实验五：构建聚类模型实现各地区经济情况聚类分析

主讲教师：严资林

实验教师：匡慈维

目录

本学期实验总体安排

本学期实验课程共 10 个学时，5 个实验项目，总成绩为 20 分。

实验项目	一	二	三	四	五
学时	2	2	2	2	2
实验内容	感知机模型	决策树模型	K近邻模型	支持向量机模型	聚类模型
分数	3	4	4	5	4
上课时间(地点)	第11周 周四 (T2102)	第12周 周六 (T2102)	第14周 周四 (T2102)	第16周 周二 (T2102)	第17周 周四 (T2102)
检查方式	提交实验截图文档	提交实验报告、工程文件			

5-6节 3&4班 ； 7-8节 1&2班

线上腾讯会议：848-8762-6539

实验任务

背景

国家统计局采用分层、多阶段、与人口规模大小成比例的概率抽样方法，在全国31个省（区、市）的1800个县（市、区）随机抽选16万个居民家庭作为调查户，调查居民的可支配收入。

居民可支配收入是指居民可用于最终消费支出和储蓄的总和，即居民可用于自由支配的收入。按照收入的来源，可支配收入包括工资性收入、经营净收入、财产净收入和转移净收入。

现有2016年31个地区的农村居民人均可支配收入情况数据，需要根据各地区的经济情况做聚类分析。

任务

使用Python自编程定义**k-均值聚类**和**层次聚类**模型，实现各地区经济情况的聚类分析。

数据说明

◆ 数据集

- 包含**31条**数据记录；
- 每条数据包含5个属性特征，分别为：

地区、工资性收入、经营净收入、财产净收入、转移净收入

	地区	工资性收入	经营净收入	财产净收入	转移净收入
2	北京	16637.5	2061.9	1350.1	2260
3	天津	12048.1	5309.4	893.7	1824.4
4	河北	6263.2	3970	257.5	1428.6
5	山西	5204.4	2729.9	149	1999.1
6	内蒙古	2448.9	6215.7	452.6	2491.7
7	辽宁	5071.2	5635.5	257.6	1916.4
8	吉林	2363.1	7558.9	231.8	1969.1
9	黑龙江	2430.5	6425.9	572.7	2402.6
10	上海	18947.9	1387.9	859.6	4325
11	江苏	8731.7	5283.1	606	2984.8
12	浙江	14204.3	5621.9	661.8	2378.1
13	安徽	4291.4	4596.1	186.7	2646.2
14	福建	6785.2	5821.5	255.7	2136.9
15	江西	4954.7	4692.3	204.4	2286.4
16	山东	5569.1	6266.6	358.7	1759.7
17	河南	4228	4643.2	168	2657.6
18	湖北	4023	5534	158.6	3009.3
19	湖南	4946.2	4138.6	143.1	2702.5
20	广东	7255.3	3883.6	365.8	3007.5
21	广西	2848.1	4759.2	149.2	2603
22	海南	4764.9	5315.7	139.1	1623.1
23	重庆	3965.6	4150.1	295.8	3137.3
24	四川	3737.6	4525.2	268.5	2671.8
25	贵州	3211	3115.8	67.1	1696.3
26	云南	2553.9	5043.7	152.2	1270.1
27	西藏	2204.9	5237.9	148.7	1502.3
28	陕西	3916	3057.9	159	2263.6
29	甘肃	2125	3261.4	128.4	1942
30	青海	2464.3	3197	325.2	2677.8
31	宁夏	3906.1	3937.5	291.8	1716.3
32	新疆	2527.1	5642	222.8	1791.3

预备知识

数据标准化

- 数据标准化就是用来消除不同量级的影响

（1）min-max标准化（归一化）

- 新数据=(原数据-最小值)/(最大值-最小值)

（2）z-score标准化（规范化）

- 新数据=(原数据- 均值)/ 标准差
- 推荐使用**sklearn.preprocessing.StandardScaler()**函数

预备知识

K-均值聚类

◆ **基本步骤：**

Step1：随机初始化**K个聚类中心**，即K个类中心向量；

Step2：对每个样本，计算其与各个类中心向量的距离(这里指 欧式距离 $d_{ij}=\sqrt{\sum_{m=1}^{n}(x_{im}-x_{jm})^2}$)，并将该样本指派给距离最小的类；

Step3：更新每个类的中心向量，更新的方法为取该类所有样本的**特征向量均值**；

Step4：直到各个类的中心向量不再发生变化为止，作为退出条件。

预备知识

层次聚类（以聚合聚类为例）

◆ **基本步骤：**

•Step1：每个样本自成一类；

•Step2：计算N个样本两两之间的**距离**，并将其中距离最近的点聚成一个小类，得到N-1个小类；

•Step3：计算余下样本点和小类间的距离，并将当前距离最近的点或小类再聚成一个类；

•Step4：重复上述过程，不断将所有样本和小类聚集成新的类，直到类的个数为**k**（人为设置的）

预备知识

评价指标——轮廓系数

- 聚类是没有标签，即不知道真实答案的预测算法，我们必须 完全依赖评价**簇内的稠密程度（簇内差异小）**和**簇间的离散程度（簇外差异大）**来评估聚类的效果。其中**轮廓系数**是最常用的聚类算法的评价指标，它能够同时衡量：

