solution sketch for TAU Foundations of Data Mining 2017/2018 MoedA

March 7, 2018

1 Solution sketch for problem 2

a The array is initialized to "F" (0) values. To insert a key x, we set $S[h_i(x)] \leftarrow T$ for $h = 1, \ldots, k$.

b If $S[h_i(x) == T$ for $i = 1, \ldots, k$ we return that x was inserted. Otherwise, we say that x was not inserted.

c False negatives are not possible. When x is inserted we set $S[h_i(x)]$ to be T for all $i = 1, \ldots, k$. There is no operation that sets cells to F. Therefore, $S[h_i(x)]$ remains T and membership query will always be positive.

d False positive are possible. For example, we insert a key A with $h_1(A) = 2$ and $h_2(A) = 3$. This sets $S[2] \leftarrow T$ and $S[3] \leftarrow T$.

We then ask for membership on a key B which happened to have $h_1(B) = 3$ and $h_2(B) = 2$. We will be a false positive.

Expression: We insert one key A, and from our assumption now exactly k cells out of m are set to T (the cells $cells(A) := \{h_i(A) \mid i = 1, \ldots, k\}$)

A second key will be false positive if and only if the set $cells(B) := \{h_i(B) \mid i = 1, \ldots, k\}$ is equal to $cells(A)$.

The probability of that is that we choose a particular unordered k tuple out of m. The number of k tuples is

$$\binom{m}{k} = \frac{m!}{k!(m-k)!}$$

and they are all equally likely so we can express the probability as $\frac{k!(m-k)!}{m!}$.

Approximate solutions that are reasonably close and well justified got almost all points.

Solution sketch for problem 3

a Node-distance pairs by increasing distance from v_1:

$$(v_1, 0)\ (v_2, 1)\ (v_3, 2)\ (v_4, 4)\ (v_5, 6)\ (v_6, 7)\ (v_7, 8)$$

b Bottom-2 all-distances sketch of node v_1 includes the nodes: v_1, v_2, v_4, v_6

c The presence is "1" for all reachable nodes from v_1. In our case, all nodes.

The two closest nodes are always in the sketch. So we can simply use a zero variance presence estimator of "1".

As hinted, the estimate is 0 for nodes not in the sketch.

For a node v in the sketch, we use the probability that it is included conditioned on the h values of other nodes. The node v It is included if its h value is less than the 2nd lowest h value among all closer nodes. (By definition of the sketch the nodes with the two lowest h values within any distance must be in the sketch, so these two nodes are in the sketch. So we can simply use the 2nd lowest h value among closer nodes than v that are in the sketch. Lets denote this value by τ_v)

Since we use $h(v) \sim U[0,1]$, the probability of being lower than $h(v) < \tau_v$ is τ_v. So the inverse probability estimate we get is $1/\tau_v$.

To show that the estimate is unbiased, we consider the "presence" estimate of a node conditioned on the h values of all other nodes. Each node has a positive probability $\tau_v > 0$ to be included in the sketch that is exactly $1/\tau_v$. Since the estimate is unbiased for any fixed set of h values for other nodes, it is unbiased.

Note: Some students proposed using the bottom-2 estimators. For that we extract the bottom-2 sketch from the ADS (two nodes with overall lowest h values). We can then use a bottom-2 estimator that is 0 for all nodes that in the sketch and positive only for the node with lowest hash.

This estimator is unbiased but is much weaker as it throws away most of the information available in the sketch. Correct use required using exponentially distributed values using the transformation $-\ln(1-h(v))$.

d The nodes in the sketch within distance at most 4 from v_1 are

v_1 (presence estimate $= 1$) , v_2 (estimate $= 1$), and v_4 (estimate $= 1/0.74$)

The node v_3 is not in the sketch and has estimate $= 0$.

e The neighborhood estimate is the sum of the presence estimate of neighborhood nodes that are in the sketch, which is $1 + 1 + 1/0.74$.

Some students extracted a bottom-2 sketch of the 4-neighborhood and used one of the cardinality estimators we used in class. This is an unbiased estimate but is weaker. Again, correct use required using exponentially distributed values using the transformation $-\ln(1-h(v))$.