Data Mining: Data

Lecture Notes for Chapter 2

Introduction to Data Mining , 2nd Edition

by

Tan, Steinbach, Kumar

1

Outline

- Attributes and Objects
- Types of Data
- Data Quality
- Similarity and Distance
- Data Preprocessing

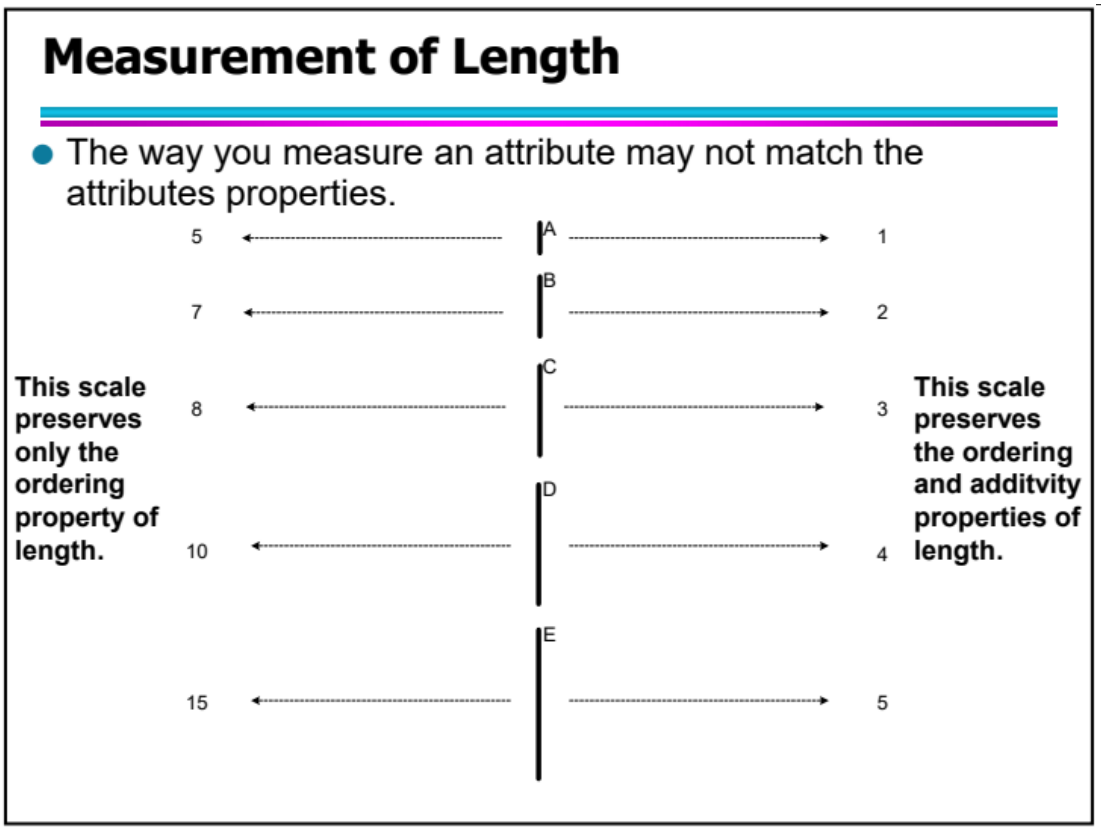
2

什么是数据?

数据对象及其属性的集合。属性是对象的属性或特征
例如:人的眼睛颜色、温度等。属性也称为
变量、字段、特征、尺寸或特征
描述一个对象的属性集合
对象也称为记录、点、案例、样本、实体或实例
Tid 退款婚姻
状态
应纳税的
收入欺诈
1 是单人 125K 否
2 不结婚 10 万不
3 无单个 70K 否
4 是已婚 12 万否
5 不离婚 95K 是的
6 不结婚 6 万不
7 是离婚 22 万否
8 没有单人 85K 是的
9 不结婚 75K 不
没有单个 90K 是 10
属性
目标

01/27/2020 4 数据挖掘导论，第 2 版
谭、斯坦贝克、卡帕特内、库马尔
更完整的数据视图
数据可能有部分
属性(对象)可能与其他属性(对象)有关系
更一般地说，数据可能有结构
数据可能不完整
稍后我们将对此进行更详细的讨论

01/27/2020 5 数据挖掘导论，第 2 版
谭、斯坦贝克、卡帕特内、库马尔
属性值
属性值是分配给特定对象属性的数字或符号。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
　的例子:高度可以用英尺或米来衡量
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⊠，但是属性值的属性可以不同



6

01/27/2020 7 数据挖掘导论，第 2 版
谭、斯坦贝克、卡帕特内、库马尔
属性的类型
名义上有不同类型的属性
　的例子:身份证号码，眼睛颜色，邮政编码

序数
示例:排名(例如，1-10 分制的薯片口味)、等级、身高{高、中、矮}
间隔
的例子:日历日期，摄氏或华氏温度。
比例
示例:开尔文温度、长度、计数、经过时间(例如，赛跑时间)

01/27/2020 8 数据挖掘导论，第 2 版
谭、斯坦贝克、卡帕特内、库马尔
属性值的属性
属性的类型取决于它拥有以下哪些属性/操作:区分度:= 顺序: < >差异是+ -
有意义:
-比率为* /
有意义的
名词性属性:显著性序数属性:显著性和顺序间隔属性:显著性、顺序和有意义
差异
比率属性:所有 4 个属性/操作

01/27/2020 9 数据挖掘导论，第 2 版
谭、斯坦贝克、卡帕特内、库马尔
比率和间隔之间的差异
10 度的温度是 5 度的两倍，这在物理上有意义吗
摄氏温度？华氏温标？开尔文标度？
考虑测量高于平均水平的高度
如果比尔的身高比平均身高高 3 英寸，鲍勃的身高比平均身高高 6 英寸，那么我们会说鲍勃是比尔的两倍高吗？
这种情况类似于温度吗？

属性类型
描述示例操作
名义属性
仅值
区分。(=，☒)
邮政编码、员工身份证号码、眼睛颜色、性别:{男性、女性}
模式，熵，偶然性关联，☒2
测试分类定性序数属性值也排序
物体。(<，>)
矿物质硬度,{好，更好，最好}，等级，街道号
中位数，
百分比、等级相关性、运行测试、符号测试
间隔时间间隔
属性，
值之间的差异是
有意义。(+，-)
日历日期，温度

摄氏还是华氏
平均值，标准偏差，皮尔森氏
相关性，t 和
测试数字定量比率比率变量，包括差异和
比率为
有意义。(*，/)
开尔文温度、货币数量、计数、年龄、质量、长度、电流
几何平均值、调和平均值、百分比变化
属性的这种分类是由于史蒂文斯

