数据挖掘导论
Introduction to Data Mining

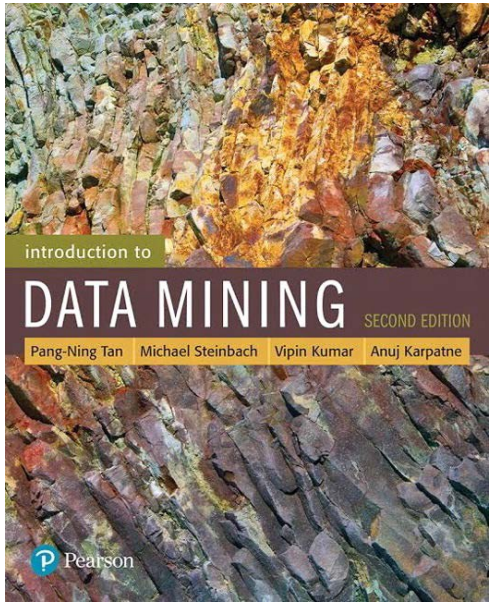
第二章：认识数据

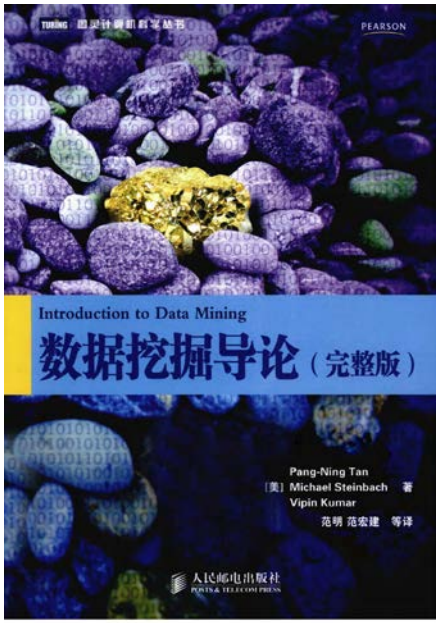
王浩
Email：haowang@szu.edu.cn

深圳大学 SHENZHEN UNIVERSITY
电子信息与工程学院

参考教材

- 数据挖掘导论（Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Addison Wesley ）





目标：

- 认识数据
- 数据获取
- 数据预处理
- 常用数据处理工具

目标：

- 认识数据
  - 属性和对象
  - 数据类型
  - 数据质量评估

属性和对象：

- 数据是对象（***data objects***）和其属性（***attributes***）的集合
- 属性是对象的性能或特征
  - 例如:人眼睛的颜色、温度等；
  - 属性有时也叫做变量，字段，特性，特征，或维。 Attribute is also known as variable, field, characteristic, feature, or dimension.
- 属性的集合可以表征一个对象
  - Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

Tid	退税	婚姻状况	年收入	是否欺诈
1	是	单身	125K	否
2	否	已婚	100K	否
3	否	单身	70K	否
4	是	已婚	120K	否
5	否	离异	95K	是
6	否	已婚	60K	否
7	是	离异	220K	否
8	否	单身	85K	是
9	否	已婚	75K	否
10	否	单身	90K	是

属性和对象-属性值：

- 属性值（***Attribute values***）是给定对象（object）的属性（attribute）的数字或符号。
- ***Attribute values*** are numbers or symbols assigned to an attribute for a particular object

- 属性和属性值的区别
  - 相同的属性可以映射到不同的属性值
    - 例如：高度可以用英尺或米来测量
  - 不同的属性可以映射到同一组值
    - 示例：ID和age的属性值是整数

属性和对象-属性的类别：

一般而言，属性可分为以下4类：

定性

- 标称(Nominal)属性
  - 标称属性的值是一些符号或实物的名称，每个值代表某种类别、编码或状态
  - 举例: 身份证号码, 实验人员标号
- 序数(Ordinal)属性
  - 序数属性可能的取值是具有意义的序列，但相继值之间的差是未知的。
  - 例如:排名(例如，午餐的味道从1-10分)，等级，身高{高，中，矮}

定量

- 区间标度(Interval)属性
  - 区间标度属性用相等的单位尺度度量。区间属性的值有序。所以，除了序列评定之外，这种属性允许比较和定量评估值之间的差。
  - 例如: 日期、摄氏或华氏温度、身高（单位：米）
- 比率标度(Ratio)属性
  - 比率标度属性的度量是比率，可以用比率来描述两个值，即一个值是另一个值的倍数，也可以计算值之间的差。
  - 例如: 开氏温度、长度，经过的时间等

从数值角度，属性又分为离散的和连续的两种。

属性和对象-属性的操作：

- 属性的类型取决于它拥有以下哪些属性/操作:
  - 差异: = ≠
  - 排序: < >
  - 加减: + -
  - 乘除: * /

  - 标称属性: = ≠
  - 序数属性: = ≠ < >
  - 区间标度属性: = ≠ < > + -
  - 比率标度: = ≠ < > + - * /

10摄氏度是5摄氏度两倍，这种说法在物理上有意义吗?

常见数据类型：

记录型数据（Record）

- ✓ 事务数据（Transaction Data）
- ✓ 数据矩阵（Data Matrix）
- ✓ 稀疏数据矩阵（Document-term Data）

有序数据（Ordered）

- ✓ 时序事务数据 （Sequential Transaction Data）
- ✓ 时间序列数据 （Time Series Data）
- ✓ 序列数据 (Sequential Data)
- ✓ 时空数据(Spatio-Temporal Data)

基于图的数据（Graph）

- ✓ 带有对象之间联系的数据
- ✓ 带有图对象的数据(天然结构为图的数据)

