



本课件仅用于教学使用。未经许可，任何单位、组织和个人不得将课件用于该课程教学之外的用途(包括但不限于盈利等)，也不得上传至可公开访问的网络环境

1

数据科学导论

Introduction to Data Science

数据分析：数据清理

陈恩红，黄振亚

Email: cheneh@ustc.edu.cn, huangzhy@ustc.edu.cn

课程主页：

<http://staff.ustc.edu.cn/~huangzhy/Course/DS2026.html>



回顾：数据分析基础

- 数据采集 Data Collection
- 数据预处理 Data Preprocessing
- 特征工程 Feature Engineering



数据预处理

- 大数据环境下的数据特点
- 为什么需要进行预处理
- 数据预处理
 - 数据清洗
 - 数据统计分析
 - 数据相关性分析
 - 数据变换



大数据环境下的数据特点—4V

数据来源多样：传感器，
IT系统，应用软件等

数据类型多样：结构化，
半结构，非结构

多样
Variety

数据分析与结果需要及时处理，
实时的结果才有价值—1秒定律

高速
Velocity

计量单位一般是TB，
甚至到了PB，EB或ZB

大量
Volume



价值
Value

TB (2^{30} KB)
PB (2^{40} KB)
EB (2^{50} KB)
ZB (2^{60} KB)

Big Data

“沙里淘金”：价值密度低，
价值深度深，带来巨大的
科学和商业价值



大数据环境下的数据特点

- 收集来的数据，是否可以直接使用？

*ratings.csv - 记事本

文件(E) 编辑(E) 格式(O) 查看(V) 帮助(H)

userId,movieId,rating,timestamp

1,1,4.0,964982703

1,3,4.0,964981247

1,6,4.0,964982224

1,47,5.0,964983815

1,50,5.0,964982931

1,70,3.0,964982400

1,101,5.0,964980868

Context:

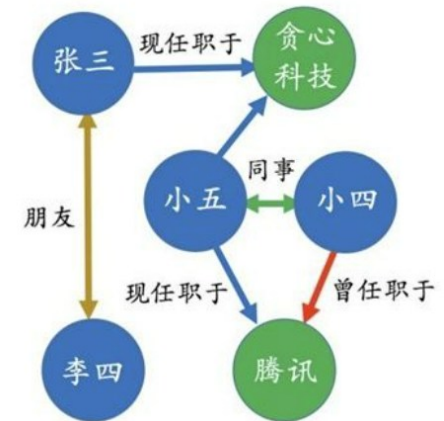
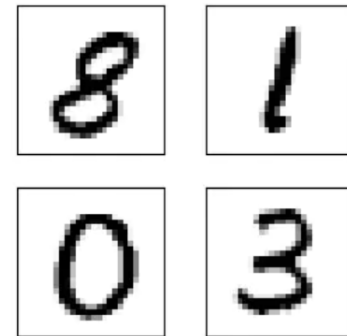
Computational complexity theory is a branch of the theory of computation in theoretical computer science that focuses on classifying computational problems according to their inherent difficulty, and relating those classes to each other. A computational problem is understood to be a task that is in principle amenable to being solved by a computer, which is equivalent to stating that the problem may be solved by mechanical application of mathematical steps, such as an algorithm.

Question:

By what main attribute are computational problems classified using computational complexity theory?

Answer:

inherent difficulty



通常情况下，直接收集的数据难以直接使用，需要对数据进行预处理



数据预处理

- 大数据环境下的数据特点
- 为什么需要进行预处理
- 数据预处理
 - **数据清洗**
 - 数据统计分析
 - 数据相关性分析
 - 数据变换



为什么进行数据预处理

□ 直接收集的数据通常是“脏的” — 数据来源不同

□ 应用需求

- 微博
- 淘宝
-



□ 收集手段

- 传感器，扫描仪
- 摄像，照相
- App收集
- 爬虫写错了
-



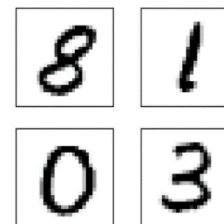
高高飞起来啊
前方高能预警
可以做成游戏的
高能来了
高高的飞起来啊！
前方高能(ノ°

□ 数据格式

- 结构化
- 半结构化
- 非结构化
-

ID	时间	内容
陈赫	08-18	天霸
邓超	08-18	我们都很好
邓超	08-18	我也不知道

```
<province>
  <name>黑龙江</name>
  <cities>
    <city>哈尔滨</city>
    <city>大庆</city>
  </cities>
</province>
```





为什么进行数据预处理

□ 直接收集的数据通常是“脏的”

