hw3

PB21051012

1.1

(1) 用户打分平均值

$\bar{r}_1 = 1.333$ $\quad \bar{r}_6 = 3.5$ $\quad \bar{r}_{11} = 3$

$\bar{r}_2 = 3$ $\quad \bar{r}_7 = 3.5$ $\quad \bar{r}_{12} = 4$

$\bar{r}_3 = 3.8$ $\quad \bar{r}_8 = 3$

$\bar{r}_4 = 2.667$ $\quad \bar{r}_9 = 4.5$

$\bar{r}_5 = 3.5$ $\quad \bar{r}_{10} = 2.5$

$$sim(1,5) = \frac{(2-4/3)(2-3.5) + (1-4/3)(3-3.5)}{\sqrt{(\frac{1}{3})^2+(\frac{1}{3})^2+(\frac{2}{3})^2}\sqrt{1.5^2+0.5^2+0.5^2+1.5^2}} = -0.456$$

$$sim(2,5) = \frac{(4-3)(2-3.5) + (2-3)(5-3.5)}{\sqrt{1^2\times 2}\sqrt{1.5^2+0.5^2+0.5^2+1.5^2}} = -0.949$$

$$sim(3,5) = \frac{(4-3.8)(5-3.5) + (4-3.8)(4-3.5) + (3-3.8)(3-3.5)}{\sqrt{0.8^2+1.2^2+0.2^2+0.2^2+0.8^2}\sqrt{1.5^2+1.5^2+0.5^2+0.5^2}} = 0.214$$

$$sim(4,5) = \frac{(1-8/3)(2-3.5) + (3-8/3)(4-3.5)}{\sqrt{(4/3)^2+(5/3)^2+(1/3)^2}\sqrt{1.5^2+1.5^2+0.5^2+0.5^2}} = 0.552$$

$$sim(6,5) = \frac{(2-3.5)(4-3.5)}{\sqrt{1.5^2+1.5^2}\sqrt{1.5^2+1.5^2+0.5^2+0.5^2}} = -0.158$$

$$sim(7,5) = \frac{(2-3.5)(3-3.5)}{\sqrt{0.5^2+0.5^2}\sqrt{1.5^2+1.5^2+0.5^2+0.5^2}} = 0.474$$

$$sim(8,5) = \frac{(4-3)(5-3.5) + (2-3)(3-3.5)}{\sqrt{1^2+1^2}\sqrt{1.5^2+1.5^2+0.5^2+0.5^2}} = 0.632$$

$$sim(9,5) = \frac{(4-4.5)(2-3.5)}{\sqrt{0.5^2+0.5^2}\sqrt{1.5^2+1.5^2+0.5^2+0.5^2}} = 0.474$$

$$sim(10,5) = \frac{(3-2.5)(2-3.5)}{\sqrt{0.5^2+0.5^2}\sqrt{1.5^2+1.5^2+0.5^2+0.5^2}} = -0.474$$

$$sim(11,5) = \frac{(5-3)(2-3.5) + (2-3)(5-3.5) + (2-3)(4-3.5) + (4-3)(3-3.5)}{\sqrt{1^2+1^2+2^2+2^2+1^2+1^2}\sqrt{1.5^2+1.5^2+0.5^2+0.5^2}} = -0.71$$

$$sim(12,5) = \frac{(5-4)(4-3.5)}{\sqrt{1^2+1^2}\sqrt{1.5^2+1.5^2+0.5^2+0.5^2}} = 0.158$$

$$Pred = 3.5 + \frac{0.214\times(3-3.8) + 0.474\times(5-4.5)}{0.214 + 0.474} = 3.596$$

2023.12.23 19:52

1.2

中国科学技术大学
University of Science and Technology of China
地址：中国 安徽 合肥市金寨路96号 邮编：230026
电话：0551-63602184 传真：0551-63631760 Http://www.ustc.edu.cn

Sample	feature1	feature2
x_1	4	1
x_2	2	3
x_3	5	4
x_4	1	0

feature1: $\mu_1 = 3$

feature2: $\mu_2 = 2$

减去均值: $X = \begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 2 & 2 \\ -2 & -2 \end{bmatrix}$