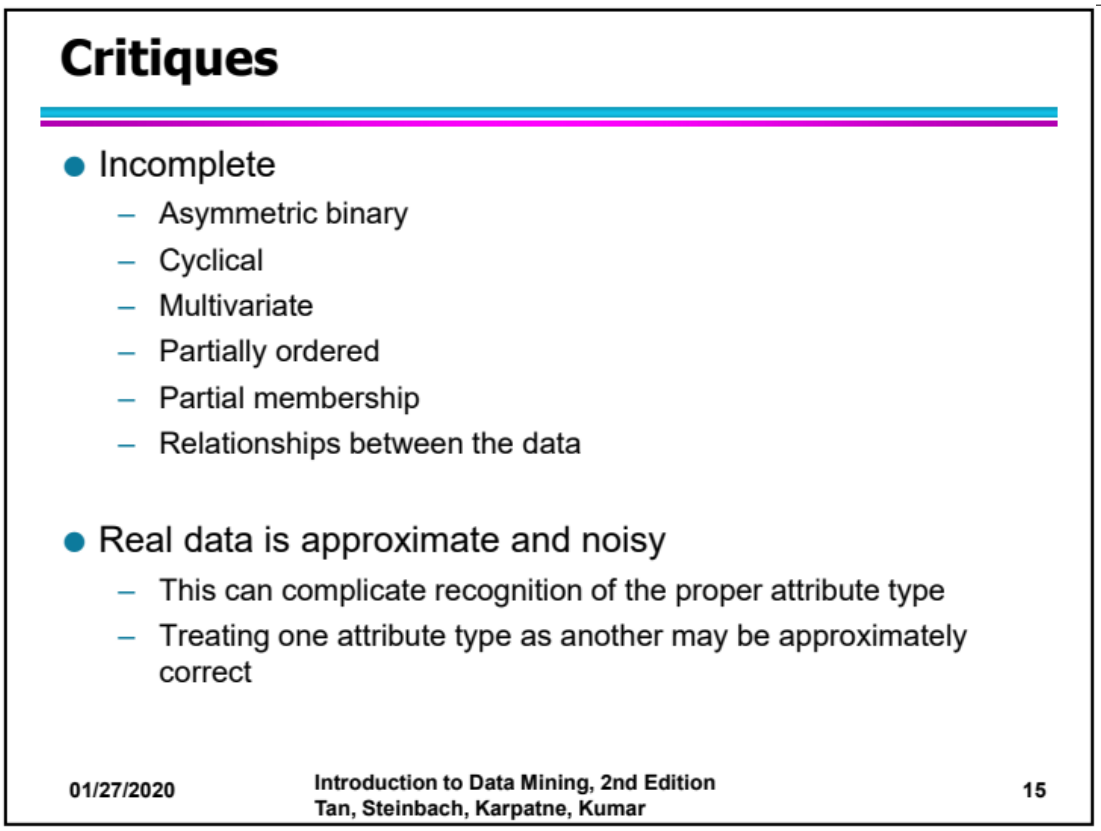
属性类型
转换注释
分类定性
如果所有员工的身份证号码
会有什么不同吗?
序数:保持顺序的变化
值，即，
新值= f(旧值),其中 f 是单调函数
包含好的、更好的最好的概念的属性同样可以用值{1，2，3}或{ 0.5，1，10}来表示。
数字量化
间隔新值= a *旧值+ b
其中 a 和 b 是常数
因此，华氏温标和摄氏温标在零值的位置和单位(度)的大小上是不同的。
比值新值= a *旧值长度可以用
米或英尺。
属性的这种分类是由于史蒂文斯

01/27/2020 12 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
离散和连续属性
离散属性
只有一组有限或可数无限的值示例:邮政编码、计数或
文件的收集
通常表示为整数变量。注意:二进制属性是离散的特例
属性
连续属性
以实数作为属性值。例如:温度、高度或重量。实际上，真实值只能被测量
用有限数量的数字表示。
连续属性通常表示为浮点变量。

01/27/2020 13 数据挖掘导论，第 2 版
谭、斯坦贝克、卡帕特内、库马尔
不对称属性
只有存在(非零属性值)被视为重要
文件中出现的□词客户交易中出现的□项目

如果我们在杂货店遇到一个朋友，我们会说以下的话吗？
“我发现我们的购买非常相似，因为我们没有购买大多数相同的东西。”
我们需要两个不对称的二元属性来代表一个普通的二元属性
关联分析使用不对称属性
不对称属性通常来自集合对象

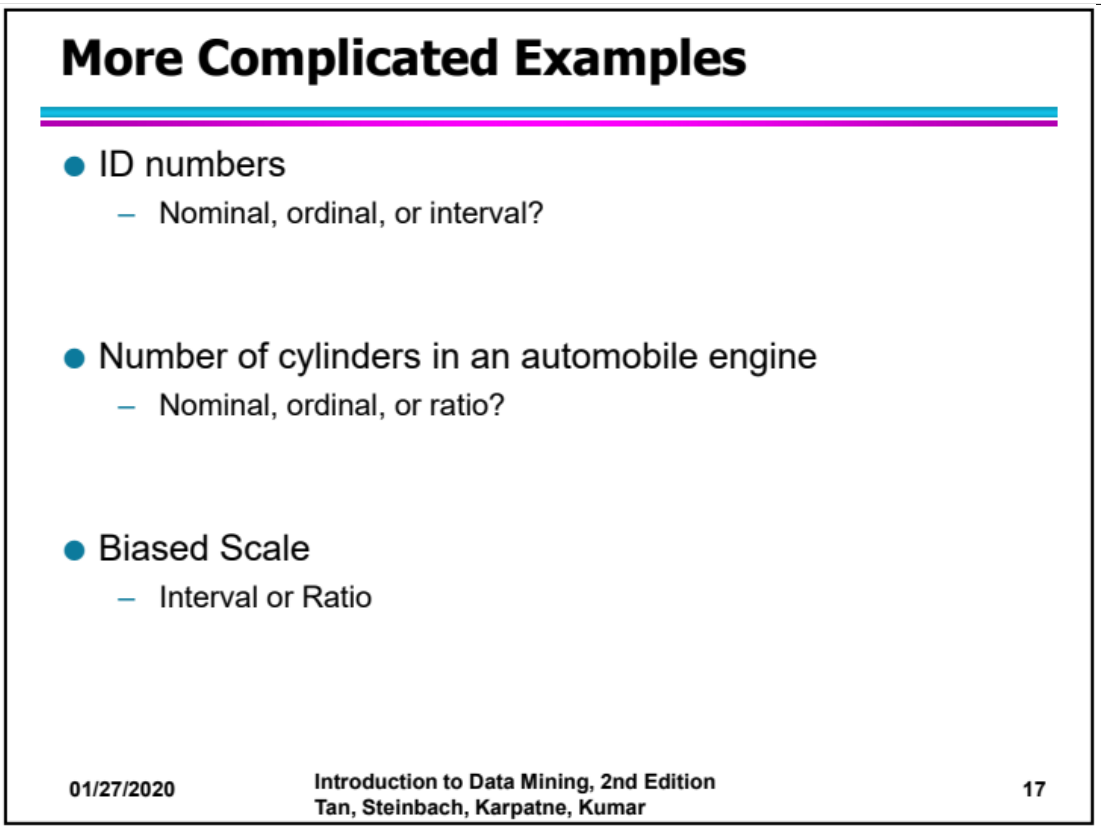
01/27/2020 14 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
一些扩展和评论
维尔曼、保罗·弗和利兰·威尔金森。"名义的、顺序的、间隔的和比率的类型是误导的."美国统计员 47，第 1 号(1993): 65-72。
莫斯特勒、弗雷德里克和约翰·图基。“数据分析和回归。统计学的第二门课程。”爱迪生-韦斯利行为科学系列:定量方法，阅读，大众。:爱迪生-韦斯利，1977 年。
对制图测量水平的再思考制图和地理信息系统 25，第 4 号(1998): 231-242。



15

01/27/2020 16 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
评论…
不是统计分析的好指南
可能会不必要地限制操作和结果
　统计分析通常是近似的
因此，☒举例来说，对序数值使用区间分析可能是合理的
变换是常见的，但不能保持规模

可以将数据转换成具有更好统计特性的新尺度
许多统计分析仅仅依赖于分布



17

01/27/2020 18 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
属性类型的关键消息
您选择的操作类型应该对您拥有的数据类型“有意义”
清晰度、顺序、有意义的间隔和有意义的比率只是数据的四个属性
您看到的数据类型(通常是数字或字符串)可能无法捕获所有属性，或者可能暗示不存在的属性
分析可能依赖于数据的这些其他属性
许多统计分析仅仅依赖于分布
很多时候，有意义的东西是由统计意义来衡量的
但最终，有意义的东西是由领域来衡量的

Types of data sets

- Record
  - Data Matrix
  - Document Data
  - Transaction Data
- Graph
  - World Wide Web
  - Molecular Structures
- Ordered
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data

01/27/2020 20 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
数据的重要特征
维度(属性数量)
高维数据带来了许多挑战
稀少
只有存在才算数
解决
模式取决于规模
大小
类型的分析可能取决于数据的大小

Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

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

21

01/27/2020 22 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
数据矩阵
如果数据对象具有相同的固定数值属性集，那么数据对象可以被视为多维空间中的点，其中每个维度代表一个不同的属性
这样的数据集可以用 m 乘 n 矩阵来表示，其中有 m 行，每个对象一行，n 列，每个属性一列
12.65 6.25 16.22 2.2 1.1
10.23 5.27 15.22 2.7 1.2
投影距离载荷厚度
y 负载
x 载荷的投影
12.65 6.25 16.22 2.2 1.1
10.23 5.27 15.22 2.7 1.2
投影距离载荷厚度
y 负载
x 载荷的投影

Document Data

- Each document becomes a 'term' vector
  - Each term is a component (attribute) of the vector
  - The value of each component is the number of times the corresponding term occurs in the document.

