Algorithmic Excursions: Topics in Computer Science II **Spring 2016**

Lecture 1 & 2 : ϵ-net and VC-dimension

Lecturer: Kasturi Varadarajan *Scribe: Sayan Bandyapadhyay*

1.1 Sampling to Preserve Geometric Information

Sampling is the process of choosing a "small" number of observations (or sample) from a population. In many applications it is expensive to study all the observations of a population and thus a "small" subset is chosen to study. A very good application is election survey, where the poll of a subset of voters are taken to predict the outcome of the election. In this section we consider a special type of sampling that preserves some property. We will use the following problem to describe this concept.

1.1.1 A Motivating Problem

We are given a set Y of points in the plane. We consider a disk D that is not known. Points in $Y \cap D$ (resp. $Y \setminus D$) are labelled $+$ (resp. $-$). The labels are not known, but can be computed. We assume that the computation of the label of a point is an expensive process. Now the goal is to compute the disk D. The interesting thing is here that we do not know how to compute D without checking the labels of all points. So, we use the following sampling technique to find a disk that "approximates" D.

1. Pick a sample $N \subseteq Y$
2. Compute label for each point in N
3. Return the smallest radius disk D_1 containing all the $+$ points in N and none of the $-$ points in N

Now it is not hard to see that some $+$ points in D might not lie inside D_1 or some $-$ points in D might lie inside D_1. To quantify the error consider the symmetric difference $D\Delta D_1 = (D \setminus D_1) \cup (D_1 \setminus D)$. Note that $Y \cap (D\Delta D_1) \subseteq Y \setminus N$, as $(D \setminus D_1) \subseteq D$ contains only $+$ points of $Y \setminus N$ and $(D_1 \setminus D)$ contains only a subset of $-$ points of Y that are not in N (see Figure 1.1). Also $D \cap D_1$ contains only $+$ points. Thus the erroneous points are the points in $D\Delta D_1$. Hence we would like to minimize the quantity $|Y \cap (D\Delta D_1)|$. In particular, for any $0 < \epsilon \leq 1$, we want $|Y \cap (D\Delta D_1)| \leq \epsilon|Y|$. Now keeping this problem in mind it is a good time to define the concept of ϵ-net which will be helpful to solve the problem.

Definition 1.1 *A subset $M \subseteq Y$ is an ϵ-net w.r.t. Δ if for any disk D' in the plane, $|Y \cap (D\Delta D')| > \epsilon|Y| \implies M \cap (D\Delta D') \neq \phi$.*

Now let us go back to our sampling algorithm where we choose the sample set N. Suppose N is an ϵ-net w.r.t. Δ, then our claim is that $|Y \cap (D\Delta D_1)| \leq \epsilon|Y|$. Suppose $|Y \cap (D\Delta D_1)| > \epsilon|Y|$. Then as $Y \cap (D\Delta D_1) \subseteq Y \setminus N$, N does not contain any point of $Y \cap (D\Delta D_1)$. But by definition of an ϵ-net w.r.t. Δ this cannot be true. Thus to solve our problem (to approximate the disk D) it is sufficient to compute an ϵ-net w.r.t. Δ. Later in this course we will see how to compute such an ϵ-net of "small" (independent of $|Y|$) size.

Our sampling technique is an example of sampling that preserves geometric information. In particular, the geometric information that we want to preserve is that for any disk D', either $|Y \cap (D\Delta D')| \leq \epsilon|Y|$ or the sample set N contains at least one point of $Y \cap (D\Delta D')$.

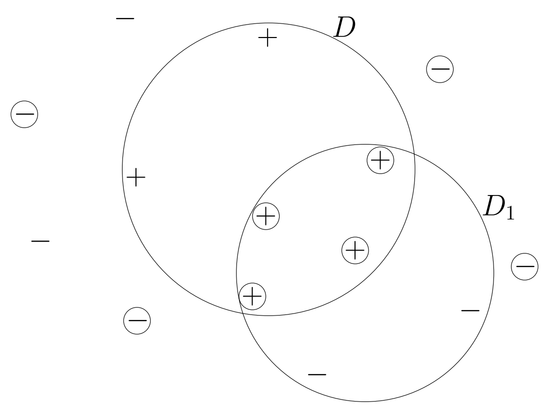

Figure 1.1: Circled $+$ and $-$ are the points of N.

