二、问题二的模型的建立与求解

1. 模型建立

由于没有目标等级函数，故需要使用无监督算法进行分类。

①聚类分析法

聚类分析的基本原则:将每个评价指标作为一个维度，得分看做该特征在该维度上的坐标。通过特征向量计算数据在高纬上的距离，距离越近则认为越相近。把相似程度较大的样品聚合为一类，另一些相似程度较大的样品聚合为另一类，关系密切的聚合到一个小的分类单位，关系疏远的聚合到一个大的分类单位，直到把所有样品聚合完毕。

(1)首先将样本中的n个点各看作一类，此时共有n类。

(2）计算类与类之间的距离，并将距离最短的两类合并成为一个新类。

(3)再重新计算类与类之间的距离，并将并将距离最短的两类合并成为一个新类。如此不断重复，直到所有的样品合为一类。

②数据处理

为了降低计算量，我们计划使用IPCA代替PCA将评价得分进行降维。

| KMO检验和Bartlett的检验 | | |
| --- | --- | --- |
| KMO值 | | 0.000 |
| Bartlett球形度检验 | 近似卡方 | 0.000 |
| df | 55.000 |
| p | NaN |
|  | | |

KMO检验的结果显示，KMO的值为nan，同时，Bartlett球形检验的结果显示，显著性P值为NaN，水平上不呈现显著性，接受原假设，也就是变量间彼此独立，则无法从中提取公因子，主成分分析无效，建议调整数据质量，程度为极不适合。故此，我们放弃降维的做法，采用计算10个品酒师评分的均值作为特征值，而不是将10个品酒师的评价都作为特征。

为了消除评分的主观性，得到一个更客观的评价结果，避免出现葡萄酒因为品酒师的个人喜恶等多种因素造成评分与实践的不匹配结果，M\_{kl}为第k个酒样品的第l个项目的十个评分，x\_{kl\_{max}}为M\_{kl}中的最大值，x\_{kl}\_{min}为最小值。故，处理后的M\_{kl\_{new}}有：

M\_{kl\_{new}}=\{x\_{kli}| x\_{kli} ∈ M\_{kl} , x\_{kli}\neq x\_{kl\_{max}}, x\_{kli} \neq x\_{kl\_{min}}\}

为了克服由于指标的纲量不同对统计分析结果带来的影响需要对原始数据进行标准化处理，以消除单位、数据大小不一致等的影响，我们进行数据特征处理。x\_i为去除最大最小值后的数据，\hat{x\_i}为第i个指标的均值，S\_i为第i个指标的标准差，x^\*\_{ij}为标准化处理以后的指标值，则有:

x^\*\_{ij}=\frac{x\_i-\hat{x\_i}}{S\_i}

2.模型求解

2.1 K-mean++

Robert Parker独一无二的葡萄酒评分体系已经成为一款新酒能否畅销的命运指挥棒。我们学习Robert Parker品酒体系使用K-mean++算法，根据每个人的评分结果将酒样自动聚为6类。

k均值算法至少有两个主要的理论缺陷：

首先，已经证明该算法的最坏情况运行时间是输入大小的超多项式。[1]

其次，与最优聚类相比，所发现的近似值相对于目标函数可能是任意差的。

k-means++ 算法通过在继续进行标准 k 均值优化迭代之前指定初始化聚类中心的过程来解决这些障碍中的第二个障碍。通过 k-means++ 初始化，该算法可以保证找到一个与最优 k 均值解具有竞争力的$ O（log k）$解。

这种方法背后的直觉是，分散出k个初始聚类中心是一件好事：从正在聚类的数据点中随机选择第一个聚类中心，然后从剩余的数据点中选择每个后续聚类中心，其概率与其与点最近的现有聚类中心的平方距离成正比。

