6.896 Sublinear Time Algorithms	April 24, 2007

Lecture 21

Lecturer: Ronitt Rubinfeld *Scribe: Ning Xie*

In this lecture, we are going to study property testing for clustering problems.

1 Clustering problems

We will be working in d-dimensional Euclidean space $\mathbb{R}^d$. For any two points x and y in $\mathbb{R}^d$, their L_2-distance (Euclidean distance) is defined to be $\text{dist}(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$. A Euclidean ball at center $x_0 \in \mathbb{R}^d$ of radius r is the set of points that within distance r from x_0, i.e., $\mathbb{B}_b(x_0) = \{x \in \mathbb{R}^d : \text{dist}(x, x_0) \leq b\}$. Let X be a set of n points in $\mathbb{R}^d$.

Definition 1 *X is* $(\mathbf{k}, \mathbf{b})$**-radius clusterable** *((k, b)-r.c. for short) is X can be partitioned into k subsets such that each subset can be covered by a Euclidean ball of radius at most b.*

Given a set of n points in $\mathbb{R}^d$, the problem of deciding if these points are (k, b)-r.c. is known to be NP-hard. Therefore it is unlikely to find any polynomial time algorithm for this problem. Can we find an efficient property testing algorithm for Clustering?

2 Testing algorithm

First we define a subset X of $\mathbb{R}^d$ of size n to be ϵ-far from (k, b)-r.c. if we need to delete at least ϵn points from X in order to make it (k, b)-r.c. We have the following surprisingly efficient testing algorithm for Clustering:

Testing Algorithm for Clustering

1 sample $m = \Theta(\frac{dk}{\epsilon} \log \frac{dk}{\epsilon})$ points from X

2 if the sampled points are (k, b)-r.c., PASS;

else FAIL

There are algorithms that find k balls with minimum radius that contain all m sampled points (known as Euclidean k-Center Problem) in time $O(m^{kd+2})$. Therefore the testing algorithm has both sample complexity and time complexity independent of n, the size of the input. The rest of this lecture is devoted to proving the following theorem, which shows the correctness of the testing algorithm.

Theorem 2 *1. If X is (k, b)-r.c., then tester passes;*

2. If X is ϵ-far from (k, b)-r.c., then $\Pr[$*tester fails*$] \geq 3/4$.

It is clear that is X is (k, b)-r.c., then the tester always passes. The non-trivial part of the proof is the second part. In order to prove the second part, we need to introduce an important concept called shatter exponent.

3 Shatter exponent

We begin with some notations.

1. $\mathbb{B}_b = \{B : B \text{ is a ball in } \mathbb{R}^d \text{ of radius at most } b\}$, i.e., $\mathbb{B}_b$ is the set of all balls in $\mathbb{R}^d$ of radius $\leq b$.

2. $\mathbb{B}_{k,b} = \{S : S \text{ is a union of at most } k \text{ balls in } \mathbb{B}_b\}$.

3. $\mathbb{B}^c_{k,b} = \{\bar{S} : S \in \mathbb{B}_{k,b}\}$.

Observation 3 *If $X \subseteq S$ for some $S \in \mathbb{B}_{k,b}$, then X is (k,b)-r.c. Conversely, if X is not (k,b)-r.c., then for all $S \in \mathbb{B}_{k,b}$, $X \nsubseteq S$. Or equivalently, if X is not (k,b)-r.c., then for all $\bar{S} \in \mathbb{B}^c_{k,b}$, $X \cap \bar{S} \neq \emptyset$. Similarly, if X is ϵ-far from (k,b)-r.c., then $\forall S \in \mathbb{B}_{k,b}$, $|X \setminus S| \geq \epsilon n$; so, $\forall \bar{S} \in \mathbb{B}^c_{k,b}$, $|X \cap \bar{S}| \geq \epsilon n$.*

Therefore, our plan is to show that, with high probability, the sample points contain at least one point of each $\bar{S} \in \mathbb{B}^c_{k,b}$. To this end, we introduce the following definitions.

Definitions

1. Let $X \subseteq \mathbb{R}^d$ be a point set and $\mathcal{S}$ be a family of subsets of X (i.e., $\mathcal{S} \subseteq 2^X$). A subset $N \subseteq X$ is an *ϵ-net on X w.r.t. $\mathcal{S}$* if for all $S \in \mathcal{S}$ such that $|X \cap S| \geq \epsilon |X|$, then $N \cap S \neq \emptyset$.

   Therefore, our hope is to show, w.h.p., the sample set is an ϵ-net on X w.r.t. $\mathbb{B}^c_{k,b}$.