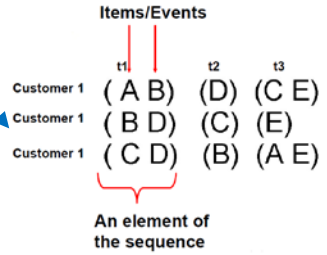
非结构化数据

…

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	0	3	0	0




```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

数据质量评估：

- 数据质量评估对数据挖掘至关重要
  - 例如1：银行用于用户信用评估的数据，若存在质量问题，可能导致银行向低信誉用户发放大量贷款；
  - 例如2：交易平台若存在质量问题，可能导致洗钱、黑产等大量违法违规问题；

1. 常见的数据质量问题有哪些?
2. 如何评估数据的问题?

数据质量问题与评估：

- 场景的数据质量问题有哪些?
  - 噪声（Noise）
  - 离群点（outliers）
  - 不一致的数据（Wrong data）
  - 虚假数据（Fake data）
  - 数据缺失 （Missing values）
  - 重复数据 （Duplicate data）

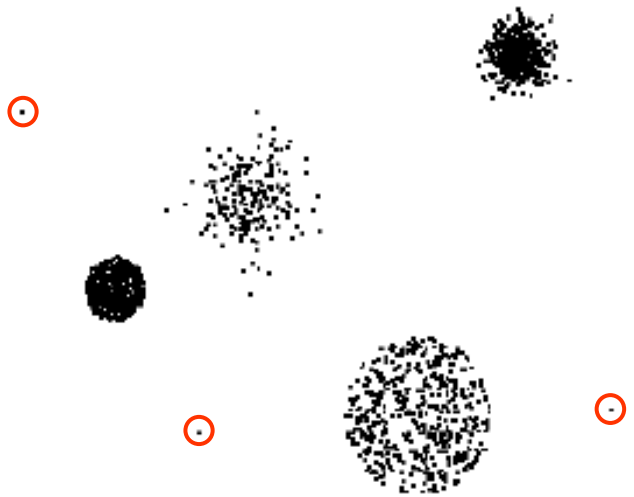
- 如何评估数据的问题?
  - 精度（precision）：同一个量重复测量值之间的接近程度；
  - 偏差（bias）：测量值与被测量数据实际分布之间的差异大小；
  - 准确率（accuracy）：测量值与实际值之间的接近程度。

数据质量问题——噪声

- 对于对象，噪声是无关的对象
- 对于属性，噪声是指对原始值的改变
  - 示例：使用手机通话时声音失真，电视屏幕上出现“雪花”
  - 下图显示了两个具有相同幅值和不同频率的正弦波,两个正弦波组合在一起,以及带有随机噪声的两个正弦波
    - 原始信号的幅值和形状失真

数据质量问题——离群点

- ***盟纲炮/开帮倦****是具有与数据集中的大多数其他数据对象截然不同的特征的数据对象*
  - **案例1：**异常值是干扰数据分析的噪声
  - **案例2：**异常值是我们分析的目标
    - 信用卡诈骗
    - 入侵检测
    - 地铁抓小偷



数据质量问题——缺失值

- 数据缺失的原因
  - 未收集到信息

    （例如，人们拒绝透露自己的年龄和体重）
  - 属性可能不适用于所有情况

    （例如，年收入不适用于儿童）
- 处理缺失值
  - 消除数据对象或变量
  - 估计缺失值并补全
    - 例如：温度的时间序列
    - 例如：普查结果
  - 在分析过程中忽略缺失值

数据质量问题——重复数据

- 数据集可以包括重复的或彼此几乎重复的数据对象
  - 合并来自异构源的数据时的主要问题
- 例如:
  - 同一个人有多个电子邮箱地址
- 数据清洗
  - 处理重复数据问题的过程
- 何时不应删除重复数据?
  - 如：两个人具有相同姓名、年龄等基本属性

目标：

- 数据获取
  - 数据检索
  - 批量数据获取
  - 网络爬虫
  - 数据筛选

数据检索：

- 最简单、最灵活的数据获取方式就是依靠检索
- 学会使用搜索引擎（非常重要且必要）
  - 百度：适合于搜索中文信息
  - Google：更适合搜索英文信息

批量数据获取：

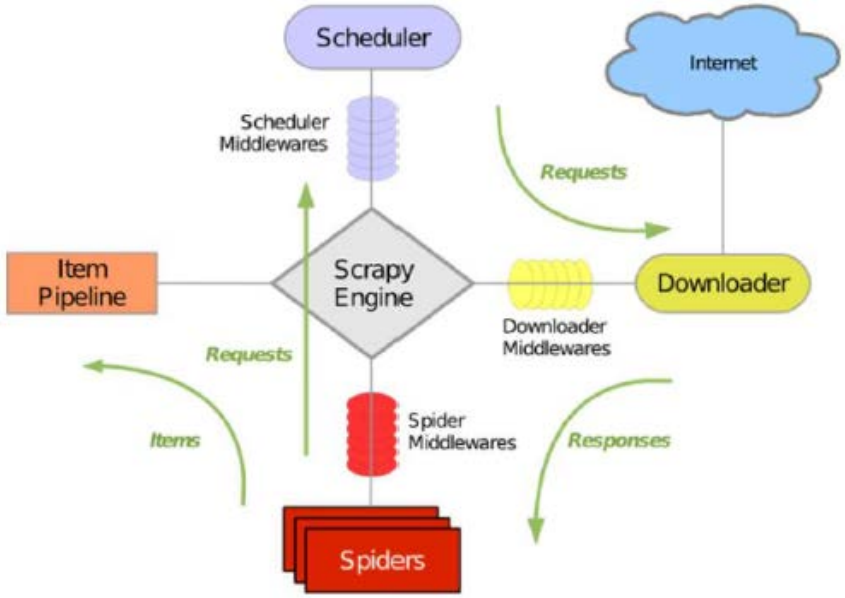
- 大量数据的获取难以手动实现，需借助爬虫程序
  - 也有可能通过交易（购买）“数据”而得
- 网络爬虫是一个自动在网上抓取数据的程序
  - 爬虫本质上就是下载特定网站网页的HTML/JSON/XML数据，并对数据进行解析、提取与存储
