TEL AVIV UNIVERSITY
Department of Computer Science
0368.3239 – Leveraging Big Data
Fall Semester, 2013/2014

Homework 2, Dec 4, 2013

- **Due on December 29, 2013.**
- **You are allowed to consult any sources or other humans, but you must write and submit the solution yourself and understand it in detail.**
- **Please submit a PDF file to the email address that will appear on the Web site. The file should be named `firstname_lastname_HW2.pdf` . Keep a copy. Include your full name and ID in the file.**

1. Consider the following streaming algorithm:

Items in set X arrive via stream. All distances, $1 \leq d(x,y) \leq M$, for $x \neq y$, $x, y \in X$. Maintain a set S of $\leq k \log M$ centers: $S = \bigcup_{1 \leq j \leq \log M} S_j$. When first point arrives, add it to all S_j.

When i'th point arrives, $i \geq 2$, consider all $1 \leq j \leq \log M$: if x has distance $\geq 2^j$, from all $c \in S_j$, and if $|S_j| < k$, add x to S_j.

- What is the running time of the algorithm?
- What is the maximum over $x \in X$ of the minimum over $c \in S$ of the distance between x and c? How does this relate to the optimal k center radius? Prove.
- In what sense is the algorithm given in class better? Worse? How much better can it be?

2. (This partially repeats things we did in class, but nevertheless, please write as accurately and rigorously as you can.)
Consider vectors in R^d.

a. Describe a $(\theta_1, (1+\epsilon)\theta_1, 1 - \frac{\theta_1}{\pi}, 1 - \frac{(1+\epsilon)\theta_1}{\pi})$ locality sensitive hash family for these vectors. That is, the probability that two vectors x and y such that the angle between them is at most θ_1 are hashed to the same bucket with probability at least $1 - \frac{\theta_1}{\pi}$, and two vectors x and y such that the angle between them is at least $(1+\epsilon)\theta_1$ are hashed to the same bucket with probability at most $1 - \frac{(1+\epsilon)\theta_1}{\pi}$.

b. Describe how to turn the family above into $\left(\theta_1, (1+\epsilon)\theta_1, (1 - \frac{\theta_1}{\pi})^k, (1 - \frac{(1+\epsilon)\theta_1}{\pi})^k\right)$ locality sensitive hash family.

c. Describe precisely and accurately an algorithm that preprocesses a set of n vectors $x_1, \ldots, x_n$ in R^d into an $O(dn^{1+\frac{1}{1+\epsilon}})$ space data structure such that given any query vector q it can perform the following in $O(dn^{\frac{1}{1+\epsilon}})$ time: If there is a vector x_j of angle at most θ_1 with q then with probability at least 99/100 we return some vector x_i such that the angle of x_i and q is at most $(1+\epsilon)\theta_1$. (You may neglect logarithmic factors in space and time bounds.)

d. Analyze your algorithm.

3. We are given n vectors $x_1, \ldots, x_n$ in R^{n^2} and a constant $\epsilon > 0$. We want to (randomly) map these vectors (using the same mapping) into n vectors $y_i = f(x_i)$ in R^{2n/ϵ^2} such that with probability at least $\frac{1}{e}$ (over the draws defining the mapping f) for each i, $(1-\epsilon)||x_i||^2 \leq ||y_i||^2 \leq (1+\epsilon)||x_i||^2$. Show how to achieve that using only $\Theta(n)$ space in addition to the input and the output vectors.

4. Let $x_1, \ldots, x_n$ be n unit vectors in R^d drawn uniformly at random from the unit sphere. Assume that $d >> \log n$. Let $\theta(x_i, x_j)$ be the angle between these vectors.

a. Prove that for any $c \geq 1$, with probability $\geq 1 - \frac{1}{n^c}$: for all i, j,

$$\cos^2(\theta(x_i, x_j)) \leq c' \frac{\log n}{d},$$

where c' is a constant that depends on the constant c. (Specify the relation between c and c'.)

b. Argue that this implies that any pair of these vectors are almost orthogonal with very high probability.