《Python数据分析升级版》第三期

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：小象
  - 新浪微博：ChinaHadoop

员工离职原因分析及预测

--梁斌

目录

- 数据分析基本概念
- 数据分析基本步骤
- Pandas简单数据分析
- 数据分析建模理论基础
- 案例分析

目录

数据分析基本概念

何谓数据分析

- 用适当的统计分析方法对收集来的大量数据进行分析，提取有用信息和形成结论对数据加以详细研究和概括总结的过程

- 数理知识
- 数据获取、加工能力
- 行业知识

数据分析基本概念

数据分析在具体业务场景的使用

- 业务逻辑清晰、目标明确
- 能够转换成适当的数据/数学/统计问题
- 有足够的数据支撑
- 熟悉模型/分析方法的局限性
- 熟悉业务场景

数据分析的目的

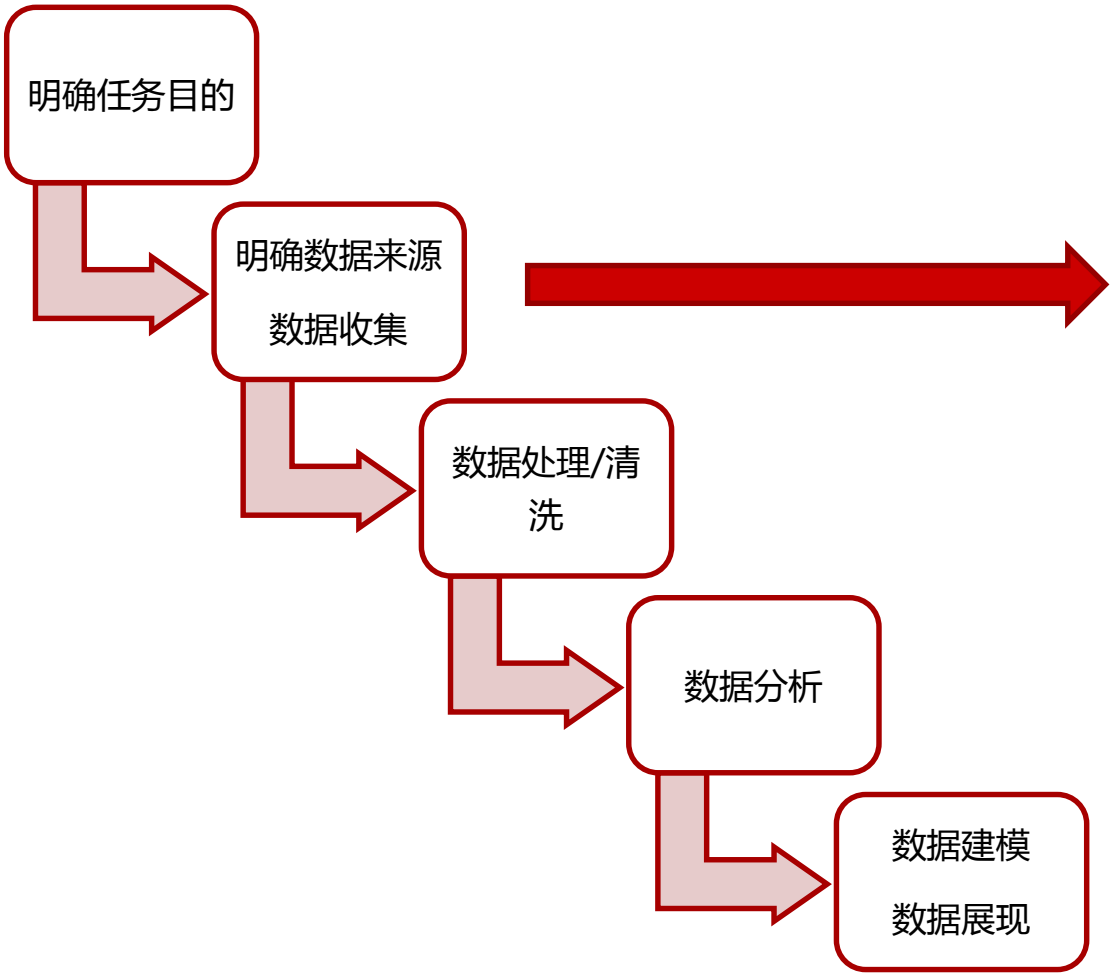
- 从数据中挖掘规律、验证猜想、进行预测

数据分析基本概念

数据科学家之路 (原图地址：点击下载)