  - 通常先定义一组入口URL，根据页面中的其他URL，深度优先或广度优先的遍历访问，逐一抓取数据

为什么要进行数据挖掘：

- 网络爬虫是什么？
  - 网络爬虫（又被称为网页蜘蛛，网络机器人，在FOAF社区中间，更经常的称为网页追逐者），是一种按照一定的规则，自动的抓取网络信息的程序或者脚本。

- 网络爬虫新手教程
  - http://cuiqingcai.com/927.html

- 可以使用大量开源爬虫工具
  - 基于Python的工具
    - Scrapy
    - Beautiful-soup



目标：

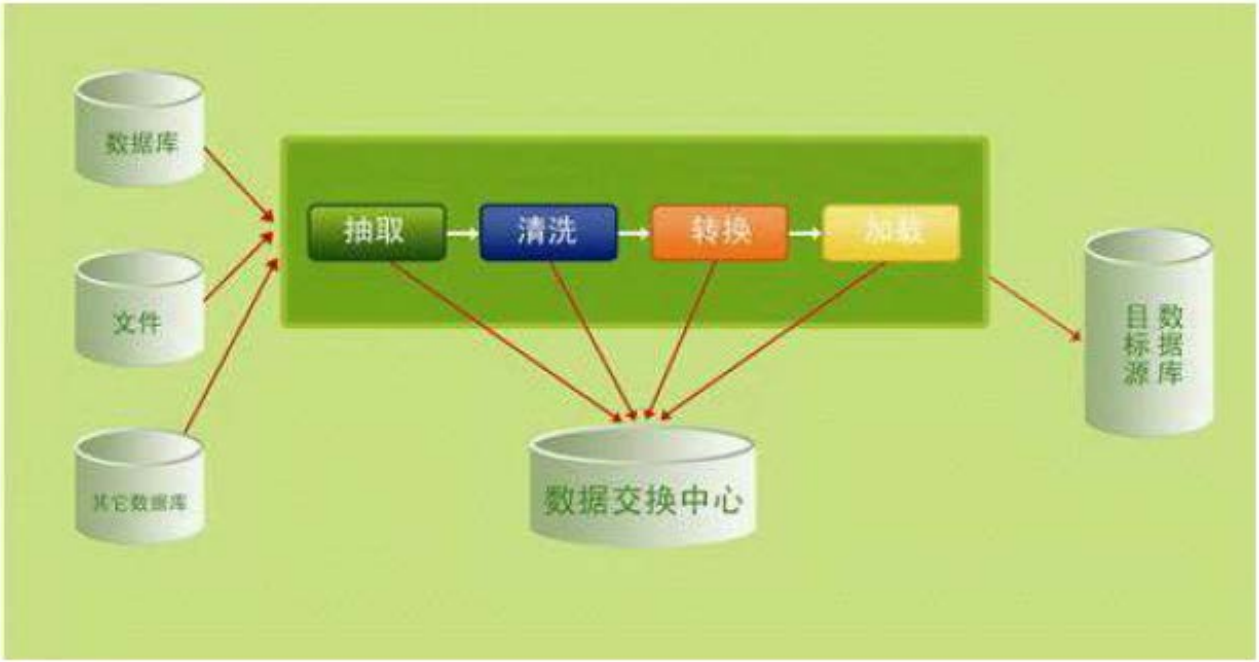
- 数据预处理
  - 各类数据质量问题处理
  - 数据集成
  - 数据变换
  - 数据归约
  - 准备建模数据集
  - 常用方法

各类数据质量问题处理：

- 如何处理各类数据质量问题?
  - 噪声（Noise）
    - 使用信号处理和图像处理的降噪技术
    - 重复测量求平均
  - 离群点（outliers）
    - 离群点检测技术
    - 使用鲁棒性算法
  - 不一致的数据（Wrong data）
    - 数据校验与更正
  - 虚假数据（Fake data）
    - 删除数据对象或属性
  - 数据缺失 （Missing values）
    - 删除数据对象或属性
    - 补全缺失值（使用固定值、平均值、中位数，算法估计）
    - 分析时忽略缺失值
  - 重复数据 （Duplicate data）
    - 去重复（deduplication）

数据集成：

- 数据集成是将多个数据源中的数据结合起来存放在一个一致的数据存储（如数据仓库）中。
  - 许多实际数据天然存放在多个数据源中，在进行数据挖掘时往往需要对其进行集成。



数据变换：

- 数据离散化(Discretization)或二值化（Binarization）
- 简单函数变换：$\log x, e^x, |x|, \ldots$
- 规范化或标准化
  - **最小-最大规范化**：对给定的数值属性A，$[\min_A, max_A]$为A规格化前的取值区间，$[new\_min_A, new\_max_A]$ 为A规范化后的取值区间，最小-最大规格化根据下式将A的值v规格化为值v' ：

$$v' = \frac{v - \min_A}{\max_A - \min_A}(\text{new\_max}_A - \text{new\_min}_A) + \text{new\_min}_A$$

  - **零-均值规范化**：对给定的数值属性A，$\bar{A}$和σ_A分别为A的平均值、标准差，零-均值规格化根据下式将A的值v规格化为值v' ：

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

数据归约：

- 属性归约

  又称为维归约、属性子集选择、特征子集选择，它通过删除不相关的或冗余的属性减小数据集，目标是找出最小属性集，使得数据在其上的概率分布尽可能地接近在原属性集上的概率分布。

  – Aggregation（聚合）：多个属性聚合为一个。如深圳近三年每月降雨量，聚合为近三年每月平均降雨量等；
  – 降维（PCA）
  – 基于决策树等方式进行属性子集选择

- 记录归约

  通过用少量记录代表或替换原有记录来减小数据集。

  – 抽样（要求：采样样本具有代表性representative，即样本与原始数据同分布）
  – 数据概化（面向属性归纳）

准备建模数据集：

- 准备建模数据集是建模前对数据的最后一个必备处理。

