HW2 BLOOM FILTER

一、实验概述

本人编写了 bloom_filter 类进行测试，其中选择了输入集为 0—999999，测试集为 1000000—2999999。数据集取得非常大是为了尽可能获得趋于理论值的报错率。哈希函数采用了助教给出的 Murmurhash 类，生成 k 个不同哈希函数的方案是把哈希函数内置的 seed 参数设置为 1—k。

二、实验结果

下表与图是实验得到结果。其中图中虚线标注了由理论值公式 $k = {}^{m}/_{n} \cdot \ln 2$ 进行计算得到的 k 值。

表 1. 测试结果表

m/n	k	报错率
2	1	0.393233
2	2	0.399058
2	3	0.46862
2	4	0.559119
2	5	0.651818
3	1	0.283552
3	2	0.236801
3	3	0.252763
3	4	0.29435
3	5	0.351687
4	1	0.220889
4	2	0.154701
4	3	0.146875
4	4	0.159671
4	5	0.184731
5	1	0.181719
5	2	0.108541
5	3	0.09165
5	4	0.091965
5	5	0.100801

m/n	*k*	*k*=1	*k*=2	*k*=3	*k*=4	*k*=5
2	1.39	0.393	0.400			
3	2.08	0.283	0.237	0.253		
4	2.77	0.221	0.155	0.147	0.160	
5	3.46	0.181	0.109	0.092	0.092	0.101

图 1.理论 k 值

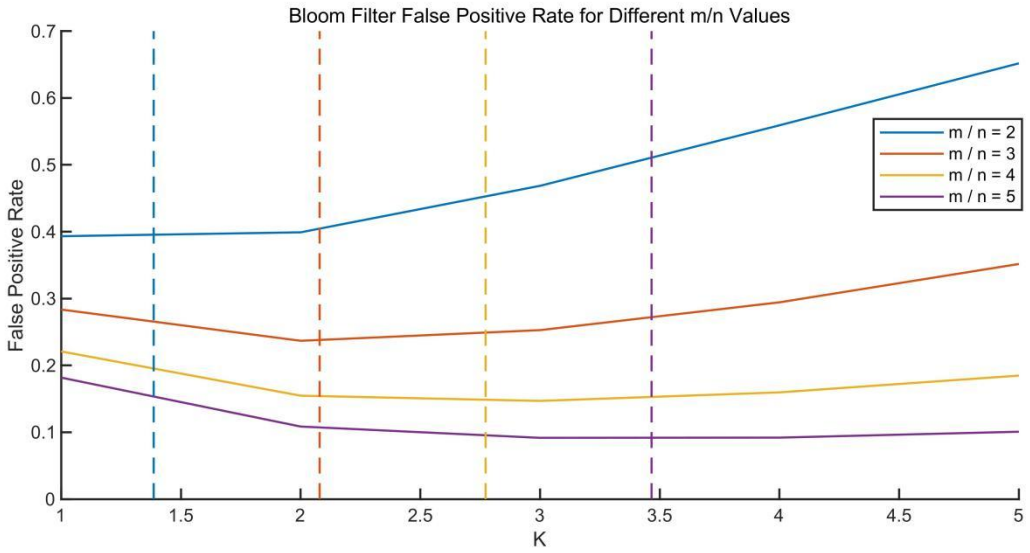

图 2. 测试结果曲线图

三、结果分析

可以发现表格中的数据与课件给出的理论值几乎一致，且图中的虚线表示的理论 k 值也与实验得出的最优 k 值符合。