0368.416701 Sublinear Time Algorithms December 7, 2009

Lecture 8

Lecturer: Ronitt Rubinfeld *Scribe: Roii Werner*

1 Lecture outline

So far, we have seen algorithms for graph related problems. In this lecture two non graph related problems are discussed:
• Clusterability of points.
• Weakly approximating edit distance.

2 Clustering

Throughout this section, we'll use the following notation:
$X \subseteq \mathbb{R}^d$, $|X| = n$

$$\text{DIST}(X, Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (L^2 \text{ distance})$$

2.1 Definitions

Definition 1 *A set of points $X \subseteq \mathbb{R}^d$ is said to be (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 $\leq b$.*

As we did with graph properties, we would like to define a notion of a set of points being "close" to (k, b)-radius clusterable.

Definition 2 *A set of points $X \subseteq \mathbb{R}^d$ is said to be ϵ-**far from** (k, b)-**radius clusterable** if at least ϵn points must be deleted from it to make it (k, b)-radius clusterable.*

2.2 The algorithm

We suggest the following tester for (k, b)-radius clusterability:
1. Sample $m = \Theta(\frac{dk}{\epsilon} \log \frac{dk}{\epsilon})$ points from X.
2. Run **exact** (k, b)-radius clustering on the sample.
   PASS if sample is (k, b)-radius clusterable,
   else **FAIL**.

Note: There exist exponential algorithms for checking (k, b)-radius clusterability exactly. Since our sample size is independent of the size of the input set, step 2 runs in constant time.

2.3 Analysis of the algorithm

Theorem 3 *For an input set of points, $X \subseteq \mathbb{R}^d$:*
1) If X is (k, b)-radius clusterable, the tester passes.
2) If X is ϵ-far from (k, b)-radius clusterable, then $Pr[\text{Tester fails}] \geq \frac{3}{4}$

The first statement is trivial (because if X is (k,b)-radius clusterable, then so is each of its subsets). The rest of our work is dedicated to proving the second statement.

Definition 4 *The following families of sets will assist us in our analysis:*
$\mathbb{B} = \{B | B \text{ is a ball of radius} \leq b \text{ in } \mathbb{R}^d\}$
$\mathbb{B}_{k,b} = \{S | S \text{ is a union of at most } k \text{ balls in } \mathbb{B}\}$
$\bar{\mathbb{B}}_{k,b} = \{S^c | S \in \mathbb{B}_{k,b}\}$

Note: If $X \subseteq S$ for some $S \in \mathbb{B}_{k,b}$ then X is (k,b)-radius clusterable.
Equivalently, if X is **not** (k,b)-radius clusterable, then
$\forall S \in \mathbb{B}_{k,b}$, $X \nsubseteq S$, so $\forall S$, $X \cap S^c \neq \emptyset$.
Moreover, if X is ϵ**-far** from (k,b)-radius clusterable, then
$\forall S \in \mathbb{B}_{k,b} \quad |X \cap S^c| \geq \epsilon n$.
Put in words, if X is ϵ**-far** from (k,b)-radius clusterable, then at least ϵn points must be removed from it to make it (k,b)-radius clusterable, which means any attempt to cover X by elements of $\bar{\mathbb{B}}_{k,b}$ must exclude at least ϵn points.

Main idea: The chosen sample hopefully contains at least one element of each $S^c \in \bar{\mathbb{B}}_{k,b}$. To formalize this idea we must first present some more definitions.

2.3.1 The shatter exponent

Definition 5 *Let $\mathbb{S}$ be a family of subsets of X.*
*We say that $N \subseteq X$ is an ϵ**-net on** X **with respect to** $\mathbb{S}$ if $\forall S \in \mathbb{S}$, $|X \cap S| \geq \epsilon n \Rightarrow N \cap S \neq \emptyset$*

Clearly, from the definition, if X is ϵ-far from (k,b)-radius clusterable and our sample is an ϵ-net on X with respect to $\bar{\mathbb{B}}_{k,b}$, then our tester fails the input (because then the sample has nontrivial intersection with each $S^c \in \bar{\mathbb{B}}_{k,b}$ which means it is not contained in any $S \in \mathbb{B}_{k,b}$).

