数据科学与工程算法基础 习题4

Author：GONGGONGJOHN

1

当哈希函数个数 $k=3$ 时，误判率

$$\eta=\left(1-e^{-\frac{3}{8}}\right)^3\approx 0.0306$$

当 $k=4$ 时，误判率

$$\eta=\left(1-e^{-\frac{1}{2}}\right)^4\approx 0.0240$$

2

误判率

$$\eta=1-e^{-\frac{\frac{m}{n}}{k}}=1-e^{-\frac{km}{n}}$$

同一个哈希函数在每组中映射到的位置是相同的，使用同一个哈希函数将每个成员映射到 k 组的效果等同于映射到 1 组的效果，显然不同于使用 k 个哈希函数（误判率更高）。

5

$$\begin{aligned}J_{1,2}&=\frac{|\{2,3\}|}{|\{1,2,3,4,5,7\}|}=\frac{1}{3}\\J_{2,3}&=\frac{|\{2\}|}{|\{2,3,4,5,6,7\}|}=\frac{1}{6}\\J_{1,3}&=\frac{|\{2,4\}|}{|\{1,2,3,4,6\}|}=\frac{2}{5}\end{aligned}$$

6

当 A_1 和 A_2 中有 k 个元素相同时，

$$J_{A_1,A_2}=\frac{k}{2m-k}$$

此时

$$
\begin{aligned}
P(|A_1 \cap A_2| = k) &= \frac{\binom{n}{k}\binom{n-k}{m-k}\binom{n-m}{m-k}}{\binom{n}{m}\binom{n}{m}} \\
&= \frac{\binom{n-m}{m-k}}{\binom{n}{m}} \cdot \frac{\binom{n}{k}\binom{n-k}{m-k}}{\binom{n}{m}} \\
&= \frac{\binom{n-m}{m-k}}{\binom{n}{m}} \cdot \frac{m!}{k!(m-k)!} \\
&= \frac{\binom{m}{k}\binom{n-m}{m-k}}{\binom{n}{m}}
\end{aligned}
$$

故

$$
\mathbb{E}(J_{A_1,A_2}) = \sum_{k=0}^{m} \frac{k}{2m-k} \cdot \frac{\binom{m}{k}\binom{n-m}{m-k}}{\binom{n}{m}}
$$

若引入广义超几何函数，则其还可以写为

$$
\mathbb{E}(J_{A_1,A_2}) = \frac{m}{2m-1} \cdot \frac{\binom{n-m}{m-1}}{\binom{n}{m}} \cdot {}_3F_2(1-2m, 1-m, 1-m; 2-2m, n+2-2m; 1)
$$

7

当两个集合的Jaccard相似度为 0 时，特征矩阵不存在全为 1 的行

故 $P(mh(S_1) = mh(S_2)) = 0$

即Min-Hashing一定能给出一个正确的估计

8

由条件可知

$$
\begin{aligned}
E(X_i) &= P(h_i(S_1) = h_i(S_2)) = JS(S_1, S_2) \\
E\left(\widehat{JS}(S_1, S_2)\right) &= \frac{1}{k}\sum_{i=1}^{k} E(X_i) = JS(S_1, S_2)
\end{aligned}
$$

而由Chernoff bound可知

$$
P\left(\left|\frac{x-\mu}{\mu}\right| > \delta\right) < 2\exp\left\{-\frac{\mu\delta^2}{4}\right\}
$$

故对任意 $k > 0$，有

$$
\begin{aligned}
P\left(\left|\widehat{JS}(S_1, S_2) - JS(S_1, S_2)\right| > \varepsilon JS(S_1, S_2)\right) &= P\left(\left|\frac{k \cdot \widehat{JS}(S_1, S_2) - k \cdot JS(S_1, S_2)}{k \cdot JS(S_1, S_2)}\right| > \varepsilon\right) \\
&< 2\exp\left\{-\frac{k \cdot JS(S_1, S_2)\varepsilon^2}{4}\right\}
\end{aligned}
$$

现取

$$\delta = 2\exp\left\{-\frac{\varepsilon^2 \cdot k \cdot JS(S_1,S_2)}{4}\right\}$$
$$k = \frac{4\ln\left(\frac{2}{\delta}\right)}{\varepsilon^2 \cdot JS(S_1,S_2)}$$

则有

$$P\left(\left|\widehat{JS}(S_1,S_2) - JS(S_1,S_2)\right| > \varepsilon JS(S_1,S_2)\right) < \delta$$

且此时 $k = \mathcal{O}(\frac{\ln(1/\delta)}{JS\cdot\varepsilon^2})$

9

（1）由特征矩阵可知 $S_1 = \{0,2,4\}, S_2 = \{0,1\}, S_3 = \{3,4\}$

因此

$$J_{S_1,S_2} = \frac{1}{4}$$
$$J_{S_1,S_3} = \frac{1}{4}$$
$$J_{S_2,S_3} = 0$$

（2）我们首先计算其特征矩阵

行号	S_1	S_2	S_3	$h_1(x) = 7x+1 \mod 6$	$h_2(x) = 11x+2 \mod 6$	$h_3(x) = 5x+2 \mod 6$
0	1	1	0	1	2	2
1	0	1	0	2	1	1
2	1	0	0	3	0	0
3	0	0	1	4	5	5
4	1	0	1	5	4	4
5	0	0	0	0	3	3

故由上表易知其可得Min-Hashing签名为

	S_1	S_2	S_3
h_1	1	1	4
h_2	0	1	4
h_3	0	1	4

11

由条件可知 $1-(1-t^r)^b = \frac{1}{2}$

故

$$t = \left(1 - 2^{-\frac{1}{b}}\right)^{\frac{1}{r}}$$