Web 信息处理与应用作业 3

傅申　PB20000051

1. 计算题

1.1. 在课件中, 我们给出了如下评分矩阵:

		Users											
		1	2	3	4	5	6	7	8	9	10	11	12
Movies	1	1		3		?	5			5		4	
	2			5	4			4			2	1	3
	3	2	4		1	2		3		4	3	5	
	4		2	4		5			4			2	
	5			4	3	4	2					2	5
	6	1		3		3			2			4	

采用基于用户 (User-based) 的评分预测方法 (同样采用 2–最近邻), 预测用户 5 对于电影 1 的评分, 并与课件中给出的基于物品的评分结果进行比较. 需写出详细计算过程.

首先, 计算各个对电影 1 有评分的用户 (1, 3, 6, 9 和 11) 和用户 5 打分的平均值:

$$\bar{r}_1 = \frac{4}{3}, \bar{r}_3 = \frac{19}{5}, \bar{r}_5 = \frac{7}{2}, \bar{r}_6 = \frac{7}{2}, \bar{r}_9 = \frac{9}{2}, \bar{r}_{11} = 3$$

然后计算他们与用户 5 之间的相似度:

$$\mathrm{sim}(1,5) = \frac{\frac{2}{3}\times -\frac{3}{2} + -\frac{1}{3}\times -\frac{1}{2}}{\frac{1}{3}\sqrt{1+4+1}\times\frac{1}{2}\sqrt{9+9+1+1}} = -\frac{1}{2}\sqrt{\frac{5}{6}} \approx -0.456$$

$$\mathrm{sim}(3,5) = \frac{\frac{1}{5}\times\frac{3}{2} + \frac{1}{5}\times\frac{1}{2} + -\frac{4}{5}\times -\frac{1}{2}}{\frac{1}{5}\sqrt{16+36+1+1+16}\times\sqrt{5}} = \frac{4}{5\sqrt{14}} \approx 0.214$$

$$\mathrm{sim}(6,5) = \frac{-\frac{3}{2}\times\frac{1}{2}}{\frac{1}{2}\sqrt{9+9}\times\sqrt{5}} = -\frac{1}{2\sqrt{10}} \approx -0.158$$

$$\mathrm{sim}(9,5) = \frac{-\frac{1}{2}\times -\frac{3}{2}}{\frac{1}{2}\sqrt{1+1}\times\sqrt{5}} = \frac{3}{2\sqrt{10}} \approx 0.474$$

$$\mathrm{sim}(11,5) = \frac{2\times -\frac{3}{2} - 1\times\frac{3}{2} - 1\times\frac{1}{2} + 1\times -\frac{1}{2}}{\sqrt{1+4+4+1+1+1}\times\sqrt{5}} = -\frac{11}{4\sqrt{15}} \approx -0.710$$

因此, 用户 5 的 2–最近邻为用户 3 和用户 9, 估计用户 5 对电影 1 的评分为:

$$\mathrm{Pred}(5,1) = \frac{7}{2} + \frac{-\frac{4}{5}\,\mathrm{sim}(3,5) + \frac{1}{2}\,\mathrm{sim}(9,5)}{\mathrm{sim}(3,5) + \mathrm{sim}(9,5)} \approx 3.596$$

基于物品的评分结果为 2.6, 因此基于用户的评分高于基于物品的评分.