Definition 6 *For a family $\mathbb{S}$ of subsets of X, $m \in \mathbb{N}$ and $A \subseteq \mathbb{R}^d$ a finite set:*
1) $\phi_{\mathbb{S}}(A) = \{A \cap S | S \in \mathbb{S}\}$
2) $\phi_{\mathbb{S}}(m) = \max_{A:|A|=m} |\phi_{\mathbb{S}}(A)|$
3) The ***shatter exponent*** *of $\mathbb{S}$:* $SE(\mathbb{S}) = \min l$ *such that* $\exists c\ \phi_{\mathbb{S}}(m) \leq cm^l$

Theorem 7 *For a sample U such that* $|U| = O(\frac{SE(\mathbb{S})}{\epsilon} \log \frac{SE(\mathbb{S})}{\epsilon})$*:*
$Pr[U \text{ is an } \epsilon - \text{net on } X \text{ with respect to } \mathbb{S}] \geq \frac{3}{4}$

Recall that our goal was to show that if the input is ϵ-far from (k,b)-radius clusterable, then our tester fails with high probability, and we've seen that if the sample is an ϵ-net on X with respect to $\bar{\mathbb{B}}_{k,b}$, then our tester does indeed fail. Put together with the above theorem, what remains for us to show is that $SE(\bar{\mathbb{B}}_{k,b}) = O(dk)$.

2.3.2 Bounding the shatter exponent of $\bar{\mathbb{B}}_{k,b}$

We begin by bounding $SE(\mathbb{B})$:
$|\phi_{\mathbb{B}}(A)|$ =The number of sets of points Y that can be covered by a ball so that no other point in $A \setminus Y$

is covered = The number of sets of points Y that can be covered by minimal balls such that no outer point is covered.
Denote by $B^*(A)$ the minimum radius ball covering A.
Fact (without proof): $\forall A \exists A' \subseteq A$ such that $B^*(A) = B^*(A')$ and $|A'| \leq d+1$.
So, for $|A| = m$, the number of such minimal balls for sets in $\phi_{\mathbb{B}}(A)$ is $\leq \binom{m}{d+1} \leq m^{d+1}$ and therefore by the above description of $\phi_{\mathbb{B}}(A)$ we get that $SE(\mathbb{B}) \leq d+1$.
We proceed to bound $SE(\mathbb{B}_{k,b})$:
As before, elements Y of $\phi_{\mathbb{B}_{k,b}}(A)$ can be uniquely determined by the minimal k balls which cover them, and by the same reasoning as above there are at most $(m^{d+1})^k) = m^{(d+1)k}$ such choices of k balls (because there are at most m^{d+1} choices for each ball independently), and so $SE(\mathbb{B}_{k,b}) \leq (d+1)k$.
Finally, we bound $SE(\bar{\mathbb{B}}_{k,b})$:
Notice that $|\phi_{\mathbb{B}_{k,b}}(A)| = |\phi_{\bar{\mathbb{B}}_{k,b}}(A)|$ (since there is a one-to-one mapping between subsets of A and their complements, and a subset of A is covered by k balls (such that no additional points are covered) if and only if its complement is covered by the complement of the union of those balls).
So it follows from the bound on $SE(\mathbb{B}_{k,b})$ that $SE(\bar{\mathbb{B}}_{k,b}) \leq (d+1)k$, and this completes the proof of theorem 3.

3 Weakly approximating Edit Distance

Definition 8 *Let A, B be strings.*
*The **Edit Distance** $ED(A,B)$ is the minimal number of character insertions/deletions/substitutions required to transform A into B (or vice versa).*

3.1 First attempt

Let's try to look at this problem from our standard viewpoint, i.e trying to determine when two strings are close to each other in "ϵ-far" terms.
That is, given input strings A, B such that $|A| = |B| = n$:
if $A = B$ output **PASS**.
if $ED(A,B) \geq \epsilon n$ output **FAIL** with high probability.
From our experience it should be clear that the solution is very straightforward: Pick $O(\frac{1}{\epsilon})$ pairs of characters at random and compare them. If any pair is different, fail, else pass.
Standard analysis shows this indeed has the desired behavior.
Our interest will therefore be in a less trivial variant of the problem.

