数据科学与工程算法 2021-22 学年第一学期期中考试试卷

姓名：　　学号：

---

1. （5 分）假设一个分类算法的预测结果混淆矩阵如表 1 所示，计算预测正例的 F_1 值。

表 1：混淆矩阵示例

预测结果	真实结果	
	正例	反例
正例	20	10
反例	20	50

2. （5 分）假设总体有 10 个个体，按照 1～10 进行编号。使用循环等距抽样抽取容量 $n = 4$ 的样本，且第一个抽中的样本编号为 7，求其余 3 个被抽中样本的编号。

3. 假设抛一枚正面朝上概率为 0.6 的硬币 1000 次，随机变量 X 定义为正面朝上的次数。
   1) （5 分）运用 Chebyshev 不等式给出事件 $X > 900$ 的概率上界；
   2) （5 分）运用 Chernoff 不等式给出事件 $X < 300$ 的概率上界（表示为自然常数 e 的表达式）。

4. 假设一个布隆过滤器的容量为 10000 位，集合中有 2000 个元素。
   1) （5 分）计算使用 2 个哈希函数时的误判率（表示为自然常数 e 的表达式）；
   2) （5 分）计算使用 5 个哈希函数时的误判率（表示为自然常数 e 的表达式）。

5. 根据表 2 中的集合表示，回答以下问题。

表 2：集合表示

集合	0	1	2	3	4
S_1	0	1	1	0	1
S_2	0	0	1	1	0

   1) （5 分）计算集合 S_1 和 S_2 的 Jaccard 相似度；
   2) （5 分）使用哈希函数 $h_1(x) = (4x + 3) \bmod 5$，计算 S_1 和 S_2 的最小哈希签名值；
   3) （5 分）使用哈希函数 $h_2(x) = (7x + 1) \bmod 5$，计算 S_1 和 S_2 的最小哈希签名值。

6. （10 分）给定数据流 ⟨1,2,3,0,2,1⟩，哈希函数 $h_1(x) = (x + 2) \bmod 4$, $h_2(x) = (2x + 1) \bmod 4$, $h_3(x) = (3x + 1) \bmod 4$ 以及 $g_1(x) = \begin{cases} +1, \text{if } x \bmod 2 = 0 \\ -1, \text{if } x \bmod 2 = 1 \end{cases}$，$g_2(x) = \begin{cases} +1, \text{if } x \bmod 4 \geq 2 \\ -1, \text{if } x \bmod 4 < 2 \end{cases}$ 和 $g_3(x) = \begin{cases} +1, \text{if } x \bmod 3 = 0 \\ -1, \text{if } x \bmod 3 \neq 0 \end{cases}$，利用 Count Sketch 估计元素 0, 1, 2, 3 在数据流中的频度。

7. （10 分）给定数据流 ⟨1,2,3,0,2,1⟩，哈希函数 $h_1(x) = (x + 1) \bmod 3$, $h_2(x) = (2x + 3) \bmod 4$, 和 $h_3(x) = (3x + 2) \bmod 4$，利用 Count-Min Sketch 估计元素 0, 1, 2, 3 在数据流中的频度。

8. ~~（10 分）$\{X_i\}_{1 \leq i \leq n}$ 为指数分布 $P(X_i = x) = \lambda e^{-\lambda x}$ 的独立同分布样本，计算参数 λ 的最大似然估计。~~

9. （5 分）对于转移概率矩阵为 $P = \begin{pmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{pmatrix}$ 的马尔可夫链，求其平稳分布。

10. （20 分）写出图 1 和图 2 中马尔可夫链的各个状态的周期，判断其是否可约，是否是非周期的。

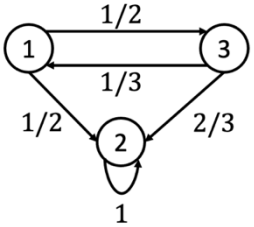

图1：马尔可夫链实例一

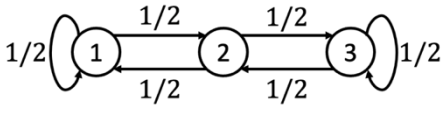

图2：马尔可夫链实例二