03-Machine Learning Fundamentals

课件：数据分析与机器学习\_2022年5月27日.pdf

试验环境链接：https://pan.baidu.com/s/1C43EqMM9oV66GLT-ikkFKg

提取码：n24t

培训录制链接：https://cisco.webex.com/recordingservice/sites/cisco/recording/playback/7594456fbf8a103abf1c00505681061a

密码：59JcU2SD

本文档所在的子目录machine learning下，有二本关于机器学习的书籍。由于机器学习也是当前非常热门的一门学科分支和职业，网上相应的书籍也非常多，可以自行搜索。

请回答下面题目

1. Python语言中关于机器学习涉及到哪些第三方工具库？请列举出来（自行在网上搜索）

机器学习使计算机从研究数据和统计数据中学习，是向人工智能（AI）方向迈进的一步。

机器学习是一个分析数据，自我学习，并预测结果的程序，所以它一定会涉及到数据的分析和处理，故人们通常也会把数据分析和可视化也列为机器学习中。

4种工具库

|  |  |
| --- | --- |
| 工具库 | 简述 |
| Numpy | 科学计算库 |
| Pandas | 数据加载，处理，和保存 |
| Matplotlib | 可视化 |
| sk-learn | 机器学习，基于Numpy和Scipy的库 |

6种工具库

|  |  |
| --- | --- |
| 工具库 | 简述 |
| NumPy | 科学计算 |
| Scipy | 科学计算和技术计算 |
| Scikit-learn | 也就是sk-learn，机器学习 |
| Tensorflow | 深度学习，是google |
| Pytorch | 机器学习，是facebook |
| Keras | 神经网络库，卷积，循环神经网络 |

10种工具库

|  |  |
| --- | --- |
| 工具库 | 简述 |
| TensorFlow | 深度学习，是google |
| Scikit-Learn | 也就是sk-learn，机器学习 |
| Numpy | 科学计算 |
| Keras | 神经网络库，卷积，循环神经网络 |
| PyTorch | 机器学习，是facebook |
| LightGBM | 机器学习库 |
| Eli5 | 机器学习框架，内置许多其它学习机器学习库 |
| SciPy | 科学计算和技术计算 |
| Theano | 多维数组计算 |
| Pandas | 数据加载，处理，和保存 |

