PRINCETON UNIV. F'07 COS 597D: THINKING LIKE A THEORIST

Lecture 12: Two Notions of ϵ-Nets and Applications

Lecturer: *Sanjeev Arora* Scribe:*Aditya Bhaskara*

1 Introduction

We first review the concepts of Vapnik-Chervonenkis dimension and ϵ-nets defined in last class. Then we look at an application to detection of failures in networks. Finally we will look at approximate nearest neighbour searching in metric spaces. Here the term ϵ-net is used in a somewhat different sense.

1.1 Review of VC-Dimension

DEFINITION 1 *A* range space *is defined by the pair $(X, \mathcal{R})$, where X is an arbitrary set and $\mathcal{R} \subseteq 2^X$.*

DEFINITION 2 *A set $Y \subseteq X$ is said to be* shattered *by $\mathcal{R}$ if for every subset T of Y, there exists an A in $\mathcal{R}$ such that $T = A \cap X$.*

DEFINITION 3 *The Vapnik Chervonenkis (VC) dimension of a range space $(X, \mathcal{R})$ is defined to be the size of the largest $Y \subseteq X$ that is shattered by $\mathcal{R}$.*

A simple example of a range space we had seen last time was when $X = [0, 1]^2$ and $\mathcal{R}$ is the set of all axis parallel rectangles contained in $[0, 1]^2$. It can easily be seen that the VC-dimension of this range space is 3. We now recall the definition of an ϵ-net.

DEFINITION 4 *Suppose $(X, \mathcal{R})$ is a range space of finite VC dimension d. A set $S \subseteq X$ is called an ϵ-net if for every $A \subseteq X$ with $|A| \geq \epsilon|X|$, we have $S \cap X \neq \phi$.*

We also recall the "sampling theorem" we proved last time.

THEOREM 1
Suppose $(X, \mathcal{R})$ is a range space of finite VC dimension d, and suppose S is a random set of m elements from X, where m satisfies

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

Then with probability at least $1 - \delta$, S is an ϵ-net for $(X, \mathcal{R})$.

The theorem is quite surprising because the bound on m depends only on ϵ and δ, and *not* on $|X|$, as is typically the case when we use any sort of Chernoff or Union bounds.

We will now look at an application of the VC dimension theory to the problem of detecting failures in networks, from a paper of Kleinberg [1].

2 Detecting Failures in Networks

Suppose we have a large network and we want to find out how 'well-connected' it is. More precisely, we want to find out if the network is remains connected when some k of the edges fail. Suppose the way we want to detect this is to have 'detectors' at some of the nodes, and check if all pairs of detectors are still connected. Naturally, the aim is to minimize the number of detectors we need to use while being able to say that the network is connected with high probability.

It is easy to see that if we want something this strong, we would need a detector at every node. For this, just consider a two-connected graph with each vertex of degree k. Now suppose all edges connected to a vertex fail. Clearly, the vertex has to be one of the detectors if we want to conclude the graph is not connected. This is true for all vertices, so we need a detector at each node.

Now say we relax the requirement and not want 'large chunks' of the network to be disconnected from each other. More precisely, we want to be able to find if removal of k edges divides the graph into two pieces each large enough.

Definition 5 *Suppose G is a graph with n vertices. A set Z of at most k edges is said to be an (ϵ, k)-separation if after removing Z, there exist two sets of vertices A, B of size at least ϵn such that no vertex in A is connected to a vertex in B.*

We will prove that for this weaker requirement, it suffices to have detectors at only a constant (i.e., depending on ϵ, k and not the size of the graph) number of nodes. In particular, we show

Theorem 2
A random set of $O\left(\frac{k}{\epsilon}\log\frac{k}{\epsilon}\right)$ nodes is with probability at least $\frac{1}{2}$ a 'detector set' for all *(ϵ, k)-separations of the underlying graph G.*

To prove this, we will bound the VC dimension of an appropriately defined range space and then appeal to the sampling theorem (Thm. 1).

Denote the vertex set of the graph by V. Define a range space $(V, \mathcal{R})$ with $R \in \mathcal{R}$ iff R is a union of connected components of $G \setminus Z$, for a set Z of at most k edges. This choice of $\mathcal{R}$ is critical. We will show that the VC dimension of this range space is at most $2k+1$, in particular that no set of size $2k+2$ can be shattered. We will need the following simple lemma.

Lemma 3
Suppose S is a set of $2l$ vertices in a connected graph $G(V, E)$. Then there exist l pairwise edge-disjoint paths $P_1, P_2, \ldots, P_l$ (in G) such that each vertex in S is an end-point of precisely one of the P_i.

