# Lab 2 Unsupervise 实验报告

何春望 PB17000075

1. 代码思想

PCA：各列减去平均值，得到均值为0的数据，然后调用numpy求协方差矩阵，并得到其特征值和特征向量。根据阈值选择前k个特征向量，然后将原始数据与对应特征向量相乘，即可得到降维投影。

kMeans：随机生成n个中心点，每次迭代先遍历数据，将其分配给最近的中心簇；然后在每个簇内计算所有点的平均值，作为新的簇中心，然后进行下一轮迭代，直到没有数据点所属的簇需要改变，即可得到n个中心点和每组数据所属的簇编号。

1. threshold

原数据共13维，按影响由大到小排序后前k维阈值如下表：

|  |  |
| --- | --- |
| k | threshold |
| 1 | 0.9980912304918977 |
| 2 | 0.9998271461166034 |
| 3 | 0.9999221050741548 |
| 4 | 0.999972322430337 |
| 5 | 0.9999846861150249 |
| 6 | 0.9999931482453601 |
| 7 | 0.9999959550599172 |
| 8 | 0.99999747814045 |
| 9 | 0.9999986059708893 |
| 10 | 0.9999993273867002 |
| 11 | 0.9999997054469667 |
| 12 | 0.9999999174607213 |
| 13 | 1.0 |

1. 分为不同类别数的结果

分别分为2~9簇进行测试，得到结果如下

使用PCA时：

|  |  |  |
| --- | --- | --- |
| k | RI | SI |
| 9 | 0.6893290166952326 | 0.5868245237414554 |
| 8 | 0.697010093315559 | 0.5220523354567715 |
| 7 | 0.7096426077572526 | 0.597366570627867 |
| 6 | 0.6850758585666222 | 0.5698101861839927 |
| 5 | 0.691931695550054 | 0.5832743149720723 |
| 4 | 0.702723290801752 | 0.5839849924882091 |
| 3 | 0.718656763791024 | 0.5880872417516695 |
| 2 | 0.670285025074589 | 0.6644828486489767 |

不使用PCA时：

|  |  |  |
| --- | --- | --- |
| k | RI | SI |
| 9 | 0.6968831333714213 | 0.5293792834180435 |
| 8 | 0.6898368564717832 | 0.5219061472604561 |
| 7 | 0.7093252078969086 | 0.5638755981125787 |
| 6 | 0.7012632514441693 | 0.5421550993884968 |
| 5 | 0.6914873357455723 | 0.5640175937193861 |
| 4 | 0.7032311305783026 | 0.5587089480903829 |
| 3 | 0.718656763791024 | 0.571138193786884 |
| 2 | 0.670285025074589 | 0.6568536504294317 |

可见k=2时轮廓系数SI取到最大值，兰德系数RI均在70%左右。

1. 结果分析

两个表格相似，是因为属性集中有一个属性占了绝大多数影响，这一点由threshold也可看出。对比源数据，可知Proline对红酒品质起到决定性作用，其值越高红酒品质越好。对数据进行二分时，临界点大约在880左右。可视化结果：