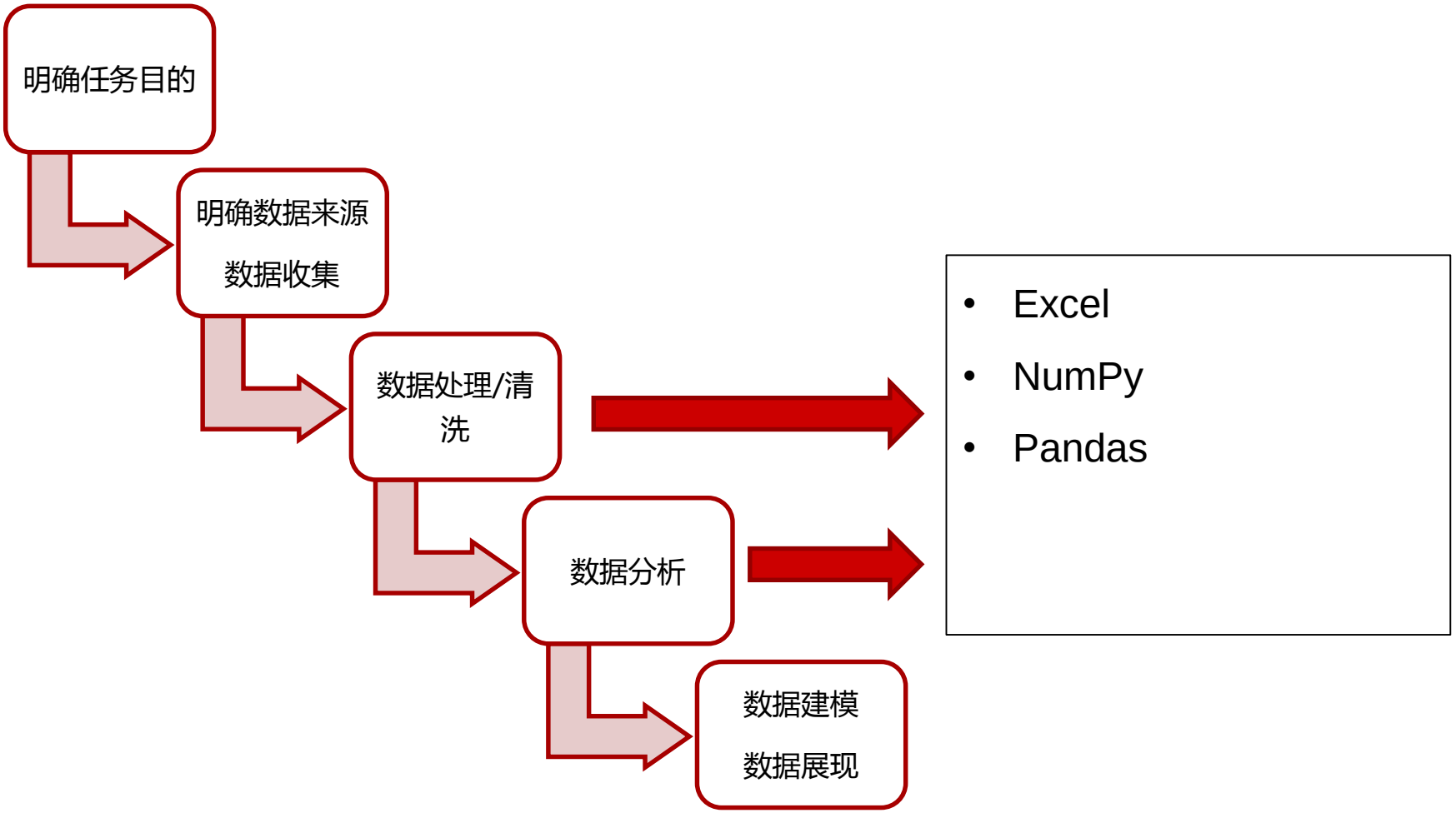
目录

数据分析基本步骤

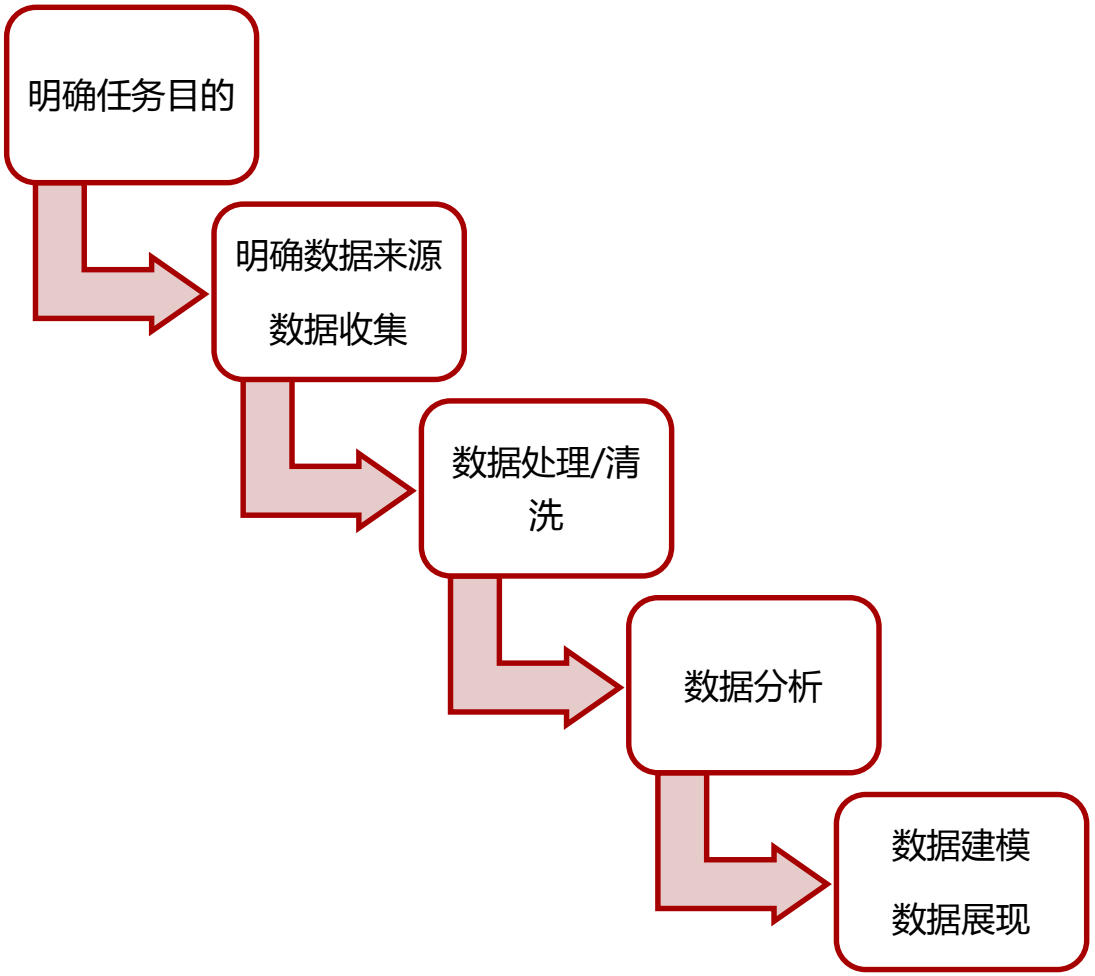


- 客户提供
- 自动化收集
  - 网络爬虫
- 手动收集
  - 调查问卷
- 途径
  - 公开信息
  - 外部数据库
  - 自有数据库
  - 调查问卷
  - 客户数据

数据分析基本步骤


明确任务目的
明确数据来源
数据收集
数据处理/清洗
数据分析
数据建模
数据展现
Excel
NumPy
Pandas

数据分析基本步骤



- 数据建模
  - scikit-learn
  - TensorFlow
  - …
- 数据展现
  - Matplotlib
  - Pandas
  - Seaborn
  - Excel
  - …

