法律声明

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

- 课程详情请咨询
  - 微信公众号：大数据分析挖掘
  - 新浪微博：ChinaHadoop

第三讲

pandas

$$y_i t = \beta' x_{it} + \mu_i + \epsilon_{it}$$

数据分析工具Pandas

--梁斌

目录

- Pandas的数据结构
- Pandas的数据操作
- Pandas统计计算和描述
- 数据的分组与聚合
- 数据清洗、合并、转化和重构
- 聚类模型：K-Means
- 实战案例：全球食品数据分析（Open Food Facts）

什么是Pandas

Pandas

- 一个强大的分析结构化数据的工具集
- 基础是NumPy，提供了高性能矩阵的运算
- 应用，数据挖掘，数据分析
  - 如，学生成绩分析、股票数据分析等。
- 提供数据清洗功能

目录

Pandas的数据结构

Series

- 类似一维数组的对象
- 通过list构建Series
  - ser_obj = pd.Series(range(10))
- 由数据和索引组成
  - 索引在左，数据在右
  - 索引是自动创建的
- 获取数据和索引
  - ser_obj.index, ser_obj.values
- 预览数据
  - ser_obj.head(n)

示例代码： 01_pandas_data_structures.ipynb

Pandas的数据结构

Series (续)

- 通过索引获取数据
  - ser_obj[idx]
- 索引与数据的对应关系仍保持在数组运算的结果中
- 通过dict构建Series
- name属性
  - ser_obj.name, ser_obj.index.name

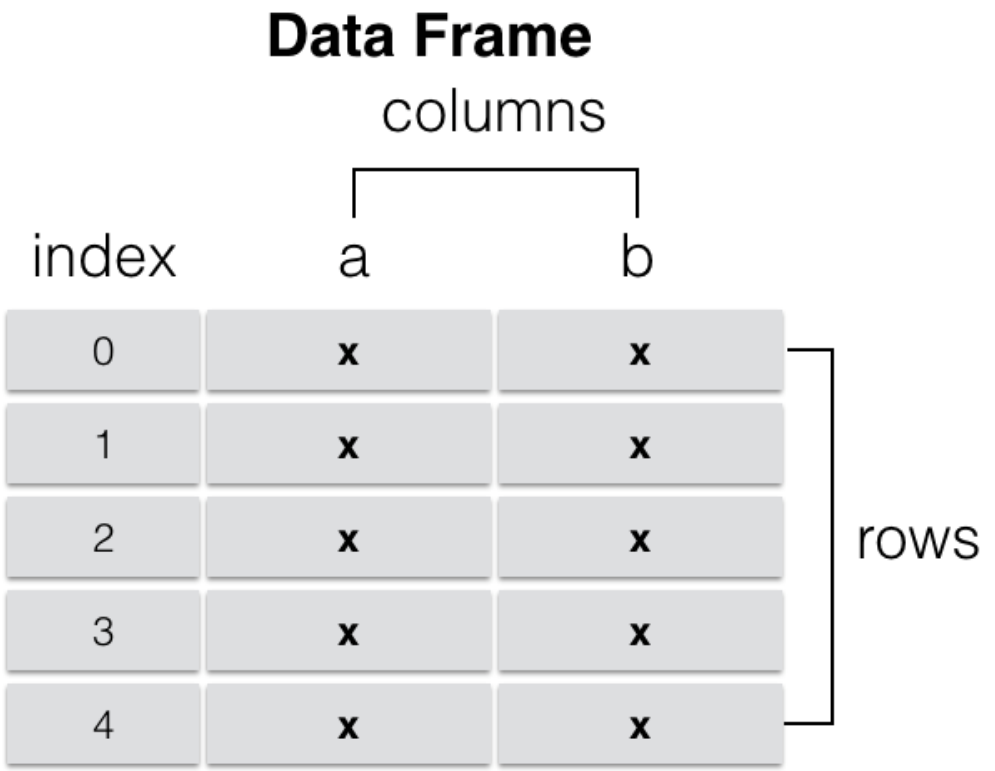
示例代码： 01_pandas_data_structures.ipynb

Pandas的数据结构

DataFrame

示例代码： 01_pandas_data_structures.ipynb

- 类似多维数组/表格数据 (如，excel, R中的data.frame)
- 每列数据可以是不同的类型，what about ndarray?
- 索引包括列索引和行索引