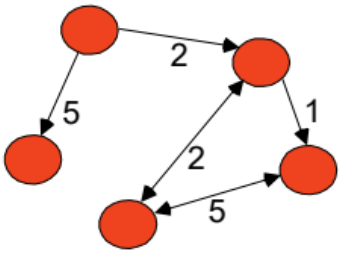
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

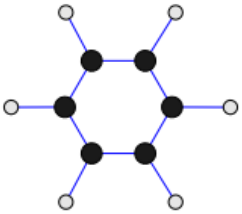
23

01/27/2020 24 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
事务数据
一种特殊类型的数据，其中
每笔交易都涉及一系列项目。例如，考虑一家杂货店。这套产品
顾客在一次购物旅行中购买的商品构成一笔交易，而购买的单个产品是商品。
可以将交易数据表示为记录数据 TID 项目
1 面包，可乐，牛奶 2 啤酒，面包
3 啤酒，可乐，尿布，牛奶 4 啤酒，面包，尿布，牛奶 5 可乐，尿布，牛奶

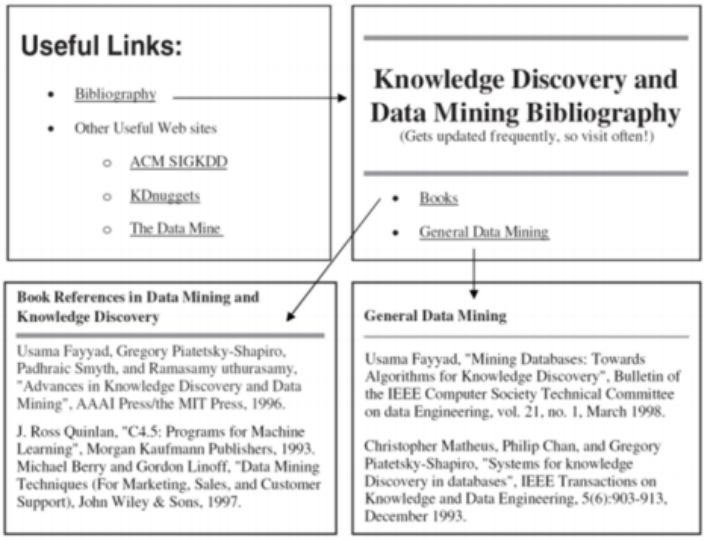
Graph Data

- Examples: Generic graph, a molecule, and webpages



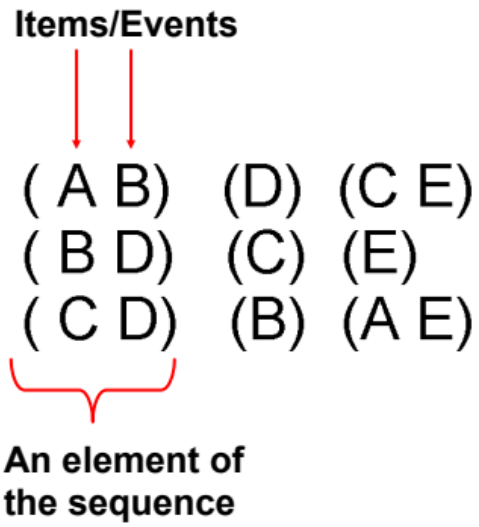


Benzene Molecule: C6H6

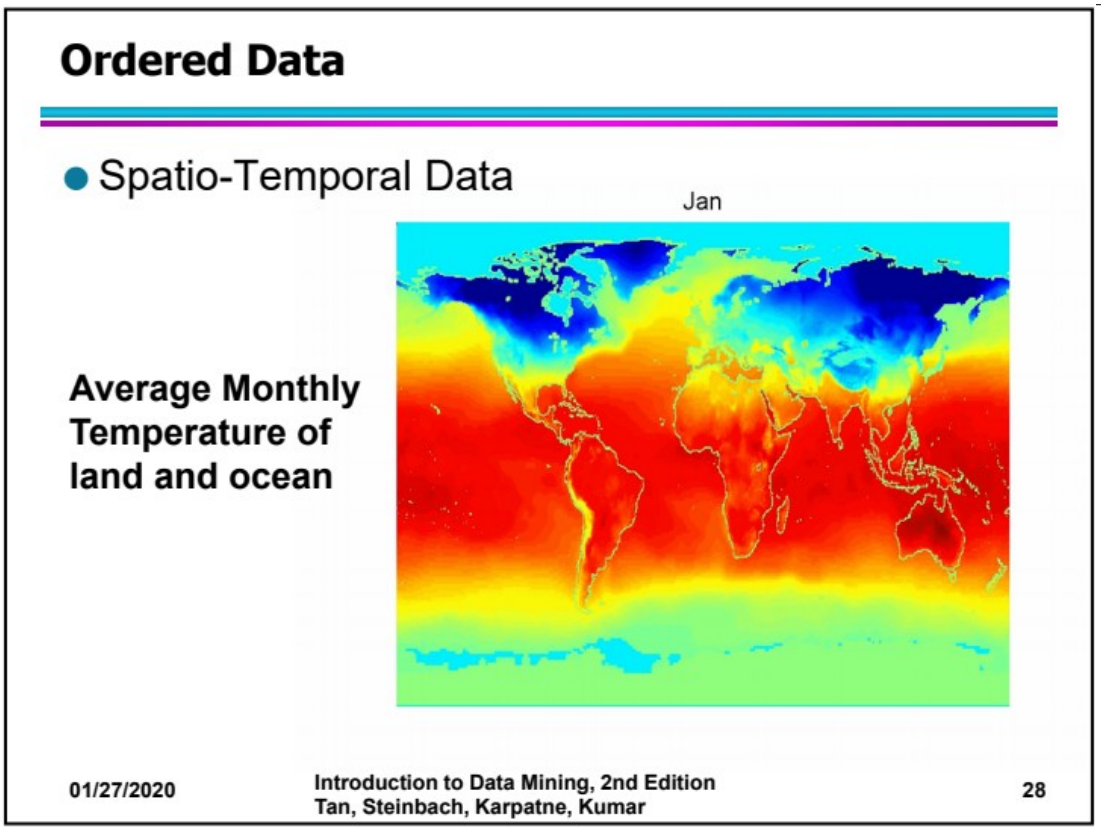


Ordered Data

- Sequences of transactions



谭、斯坦贝克、卡帕特内、库马尔
有序数据
基因组序列数据



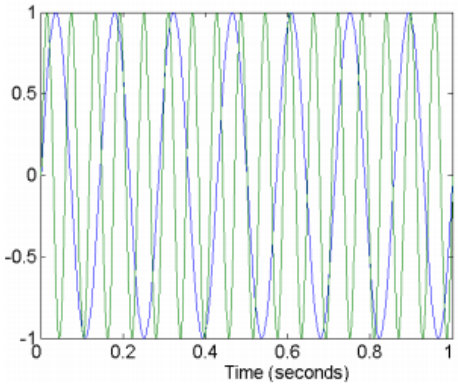
28

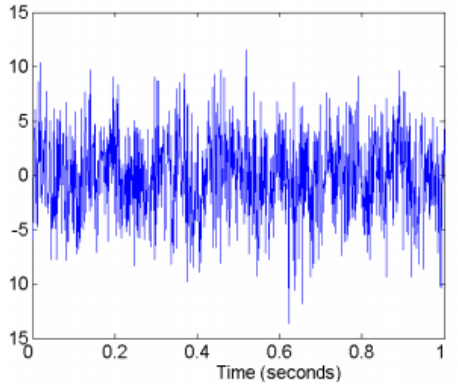
01/27/2020 29 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
数据质量
糟糕的数据质量对许多数据处理工作产生了负面影响
“最重要的一点是，糟糕的数据质量是一场正在上演的灾难。
糟糕的数据质量会使典型公司损失至少百分之十(10%)的收入；20%可能是一个更好的估计。"
托马斯·莱德曼，《管理评论》，2004 年 8 月
数据挖掘示例:使用不良数据建立了一个用于检测贷款风险人群的分类模型
一些有信用的候选人被拒绝贷款更多的贷款被给予违约的个人

