TEL AVIV UNIVERSITY
Department of Computer Science
0368.3239 – Foundations of Data Mining
Fall Semester, 2017/2018

Homework 2, November 20, 2017

- **Due on Tuesday December 4 23:59 IST.**
- **Submission instructions: We are using** `https://gradescope.com`**. Gradescope entry code for the course is: MYVDZ5**
  Please prepare a PDF file with each problem starting on a new page. When uploading, you will need to indicate locations of each problem/section.
- **You may consult any sources or people but you must write and submit the solution yourself and state your collaborators.**

1. Consider the linear sketch of size 3 defined by the following matrix

$$M = \begin{bmatrix} 1 & 1 & 1 & 1 & \dots & 1 \\ 1 & 2 & 3 & 4 & \dots & n \\ 1 & 4 & 9 & 16 & \dots & n^2 \end{bmatrix} .$$

We maintain $s = Mb$ for a *non-negative* vector b of length n. Let $s = (s_0, s_1, s_2)$.

a. (10 points) Show how to use s to detect if b contains exactly one non-zero entry (without error) and return this entry in case b indeed contains exactly one non-zero entry.

b. (10 points) Describe a linear sketch that returns a positive element from b uniformly at random with constant probability. (You may still assume that b is non-negative so you can use your sketch in (a).) State the query algorithm precisely. Give a lower bound on the success probability of your sketch. What is the size of your sketch ?

   Hint: Design a simple sketch that samples a random subset of 2^j elements from b in expectation for $j = 0, 1, \dots, \log n$, and compose it with sketches from the first part of this question.

You may use a fully random hash functions and assume that n is a power of 2 for simplicity.

2. A count-min sketch is designed to approximate the values of the entries of the vector b which we sketch that is initially all 0 and when provided with positive updates of the form (i, x) ($b_i := b_i + x$).

The sketch has two parameters (ϵ, δ) and consists of an $d \times w$ array count of counters where $w = \lceil \frac{e}{\epsilon} \rceil$ and $d = \lceil \ln(1/\delta) \rceil$. Let $h_1, \dots, h_d$ independent hash functions from a pairwise independent family of hash functions mapping $\{1, 2, \ldots, n\}$ to $\{1, \ldots, w\}$. We initialize the counters to 0, and we process an update (i, x) by setting

$$\text{For } j = 1, \ldots, d : \ \ \text{count}[j, h_j[i]] \leftarrow \text{count}[j, h_j[i]] + x \ .$$

Assuming that all updates are positive, we approximate the current value of b_i by

$$\hat{b}_i := \min_{j=1,\ldots,d} \text{count}[j, h_j[i]] \ .$$

a. (10 points) The count-min sketch is in fact a linear sketch of the form $s = Mb$. Define precisely the matrix M. What is the length of s ?

b. (5 points) Prove that $b_i \le \hat{b}_i \le b_i + \epsilon ||b||_1$ with probability $1 - \delta$ (where $||b||_1 = \sum_{i=1}^{n} |b_i|$). **Hint:** Use Markov inequality.

   To allow both positive and negative updates we add a few more rows to the array so now we have $d' \ge d$ rows and w columns. The update procedure is defined as before in all d' rows but our estimate is

$$\hat{b}_i = \text{median}_{j=1,\ldots,d'} \text{count}[j, h_j[i]] \ .$$

c. (5 points) How many rows d' do we need in order to guarantee that $b_i - 2\epsilon ||b||_1 \le \hat{b}_i \le b_i + 2\epsilon ||b||_1$ with probability at least $1 - \delta$? **Hint:** Use Markov and Chernoff bounds.

3. **Estimating weighted Jaccard similarity**. Consider two nonnegative vectors

$$V = (v_1, \ldots, v_n) \tag{1}$$
$$U = (u_1, \ldots, u_n) \ . \tag{2}$$

The weighted Jaccard similarity of the two vectors is defined as

$$J(U, V) := \frac{|| \min\{U, V\}||_1}{|| \max\{U, V\}||_1} \ ,$$

where

$$|| \min\{U, V\}||_1 := \sum_{i=1}^{n} \min\{v_i, u_i\}$$
$$|| \max\{U, V\}||_1 := \sum_{i=1}^{n} \max\{v_i, u_i\} \ .$$

We are interested in estimating $J(U, V)$ from respective coordinated bottom-k samples $S(V)$ and $S(U)$ of the vectors. Assume $k \ge 3$.

Note: *coordinated* means the samples are computed using the same hash function.

Note: By a bottom-k sample of a vector V we refer to a bottom-k sample of the set of the key-value pairs $\{(i, v_i)\}$ that include all $i = 1, \ldots, n$ where $v_i > 0$.

For simplicity, assume we use pps ("priority") sampling ($r(i, v_i) = h(i)/v_i$ for $h \sim U[0, 1]$).

Given $S(V)$, $S(U)$, and h

a. (8 points) Give a nonnegative unbiased estimator to $||\min\{U, V\}||_1$

b. (8 points) Give a nonnegative unbiased estimator to $||\max\{U, V\}||_1$

c. (4 points) Compare the quality of your $||\max\{U, V\}||_1$ estimate to the quality of an estimate obtained directly from a bottom-k sample $S(A)$ of the vector $A = \max\{U, V\}$. In which case we can get a better estimate? Substantiate your claim.

Note: For concreteness, you may consider the conditioned inverse probability estimate shown in class.

4. **Difference estimation from samples.** Consider a vector $(w_1, \ldots, w_n)$ with nonnegative entries $w_i \geq 0$ that is pps sampled with some $\alpha > 0$. So that the ith entry is sampled (that is, (i, w_i) is included in the sample S) independently with probability $\min\{1, \alpha w_i\}$.

a. (5 points) Assume that we know that $w_1 \geq w_2 > 0$. Find an unbiased nonnegative estimator for $w_1 - w_2$.

b. (5 points) What is the variance of your estimator? (expressed in terms of α, w_1 and w_2)

c. (5 points) What can you say on w_i when the sample does not include the ith entry? (specify the tightest range you can)

d. (5 points) We now assume that we know that $w_1 \geq w_2 \geq 0$ (allowing the case $w_2 = 0$). Specify an unbiased estimator (that can be negative) for $w_1 - w_2$. Is there an unbiased nonnegative estimator? (prove).

5. **Difference estimation from hash-based samples.** Consider pps sampling as in the previous question. Except that this time we use a random hash function $h(i) \sim U[0, 1]$ (assume $h(i)$ and $h(j)$ are independent when $i \neq j$) to perform the sampling: (i, w_i) is included in the sample iff $h(i) \leq \alpha w_i$. We assume that $h \sim H$ is available to us with the sample S. In the following, assume that we know that $w_1 \geq w_2$.

a. (5 points) What can we say about the value of w_i when the ith entry is not sampled? (specify the tightest range you can)

b. (5 points) What can we say on $w_1 - w_2$ given a sample S? (specify the tightest range you can)

c. (10 points) Assuming $w_2 > 0$, can you find an unbiased nonnegative estimator for $w_1 - w_2$ with lower variance than in the previous question? If so, state it and analyse the variance. You may assume for simplicity that $w_1 \leq 1/\alpha$.

d. (Extra credit 10 points) Can you find an unbiased nonnegative estimator for $w_1 - w_2$ when allowing $w_2 = 0$? You may assume for simplicity that $w_1 \geq 1/\alpha$. Prove that it is unbiased for all $w_2 \geq 0$.