- 而在建模之前，还需要考虑到数据的正负样本的比例。通常，对于正样本稀疏的数据，选用15%~30%的负样本来建模。
  - 例如：在建立欺诈检测模型时，欺诈记录的数据占比例很小。如果直接用这样的数据进行建模，那么，预测结果为无欺诈的可能性将会很高。但是这样得到的模型用处不大甚至完全无用。

- 为了评估模型，一般将建模数据集分三个部分，即训练集（training set）、验证集（validation set）、测试集（test set）。将数据的训练集作为最初用于建立模型的数据，用验证集来优化模型，用测试集来评估模型。

目标：

- 数据预处理
  - 各类数据质量问题处理
  - 数据集成
  - 数据变换
  - 数据归约
  - 准备建模数据集
  - 常用方法

常用方法:

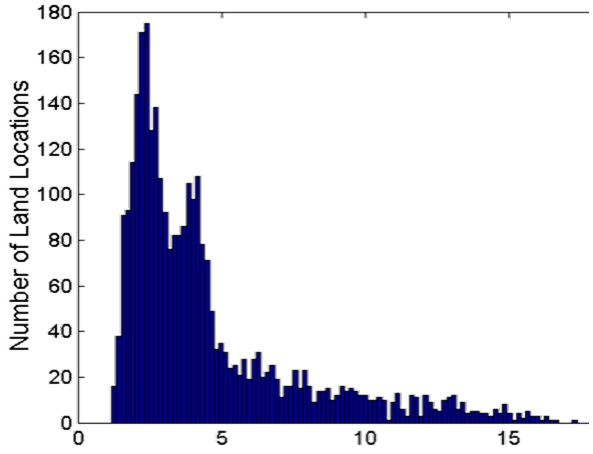
- 聚集
- 抽样
- 离散化和二元化
- 属性转换
- 降维/维归约
- 特征子集选择
- 特征创建

数据聚集

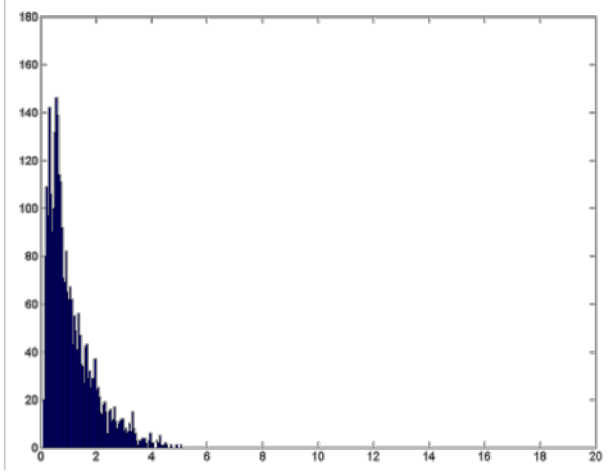
- 将两个或多个属性（或对象）组合为单个属性（或对象），对应属性（记录）归约。
- 目的
  - 数据量缩减
    - 减少属性或对象的数量
  - 规模变化
    - 城市聚合成地区、州、国家等
    - 天数聚合为周、月或年
  - 更“稳定”的数据
    - 聚合数据的变异性(variability)较小

数据聚集–例子

- 该示例基于1982年至1993年期间澳大利亚的降水量。
  - 平均月降水量的标准差直方图（澳大利亚国土按经纬度0.5◦乘以0.5◦大小分成 3,030 个网格）。
  - 平均年降水量的标准差直方图。
- 平均年降水量比平均月降水量变异性小。
- 所有降水量的测量（以及它们的标准差）都以厘米（cm)为单位。



平均月降水量的标准差直方图



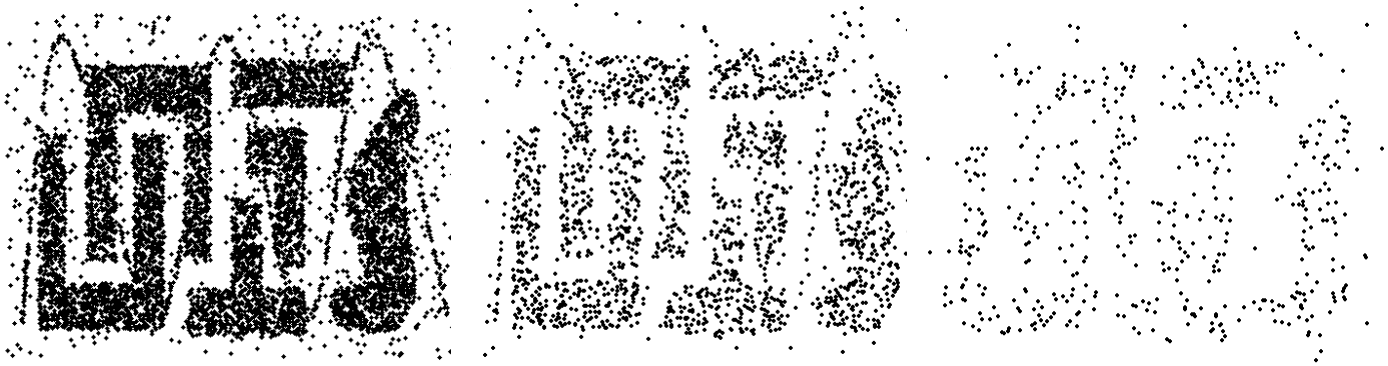
平均年降水量的标准差直方图

抽样

- 抽样是一种选择数据对象子集进行分析的常用方法。
  - 统计学家经常抽样，因为获取感兴趣的整个数据集过于昂贵或耗时。
  - 数据挖掘中进行抽样通常是由于处理所有数据所需要的内存或时间方面的计算成本太高。

- 有效采样的关键原则如下:
  - 如果样本具有代表性，那么使用样本几乎与使用整个数据集一样有效
  - 如果样本具有与原始数据集大致相同的属性，则样本具有代表性

样本容量



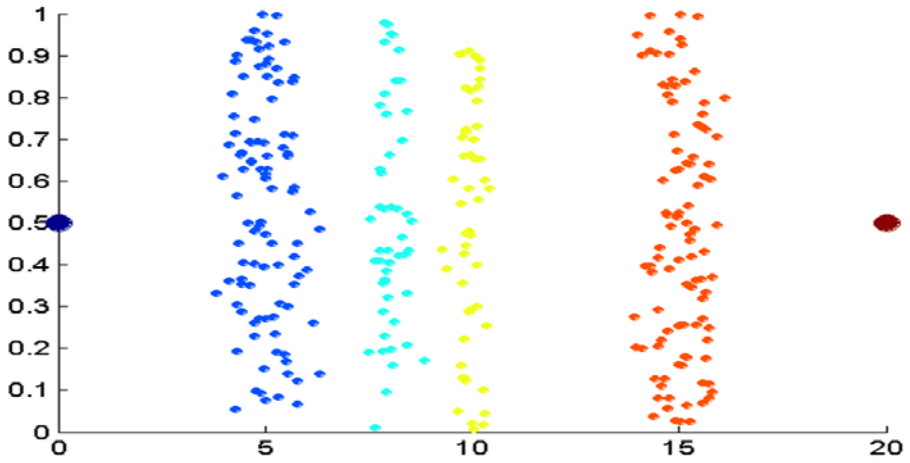
8000 个点　　2000 个点　　500 个点

抽样方法

- 简单随机抽样
  - 选取任何特定项的概率相等
  - 无放回抽样
    - 每个选中项将被立即从总体中删除
  - 有放回抽样
    - 对象被选中时不从总体中删除
    - 相同的对象可能被多次抽出
- 分层抽样
  - 将数据分割成几个组；然后从每个组中随机抽取样本

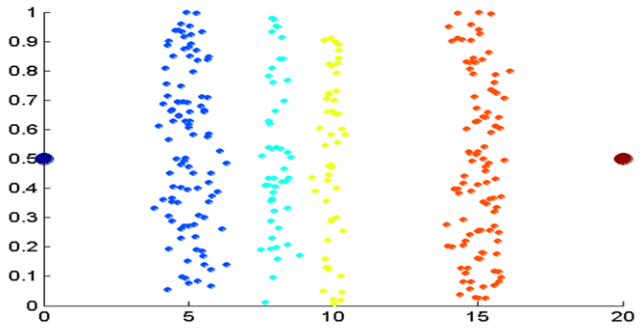
离散化

- 离散化是将连续属性转换为序数属性的过程
  - 可能无限多的值映射到少数类别中
  - 在无监督和有监督的问题设定中都会使用到离散化

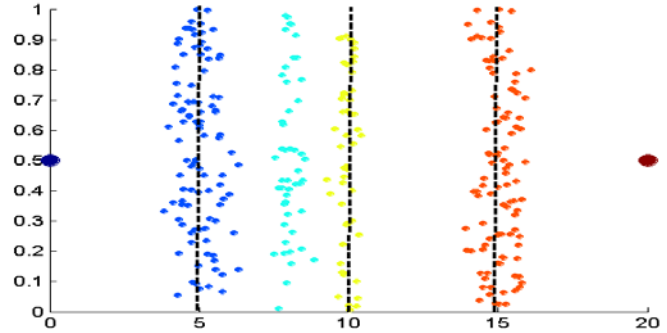


离散化：数据由四组点和两个异常值组成。

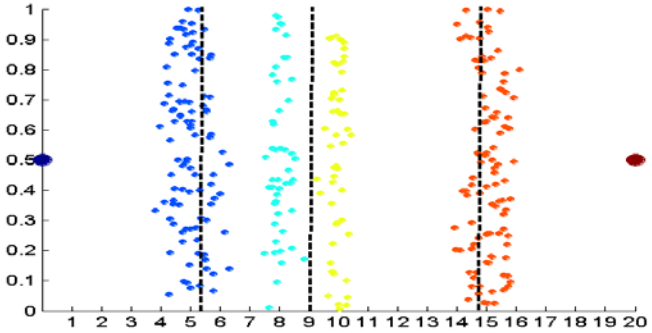
无监督离散化



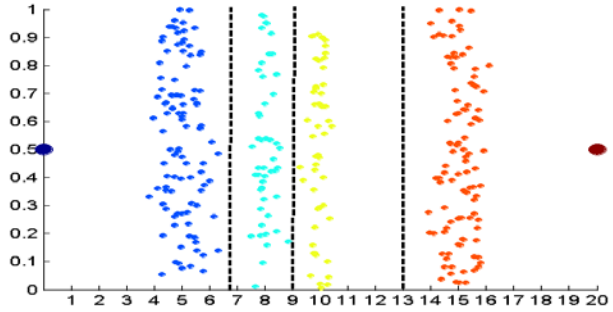
原始数据



用于获得4个值的等宽离散化方法。



