HXH 2022.10.28

1. Homogeneous Network Embedding 介绍word2vec应用在同质图

给定一个图 $G=(V,E)$

条件独立性假设

目标： $\arg\max_{\theta} \prod_{v\in V} \prod_{c\in N(v)} p(c|v;\theta)$ (1)

计算时，把相乘变成转换成对数函数：

$\arg\max_{\theta} \sum_{v\in V} \sum_{c\in N(v)} \log p(c|v;\theta)$

$N(v)$ 表示节点v的邻域

$p(c|v;\theta)$ 表示给定v时，有c的条件概率

2. Heterogeneous Network Embedding (metapath2vec) 处理异质图

① Heterogeneous Skip-gram

因为考虑了节点类型

给定：$G=(V,E,T)$，with $|T_V|>1$

$V \to T_V$，$E \to T_E$，$|T_V|+|T_E|>2$

最大化存在异质上下文 $N_t(v)$ 的概率：

$$\arg\max_{\theta} \sum_{v\in V} \sum_{t\in T_V} \sum_{c_t\in N_t(v)} \log p(c_t|v;\theta) \quad (2)$$

类型

"属于该类型的邻居(所以说是Heterogeneous)"

$N_t(v)$： v的邻居，节点类型为第t种

$$p(c_t|v;\theta) = \frac{e^{X_{c_t}\cdot X_v}}{\sum_{u\in V} e^{X_u\cdot X_v}}$$

softmax

注意：这里未考虑节点类型 $(t\in T_V)$

X_v： 矩阵X的第v行，表示节点v的向量表示。

(注意：不同类型的节点将会嵌入到同一维度空间的向量)

② 负采样

给定一个负采样样本量 M

详细理解见负采样笔记

(2)式变成：$\log\sigma(X_{c_t}\cdot X_v) + \sum_{m=1}^{M} E_{u^m\sim p(u)}[\log\sigma(-X_{u^m}\cdot X_v)]$

其中 $\sigma(x) = \frac{1}{1+e^{-x}}$

$p(u)$：预先定义的分布

3. 基于元路径的随机游走 能够确保不同类型的节点之间的关系嵌入至 skip-gram 中，建立每个节点的异构邻域

给定异质网络：$G=(V,E,T)$

元路径框架：$\mathcal{P}: V_1 \xrightarrow{R_1} V_2 \xrightarrow{R_2} \cdots V_t \xrightarrow{R_t} V_{t+1} \cdots \xrightarrow{R_{l-1}} V_l$ V_1 与 V_l 同一种类型

第 t 步的转移概率被定义为：

$$p(v^{i+1}|v^i,\mathcal{P}) = \begin{cases} \frac{1}{|N_{t+1}(v_t^i)|} & (v^{i+1},v_t^i)\in E,\ \phi(v^{i+1})=t+1 \\ 0 & (v^{i+1},v_t^i)\in E,\ \phi(v^{i+1})\neq t+1 \\ 0 & (v^{i+1},v_t^i)\notin E \end{cases}$$

邻居节点不是 t+1 类型

其中 $v_t^i \in V_t$.

$N_{t+1}(v_t^i)$ 表示 v_t^i 节点的邻域中属于 V_{t+1} 类型的节点（根据定义的元路径）

此外：元路径第一个节点类型与最后一个节点类型一致
如 $p(v^{i+1}|v_t^i) = p(v^{i+1}|v_1^i)$, if $t=l$

4. metapath2vec++ 同时实现异构网络中结构和语义关联的建模

之前：给定：$G=(V,E,T)$. with $|T_V|>1$

最大化存在异质上下文 $N_t(v)$ 的概率：

$$\arg\max_{\theta} \sum_{v\in V}\sum_{t\in T_V}\sum_{c_t\in N_t(v)} \log p(c_t|v;\theta) \quad (2)$$

$N_t(v)$：v 的邻居、节点类型为第 t 种

$$p(c_t|v;\theta) = \frac{e^{X_{c_t}\cdot X_v}}{\sum_{u\in V} e^{X_u\cdot X_v}}$$

softmax 在 softmax 中未考虑每个节点的类型.