协方差矩阵: $C = \frac{1}{N-1} X^T X = \frac{1}{3}\begin{pmatrix} 10 & 6 \\ 6 & 10 \end{pmatrix} = \begin{pmatrix} \frac{10}{3} & 2 \\ 2 & \frac{10}{3} \end{pmatrix}$

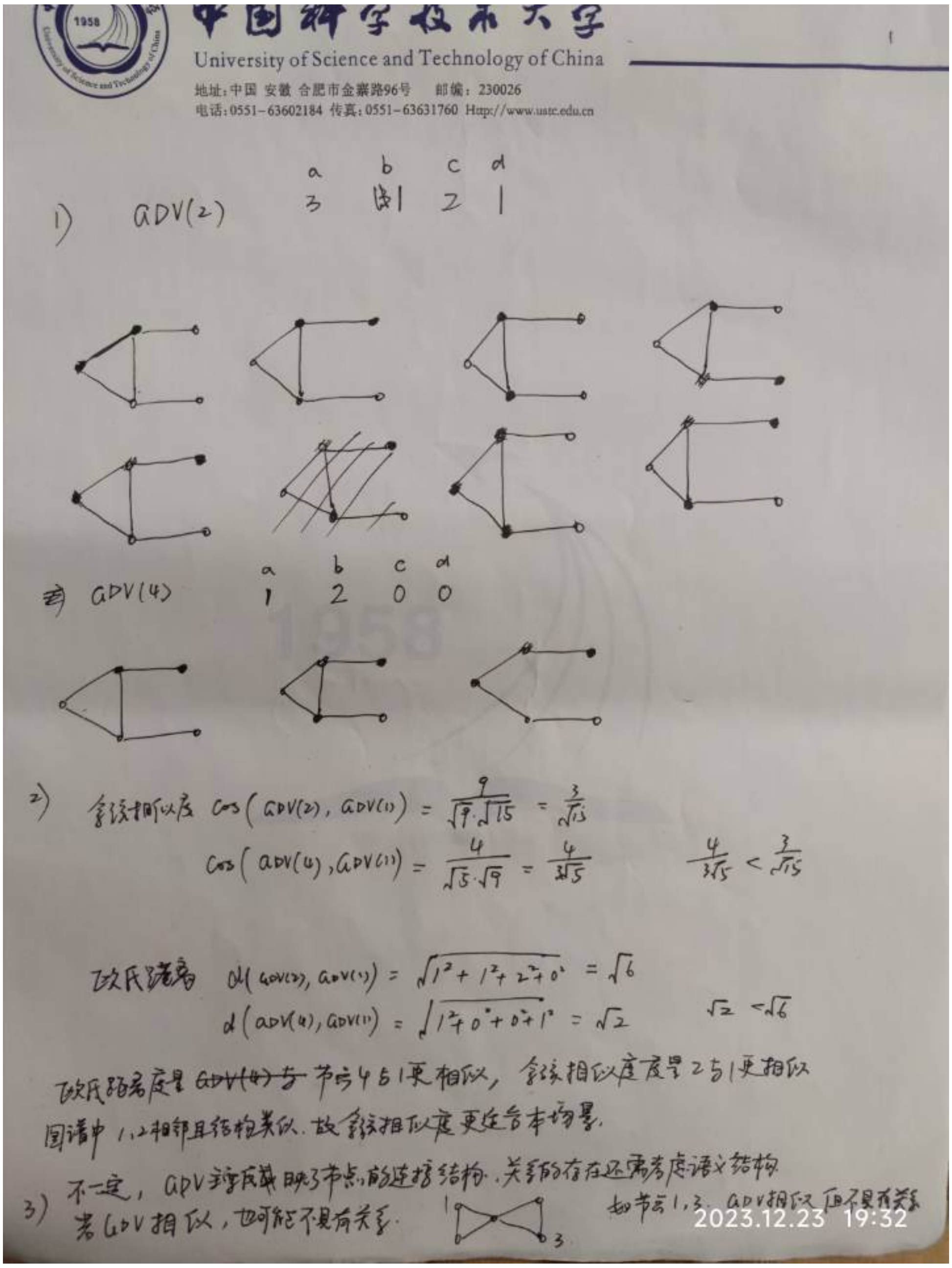
$|\lambda E - C| = \begin{vmatrix} \lambda - \frac{10}{3} & -2 \\ -2 & \lambda - \frac{10}{3} \end{vmatrix} = (\lambda - \frac{10}{3})^2 - 4 = 0 \Rightarrow \lambda = \frac{4}{3} / \frac{16}{3}$