用于获得4个值的等频率离散化方法。



用于获得4个值的K均值方法。

监督离散化

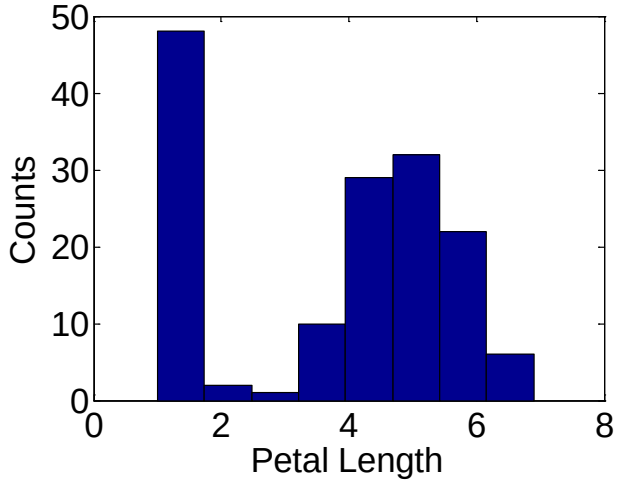
- 如果自变量和因变量只有很少的值，那么许多分类算法可以表现得很好
- 使用Iris数据集举例说明了进行离散化的有效性

- Iris植物数据集
  - 可从UCI机器学习库获取 http://www.ics.uci.edu/~mlearn/MLRepository.html
  - 来自统计学家
  - 三种花卉类型（类别）：
    - Setosa
    - Versicolour
    - Virginica
  - 四个属性
    - 萼片宽度和长度
    - 花瓣宽度和长度

Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.

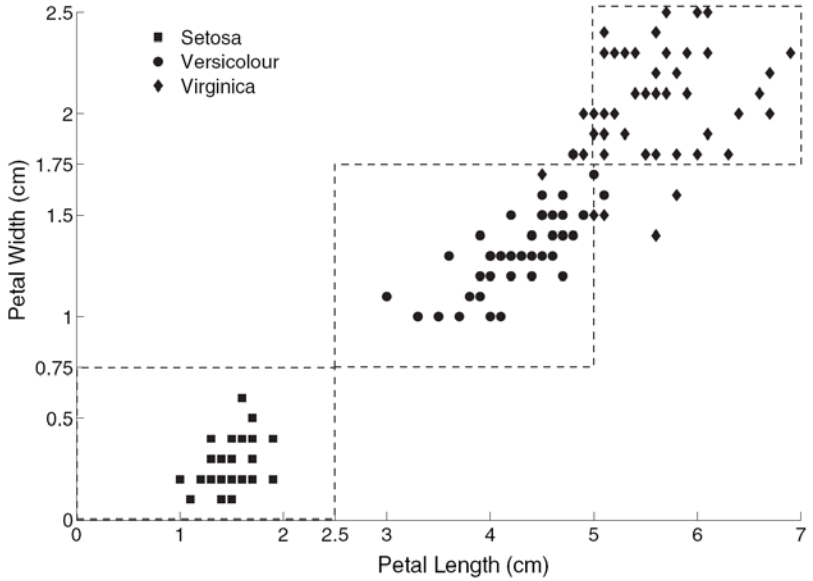
离散化：Iris例子...

- 我们如何知道什么是最好的离散化?
  - 无监督离散化:在数据值中查找断点
    - 例如：花瓣长度



  - 监督离散化:使用类标签查找断点

离散化：Iris例子



花瓣宽度小或花瓣长度小意味着是Setosa 。
花瓣宽度中等或花瓣长度中等意味着是Versicolour 。
花瓣宽或花瓣长意味着是Virginica。

二元化

- 二元化将连续或分类属性映射为一个或多个二元变量
- 通常用于关联分析
- 通常将连续属性转换为分类属性，然后将分类属性转换为一组二元属性
  - （关联分析需要不对称的二元属性）
  - 示例：one-hot encoding

分类值	整数值	x_1	x_2	x_3
awful	0	0	0	0
poor	1	0	0	1
OK	2	0	1	0
good	3	0	1	1
great	4	1	0	0

属性变换

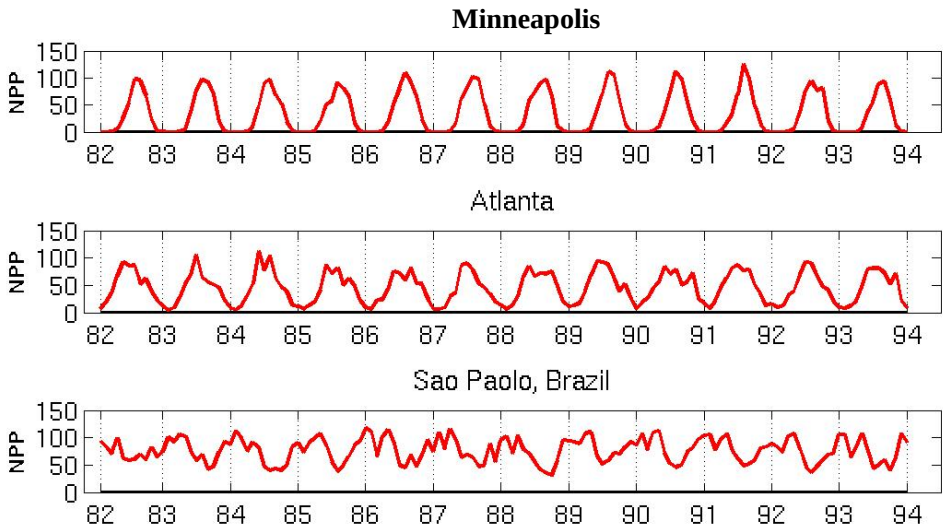
- 属性变换是一个函数，它将给定属性的整组值映射到一组新的替换值，以便可以用一个新值标识每个旧值
  - 简单函数变换: log(x), e^x, |x|等
  - 规范化或标准化
    - 指在发生频率、平均值、方差和范围方面调整属性差异的各种技术
    - 去除不需要的、常见的信号，例如季节性
  - 在统计学中，标准化是指减去均值，除以标准差
    - 零-均值规范化/z-score标准化：对给定的数值属性A，$\bar{A}$和σ_A分别为A的平均值、标准差，零-均值规范化根据下式将A的值v规范化为值v' ：

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

    - 最小-最大规范化：

将数据值映射到[0, 1] $x^* = \frac{x - min}{max - min}$ 映射到[-1, 1] $x^* = 2\frac{x - min}{max - min} - 1$

例子：植物生长的样本时间序列



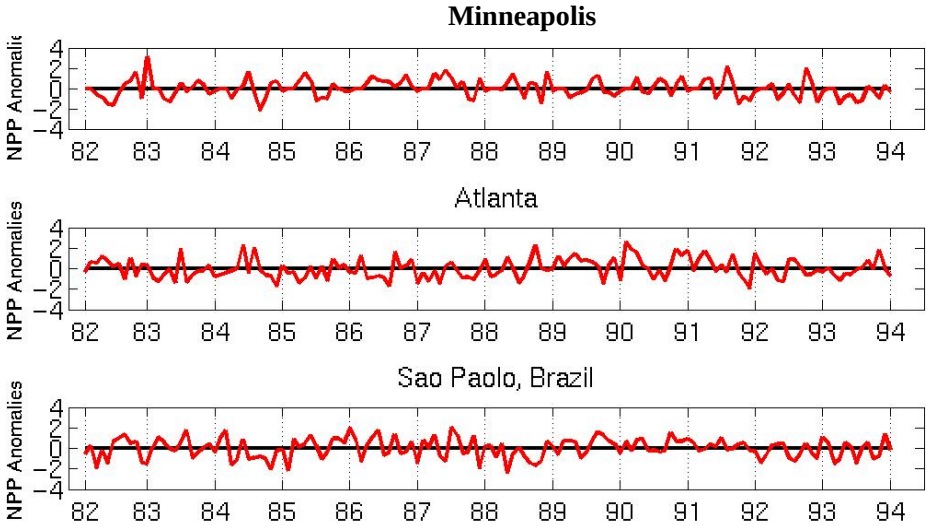
时间序列之间的相关性

	Minneapolis	Atlanta	Sao Paolo
Minneapolis	1.0000	0.7591	-0.7581
Atlanta	0.7591	1.0000	-0.5739
Sao Paolo	-0.7581	-0.5739	1.0000

* 净初级生产力（Net Primary Production， NPP）是生态系统科学家用来衡量植物生长的指标。

例子：植物生长的样本时间序列