确切的算法如下：

1.在数据点中随机选择一个中心。

2.对于尚未选择的每个数据点 x，计算 D（x），即 x 与已选择的最近中心之间的距离。

3.使用加权概率分布随机选择一个新数据点作为新中心，其中选择概率与 D（x）2 成正比的点 x。

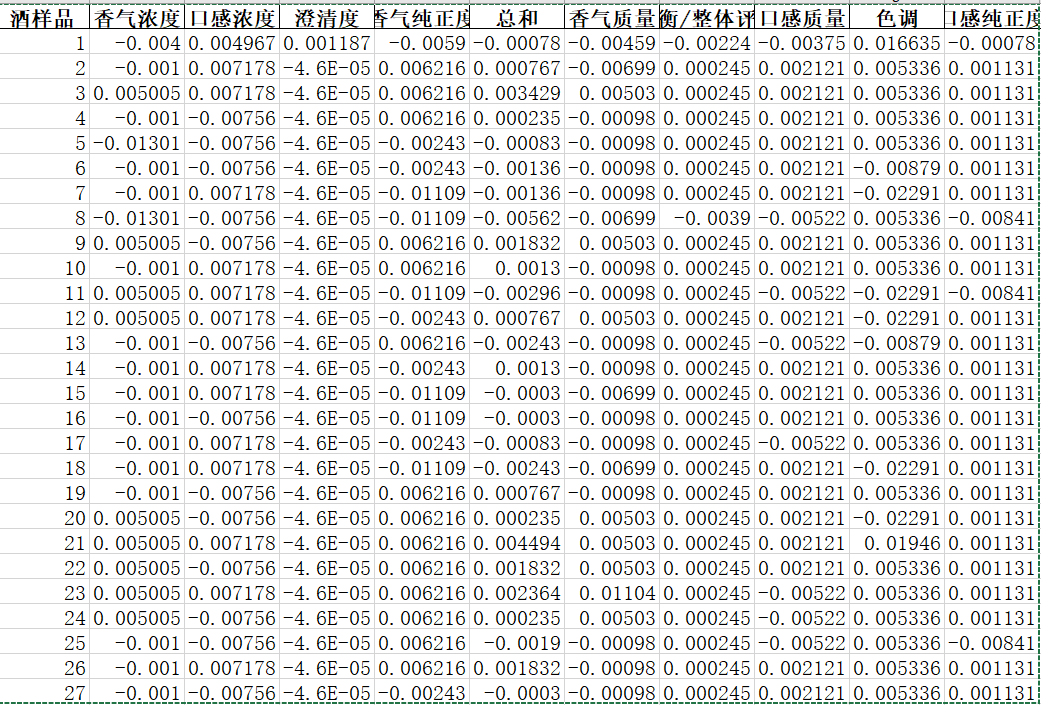
4.重复步骤 2 和 3，直到选择了 k 个中心。

5.已选择初始中心，请继续使用标准 k 均值聚类。

这种播种方法显著改善了 k 均值的最终误差。虽然算法中的初始选择需要额外的时间，但 k 均值部分本身在此种子设定后收敛得非常快，因此算法实际上降低了计算时间。作者使用真实和合成数据集测试了他们的方法，并且通常获得了2倍的速度提高，对于某些数据集，误差提高了近1000倍。在这些模拟中，新方法在速度和误差方面几乎总是至少与普通k均值一样好。

此外，作者为他们的算法计算了近似比。k-means++ 算法保证期望值中的近似比 O（log k）（在算法的随机性上），其中{\displaystyle k}k是使用的聚类数。这与普通 k 均值相反，后者可以生成任意比最优值更差的聚类。[6] k-means++ 相对于任意距离的性能的推广在 中提供。[7]

