**西安电子科技大学**

**安全前沿讨论班（I） 课程实验报告**

**实验名称 使用K-Means聚类进行青少年市场细分**

学 院 网络与信息安全学院 班 级 1618019

姓 名 曹寅峰 学 号 16020610025

实验日期 年 月 日

|  |
| --- |
| 指导教师评语：  指导教师：  年 月 日 |
| **实验报告内容基本要求及参考格式**   1. 实验目的   通过本次实验，掌握K-Means聚类方法，了解如何使用Python中的K-Means模块进行数据分析，加深对pandas库、numpy库、sklearn库的了解，掌握这些库中常见的数据分析方法，理解K-Means进行数据分析的基本步骤。  1. 任务1：利用pandas库加载数据并查看数据；  2. 任务2：处理数据中的缺失值并对数据进行标准化；  3. 任务3：应用K-Means模块训练模型并对结果进行分析。   1. 实验环境   Anaconda   1. 实验基本原理及步骤（或方案设计及理论计算） 2. K-Means算法的原理和流程； 3. 在特征空间中选择𝑘个点作为初始质心（通常是随机选择） 4. 计算每个样本到各个簇质心的距离（欧式距离等距离度量） 5. 将样本分配到距离自己最近的质心所代表的簇 6. 以簇内样本的均值更新各簇的质心 7. 重复2-4优化过程，直至样本的分配和质心不再改变，（或迭代次数超过某个值） 8. 质心距离度量方式，如欧氏距离、曼哈顿距离、马氏距离、海明距离、余弦相似度和杰卡德距离等；   **欧式距离：**欧几里得距离或欧几里得度量是[欧几里得空间](https://baike.baidu.com/item/%E6%AC%A7%E5%87%A0%E9%87%8C%E5%BE%97%E7%A9%BA%E9%97%B4" \t "_blank)中两点间“普通”（即直线）[距离](https://baike.baidu.com/item/%E8%B7%9D%E7%A6%BB" \t "_blank)。使用这个距离，欧氏空间成为[度量空间](https://baike.baidu.com/item/%E5%BA%A6%E9%87%8F%E7%A9%BA%E9%97%B4" \t "_blank)，这也是我们最早接触的距离公式，比较好理解  **曼哈顿距离：**表示两个点在标准坐标系上的绝对轴距之和  /var/folders/c9/zw08ghqn5k75y25byzqc7sdc0000gn/T/com.microsoft.Word/WebArchiveCopyPasteTempFiles/70  图中红线代表曼哈顿距离，绿色代表欧氏距离，也就是直线距离，而蓝色和黄色代表等价的曼哈顿距离。曼哈顿距离——两点在南北方向上的距离加上在东西方向上的距离，即d（i，j）=|xi-xj|+|yi-yj|  **马氏距离：**马氏距离是由印度统计学家马哈拉诺比斯提出的，表示数据的协方差距离。它是一种有效的计算两个未知样本集的相似度的方法。与欧氏距离不同的是它**考虑到各种特性之间的联系，**即独立于测量尺度。 对于一个均值为  ，协方差矩阵为Σ的多变量矢量  ，其马氏距离为  **海明距离：**在信息编码中，两个合法代码对应位上编码不同的位数称为码距，又称海明距离。举例如下：10101和00110从第一位开始依次有第一位、第四、第五位不同，则海明距离为3。  **余弦相似度：**余弦距离，也称为余弦相似度，是用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小的度量。  余弦值越接近1，就表明夹角越接近0度，也就是两个向量越相似，这就叫"余弦相似性"  **杰卡德相似度系数：**Jaccard（杰卡德）相似性系数主要用于计算符号度量或布尔值度量的样本间的相似度。若样本间的特征属性由符号和布尔值标识，无法衡量差异具体值的大小，只能获得“是否相同”这样一种结果，而Jaccard系数关心的是样本间共同具有的特征。  **杰卡德距离：**与Jaccard系数相反的概念是Jaccard距离，用两个集合中不同元素所占元素的比例来衡量两个集合（样本）的区分度，可用如下公式表示：  /var/folders/c9/zw08ghqn5k75y25byzqc7sdc0000gn/T/com.microsoft.Word/WebArchiveCopyPasteTempFiles/Center  **皮尔逊相关系数：**Pearson 相关系数是用协方差除以两个变量的标准差得到的，虽然协方差能反映两个随机变量的相关程度（协方差大于0的时候表示两者正相关，小于0的时候表示两者负相关），但其数值上受量纲的影响很大，不能简单地从协方差的数值大小给出变量相关程度的判断。为了消除这种量纲的影响，于是就有了相关系数的概念。  当两个变量的方差都不为零时，相关系数才有意义，相关系数的取值范围为[-1,1]   1. 聚类结果评价指标。   对于一个聚类的结果，有两种评价的指标，一种是外部指标，一种是内部指标。  /var/folders/c9/zw08ghqn5k75y25byzqc7sdc0000gn/T/com.microsoft.Word/WebArchiveCopyPasteTempFiles/watermark,type_ZmFuZ3poZW5naGVpdGk,shadow_10,text_aHR0cHM6Ly9ibG9nLmNzZG4ubmV0L3dlaXhpbl8zODA5MTgzMA==,size_16,color_FFFFFF,t_70  聚类衡量2  聚类衡量3  实际上，对于不同的问题，选择不同的方法，要使用不同的聚类指标进行评价   1. 实验数据记录（或仿真及软件设计）     首先导入CSV文件  查看维度    统计毕业年份的信息，注意设置不要忽略NA    利用describe函数可以查看统计信息，包括最值，均值，标准差，数据类型等  为了筛选无效成绩，只保留13-20的年龄段  再通过虚拟编码，把gender转换成三个二元变量，这样就可以计算欧式距离了  将处理后的数据进行拼接  查看均值  再把之前的无效年龄用平均年龄代替，例如最右下角的17.252  查看当前的列名  进行数据标准化，方便计算  查看标准化的结果，这里和同学的同一份代码均值结果不同，可能是mac下python3.7版本中忽略了某些过小值导致，但方差是一致的。  最后进行训练，设置分为5类（k=5）   1. 实验结果分析及回答问题（或测试环境及测试结果）   我们成功将青少年分为5类  其中每一个类的中心是如下特征，代表青少年的不同偏好种类  也可以将k设置为2，这样就变成了二分类问题，可以从gender看出，这次性别拥有相当大的区别。  同理设置k=3  最后我们并没有进行定量分析，而是进行定性分析，查看各个种类间区别，这是由于我们的目的——定向营销决定的，我们更关心青少的的种类  总结，k聚类有如下优缺点：  优点：  1、解决聚类问题的经典算法，简单、快速  2、当处理大数据集时，该算法保持可伸缩性和高效率  3、当簇近似为高斯分布时，它的效果较好  缺点：  1、在簇的平均值可被定义的情况下才能使用，可能不适用于某些应用  2、必须实现给出k（要生成簇的数目），而且对初值敏感，即对于不同的初值，可能会导致不同结果  3、不适合非凸形状的簇或者大小差别很大的簇  4、对噪声和孤立点敏感  期间遇到了同一份代码在两种平台运行结果不同的问题，初步判定是由于数据过小，mac平台忽略的问题。  通过这次实验，了解到了简单聚类算法与评价方法，在很多日常生活中都可以试用。  **同时这应该也是本次课程最后一次报告了，个人感觉安全前沿讨论班这门课程内容翔实，收获较大。了解到了jupyter这个强大的工具，也锻炼了我的python编程能力，这在今后的各个领域都能用到，如果能早一点学就更好了。**  **同时感谢老师的辛苦备课，批改作业。祝王利娟老师日后工作顺利，身体健康！**  **曹寅峰** |