□ 不完整

- 有些数据属性的值丢失或不确定
- 缺失必要的信息，例：缺失学生成绩
-

□ 不准确

- 数据错误，属性值错误，例：成绩 = -10
- 噪声数据：包含孤立（偏离期望）的离群
-

□ 不一致

- 数据结构有较大差异，例，编码或者命名上存在差异
- 数据需求改动，例，评价等级：“百分制”与“A, B, C”
- 存在数据重复和信息冗余现象
-

学号 Sno	课程号 Cno	成绩 Grade
200215121	1	92
200215121	2	85
200215121	3	88
200215122	2	90
200215122	3	80



为什么进行数据预处理

- 真实应用的数据是“脏的”——举例
 - 滥用缩写词 例：中科大，科大，中国科大，USTC
 - 数据中的内嵌控制信息 例：路程=速度*时间
 - 重复记录
 - 缺失值
 - 拼写变化与时态，例：propose, proposed, proposing
 - 不同的计量单位
 - 噪声
 - **UGC**数据，例：弹幕(短文本)



为什么进行数据预处理

- 数据错误的不可避免性
 - 数据输入和获得过程数据错误的不可避免性
 - 数据集成所表现出来的错误
 - 数据传输过程所引入的错误
- 没有高质量的数据，就没有高质量的结果
 - 高质量的决策必须依赖高质量的数据
 - 例如，数据重复或者缺失将会产生不正确的分析结果，误导决策
- 数据预处理是进行大数据的分析和挖掘的工作中占工作量最大的一个步骤（80%）



数据预处理

- 大数据环境下的数据特征
- 为什么需要进行预处理
- 预处理的基本方法
 - 数据清理
 - 数据集成
 - 数据变换



数据预处理：数据清理

- 数据清理的目标
 - 解决数据质量问题
 - 让数据更适合分析、建模
- 数据清理基本任务
 - 处理缺失值
 - 清洗噪声数据
 - 纠正不一致数据
 - 根据需求进行清理
 -

ID	住址	学历	单位	专业	收入
01	A区	本科	A	CS	C
02	B区	本科	B	EE	C
03	A区	本科	A	CS	C
04	A区	硕士	C	CS	B
05	A区	博士	A	DS	A
...	...	...	...	...	...

ID	住址	学历	单位	专业	收入
01	A区	本科	A	CS	C
02	B区	本科	B	EE	C
03	A区	本科	A	CS	0
04	A区		C	CS	B
	A区	博士	A	DS	
...	...	...	...	...	...



数据清理-处理缺失值

□ 造成数据缺失的原因

- 信息无法获取，或获取代价大。

 - 反爬虫，加密

- 信息遗漏

 - 需求不明确

 - 采集故障，存储故障，传输故障

 - 人为因素

- 数据的某些属性不可用，或不存在（与设计有关）

 - 如：学生的收入，老师的成绩等



数据清理-处理缺失值

□ 数据缺失的类型

- 完全随机缺失：不依赖任何属性/变量，不影响样本的无偏性
- 随机缺失：缺失自身无关，与其他完全属性/变量有关系
 - 体检“老年人血压”漏检缺失：“血压值”与“年龄”
 - 问卷“某人群未填收入”缺失：“收入”与“人群背景”的关系
- 非随机缺失：数据缺失与属性/变量自身的取值有关
 - 疾病诊断缺失：重大疾病患者拒绝提供诊断结果

ID	住址	学历	单位	专业	收入
01	A区	本科	A	CS	C
02	B区	本科	B	EE	C
03	A区	本科	A	CS	0
04	A区		C	CS	B
	A区	博士	A	DS	
...	...	...	...	...	...



数据清理-处理缺失值

□ 处理缺失数据的方法：首先确认缺失数据的影响

□ 数据删除（可能丢失信息，或改变分布）

- 删除数据
- 删除属性
- 改变权重

□ 数据填充

■ 特殊值填充

- 样本/属性的均值、中位数、众数填充
- 空值填充，不同于任何属性值。例，NLP词表补0，DL补mask

■ 预测：使用最可能的数据填充

- K最近距离法（KNN）
- 利用回归等估计方法
- 大模型等



模型预测：建立模型预测缺失值



数据清理-处理缺失值

□ K最近距离法

- 完整数据中找到**1个**与它最相似的样例，然后用该样本的值来进行填充
- 根据相关分析(距离)来确定距离缺失数据样本的最近**K个**样本
- 将这K个值加权平均估计样本缺失数据

□ 模型法：回归法

- 基于**数据集**，建立回归模型
- 将已知属性值代入模型来估计未知属性值，以此预测值填充

ID	住址	学历	单位	专业	收入
01	A 区	本科	A	CS	C
02	B 区	本科	B	EE	C
03	A 区	本科	A	CS	0
04	A 区		C	CS	B
	A 区	博士	A	DS	
...	...	...	...	...	...