下表是经过数据处理后的数据。



使用Python 进行K-mean++ 聚类。

### 分析步骤 1. 根据字段进行聚类类别差异性分析; 2. 根据聚类汇总分析各聚类类别的频数; 3. 根据数据集聚类标注可以知道每一个样本数据被分到哪个类别; 4. 聚类中心坐标可以用于分析各样本与中心点的距离; 5. 对分析进行综述。

### 聚类分析结果 输出结果1：字段差异性分析

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 聚类类别（平均值±标准差） | | | | | | F | P |
| 类别1(n=10) | 类别2(n=8) | 类别3(n=4) | 类别6(n=2) | 类别4(n=2) | 类别5(n=1) |
| 澄清度2 | 3.01±0.032 | 3.0±0.0 | 3.0±0.0 | 3.0±0.0 | 3.0±0.0 | 3.0±nan | 0.294 | 0.911 |
| 色调2 | 5.96±0.853 | 6.0±0.0 | 2.5±1.0 | 7.0±1.414 | 2.0±0.0 | 6.0±nan | 24.596 | 0.000\*\*\* |
| 香气浓度2 | 5.75±0.635 | 6.5±0.535 | 6.25±0.5 | 7.0±0.0 | 7.0±0.0 | 4.0±nan | 6.671 | 0.001\*\*\* |
| 口感浓度2 | 4.77±0.998 | 5.0±1.069 | 5.5±1.0 | 6.0±0.0 | 5.0±1.414 | 4.0±nan | 0.863 | 0.522 |
| 香气纯正度2 | 4.06±0.755 | 4.875±0.354 | 3.5±1.0 | 5.0±0.0 | 4.5±0.707 | 3.0±nan | 3.768 | 0.014\*\* |
| 口感持久性2 | 5.5±0.527 | 5.375±0.518 | 5.75±0.5 | 6.0±0.0 | 6.0±0.0 | 5.0±nan | 1.263 | 0.316 |
| 口感质量2 | 15.16±1.362 | 15.25±1.389 | 14.5±1.732 | 16.0±0.0 | 16.0±0.0 | 13.0±nan | 0.994 | 0.445 |
| 香气质量2 | 11.48±0.865 | 13.25±1.488 | 11.5±1.0 | 14.0±0.0 | 14.0±0.0 | 10.0±nan | 5.548 | 0.002\*\*\* |
| 平衡/整体评价2 | 8.94±0.19 | 9.0±0.0 | 9.0±0.0 | 9.0±0.0 | 9.0±0.0 | 8.0±nan | 12.201 | 0.000\*\*\* |
| 口感纯正度2 | 3.88±0.316 | 4.0±0.0 | 3.75±0.5 | 4.0±0.0 | 4.0±0.0 | 3.0±nan | 2.589 | 0.056\* |
| 总和2 | 68.51±1.43 | 72.25±1.282 | 65.25±1.258 | 77.0±1.414 | 70.5±0.707 | 59.0±nan | 41.505 | 0.000\*\*\* |
| 注：\*\*\*、\*\*、\*分别代表1%、5%、10%的显著性水平 | | | | | | | | |

**图表说明：**

上表展示了定量字段差异性分析的结果，包括均值±标准差的结果、F检验结果、显著性P值。  
● 分析每个分析项是否小于0.05或者0.01（根据检验标准要求，严格的话使用0.01）;  
● 若呈显著性，拒绝原假设，说明两组数据之间存在显著性差异，可以根据均值±标准差的方式对差异进行分析，反之则表明数据不呈现差异性。