1.2. 给定 4 个二维向量: $x_1 = (4, 1), x_2 = (2, 3), x_3 = (5, 4), x_4 = (1, 0)$.
现需要使用主成分分析将特征空间降为一维. 请使用主成分分析法计算其主成分, 并计算降维后每个样本点的新特征表示. (提示: 可使用 $(X - \bar{X})^{\mathrm{T}}(X - \bar{X})$ 计算协方差矩阵, 其两个特征值分别为 16 与 4)

因为样本的均值为 $\bar{x} = (3, 2)$, 所以对样本标准化后得到:

$$\boldsymbol{X} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \\ 2 & 2 \\ -2 & -2 \end{pmatrix}$$

其协方差矩阵为:

$$\boldsymbol{C} = \begin{pmatrix} \mathrm{cov}(x^{(1)}, x^{(1)}) & \mathrm{cov}(x^{(1)}, x^{(2)}) \\ \mathrm{cov}(x^{(2)}, x^{(1)}) & \mathrm{cov}(x^{(2)}, x^{(2)}) \end{pmatrix} = \frac{1}{3}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \frac{1}{3}\begin{pmatrix} 10 & 6 \\ 6 & 10 \end{pmatrix}$$

求协方差矩阵的特征值:

$$\det(\boldsymbol{C} - \lambda\boldsymbol{I}) = 0 \Leftrightarrow \begin{vmatrix} 10 - 3\lambda & 6 \\ 6 & 10 - 3\lambda \end{vmatrix} = (10 - 3\lambda)^2 - 36 = 0 \Leftrightarrow \lambda_1 = \frac{4}{3}, \lambda_2 = \frac{16}{3}$$

两个特征值对应的单位特征向量分别为:

$$\boldsymbol{w}_1 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right), \boldsymbol{w}_2 = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)$$

因为要降到一维, 所以取 $\boldsymbol{w}_2$ 作为主成分, 降维后的样本为:

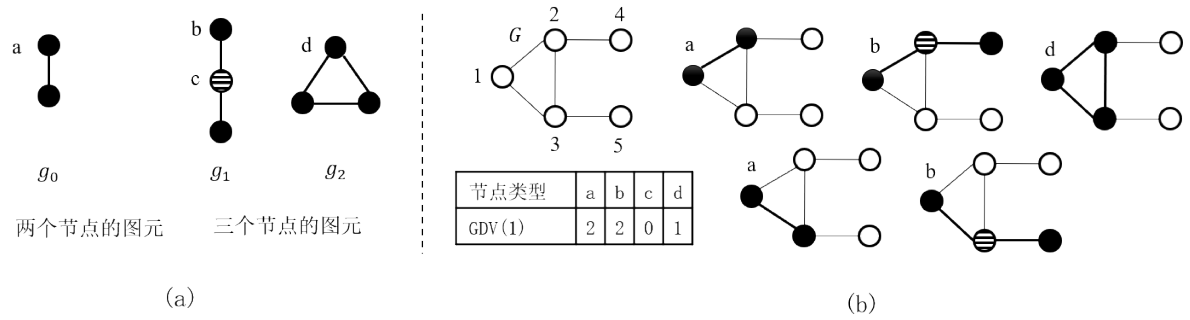
$$\boldsymbol{W}^* = (\boldsymbol{w}_2) = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)^{\mathrm{T}}$$

$$\boldsymbol{X}' = \boldsymbol{X}\boldsymbol{W}^* = \begin{pmatrix} 1 & -1 \\ -1 & 1 \\ 2 & 2 \\ -2 & -2 \end{pmatrix}\begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} \sqrt{2} \\ -\sqrt{2} \\ 0 \\ 0 \end{pmatrix}$$

1.3. 在知识图谱中, 属性对于描述实体及其之间的关系有着重要的作用. 由于属性一般表示为边的结构, 因此, 采用局部子图特征描述实体成为一种常见的思路. 其中, 图元 (Graphlets) 是一种局部子图特征描述的启发式方法, 它考虑了不同数目结点构成子图的同构性, 同时对于节点的类型进行了区分.

如下图 (a) 所示, 其中字母 a, b, c, d 分别表示不同的节点类型. 显然, 由两个节点构成的图元只有 g_0, 由于两个节点的等价性, 因此只包含一种类型的结点 a. 而由三个节点构成的图元包括两种结构 g_1 和 g_2. 其中, g_1 包含两种不同类型的结点, 分别为位于中心的 c 和位于两侧的 b. 而与 g_0 类似, g_2 只包含一种类型的结点 d.

