AI3602 Data Mining: Homework 7

Xiangyuan Xue (521030910387)

1. Assume that every hash function maps a key to the n buckets with equal probability. Then the probability that a key is hashed to a specific bucket is $\frac{1}{n}$. Note that there are m keys and k functions, so the probability that a bucket is empty can be specified as $\left(1-\frac{1}{n}\right)^{km}$. If a false positive occurs, all the k buckets should contain at least one key, so the probability of making a mistake should be

$$p(k) = \left[1 - \left(1 - \frac{1}{n}\right)^{km}\right]^k \approx \left(1 - e^{-\frac{km}{n}}\right)^k$$

where the approximation is based on the assumption that n is large enough. To minimize the probability of making a mistake, we define the logarithm function $f(k)$ as

$$f(k) = \log p(k) = k \ln\left(1 - e^{-\frac{km}{n}}\right) = -\frac{n}{m} \ln\left(e^{-\frac{km}{n}}\right) \ln\left(1 - e^{-\frac{km}{n}}\right)$$

According to the symmetric properties of the function, we know that the minimum value is achieved when $e^{-\frac{km}{n}} = \frac{1}{2}$, namely $k = \frac{n}{m} \ln 2$. In practice, using $k = \lceil \frac{n}{m} \ln 2 \rceil$ is preferred since k should be a positive integer.

2. In the given stream, the element 1 appears 3 times, element 2 appears 2 times, element 3 appears 2 times, and element 4 appears 2 times. Therefore, the surprise number is

$$\sum_{i=1}^{4} m_i^2 = 3^2 + 2^2 + 2^2 + 2^2 = 21$$

Similarly, the third moment is specified as

$$\sum_{i=1}^{4} m_i^3 = 3^3 + 2^3 + 2^3 + 2^3 = 51$$

3. (a) We can design an estimation algorithm based on DGIM method. The item purchased and the purchase price will be used. For each item, we treat the purchase price as multiple binary streams, where each stream represents one bit of the price. Then we can use DGIM to count the number of ones in each stream and compute the total price by a weighted sum. The average price can be estimated by dividing the total price by the item count. The coming new items can be maintained following DGIM method, and the estimation can be updated accordingly.

   (b) We can solve the problem with Flajolet-Martin approach. The customer's ID and the purchase price will be used. We maintain two streams, one for all the customers

and the other for the customers who made a purchase of at least 50 dollars. Then we can use Flajolet-Martin to estimate the number of distinct customers. The fraction of customers who made a purchase of at least 50 dollars can be estimated by the ratio of two estimated values. The coming new items can be maintained following Flajolet-Martin approach, and the estimation can be updated accordingly.

(c) We can solve the problem with Flajolet-Martin approach. The customer's ID and the item purchased will be used. For each item, we maintain a stream for the customers who purchased it. Then we can use Flajolet-Martin to estimate the number of distinct customers for each item. Once an estimated value reaches 10, we increment the counter. The fraction of items purchased by at least 10 customers can be estimated by the ratio of the counter and the number of items. The coming new items can be maintained following Flajolet-Martin approach.