2. For any $A \subseteq \mathbb{R}^d$, define the projection of $\mathcal{S}$ on A to be $\varphi_{\mathcal{S}}(A) = \{A \cap S : S \in \mathcal{S}\}$.

3. For each $m \in \mathbb{N}$, define $\Phi_{\mathcal{S}}(m)$ to be the maximum size of a projection of $\mathcal{S}$ on a set of size m: $\Phi_{\mathcal{S}}(m) = \max_{A:|A|=m} |\varphi_{\mathcal{S}}(A)|$.

4. The **shatter exponent** of $(X, \mathcal{S})$, denoted $\mathrm{SE}(X, \mathcal{S})$[1], is the minimum ℓ such that there exists a constant c with $\Phi_{\mathcal{S}}(m) \leq cm^\ell$ for all $m \geq 2$.

The importance of shatter exponent is demonstrated in the following theorem.

Theorem 4 (Haussler and Welzl, 1987) *Let X be a set of points in $\mathbb{R}^d$ and let $\mathcal{S}$ be a family of subsets of X. Let U be a random sample set from X of size*

$$|U| \geq \Theta\left(\frac{SE(X,\mathcal{S})}{\epsilon} \log \frac{SE(X,\mathcal{S})}{\epsilon}\right),$$

then $\Pr[U$ *is an ϵ-net on X w.r.t. $\mathcal{S}$*$] \geq 3/4$.

We are not going to prove this theorem here, interested readers are referred to the original paper. Combining our previous observations with this theorem, our task of proving Theorem 2 is thus reduced to proving that $\mathrm{SE}(\mathbb{B}^c_{k,b}) = O(kd)$.

4 Bounding the shatter exponent

We will begin with a simple fact for d-dimensional Euclidean space. Consider the simplest one dimensional case. Let A be a set of points on a line. Define $B^*(A)$ to be the minimum segment that covers A. Then there are two points a and b in A, namely the two endpoints on the line, such that the minimum segment that covers $\{a, b\}$ is the same segment that covers A. Generalizing this observation to higher dimension, for a point set A in $\mathbb{R}^d$, define $B^*(A) = \{$minimum radius ball in $\mathbb{R}^d$ that covers $A\}$, then we have

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

[1] The shatter exponent is a lower bound on the corresponding VC-dimension of $(X, \mathcal{S})$. The VC-dimension (Vapnik-Chervonekis Dimension), a very important notion in machine learning theory, is defined as follows. For any $A \subseteq X$, we say A is shattered if $\varphi_{\mathcal{S}}(A) = 2^A$, i.e., the projection of $\mathcal{S}$ on A includes all possible subsets of A. The VC-dimension of $(X, \mathcal{S})$ is the cardinality of the largest set A that it shatters: VC-dim $= \max\{|A| : A \text{ is shattered}\}$. By definition, for every $m \leq$ VC-dim$(X, \mathcal{S})$, $\Phi_{\mathcal{S}}(m) = 2^m$. It can be shown that, for all $(X, \mathcal{S})$, $\mathrm{SE}(X, \mathcal{S}) \leq$ VC-dim$(X, \mathcal{S})$.

It follows that the total number of different minimum radius covering balls for X is at most $\binom{n}{d+1} < n^{d+1}$.

Our next observation is, for clustering purposes, it is enough to argue in terms of *minimum radius covering balls* w.r.t. set X. That is, all of our previous discussions on the relation between X being (k,b)-r.c. and b-radius balls in $\mathbb{R}^d$ still apply if we replace the general balls with minimum radius covering balls w.r.t. X. Combining this observation with Fact 5 gives

$$\begin{aligned}\Phi_{\mathbb{B}_{1,b}}(n) &= \text{the number of minimum radius covering balls w.r.t. } X \text{ of radius at most } b \\ &\leq \text{the number of minimum radius covering balls w.r.t. } X \\ &< n^{d+1}.\end{aligned}$$

Therefore $\mathrm{SE}(X, \mathbb{B}_{1,b}) < (d+1)$ by definition. Since $\mathbb{B}_{k,b}$ includes unions of k balls, $\mathrm{SE}(X, \mathbb{B}_{k,b}) < k(d+1)$. Finally note that $A \subseteq X$ is in $\mathcal{S}$ if and only if $X \setminus A$ is in $\bar{\mathcal{S}}$. Therefore, $\mathrm{SE}(X, \mathbb{B}^c_{k,b}) = \mathrm{SE}(X, \mathbb{B}_{k,b}) < k(d+1) = O(kd)$. This complete the proof of Theorem 2.