2020 美赛 C 题参考解题思路

C 题总的来说是属于数据挖掘的数学建模问题。数据类型的建模常用工具可以使用 SPSS，当然另外较为常用的工具可以使用 SQL Server 或者 Python（可以更方便的做文本分析）。

• 第一问

第一问可以看为数据挖掘的准备工作，可以看作数据预处理。最主要的将不蕴含或者蕴含较少实际意义的数据作剔除或者合理处理。其中不可忽视的数据是 vine 和 verified_purchase。因为这两项数据大大程度的影响了评级和评论的重要程度，根据现实生活的情形作合理处理即可。

• 第二问

a．本题需要做的工作是通过客户本次购物的评级以及评论来对产品进行一个综合评价，当然单纯的评级已经是客户自己量化好的数据了，本次的难点是将定性的评价量化。涉及数据挖掘中的文本分析，可以使用 LDA 模型（将文本向量化）。其中推荐一种统计方法 TF-IDF，可以自己建立较为科学的几类评论来代表不同等级的评价（量化评论），在通过 TF-IDF 算法将评论分类量化。当然此量化方法导致的量化数据并非连续的。一可以参考以下方法，同样也需要建立较为科学的几类评论来代表不同等级的评价，然后利用余弦相似性算法或者简单共有词算法等计算评论与你所建立的评论之间的相似度，通过相似度来量化评论。量化过后与评级一同采用简单的综合评价方法（例如线性加权）得到综合评价的结果。需要多注意的是亚马逊 vine Voices 的评论是否更具有说明力。另外也需要注意的是，评论标题与内容有可能存在不符。

b．本题主要是考虑评价和时间的关系，可以利用拟合来完成。根据 a 每一次的购买都会有一个综合评价的数值或者分数，因此可以将每一天看作一个时间单位，拟合平均评价值和时间的关系，从而说明声誉的趋势。如若样本数据太大，也可以进一步的将大样本转化为小样本再进行拟合。

c．本题需要基于产品声誉和时间的关系，才能找到潜在成功或者失败的评级评论的度量组合，需要对声誉时间的函数关系中**拐点和峰值**左右时间一定范围内的数据进行挖掘分析，找到基于文本的度量值和基于评级的度量值的组合。

d．挖掘时间节点前的星级评价对时间节点后的评论是否存在联系，可以巧妙的利用相关性分析解决此问题。例如仅考虑每一条评论与此前一周的评级的关系，即可提取数据样本，作处理后利用 spss 作相关性分析。

e．与 d 类似，挖掘时间节点前的评论与后续的评级之间存在多大的关联，可以利用 LDA 模型对此前的评论作词频统计，从而提取评级与评论的数据样本。作处理后再进行相关性分析。

- **第三问**

总结本次的分析与结果，需说明数据处理以及分析过程的合理性，最终阐述结论，根据文本评论分析一些潜在的问题，并提出建设性的建议。