数据挖掘
第2章 数据

教师：王东京
学院：计算机学院
邮箱：dongjing.wang@hdu.edu.cn

本章内容

属性和对象 Attributes and Objects

数据类型 Types of Data

数据质量 Data Quality

相似度和距离 Similarity and Distance

数据预处理 Data Preprocessing

什么是数据?

数据（集）是**数据对象（Object）**及其**属性（Attributes）**的合集

属性是对象的性质或者特性

- Examples: eye color of a person, temperature, etc.
- 属性有时也叫做变量、特性、字段、特征或维。

属性的集合能够描述/刻画一个对象

- 数据对象有时也叫做记录、点、向量、模式、事件案例、样本、观测或实体。

Attributes

Objects

Tid	Refund 退款	Marital Status 婚姻状态	Taxable Income 应纳税收入	Cheat 是否存在欺诈
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

数据的更完整的视图

数据可能有多部分组成

属性（对象）可能与其他属性（对象）有关系

更一般而言，数据可能具有结构

数据可能不完整

稍后我们将详细讨论

2.1.1 属性值及其度量

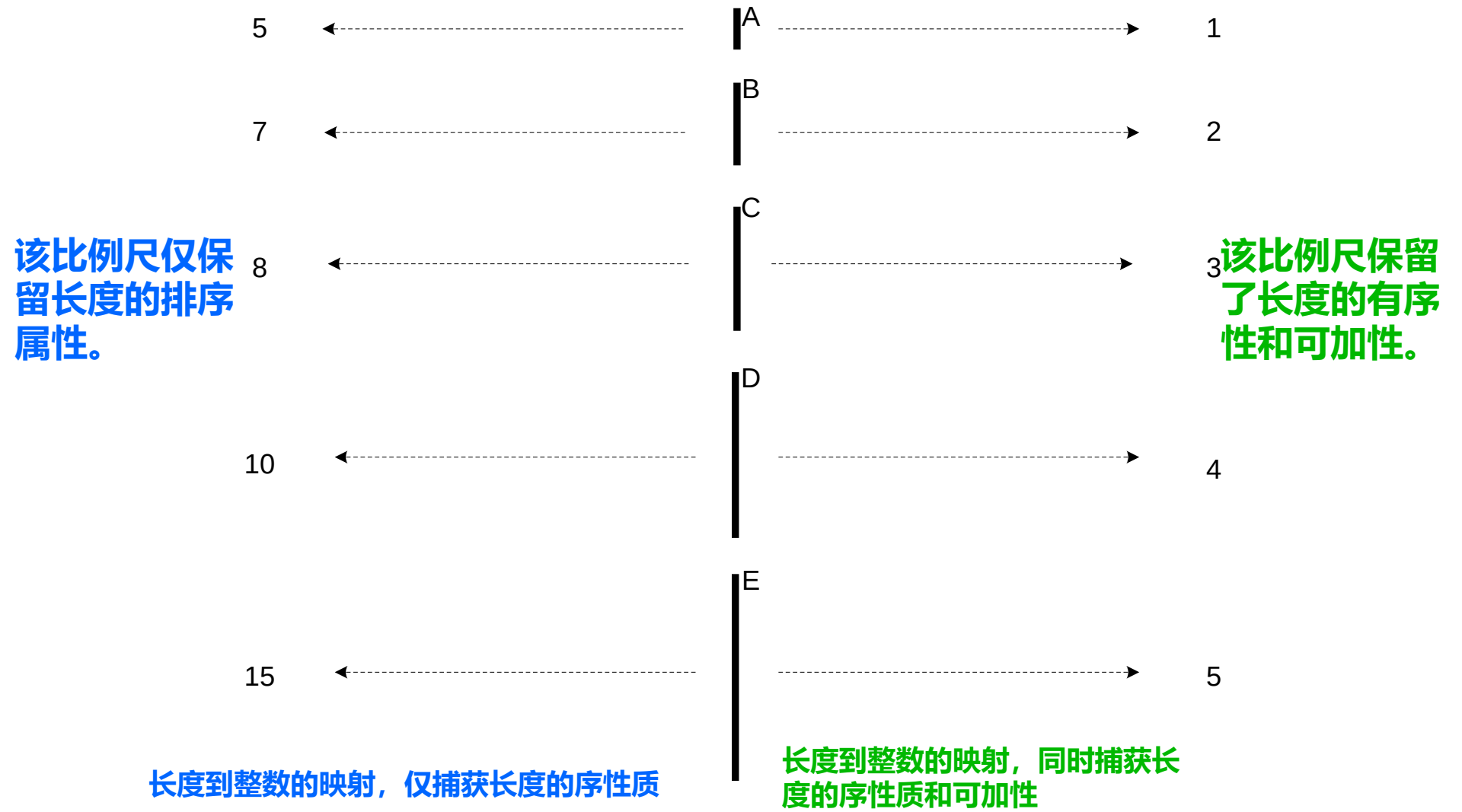
属性值（Attribute Values）是分配给特定对象的属性的数字或符号

属性和属性值之间的区别

- 相同的属性可以映射到不同的属性值
  - 示例：高度可以厘米或米为单位
- 可以将不同的属性映射到同一组值
  - 示例：ID和Age的属性值是整数
  - **但是属性值的属性可以不同**

长度的测量

测量属性的方式可能与属性属性不匹配。



该比例尺仅保留长度的排序属性。

该比例尺保留了长度的有序性和可加性。

长度到整数的映射，仅捕获长度的序性质

长度到整数的映射，同时捕获长度的序性质和可加性

属性类型

类型的属性多种多样

- **标称**（Nominal）
  - Examples: ID numbers, eye color, zip codes
- **序数**（Ordinal）
  - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}
- **区间**（Interval）
  - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
- **比率**（Ratio）
  - Examples: temperature in Kelvin, length, counts, elapsed time (e.g., time to run a race)

属性值的特性（Properties of Attribute Values）

属性的类型取决于它拥有以下哪些特性/操作:

- 相异性（distinctness）： = ≠
- 序（order）： < >
- 加减法（差值有意义） + -
  - meaningful differences
- 乘法（比率有意义） * /
  - Meaningful ratio

属性值的特性（Properties of Attribute Values）

属性的类型取决于它拥有以下哪些特性/操作:

- 相异性（distinctness）： = ≠
- 序（order）： < >
- 加减法（差值有意义） + -
  - meaningful differences
- 乘法（比率有意义） * /
  - Meaningful ratio
- **标称**（Nominal）属性：distinctness
- **序数**（Ordinal）属性：distinctness & order
- **区间**（Interval）属性：distinctness, order & meaningful differences
- **比率**（Ratio）属性: all 4 properties/operations

比率（Ratio）与区间（Interval）的区别

考虑测量身高值

- 如果小明的身高是2.29米，而小红的身高是1.145米，那么我们能说小明的身高是小红的两倍吗?

考虑测量高于平均水平的高度

- 如果小明的身高比平均水平高10厘米，而小红的身高比平均水平高20厘米，那么我们能说小红的身高是比小明的两倍吗?

不同的属性类型

属性类型		描述	例子	操作
分类的（定性的）	标称	其属性值只提供足够的信息以区分对象。这种属性值没有实际意义。	颜色、性别、产品编号	众数、熵、列联相关。
	序数	其属性值提供足够的信息以区分对象的序。	成绩等级（优、良、中、及格、不及格）、年级（一年级、二年级、三年级、四年级）	中值、百分位、秩相关、符号检验。
数值的（定量的）	区间	其属性值之间的差是有意义的。	日历日期、摄氏温度	均值、标准差、皮尔逊相关
	比率	其属性值之间的差和比率都是有意义的。	长度、时间和速度	几何平均、调和平均、百分比变差

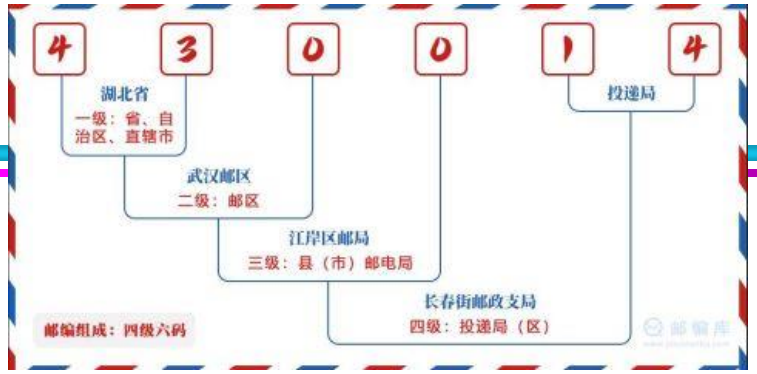
每种属性类型拥有其上方属性类型上的所有性质和操作。

表 2-3　定义属性层次的变换

属性类型		变　换	注　释
分类的（定性的）	标称	任何一对一变换，例如值的一个排列	如果所有雇员的 ID 号都重新赋值，不会出现任何不同
	序数	值的保序变换，即 新值=f(旧值), 其中 f 是单调函数	包括好、较好、最好的属性可以完全等价地用值{1, 2, 3}或用{0.5, 1, 10}表示
数值的（定量的）	区间	新值 =a*旧值+b, 其中 a、b 是常数	华氏和摄氏温度的零度的位置不同，1度的大小（即单位长度）也不同
	比率	新值=a*旧值	长度可以用米或英尺度量

允许的变换（不改变属性意义）

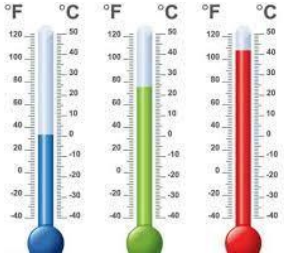
离散和连续属性



离散（Discrete）属性

- 只有一组有限或无数个值
- 示例：邮政编码，一组文档中的单词集或单词数
- 通常表示为整数变量。
- 注意：二进制属性是离散属性的特例

连续（Continuous）属性



- 将实数作为属性值
- 例如：温度、高度或重量。
- 实际上，只能使用有限数量的数字来测量和表示实际值。
- 连续属性通常表示为浮点变量。

不对称属性 Asymmetric Attributes

仅存在（presence）（非零属性值）被视为重要/有意义

- 文件中的字词
- 客户交易中存在的商品

如果我们在商店遇到一个朋友，我们会说以下话吗?

