

第 4 章课后习题参考

- 第 1 题：略
- 第 2 题：该方法的误判率约为 $(1 - e^{-\frac{km}{n}})^k$ 。和使用 k 个哈希函数相比，一方面哈希函数范围缩小，某一位被某个元素置为 1 的概率增加；另一方面由于每组只使用一个哈希函数，某非成员因 k 个分组同时被置为 1 导致误判的概率减小。两方面综合考虑，与使用 k 个哈希函数的整体误判率基本相当。

- 第 5 题：略

- 第 6 题：

假设 A_1 和 A_2 的交集大小为 a ，则 $JS(A_1, A_2) = \frac{a}{2m - a}$

因此，我们得到 $E(JS(A_1, A_2)) = \sum_a P(a) JS(A_1, A_2)$ ，其中 $P(a) = \frac{\binom{m}{a} \binom{n-m}{m-a}}{\binom{n}{m}}$

但是，我们要注意分两种 ($m \leq \frac{n}{2}$ 和 $m > \frac{n}{2}$) 情况讨论，最终结果如下

$$E(JS(A_1, A_2)) = \begin{cases} \sum_{a=0}^m \frac{\binom{m}{a} \binom{n-m}{m-a}}{\binom{n}{m}} \frac{a}{2m-a}, & m \leq \frac{n}{2} \\ \sum_{a=2m-n}^m \frac{\binom{m}{a} \binom{n-m}{m-a}}{\binom{n}{m}} \frac{a}{2m-a}, & m > \frac{n}{2} \end{cases}$$

- 第 7 题：课本中定理 4.1 的特殊情况
- 第 8 题： X_i 为伯努利随机变量且 $P(X_i = 1) = JS(S_1, S_2)$ ，然后带入 Chernoff 不等式可以证明
- 第 9 题：参考课本中的计算过程（注意例 4.10 的错误），容易出现计算错误
- 第 11 题：略