01/27/2020 30 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
数据质量…
什么样的数据质量问题？我们如何发现数据的问题？我们能为这些问题做些什么？
数据质量问题的例子:
噪声和异常值缺失值重复数据错误数据虚假数据

Noise

- For objects, noise is an extraneous object
- For attributes, noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen





Two Sine Waves **Two Sine Waves + Noise**

31

01/27/2020 32 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
离群值是数据对象，其特征与数据集中的大多数其他数据对象大不相同
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
原因?
极端值

01/27/2020 33 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
缺失值
价值缺失的原因
没有收集信息
(例如，人们拒绝给出他们的年龄和体重)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
示例:温度时间序列⌧示例:人口普查结果
在分析过程中忽略缺失值

01/27/2020 34 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔

缺少值…
完全随机失踪(MCAR)
值的缺失与属性无关。根据属性填写值。分析总体上可能是无偏的
随机缺失
缺失与其他变量相关。根据其他值填写值几乎总是会在分析中产生偏差
非随机缺失(MNAR)
缺失与未观察到的测量有关，信息性或不可忽略的缺失
无法从数据中了解情况

01/27/2020 35 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
重复数据
数据集可能包括重复或几乎重复的数据对象
合并来自不同来源的数据时的主要问题
示例:
同一个人有多个电子邮件地址
数据清理
处理重复数据问题的过程
何时不应删除重复数据?

01/27/2020 36 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
相似性和相异性度量
相似性度量
两个数据对象相似程度的数值度量。当物体越来越相似时就越高。经常落在[0，1]的范围内
相异度量
两个数据对象有多不同的数值度量
当对象更相似时下限最小相异度通常为 0 上限变化
邻近指的是相似或不同

Similarity/Dissimilarity for Simple Attributes

The following table shows the similarity and dissimilarity between two objects, x and y, with respect to a single, simple attribute.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d - \lvert x - y \rvert / (n-1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Euclidean Distance

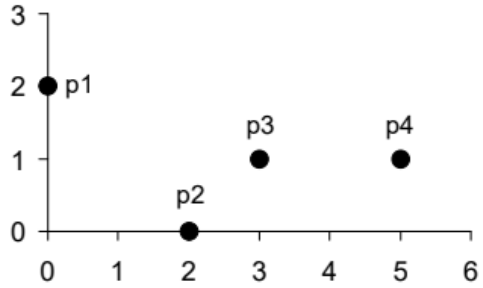
- Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

where n is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects $\mathbf{x}$ and $\mathbf{y}$.

- Standardization is necessary, if scales differ.

Euclidean Distance



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

Where r is a parameter, n is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects $\boldsymbol{x}$ and $\boldsymbol{y}$.

谭、斯坦贝克、卡帕特内、库马尔
闵可夫斯基距离:例子
r = 1。城市街区(曼哈顿，出租车，L1 标准)距离。
二进制向量的一个常见例子是汉明距离，它只是两个二进制向量之间不同的位数
r = 2。欧几里得距离
r　　.距离。
这是向量的任何分量之间的最大差异
不要将 r 与 n 混淆，即所有这些距离都是为所有尺寸定义的。

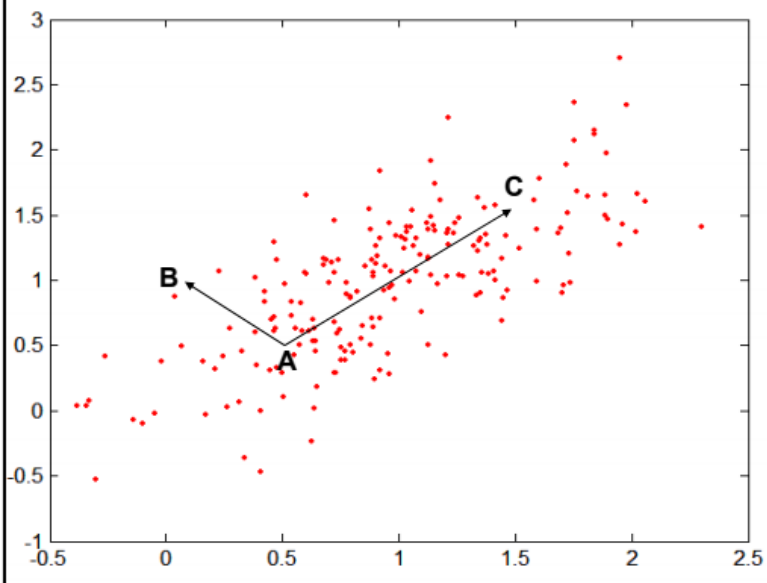
01/27/2020 42 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
闵可夫斯基距离
距离矩阵
点 x y
p1 0 2
p2 2 0
p3 3 1
p4 5 1
L1 p1 p2 p3 p4
p1 0446
p2 4024
p3 4202
p4 6420
L2 p1 p2 p3 p4
p1 0 2.828 3.162 5.099
p2 2.828 0 1.414 3.162
p3 3.162 1.414 0 2
p4 5.099 3.162 2 0
L p1 p2 p3 p4
p1 0235
p2 2013
p3 3102
p4 5320

Mahalanobis Distance

$$\text{mahalanobis}(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T \Sigma^{-1} (\mathbf{x}-\mathbf{y})$$

Σ is the covariance matrix

Mahalanobis Distance



Covariance Matrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4

01/27/2020 45 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
距离的共同性质
距离，如欧几里德距离，有一些众所周知的属性。
1.仅当 x = y 时，所有 x 和 y 的 d(x，y)　0 和 d(x，y) = 0(正定性)
2.所有 x 和 y 的 d(x，y) = d(y，x)(对称)3。所有点 xy 和 z(三角形不等式)的 d(xz)　d(xy) + d(yz),其中 d(x，y)是点(数据对象)x 和 y 之间的距离(相异度)
满足这些属性的距离是一个度量

01/27/2020 46 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
相似性的共同性质
相似之处，也有一些众所周知的属性。
1.仅当 x = y 时，s(xy) = 1(或最大相似性)(不总是成立，例如，余弦)2。所有 x 和 y 的 s(x，y) = s(y，x)(对称)
其中 s(x，y)是点(数据对象)之间的相似性，x 和 y

01/27/2020 47 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
二元向量之间的相似性
常见的情况是对象 x 和 y 只有二进制属性
使用以下数量计算相似性
f01 =其中 x 为 0，y 为 1 的属性数 f10 =其中 x 为 1，y 为 0 的属性数 f00 =其中 x 为 0，y 为 0 的属性数 f11 =其中 x 为 1，y 为 1 的属性数
简单匹配和 Jaccard 系数 SMC =匹配数/属性数= (f11 + f00) / (f01 + f10 + f11 + f00)
j = 11 个匹配项的数量/非零属性的数量= (f11) / (f01 + f10 + f11)