在本培训学习中，我们先从简单的入手，即涉及到4种工具库。

1. 机器学习会用到大量数据运算，特别是矩阵运算。请用Python演示一个4x4矩阵的加法，减法，乘法（包括矩阵点乘和矩阵相乘）

在矩阵的运算中，涉及到同维度矩阵的运算和不同维度矩阵的运算，甚至矩阵和数值之间的运算，大致可分为如下三种情况：

1，4x4的矩阵和4x4的矩阵运算

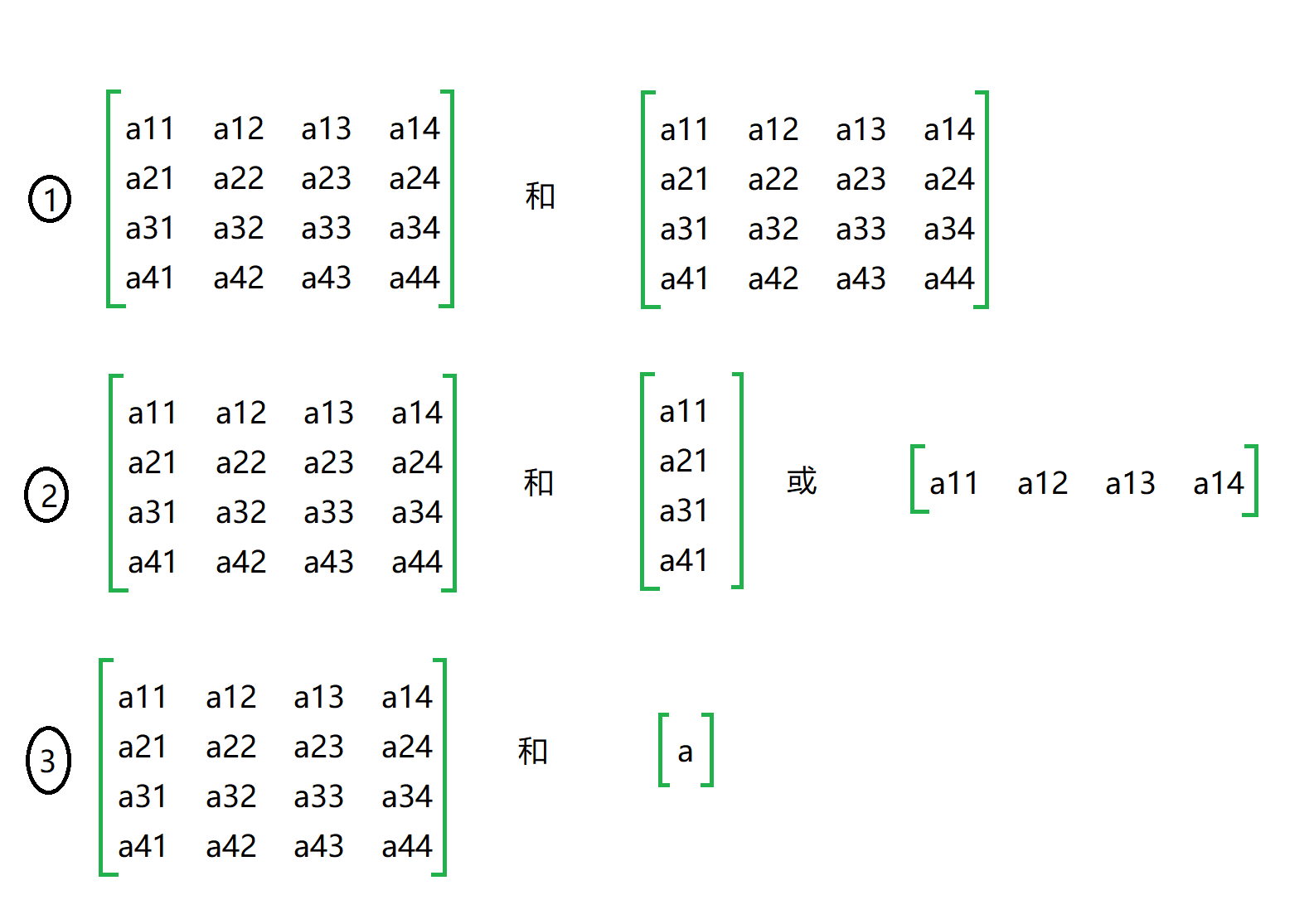
对应位置上的元素进行运算

2，4x4的矩阵和4x1的矩阵或1x4的矩阵

如果是4x1矩阵，则按列进行运算，如果是1x4矩阵，则按行进行运算

3，4x4的矩阵和数值，即1x1的矩阵

数值与所有的元素进行运算



但对于矩阵乘法，分二种情况：相乘和点乘。相乘应用上面规则，点乘是应用线性代数中的矩阵乘法规则，即[mxn] X [nxr] = [mxr]的规则，但是比它有所扩展，具体看案例