Pandas的数据结构

示例代码： 01_pandas_data_structures.ipynb

DataFrame

- 通过ndarray构建DataFrame
- 通过dict构建DataFrame
- 通过列索引获取列数据（Series类型 ）
  - df_obj[col_idx] 或 df_obj.col_idx
- 增加列数据，类似dict添加key-value
  - df_obj[new_col_idx] = data
- 删除列
  - del df_obj[col_idx]

Pandas的数据结构

索引对象Index

- Series和DataFrame中的索引都是Index对象
- 不可变(immutable)
  - 保证了数据的安全
- 常见的Index种类
  - Index
  - Int64Index
  - MultiIndex，“层级”索引
  - DatetimeIndex，时间戳类型

示例代码： 01_pandas_data_structures.ipynb

目录

Pandas的数据操作

索引操作

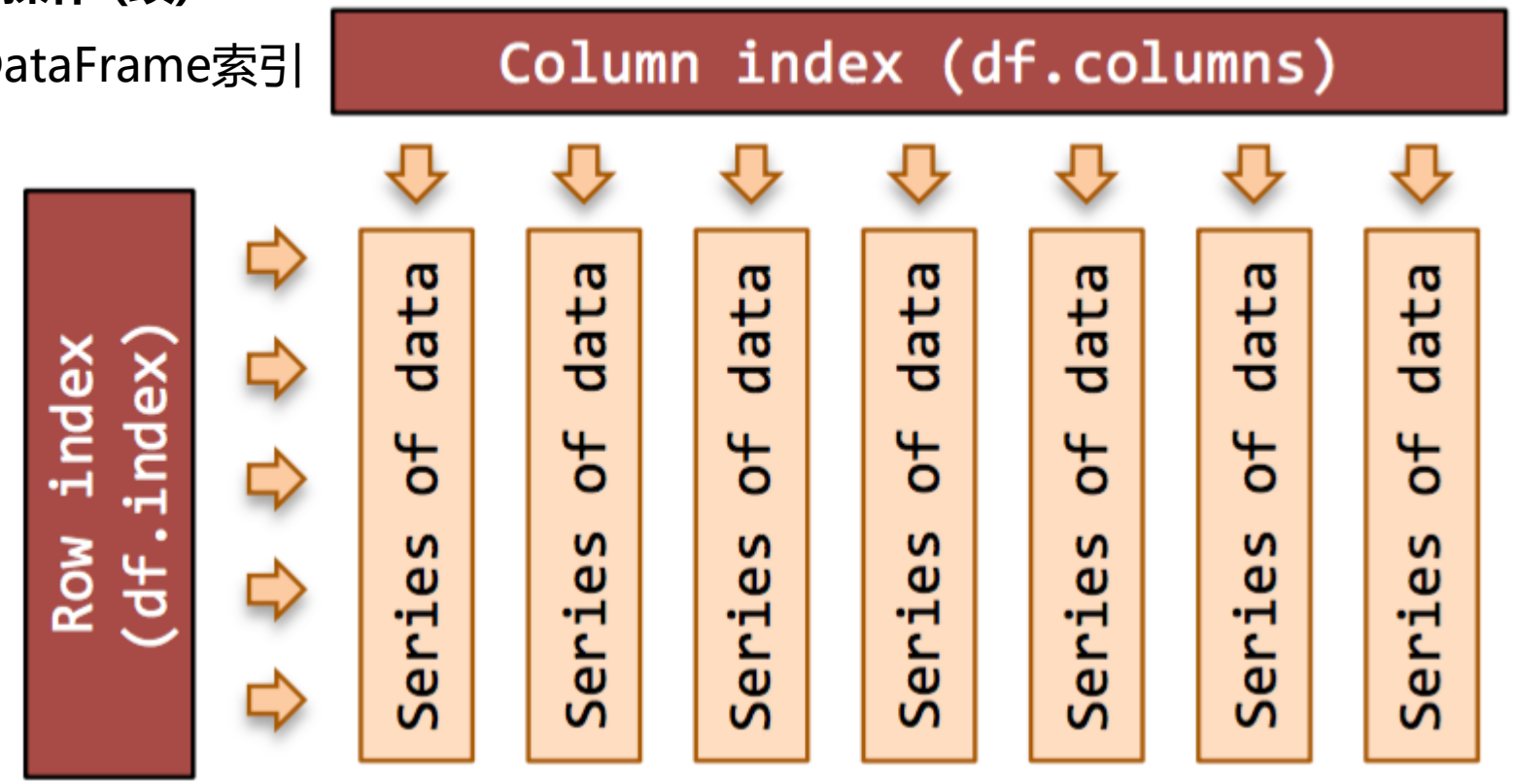
- Series索引
  - 行索引，ser_obj[‘label’], ser_obj[pos]
  - 切片索引，ser_obj[2:4], ser_obj[‘label1’:’label3’]
    - 注意，按索引名切片操作时，是包含终止索引的。
  - 不连续索引，ser_obj[[‘label1’,’label2’, ‘label3’]]
    ser_obj[[pos1, pos2, pos3]]
  - 布尔索引