《Python数据分析升级版》第三期

课程介绍

《Python数据分析升级版III》课程安排				
	标题	内容	实战案例	上课时间
第一课	工作环境准备及数据分析基础	1. 课程介绍 2. 工作环境准备 3. 数据分析中常用的Python技巧 3. 科学计算库NumPy 4. 使用Pandas进行简单的数据分析	1. 世界幸福指数报告分析	2017/10/11 20:00-22:00
第二课	Pandas进阶及统计分析	1. Pandas进阶及技巧 2. 数据合并、分组及比较 3. 透视表 4. 常用的统计分布 5. 使用Python进行假设检验	2. 麦当劳菜单营养成分分析	2017/10/14 15:00-17:00
第三课	数据展示及可视化	1. 数据可视化的概念及准则 2. 基本图表的绘制及应用场景 -- 散点图，柱状图，线图 3. 数据分析常用图表的绘制 -- 多图绘制，直方图，盒子图，热图 4. 动画及交互式渲染 5. Pandas及Seaborn制图	3. 宠物小精灵数据分析及展示	2017/10/15 15:00-17:00
第四课	Python机器学习(1)	1. 机器学习基本概念与流程 2. Python机器学习库scikit-learn 3. 机器学习常用算法介绍及演示(1) -- KNN, 线型回归，逻辑回归，SVM，决策树	4-1. 根据可穿戴设备识别用户行为	2017/10/21 15:00-17:00
第五课	Python机器学习(2)	1. 模型评价指标及模型选择 2. 交叉验证 3. 机器学习常用算法介绍及演示(2) -- 朴素贝叶斯，随机森林，GBDT	4-2. 根据可穿戴设备识别用户行为	2017/10/22 15:00-17:00
第六课	图像数据处理及分析	1. 图像数据操作 2. 常用的图像特征描述 3. K-Means聚类及图像压缩	5-1. 根据海报预测电影分类	2017/11/04 15:00-17:00
第七课	神经网络及深度学习	1. 人工神经网络 2. 深度学习 3. TensorFlow框架学习及使用 4. TensorFlow实现卷积神经网络 (CNN)	5-2. 根据海报预测电影分类	2017/10/28 15:00-17:00
第八课	文本数据处理及分析	1. Python文本数据处理 -- 正则表达式结合Pandas使用技巧 2. 自然语言处理及NLTK 3. 文本特征及分类 4. 主题模型及LDA	6. 豆瓣影评数据分析	2017/11/05 15:00-17:00
第九课	社交网络分析	1. 图论简介 2. 网络的操作及可视化 3. 网络分析 4. 课程总结	7. “权利的游戏”人物关系分析	2017/11/06 15:00-17:00

数据处理及分析

机器学习及建模

非结构化数据分析

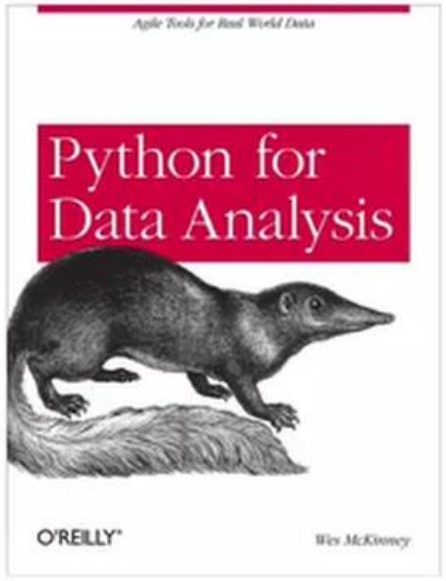
目录

- 数据分析基本概念
- 数据分析基本步骤
- Pandas简单数据分析
- 数据分析建模理论基础
- 案例实战

Pandas简单数据分析

Pandas

- 2008年由Wes McKinney创建
- 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



Pandas简单数据分析

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)