$\begin{pmatrix} \frac{10}{3} & 2 \\ 2 & \frac{10}{3} \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = \frac{16}{3}\begin{pmatrix} a \\ b \end{pmatrix} \Rightarrow a = b$

单位特征向量 $w_1 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$

$X' = X \cdot w_1 = \begin{pmatrix} 0 \\ 0 \\ 2\sqrt{2} \\ 2\sqrt{2} \end{pmatrix}$

1.3



中国科学技术大学
University of Science and Technology of China
地址：中国 安徽 合肥市金寨路96号　邮编：230026
电话：0551-63602184 传真：0551-63631760 Http://www.ustc.edu.cn

1) GDV(2)

a	b	c	d
3	1	2	1

GDV(4)

a	b	c	d
1	2	0	0

2) 余弦相似度 $\cos(GDV(2), GDV(1)) = \frac{9}{\sqrt{9}\cdot\sqrt{15}} = \frac{3}{\sqrt{15}}$

$\cos(GDV(4), GDV(1)) = \frac{4}{\sqrt{5}\cdot\sqrt{9}} = \frac{4}{3\sqrt{5}}$　　$\frac{4}{3\sqrt{5}} < \frac{3}{\sqrt{15}}$

欧氏距离 $d(GDV(2), GDV(1)) = \sqrt{1^2+1^2+2^2+0^2} = \sqrt{6}$

$d(GDV(4), GDV(1)) = \sqrt{1^2+0^2+0^2+1^2} = \sqrt{2}$　　$\sqrt{2} < \sqrt{6}$

欧氏距离度量 ~~GDV(4)与~~ 节点4与1更相似，余弦相似度度量2与1更相似
图谱中1,2相邻且结构类似，故余弦相似度更适合本场景.

3) 不一定，GDV主要反映了节点的连接结构，关系的存在还需考虑语义结构
若GDV相似，也可能不是有关系. 　如节点1,3 GDV相似 但不是有关系

2023.12.23 19:32

2.1

基于用户的协同过滤由于物品数量远大于评论数量，矩阵过于稀疏不太适合
基于内容的推荐由于针对某一作品的评论难以全面反应用户的喜好，不太合适
基于项目的协同过滤由于项目的稳定性，可以通过此方法进行协同过滤，故更合适

2.2

引入动态注意力机制，使模型能够在每一时刻关注对当前话题最重要的部分。这有助于模型在处理交错话题时更灵活地适应不同的关注点。
考虑使用场景切换检测方法，以检测对话中不同话题之间的切换点。这可以是基于阈值的方法，也可以是监督学习的方法，如使用深度学习模型来学习场景切换的模式。
实现对话状态追踪，理解对话者的意图，即使在多个话题之间切换时也能够保持上下文的一致性。

2.3

对于一组给定的数据 $v_1, \ldots, v_n, v_i \in \mathbb{R}^m$ ，中心化后 $x_1, x_2, \ldots, x_n = v_1 - \mu, v_2 - \mu, \ldots, v_n - \mu$ ，其中 $\mu = \frac{1}{n}\sum_{i=1}^{n} v_i$ 。

由于向量内积的几何意义可以表示为第一个向量投影到第二个向量的长度，因此向量 x_i 在 w （单位方向向量）上的投影坐标可以表示为 $\langle x_i, w\rangle = x_i^T w$ 。所以目标是找到一个投影方向 w ，使得 $x_1, x_2, \ldots, x_n$ 在 w 上的投影长度（方差）尽可能大。

由于中心化后，投影的均值为 0 ，（因为中心化后 $\mu' = \frac{1}{n}\sum_{i=1}^{n} x_i^T w = (\frac{1}{n}\sum_{i=1}^{n} x_i^T)w = (\frac{1}{n}\sum_{i=1}^{n}(v_i - \mu)^T)w = 0$ ），因此投影后的方差就可以写为

$$
\begin{aligned}
D(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i^T w)^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i^T w)^T(x_i^T w) \\
&= \frac{1}{n}\sum_{i=1}^{n} w^T x_i x_i^T w \\
&= w^T(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T)w
\end{aligned}
$$

