Assignment 4 for Large Scale Data Mining

Name: Zijian Zhang
Matrikelnr.: 3184680

May 22, 2016

1

$k = 3$, False positive rate is:

$$\lim_{m\to\infty}(1-\frac{1}{8m})^{3n} = 1-e^{-\frac{3n}{8m}}$$

$k = 4$, False positive rate is:

$$1-e^{-\frac{n}{2m}}$$

2

2.1

False positive probability within each hash function:

$$1-\lim_{\to\infty}(1-\frac{k}{n})^{m} = 1-e^{-\frac{km}{n}}$$

False positive probability using all k hash functions:

$$(1-e^{-\frac{km}{n}})^{k}$$

False positive probability using n bit for all of k hash functions:

$$(1-e^{-\frac{km}{n}})^{k}$$

thus they are the same.

2.2

False positive rate is:

$$(1-e^{-\frac{km}{n}})^{k}$$

the differential of k is

$$\frac{d}{dk}((1-e^{-\frac{km}{n}})^k) \tag{1}$$

$$= (ln(1-e^{-\frac{km}{n}}) + \frac{mke^{-\frac{km}{n}}}{1-e^{-\frac{km}{n}}})(1-e^{-\frac{km}{n}})^k \tag{2}$$

Let the differential be zero, thus

$$(2) = 0$$

$$ln(1-e^{-\frac{km}{n}}) = -\frac{mk}{n}\frac{e^{-\frac{km}{n}}}{1-e^{-\frac{km}{n}}}$$

Notice that k always occurs with form of $-\frac{mk}{n}$. Let $x = -\frac{mk}{n}$ and get differential of x for both sides, thus:

$$-\frac{e^x}{1-e^x} = \frac{(e^x + xe^x)(1-e^{-x}) + xe^x}{1-e^x}$$

$$x = -1$$

$$or$$

$$x = -ln(2)$$

Obviously, $x = -1$ does not pass. thus $x = -ln(2)$, i.e.

$$k = -\frac{nln(2)}{m}$$

which leads to the minimum false positive rate.

3

3.1

(a) largest length of tail 0 is 0, thus number of distinct element is $2^0 = 1$
(b) largest length of tail 0 is 1, thus number of distinct elements is $2^1 = 2$
(c) largest length of tail 0 is 4, thus number of distinct elements is $2^4 = 16$

3.2

If hash function is of form $h(x) = ax + bmod2^k$, should a better not be of form 2^t where $t \in \mathbb{N}$, because it introduces extra tail 0s thus the number of tail 0s totally determined by b.

4

4.1

surprise number= $3^2+2^2+2^2+2^2 = 21$, the third moment= $3^3+2^3+2^3+2^3 = 51$

4.2

$X_0.value = 2$, $X_1.value = 3$, $X_2.value = 2$, $X_3.value = 2$, $X_4.value = 1$, $X_5.value = 1$, $X_6.value = 2$, $X_7.value = 1$, $X_8.value = 1$

4.3

minimum possible surprise number= $\frac{n^2}{m}$
maximum possible surprise number= $m - 1 + (n - m + 1)^2$