01/27/2020 48 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
SMC 对 Jaccard:示例
x = 1 0 0 0 0 0 0 0 0 0
y = 0 0 0 0 0 1 0 1 0 1
f01 = 2(其中 x 为 0，y 为 1 的属性数)f10 = 1(其中 x 为 1，y 为 0 的属性数)f00 = 7(其中 x 为 0，y 为 0 的属性数)f11 = 0(其中 x 为 1，y 为 1 的属性数)
SMC =(F11+f00)/(f01+F10+F11+f00)=(0+7)/(2+1+0+7)= 0.7
j =(F11)/(f01+F10+F11)= 0/(2+1+0)= 0

01/27/2020 49 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
余弦相似性
如果 d1 和 d2 是两个文档向量，那么 cos(d1，d2)= <d1，d2> / ||d1|| ||d2||，
其中< d1，d2 >表示矢量的内积或矢量点积，dand 和|| d ||是矢量 d 的长度。例如:
d1 = 3205000200
d2 = 1 0 0 0 0 0 0 1 0 2
< d1，D2 > = 3 * 1+2 * 0+0 * 0+5 * 0+0 * 0+0 * 0+0 * 0+2 * 1+0 * 0+0 * 2 = 5 | D1 | | =(3 * 3+2 * 2+0 * 0+5 * 5+0 * 0+0 * 0+0 * 0+2 * 2+0 * 0+0 * 0)0.5 =(42)0.5 = 6.481 | | D2 | | =(1 * 1+0 * 0+0 * 0)

Extended Jaccard Coefficient (Tanimoto)

- Variation of Jaccard for continuous or count attributes
  - Reduces to Jaccard for binary attributes

$$EJ(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \mathbf{x} \cdot \mathbf{y}}$$

Correlation measures the linear relationship between objects

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{covariance}(\mathbf{x}, \mathbf{y})}{\text{standard\_deviation}(\mathbf{x}) * \text{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x \, s_y}, \quad (2.11$$

where we are using the following standard statistical notation and definitions

$$\text{covariance}(\mathbf{x}, \mathbf{y}) = s_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y}) \quad (2.12$$

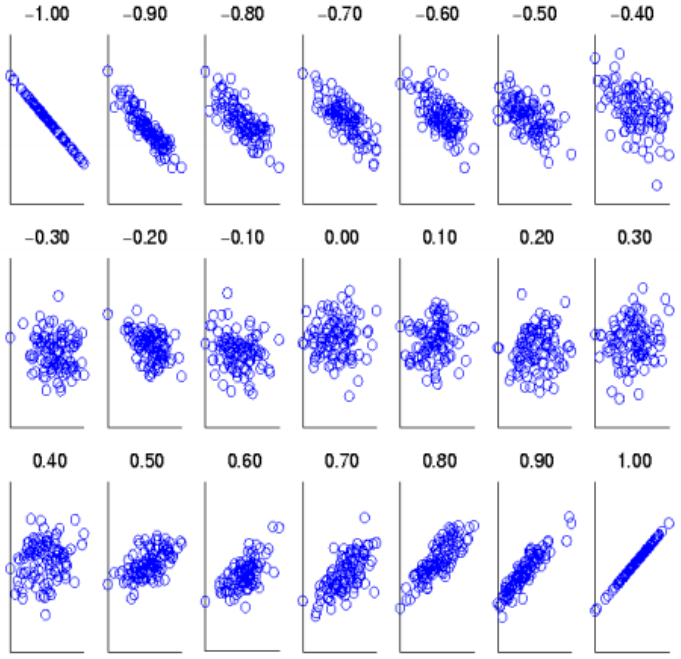
$$\text{standard\_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2}$$

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

$$\overline{x} = \frac{1}{n} \sum_{k=1}^{n} x_k \text{ is the mean of } \mathbf{x}$$

$$\overline{y} = \frac{1}{n} \sum_{k=1}^{n} y_k \text{ is the mean of } \mathbf{y}$$

Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1.

52

Drawback of Correlation

- **x** = (-3, -2, -1, 0, 1, 2, 3)
- **y** = (9, 4, 1, 0, 1, 4, 9)

$y_i = x_i^2$

- mean(**x**) = 0, mean(**y**) = 4
- std(**x**) = 2.16, std(**y**) = 3.74

- corr = (-3)(5)+(-2)(0)+(-1)(-3)+(0)(-4)+(1)(-3)+(2)(0)+3(5) / (6 * 2.16 * 3.74)
  = 0

53

谭、斯坦贝克、卡帕特内、库马尔
邻近度量的比较
应用领域
相似性度量往往特定于属性和数据的类型
记录数据、图像、图形、序列、三维蛋白质结构等。往往有不同的衡量标准
然而，人们可以谈论各种您希望邻近度测量具有的属性
对称是常见的
对噪音和异常值的容忍度是另一种发现更多类型模式的能力？许多其他可能
衡量标准必须适用于数据，并产生符合领域知识的结果

01/27/2020 55 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
基于信息的度量
信息论是一个发展完善、应用广泛的基础学科
一些相似性度量是基于信息论的
不同版本的互信息最大信息系数及其相关
措施
一般情况下，可以处理非线性关系，计算起来既复杂又耗时

01/27/2020 56 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
信息和概率
信息与事件的可能结果相关
信息的传输、硬币的翻转或数据的测量
结果越确定，包含的信息就越少，反之亦然
例如，如果一枚硬币有两个头像，则头像的结果不提供任何信息
更定量地说，信息与结果的概率有关
:结果的概率越小，它提供的信息就越多，反之亦然
熵是常用的度量

01/27/2020 57 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
熵
用于
变量(事件)，X，
对于 n 个可能的值(结果)，x1，x2 …，xn 每个结果具有概率 p1，p2 …，pn X，H(X)的熵由下式给出
$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$
□
□□□
熵介于 0 和 log2n 之间，以位为单位
因此，熵是衡量平均来说代表一个 X 的观测值需要多少位的尺度

01/27/2020 58 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
熵示例

对于正面概率为 p、反面概率为 q = 1 - p 的硬币
$H\ p\ \log\ p\ q\ \log\ q$
p= 0.5，q = 0.5(公平硬币)H = 1 p = 1 或 q = 1，H = 0
公平的四面骰子的熵是多少?

01/27/2020 59 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
样本数据的熵:示例
最大熵是 log25 = 2.3219
头发颜色计数 p -plog2p
黑色 75 0.75 0.3113
棕色 15 0.15 0.4105
金发 5 0.05 0.2161
红色 0 0.00 0
其他 5 0.05 0.2161
总计 100 1.0 1.1540

01/27/2020 60 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
样本数据的熵
假设我们有
某个属性的观察数(m)，X，例如，班上学生的头发颜色，其中有 n 个不同的可能值，第 I 类的观察数为 mi。那么，对于这个样本
$H\ X\ \ m\ m\ \ \log\ m\ m$
对于连续数据，计算更加困难

01/27/2020 61 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
交互信息
一个变量正式提供另一个变量的信息，$I\ X\ Y\ \ H\ X\ H\ Y\ H\ X\ Y$，其中
H(X，Y)是 X 和 Y 的联合熵，
$Y\ \ pij\log\ pij\ H\ X$
其中，π 是 X 的第 I 个值和 Y 的第 j 个值同时出现的概率
对于离散变量，这很容易计算。离散变量的最大互信息是
log2(最小值(nX，nY)，其中 nX (nY)是 X (Y)的值的数量

01/27/2020 62 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
互信息示例
学生身份
p -plog2p 计数
本科 45 0.45 0.5184
梯度 55 0.55 0.4744

总计 100 1.00 0.9928
分数 p -plog2p
A 35 0.35 0.5301
b5 0.50 0.5000
C 15 0.15 0.4105
总计 100 1.00 1.4406
学生身份
分数 p -plog2p
本科 5 0.05 0.2161
本科 B 30 0.30 0.5211
本科 C 10 0.10 0.3322
甲级 30 0.30 0.5211
学士学位 20 0.20 0.4644
丙级 5 0.05 0.2161
总计 100 1.00 2.2710
学生身份和年级的相互信息= 0.9928 + 1.4406 - 2.2710 = 0.1624