数据清理-处理缺失值

□ 大模型做数据填充

- 提示词工程
- 大模型数据读取和代码
- 填充结果的合理性与真实性
- 计算成本与效率
- 与数据分布的适配性
- 结果验证与后处理
- 数据隐私与安全风险

ID	住址	学历	单位	专业	收入
01	A区	本科	A	CS	C
02	B区	本科	B	EE	C
03	A区	本科	A	CS	0
04	A区		C	CS	B
	A区	博士	A	DS	
...	...	...	...	...	...



数据清理-清洗噪声

- 噪声是测量误差的随机部分
 - 包括错误值，或偏离期望的孤立点值
 - 需要对数据进行平滑
- 常用的处理方法
 - 回归(Regression)
 - 让数据适应回归函数来平滑数据
 - 识别离群点，常用聚类方法
 - 监测并且去除孤立点

ID	住址	学历	单位	专业	收入
01	A区	本科	A	CS	C
02	B区	本科	B	EE	C
03	A区	本科	A	CS	0
04	A区		C	CS	B
	A区	博士	A	DS	
...	...	...	...	...	...

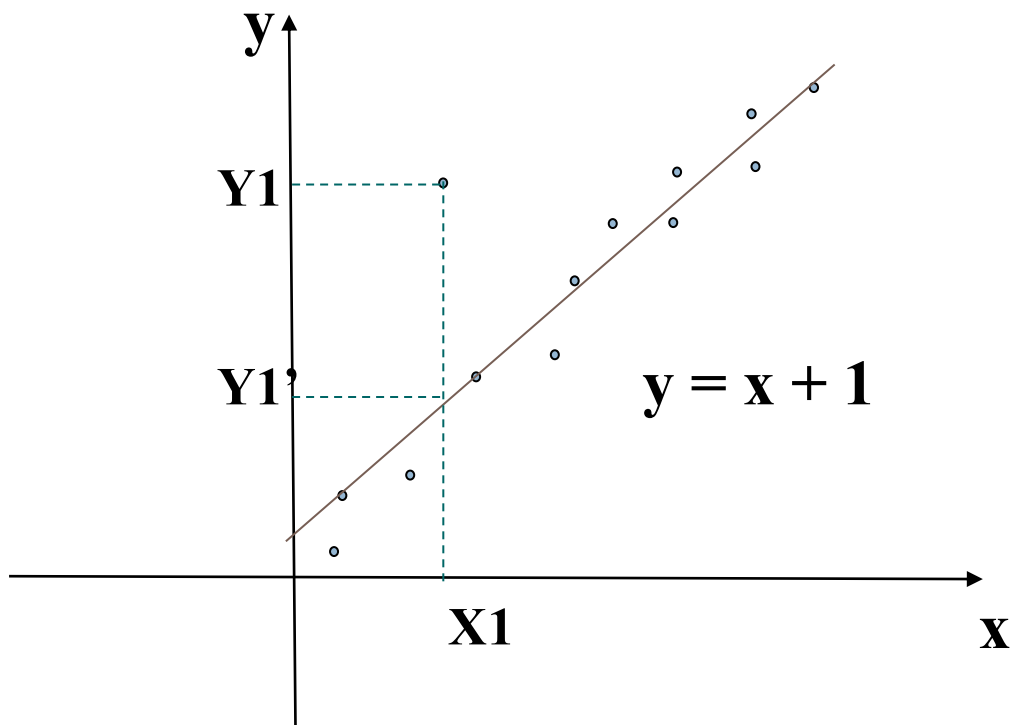


数据清理-清洗噪声

19

□ 回归：让数据适应回归函数来平滑数据

- 通过线性回归模型，对不符合回归的数据进行平滑处理
- 用某些属性预测其他属性



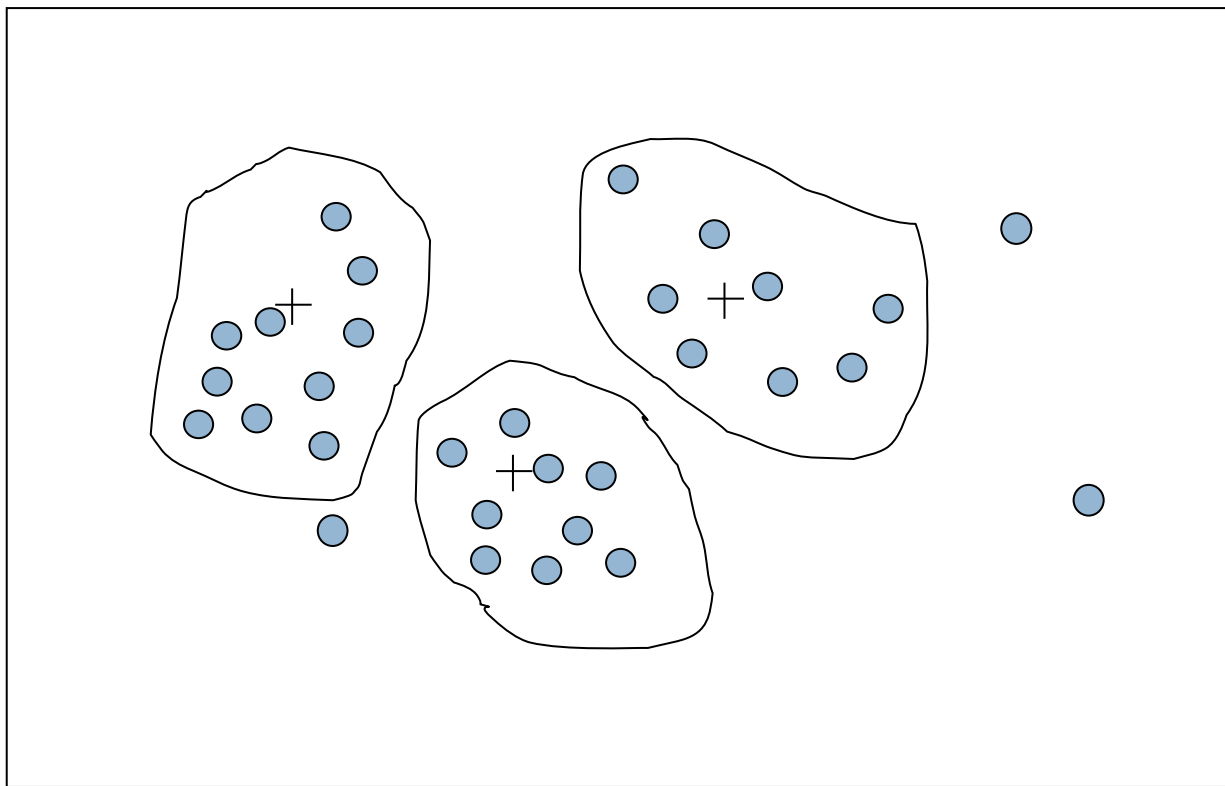
ID	住址	学历	单位	专业	收入
01	A区	本科	A	CS	C
02	B区	本科	B	EE	C
