新媒体大数据分析 2024 秋　　第二次作业（书面作业）

---

1. 现有一组向量，$x_1 = (3,4)^T, x_2 = (5,6)^T, x_3 = (2,2)^T, x_4 = (8,4)^T$，分别求以下距离：

(a) x_1 和 x_2 的马氏距离（*Mahalanobis Distance*）

(b) x_2 和 x_3 的欧氏距离（*Euclidean Distance*）

(c) $p = 3$ 时 x_3 和 x_4 的明氏距离（*Minkowski Distance*）

2. 现有数据组 $A: \{1,2,3,4,5\}$ 和数据组 $B: \{1,3,5,7,9\}$，请分别计算 A 和 B 的以下相似性度量：

(a) 余弦相似度

(b) *Jaccard* 系数

3. 请对以下一组数据 $[10,20,30,40,50,60,70,80,90,100]$ 进行归一化操作，分别计算出其按以下方法归一化的结果：

(a) *min-max* 归一化

(b) *z-score* 归一化

4. 现有一组二维数据：$x_1 = (-1,-2)^T, x_2 = (-1,0)^T, x_3 = (0,0)^T, x_4 = (2,1)^T, x_5 = (0,1)^T$，使用主成分分析 (PCA) 将这组数据降至一维。

5. 路透社的 $811,400$ 份文档中 "*car*"、"*auto*"、"*insurance*" 和 "*best*" 这四个单词的频次以及这四个词在 ***Doc1***、***Doc2*** 和 ***Doc3*** 这 *3* 个文档的频次如下表所示。

Term	**df**	**tf@Doc1**	**tf@Doc2**	**tf@Doc3**
car	18,871	34	8	32
auto	3,597	3	24	0
insurance	19,167	0	51	6
best	40,014	18	0	13

(a) 计算关于这四个单词的三个文件的 *TF-IDF* 的值

(b) 试采用欧式归一化方法（即向量各元素平方和为 1），得到处理后的各文档向量化表示，其中每个向量为 4 维，每一维对应 1 个词项

(c) 基于 (b) 中得到的向量化表示，对于查询向量 $(1,0,1,0)^T$，计算 *3* 篇文档的得分并进行排序

员工编号	部门	工龄 (年)	是否有晋升	是否离职
1	销售	3	否	是
2	销售	6	是	是
3	技术	5	否	是
4	技术	7	是	否
5	人力资源	2	否	是
6	人力资源	8	是	否

6. 上图为某公司员工的部分离职情况:

(a) 计算目标变量“是否离职”的熵

(b) 计算特征“是否有晋升”对目标变量“是否离职”的信息增益

(c) 如果分割特征“工龄”, 最大化信息增益的分割点是多少年？

7. 假设我们有 *5* 个项目的真实相关性和两种算法的预测相关性排序如下:

项目	真实相关度	算法 1 排序	算法 2 排序
item1	3	1	2
item2	2	3	1
item3	1	2	3
item4	0	4	4
item5	4	5	5

请完成以下任务:

(a) 计算算法 *1* 和算法 *2* 的 *NDCG@5*, 哪个算法排序得更好？

(b) 计算算法 *1* 和算法 *2* 的 *Spearman Rank Correlation*, 哪个算法排序得更好？