一点点数据处理

Designed by HK-SHAO 2021-11-02

总览

1. 数据的维度
2. 数据的格式
3. 数据的清洗
4. 数据的分析
5. 数据的可视化

#0 什么是数据

数据跟信息和信号的区别和联系是什么?

1. 信息:
   - 音讯、消息、通讯系统传输和处理的对象，泛指人类社会传播的一切内容。它能够消除随机和不确定，是创建一切宇宙万物的最基本单位。
2. 信号
   - 表示消息的物理量，如电信号可以通过幅度、频率、相位的变化来表示不同的消息。除此之外还有化学信号、生物信号，但本质也是物理信号。
3. 数据:
   - 信息的表现形式和载体，可以被存储到某种介质当中。数据是事实或观察的结果，用于表示客观事物的未经加工的原始素材。

#1 数据的维度

什么是数据（集）的维度？数据有哪些维度？

1. 0维的数据
    - 单个信息（数字、布尔值、字符串等）
2. 1维的数据
    - 序列：例如一系列信息
3. 2维的数据
    - 矩阵：例如一张Excel表格
4. 3维的数据
    - 张量：可被映射到空间中

#3 数据的格式

数据的编码是什么？数据有哪些格式？数据与文件的联系？

数据编码

1. 二进制编码
    - 多媒体文件、可执行程序、PDF文档、压缩包等
2. 字符编码：一种特殊的二进制编码
    - TXT文件、SVG、HTML网页、源代码等

数据格式

- 常用的数据交换格式：JSON、XML、YAML、CSV、SQL

#4 数据的清洗

发现并纠正数据文件中可识别的错误，包括检查数据一致性，处理无效值和缺失值等。一般由计算机程序自动完成，而不建议人工进行。

1. 缺失值
   数据集中缺失的部分
2. 无效值
   数据值不符合类型或不符合规则
3. 离群点
   数值与群体表现离群，可能出现异常

#5 数据的分析

1. 最大值、最小值、平均值、众数等
2. 线性拟合、多项式拟合、指数或对数曲线拟合
3. S形曲线拟合（Sigmoid 函数）、逻辑斯蒂回归（Logistic 回归）
4. T检验、卡方检验（独立性检验）、F检验
5. 机器学习方法

#6 数据的可视化

- Apache ECharts（echarts.apache.org）
- Data-Driven Documents （d3js.org）

谢谢朋友们