Proof: Consider l paths $P_1, P_2, \ldots, P_l$ in G such that each vertex in S is an endpoint of precisely one of the P_i, and such that sum of lengths of the paths P_i is minimized. We claim two paths P_i and P_j can have at most one vertex in common. This is because if say P_i (end points i_1, i_2) and P_j (end points j_1, j_2) have two common vertices, we could find paths from i_1 to j_1 and i_2 to j_2 with total length less than sum of lengths of P_i and P_j. This contradicts the minimality of the sum of lengths of P_i. Thus P_i's are pairwise edge-disjoint. $\square$

It is now rather easy to see that no subset of V of size $2k+2$ can be shattered. Consider an $S \subseteq V$ of size $2k+2$, and find paths $P_1, P_2, \ldots, P_{k+1}$ as guaranteed by Lemma 3. Suppose the endpoints of P_i are s_{2i-1}, s_{2i}. We claim that there is no $R \in \mathcal{R}$ such that $R \cap V$ is $\{s_1, s_3, \ldots, s_{2k+1}\}$. R is in $\mathcal{R}$ implies there is a set Z of at most k edges such that R is the union of some of the connected components of $G \setminus Z$. Now suppose s_{2i-1} is in R and s_{2i} is not (for all i). Then at least one edge of P_i must be in Z. Since P_i are edge disjoint, it follows that $|Z| \geq k+1$, a contradiction.

This completes the proof of Theorem 2. We will now look at metric spaces and approximate nearest neighbor searching, where a different notion of ϵ-net is used.

3 Nearest Neighbors in Metric Spaces

We start be recalling the definition of a metric space.

Definition 6 *A* metric space *is defined by a pair (X, d), where X is a set of* points *and $d : X \times X \to \mathbb{R}^+ \cup \{0\}$ is a* distance function *that satisfies the following properties.*

1. *$d(x, y) = d(y, x)$ for all $x, y \in X$*

2. *$d(x, y) = 0$ iff $x = y$*

3. *$d(x, y) + d(y, z) \geq d(x, z)$ for all $x, y, z \in X$.*

Some typical examples of metric spaces are $\mathbb{R}^n$ with ℓ_p norm, shortest paths in graphs, and so on.

The problem of Nearest neighbor searching in metric spaces is a central one, with lots of applications. Formally the problem is the following: we are given a set of points $A \subseteq X$, and a query point q. The goal is to find a point $a \in A$ minimizing $d(a, q)$. The aim of course, is to do this in time much smaller than $|A|$, preferably polylog($|A|$).

Later in this section we will see a 3-factor approximation algorithm due to [2] for this problem. We will now define some of the terms needed to specify the algorithm and analyze it.

3.1 ϵ-Nets in Metric Spaces and Dimension

Definition 7 *Given a metric space (X, d), a subset Y of X is said to be an ϵ-net if*

1. *For $a, b \in Y$, we have $d(a, b) \geq \epsilon$.*

2. *For all $x \in X$, there exists an $a \in Y$ such that $d(x, a) < \epsilon$.*

For every finite metric space X and every $\epsilon > 0$, it is clear that there exists an ϵ-net. It can easily be found by the following greedy algorithm: Start with a single point in the net N. While there exists a point $x \in X$ such that $\min_{a \in N} d(a, x) > \epsilon$, add x to N, and repeat. Clearly since X is finite this terminates, and we are left with an ϵ-net at the end. Using Zorn's lemma one can argue this even for infinite metric spaces, but we will not go into the proof.

Another 'natural' parameter of metric spaces is the *dimension.* For say $\mathbb{R}^d$ with any ℓ_p norm, it is *clear* that the dimension is d, but how can one define it for arbitrary metric spaces?

A first attempt would be to define the dimension using the intuition that $|B(x,r)| \approx r^d$, where $B(x,r)$ denotes the ball of radius r around the point x (i.e., the set of points at a distance $< r$ from x). A trouble with this is that the choice of r could critically affect the value of the dimension. A notion that is more commonly used is the following.

Definition 8 *(Doubling Dimension) The* doubling dimension *of a metric space* (X,d) *is defined as the smallest* k *such that every subset* $S \subseteq X$ *can be covered by* 2^k *sets of diameter at most half the diameter of* S*.*

It can be shown that replacing the phrase 'sets of diameter' everywhere in the definition above by 'balls of radius' would change the value by at most a factor of 2.[1] Also, it is clear that for the case of $\mathbb{R}^d$, the doubling dimension is $\Theta(d)$.

Consider constant degree expanders, say d-uniform $(\alpha, d/2)$-vertex expanders for some constant α. Given such an expander on n vertices, we can show that the doubling dimension of the metric space with the shortest path metric is $\Omega(\log n)$. Another important notion in metric spaces is that of the *aspect ratio*, which we define now.