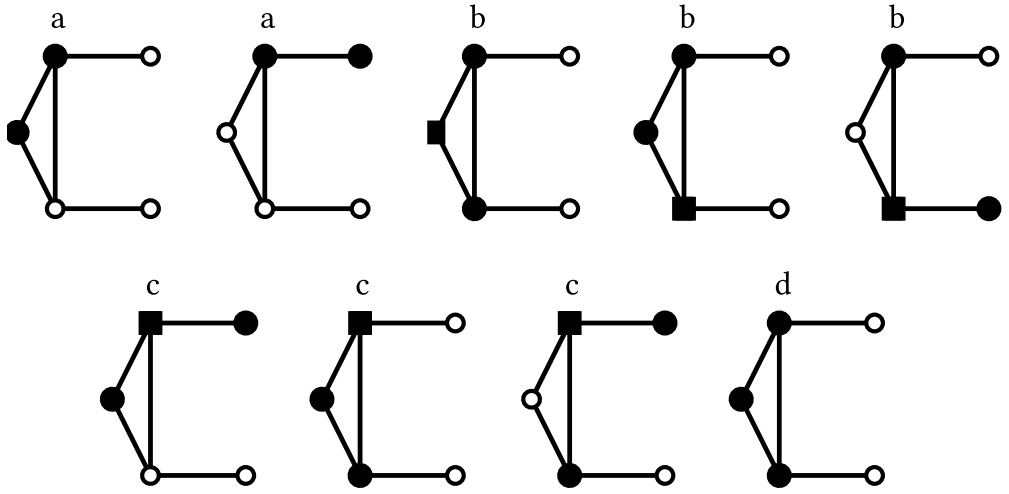
在图元定义的基础之上, 我们可以通过图元度向量 (Graphlet Degree Vector, 简称 GDV) 来描述特定实体的局部子图特征. 如下图 (b) 所示, 在只考虑两个和三个节点组成的图元的情况下, 对于节点 1 而言, 根据它在不同图元中具有不同类型的出现次数, 可知节点 1 对应的 GDV(1) = [2, 2, 0, 1].



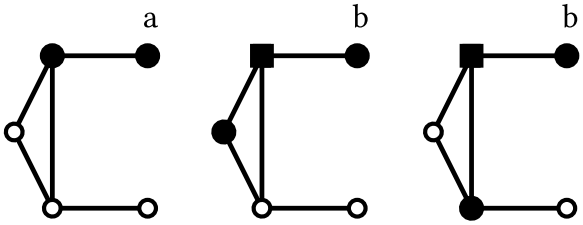
基于上述描述, 回答以下问题:

1) 试写出 GDV(2) 和 GDV(4), 并给出必要的图示.
2) 请根据欧氏距离和余弦相似度, 基于 GDV 分析节点 1 的特征与节点 2, 节点 4 中哪个更相似, 并分析哪一种相似度量更适合本场景, 为什么?
3) 结合知识图谱概念部分内容, 试讨论具有相似 GDV 描述的两个实体是否一定具有某种关系并构成三元组, 为什么? (问答题, 言之有理即可)

1) GDV(2) = [2, 3, 3, 1], 如下图所示



GDV(4) = [1, 2, 0, 0], 如下图所示



2) 首先计算欧式距离:

$$d_{\text{Euclidean}}(1,2) = \|\text{GDV}(1) - \text{GDV}(2)\|_2 = \sqrt{11}$$
$$d_{\text{Euclidean}}(1,4) = \|\text{GDV}(1) - \text{GDV}(4)\|_2 = \sqrt{2}$$

因为 $d_{\text{Euclidean}}(1,2) > d_{\text{Euclidean}}(1,4)$, 所以, 在欧氏距离度量下, 可以认为节点 1 的特征与节点 4 更相似.

