6.896 Sublinear Time Algorithms October 28, 2004

Lecture 13: Clustering

Lecturer: Rahul Sami *Scribe: Jonathan Kelner*

1 (k,b)-radius clustering

In this lecture, we shall discuss sublinear time algorithms for two types of clustering problems. The first one that we shall consider is (k,b)-radius clustering.

Throughout these notes, let X be a set of n points in $\mathbb{R}^d$, and let distances be measured in the L_2 metric, so that $\text{dist}(x,y) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2}$.

Definition 1 *We say that X is* (k,b)-radius clusterable *if X can be partitioned into k subsets such that each subset can be covered by a Euclidean ball of radius at most b.*

Following our usual pattern, we shall require a notion of the distance that an instance is from being (k,b)-radius clusterable:

Definition 2 *We say that X is* ϵ-far from being (k,b)-radius clusterable *if at least ϵn of the points of X must be deleted to make it (k,b)-radius clusterable.*

The main goal of the first section of these notes will be to prove the following theorem:

Theorem 3 (Alon, Dar, Parnas, Ron) *There exists an algorithm that uses $\Theta(\log \frac{dk}{\epsilon})$ queries such that:*

1. *If X is (k,b)-clusterable then the algorithm will output "yes."*
2. *If X is ϵ-far from being (k,b)-radius clusterable, then the algorithm will output "no" with probability at least $2/3$.*

While the analysis is slightly complicated, the algorithm itself is very simple:

1. Sample $m = \Theta(\frac{dk}{\epsilon}\log(\frac{dk}{\epsilon}))$ points from X.
2. Run an exact clustering algorithm on the sample. If the sample is (k,b)-radius clusterable, output "yes," otherwise output "no."

The second step of the algorithm requires us to run an exact clustering algorithm on the sample, which is quite slow. That is why we have only asserted that this algorithm has a small query complexity and not that it has a fast running time.

1.1 Nets and Shatter Exponents

Before we can prove the correctness of our algorithm, we shall require a few definitions about collections of sets. Throughout this section, let $S = \{S_i, S_i \subseteq \mathbb{R}^d\}$ be a collection of subsets of $\mathbb{R}^d$, let $X \subseteq \mathbb{R}^d$ be a finite subset of Euclidean space, and let $\epsilon \in (0,1)$ be a real number between 0 and 1.

Definition 4
1. *A subset $N \subseteq X$ is called an* ϵ-net on X with respect to S *if for all S_i such that $|X \cap S_i| \geq \epsilon|X|$, the intersection $N \cap S_i \neq \emptyset$.*
2. *For $A \subseteq \mathbb{R}^d$, let $\phi_S(A) = \{A \cap S_i | S_i \in S\}$.*
3. *For $m \in \mathbb{Z}$, let $\phi_S(m) = \max_{A,|A|=n} |\phi_S(A)|$.*
4. *We define the* shatter exponent of S *to be the smallest integer $SE(S)$ such that, for some fixed constant c, $\phi_S(m) \leq cm^{SE(S)}$ for all m.*

Example 1 *Let $S = \{B | B$ is a ball in $\mathbb{R}^d\}$. Given X of cardinality n in $\mathbb{R}^d$,*

$$\phi_S(X) = \{A \subseteq X | A = X \cap B \text{ for some ball } B\}.$$

For any $A \subseteq X$, let $B^*(A)$ be the minimum radius ball in $\mathbb{R}^d$ that covers A.

Fact 5 *For any such A, there exists a subset $A' \subseteq A$ such that $B^*(A') = B^*(A)$, and $|A'| \leq d+1$.*

This means that each $B^*(A)$ is determined by at most $d+1$ of the points of X, which implies that there cannot be more than $\binom{n}{d+1}$ distinct such sets. This means that $\phi_S(X)$ consists of at most $\binom{n}{d+1} \leq n^{d+1}$ sets, so the shatter exponent of S is less than or equal to $d+1$.

Example 2 *Let $\mathbb{B}_k = \{B_k | B_k$ is the union of k balls in $\mathbb{R}^d\}$. It is easy to see that $|\phi_{\mathbb{B}_k}| \leq (|\phi_S(d)|)^k \leq n^{(d+1)k}$, so the shatter exponent of $\mathbb{B}_k$ is at most $(d+1)k$.*

We shall make use of shatter exponents through the following theorem:

Theorem 6 (Hausler, Welzl) *Let S, X, and ϵ be as above. If you pick a sample U from X of size $\frac{8SE(S)}{\epsilon}\log(\frac{8SE(S)}{\epsilon})$, then U will be an ϵ-net on X with respect to S with probability at least $2/3$.*

1.2 Proof of Correctness of the Algorithm

If X is (k,b)-clusterable, the sample will be (k,b)-clusterable as well, so the algorithm clearly always outputs "yes" whenever the input can be clustered. It thus just remains to show that it rejects with probability at least 2/3 whenever the input is ϵ-far from being (k,b)-clusterable.