“我发现我们的购物非常相似，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进制属性来表示一个普通的二进制属性

- 关联分析使用非对称属性

非对称属性通常由集合中对象引起

属性类型的关键信息

用户选择的操作类型对于其拥有的数据类型应该是“有意义的”

- 相异性、序、加法（有意义的差值）、乘法（有意义的比率）只是数据的属性中的四个
- 用户所见到的数据类型（通常是数字或字符串）可能无法捕获所有属性，也可能会暗示不存在的属性
- 分析可能取决于数据的这些其他属性
  - 许多统计分析仅取决于分布
- 很多时候“什么是有意义的”是通过统计显着性来衡量的
- 但是最后，有意义与否的是由应用领域衡量的

单选题 1分

设置

说100°（度）的温度是50°的温度的两倍在物理上有意义吗？如果有，那是在什么情况下？

A 有，摄氏温度 ℃

B 有，华氏温度 ℉

C 有，绝对温度 K

D 没有意义

提交

2.1.2 数据集的类型 Types of data sets

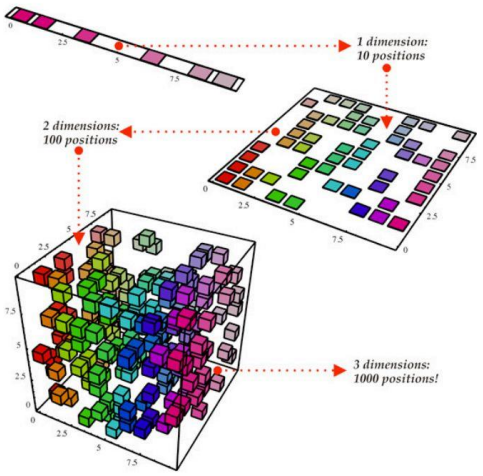
记录数据 Record
- Data Matrix
- Document Data
- Transaction Data

图形数据 Graph
- World Wide Web
- Molecular Structures

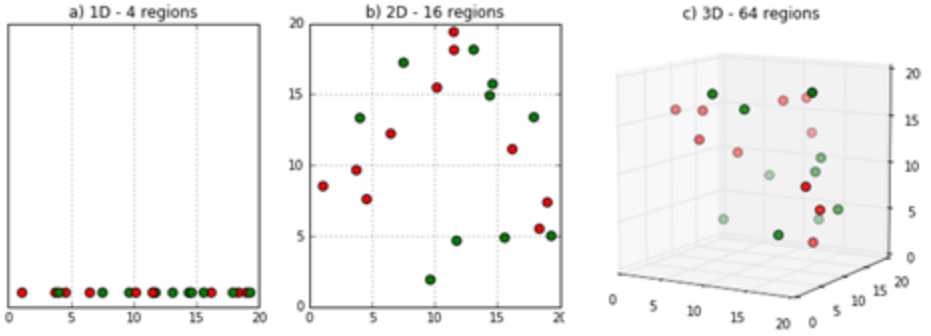
有序数据 Ordered
- Spatial Data
- Temporal Data
- Sequential Data
- Genetic Sequence Data

数据集的重要特性

- 维度 Dimensionality（属性数量）
  - 高维数据带来许多挑战

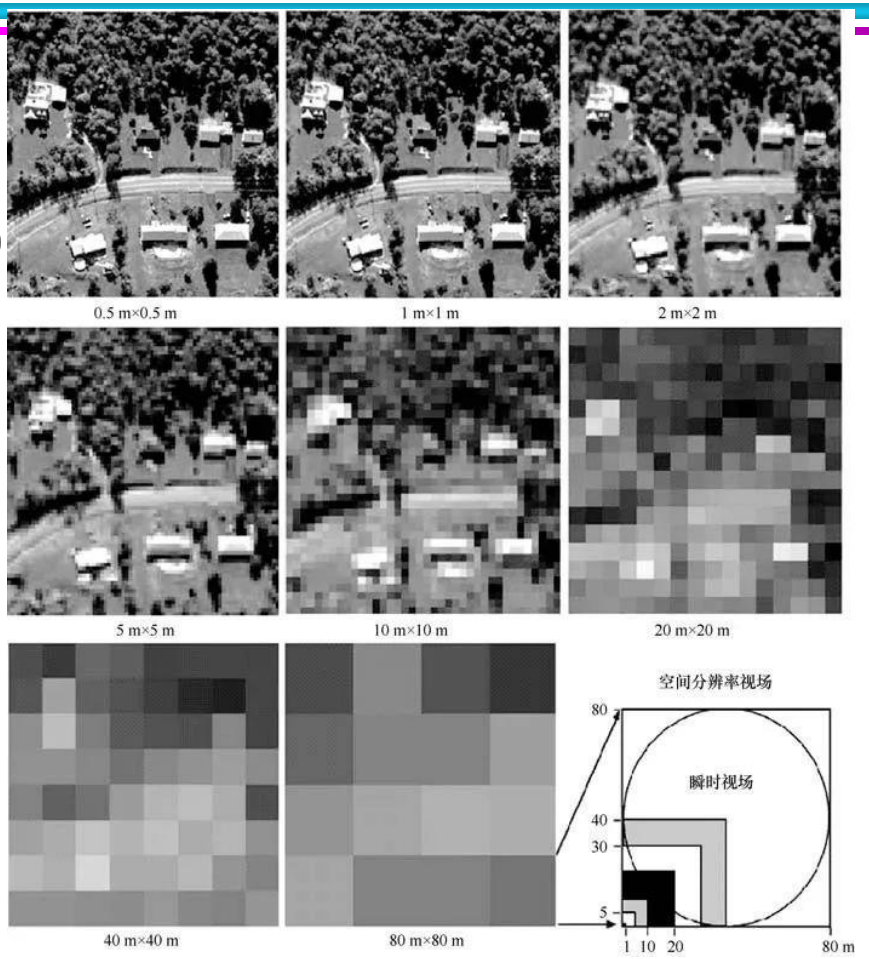


- 稀疏性 Sparsity
  - 只有存在（presence）才重要



数据集的重要特性

- 解析度（Resolution）
  - 模式取决于规模（scale）



- 大小
  - 分析类型可能取决于数据大小

记录数据Record Data

包含记录集合的数据，每个记录包含一组固定的属性

Tid	Refund 退款	Marital Status 婚姻状态	Taxable Income 应纳税收入	Cheat 是否存在欺诈
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

数据矩阵 Data Matrix

如果数据对象具有相同的固定数值属性集，则可以将数据对象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* x *n*矩阵表示

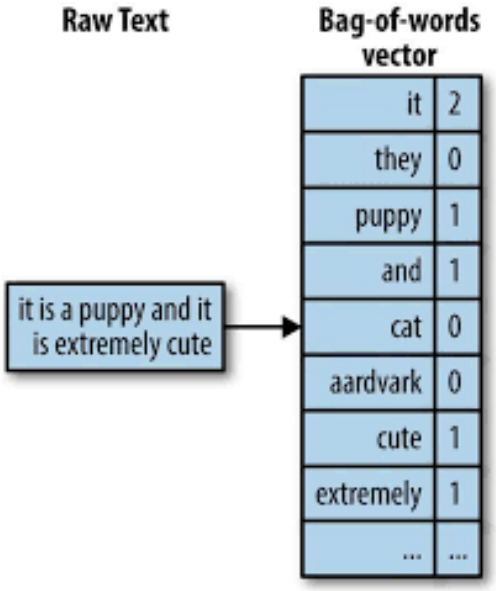
- 其中有 *m* 行，每个对象一个
- *n* 列，每个属性一个。

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

文档数据 Document Data

每个文档都成为一个“术语（term）”向量

- 每个term都是向量的组成部分（属性）
- 每个组件的值是相应术语在文档中出现的次数。



	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

事务数据 Transaction Data

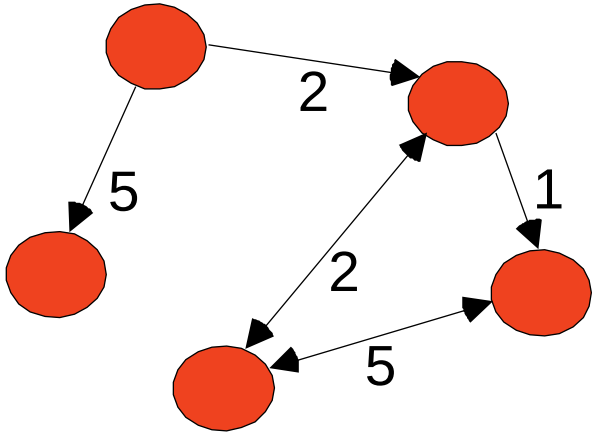
一种特殊类型的数据，其中

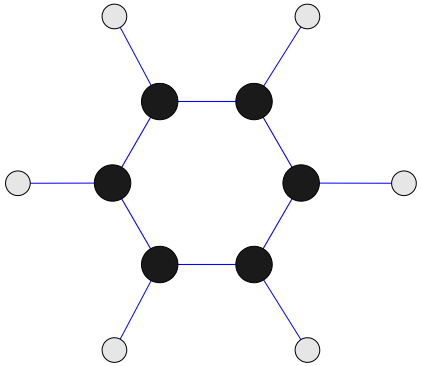
- 每个交易都涉及一组项目（item）。
- 例如，考虑一家杂货店。 客户在一次购物中购买的一组产品构成一项事务（交易），而购买的单个产品就是这些项目。
- 可以将交易数据表示为记录数据（record data）

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

图数据 Graph Data

示例：通用图，分子结构图和网页链接图





苯分子

Benzene Molecule: C6H6

Useful Links:

- Bibliography
- Other Useful Web sites
  - ACM SIGKDD
  - KDnuggets
  - The Data Mine

Knowledge Discovery and Data Mining Bibliography

(Gets updated frequently, so visit often!)

- Books
- General Data Mining

Book References in Data Mining and Knowledge Discovery

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.
Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support), John Wiley & Sons, 1997.