1.2 VC-dimension

Definition 1.2 *A set system (or a range space) $\mathcal{S}$ is a pair $(X, \mathcal{R})$, where X is a finite or infinite ground set, and $\mathcal{R}$ is a finite or infinite family of subsets of X. Each element of $\mathcal{R}$ is called a range.*

An example of a set system is $(X_1, \mathcal{R}_1)$, where X_1 is the real line and each element of $\mathcal{R}_1$ is an interval. Another example could be the pair $(X_2, \mathcal{R}_2)$, where X_2 is the plane and each element of $\mathcal{R}_2$ is the symmetric difference of two disks.

Now consider a range space $\mathcal{S} = (X, \mathcal{R})$. Given $Y \subseteq X$, $\mathcal{R}_Y$, the projection of $\mathcal{R}$ onto Y is $\{Y \cap r | r \in \mathcal{R}\}$. Projection of $\mathcal{S}$ onto Y is $(Y, \mathcal{R}_Y)$. For example, again consider the set system $(X_1, \mathcal{R}_1)$. Let $Y = \{a, b, c\}$ such that $a < b < c$. Then $R_Y = \{\phi, \{a\}, \{b\}, \{c\}, \{a, b\}, \{b, c\}, \{a, b, c\}\}$. R_Y does not contain $\{a, c\}$, as any interval that contains a and c must also contains b. For a range space $(X, \mathcal{R})$, a subset $Y \subseteq X$ is said to be completely shattered if R_Y is the collection of all subsets of Y. For the set system $(X_1, \mathcal{R}_1)$, any two point subset is completely shattered.

Definition 1.3 *Vapnik-Chervonenkis dimension (VC-dimension) of a set system $\mathcal{S} = (X, \mathcal{R})$ is the largest integer m for which there is a set $Y \subseteq X$ of size m that is completely shattered. If such a largest integer does not exist, VC-dimension is ∞.*

For the set system $(X_1, \mathcal{R}_1)$, it is not possible to completely shatter any three points subset and hence from our previous discussion the VC-dimension is 2. For the range system with the plane as the ground set and halfplanes as the ranges, one can show that the VC-dimension is 3. Also for the range system with the plane as the ground set and convex sets as the ranges, VC-dimension is ∞. For any m, one can select a set of m points in convex positions that is completely shattered.

Definition 1.4 *Given a range space $\mathcal{S} = (X, \mathcal{R})$ its shatter function $\pi_{\mathcal{S}} : \mathbb{N} \to \mathbb{N}$ is defined as*

$$\pi_{\mathcal{S}}(m) = \max_{B \subset X : |B| = m} |\mathcal{R}_B|$$

For our example set system $\mathcal{S}' = (X_1, \mathcal{R}_1)$, $\pi_{\mathcal{S}'}(0) = 1$, $\pi_{\mathcal{S}'}(1) = 2$, $\pi_{\mathcal{S}'}(2) = 4$, and $\pi_{\mathcal{S}'}(3) = 7$. One intereseting question in this context is, “Is shatter function of a set system polynomially bounded?”. For example, $\pi_{\mathcal{S}'}(m) = O(m^2)$. Indeed, for any finite set of points, a subset that can be generated by an interval is uniquely identified by the maximum and the minimum point of that subset. Thus for a set of m points $O(m^2)$ distinct subsets can be generated. In general, the following lemma gives a bound on the shatter function.