1. 在Python代码的调试中，我们经常采用print输出值查看的方法，请问在Python中，print有几种格式化输出？。

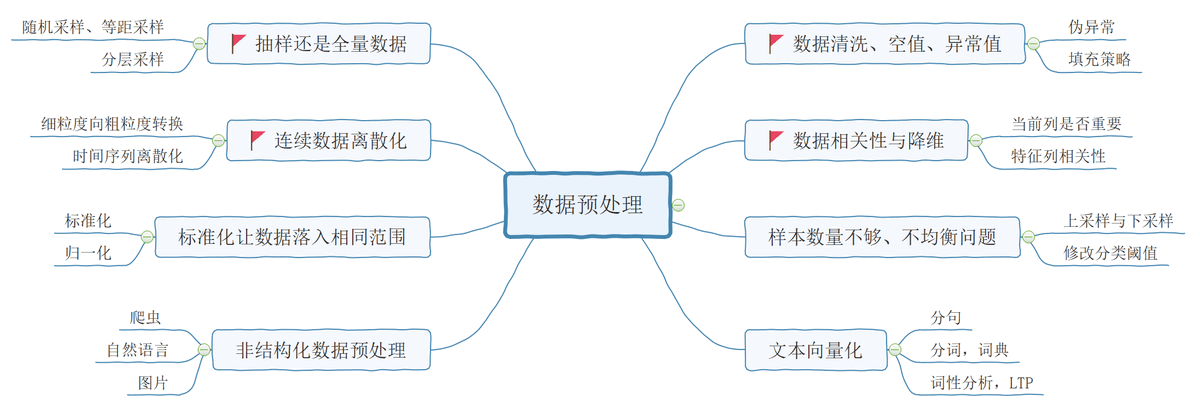
在Python中，print有三种格式化输出方式：

|  |  |  |
| --- | --- | --- |
| 方式 | 格式 | 案例 |
| 1 | print()常用方法 | print("Hello World") |
| 2 | format()函数 | print('{0} {1}'.format('hello','world')) |
| 3 | f-string | print(f'My lucky number is {number}') |
| 参考网址 | https://blog.csdn.net/qq\_45869700/article/details/121178936 | |

1. 请列举出在Python中机器学习的7大步骤。

|  |  |
| --- | --- |
| 步骤 | 描述 |
| 1 | 获取数据源：自有数据、购买数据例如万得、爬虫爬取数据 |
| 2 | 清洗与预处理: (最耗时，空值、异常、抽样、相关性分析) |
| 3 | 拆分数据集：训练集 75% | 测试集 (未来) 25% |
| 4 | 特征工程 (最重要)：标准化、归一化 |
| 5 | 训练模型：选择和训练模型 (聚类、回归、分类) |
| 6 | 模型调优：超参数，特征工程 |
| 7 | 模型保存与加载：参数的保存、迁移学习 |