然后计算余弦相似度:

$$\cos(\theta_{1,2}) = \frac{\text{GDV}(1) \cdot \text{GDV}(2)}{\|\text{GDV}(1)\|_2 \|\text{GDV}(2)\|_2} = \frac{11}{3\sqrt{23}}$$
$$\cos(\theta_{1,4}) = \frac{\text{GDV}(1) \cdot \text{GDV}(4)}{\|\text{GDV}(1)\|_2 \|\text{GDV}(4)\|_2} = \frac{6}{\sqrt{115}}$$

因为 $\cos(\theta_{1,2}) > \cos(\theta_{1,4})$, 所以, 在余弦相似度度量下, 可以认为节点 1 的特征与节点 2 更相似.

余弦相似度度量更适合本场景. 在欧氏距离度量下, 对于两个 GDV 分布类似的节点, 如果它们的 GDV 模长差异很大, 得到的欧氏距离也会很大, 而余弦相似度就不会受到模长影响. 比如题中的节点 2, 它与节点 1 之间有更多的相同属性, 但是在欧式距离度量下相似度却很低.

3) 具有相似 GDV 描述的两个实体并不一定具有某种类似的关系或构成三元组, 因为 GDV 主要反映了局部结构的相似性, 而知识图谱中的关系通常涉及全局信息和语义含义. 两个实体可能在局部结构上相似, 但其全局语义关联可能不同.

2. 问答题 (言之有理即可)

2.1. 你需要为一个新上线的在线电影平台设计一个推荐系统. 这个平台包含了超过一百万部电影, 但只有约一万条评分记录. 你需要从基于用户 (User-based) 的协同过滤、基于项目 (Item-based) 的协同过滤和基于内容 (Content-based) 的推荐三种方法中选择一种, 请问哪一种方法最合适, 为什么? 该方法在应用中存在何种问题, 如何解决?

基于内容的推荐方法最为合适, 因为在该情况下, 评分记录相比于电影数量过少, 数据过于稀疏, 无法有效地利用基于用户和基于项目的协同过滤方法.

这种方法也存在一些问题, 比如:

- 需要对电影和评价进行分析, 以选择合适的特征. 这需要相关邻域的专业知识和对用户偏好的深入理解.
- 该方法只能根据用户现有兴趣进行推荐, 无法发现用户潜在的兴趣.
- 存在用户冷启动问题, 即无法为新用户进行推荐.

2.2. 与搜索任务类似, 对话理解任务同样往往需要借助上下文信息来准确理解对话者的真实意图. 然而, 对话理解任务往往存在一类特殊场景, 即对话双方同时对两个乃至多个话题展开讨论. 从而导致了对话主题上下文情境的 "交错" 现象. 在这一场景下, 传统依赖时间等切分会话的方法都会失灵. 那么, 如何在话题 "交错" 情况下有效切分和获取上下文信息, 并确保准确理解对话者意图呢?

- 可以引入话题建模方法来识别对话中的主题, 并根据主题切分对话.
- 使用注意力机制, 确保模型在理解和生成响应时能够集中关注对话中最相关的部分. 这有助于应对话题交错情况, 使模型更加灵活地处理上下文信息.
- 采用多任务学习的方法, 同时处理多个话题或任务, 促使模型更全面地理解对话, 以更好地应对话题交错的情况.

2.3. 主成分分析的一个重要前提是: 最大特征值对应的特征向量可以最大化投影方差. 为什么这一前提是成立的?

