Shanghai Jiao Tong University

Assignment3 Concept Questions

Student name: *520030910342 Jiyu Liu*

Course: *Data Mining* – Professor: *Liyao Xiang*

Date: *May 20, 2023*

1 Concepts Questions

1.1 Qustion 1

Given the size m of the key set and the length of n of the bit strings, letting k be the number of independent hash functions, we can get that the false positive possibility

$$FP(k) = (1 - e^{-km/n})^k$$

.

The logarithm of this equation is

$$ln(FP(k)) = kln(1 - e^{-km/n})$$

And the derivative of the above equation is that

$$\frac{\mathrm{d}ln(FP(k))}{\mathrm{d}k} = ln(1 - e^{-km/n}) + \frac{km}{n}\frac{e^{-km/n}}{1 - e^{-km/n}}$$

Let x be $e^{-km/n}$, $x \in (0,1)$ $(k,m,n > 0)$. Then we can get the equation as below:

$$\frac{\mathrm{d}ln(FP(x))}{\mathrm{d}x} = ln(1 - x) - ln(x)\frac{x}{1 - x}$$

Letting the derivative be zero, we can get that $x = \frac{1}{2}$. So $e^{-km/n} = \frac{1}{2}$. Therefore

$$k = ln2\frac{n}{m}$$

1.2 Question 2

In this stream, 1 occurs three times, 2 occurs two times, 3 occurs two times, 4 occurs 2 times. According to the definition the second moment for the stream is $3^2 + 2^2 + 2^2 + 2^2 = 9 + 4 + 4 + 4 = 21$. Similarly, the third moment for the stream is $3^3 + 2^3 + 2^3 + 2^3 = 27 + 8 + 8 + 8 = 51$.

1.3 Question 3

(a) For each item, estimate the average purchase price.

Key attribute: The item purchased

To estimate the average purchase price for each item, we would need to sample the data in a way that includes each item proportionally to its presence in the full dataset. In this way, the sample maintains the distribution of purchase prices for each item.

(b) Estimate the fraction of customers who made a purchase of $50 or more.

Key attributes: the Customer's ID and the Purchase Price

To estimate the fraction of customers who made a purchase of $50 or more, we need to maintain the distribution of customers' IDs and purchase prices. (c) Estimate the fraction of items that were purchased by at least 10 customers. Key attributes: the Customer's ID and the item purchased

To estimate the fraction of items that were purchased by at least 10 customers, we need to maintain the distribution of customers' IDs and the item purchased.