1. 机器学习是需要从大量的用户数据中找出普通规律，但是数据来源的多样化，用户数据可能存在问题，需要进行预处理，请列出出数据预处理的8大策略。

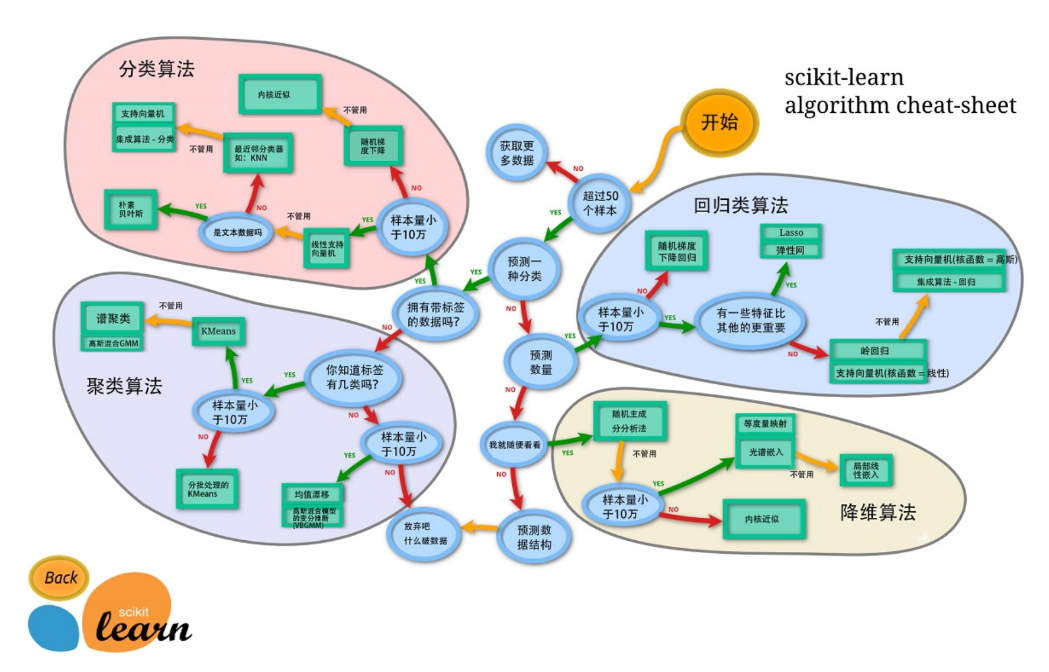


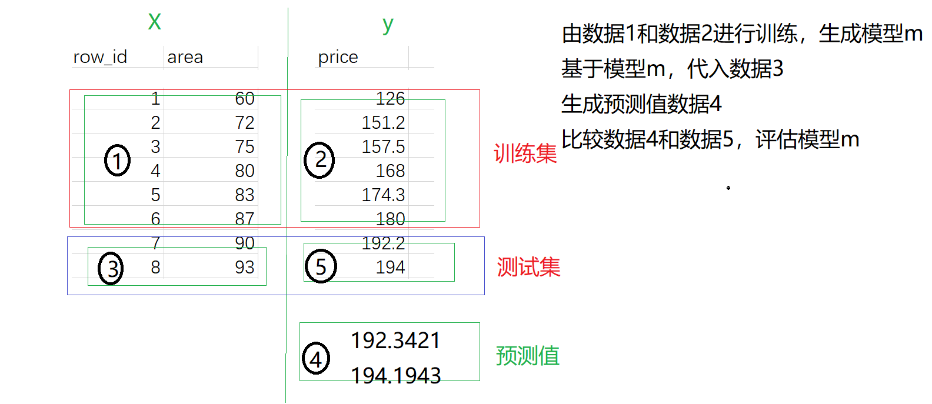
1. 本培训使用sk-learn机器学习库，在sk-learn中有几种常用的算法？回归类算法要对数据如何处理才能满足回归要求？

机器学习库sk-learn有4大类算法：回归类算法，分类算法，聚类算法，降维算法

常规类机器学习主要是回归类算法，又可细分为线性回归，多项式回归（二次多项式，三次多项式）和曲线回归

本培训视频主要涉及到线性回归，二次多项式回归，以及聚类





1. 针对问题6的各种算法，列出算法结果好坏的评估指标。

每一种算法都是针对不同的应用场景，也有对应的算法结果好坏的评估指标

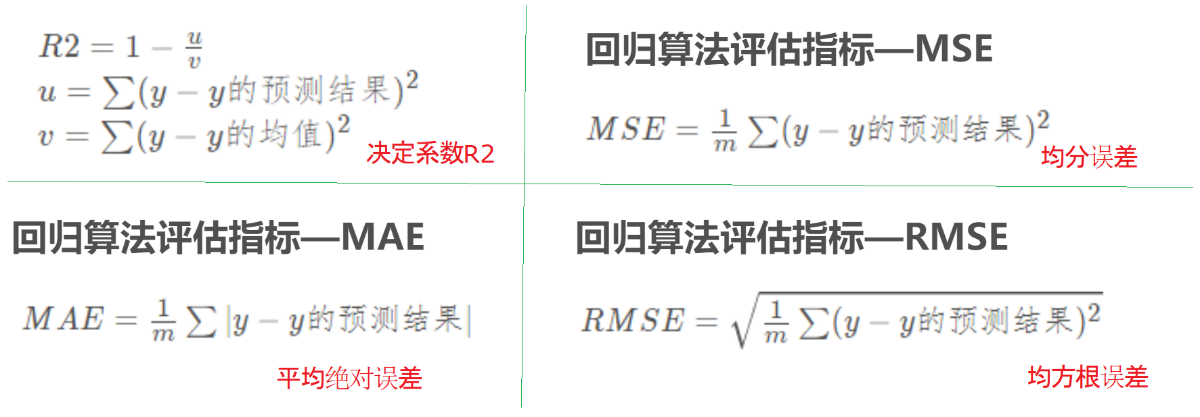
|  |  |  |
| --- | --- | --- |
| 算法 | 指标 |  |
| 回归类算法 | R方范围[-inf,1] (真实值与预测 ),MSE,MAE,RMSE |  |
| 聚类算法 | 轮廓系数 SSE(误差平方和) |  |
| 分类算法 | 准确率,查准率,查全率,F1-score,PR曲线,ROC曲线和AUC指标 |  |
| 参考网址 | <https://blog.csdn.net/qq_43214331/article/details/120099650> | |