示例代码： 02_pandas_data_process.ipynb

Pandas的数据操作

索引操作 (续)

- DataFrame索引



示例代码： 02_pandas_data_process.ipynb

Pandas的数据操作

索引操作 (续)

- DataFrame索引
  - 列索引
    - df_obj[‘label’]
  - 不连续索引
    - df_obj[[‘label1’ , ‘label2’]]

示例代码： 02_pandas_data_process.ipynb

Pandas的数据操作

索引操作总结

- Pandas的索引可归纳为3种
- .loc，标签索引
- .iloc，位置索引
- .ix，标签与位置混合索引
  - 先按标签索引尝试操作，然后再按位置索引尝试操作
- 注意
  - DataFrame索引时可将其看作ndarray操作
  - 标签的切片索引是包含末尾位置的

示例代码： 02_pandas_data_process.ipynb

Pandas的数据操作

运算与对齐

- 按索引对齐运算，没对齐的位置补NaN
  - Series 按行索引对齐
  - DataFrame按行、列索引对齐
- 填充未对齐的数据进行运算
  - 使用add, sub, div, mul
  - 同时通过fill_value指定填充值
- 填充NaN
  - fillna

示例代码： 02_pandas_data_process.ipynb

Pandas的数据操作

函数应用

- 可直接使用NumPy的ufunc函数，如abs等
- 通过apply将函数应用到行或列上
  - 注意指定轴的方向，默认axis=0
- 通过applymap将函数应用到每个数据上

示例代码： 02_pandas_data_process.ipynb

Pandas的数据操作

排序

- sort_index，索引排序
  - 对DataFrame操作时注意轴方向
- 按值排序
  - sort_values(by= ‘label’)

示例代码： 02_pandas_data_process.ipynb

Pandas的数据操作

处理缺失数据

- 判断是否存在缺失值
  - ser_obj.isnull(), df_obj.isnull()
- dropna
  - 丢弃缺失数据
- fillna
  - 填充缺失数据

示例代码： 02_pandas_data_process.ipynb

目录

Pandas统计计算和描述

常用的统计计算

- sum, mean, max, min...
- axis=0 按列统计，axis=1按行统计
- skipna 排除缺失值， 默认为True
- idmax, idmin, cumsum

