数据科学导论

Introduction to Data Science

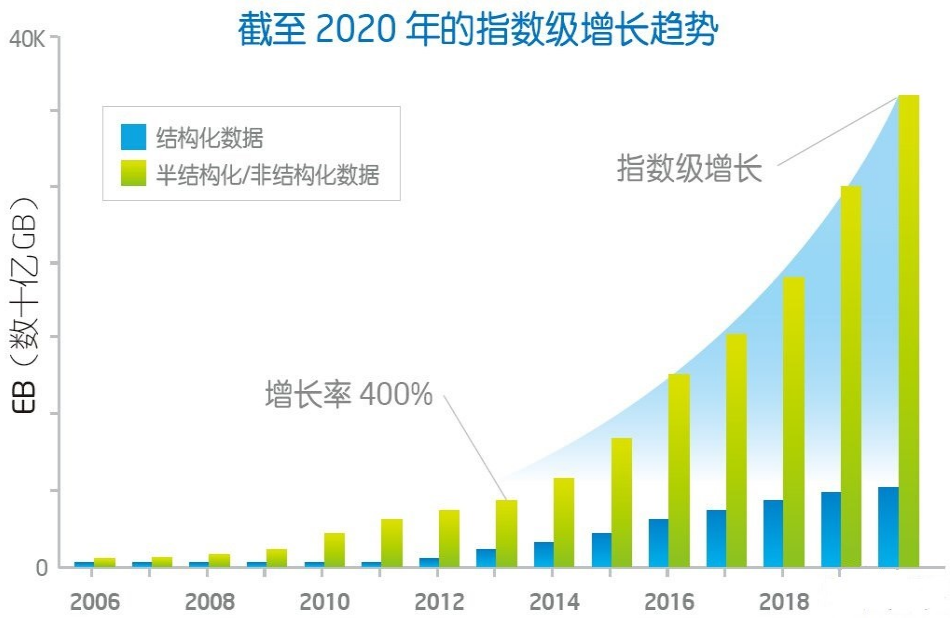
第二章 数据分析基础

黄振亚，陈恩红
Email: huangzhy@ustc.edu.cn, cheneh@ustc.edu.cn

课程主页：
http://staff.ustc.edu.cn/~huangzhy/Course/DS2024.html

回顾：数据分析基础

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

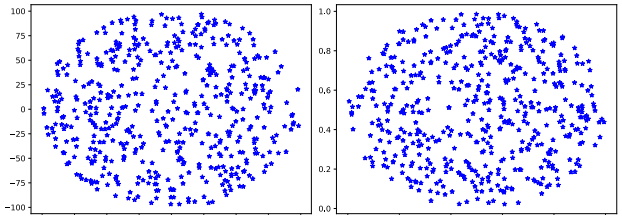


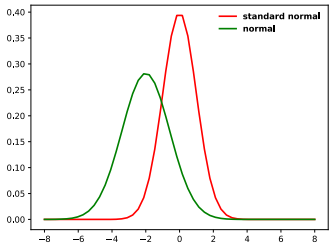
数据预处理

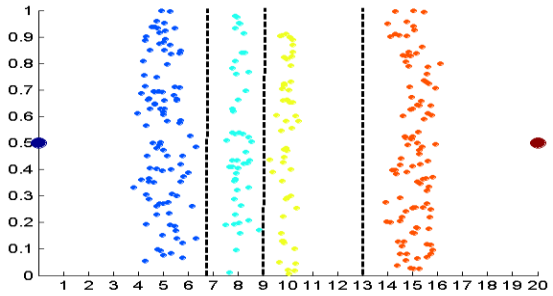
- 大数据环境下的数据特征
- 为什么需要进行预处理
- 预处理的基本方法
  - 数据清理
  - 数据集成
  - 数据变换
  - 数据规约

数据预处理：数据变换

- 数据变换的目的是将数据转换成适合分析建模的形式
  - 前提条件：尽量不改变原始数据的规律
  - 数据规范化
    - 最小-最大规范化
    - z-score规范化
    - 小数定标规范化
  - 数据离散化
    - 非监督离散化
    - 监督离散化

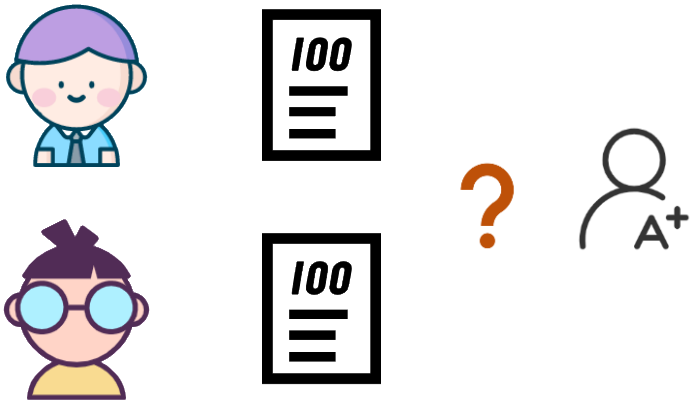






数据预处理：数据变换

- **数据规范化**
  - **目的：将不同数据（属性）按一定规则进行缩放，使它们具有可比性**
  - 例如，我们需要考察学生A和学生B的某门课程成绩。A的考试满分是100分（及格60分），B的考试满分是150分（及格90分）。显然，A和B的100分代表着完全不同的含义。



