名称：基于北京地铁IC卡出行记录的乘客画像方法设计

现有数据6个月的IC记录：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| IC卡号 | 进站时间 | 进站点(First) | 离站时间 | 终点站（Last） |
| 001 | 201906010930 | A | 201906011030 | B |
| … | … | … | … | … |

初步分析可得某用户数据

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 首次进站平均时间 | 终次离站平均时间 | 总出行记录数 | 日均出行次数 | 周均出行天数 | 出行记录时长 | 出行时始发站点的概率 | 终点站点的概率 |
| T（in） | T(out) | N | N（day） | Q(week) | L | P(First) | P（Last） |

**一.住址与工作地大致区域及距离的判断：**

1.选择乘客总出行记录中频度最高的2个站点

2.每天的第一次出行记录的初站点与最后出行记录的终点站相同🡪住址

第一次出行记录的终点站与最后出行记录的初站点相同🡪工作地

3.住址与工作地的距离：

（1）注册百度地图开发密钥，python爬虫+密钥+地点输入获取json数据，从json数据中读取各地点精度，维度，然后计算距离

（2）距离公式：

|  |
| --- |
| （1）地球球心为原点，球心与赤道上0经度点的连线为X轴，球心与赤道上东经90度点的连线为Y轴，球心与北极点的连线为Z轴：  x=R×cosα×cosβ R为地球半径，约等于6371km  y=R×cosα×sinβ α为纬度，北纬取+，南纬取-  z=R×sinα β为经度，东经取+，西经取- |
| （2）求至直线距离，即弦长：  L=[(x1-x2)^2+(y1-y2)^2+(z1-z2)^2]^0.5 |
| （3）弦长转弧长：  C=arcsin(L/2R)\*Pi\*R/90 |

**二.出行模式的判断：**

变量：首次进站时间，终次离站时间，日出行次数，周出行天数，出行记录时长

算法选取：k-means算法：

1.确认簇类个数k

2.初始化簇类中心点µ

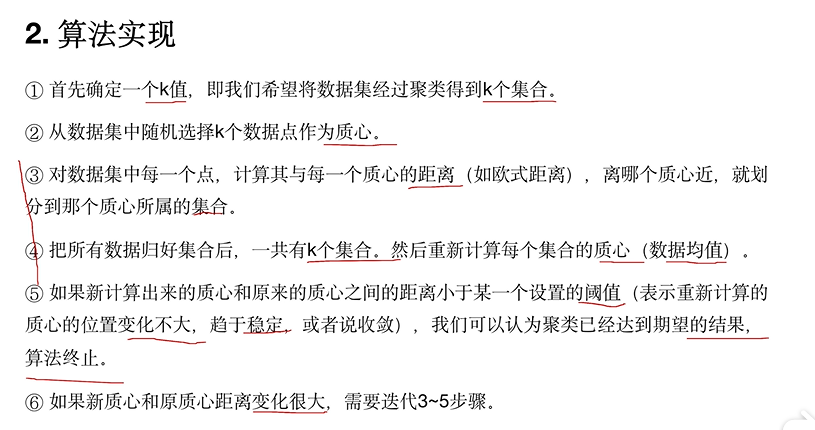
3.以各样本点离簇类中心的距离平方和越来越小为原则，不断迭代更新簇类中心点的位置

|  |
| --- |
| J为代价函数，c为样本点，µ为簇类个数。目标为求min(J) |

簇类个数决定问题：采用肘部原则，设置簇类个数为1-7进行训练，画出代价和簇类个数坐标图，取肘部的簇类个数。

预期结果举例：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 出行模式 | 日均出行次数 | 周均出行天数 | 平均出行时长 | 初次出行平均时间 | 末次出行平均时间 | 人数占比 |
| 1 |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |
| 3 |  |  |  |  |  |  |



**三.乘客身份大致判断：**

1.每个工作日，时间相对固定（+-15分钟），且始末站点相同出行记录🡪硬性工作制人员

2.每周有5-6天有时间点相对幅度较大（+-60分钟），且始末站点相同出行记录🡪弹性工作制人员（例如996，995的互联网行业人员）

3.每个工作日有固定出行记录且第一次出行记录的结束时间<7点🡪普通学生（学校开课时间一般是7：00）

**四.基于时空分析的乘客爱好判断：**

在高速发展，生活压力大的繁华大都市，每个人都必须为美好生活而奋斗拼搏，成年人每天的大部分时间是在忙碌的工作中度过的，学生的生活重心在学习上。只有在闲暇之余，我们才可能抽挤出时间用在个人爱好上。

从信息论的角度来看，某件事物发生的频率越低，其中蕴含的信息量越大。

|  |
| --- |
| f(P)对应P事件的信息量，p为该事件的概率。概率越低，信息量越大 |

正所谓偶然中存在着必然，一个人的爱好可以通过其占比频率较低的个人出行记录所对应的事件来判断。通过基于时空分析的方法，例如A时段B地正在举办一场音乐会，而X乘客在该时段，该地附近的P站点有出站记录，且其出现在该站点的概率小于平均站点的概率，可判定该乘客参加了该音乐会，其喜欢音乐。

事件🡪时间，空间，人物🡪个人爱好

问题：如何判定该乘客参与了某活动？

步骤：1.空间符合：python爬虫，爬取大麦网各种活动的时间，地点，所属爱好类型；

爬取北京各地铁公交站点和大麦网各活动地点的经纬度

通过经纬度计算直线距离，记录下距离各活动地点前四的地铁站点

将以上内容建表，存储在mysql数据库中

完成以上内容，在空间上的问题解决

2.人物和时间符合：①在活动开始前乘客IC卡有在该活动场地距离前四某站点出站的记录

②在活动结束后乘客IC卡有在该活动场地距离前四某站点入站的记录

③在活动进行期间乘客IC卡无出行记录

3.空间，时间，人物都满足的乘客最后筛选：

判断该活动地点的附近站点是否是该乘客出行的常用站点，记录乘客出行记录里各站点的频度，大于平均频度标记为常用站点，反之为不常用站点。排除掉是常用站点的乘客，为余下的乘客标记上个人爱好。

数学表示：