统计描述

- describe 产生多个统计数据

示例代码： 03_pandas_stats.ipynb

Pandas统计计算和描述

方法	说明
count	非NA值的数量
describe	针对Series或各DataFrame列计算汇总统计
min、max	计算最小值和最大值
argmin、argmax	计算能够获取到最小值和最大值的索引位置（整数）
idxmin、idxmax	计算能够获取到最小值和最大值的索引值
quantile	计算样本的分位数（0到1）
sum	值的总和
mean	值的平均数
median	值的算术中位数（50%分位数）
mad	根据平均值计算平均绝对离差
var	样本值的方差
std	样本值的标准差

Pandas统计计算和描述

方法	说明
skew	样本值的偏度（三阶矩）
kurt	样本值的峰度（四阶矩）
cumsum	样本值的累计和
cummin、cummax	样本值的累计最大值和累计最小值
cumprod	样本值的累计积
diff	计算一阶差分（对时间序列很有用）
pct_change	计算百分数变化

目录

Pandas层级索引

层级索引（hierarchical indexing）

- MultiIndex对象
- 选取子集
  - 外层选取 ser_obj[‘outer_label’]
  - 内层选取 ser_obj[:, ‘inner_label’]
- 常用于分组操作、透视表的生成等
- 交换分层顺序
  - swaplevel()
- 排序分层
  - sortlevel()

示例代码： 04_pandas_multi_index.ipynb

Pandas层级索引

层级索引（续）

		0	1	2	3
bar	one	-1.133800	0.548640	1.109034	0.643708
	two	-0.792654	0.518681	-0.611958	0.913413
baz	one	0.775624	-2.520829	-0.472691	-0.557803
	two	0.190005	0.435193	1.635680	1.584821
foo	one	-0.592235	-0.361735	1.336444	-1.280014
	two	-1.016622	1.409086	0.114743	0.408211
qux	one	0.662941	-1.258482	-0.373214	-0.974658
	two	-0.931004	0.596507	0.148323	0.475039

示例代码： 04_pandas_multi_index.ipynb

Pandas分组与聚合

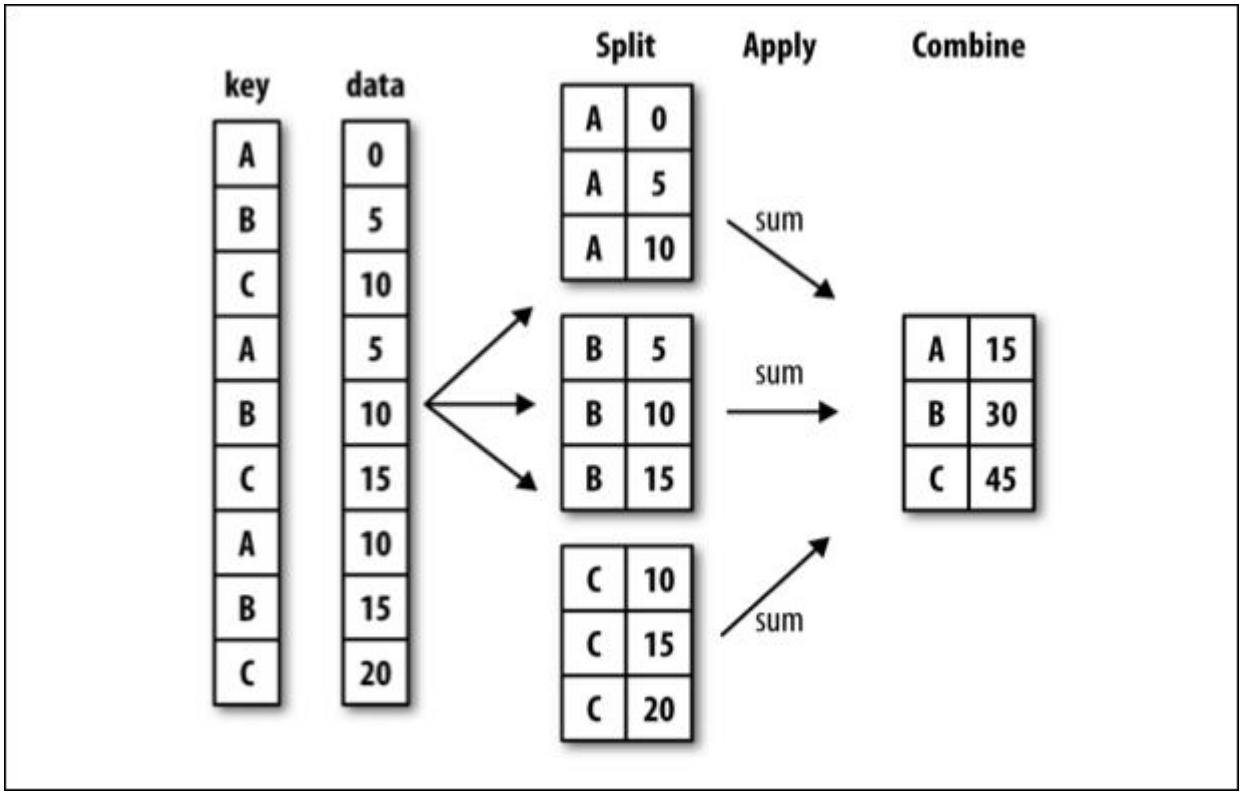
分组 (groupby)