Lemma 1.5 *Suppose a set system $\mathcal{S} = (X, \mathcal{R})$ has VC-dimension $d < \infty$. Then*

$$\pi_{\mathcal{S}}(m) \leq \binom{m}{0} + \binom{m}{1} + \ldots + \binom{m}{d}$$

1.3 ϵ-net

Previously, we have seen the definition of ϵ-net w.r.t. Δ operator. In this section, we generalize that definition for any finite range space, i.e range space with finite ground set.

Definition 1.6 *Let $\mathcal{S} = (X, \mathcal{R})$ be a finite range space. For $0 < \epsilon < 1$, $N \subseteq X$ is said to be an ϵ-net if for any $r \in \mathcal{R}$ such that $|r| > \epsilon|X|$, $N \cap r \neq \phi$.*

Consider a range space in the real line with 16 points as the ground set X, and each range is the intersection of X and an interval. Let $\epsilon = \frac{1}{4}$. Now it is easy to see that if we take every fourth point from a sorted ordering of the points in X w.r.t. their values, we get an ϵ-net. In general, we need to pick every $\epsilon|X|^{th}$ point. Hence the size of the ϵ-net would be $O(\frac{1}{\epsilon})$. For general range spaces it is not straightforward if one can get an ϵ-net of size $O(\frac{1}{\epsilon})$. In the following lemma we prove a weaker bound for general range spaces.

Lemma 1.7 *Let $\mathcal{S} = (X, \mathcal{R})$ be a finite range space and $0 < \epsilon < 1$. Then $\mathcal{S}$ has an ϵ-net of size $O(\frac{1}{\epsilon} \ln |\mathcal{R}|)$.*

Proof: We give a probabilistic proof for this lemma. Let $N \subseteq X$ be chosen by sampling uniformly from X, $\frac{c}{\epsilon} \ln |\mathcal{R}|$ points, independently and with replacement, where $c > 0$ is a suitable constant. Note that it is sufficient to show that N is an ϵ-net with probaility > 0. Indeed, if there is no ϵ-net of size $O(\frac{1}{\epsilon} \ln |\mathcal{R}|)$, the probability that N is an ϵ-net is 0.

For $r \in \mathcal{R}$, let B_r be the event $r \cap N = \phi$. Now consider any r such that $|r| > \epsilon|X|$. Then the probability that a particular point in N does not belong to r is at most $1 - \frac{\epsilon|X|}{|X|} = 1 - \epsilon$. As all the $\frac{c}{\epsilon} \ln |\mathcal{R}|$ points in N are chosen independent of each other,

$$\begin{aligned} Pr[B_r] &\leq (1-\epsilon)^{\frac{c}{\epsilon} \ln |\mathcal{R}|} && (1.1) \\ &\leq e^{-c \ln |\mathcal{R}|} \quad \text{(as } 1 + x \leq e^x\text{)} && (1.2) \\ &= \frac{1}{|\mathcal{R}|^c} && (1.3) \end{aligned}$$

Then the probability that for at least one range r with $|r| > \epsilon|X|$, $r \cap N = \phi$ is,

$$\begin{aligned} Pr[\bigcup_{r \in \mathcal{R}: |r| > \epsilon|X|} B_r] &\leq \sum_{r \in \mathcal{R}: |r| > \epsilon|X|} Pr[B_r] \quad \text{(by union bound)} && (1.4) \\ &\leq \frac{|\mathcal{R}|}{|\mathcal{R}|^c} && (1.5) \\ &= \frac{1}{|\mathcal{R}|^{c-1}} && (1.6) \\ &< 1 \quad \text{(if } c \geq 2\text{)} && (1.7) \end{aligned}$$

Thus the probability that for any range r with $|r| > \epsilon|X|$, $r \cap N \neq \phi$ is > 0. Hence N is an ϵ-net with probaility > 0. ■

One might be interested in improving the bound in Lemma 1.7. Actually, this is possible for the range spaces with finite VC-dimension. In particular, one can show that for a range space $(X, \mathcal{R})$ with finite VC-dimension, there is an ϵ-net whose size is independent of $|\mathcal{R}|$.