如何用一个同等的标准来比较A与B的成绩数据呢？

数据变换-规范化

- **最小-最大规范化**
  - 对原始数据进行线性变换。把数据A的观察值v从原始的区间[$\min_A$, $\max_A$]映射到新区间 [new_min_A, new_max_A]
    - 0-1规范化又称为归一化

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - 数理依据：

$$\frac{v' - new\_min_A}{new\_max_A - new\_min_A} = \frac{v - min_A}{max_A - min_A}$$

数据变换-规范化

- 最小-最大规范化

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - 例：假设某属性规范化前的取值区间为[－100，100]，规范化后的取值区间为[0，1]，采用最小-最大规范化 66，得

$$v' = \frac{66 - (-100)}{100 - (-100)}(1 - 0) + 0 = 0.83$$

快速练习：采用最小-最大规范化 -80？

数据变换-规范化

假设A的课程成绩为70分（0-100分），B的课程成绩为110分（0-150分），采用最小-最大规范化来比较A和B的成绩

A

取值区间为[0,100]，规格后的取值空间为[0,1]，采用最小-最大规范70后为0.7

B

取值区间为[0,150]，规格后的取值空间为[0,1]，采用最小-最大规范110后为0.73

用最小-最大规范化后得出B的成绩更好

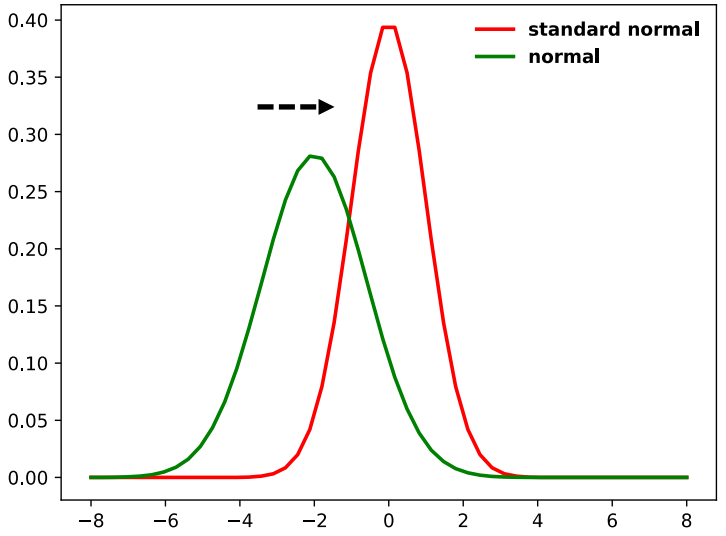
数据变换-规范化

- z-score规范化
  - 最大最小值未知，或者离群点影响较大时，假设数据服从正态分布
    - 某一原始数据(v)与原始均值的差再除以标准差，可以衡量某数据在分布中的相对位置

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

  - 例：假设某属性的平均值、标准差分别为80、25，用z-score规范化 66

$$v' = \frac{66 - 80}{25} = -0.56$$



数据变换-规范化

- z-score规范化
  - 例：假设学生的成绩分布符合正态分布，某素质课考试的平均分为73分，标准差为7分，A得78分；实践课考试的平均分为80分，标准差为6.5分，A得83分。那么A的哪一门考试成绩比较好？

采用z-score规范化得出A的素质课成绩要优于实践课成绩

数据变换-规范化

- 小数定标规范化
  - 通过移动小数点的位置来进行规范化。小数点移动多少位取决于属性A的取值中的最大绝对值。

$$v' = \frac{v}{10^j}$$

其中，j 是使 Max(|v'|)<1的最小整数

  - 比如属性A的取值范围是-999到88，那么最大绝对值为999，小数点就会移动3位，即新数值=原数值/1000。那么A的取值范围就被规范为-0.999到0.088。

数据变换-规范化

- 小结

	优点	缺点	适用场景
最小-最大规范化	保留了原始数据中存在的关系，是消除量纲和数据取值范围影响的最简单方法	对最大最小值敏感，新数据加入时，可能改变最大最小值，需重新计算	适用于原始数据不存在很大/很小的一部分数据的时候
z-score规范化	算法简单方便，结果方便比较，应用于数值型的数据，且不受数据量级的影响	总体平均值和方差不一定可知，在一定程度上要求数据分布，结果没有具体意义，只用于比较	适用于最大最小值未知，或者离群点影响较大的时候
小数定标规范化	算法实现简单	不适用于不同含义数据的比较，无实际意义	使用含义相同的数据，且最大最小相差较大