# 摘要

我们首先对收集到的数据进行初步过滤，然后进行汇总分析。我们基于街道将北京市分割成若干个区域，对每一个区域进行统计分析，有一个大致的了解。对于医疗推荐，我们找到了北京市多家影响力高的知名医院，我们构建了医院-地点矩阵，针对具体的地点，不同医院不同时间段的就医数量构成一个地点的就医向量，然后对其进行主成分分析降维，得到互相独立的一个向量，然后再使用高斯混合模型来对所有地区分类，考虑到各个地区的人数不同，所以再使用余弦相似度来度量地区之间的相似程度，由此划分区域，再根据每个区域的进行最大似然估计来得到医院的推荐排名。

# 介绍

我们的目标是对北京市的每一块区域进行就医医院的推荐，我们有以下的创新：

1. 不同于传统的矩阵划分区域，我们直接使用了街道进行划分，更准确。
2. 我们基于GMM对每块地区进行分类，由于中心极限定理，我们从概率角度对地区的就医偏好进行了建模，与生活实际较为贴切。
3. 我们同时考虑到了不同地区生活的人数不同，这会影响到不同地区的就医情况，所以我们采用余弦相似度来度量区域的相近程度。
4. 得益于GMM的概率模型性质，可以看作对区域进行了软聚类（soft clustering），这使得我们能够对每一个地区的就医偏好进行排序建模。

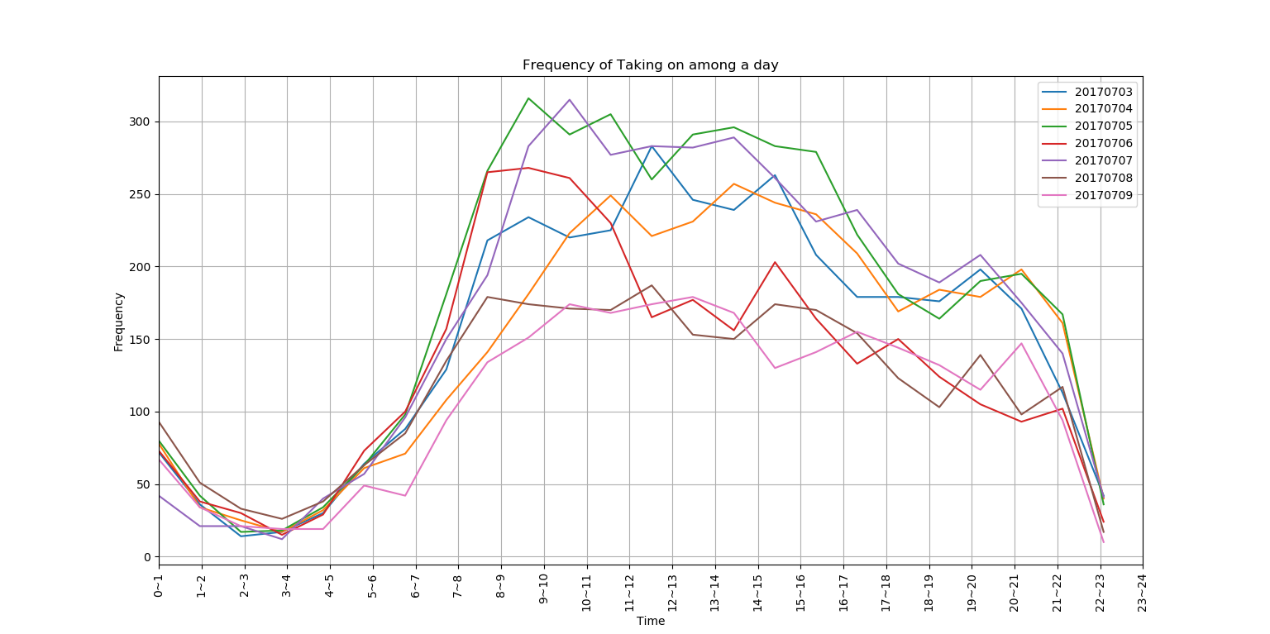
# 数据准备和初步分析

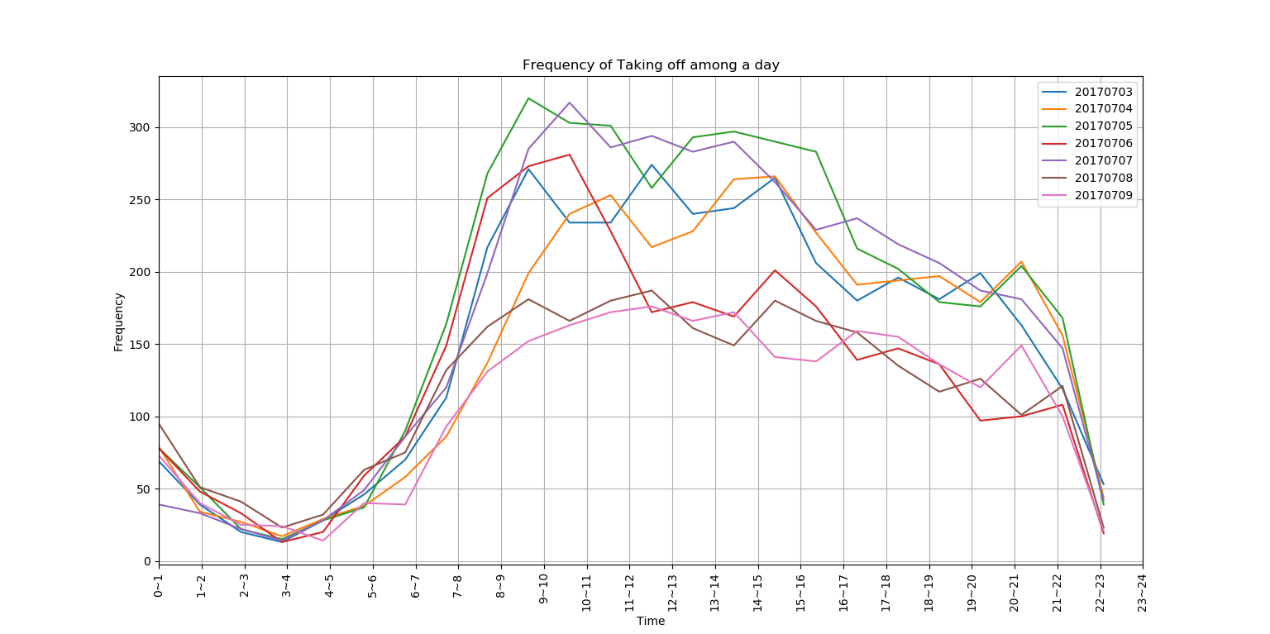
数据来源，我们的数据来自于北京城市规划院，给我们提供了整个北京市范围内的出租车GPS数据，数据格式如下表所示，可以看到原始数据（raw data）是没有列标记的，在这个阶段我们将通过合理的分析确定列所代表的意义。数据的时间范围为2017年7月3号到2017年7月9号一共七天，其中3号为星期一。原始数据总共有 条，范围是整个北京市。