方差分析的结果显示:  
对于变量澄清度2，显著性P值为0.911，水平上不呈现显著性，不能拒绝原假设，说明变量澄清度2在聚类分析划分的类别之间不存在显著性差异；  
对于变量色调2，显著性P值为0.000\*\*\*，水平上呈现显著性，拒绝原假设，说明变量色调2在聚类分析划分的类别之间存在显著性差异；  
对于变量香气浓度2，显著性P值为0.001\*\*\*，水平上呈现显著性，拒绝原假设，说明变量香气浓度2在聚类分析划分的类别之间存在显著性差异；  
对于变量口感浓度2，显著性P值为0.522，水平上不呈现显著性，不能拒绝原假设，说明变量口感浓度2在聚类分析划分的类别之间不存在显著性差异；  
对于变量香气纯正度2，显著性P值为0.014\*\*，水平上呈现显著性，拒绝原假设，说明变量香气纯正度2在聚类分析划分的类别之间存在显著性差异；  
对于变量口感持久性2，显著性P值为0.316，水平上不呈现显著性，不能拒绝原假设，说明变量口感持久性2在聚类分析划分的类别之间不存在显著性差异；  
对于变量口感质量2，显著性P值为0.445，水平上不呈现显著性，不能拒绝原假设，说明变量口感质量2在聚类分析划分的类别之间不存在显著性差异；  
对于变量香气质量2，显著性P值为0.002\*\*\*，水平上呈现显著性，拒绝原假设，说明变量香气质量2在聚类分析划分的类别之间存在显著性差异；  
对于变量平衡/整体评价2，显著性P值为0.000\*\*\*，水平上呈现显著性，拒绝原假设，说明变量平衡/整体评价2在聚类分析划分的类别之间存在显著性差异；  
对于变量口感纯正度2，显著性P值为0.056\*，水平上不呈现显著性，不能拒绝原假设，说明变量口感纯正度2在聚类分析划分的类别之间不存在显著性差异；  
对于变量总和2，显著性P值为0.000\*\*\*，水平上呈现显著性，拒绝原假设，说明变量总和2在聚类分析划分的类别之间存在显著性差异；

**输出结果2：**

上表展示了模型聚类的结果，包括频数，所占百分比。

**聚类汇总**

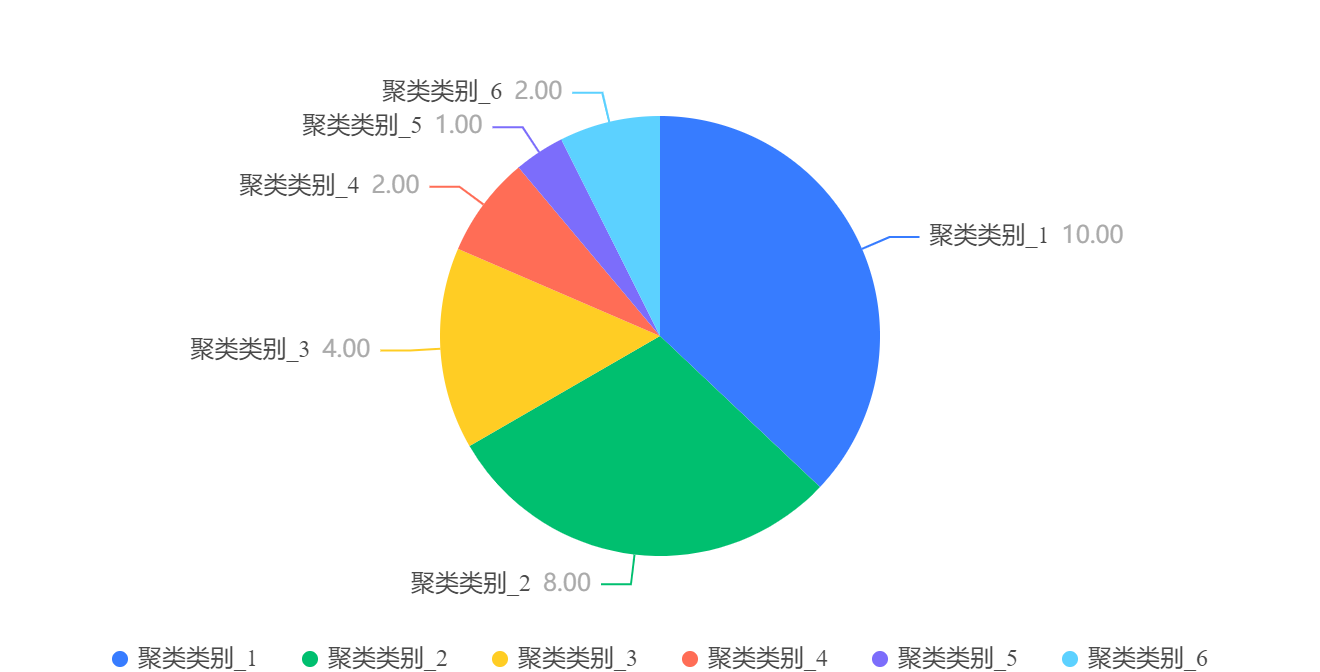
|  |  |  |
| --- | --- | --- |
| 聚类类别 | 频数 | 百分比% |
| 聚类类别\_1 | 10 | 37.037% |
| 聚类类别\_2 | 8 | 29.63% |
| 聚类类别\_3 | 4 | 14.815% |
| 聚类类别\_4 | 2 | 7.407% |
| 聚类类别\_5 | 1 | 3.704% |
| 聚类类别\_6 | 2 | 7.407% |
| 合计 | 27 | 100.0% |