最大信息系数
Reshef、David N、Yakir A. Reshef、Hilary K. Finucane、Sharon R. Grossman、Gilean McVean、Peter J. Turnbaugh、Eric S. Lander、Michael Mitzenmacher 和 Pardis C. Sabeti。"在大数据集中检测新的关联."science 334，no. 6062 (2011): 1518-1524。
将互信息应用于两个连续变量
考虑变量可能归入离散类别 nX × nY ≤ N0.6，其中
nX 是 x 的数值数， nY 是 y 的数值数
N 是样本(观察值、数据对象)的数量
计算相互信息
用 log2 归一化(最小值(nX，nY)
获取最高价值

结合相似性的一般方法
有时属性有许多不同的类型，但需要整体的相似性。
1:对于第个属性，计算相似度 sk(x，y)，范围为[0，1]。
2:为 kth 属性定义一个指标变量k，如下所示:
如果 kth 属性是非对称属性，k = 0，并且
两个对象的值都为 0，或者如果其中一个对象缺少 kth 属性k = 1 的值，则为
3.计算

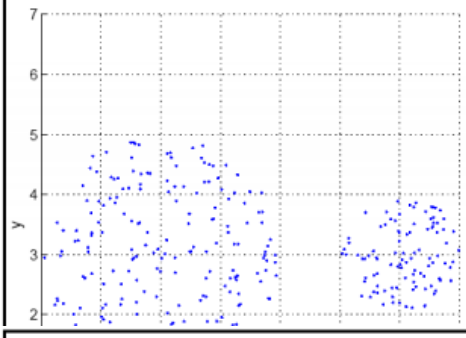
使用权重组合相似性
可能不想对所有属性一视同仁。使用非负权重ω
$similarity(\mathbf{x}, \mathbf{y})$ ∑ ∑

还可以定义距离的加权形式

01/27/2020 66 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
密度
测量数据对象在指定区域内相互接近的程度
密度的概念与接近度密切相关。密度的概念通常用于聚类，并且
异常检测示例:
欧几里德密度
　欧几里德密度=每单位体积的点数
概率密度
　估计了数据的分布情况
基于图的密度⊠连通性

Euclidean Density: Grid-based Approach

- Simplest approach is to divide region into a number of rectangular cells of equal volume and define density as # of points the cell contains



0	0	0	0	0	0	0
0	0	0	0	0	0	0
4	17	18	6	0	0	0
14	14	13	13	0	18	27
11	18	10	21	0	24	31

Euclidean Density: Center-Based

- Euclidean density is the number of points within a specified radius of the point

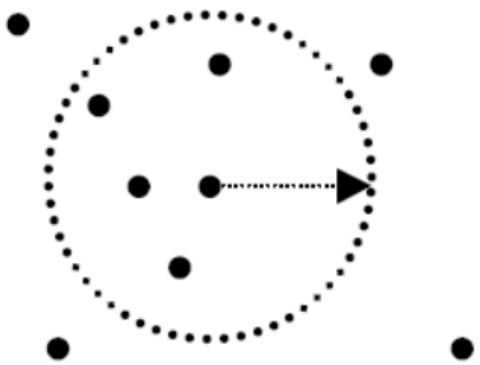


Illustration of center-based density.

Data Preprocessing

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation

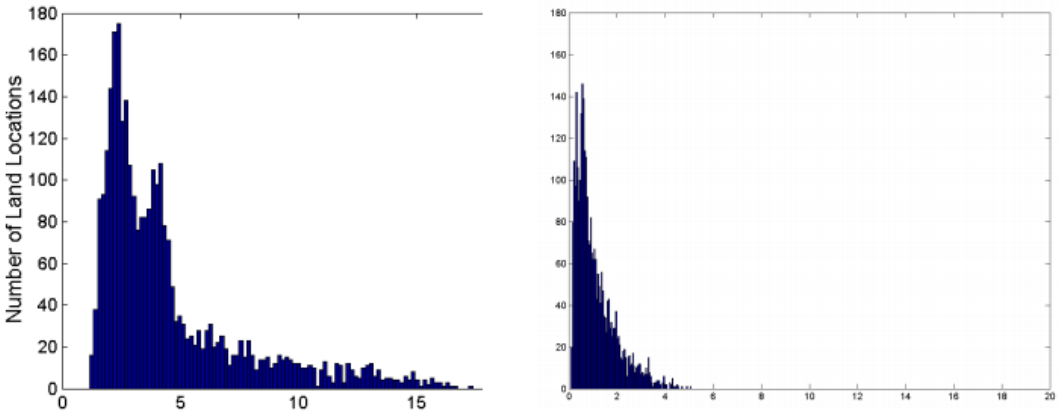
69

01/27/2020 70 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
聚合
将两个或多个属性(或对象)组合成一个属性(或对象)
目的
数据整理
减少属性或对象的数量
规模变化
城市聚集成地区、州、国家等。☒日累计为周、月或年
更多“稳定”的数据
汇总数据的可变性更小

01/27/2020 71 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
例子:澳大利亚的降水
这个例子是基于 1982 年至 1993 年澳大利亚的降雨量。
下一张幻灯片显示
澳大利亚 3，030±0.5×0.5 网格单元的月平均降水量标准偏差直方图，以及
同一地点年平均降水量标准差的直方图。
平均年降雨量的可变性小于平均月降雨量。所有降水测量值(及其标准偏差)均以厘米为单位。

Example: Precipitation in Australia ...

Variation of Precipitation in Australia



Standard Deviation of Average Monthly Precipitation

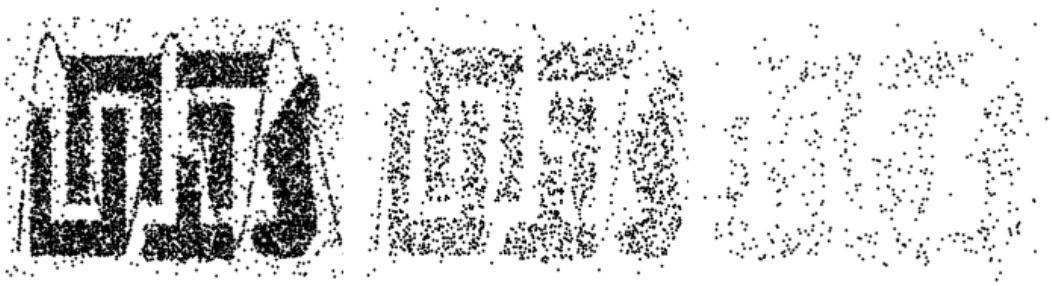
Standard Deviation of Average Yearly Precipitation

72

01/27/2020 73 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
抽样
取样是用于数据简化的主要技术。
它通常用于数据的初步调查和最终数据分析。
统计人员经常进行抽样，因为获取整套感兴趣的数据过于昂贵或耗时。
采样通常用于数据挖掘，因为处理整个感兴趣的数据集过于昂贵或耗时。

01/27/2020 74 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
取样…
有效取样的主要原则如下:
如果样本具有代表性，那么使用样本几乎和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(感兴趣)，则该样本具有代表性

Sample Size



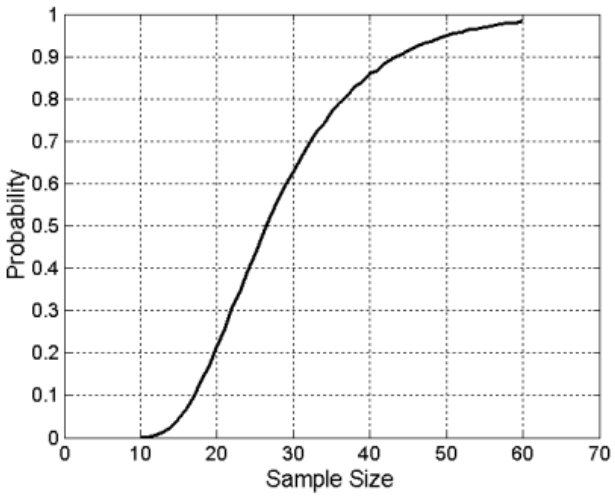
8000 points　　2000 Points　　500 Points

75

01/27/2020 76 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
抽样类型
简单随机抽样
选择任何特定项目的可能性是相等的
不替换取样
　:当每一个项目被选中时，它就被从人口中删除
补替抽样法
　对象不会从总体中移除，因为它们是为样本选择的。
　在取样与替换时，同一物体可以被拾取不止一次
分层抽样
将数据分成几个分区；然后从每个分区中随机抽取样本