使用每月Z Score进行标准化：减去月平均值，然后除以月标准差

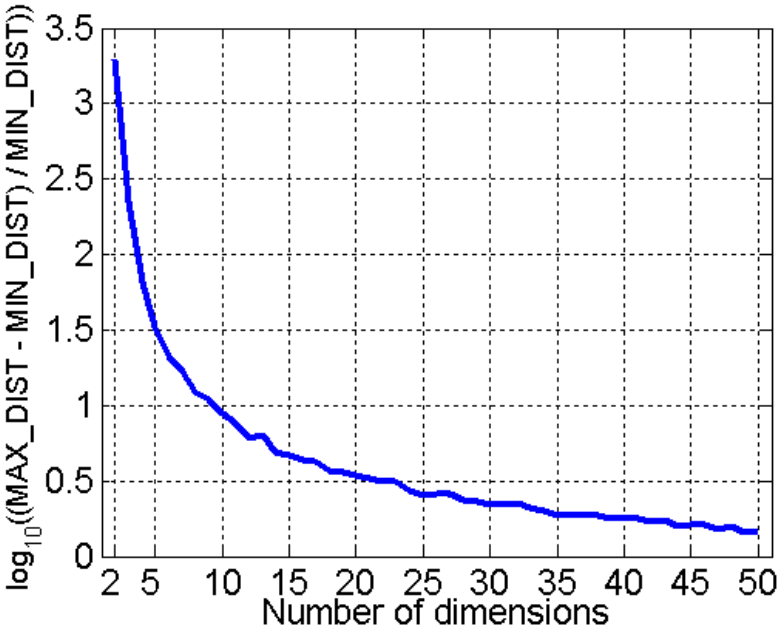


时间序列的相关性

	Minneapolis	Atlanta	Sao Paolo
Minneapolis	1.0000	0.0492	0.0906
Atlanta	0.0492	1.0000	-0.0154
Sao Paolo	0.0906	-0.0154	1.0000

维度诅咒

- 当维度增加时，数据在它所占据的空间中变得越来越稀疏
- 点之间的密度和距离的定义对于聚类和离群点检测至关重要，但当维度增加时，这二者变得越发难以衡量
- (Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful)



- **随机产生500个点，计算任意一对点之间的最大距离和最小距离之差**

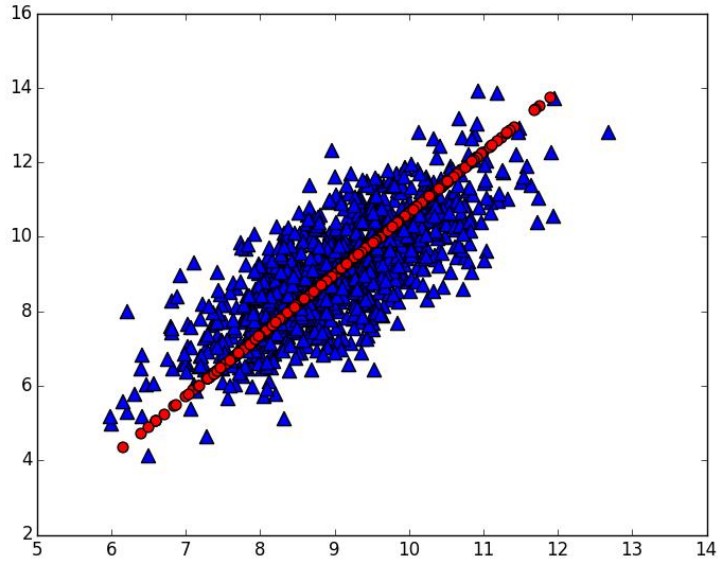
降维技术

- 目的:
  - 避免维度诅咒
  - 减少数据挖掘算法所需的时间和内存量
  - 使数据更容易可视化
  - 可能有助于消除不相关的特征或减少噪音

- 技术
  - 主成分分析（Principal Components Analysis, PCA)
  - 奇异值分解（Singular Value Decomposition, SVD）
  - 其他：监督和非线性技术

PCA

PCA： 主成分分析， 降维



降维问题的优化目标：将一组 N 维向量降为 m 维，其目标是选择 m 个单位正交基，使得原始数据变换到这组基上后，各变量两两间协方差为 0，而变量方差则尽可能大（在正交的约束下，取最大的 m 个方差）。

PCA

PCA： 主成分分析， 降维

PCA

输入： 样本集 $D = \{x_1, x_2, \ldots, x_n\}$;低维空间维数m

过程：

1：对所有样本进行中心化： $x_i \leftarrow x_i - \frac{1}{n}\sum_{i=1}^{n} x_i$

2：计算样本的协方差矩阵 XX^T

3：对协方差矩阵 XX^T 做特征值分解

4：取最大的m个特征值所对应的单位特征向量 $w_1, w_2, \ldots, w_m$

输出： 投影矩阵 $W = (w_1, w_2, \ldots, w_m)$

https://blog.csdn.net/qq_38262266

特征子集选择

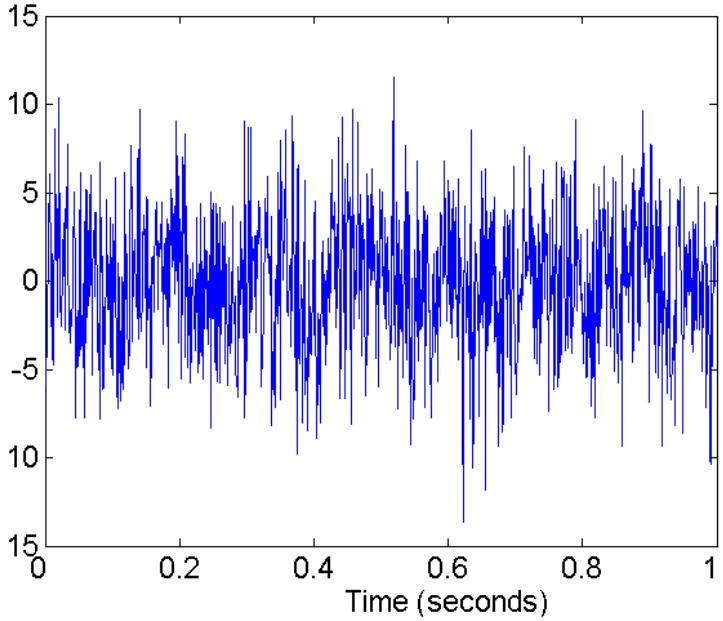
- 降低数据维数的另一种方法
- 冗余特征
  - 复制一个或多个其他属性中包含的大部分或全部信息
  - 示例：产品的购买价格和已支付的销售税金额
- 无关特征
  - 不包含对当前的数据挖掘任务有用的信息
  - 示例：学生ID通常与预测学生GPA的任务无关
- 发展了许多技术，特别是分类技术

特征创建

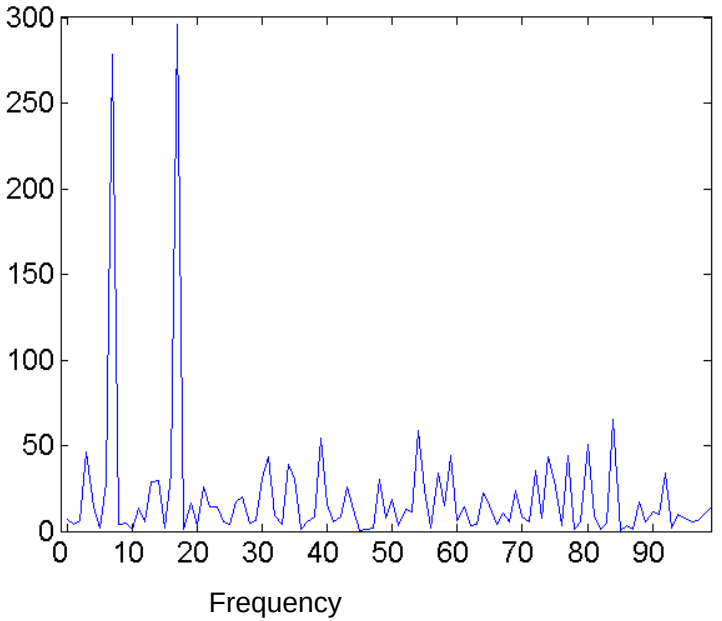
- 创建可以比原始属性更有效地捕获数据集中重要信息的新属性

- 一般分为三种方法:
  - 特征提取
    - 示例：从图像中提取边缘
  - 特征构建
    - 示例：将质量除以体积以获得密度
  - 将数据映射到新空间
    - 示例：傅里叶变换

特征创建——举例



傅里叶变换



时域：两个正弦波+噪声

频域：功率频谱

相似性和相异性的度量

- 相似性度量
  - 两个数据对象的相似程度的数值度量
  - 当对象更相似时，值更高。
  - 通常在[0,1]范围内
- 相异性度量
  - 两个数据对象不同程度的数值度量
  - 当对象更相似时，值更低