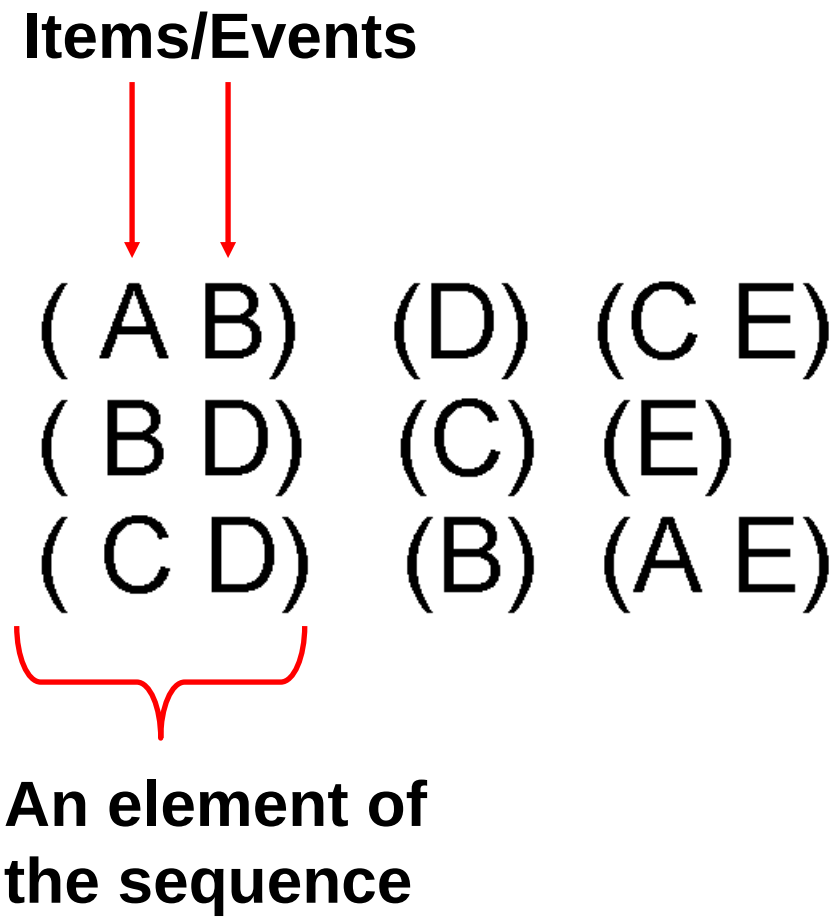
General Data Mining

Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on data Engineering, vol. 21, no. 1, March 1998.

Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

有序数据 Ordered Data

事务（transactions）序列数据



有序数据

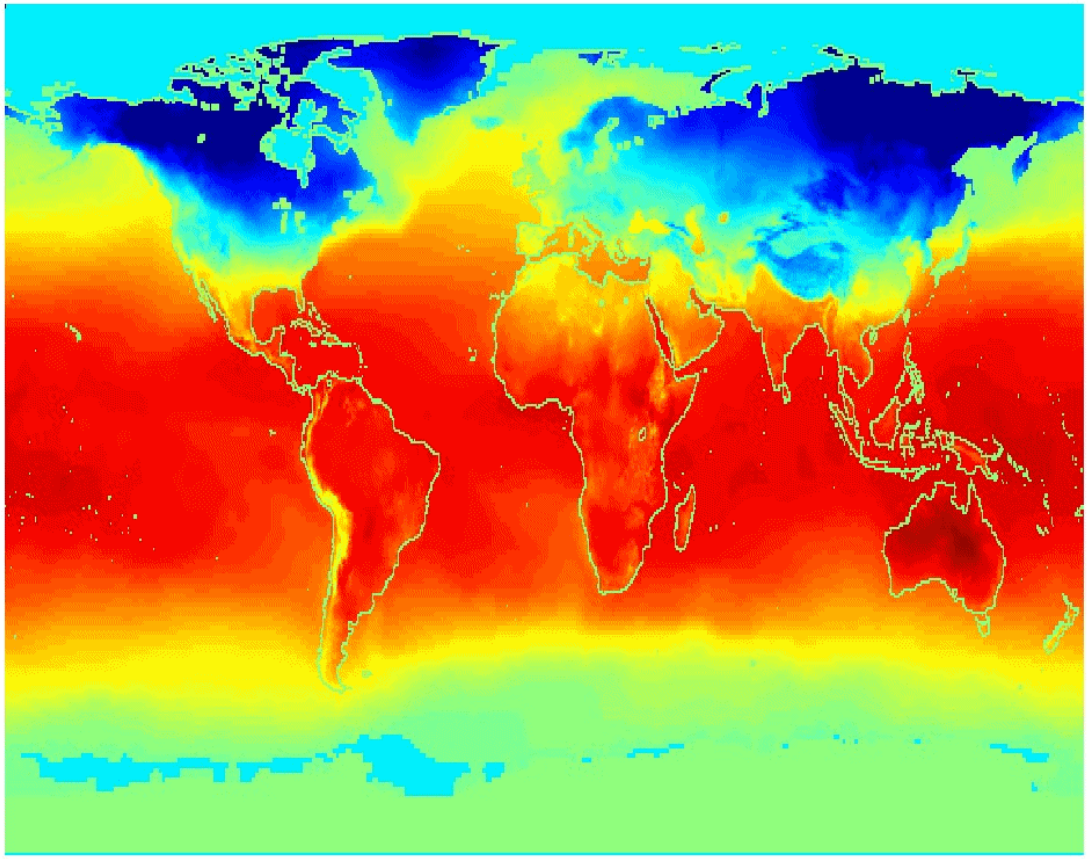
基因组序列数据

GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG

有序数据

时空数据 Spatio-Temporal Data

陆地与大洋的温度变化（月平均值）



单选题 1分

设置

化学分子中各个原子之间的关系用哪种数据表示最合适?

A 矩阵数据 Matrix

B 有序数据 Ordered Data

C 图结构数据 Graph

D 事务数据 Transaction Data

提交

数据质量 Data Quality

低质量数据会对许多数据处理工作产生负面影响

“最重要的一点是，糟糕的数据质量是一场不断发展的灾难。

- 低质量数据使很多公司损失了至少10~20％的收入。”

Thomas C. Redman, DM Review, August 2004

数据挖掘示例：使用低质量数据建立用于检测存在贷款风险的人的分类模型

- 一些信誉良好的候选人被拒绝贷款
- 向容易违约的个人提供了更多贷款

数据质量 …

数据质量问题有哪些?

如何发现数据问题?

如何解决这些问题?

数据质量问题的示例:

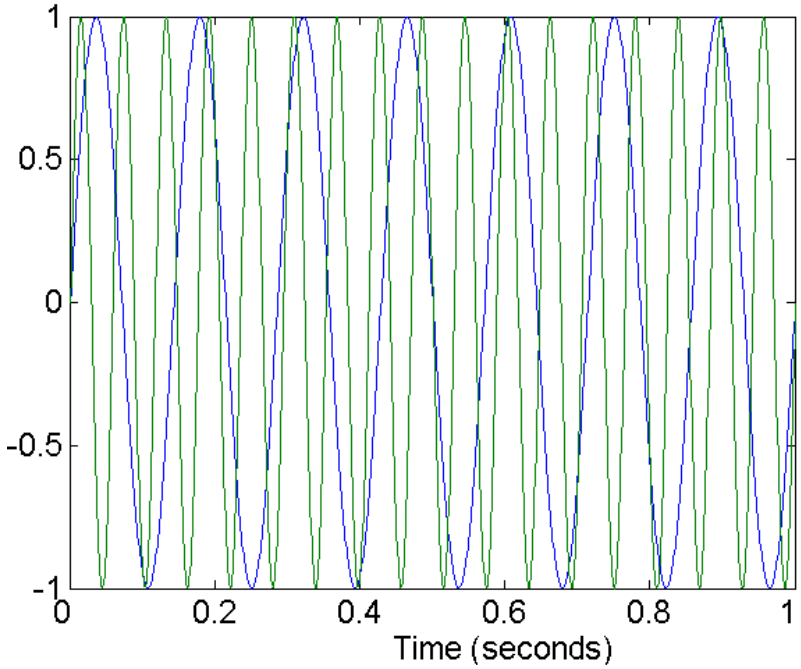
- 噪声和离群值
- 缺失值
- 数据重复
- 数据错误
- 伪数据

噪声 Noise

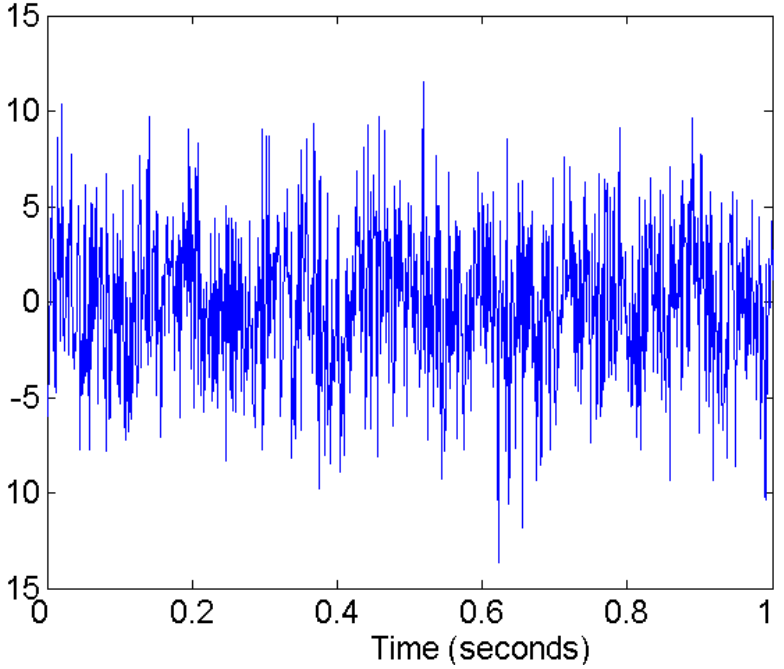
对于对象来说，噪声是多余的/无用的对象

对于属性，噪声是指原始值被修改

- 例如：通话不畅时人的声音失真，电视屏幕上出现“雪花”



Two Sine Waves



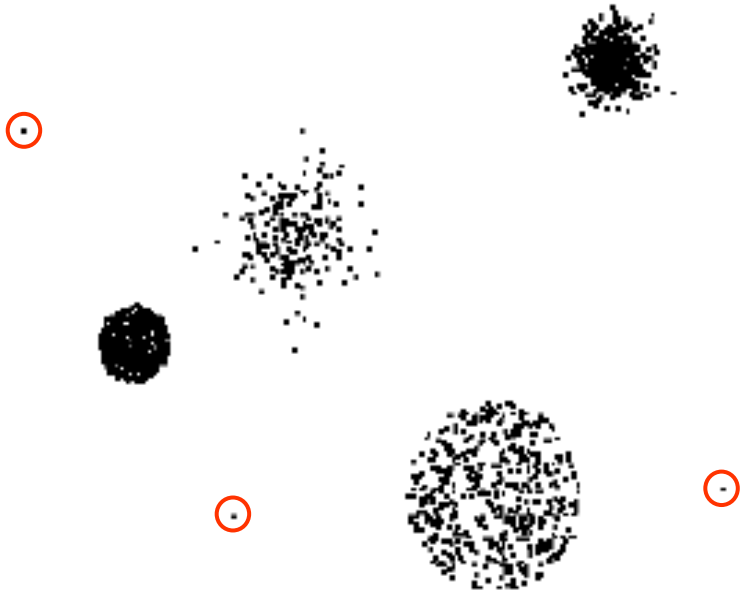
Two Sine Waves + Noise

离群点/异常点 Outliers

离群值是具有与数据集中大多数其他数据对象明显不同的特征的数据对象

- **情况1：**异常值是干扰数据分析的噪声
- **情况2：**离群值有时是我们分析的目标
  - Credit card fraud
  - Intrusion detection

原因?