SERIES

index	element
0	**1**
1	**2**
2	**3**
3	**4**
4	**5**

Pandas简单数据分析

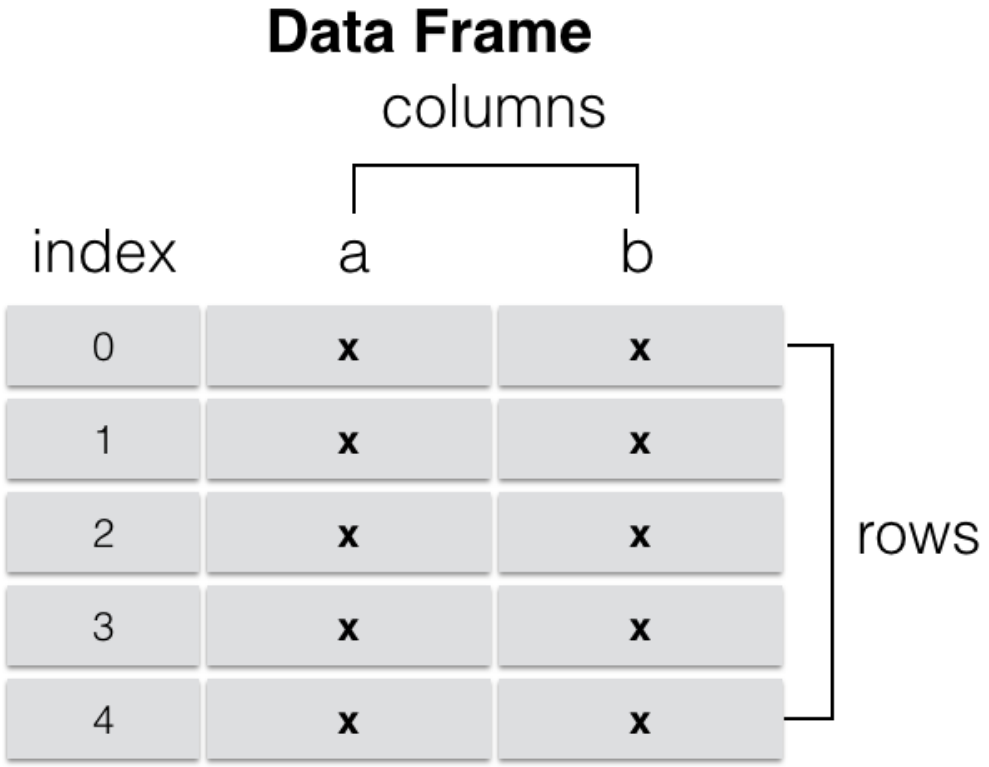
Series (续)

- 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

Pandas简单数据分析

DataFrame

- 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
- 索引包括列索引和行索引



Pandas简单数据分析

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]

索引操作

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

Pandas简单数据分析

排序

- sort_index，索引排序
- 按值排序，sort_values(by='label')

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

Pandas简单数据分析

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简单数据分析

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

Pandas简单数据分析

- 处理缺失数据
  - dropna() 丢弃缺失数据
  - fillna() 填充缺失数据
- 数据过滤
  - df[filter_condition] 依据filter_condition对数据进行过滤

Pandas简单数据分析

- Pandas提供了内建的绘图功能（基于matplotlib）
- plot(kind, x, y, title, figsize)

  x, y 横纵坐标对应的数据列

  title图像名称

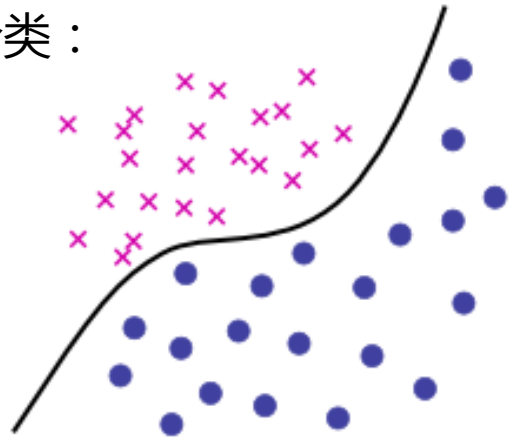
  figsize图像尺寸
- 保存图片

  plt.savefig()
- 更多例子请参考：https://pandas.pydata.org/pandas-docs/stable/visualization.html

