八、问题四的模型与建立

8.1 问题分析

在本题中，自变量为婴儿的整晚睡眠时间，睡醒次数，以及婴儿的入睡方式，因变量为婴儿的睡眠质量，分为优，良，中，差四类，需要根据综合评判，得到对应的婴儿睡眠治疗的指标。

应用评价型模型的难点在于，婴儿的年龄不相同，入睡方式不同，且数据繁多，难以统一标准，因此，本文采用了无监督分类模型的 Kmeans，并规定分成的簇的数量为 4，即对应优，良，中，差，共 4 类，再根据 4 堆簇的中心点指标，去判断该簇属于优，良，中，差中的一类。

8.2 Kmeans 算法

8.2.1 Kmeans 算法原理

已知数据集$(x_1, x_2, \dots x_n)$，Kmeans 聚类要把这n个数据划分到k个集合中$(k \le n)$，使得组内平方和最小，它的目标是找到使得下式满足的聚类S_i

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2 \qquad (4-1)$$

其中μ_i是S_i中所有点的均值。

8.2.2 Kmeans 算法步骤

（1） 对数据集进行标准化和归一化，避免均值和方差大的数据对聚类产生决定性影响

（2）选择初始化的k个样本作为初始聚类中心$a = a_1, a_2 \dots a_k$

（3）针对数据集中每个样本x_i，计算它到k个聚类中心的距离，并将其分到距离最小的聚类中心所对应的类中

（4）针对每个类别$a_j (j = 1,2..k)$，重新计算它的聚类中心$a_j = \sum_{x \in S_i} x$，即属于该类的所有样本的质心

（5）重复(3)(4)步，知道达到某个中止条件（迭代次数，可允许最小误差等）

其伪代码为：

```
获取数据 n 个 m 维的数据
随机生成 K 个 m 维的点
while(t)
    for(int i=0;i < n;i++)
        for(int j=0;j < k;j++)
            计算点 i 到类 j 的距离
    for(int i=0;i < k;i++)
        1. 找出所有属于自己这一类的所有数据点
        2. 把自己的坐标修改为这些数据点的中心点坐标
end
```

8.2.3 Kmeans 模型求解

通过 Python 的可视化库 matplotlib 可以直观地展现出聚类散点图，由于变量数大于 2 个，即取主成分分析(PCA)降维后前两个主成分来绘制散点图，在一定程度上可查看聚类效果。

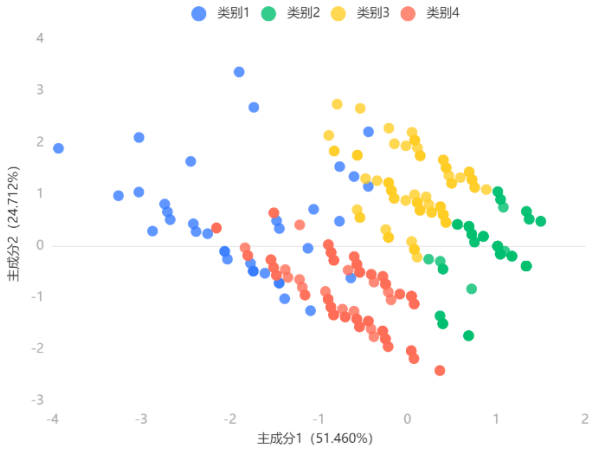

图 4.1 聚类散点图

表 4.1 聚类中心点坐标

聚类种类	中心值_整晚睡眠时间小时制	中心值_睡醒次数	中心值_入睡方式
1	8.972222222222223	5.138888888888889	1.8333333333333335
2	11.261111111111111	0.32592592592592795	3.8962962962962955
3	9.627551020408164	1.7551020408163265	4.193877551020408
4	9.704166666666667	1.4083333333333334	1.4416666666666653

根据附件可知，研究的婴儿年龄均为 1~3 个月大小，并通过查阅相关资料表 4.2，可以知道睡眠质量优，良，中，差分别为聚类种类中的 1，3，2，4。

表 4.2 0~1 岁婴儿睡眠规律图

0~1岁宝宝睡眠规律图	年龄	就寝时间	睡眠潜伏期	白天清醒时间	白天小睡时间	夜晚睡眠时间	夜晚夜醒时间	夜奶次数
	0-4周	20:00	>60分钟	1小时	1.5-2.5小时/次 4次	2小时/段 4段	3次	3-4次
	4周-3月	20:20	30-40分钟	2-2.5小时	1.5小时/次 4次	2-2.5小时/段 4段	3次	2-3次
	3月-6月	20:50	20-30分钟	2.5-3小时	1.5小时/次 3-4次	2.5-3.5小时/段 3-4段	2-3次	1-2次
	6月-12月	20:50	<20分钟	3-4小时	1.5小时/次 2-3次	3.5-4.5小时/段 2-3段	1-2次	0-1次

8.3 KNN 预测

和问题二预测同理，这里仍然使用有监督的分类算法 KNN，以母亲的身体指标，心理指标为自变量，将婴儿的睡眠的睡眠质量作为因变量，并枚举超参数寻优

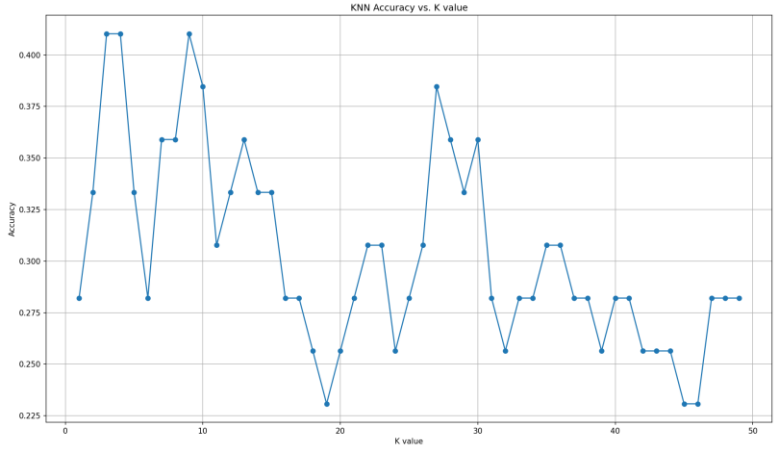

图 4.2 KNN 中 K 值寻优时的测试集准确率

得到最后 20 组（编号 391-410 号）婴儿的综合睡眠质量如下：

表 4.3 编号 391-410 号婴儿的综合睡眠质量预测结果表

编号	391	392	393	394	395	396	397	398	399	400
分类	中	中	差	中	中	优	差	中	中	中
编号	401	402	403	404	405	406	407	408	409	410
分类	良	良	良	差	中	优	中	中	中	中