缺失值 Missing Values

缺失值的原因

- 没有收集信息（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
  - 示例：温度的时间序列
  - 示例：结果
- 在分析过程中人口普查忽略缺失值

缺失值 Missing Values ...

完全随机遗失（Missing completely at random, MCAR）
- 值的缺失与属性无关
- 根据该属性填充/补充值
- 整体分析可能不会产生偏差（bias）

随机丢失（Missing at Random, MAR）
- 数据缺失与其他变量有关
- 根据其它值填充/补充值
- 大部分情况下，对数据的整体分析会产生偏差

非随机丢失（Missing Not at Random, MNAR）
- 数据缺失与未观察到的测量有关
- 导致有价值信息或不可忽视数据的缺失

无法从数据中了解数据缺失情况

重复数据 Duplicate Data

数据集可能包含重复的数据对象，或几乎互相重复的数据对象

- 合并来自异构源的数据时的主要问题

例子：

- 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

什么时候不应该删除重复数据?

2.2 相似度和相异度度量

相似度度量 Similarity measure

- 两个数据对象的相似程度的数值度量。
- 当对象更相似时，该值较高。
- 通常落在[0,1]范围内（非负）

2.2 相似度和相异度度量

相似度度量 Similarity measure

- 两个数据对象的相似程度的数值度量。
- 当对象更相似时，该值较高。
- 通常落在[0,1]范围内（非负）

相异度/差异度量 Dissimilarity measure

- 两个数据对象有多不同的数值度量
- 当物体更相似时降低
- 最小差异度通常为0
- 上限有所不同

邻近度（Proximity）是指相似或不相似

简单属性的相似度和差异度

下表显示了关于单个简单属性上的两个对象 x 和 y 之间的相似性和不相似性。

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

欧几里得距离 Euclidean Distance

Euclidean Distance

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

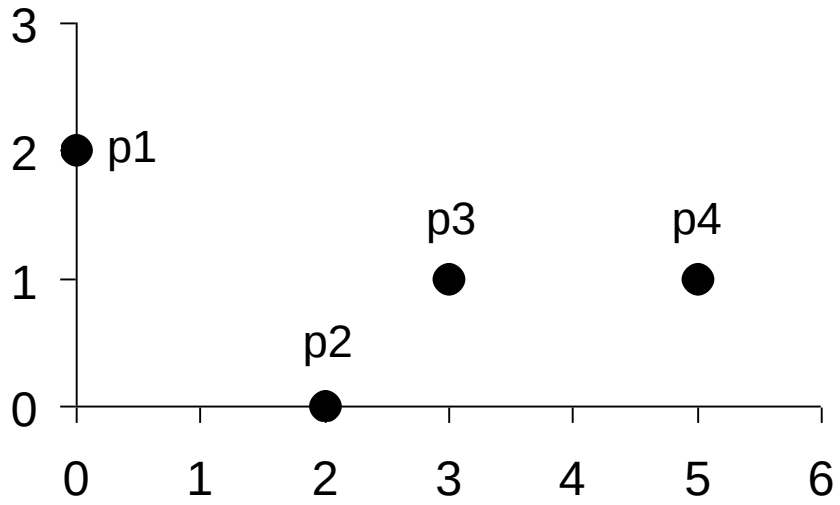
where *n* is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects $\mathbf{x}$ and $\mathbf{y}$.

其中*n* 是维数（属性）的数量， x_k 和 y_k 分别是数据对象 $\mathbf{x}$ 和 $\mathbf{y}$ 的第 *k* 个属性（分量）。

如果尺度(scales)不同，则需要标准化(Standardization)

Euclidean Distance

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



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

距离矩阵 Distance Matrix

闵可夫斯基距离 Minkowski Distance

欧氏距离的泛化与推广（generalization）

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

其中 r 是参数，n 是维数（属性），x_k 和 y_k 分别是数据对象 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的第 k 个属性（分量）。

闵可夫斯基距离示例

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

r = 1. 城市街区(曼哈顿, 出租车, L_1 范数norm) 距离。

- 二进制向量的一个常见示例是汉明距离，该距离是两个二进制向量之间不同的位数

r = 2. Euclidean distance

$r \to \infty$. “上确界supremum” (L_{max}范数, L_∞范数) 距离.

- 对象属性之间的最大距离（This is the maximum difference between any component of the vectors）

不要混淆*r* 和 *n*，所有的这些距离都是对*n*的所有值定义的。

Minkowski Distance

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

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

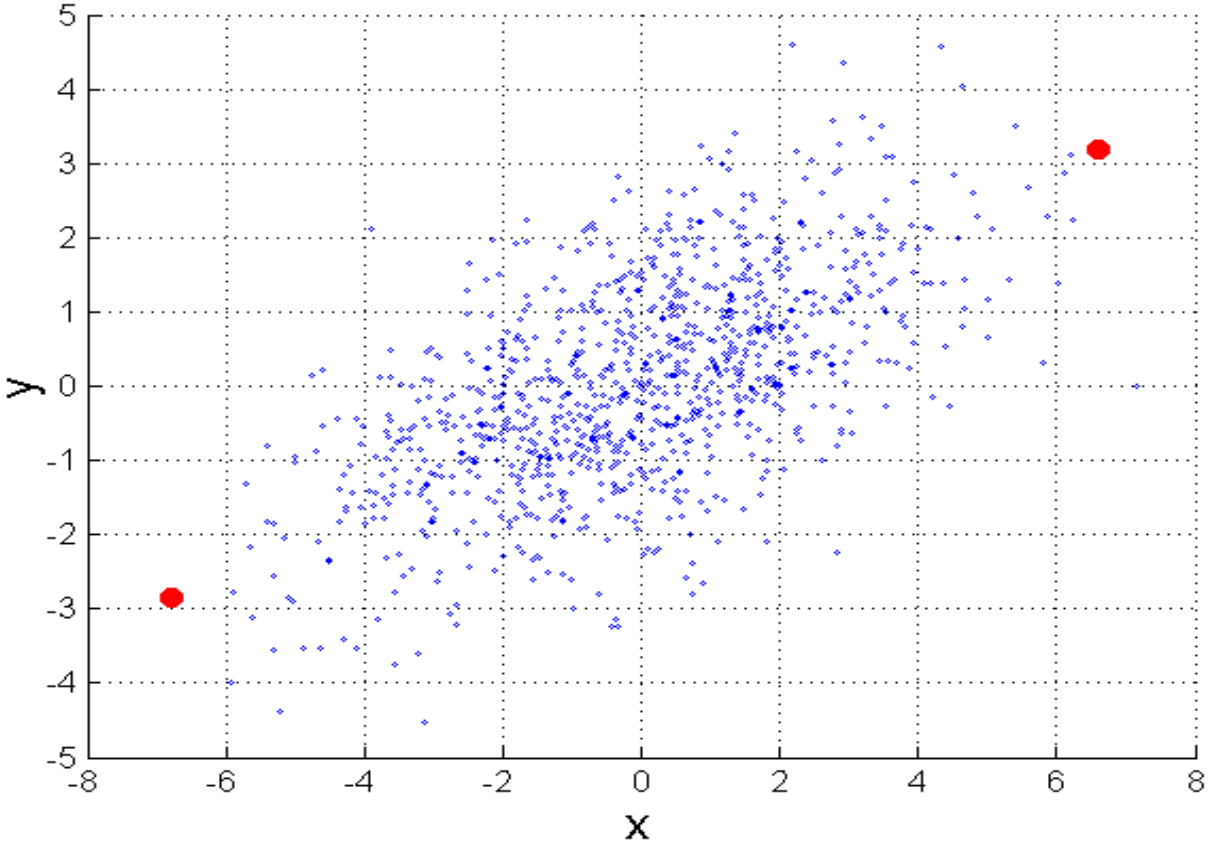
L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