- 对数据集进行分组，然后对每组进行统计分析
- SQL能够对数据进行过滤，分组聚合
- pandas能利用groupby进行更加复杂的分组运算
- 分组运算过程
  - split->apply->combine
  - 拆分：进行分组的根据
  - 应用：每个分组运行的计算规则
  - 合并：把每个分组的计算结果合并起来

Pandas分组与聚合

分组 (续)

- 分组运算过程
  - split->apply->combine



Pandas分组与聚合

分组 (续)

- GroupBy对象：DataFrameGroupBy，SeriesGroupBy
- GroupBy对象没有进行实际运算，只是包含分组的中间数据
- 对GroupBy对象进行分组运算/多重分组运算，如mean()
  - 非数值数据不进行分组运算
- size() 返回每个分组的元素个数

示例代码： 05_pandas_groupby.ipynb

Pandas分组与聚合

分组 (续)

- 按列名分组
  - obj.groupby(‘label’)
- 按列名多层分组
  - obj.groupby([‘label1’ , ‘label2’])->多层dataframe
- 按自定义的key分组
  - obj.groupby(self_def_key)
  - 自定义的key可为列表或多层列表
- unstack可以将多层索引的结果转换成单层的dataframe

示例代码： 05_pandas_groupby.ipynb

Pandas分组与聚合

分组 (续)

- GroupBy对象支持迭代操作
  - 每次迭代返回一个元组 (group_name, group_data)
  - 可用于分组数据的具体运算
- GroupBy对象可以转换成列表或字典
- Pandas也支持按列分组
- 其他分组方法
  - 通过字典分组
  - 通过函数分组，函数传入的参数为行索引或列索引
  - 通过索引级别分组

示例代码： 05_pandas_groupby.ipynb

Pandas分组与聚合

聚合 (aggregation)

- 数组产生标量的过程，如mean()、count()等
- 常用于对分组后的数据进行计算
- 内置的聚合函数
  - sum(), mean(), max(), min(), count(), size(), describe()
- 可自定义函数，传入agg方法中
  - grouped.agg(func)
  - func的参数为groupby索引对应的记录

示例代码： 05_pandas_groupby.ipynb

Pandas分组与聚合

聚合 (续)