其中 $\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$ 如果熟悉概率论的话，不难看出是样本的协方差矩阵；我们将其写作 Σ 。w 是单位向量，所以 $w^T w = 1$ 。因此我们要求最大化投影方差，可以表示为

$$
\begin{aligned}
&\max w^T \Sigma w \\
&\mathrm{s.t.} w^T w = 1
\end{aligned}
$$

求解该优化问题，我们使用拉格朗日乘子法，

$$
\begin{aligned}
&F(w) = w^T \Sigma w - \lambda(w^T w - 1) \\
&令 \frac{\partial F}{\partial w} = 0, \\
&得 \Sigma w - \lambda w = 0
\end{aligned}
$$

此时$D(x) = \lambda$。据线性代数的知识，投影后的方差就是协方差矩阵的特征值，要求的最大化方差也就是协方差矩阵的最大特征值，最佳投影方向就是最大特征值对应的特征向量。

2.4

考虑多模态属性的时效性。在一些场景中，实体的属性可能会随着时间变化。
选择包含更多信息、更丰富多样性的多模态属性。这可以通过分析属性的内容、多模态数据的质量以及数据的多样性来确定。
选择那些与实体关联度高、最具代表性的多模态属性。例如，对于一个人物实体，可能有多张照片，但选择与实体最相关、最能代表其外貌的照片是比较合理的。
考虑下游任务的导向性。如果知识图谱用于某一具体的任务，比如人脸识别，那么选择最具代表性和辨识度的人脸图像可能是关键的。

2.5

独立级联模型（ICM）下的证明：

考虑集合覆盖问题，该问题是 NP-Hard 的。给定一个集合族以及目标集合，找到最小数量的集合，使得它们的并集包含目标集合。现在我们将集合覆盖问题规约到 IC 模型的信息传播问题上。给定一个集合覆盖问题实例：集合族 {S1, S2, ..., Sn} 和目标集合 U，我们可以构建一个对应的 IC 模型网络。

- 每个集合 Si 对应网络中的一个节点 Vi。
- 每个元素 e 属于目标集合 U 对应网络中的一个节点 Ue。

构建边的规则如下：

- 如果元素 e 属于集合 Si，则存在一条边连接节点 Ue 和节点 Vi。

定义激活概率：

- 对于每个节点 Vi，设置其激活概率为 p，其中 p 是集合 Si 的代价除以集合 Si 中包含的元素数量。

目标是在这个网络中找到最小数量的节点，使得它们的激活能够覆盖整个目标集合 U。

如果我们能够找到 IC 模型中的最大激活集合，它所包含的节点数正好就是集合覆盖问题的最小集合数量。因此，IC 模型信息传播最大化问题是 NP-Hard 的。

线性阈值模型（LTM）下的证明：

考虑加权集合覆盖问题，该问题是 NP-Hard 的。给定一个加权集合族以及目标集合，找到最小权重的集合，使得它们的并集包含目标集合。现在我们将加权集合覆盖问题规约到 LT 模型的信息传播问题上。

给定一个加权集合覆盖问题实例：加权集合族 {S1, S2, ..., Sn}，每个集合 Si 都有一个权重 w(Si)，以及目标集合 U，我们可以构建一个对应的 LT 模型网络。

- 每个集合 Si 对应网络中的一个节点 Vi。
- 每个元素 e 属于目标集合 U 对应网络中的一个节点 Ue。

构建边的规则如下：

- 如果元素 e 属于集合 Si，则存在一条带有权重 w(Si) 的边连接节点 Ue 和节点 Vi。

定义阈值：

- 对于每个节点 Vi，设置其阈值为 t，其中 t 是集合 Si 的权重除以集合 Si 中包含的元素数量。

目标是在这个网络中找到最小权重的节点，使得它们的激活能够覆盖整个目标集合 U。

如果我们能够找到 LT 模型中的最小激活权重集合，它的总权重正好就是加权集合覆盖问题的最小权重。因此，LT 模型信息传播最大化问题是 NP-Hard 的