数据挖掘与商务分析

第 1 次平时作业--参考答案

1. 数据规范化练习题：
某示例数据集记录了某单位成年人的身高和体重数据。身高范围为 1.4-1.9 m，体重范围为 40-90kg。请结合表 1，回答下列问题：

表 1. 体型数据表

ID	身高（m）	体重（kg）
1	1.70	50
2	1.60	50
3	1.70	60
4	1.65	45
……	……	……

1） 请采用 Min-Max 规范化，将表 1 显示用户的身高和体重数据规范化到[0,1].

解答：由于身高范围为 1.4-1.9 m，体重范围为 40-90kg，根据 Min-Max 规范化公式可以得到：

ID	规范化身高	规范化体重
1	0.6	0.2
2	0.4	0.2
3	0.6	0.4
4	0.5	0.1

2） 基于（1）中规范化的数据，分别计算用户 1（ID=1）与用户 2、3、4 的欧几里得距离。

解答：根据规范化的数据和相似性计算公式，得到：
distance(1,2)=0.2; distance(1,3)=0.2; distance(1,4)≈0.141;

3） 若要为用户 1 推荐“最相似用户”，你会推荐谁?

解答：根据 2），由于 distance(1,4)最小，因此推荐用户 4.

4） 请以上述“相似用户推荐”为例，说明数据规范化的必要性。

解答：若不进行规范化，身高和体重的数值量级差异较大，欧几里得距离计算中体重差异将主导结果，导致“相似性”判断失真。

2. 数据缺失值处理练习题：

1) 在例题 1 中若用户 1 (ID=1) 的体重值是缺失的，你会如何处理?

解答：首先判断数据缺失的原因。如果是完全随机缺失，且数据样本总量较大的情况下，后续数据分析可以忽略该个体数据；如果是非完全随机缺失，可以根据缺失的具体情况进行缺失值填补。

2) 如果允许在用户 2、3、4 之间使用最相似的两名用户体重均值进行缺失值处理，请问填充后用户 1 的体重是多少?

解答：由于用户 1 的体重是缺失的，只能根据身高判断用户间的相似性，从给出的数据来看，与用户 1 在身高上最相似的是用户 3（规范化后的身高为 0.6），其次是用户 4（规范化后的身高是 0.5），因此可以将两者的体重的均值—52.5 公斤来填充用户 1 的体重。

3) 考虑到用户之间的欧几里得距离，如果使用欧几里得距离倒数作为权重加权，请问填充后用户 1 的体重是多少?

解答：根据规范化后身高的取值，我们可以发现用户 1 和用户 2，3，4 的距离为分别为：0.2，0，0.1。因此，可以直接用用户 3 的体重 60 公斤进行替代便可。

3. 数据离散化练习题：

分别利用等宽离散化和等深离散化将表 2 中的属性“年收入(万元)”转换为“低收入”、“中收入”和“高收入”三档。

表 2. 个人信息表

ID	年龄(岁)	性别	年收入(万元)
1	25	男	10
2	27	女	25
3	30	男	30
4	45	女	60
5	28	男	40
6	32	男	20
7	52	男	50
8	35	女	30

9	55	男	100
10	48	女	120

解答：等宽离散化将取值范围划分为等宽的若干区间，其中区间长度为（120-10）/3≈36.67，因此{1,2,3,5,6,8}被划入低收入区间，{4,7}划入中收入区间，{9,10}划入高收入区间。

等深离散化则要求将样本按收入值从小到大排序后，尽量让每个区间包含相同数量的样本。共有 10 个样本，分为三个区间，每个区间 3-4 人。将年收入按照从小到大排序得到：10，20，25，30，30，40，50，60，100，120。因此根据该规则 {1,2,6} 被划入低收入区间; {3,5,8}被划入中收入区间，{4,7,9,10}被划入高收入区间。

作业说明：

1. 本次作业不要求编程，给出计算过程或相应的文字表述；
2. 所有题目请独立完成，如果发现雷同或者近似雷同作业，记 0 分；
3. 本次作业截止时间：2025/09/30　23:59:59