由于回归类算法是我们的主要方向，下面重点讨论回归类算法的评估指标

MAE：平均绝对误差

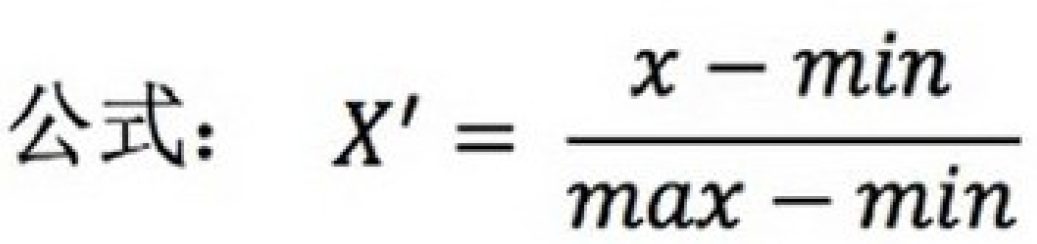
MSE：均方误差， RMSE：均方根误差

R2： 决定系数，用来描述正确率/准确率



1. 什么是归一化？什么是标准化？在机器学习的过程中，它们有什么作用？

用户数据来源多处，表达的物理意义和量纲都是不一样的，有些表示的是时间，有些表示的距离，为了减少数据量纲的影响，我们需要通过对原始数据进行变换，把数据映射到(0~1)之间。



根据公式，当前列最大值归一化为1，当前列最小值归一化为0。

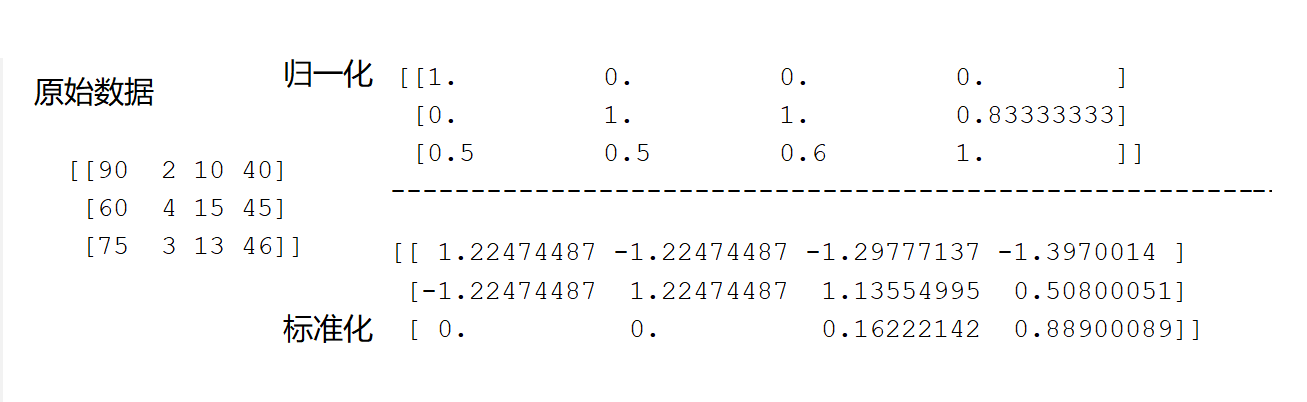
缺点是对异常点的包容性差 (如果异常点影响了最大值或者最小值，则会造成整体数值计算