  - 相异性最小时通常为0
  - 上限不同
- 邻近度是指相似性或相异性

简单属性的相似度和相异度

- 两个对象（x和y）简单属性之间的相似度和相异度；

属性类型	相异度	相似度
标称的	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
序数的	$d = \lvert x - y \rvert / (n-1)$ （值映射到整数0到n-1,其中n是值的个数）	$s = 1 - d$
区间或比率的	$d = \lvert x - y \rvert$	$s = -d,\ s = \frac{1}{1+d},\ s = e^{-d},$ $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

数据对象之间的相异度：

数据对象之间的相异度

欧式距离（Euclidean distance）：$$d(\mathbf{x},\mathbf{y})=\sqrt{\sum_{k=1}^{n}(x_k-y_k)^2}$$

闵可夫斯基距离（Minkowski distance）：$$d(\mathbf{x},\mathbf{y})=\left(\sum_{k=1}^{n}|x_k-y_k|^r\right)^{1/r}$$

其中 $r=1: l_1-$norm(城市街区距离)，$r=2: l_2-$norm（欧式距离），$r=\infty: l_\infty-$norm（上确界距离）。

距离度量具有的重要性质:

1. *非负性*：$d(\boldsymbol{x},\boldsymbol{y})\geq 0$ for all x and y and $d(\boldsymbol{x},\boldsymbol{y})=0$ iff $\boldsymbol{x}=\boldsymbol{y}$.
2. *对称性*：$d(\boldsymbol{x},\boldsymbol{y})=d(\boldsymbol{y},\boldsymbol{x})$ for all $\boldsymbol{x}$ and $\boldsymbol{y}$.
3. *三角不等式*：$d(\boldsymbol{x},\boldsymbol{z})\leq d(\boldsymbol{x},\boldsymbol{y})+d(\boldsymbol{y},\boldsymbol{z})$for all points $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$.

数据对象之间的相似度：

◘ 数据对象之间的相似度

数据对象之间的相似度通常有非负性和对称性：

1. $s(\boldsymbol{x}, \boldsymbol{y}) = 1$ (or maximum similarity) only if $\boldsymbol{x} = \boldsymbol{y}$.
   (does not always hold, e.g., cosine)
2. $s(\boldsymbol{x}, \boldsymbol{y}) = s(\boldsymbol{y}, \boldsymbol{x})$ for all $\mathbf{x}$ and $\mathbf{y}$. (Symmetry)

- 二元数据的相似性度量

  简单匹配系数 Simple Matching Coefficients:

  SMC =匹配数/属性数

  $= (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$