Sample Size

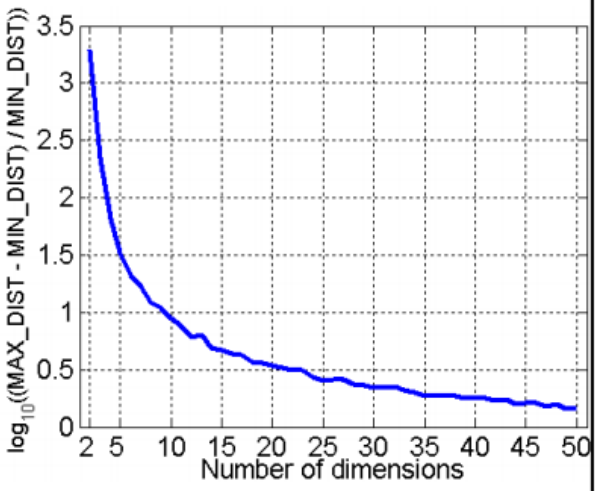
- What sample size is necessary to get at least one object from each of 10 equal-sized groups.



77

Curse of Dimensionality

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies
- Definitions of density and distance between points, which are critical for clustering and outlier detection, become less meaningful



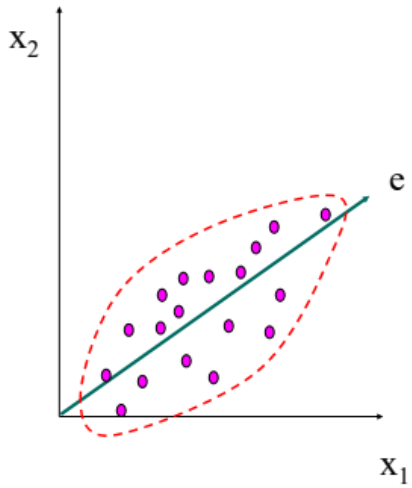
- **Randomly generate 500 points**
- **Compute difference between max and min distance between any pair of points**

78

谭、斯坦贝克、卡帕特内、库马尔
降维
目的:
避免维数灾难
减少数据挖掘算法所需的时间和内存
允许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
主成分分析奇异值分解
其他:监督和非线性技术

Dimensionality Reduction: PCA

- Goal is to find a projection that captures the largest amount of variation in data



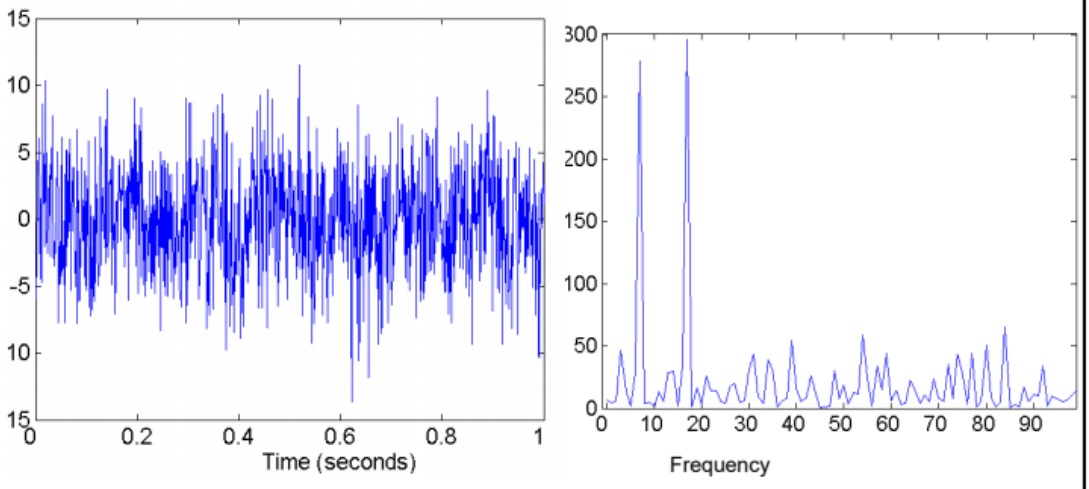
Dimensionality Reduction: PCA

01/27/2020 82 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
特征子集选择
降低数据维数的另一种方法是冗余特征
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:产品的购买价格和支付的销售税金额
无关特征
不包含对手头的数据挖掘任务有用的信息
例子:学生的身份通常与预测学生的平均绩点无关
开发了许多技术，尤其是分类技术

01/27/2020 83 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
特征创建
创建新的属性，可以比原始属性更有效地捕捉数据集的重要信息
三种通用方法:
特征抽出
示例:从图像中提取边缘
特征构造
的例子:将质量除以体积得到密度
将数据映射到新空间
的例子:傅立叶和小波分析

Mapping Data to a New Space

- Fourier and wavelet transform

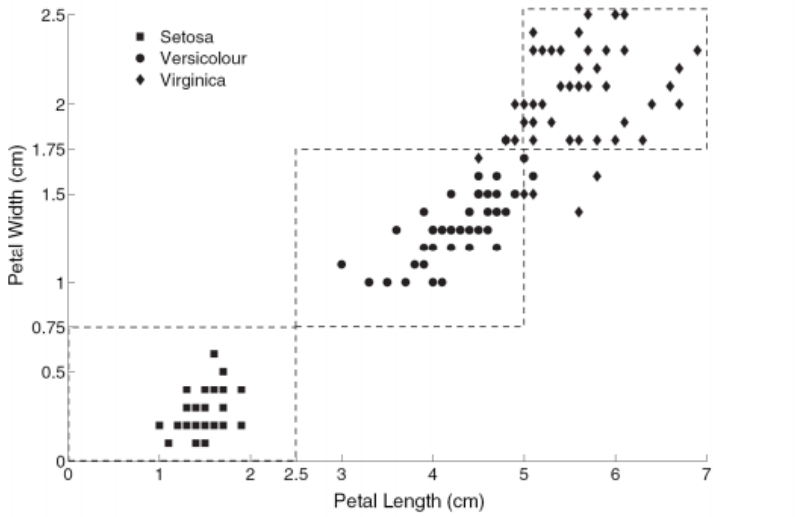


Two Sine Waves + Noise　　Frequency

01/27/2020 85 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
离散化
离散化是将连续属性转换为有序属性的过程
潜在的无限数量的值被映射到少数类别中
离散化通常用于分类
如果自变量和因变量都只有几个值，许多分类算法工作得最好
我们举例说明了使用 Iris 数据集进行离散化的有用性

01/27/2020 86 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
虹膜样本数据集
鸢尾植物数据集。
可以从 http://www.ics.uci.edu/~mlearn/MLRepository.html 的 UCI 机器学习资源库获得
来自统计学家道格拉斯·费希尔的三种花卉类型(类):
·塞托萨
杂色⊠弗吉尼亚
四个(非类)属性⊠萼片宽度和长度
花瓣宽度和长度。罗伯特·莫赫伦布鲁克。美国农业部
NRCS。1995.东北湿地植物区系:植物物种野外办公室指南。宾夕法尼亚州切斯特东北国家技术中心。
由美国农业部 NRCS 湿地科学研究所提供。