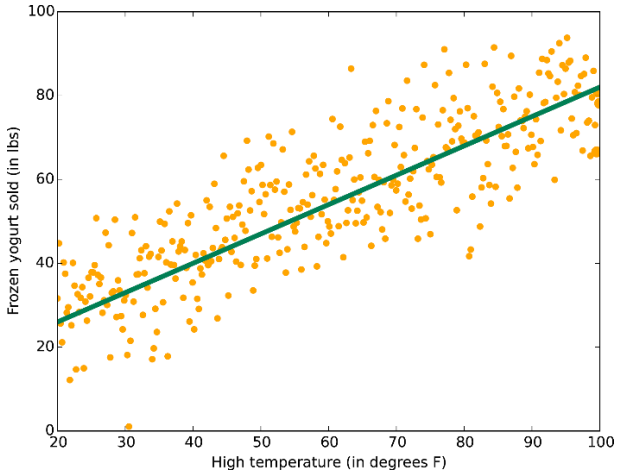
目录

数据分析建模理论基础

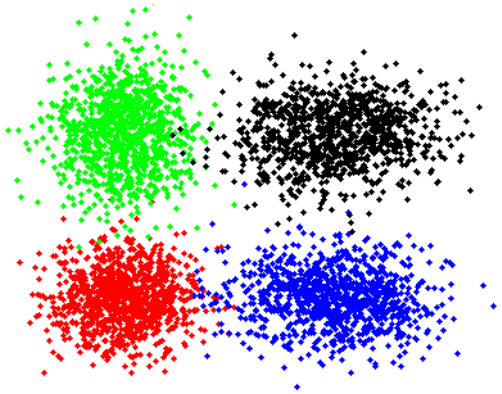
任务分类：



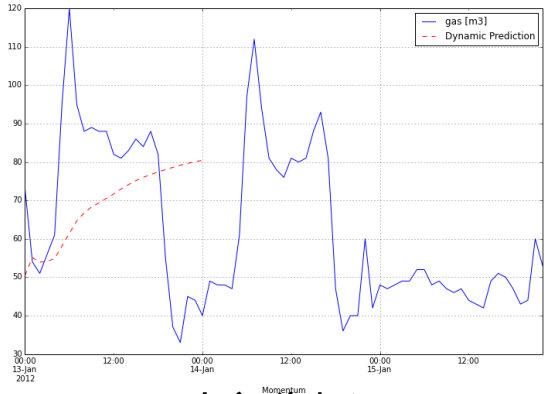
分类



回归



聚类



时序分析

数据分析建模理论基础

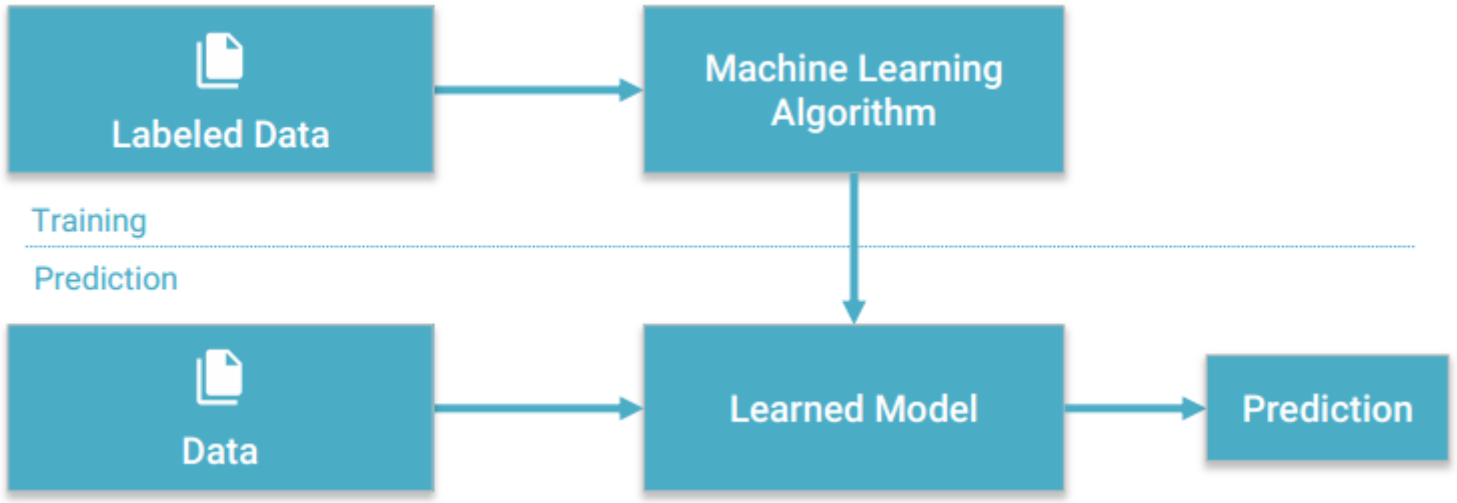
机器学习：问题描述

- “学习” 问题通常包括n个样本数据（训练样本），然后预测未知数据（测试样本）的属性
- 每个样本包含的多个属性（多维数据）被称作“特征”
- 分类：
  - 监督学习，训练样本包含对应的“标签”，如识别问题
    - 分类问题，样本标签属于两类或多类（离散）
    - 回归问题，样本标签包括一个或多个连续变量（连续）
  - 无监督学习，训练样本的属性不包含对应的“标签”，如聚类问题