Definition 9 *Given a metric space* (X,d)*, the aspect ratio* Δ *is defined by*

$$\Delta = \frac{\mathrm{Diam}(X)}{\min_{x,y\in X} d(x,y)}$$

We now give a bound on the size of a metric space in terms of the aspect ratio and the doubling dimension.

Lemma 4
Suppose (X,d) *is a metric space with aspect ratio* Δ *and doubling dimension* m*. Then* $|X| \le \Delta^{O(m)}$.

Proof: We may assume w.l.o.g. that $\min_{x,y\in X} d(x,y) = 1$ (we can rescale d appropriately), so the diameter is Δ. Thus X can be covered by at most 2^m sets of diameter $\Delta/2$. Repeating this $\log \Delta$ times, we get X covered by at most $2^{m \log \Delta}$ sets of diameter < 1, and each such set has at most one point. This gives the desired bound. □

Note that this immediately implies that if the diameter of a metric space is D and doubling dimension is m, the size of an ϵ-net is $\left(\frac{D}{\epsilon}\right)^{O(m)}$ (which exactly corresponds to the natural grid in $\mathbb{R}^d$). We now present the algorithm of [2].

3.2 A 3-factor Approximation to Nearest-Neighbor Search

We will assume the underlying metric space is (X,d), the set of points is S and the query point is q. The algorithm will return a $s \in S$ such that $d(q,s) \le 3 \min_{t\in S} d(q,t)$. We also assume that $\min_{x,y\in S} d(x,y) = 1$, and $\mathrm{Diam}(S) = \Delta = 2^k$ (say).

[1] There is a technical issue here. If one needs $B(x,r) \le 2^d B(x,r/2)$ for all x, r, then this is no longer true. All we say is that there are 2^d balls (different centers) that cover $B(x,r)$. See [2] for more on comparision between notions of dimension.

The algorithm maintains a data structure quite similar to quad-trees that are used in orthogonal range searching. But one crucial difference is that a point not only maintains children in it's 'cell', but also some from neighboring cells.

More formally, we maintain ϵ-nets of different 'scales'. Let Y_i be a 2^i-net for S, for $1 \leq i \leq k$. Further, for $y \in Y_i$, we maintain

$$L_{y,i} = \{z \in Y_{i-1} \ : \ d(y,z) \leq 7 \times 2^i\}$$

The 7 is for the analysis to work out. The algorithm, given a query point q is the following.

1. Set y to the unique point in Y_k, and $i = k$.
2. Find the point in $L_{y,i}$ that is closest to q (say this is z).
3. If $d(q,z) > 3 \times 2^{i-1}$, return the current y; else set $y = z$, decrement i and continue with Step (2).

Theorem 5
The algorithm above gives a 3-factor approximation to the Nearest-neighbor problem.

Proof: We now proceed with the analysis. Suppose the algorithm returns y and i is the value of the variable at that point. This means we have $d(q,y) \leq 3 \times 2^i$, but $d(q, L_{y,i}) > 3 \times 2^{i-1}$. Suppose a is the true optimum, i.e., $d(q,a) = \min_{s \in S} d(q,s)$. Consider the point p in the 2^{i-1}-net Y_{i-1} that is closest to a. We first argue that $p \in L_{y,i}$.

Clearly we have $d(a,p) \leq 2^{i-1}$. Thus $d(p,y) \leq d(p,a) + d(a,q) + d(q,y) \leq d(p,a) + 2d(q,y)$, by definition of a. Combining with $d(q,y) \leq 3 \times 2^i$, we get $d(p,y) < 7 \times 2^i$, thus $p \in L_{y,i}$.

It is now easy to complete the argument. By the above and the fact that $d(q, L_{y,i}) > 3 \times 2^{i-1}$, it follows that $d(q,a) \geq d(q,p) - d(p,a) \geq 3 \times 2^{i-1} - 2^{i-1} \geq 2^i$. This shows that the algorithm gives a 3-approximation. □

Observe that assuming the data structures (the Y_i and $L_{y,i}$) are constructed, the query time is just proportional to k, i.e., $\log \Delta$. This is because the $L_{y,i}$ have a size dependent just on the doubling dimension, which we assumed is a constant (however the dependance is exponential in the dimension, which is a problem hard to avoid). The paper of [2] proceeds to give a $(1+\epsilon)$-factor approximation, but we do not go into this.

References

[1] J. Kleinberg. *Detecting a Network Failure.* Proceedings of 41st IEEE Symposium on Foundations of Computer Science, 2000.

[2] R. Krauthgamer, J. R. Lee. *Navigating nets: Simple algorithms for proximity search.* Proceedings of ACM Symposium on Discrete Algorithms, 2004.