  Jaccard系数 Jaccard Coefficients:

  J = “11”匹配项的数量/非零属性的数量

  $= (f_{11}) / (f_{01} + f_{10} + f_{11})$

- 余弦相似度: $\cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{<\boldsymbol{x}, \boldsymbol{y}>}{||\boldsymbol{x}|| \, ||\boldsymbol{y}||}$

数据对象之间的相似度：

数据对象之间的相似度

- 广义Jaccard系数： $\mathrm{EJ}(\boldsymbol{x},\boldsymbol{y}) = \dfrac{<\boldsymbol{x},\boldsymbol{y}>}{||\boldsymbol{x}||_2^2 + ||\boldsymbol{y}||_2^2 - <\boldsymbol{x},\boldsymbol{y}>}$

- Pearson's系数:

$$\mathrm{corr}(\mathbf{x},\mathbf{y}) = \frac{\mathrm{covariance}(\mathbf{x},\mathbf{y})}{\mathrm{standard\_deviation}(\mathbf{x}) * \mathrm{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x\ s_y}$$

$$\mathrm{covariance}(\mathbf{x},\mathbf{y}) = s_{xy} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})(y_k - \overline{y})$$

$$\mathrm{standard\_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})^2}$$

$$\mathrm{standard\_deviation}(\mathbf{y}) = s_y = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k - \overline{y})^2}$$

常用数据挖掘工具

常用数据挖掘工具

- Orange
- Weka
- KNIME
- RapidMiner
- Tanagra
- Scikit-learn
- Tensorflow
- Pytorch
- Matlab
- Mahout
- MLlib
- R

数学符号

a, A表示标量

$\boldsymbol{a}, \boldsymbol{A}$分别表示向量和矩阵

l_2-norm: $||\boldsymbol{x}||_2^2 = x_1^2 + x_2^2 + \cdots + x_n^2$
l_1-norm: $||\boldsymbol{x}||_1 = |x_1| + |x_2| + \cdots + |x_n|$
l_∞-norm: $||\boldsymbol{x}||_\infty = \max(|x_1|, |x_2|, \ldots, |x_n|)$

Inner product: $< \boldsymbol{x}, \boldsymbol{y} > = \boldsymbol{x}^T \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i$

$\nabla f(\cdot)$: f可导
$\partial f(\cdot)$: f可微