距离矩阵

马氏距离 Mahalanobis Distance

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

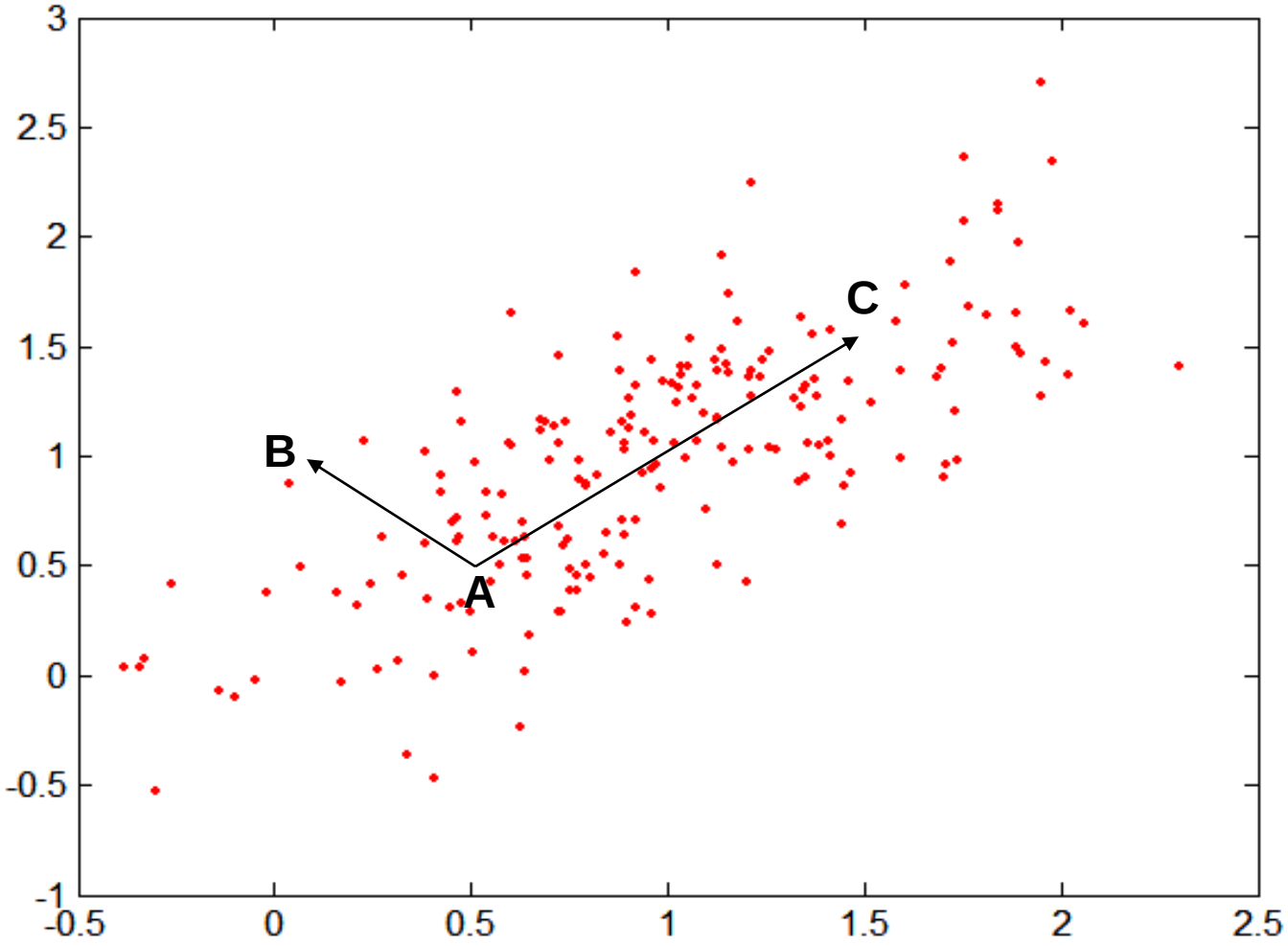


Σ 是协方差矩阵（covariance matrix）

红点之间的欧氏距离是 14.7,

马氏距离是 6.

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

距离的公共属性

Distances, such as the Euclidean distance, have some well known properties.

1. $d(\mathbf{x}, \mathbf{y}) \geq 0$ for all x and y and $d(\mathbf{x}, \mathbf{y}) = 0$ only if $\mathbf{x} = \mathbf{y}$. (非负性 Positive definiteness)
2. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$. (对称性 Symmetry)
3. $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ for all points $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$. (三角不等式 Triangle Inequality)

where $d(\mathbf{x}, \mathbf{y})$ is the distance (dissimilarity) between points (data objects), $\mathbf{x}$ and $\mathbf{y}$.

A distance that satisfies these properties is a metric（度量）

反例：集合的相似度度量

相似度的公共属性

相似度 Similarities, also have some well known properties.

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

where $s(\mathbf{x}, \mathbf{y})$ is the similarity between points (data objects), $\mathbf{x}$ and $\mathbf{y}$.

二元数据的相似性度量

$\mathbf{x}$ 和 $\mathbf{y}$ 是 2 个对象，都由 *n* 个二元属性（binary attributes）组成。

可按照如下四个部分比较其相似度

f_{01} = $\mathbf{x}$ was 0 and $\mathbf{y}$ was 1 的属性数量

f_{10} = $\mathbf{x}$ was 1 and $\mathbf{y}$ was 0 的属性数量

f_{00} = $\mathbf{x}$ was 0 and $\mathbf{y}$ was 0 的属性数量

f_{11} = $\mathbf{x}$ was 1 and $\mathbf{y}$ was 1 的属性数量

简单匹配系数 Simple Matching Coefficients (SMC) 和 杰卡德系数 Jaccard Coefficients

SMC = number of matches / number of attributes

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

J = number of 11 matches / number of non-zero attributes

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

SMC versus Jaccard: Example

x = 1 0 0 0 0 0 0 0 0 0

y = 0 0 0 0 0 0 1 0 0 1

f_{01} = 2 (the number of attributes where **x** was 0 and **y** was 1)

f_{10} = 1 (the number of attributes where **x** was 1 and **y** was 0)

f_{00} = 7 (the number of attributes where **x** was 0 and **y** was 0)

f_{11} = 0 (the number of attributes where **x** was 1 and **y** was 1)

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

= (0+7) / (2+1+0+7) = 0.7

J = $(f_{11}) / (f_{01} + f_{10} + f_{11})$ = 0 / (2 + 1 + 0) = 0

余弦相似度 Cosine Similarity

如果 $\mathbf{d}_1$ 和 $\mathbf{d}_2$ 是两个文档向量，那么

$$\cos(\mathbf{d}_1, \mathbf{d}_2) = \langle \mathbf{d}_1, \mathbf{d}_2 \rangle / \|\mathbf{d}_1\| \|\mathbf{d}_2\|,$$

其中 $\langle \mathbf{d}_1, \mathbf{d}_2 \rangle$ 是 $\mathbf{d}_1$ 和 $\mathbf{d}_2$ 的向量内积（inner product）或者向量点乘（vector dot product），$\|\mathbf{d}\|$ 是向量 $\mathbf{d}$ 的长度 。

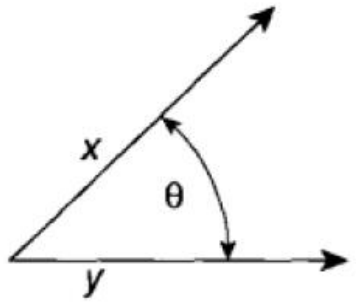


图 2-16　余弦度量的几何解释

示例:

$\mathbf{d}_1 = 3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0$

$\mathbf{d}_2 = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2$

$\langle \mathbf{d}_1, \mathbf{d2} \rangle = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$

$\|\mathbf{d}_1\| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

$\|\mathbf{d}_2\| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.449$

$\cos(\mathbf{d}_1, \mathbf{d}_2) = 0.3150$

广义 Jaccard 系数（Tanimoto系数）

Jaccard 系数的变体，适用于对象的属性为连续（continuous）值或者个数（count）

- 在二元属性（binary）情况下规约Jaccard系数

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

相关性（Correlation）度量对象之间的线性关系

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

where we are using the following standard statistical notation and definitions

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

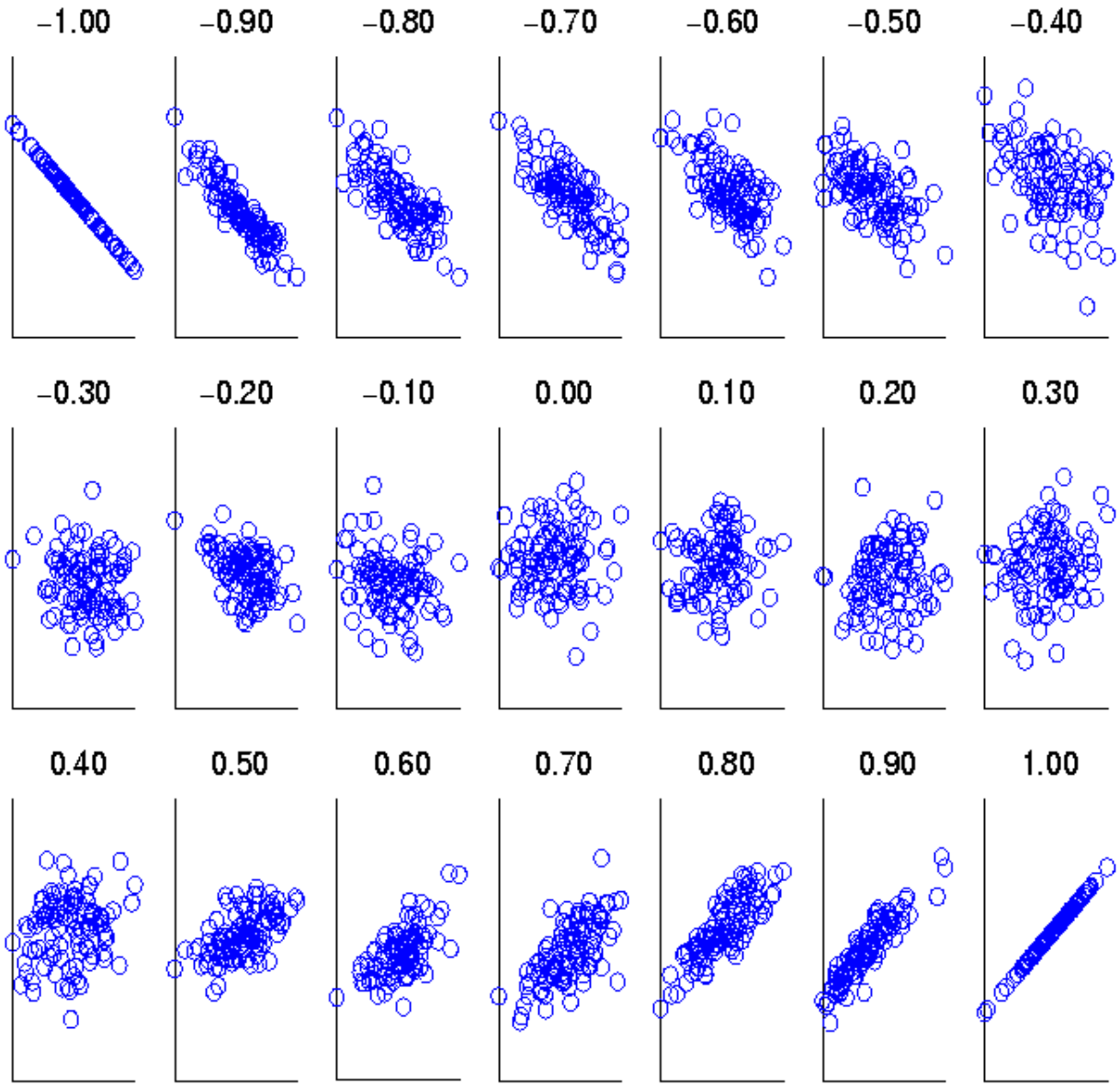
$$\text{standard\_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2}$$