数据分析建模理论基础

定义

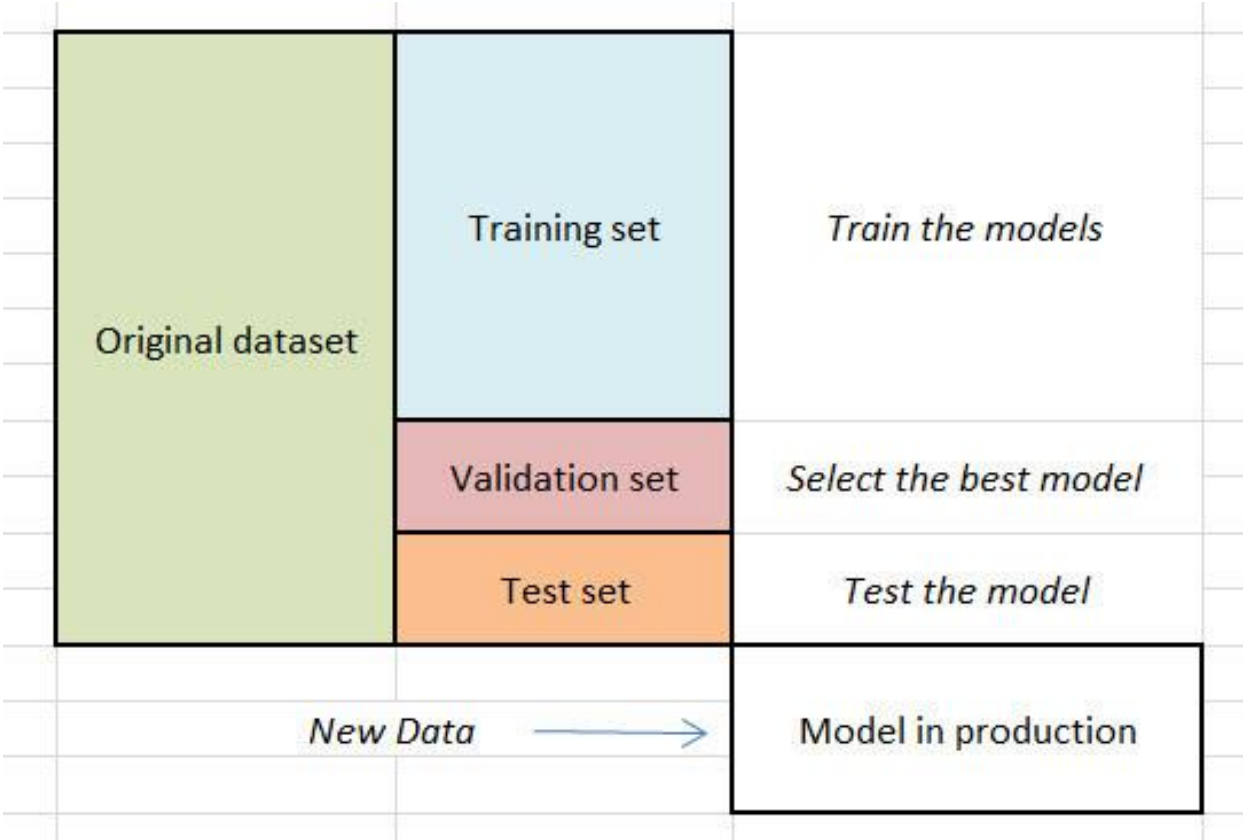
- Machine Learning is a type of Artificial Intelligence that provides computers with the ability to learn without being explicitly programmed.
- Provides various techniques that can learn from and make predictions on DATA.



数据分析建模理论基础

机器学习

- 训练集 vs 验证集 vs 测试集



目录

- 数据分析基本概念
- 数据分析基本步骤
- Pandas简单数据分析
- 数据分析建模理论基础
- 案例实战

案例实战

员工离职原因分析及预测

实战案例

项目名称：员工离职原因分析及预测

请参考相应的配套代码

1. 包含项目详细步骤的Jupyter演示版代码
2. 真实项目的程序代码

《Python数据分析升级版》第三期

疑问

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

小象问答邀请 @Robin_TY 回答问题

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