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| JKSX | 6.58E+11 | 浜珺U7413 | 2017/7/2 22:53 | 116.6437 | 40.11923 | 25 | 0 | 1 | 102 | N |
| JKSX | 6.59E+11 | 浜珺V0630 | 2017/7/2 22:53 | 116.4333 | 39.92167 | 18 | 1 | 1 | 28 | N |
| JKSX | 6.58E+11 | 浜珺R4499 | 2017/7/2 22:52 | 116.5114 | 40.019 | 0 | 0 | 1 | 0 | CZ |
| JKSX | 6.58E+11 | 浜珺U6513 | 2017/7/2 22:53 | 116.2431 | 40.00492 | 49 | 0 | 1 | 60 | N |
| JKSX | 6.58E+11 | 浜珺R5583 | 2017/7/2 22:52 | 116.1859 | 40.04455 | 0 | 0 | 1 | 188 | CZ |

首先我们需要判断每一列所代表的含义，我们的模型所需要的特征有：出租车唯一标识号、时间、经度、纬度、上车点下车点。初步分析，容易看出车牌号一列可以作为一辆车的唯一标识号，时间列也非常容易判断，经纬度分别为后两列，符合北京市的地理位置，最后一列我们判断为是状态列，即表示车辆上是否有乘客，用0和1分别表示无乘客和有乘客，通过对该列的分析我们可以得到每一个上车点和下车点。

为了验证我们对该列的判断，我们分析得到上车点和下车点后，绘制了随时间变化的上车次数和下车次数的折线图，如图所示，可以看出结果非常符合常识，白天的时候上下车次数明显高于夜间并且上下车曲线大致吻合，由此我们可以进一步确认该列的确就是状态列。





然后我们需要对数据进行清洗。我们发现，该数据集中有1950年的数据或者0经纬度的数据存在，这些都是明显错误需要剔除的数据。所以我们基于以下几条标准对数据进行了清洗：

* 经纬度越界
* 经纬度或者时间没有发生变化的车辆，即根本没有进行移动的车辆，这对我们的研究目标没有提供帮助，所以我们选择剔除掉。
* 全天处于空驶状态的车辆，这说明该辆车没有乘客搭乘，也不会对我们的研究提供帮助。

经过过滤并且我们提取了上车点和下车点的数据按车辆和时间进行排序，总共有 条，格式如下表所示，并绘制出OD点对如图所示。

# 研究方法

我们将整个北京市按照主要道路进行划分，这样更贴合实际，划分后的北京市如图所示。

我们首先选取了若干个北京市知名的医院，收集它们的经纬度信息，作为我们在这个研究中的推荐对象，这些医院的信息如下表所示：

我们首先统计每一个区域的不同时间段到各个医院的车辆数，记为，代表第k个时间段内从第i个区域到第j个医院的车辆数目，对每一个时间段构建区域医院矩阵，这些矩阵构成就医向量空间，



然后再将这些矩阵排列为列向量，就可以得到区域的特征矩阵



该矩阵的每一列就是对应的区域的特征向量，一共有7天，每天划分为24小时，一共有m家医院，共有 个维度，可以看出该特征向量维数非常高，所以我们需要采取一定的方法对其进行降维。

我们采用主成分分析法，经过降维后提取的特征可以看作是互相独立的随机变量，为高斯混合模型的建立提供条件。

## 区域软聚类

PCA是一种常用的降维方法，它利用正交变换把由线性相关的变量表示的观测数据转换为少数几个由线性无关变量表示的数据，这些线性无关的变量就叫做主成分。方差随维数变化图如下，我们决定保留90%的特征，达到了降维效果。提取主成分后的数据将会用于我们进一步的分析。我们记降维后的第i个区域的特征向量为，p为降维后得到的主成分个数。

### 高斯混合模型

我们可以认为经过降维后的特征向量具有如下形式的概率密度函数：



其中c是区域的特征向量，theta为该概率的参数向量，K是我们将要划分出来的类的个数，\phi是正态分布的密度函数，即 ，\alpha\_k就代表了该区域属于第k个类的概率。为了估计该概率密度函数的参数，由于MLE算法在这里无法有效求解，我们使用EM算法来近似估算，简而言之就是通过求期望和极大进行迭代来使得似然函数尽量大。

由此我们就可以得到区域i属于类别k的概率。

## 划分区域并构建推荐方案