- 应用多个聚合函数
  - 同时应用多个函数进行聚合操作，使用函数列表
  - 对不同的列分别作用不同的聚合函数，使用dict

示例代码： 05_pandas_groupby.ipynb

Pandas分组与聚合

聚合 (续)

- 常用的内置聚合函数

函数名	说明
count	分组中非NA值的数量
sum	非NA值的和
mean	非NA值的平均值
median	非NA值的算术中位数
std、var	无偏（分母为n - 1）标准差和方差
min、max	非NA值的最小值和最大值
prod	非NA值的积
first、last	第一个和最后一个非NA值

数据的分组运算

分组运算

- 原因:
  - 聚合运算改变了原始数据的shape
  - 如何保持原始数据的shape?
    - 使用merge的外连接，比较复杂
    - transform
- transform的计算结果和原始数据的shape保持一致
  - 如：grouped.transform(np.mean)
  - 也可传入自定义函数

示例代码： 06_pandas_grouped_apply_transform.ipynb

数据的分组运算

分组运算 (续)

- grouped.apply(func)
  - func函数在各分组上调用，然后结果通过pd.concat组装到一起
  - 产生层级索引
    - 外层索引是分组名
    - 内层索引是df_obj的行索引
  - 禁止层级索引, group_keys=False
- apply可以用来处理不同分组内的缺失数据填充
  - 如：填充该分组的均值

示例代码： 06_pandas_grouped_apply_transform.ipynb

目录

数据清洗

- 数据清洗是数据分析关键的一步，直接影响之后的处理工作
- 数据需要修改吗？有什么需要修改的吗？数据应该怎么调整才能适用于接下来的分析和挖掘？
- 是一个迭代的过程，实际项目中可能需要不止一次地执行这些清洗操作
- 处理缺失数据
  - pd.fillna()，pd.dropna()

数据连接

pd.merge

- 根据单个或多个键将不同DataFrame的行连接起来
- 类比数据库的连接操作 (第三课)
- 默认将重叠列的列名作为“外键”进行连接
  - on显示指定“外键”
  - left_on，左侧数据的“外键”
  - right_on，右侧数据的“外键”
- 默认是“内连接”(inner)，即结果中的键是交集

示例代码： 07_data_merge.ipynb

数据连接

pd.merge (续)

- how指定连接方式
- “外连接”(outer)，结果中的键是并集
- “左连接”(left)
- “右连接”(right)
- 处理重复列名
  - suffixes，默认为_x, _y
- 按索引连接
  - left_index=True或right_index=True

示例代码： 07_data_merge.ipynb

数据合并

pd.concat

- 沿轴方向将多个对象合并到一起
- NumPy的concat
  - np.concatenate
- pd.concat
  - 注意指定轴方向，默认axis=0
  - join指定合并方式，默认为outer
  - Series合并时查看行索引
  - DataFrame合并时同时查看行索引和列索引

示例代码： 08_data_concat.ipynb

数据重构

重构

- stack
  - 将列索引旋转为行索引，完成层级索引
  - DataFrame->Series
- unstack
  - 将层级索引展开
  - Series->DataFrame
  - 默认操作内层索引，即level=-1

示例代码： 09_data_reshape.ipynb

数据重构

重构

- stack
  - 将列索引旋转为行索引，完成层级索引
  - DataFrame->Series
- unstack
  - 将层级索引展开
  - Series->DataFrame
  - 默认操作内层索引，即level=-1