偏差)，因此这种方式稳定性较差，适合比较精确数据模型

针对归一化的缺点，为了更好地消除异常点的影响，引入标准化概念。

Chart

Description automatically generated with medium confidence 其中：分母σ为标准差，分子mean为平均值



1. 请使用提供的数据house.csv，使用回归类算法找出房价和其它自变量之间的关系。（按照7大步骤的顺序），并输出评估指标。

导入house.csv数据后显示如下：

线性回归算法本质上是线性代数中的线性方程组，根据变量求系数

y = WX + b

其中y, b为一维矩阵，W, X为二维矩阵，根据X和y，求W和b

W，我们称为权重，就是线性方程的斜率，b，我们称为偏置，就是线性方程的截距

如果有二个自变量x1和x2，则可展开为：

y = w1x1 + w2x2 + b

根据x1, x2和y，求w1，w2和b

1. 回归类算法的本质就是数据拟合，数据拟合有几种情况，请用图示说明。

数据拟合有三种情况，分别是欠拟合、合适拟合和过拟后

Chart, scatter chart

Description automatically generated

我们可以通过对数据进行可视化，大致确定采用什么模型进行数据拟合

比如，对boston房价数据进行可视化

1. 请使用sk-learn库提供的boston数据，分析房价和其它自变量之间的关系，并输出评估指标。

我们按照最简单的线性回归模型对boston房屋价格数据进行分析。

步骤参考问题8，只是数据来源来同。

通过对比测试集的预测值和真实值，以及决定系数R2，我们发现这个模型的准确率并不是太高，接下来我们就要进行数据对比和分析，对预测模型进行调优。

Text

Description automatically generated

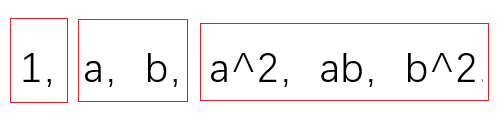
1. 重复问题11，如何提高评估指标？并用新的方法重新训练，比较评估指标是否提高？（提示：采用升维的方式，使用多项式拟合）

本题是针对问题11的模型准确率不高，对模型进行调优。

通过问题10中的可视化，我们可以发现boston中13个特征值和目标值并不是线性关系，所以采用的线性回归模型不适合。

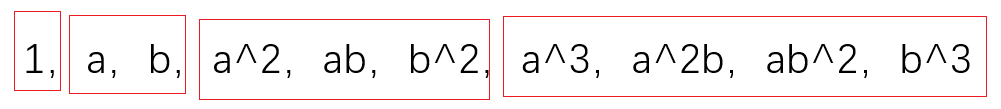
二次多项式：

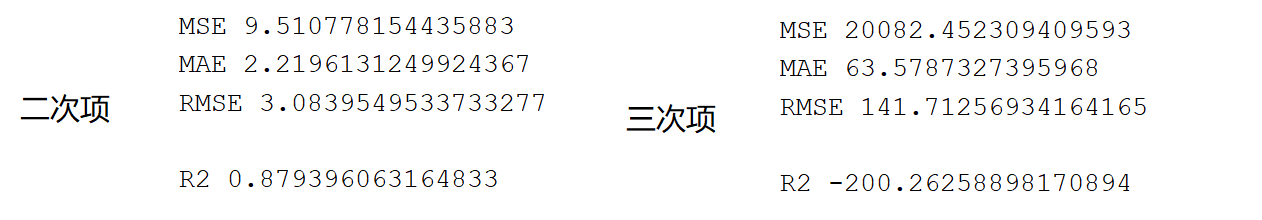
最高项是二次，比如有a，b二个特征，其二次多项式则为



三次多项式：

最高项是三次，同时a，b二个特征的三次多项式则为





1. 什么是K-means算法，它属于机器学习中的哪一类？

K-means是聚类算法中的一种，其中K表示类别数，means表示均值。顾名思义K-means

是一种通过均值对特征数据点进行聚类的算法。

K-means算法通过预先设定的K值及每个类别的初始质心对相似的数据点进行划分，并通过划分后的均值迭代优化获得最优的聚类结果。

K-means选取的初值并不重要，只是影响迭代次数，不会影响最后的结果

1. 请使用提供的数据mobile.xls，对用户数据进行分析，并结合可视化方法，判断这些用户生活在什么社区圈？

要求：

1，以mobile.xls为数据来源

2，数据进行归一化或标准化处理

3，先对数据进行聚类

4，对每一类数据进行可视化，给出结论

可能商圈的分析：

1. 商业区：日均人流量大，工作日上班、凌晨、周末的人均停留时间相对较短。

2. 住宅区：凌晨、周末人均停留时间相对较长，而工作日上班时间人均停留较短，并

且日均人流量较少。

3. 工作区：工作日上班人均停留时间较长，凌晨、周末人均停留时间相对较短。

样本数据的分析：

基站编号只是对手机基站的唯一表示，跟商圈无关，需要排除

停留时间和人均流量单位不一致，需要进行数据进行标准化

处理步骤分二步：

1. 根据题目要求，利用聚类算法，对数据进行处理，分成三类。

这一步属于数据分析和处理，主要是利用sk-learn中的K-means对数据进行聚类，分成三类，给出类别编号。

1. 对第一步生成的类别编号，判断可能的商圈。

这一步是对每一类的数据进行可视化，根据可视化的图形特点，就可判断出商圈。