**Thinking1 什么是Graph Embedding，都有哪些算法模型**

* factorization methods （图因式分解机）
* random walk techniques（随机游走）
* deep learning（深度学习）

**Thinking2 如何使用Graph Embedding在推荐系统，比如NetFlix 电影推荐，请说明简要的思路**

Embedding 在数学上表示一个映射关系， F: X -> Y， 也就是一个函数。函数具有两个性质：injective 和 structure-preserving。Injective，即我们所说的单射函数，对于每个 Y 只有唯一的 X 对应，反之亦然；structure-preserving，结构保存，比如在 X 所属的空间上 X1 < X2，那么映射后在 Y 所属空间上 Y1 < Y2。

简单点说，深度学习中，Embedding 特指用一个低维度向量表示一个实体，可以是一个词（Word2Vec），可以是一个物品（Item2Vec），亦或者网络关系中的节点（Graph Embedding）。

Embedding 主要的三个应用方向：

* 在深度学习网络中使用 Embedding 层，将高维稀疏特征向量转换成低维稠密特征向量，从而减少后续模型参数量，后续模型可以是深度学习模型，或者传统的机器学习模型；
* 作为预训练技术，直接使用别人训练好的 Embedding 向量，与其他特征向量一同输入后续模型进行训练，例如 Word2Vec；
* 通过计算用户和物品的 Embedding 相似度，Embedding 可以直接作为推荐系统或计算广告系统的召回层或者召回方法之一，例如 NetFlix电影推荐系统。

Graph Embedding 用低维、稠密、实值的向量表示网络中的节点。Word2Vec 通过序列（sequence）式的样本：句子，学习单词的真实含义。仿照 Word2Vec 思想而生的 Item2Vec 也通过商品的组合去生成商品的 Embedding，这里商品的组合也是序列式的，我们可以称他们为“Sequence Embedding”。然而，在更多场景下，数据对象之间更多以图/网络的结构呈现。

Graph Embedding 之所以能好于 “Sequence Embedding”，是因为 Graph Embedding 能够生成一些“不存在”的序列。Node2Vec 所体现的网络的同质性和结构性在推荐系统中也是可以被很直观的解释的。同质性相同的电影很可能是同品类、同属性、或者经常被一同观看的电影，而结构性相同的电影则是各类型最受欢迎影片等拥有类似趋势或者结构性属性的电影。毫无疑问，二者在推荐系统中都是非常重要的特征表达。由于 Node2Vec 的这种灵活性，以及发掘不同特征的能力，甚至可以把不同 Node2Vec 生成的 embedding 融合共同输入后续深度学习网络，以保留电影的不同特征信息。

**Thinking3 数据探索EDA都有哪些常用的方法和工具**

A. 检查数据：

- 是否有缺失值？   
- 是否有异常值？   
- 是否有重复值？   
- 样本是否均衡？  
- 是否需要抽样？   
- 变量是否需要转换？   
- 是否需要增加新的特征？

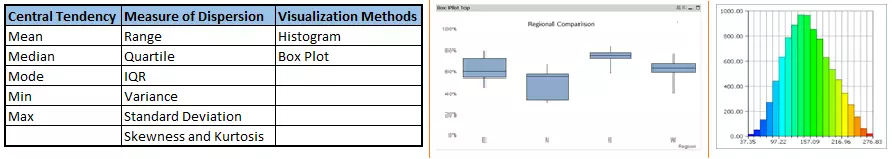
注：数据类型分为数值型，类别型，文本型，时间序列等。这里主要指的是数值型（定量数据）和类别型（定性数据），其中数值型又可以分为连续型和离散型。

B. 使用描述统计量和图表对数据进行描述：

1）连续变量：

常见的描述统计量：平均值，中位数，众数，最小值，最大值，四分位数，标准差等

图表：频数分布表（需进行分箱操作），直方图，箱线图（查看分布情况）



2）无序型离散变量：

常见的描述统计量：各个变量出现的频数和占比

图表：频数分布表（绝对频数，相对频数，百分数频数），柱形图，条形图，茎叶图，饼图

3）有序型离散变量：

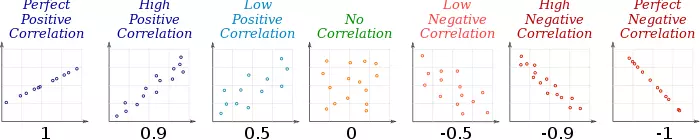
常见的描述统计量：各个变量出现的频数和占比

图表：频数分布表，堆积柱形图，堆积条形图（比较大小）

C. 考察变量之间的关系：

1）连续变量和连续变量（Continuous & Continuous）：

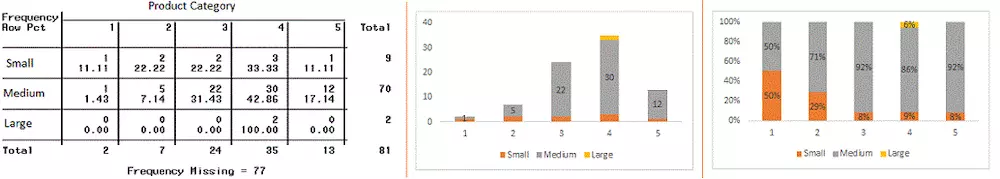
对于连续变量与连续变量之间的关系，可以通过散点图进行查看。对于多个连续变量，可使用散点图矩阵，相关系数矩阵，热图。



量化指标：皮尔逊相关系数（线性关系），互信息（非线性关系）

2）离散变量和离散变量（Discrete & Discrete）：

对于离散变量与离散变量之间的关系，可以通过交叉分组表，复合柱形图，堆积柱形图，饼图进行查看。对于多个离散变量，可以使用网状图，通过各个要素之间是否有线条，以及线条的粗线来显示是否有关系以及关系的强弱。



量化指标：卡方独立性检验--->Cramer’s φ (Phi) or Cramer’s V

3）离散变量和连续变量（Discrete & Continuous）：

对于离散变量和连续变量之间的关系，可以使用直方图，箱线图，小提琴图进行查看，将离散变量在图形中用不同的颜色显示，来直观地观察变量之间的关系。

量化指标：独立样本t检验中的t统计量和相应的p值（两个变量），单因素方差分析中的η²（三个变量及以上）

D. 其他

1）检查数据的正态性：直方图，箱线图，Q-Q图（Quantile-Quantile Plot ）

直方图，箱线图：看图形是否对称

Q-Q图：比较数据的分位数与某个理论分布的分位数是否匹配