\left|\frac{|P \cap H|}{|P|} - \frac{|S \cap H|}{|S|}\right| < \varepsilon.$$

This is a contradiction, so x is an ε-approximate center point for P. This algorithm is an example of how the VC-dimension helps us to keep sample sizes small in geometric algorithms.