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

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

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

相关度可视化



解释相关度从-1到1的散布图

非线性相关

$\mathbf{x} = (-3, -2, -1, 0, 1, 2, 3)$

$\mathbf{y} = (9, 4, 1, 0, 1, 4, 9)$

$y_i = x_i^2$

$\text{mean}(\mathbf{x}) = 0, \text{mean}(\mathbf{y}) = 4$

$\text{std}(\mathbf{x}) = 2.16, \text{std}(\mathbf{y}) = 3.74$

$\text{corr} = (-3)(5)+(-2)(0)+(-1)(-3)+(0)(-4)+(1)(-3)+(2)(0)+3(5) / (6 * 2.16 * 3.74)$

$= \mathbf{0}$

如果相关度为0，则两个数据对象的属性之间不存在线性关系。然而，仍然可能存在非线性关系。

基于信息（Information）的度量

信息论（Information theory）是一个较为完善且基础的学科，具有广泛应用

部分相似性度量策略的基础是信息论

- 各种版本的互信息（Mutual information）
- 最大信息系数（Maximal Information Coefficient,MIC）和相关度量
- 通用，可以处理非线性（non-linear）关系
- 计算起来可能很复杂且耗时

信息与概率论（Probability）

信息与事件的可能结果有关

- 消息的传输，硬币的翻转或数据段的测量

结果越确定，其包含的信息越少，反之亦然

- 例如，如果硬币有两个正面，那么正面的结果将不提供任何信息
- 从数量上讲，信息与结果的可能性相关
  - 结果的可能性越小，它提供的信息越多，反之亦然
- 熵是常用的量度

熵 Entropy

对于

- 一个变量(事件), X,
- 有 n 种可能的取值 (结果), $x_1, x_2 \ldots, x_n$
- 每种结果有对应的概率, $p_1, p_2 \ldots, p_n$
- X 的熵 $H(X)$ 定义为

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

熵介于0和$\log_2 n$之间，用比特来衡量

- 因此熵是一种衡量“表示X平均需要多少比特”的指标

Entropy Examples

有一个硬币，正面朝上的概率为 p，反面朝上的概率为$q = 1 - p$

$$H = -p \log_2 p - q \log_2 q$$

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

- For p= 0.5, q = 0.5 (fair coin) H = log2=1
- For p = 1 or q = 1, H = 0

单选题 1分

设置

计算一个均匀的筛子（6个面）的熵？和均匀的硬币的熵（2个面）相比哪个更大?

A 筛子

B 硬币

C 一样大

提交

Entropy for Sample Data

假设我们有如下例子：

- 大量（假定 m 个）对某个属性 X 的观察值，比如，班里同学的头发的颜色
- 有 n 种可能的取值
- 第 i 个类别里面的值的数量为 m_i
- 那么，熵为

$$H(X) = -\sum_{i=1}^{n} \frac{m_i}{m} \log_2 \frac{m_i}{m}$$

对于连续的数据，计算过程会更加困难

Entropy for Sample Data: Example

Hair Color	Count	p	$-p\log_2 p$
Black	75	0.75	0.3113
Brown	15	0.15	0.4105
Blond	5	0.05	0.2161
Red	0	0.00	0
Other	5	0.05	0.2161
Total	100	1.0	1.1540

Maximum entropy is

$\log_2 5 = 2.3219$

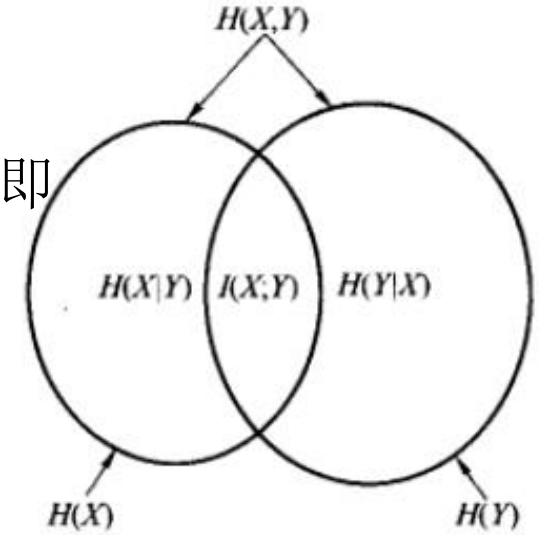
$$H(X) = -\sum_{i=1}^{n} \frac{m_i}{m} \log_2 \frac{m_i}{m}$$

互信息 Mutual Information

一个变量（variable）能够为另外一个变量所提供的的信息，是衡量随机变量之间相互依赖程度的度量

正式定义为 $I(X,Y) = H(X) + H(Y) - H(X,Y)$, 其中 $H(X,Y)$ 是 X and Y 的联合熵（joint entropy）

表示X和Y一起发生的时的信息熵或产生的信息量，即X和Y一起发生时的确定度。



$$H(X,Y) = -\sum_i \sum_j p_{ij} \log_2 p_{ij}$$

其中 p_{ij} 是 X 的第 i 个变量与 Y 的第 j 个变量共同出现的概率。

对于离散变量，很容易计算

互信息

随机变量：X表示今天多云，Y表示明天下雨

互信息：“已知今天多云，明天下雨的不确定性”与“不知道今天是否多云，明天下雨的不确定性”之差

互信息

随机变量：X表示今天多云，Y表示明天下雨

互信息：“已知今天多云，明天下雨的不确定性”与“不知道今天是否多云，明天下雨的不确定性”之差

$$I(X;Y) = H(X) - H(X|Y).$$

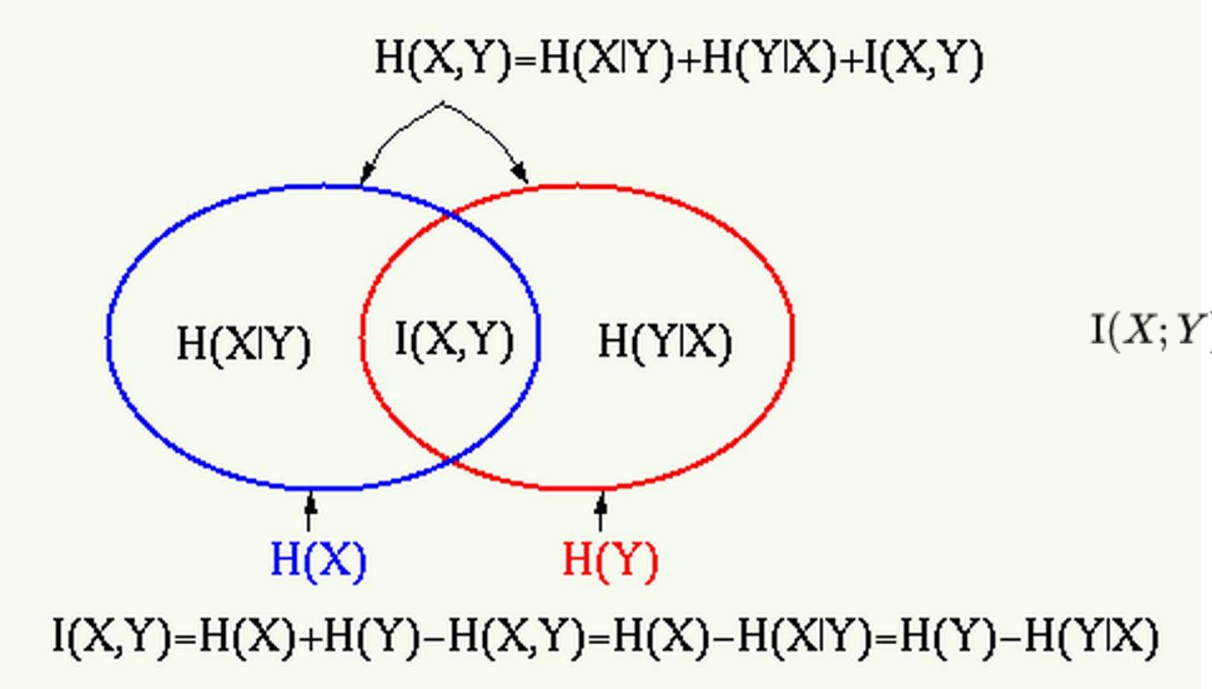
$H(X)$ 是 X 的信息熵， $H(Y|X)$ 是已知 X 情况下，Y带来的信息熵（条件熵）。

从概率角度，互信息是由随机变量 X, Y 的联合概率分布 $p(x,y)$ 和边缘概率分布 $p(x), p(y)$ 得出。

$$\mathrm{I}(X;Y) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p(x,y) \log \left(\frac{p(x,y)}{p(x)\, p(y)} \right),$$

互信息

随机变量：X表示今天多云，Y表示明天下雨



$$
\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X) \\
&= H(X) + H(Y) - H(X,Y) \\
&= H(X,Y) - H(X|Y) - H(Y|X)
\end{aligned}
$$

互信息

$$I(X;X) = H(X) - H(X|X) = H(X)$$

性质

- 1.非负性（证明来自Jeson 不等式)

$$\mathrm{I}(X;Y) \geq 0$$

- 2. 对称性

$$\mathrm{I}(X;Y) = \mathrm{I}(Y;X)$$

- 3. 与条件熵和联合熵的关系

$$\begin{aligned}\mathrm{I}(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(X,Y) - H(X|Y) - H(Y|X)\end{aligned}$$

互信息示例

Student Status	Count	p	$-p\log_2 p$
Undergrad	45	0.45	0.5184
Grad	55	0.55	0.4744
Total	100	1.00	0.9928

Grade	Count	p	$-p\log_2 p$
A	35	0.35	0.5301
B	50	0.50	0.5000
C	15	0.15	0.4105
Total	100	1.00	1.4406

Student Status	Grade	Count	p	$-p\log_2 p$
Undergrad	A	5	0.05	0.2161
Undergrad	B	30	0.30	0.5211
Undergrad	C	10	0.10	0.3322
Grad	A	30	0.30	0.5211
Grad	B	20	0.20	0.4644
Grad	C	5	0.05	0.2161
Total		100	1.00	2.2710

Mutual information of Student Status and Grade = 0.9928 + 1.4406 - 2.2710 = 0.1624

最大信息系数 Maximal Information Coefficient

Reshef, David N., Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher, and Pardis C. Sabeti. "Detecting novel associations in large data sets." *science* 334, no. 6062 (2011): 1518-1524.

