数据科学导论 2022秋 第二次作业

第一题 - 因果分析

TA 最近得到了一个任务。抖腿公司会在短视频间隙向用户推送一些广告，公司很关心这些广告的点击率，于是分配了TA一个任务：提供用户点击或未点击广告的**历史记录**，分析哪些因素会是影响用户点击广告的**原因**。

TA 发现主要有两种进行因果分析的模型：虚拟事实模型（Counterfactual Model, Rubin 1978)和因果图模型（Causal Diagram, Pearl 1995)，但是TA不清楚这些模型是如何计算因果作用的。请你帮帮TA。

请**任选一种**模型，**简述**这种模型的**主要思想**，并说说**在该实际场景**下如何应用。

（可选题）关联分析

TA经常去C楼超市买吃的，买一包虾片之后总喜欢再买一袋酸奶。TA比较好奇，如果顾客群体中有诸如此类的购买习惯的共性，是否可以从超市的购买记录中挖掘出来呢？TA得到了超市的一些交易记录，其中不同的字母代表不同的物品。请你帮帮TA挖掘数据中的这种关联。

购买记录ID	购买的物品
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y}
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I, E}

在数据挖掘中，这种任务叫做**关联规则学习**或者叫**关联分析**。请自学有关的知识(如关联分析算法)，了解基本概念（如频繁项集、关联规则等）和Apriori算法。

我们设定最小支持度为0.6，最小置信度为0.8。请找出所有的频繁项集(满足支持度)，以及满足如下条件的强关联规则： Buys(item1) AND Buys(item2) ==> Buys(item3) (同时满足置信度和支持度)。

（可选题）聚类

TA最近学习了一些聚类算法，对于泰坦尼克号上乘客的生存情况分析有了新的灵感。他好奇轮船上的最终生存与否和乘客本身的信息的关系，是不是能使用聚类方法找出来，你能帮帮TA吗？

请观察titanic文件夹中的train.csv和readme.md，了解乘客都有什么基本信息**（与作业1可选题相同）**，思考怎么抽象成为聚类算法问题。

阅读材料， **Extra Section部分**（可以略过代码只看文字和图片）。简述他做了什么数据处理，采用了什么聚类算法，效果分别怎么样呢？

作业要求

- **严禁抄袭**!

- 上传**单个pdf文件**至网络学堂。
- 推荐使用 markdown 或 latex 写作业。