3.2 The goal

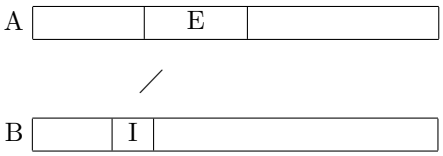
We wish to devise an algorithm with the following behavior:
Input: A, B such that $|A| = |B| = n$.
if $ED(A,B) \leq n^{\alpha}$ output **PASS** with probability $\geq \frac{3}{4}$.
if $ED(A,B) > cn$ output **FAIL** with probability $\geq \frac{3}{4}$.

Theorem 9 *The following bounds hold for the above stated problem:*
$\tilde{O}(n^{\frac{\alpha}{2}})$ *for* $\alpha \leq \frac{2}{3}$.
$\tilde{O}(n^{2\alpha-1})$ *for* $\frac{2}{3} < \alpha < 1$.
$\Omega(n^{\frac{\alpha}{2}})$.

Note: $\tilde{O}(g(n)) = O(g(n)\log^k(g(n)))$ for some k.

3.3 Matching substrings

Definition 10 *A string E of length n ϵ-**matches** a string I of the same length if $HD(E, I) \leq \epsilon n$ (where HD is Hamming Distance - the number of substitutions required to get from E to I).*



The substring matching problem:
Given:
• String $E = E_1 \dots E_m \quad E \subseteq A$
• Block $I = I_1 \dots I_b \quad I \subseteq B$
• ϵ
Find: (in $O(n^{\frac{\alpha}{2}})$ queries)
All shifts $t \leq 2n^{\alpha} \equiv u$ such that $E_{t+1} \dots E_{t+b}$ ϵ-matches I.
Note: We assume $b << \sqrt{u}$.

3.3.1 The Ruler Procedure

The ruler procedure is an efficient procedure to find these approximately matching substrings.
The idea is to define a "ruler" on the strings, where "centimeters" are units of size $\sqrt{u}$ and "millimeters" go from 1 to $\sqrt{u}$.
More formally, if we define:
$Q_A = \{\sqrt{u}, 2\sqrt{u}, \dots, u\}$
$Q_B = \{1, 2, 3, \dots, \sqrt{u}\}$
Then $\forall t \in \{0, \dots, u-1\}$ $\exists i \in Q_A, j \in Q_B$ such that $i - j = t$:
$cen \leftarrow \lceil \frac{t}{\sqrt{u}} \rceil$
$mil \leftarrow t \bmod \sqrt{u}$
$i \leftarrow cen \cdot u$
$j \leftarrow \sqrt{u} - mil$
Notice that now by comparing $E[i]$ to $I[j]$ we compare one character of a t-shifted block of size b in E to the corresponding character in I.
Now, to approximately compare t-shifts of length b in E to I, we may proceed as follows:
Sample $l = \Theta(\log n)$ numbers $m_1, \dots, m_l \in \{0, \dots, b - \sqrt{u}\}$.
For $i \in Q_A, j \in Q_B$ such that $i - j = t$, consider the **"fingerprint"** given by:
$\{i + m_1, i + m_2 \dots, i + m_l\}$
$\{j + m_1, j + m_2 \dots, j + m_l\}$
Comparing the first set of indices in E to the second set of indices in I determines if this t-shift of E approximately matches I with high probability.

3.3.2 Efficiently implementing the ruler procedure

We want to efficiently determine when the ruler mark i in E generates the same fingerprint as the ruler mark j in I.
To do this, we maintain a binary search tree, keyed by the values of the fingerprints. Each leaf in the tree will point to all the indices of E and all the indices of I which resulted in that fingerprint.