- （1） 样本与其自身所在的簇中的其他样本的相似度，等于样本与同一簇中所有其他点之间的平均距离，这个距离记作**a**。
- （2）样本与其他簇中的样本的相似度，等于样本与下一个最近的簇中的所有点之间的平均距离，这个距离记作**b**。
- （3）单个样本的轮廓系数计算为：

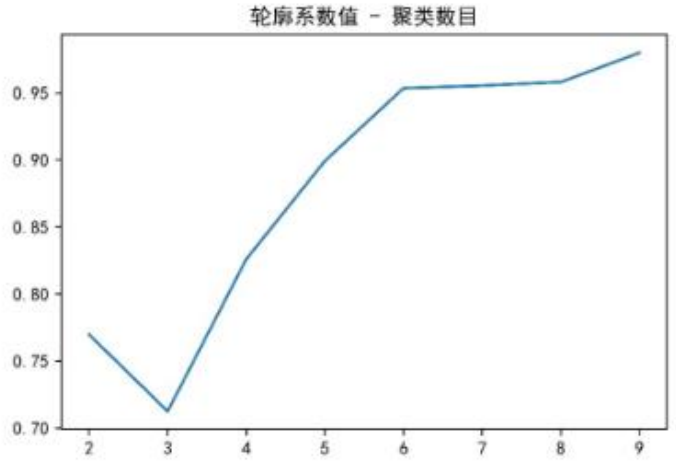
$$S = \frac{\mathrm{b} - a}{\max(a, b)}$$

预备知识

评价指标——轮廓系数

在sklearn.metrics模块中有用来评价聚类结果的指标，如下表：

方法名	真实值	最佳值	Sklearn函数
轮廓系数评价法	不需要	畸变程度最大	sihouette_score



说明：由轮廓系数图可以看出，聚类数目为2-3和3-4时，平均畸变程度较大，所以k取2/3/4都可以。

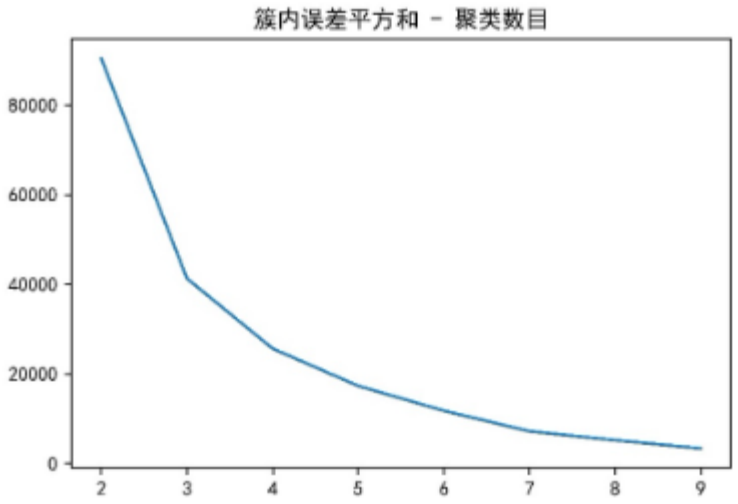
预备知识

评价指标——簇内误差平方和

对于一个簇来说，所有样本点到质心的距离之和越小，我们就认为这个簇中的样本越相似，簇内差异就越小。簇内误差平方和

$$SSE=\sum_{j=0}^{k}\sum_{i=1}^{n}(x_i-\mu_i)^2$$

其中 x 表示簇中的一个样本点，μ 表示簇中的质心（也即簇内样本的均值）， n表示每个样本点中的特征数目，i表示组成点x的每个特征，k表示簇的个数。



说明：随着聚类数k的增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，那么误差平方和SSE自然会逐渐变小。并且，当k小于真实聚类数时，由于k的增大会大幅增加每个簇的聚合程度，故SSE的下降幅度会很大，而当k到达真实聚类数时，再增加k所得到的聚合程度回报会迅速变小，所以SSE的下降幅度会骤减，然后随着k值的继续增大而趋于平缓。所以也称为**手肘法**。

实验步骤

◆ 实验步骤 （使用Python自编程）

1、准备数据

✓读取数据，将数据进行规范化处理

2、定义模型

✓根据算法定义，完成k-均值/层次聚类模型的编程

3、训练模型

✓训练模型，调整k值，根据评价指标找到合适的k值

4、评估模型

✓根据合适的k值，对数据做聚类，并评估模型

实验要求

1、数据集标准化处理；

2、使用python自编程完成k-均值和层次聚类的模型定义；

3、记录调参过程和结果，根据评价指标，选出最合适的K值；

4、使用轮廓系数评价法和簇内误差平方和指标来评价模型。

5、用图/表格形式展示出最后的分类结果，以及质心（k-means方法需要）。

提交方式

实验报告提交至平台 http://grader.tery.top:8000/#/courses

注意：

- 1、用户名、密码默认均为学号（若之前有修改过密码的，请用新密码登陆）；
- 2、请提交到相应的条目「2022春统计机器学习」课程 - 实验五；
- 3、提交截止时间：下周四 晚24点前；
- 4、文件夹&压缩包命名要求： 学号_姓名_统计机器学习实验五
- 5、提交内容：实验报告(.pdf文件)+代码(.py文件)，一起打包为zip格式压缩包。

统计机器学习实验

同学们，请开始实验吧！