**图表说明：**

**智能分析**

聚类分析的结果显示，聚类结果共分为6类，  
聚类类别\_1的频数为10，所占百分比为37.037%；  
聚类类别\_2的频数为8，所占百分比为29.63%；  
聚类类别\_3的频数为4，所占百分比为14.815%；  
聚类类别\_4的频数为2，所占百分比为7.407%；  
聚类类别\_5的频数为1，所占百分比为3.704%；  
聚类类别\_6的频数为2，所占百分比为7.407%。

**输出结果3：聚类汇总图**



**图表说明：**

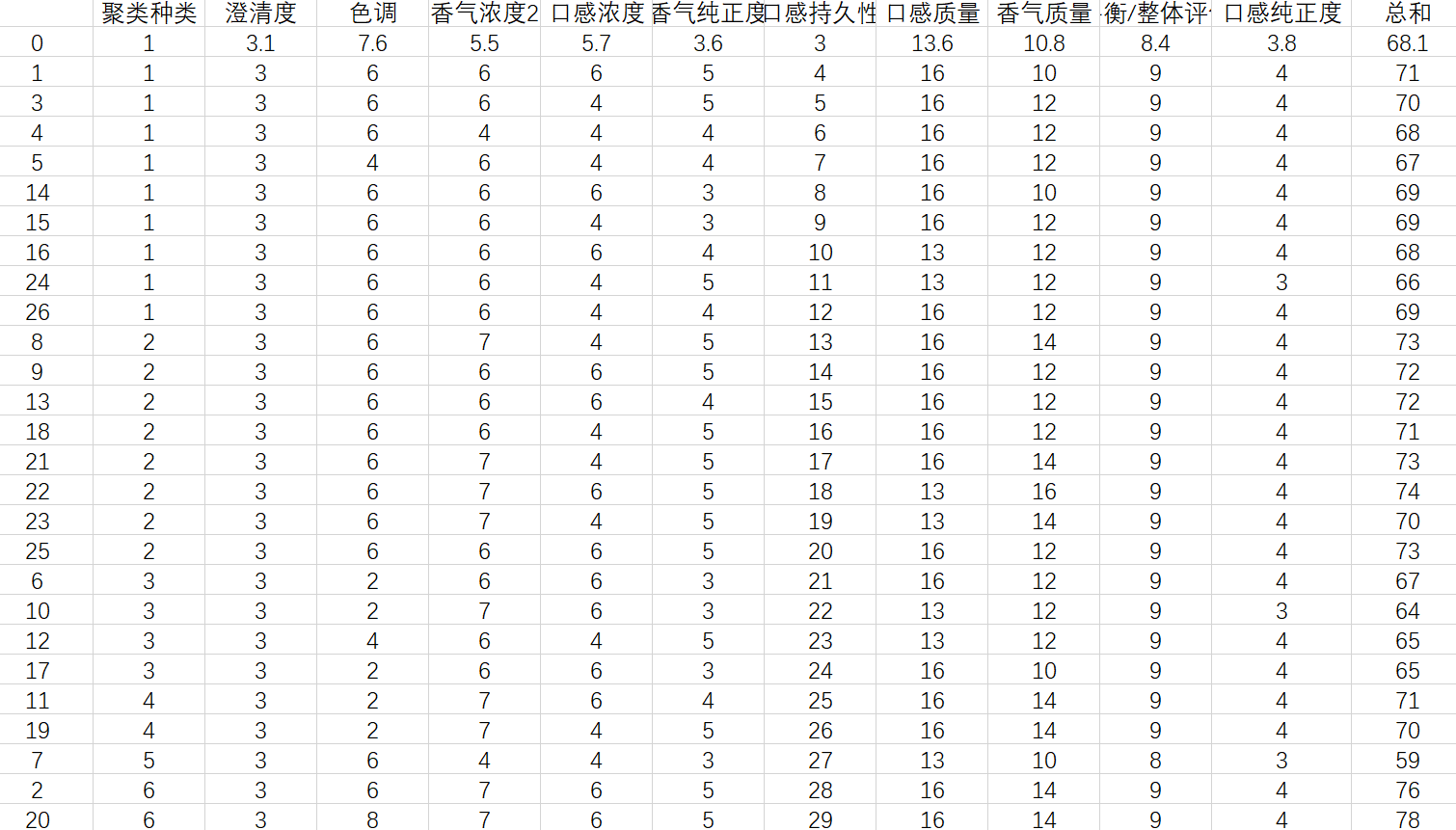
上表格展示了模型聚类结果的部分数据聚类标注，其为预览结果，只显示综合排序的前10条数。