03	A区	本科	A	CS	0
04	A区		C	CS	B
	A区	博士	A	DS	
...	...	...	...	...	...



数据清理-清洗噪声

20

- 识别离群点：聚类分析检测离群点，消除噪声
 - 聚类将类似的值聚成簇
 - 落在簇集合之外的值被视为离群点





数据清理-根据需求清理数据

在特定的应用任务中，根据目标不同，需要特殊的数据清理方法

推荐系统

- 通用推荐问题
- 冷启动问题

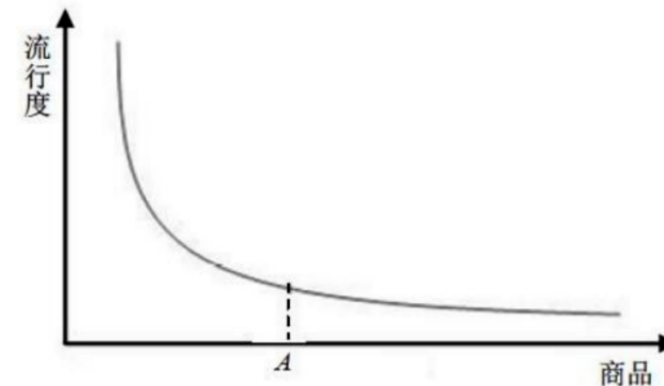
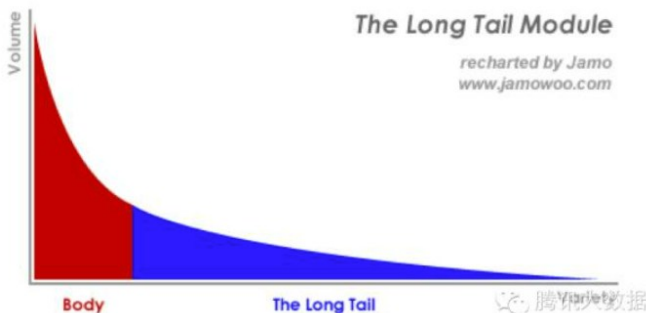
教育大数据

社交网络

POI任务：Point of interest

Dataset	Douban Book	Yelp
# Users	6,576	25,783
# Items	20,547	33,105
# Ratings	326,419	727,259
Rating Sparsity	99.76%	99.91%
Avg. friends of each user	6.0	3.8
# Users without friends	1,314	10,867

Table 1: The statistics of two datasets.



Li Wang, Zhenya Huang, Qi Liu, Enhong Chen, Preference-Adaptive Meta-Learning for Cold-Start Recommendation, IJCAI'2021.