To this end, let $\overline{\mathbb{B}_{k,b}} = \{S \,|\, \text{the complement of } S \text{ is the union of } k \text{ balls in } \mathbb{R}^d \text{ of radius} \leq b\}$. It is easy to see that $SE(\overline{\mathbb{B}_{k,b}}) \leq SE(\mathbb{B}_k) \leq n(d+1)$.

If X is ϵ-far from being (k,b)-radius clusterable, then no set of k balls of radius b covers $(1-\epsilon)n$ points of X. This means that for all elements $\overline{B_{k,b}}$ of $\overline{\mathbb{B}_{k,b}}$, we have

$$|X \cap \overline{B_{k,b}}| \geq \epsilon n. \tag{1}$$

By Theorem 6, the sample that we select is an ϵ-net of X with respect to $\overline{\mathbb{B}_{k,b}}$ with probability at least 2/3. Assume that this occurs. If the algorithm accepts, this sample is also (k,b)-clusterable. However, this means that there is some $B_{k,b} \in \mathbb{B}_{k,b}$ containing the sample. For these balls, the intersection of the sample with $\overline{B_{k,b}} = \emptyset$. Since we have assumed that the sample is an ϵ-net, this implies that $|X \cap \overline{B_{k,b}}| \leq \epsilon n$. It thus follows by Equation 1 that the original set is not (k,b)-radius clusterable, which is what we wanted.

2 A Different Clustering Question

In this section, we study another notion of clustering. Suppose that you are given a finite set X and any metric on its points. Given a set of k "centers," $c_1, \ldots, c_k \in X$, let $f_{c_1,\ldots,c_k}(x) = \min_i \text{dist}(x, c_i)$, and let

$$E_X(f_{c_1,\ldots,c_k}) = \frac{1}{|X|}\sum_{x \in X} f_{c_1,\ldots,c_k}(x).$$

Our goal is to find a set of c_i that minimizes $E_X(f_{c_1,\ldots,c_k})$. The algorithm that we shall employ is the obvious one: pick a sample $S = s_1, \ldots, s_m$ from X and find the $c_1, \ldots, c_k \in S$ that minimize $E_S(f_{c_1,\ldots,c_k})$.

Theorem 7 (Mishra, Oblinger, Pitt) *Let M be the maximum distance between two points in X, let $\hat{f} = f_{c_1,\ldots,c_k}$ be the minimum distance function for the set of centers that our algorithm selects, and let f^* be the minimum distance function for the optimal set of centers. If we pick*

$$m = O\left(\frac{M^2}{\epsilon^2}(k \log n + \ln \frac{2}{\delta})\right)$$

points to be in our sample, then the algorithm will select a set of centers such that

$$E_X(\hat{f}) \leq 2E_X(f^*) + 3\epsilon$$

with probability at least $1 - \delta$.

Lemma 8 *Let F be a set of functions on X with $0 \leq f(x) \leq m$ for all $f \in F$ and all $x \in X$. Let $S = \{s_1, \ldots, s_m\}$ be a set of samples from X with*

$$m \geq \frac{M^2}{2\epsilon^2}\left(\ln |F| + \ln\left(\frac{2}{\delta}\right)\right).$$

Then

$$\Pr[\exists f \in F |\, |E_X(f) - E_S(f)| \geq \epsilon] \leq \delta.$$

Proof The proof is just like the proof of the Chernoff bound. ■

Lemma 9 *Let $F_S = \{f_{c_1,\ldots,c_k} | c_1, \ldots, c_k \in S\}$, and let $F_X = \{f_{c_1,\ldots,c_k} | c_1, \ldots, c_k \in X\}$. If $f^* \in F_X$ minimizes $E_S(f^*)$, then there exists $\hat{f} \in F_S$ such that $E_S(\hat{f}) \leq 2E_S(f^*)$.*

Proof To see this, take the centers of the the clusters with centers in S to be the closest sample points to the optimal centers in X. Let s_1 be an arbitrary point in S, let $x \in X$ be the closest of the optimal centers in X to s_1, and let $s_2 \in S$ be the closest point in S to x. Then we have

$$d(s_1, s_2) \leq d(s_1, x) + d(s_2, x) \leq 2d(x, s_1),$$

from which the lemma follows. ■

We are now ready to prove the correctness of the algorithm.
Proof [Proof of Theorem 7] We shall apply the above lemmas to the problem at hand. Note that when we apply Lemma 8, we have that $|F| = \Theta(n^k)$ since there are $\binom{n}{k}$ ways to select k centers from n points.

We now have the following three easy assertions:

1. By Lemma 8 , $E_X(\hat{f}) \leq E_S(\hat{f}) + \epsilon$.
2. By Lemma 9, $E_S(\hat{f}) \leq 2E_S(f^*)$.
3. By Lemma 8, $E_S(f^*) \leq E_X(f^*) + \epsilon$.

Putting these three statements together yields

$$E_X(\hat{f}) \leq 2E_X(f^*) + 3\epsilon,$$

as desired. ■