对于 m 维随机变量 $\boldsymbol{x}$, 设其协方差矩阵为 Σ. 假设其主成分为 $y = \boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{x}$, 则其方差为

$$\mathrm{var}(y) = \boldsymbol{\alpha}^{\mathrm{T}}\Sigma\boldsymbol{\alpha}$$

要使这个方差达到最大, 等价于求解约束优化问题

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} \boldsymbol{\alpha}^{\mathrm{T}}\Sigma\boldsymbol{\alpha} \\ &\text{s.t. } \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha} = 1 \end{aligned}$$

定义拉格朗日函数

$$L(\boldsymbol{\alpha}_1, \lambda) = \boldsymbol{\alpha}^{\mathrm{T}}\Sigma\boldsymbol{\alpha} - \lambda\left(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha} - 1\right)$$

其中, λ 是拉格朗日乘子. 将拉格朗日函数对 $\boldsymbol{\alpha}$ 求导, 得到

$$\frac{\partial L^{'}}{\partial \boldsymbol{\alpha}} = 2\Sigma\boldsymbol{\alpha} - 2\lambda\boldsymbol{\alpha} = 0$$

因此 λ 是 Σ 的特征值, $\boldsymbol{\alpha}$ 是对应的单位特征向量. 方差为

$$\mathrm{var}(y) = \boldsymbol{\alpha}^{\mathrm{T}}\Sigma\boldsymbol{\alpha} = \boldsymbol{\alpha}^{\mathrm{T}}\lambda\boldsymbol{\alpha} = \lambda$$

因此, 最大特征值对应的特征向量可以最大化投影方差.

2.4. 属性多模态知识图谱往往在传统知识图谱的基础上直接为实体添加多模态属性 (如图片). 然而, 多模态属性丰富多样的表达形式带来了如何挑选一条最合适的多模态属性 (例如, 一张最具有代表性的照片) 的问题. 请结合知识图谱的概念或下游任务, 谈谈你认为合理的挑选多模态属性的原则.

我认为合理的挑选原则如下:

- 选择最具有代表性和信息丰富性的多模态属性, 以确保所选属性能够充分反映实体的特征.
- 选择与实体关联强烈的多模态属性, 以确保属性能够与实体的概念紧密相连.
- 多模态属性之间应该一致, 确保它们在表达实体特征时相互协调.
- 考虑知识图谱的具体应用需求, 选择适用于下游任务的多模态属性.

2.5. 试证明在独立级联 (Independent Cascade Model) 模型和线性阈值 (Linear Threshold) 模型条件下, 信息传播最大化问题是 NP-Hard 的. (两个模型对应的规约情况都要证明)

对于独立级联模型

考虑集合覆盖判定问题: 给定一组子集 $\mathcal{S} = \{S_1, S_2, ..., S_m\}$ 和全集 $U = \{u_1, u_2, ..., u_n\}$, 判定是否有 k 个子集的并集等于 U (可以假设 $k < n < m$).

给定任意集合覆盖判定问题的实例, 可以定义一个对应的有向二分图, 其包含 $n + m$ 个节点: m 个表示子集 S_i 的节点和 n 个表示全集 U 的节点. 如果 $u_j \in S_i$, 则有从 u_j 指向 S_i 的边, 其激活概率为 1. 这样, 这个集合覆盖判定问题就被规约为了一个信息传播最大化问题: 判定是否存在 k 个节点 (记为 A), 使得 $f(A) \geq n + k$.

因为集合覆盖判定问题是 NP-Complete 的, 所以信息传播最大化问题是 NP-Hard 的.

对于线性阈值模型

考虑顶点覆盖问题: 给定一个 n 个顶点的无向图 $G = (V, E)$ 和整数 k, 判定是否存在 k 个顶点的集合 S, 使得 G 中的所有边都至少有一个端点在 S 中.

通过将 G 中的所有无向边转化为有向双向边, 可以把顶点覆盖问题规约为信息传播最大化问题: 如果 G 中存在 k 个顶点的覆盖 S, 那么可以断定 $\max(f(S)) = n$. 反之, 如果 $\max(f(S)) = n$, 则 S 是 G 的顶点覆盖.

因为顶点覆盖问题是 NP-Complete 的, 所以信息传播最大化问题是 NP-Hard 的.