示例代码： 09_data_reshape.ipynb

数据转换

处理重复数据

- duplicated() 返回布尔型Series表示每行是否为重复行
- drop_duplicates() 过滤重复行
  - 默认判断全部列
  - 可指定按某些列判断

map

- Series根据map传入的函数对每行或每列进行转换

数据替换

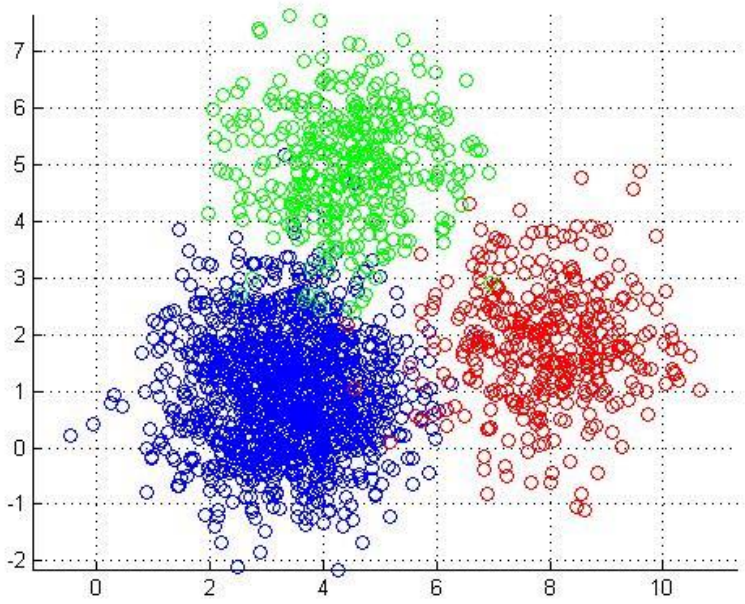
- replace

示例代码： 10_data_transform.ipynb

目录

K-Means

- 聚类（clustering）属于无监督学习（unsupervised learning）
- 无类别标记
- 在线demo http://syskall.com/kmeans.js/



K-Means

- 数据挖掘十大经典算法之一
- 算法接收参数k；然后将样本点划分为k个聚类；同一聚类中的样本相似度较高；不同聚类中的样本相似度较小
- 算法思想：

  以空间中k个样本点为中心进行聚类，对最靠近它们的样本点归类。通过迭代的方法，逐步更新各聚类中心，直至达到最好的聚类效果

示例代码： lect03_kmeans

K-Means

- 算法描述：

  1. 选择k个聚类的初始中心

  2. 在第n次迭代中，对任意一个样本点，求其到k个聚类中心的距离，将该样本点归类到距离最小的中心所在的聚类

  3. 利用均值等方法更新各类的中心值

  4. 对所有的k个聚类中心，如果利用2,3步的迭代更新后，达到稳定，则迭代结束。

- 优点：

  速度快，简单

- 缺点：

  最终结果和初始点的选择相关，容易陷入局部最优，需要给定k值

示例代码： lect03_kmeans

目录

实战案例

项目介绍

- https://www.kaggle.com/openfoodfacts/world-food-facts

项目任务

- 统计各国家食物中的食品添加剂种类个数

涉及知识点

- 掌握Pandas的数据操作和分析

示例代码：lect03_proj

参考

- 10分钟了解Pandas

  http://pandas.pydata.org/pandas-docs/stable/10min.html
- Pandas的索引操作

  http://pandas.pydata.org/pandas-docs/stable/indexing.html
- Pandas处理缺失数据

  http://pandas.pydata.org/pandas-docs/stable/missing_data.html
- Pandas绘图

  http://pandas.pydata.org/pandas-docs/version/0.18.1/visualization.html
- Pandas高级索引/层级索引

  http://pandas.pydata.org/pandas-docs/stable/advanced.html

参考

- 《Python for Data Analysis》
- Pandas中的GroupBy

  http://pandas.pydata.org/pandas-docs/stable/groupby.html
- Pandas透视表

  http://pandas.pydata.org/pandas-docs/stable/reshaping.html
- k-means

  https://en.wikipedia.org/wiki/K-means_clustering
- k-means算法及示例

  http://www.saedsayad.com/clustering_kmeans.htm

疑问

- 问题答疑： http://www.xxwenda.com/
  - 可邀请老师或者其他人回答问题

小象问答 @Robin_TY

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：小象
- 新浪微博：ChinaHadoop