将互信息应用于两个连续变量

考虑将变量归类为离散类别（离散化）：

- $n_X \times n_Y \leq N^{0.6}$ where
  - n_X 是 X 离散化之后取值的个数
  - n_Y 是 Y 离散化之后取值的个数
  - N 是样本的个数 (observations, data objects)

计算互信息

- 归一化：Normalized by $\log_2(\min(n_X, n_Y))$

取最高值

结合相似度（similarity）的常用策略

有时属性有许多不同的类型，但是需要整体相似度。

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

结合相似度（similarity）的常用策略

有时属性有许多不同的类型，但是需要整体相似度。

1: 对于第 k^{th} 个属性, 计算相似度 $s_k(\mathbf{x}, \mathbf{y})$, 相似度范围为[0, 1].

2: 对于第k^{th}个属性，定义一个指示变量（indicator），δ_k:

如果第k^{th}个属性是非对称（asymmetric attribute）属性且对应对象的属性值均为0，或者其中一个对象的第k^{th}个属性值缺失，则 $\delta_k = 0$

否则 $\delta_k = 1$

3. 计算

$$\text{similarity}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^{n} \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^{n} \delta_k}$$

不同相似度策略的权重

如果需要区分不同属性的重要性.

- 使用非负权重 ω_k

- $similarity(\mathbf{x},\mathbf{y}) = \frac{\sum_{k=1}^{n} \omega_k \delta_k s_k(\mathbf{x},\mathbf{y})}{\sum_{k=1}^{n} \omega_k \delta_k}$

闵可夫斯基距离可以修改为:

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

密度 Density

测量指定区域中数据对象彼此靠近的程度

密度的概念与邻近度（proximity）的概念紧密相关

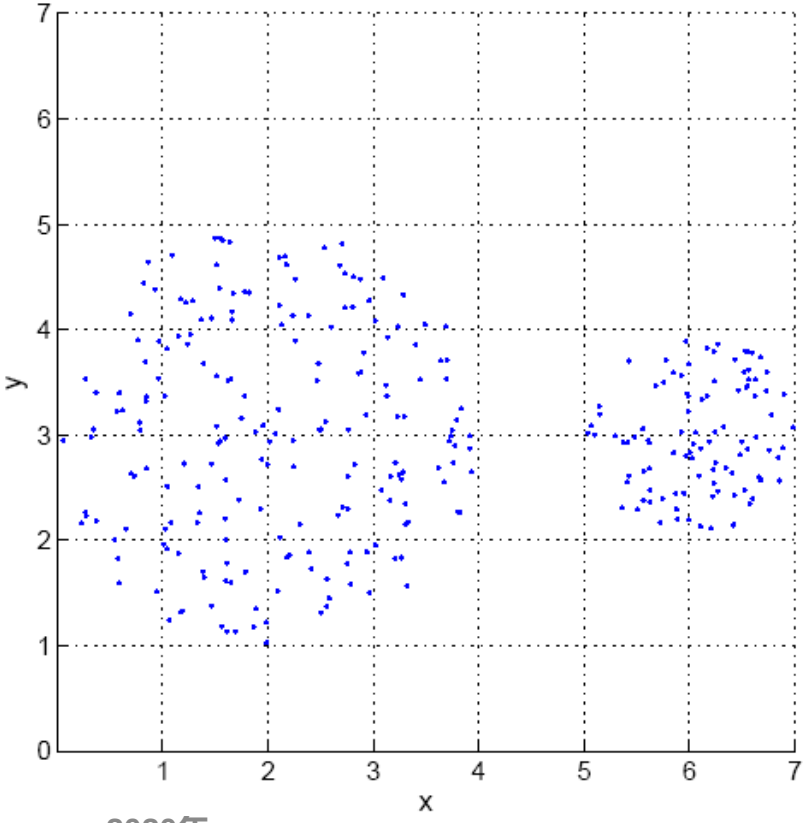
密度的概念通常用于聚类和异常检测

例子:

- 欧氏密度
  - 欧氏密度=每单位体积的点数
- 概率密度
  - 估计数据的分布情况
- 基于图的密度
  - 连接性

欧几里得密度: 基于网格的方法

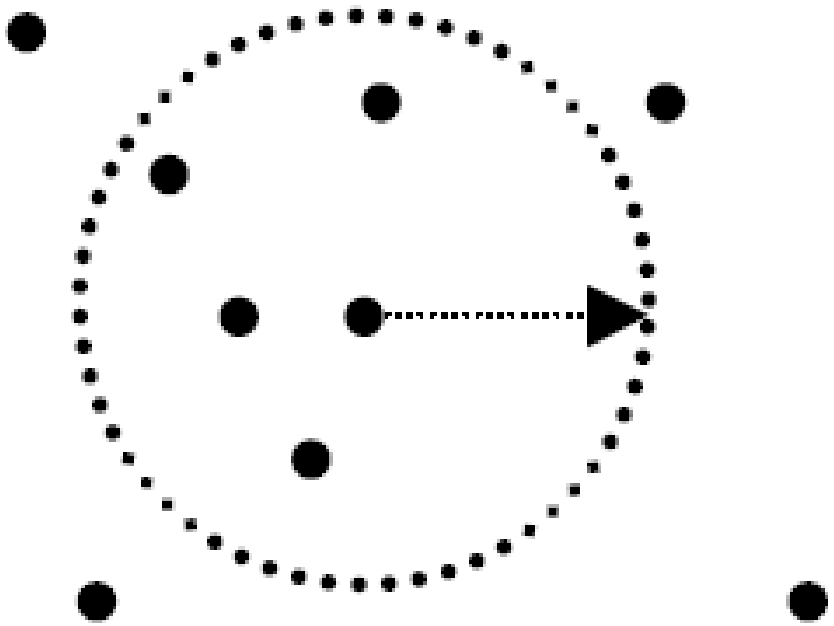
最简单的方法是将区域划分为多个等体积的矩形单元，并将密度定义为该单元包含的点数



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
3	20	14	4	0	0	0
0	0	0	0	0	0	0

欧几里得密度: 基于中心的方法

欧几里得密度是点的指定半径内的点数



基于中心的密度 （center-based density）展示

2.3 数据预处理 Data Preprocessing

聚合 Aggregation

抽样 Sampling

维度规约 Dimensionality Reduction

特征子集选择 Feature subset selection

特征创建 Feature creation

离散化和二元化 Discretization and Binarization

变换 Attribute Transformation

2.3.1 聚合 Aggregation

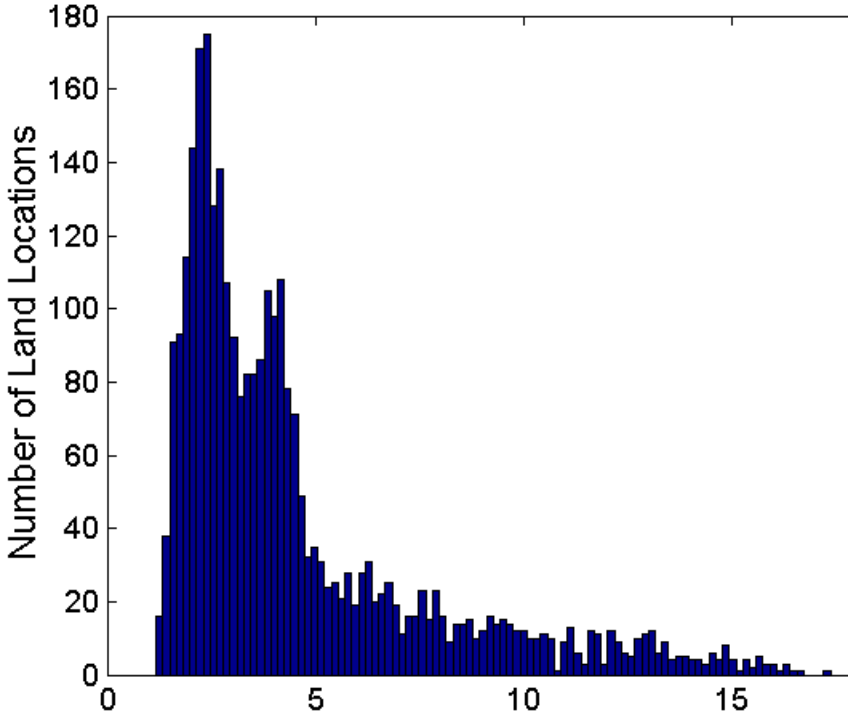
将两个或多个属性（或对象）组合为一个属性（或对象）

目的

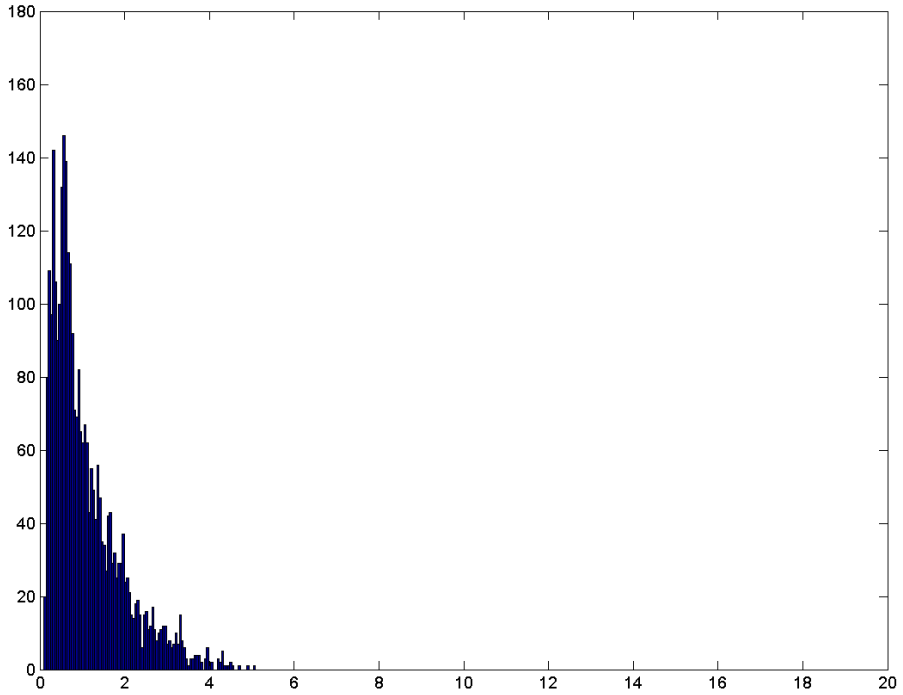
- 数据缩减
  - 减少属性或对象的数量
- 规模变化
  - 城市汇总为地区，州，国家等。
  - 天数汇总成周，月或年
- 更多“稳定”数据
  - 汇总数据往往具有较小的可变性