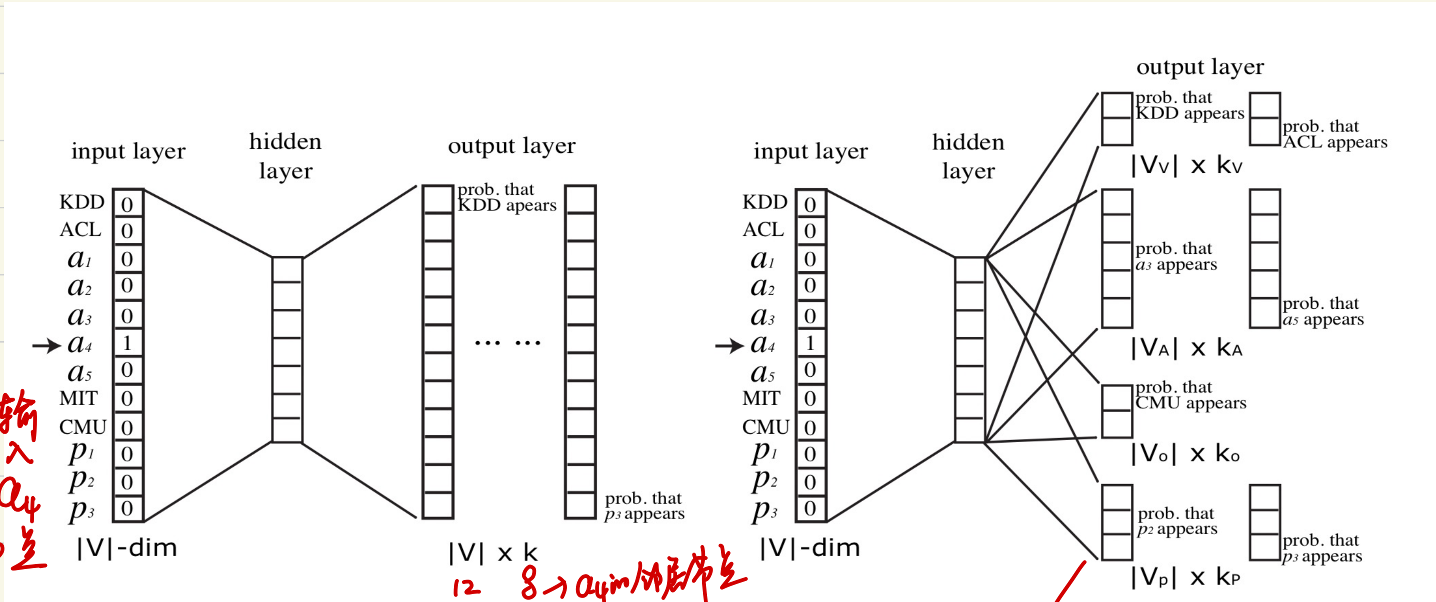
X_v：矩阵 X 的第 v 行，表示节点 v 的向量表示.

改进：根据具体的节点类型t来调整 $p(c_t|v;\theta)$ 更新softmax

$$p(c_t|v;\theta)=\frac{e^{X_{c_t}\cdot X_v}}{\sum_{u_t\in V_t} e^{X_{u_t}\cdot X_v}}$$

也就是说：在归一化过程中，会根据固定类型的节点来调整

如此操作，metapath2vec++ 实际上是在skip-gram模型的输出层考虑了节点v的每个邻居节点的节点类型



输入 a_4 节点

12　8→a_4的邻居节点

(b) Skip-gram in *metapath2vec*, node2vec, & DeepWalk

(c) Skip-gram in *metapath2vec++*

输出4个，分别对应 a_4 的4种邻居节点类型 (V, A, O, P)

$V=V_V\cup V_A\cup V_O\cup V_P$

$k=k_V+k_A+k_O+k_P$

负采样改写.

给定一个负采样样本量 M

(2)式更新：$\log\sigma(X_{c_t}\cdot X_v)+\sum_{m=1}^{M}E_{u^m\sim P(u)}[\log\sigma(-X_{u^m}\cdot X_v)]$

其中 $\sigma(x)=\frac{1}{1+e^{-x}}$

$P(u)$：预先定义的分布

每种类型对应一个分布

$$O(X)=\log\sigma(X_{c_t}\cdot X_v)+\sum_{m=1}^{M}E_{u_t^m\sim P_t(u_t)}[\log\sigma(-X_{u_t^m}\cdot X_v)]$$

总结：　比较目标函数

metapath2vec：
$$\log\sigma(X_{c_t}\cdot X_v)+\sum_{m=1}^{M}E_{u^m\sim p(u)}[\log\sigma(-X_{u^m}\cdot X_v)]$$

metapath2vec++：
$$\log\sigma(X_{c_t}\cdot X_v)+\sum_{m=1}^{M}E_{u_t^m\sim p_t(u_t)}[\log\sigma(-X_{u_t^m}\cdot X_v)]$$

metapath2vec 异质体现在序列采样策略上.

metapath2vec++ 的异质不仅体现在序列采样策略上，还体现在负采样（目标函数）.