Discretization: Iris Example



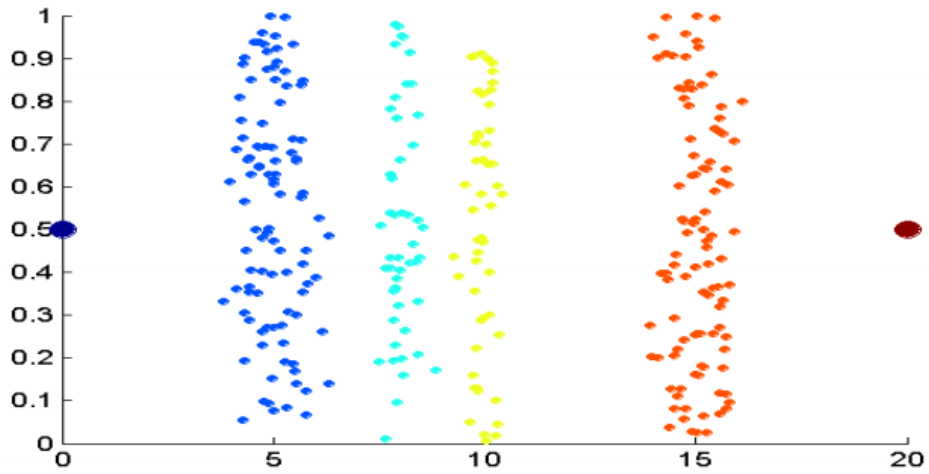
Petal width low or petal length low implies Setosa.
Petal width medium or petal length medium implies Versicolour.
Petal width high or petal length high implies Virginica.

87

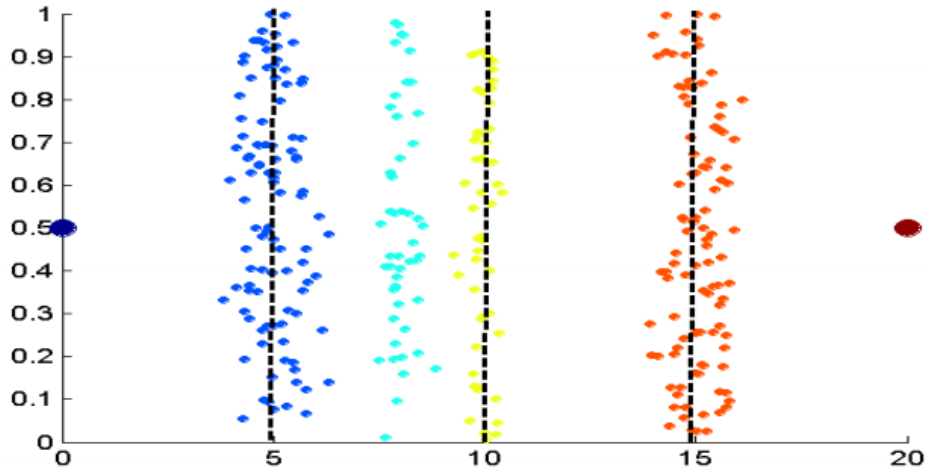
01/27/2020 88 数据挖掘导论，第二版

谭、斯坦贝克、卡帕特内、库马尔
离散化:虹膜示例…
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
　示例:花瓣长度
监督离散化:使用类标签来查找断点
0 2 4 6 8
10
20
30
40
50
花瓣长度
计数

Discretization Without Using Class Labels

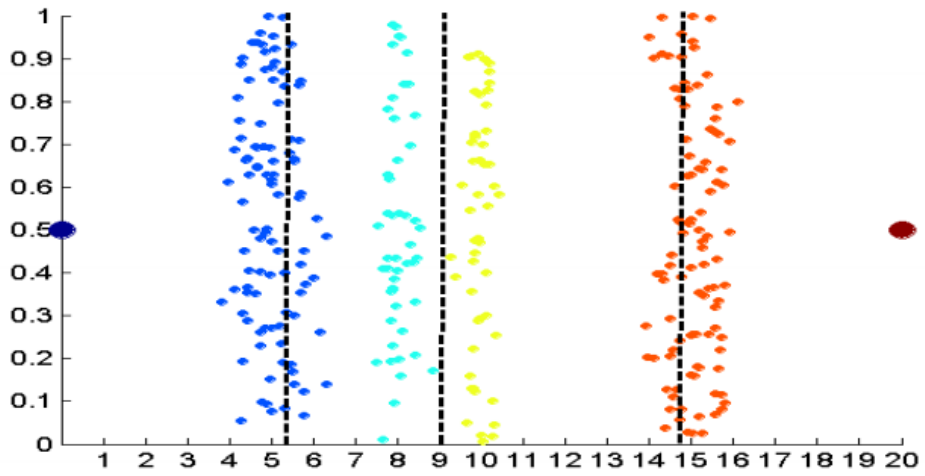


Discretization Without Using Class Labels



Equal interval width approach used to obtain 4 values.

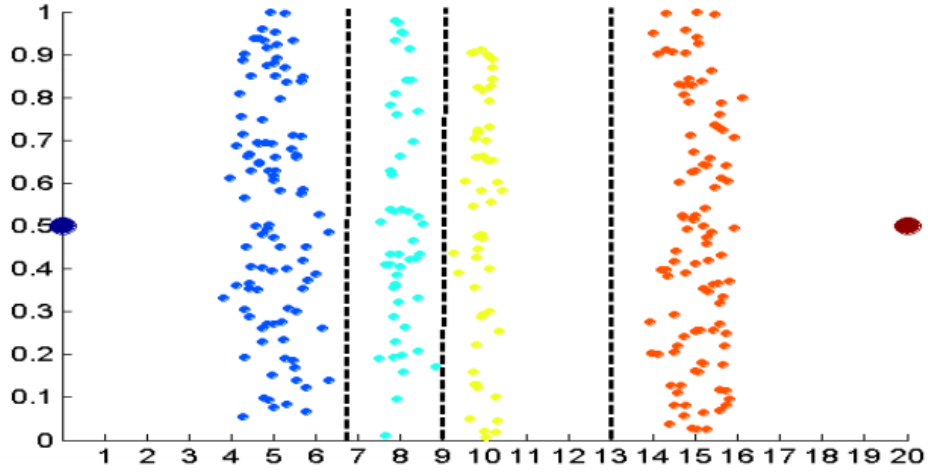
Discretization Without Using Class Labels



Equal frequency approach used to obtain 4 values.

91

Discretization Without Using Class Labels



K-means approach to obtain 4 values.

92

谭、斯坦贝克、卡帕特内、库马尔
二值化
二进制化将连续或分类属性映射到一个或多个二进制变量中
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进制属性
关联分析需要不对称的二元属性
示例:眼睛颜色和高度测量为{低、中、高}

01/27/2020 94 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
属性转换
属性转换是一种将给定属性的整组值映射到一组新的替换值的功能，这样每个旧值都可以用一个新值来标识
简单函数:xk，log(x)，ex，|x|规格化
提到了各种技术来适应不同属性之间在出现频率、平均值、方差、范围
去掉不需要的公共信号，例如季节性
在统计学中，标准化是指减去平均值，再除以标准差

01/27/2020 95 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
示例:植物生长的样本时间序列
时间序列之间的相关性
明尼阿波利斯
明尼阿波利斯亚特兰大圣保罗
明尼阿波利斯 1.0000 0.7591 -0.7581
亚特兰大 0.7591 1.0000 -0.5739
圣保罗-0.7581 -0.5739 1.0000
时间序列之间的相关性
净初级
产量是生态系统科学家用来衡量植物生长的一个指标。

01/27/2020 96 数据挖掘导论，第二版
谭、斯坦贝克、卡帕特内、库马尔
季节性是相关性的主要原因
时间序列之间的相关性
明尼阿波利斯
使用月度 Z 值进行标准化:减去月度平均值，除以月度标准偏差
明尼阿波利斯亚特兰大圣保罗
明尼阿波利斯 1.0000 0.0492 0.0906
亚特兰大 0.0492 1.0000 -0.0154
圣保罗 0.0906 -0.0154 1.0000
时间序列之间的相关性