Data Mining Homework 03

Qiu Yihang

May 2023

1 #(Independent Hash Functions) in Bloom Filter

Solution. The optimal k should give a minimal false positive probability.

The fraction of false positive 1 on a certain bit in the vector B is $\left(1-e^{-\frac{km}{n}}\right)$.

We only get a false positive result when all k functions gives a false positive.

Thus, the probability for false positive is

$$\mathbf{Pr}\left[\texttt{false positive}\right] = \left(1-e^{-\frac{km}{n}}\right)^k$$

$$\min\ \mathbf{Pr}\left[\texttt{false positive}\right] \iff \frac{\partial}{\partial k}\left(1-e^{-\frac{km}{n}}\right)^k = 0$$

We have

$$\begin{aligned}\frac{\partial}{\partial k}\left(1-e^{-\frac{km}{n}}\right)^k = 0 &\iff \left(1-e^{-\frac{km}{n}}\right)^k\left[\ln\left(1-e^{-\frac{km}{n}}+\frac{m}{n}\frac{k}{e^{\frac{km}{n}}-1}\right)\right]=0\\ &\iff \text{either } 1-e^{-\frac{km}{n}}=0 \text{ or } \ln\left(1-e^{-\frac{km}{n}}+\frac{m}{n}\frac{k}{e^{\frac{km}{n}}-1}\right)=0\\ &\iff k=0\ (\textit{discarded}) \text{ or } \frac{n}{m}\ln 2\end{aligned}$$

Therefore, the optimal k is $\frac{n}{m}\ln 2$. ■

2 Moments

Solution. The frequencies of values for stream $3,1,4,1,3,4,2,1,2$ are as follows.

Value	1	2	3	4
Frequency	3	2	2	2

Thus,

the second moment, i.e. the surprise number, is $3^2 + 2^2 + 2^2 + 2^2 = 21$, and

the third moment is $3^3 + 2^3 + 2^3 + 2^3 = 51$. ■

3 Problem 03

(a). The key attribute should be (the item purchased, the purchase price). Sample a decent amount of samples for each item, and calculate the average price for them.

(b). The key attribute should be (the customer's ID, the purchase price.) We should sample randomly under a uniform distribution.

(c). The key attribute should be (the item purchased, the customer's ID.) We should sample randomly under a uniform distribution.