**输出结果5：聚类中心点坐标**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 聚类种类 | 中心值\_澄清度2 | 中心值\_色调2 | 中心值\_香气浓度2 | 中心值\_口感浓度2 | 中心值\_香气纯正度2 | 中心值\_口感持久性2 | 中心值\_口感质量2 | 中心值\_香气质量2 | 中心值\_平衡/整体评价2 |
| 1 | 3.01 | 5.96 | 5.75 | 4.77 | 4.06 | 5.5 | 15.16 | 11.48 | 8.94 |
| 2 | 3 | 6 | 6.5 | 5 | 4.875 | 5.375 | 15.25 | 13.25 | 9 |
| 3 | 3 | 2.5 | 6.25 | 5.5 | 3.5 | 5.75 | 14.5 | 11.5 | 9 |
| 4 | 3 | 2 | 7 | 5 | 4.5 | 6 | 16 | 14 | 9 |
| 5 | 3 | 6 | 4 | 4 | 3 | 5 | 13 | 10 | 8 |
| 6 | 3 | 7 | 7 | 6 | 5 | 6 | 16 | 14 | 9 |

我们可以明显的发现，算法自动将评分接近的样本酒聚为一类，随着聚类种类编号数的上升，酒样的评分越高。

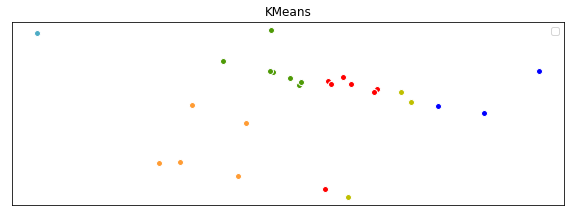
红酒：



白酒：



这个是聚类后各个特征点在高维空间到二维坐标空间的投影，不同颜色表示不同类。



层次分析法

1. Arthur, D.; Vassilvitskii, S. (2006). "How slow is the k-means method?". Proceedings of the twenty-second annual symposium on Computational geometry. ACM New York, NY, USA. pp. 144–153.

6.Kanungo, T.; Mount, D.; Netanyahu, N.; Piatko, C.; Silverman, R.; Wu, A. (2004), "A Local Search Approximation Algorithm for k-Means Clustering" (PDF), Computational Geometry: Theory and Applications, 28 (2–3): 89–112, doi:10.1016/j.comgeo.2004.03.003, archived from the original (PDF) on 2006-02-09.

7. Nielsen, Frank; Nock, Richard (2013), Total Jensen divergences: Definition, Properties and k-Means++ Clustering, arXiv:1309.7109, Bibcode:2013arXiv1309.7109N.