示例: 澳大利亚的降水量 1982~1993

降水量变化



月平均降水量的标准差



年平均降水量的标准差

2.3.2 抽样 Sampling

采样是用于数据缩减的主要技术。

- 它通常用于数据的初步调查和最终数据分析。

统计人员经常进行采样，因为获取整个相关的数据集的成本和耗时过高。

采样通常用于数据挖掘，因为处理整个数据集成本过高或者过于耗时。

抽样 Sampling …

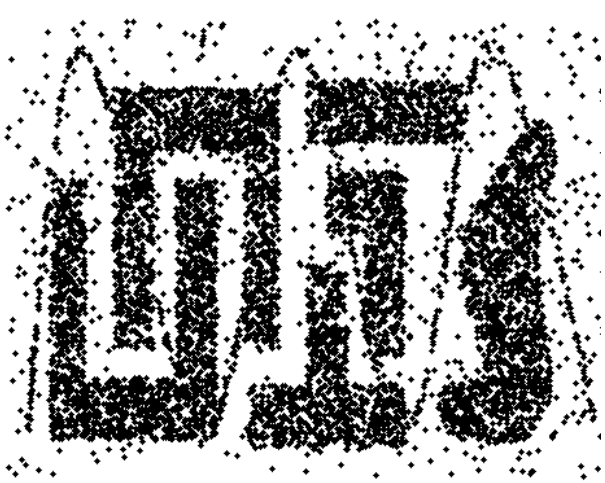
有效抽样的关键原则:

- 如果一个采样样本是有代表性的（representative），那么使用这个样本能够取得和使用原始数据集几乎一样的效果和结果
- 如果一个采样和整个原始数据集的特性近乎相同，那么这个采样是有代表性的（representative）

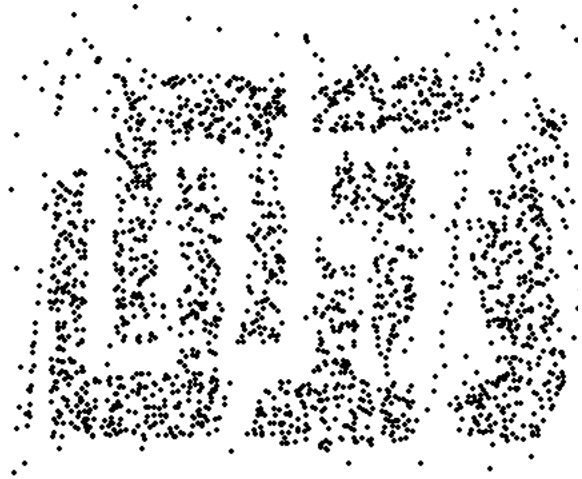
抽样方法

- 简单随机抽样（Simple Random Sampling）
  - 选择任何特定项目的概率是均等的
  - 无放回抽样：Sampling without replacement
    - 选择每个项目后，将其从总体数据集中删除
  - 有放回抽样：Sampling without replacement
    - 在抽到某个项目后，不会将其从总体中删除。
    - 在有放回抽样中，可以多次抽取同一个对象
- 分层抽样（Stratified sampling）
  - 将数据分成几个组； 然后从每个组中随机抽取样本

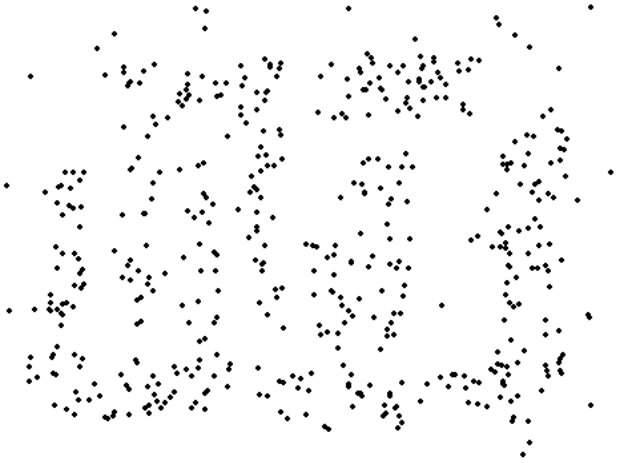
样本容量 Sample Size



8000 points



2000 Points

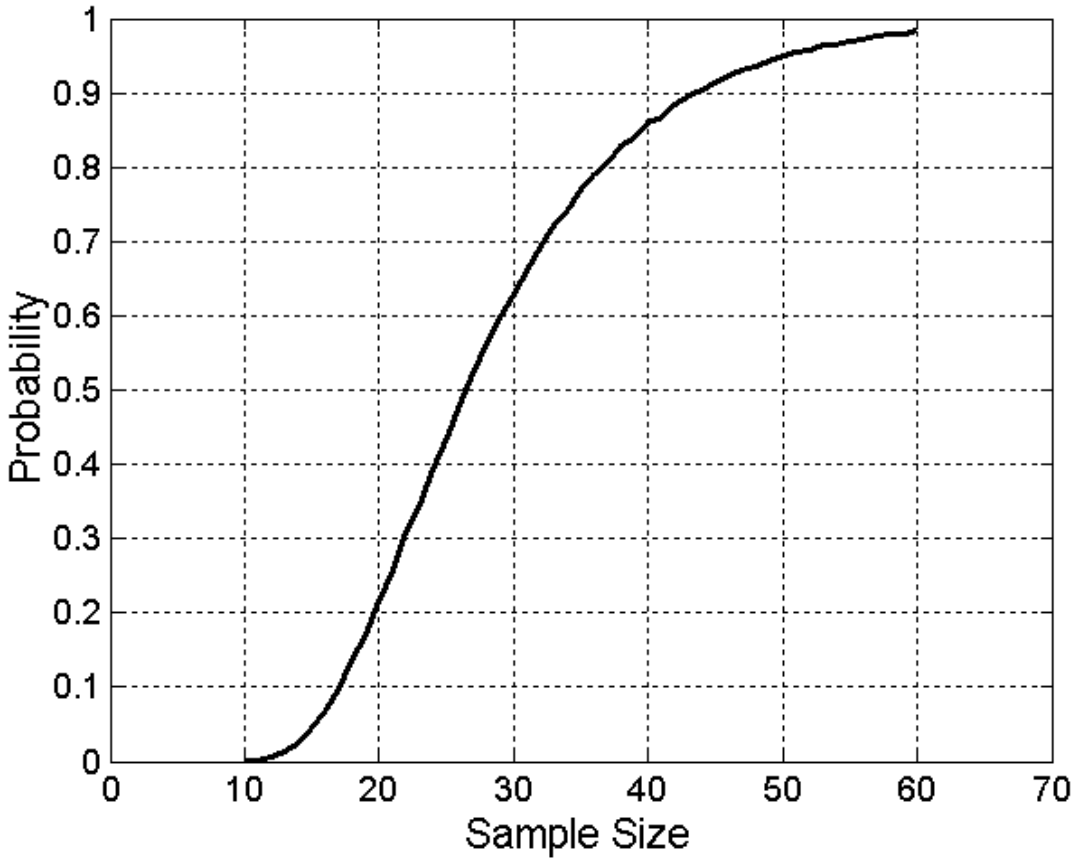


500 Points

样本容量 Sample Size

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

(a) 点的 10 个组



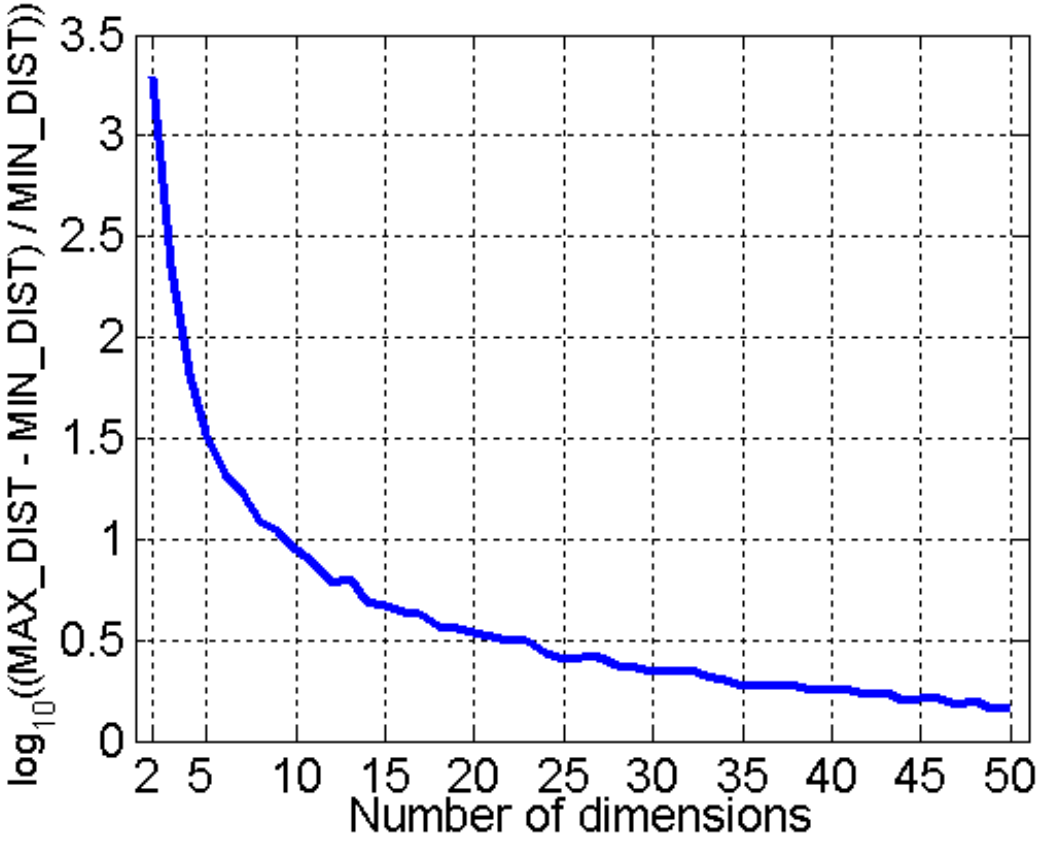
(b) 样本包含所有 10 个组中点的概率

2.3.3维度规约 Dimensionality Reduction

维度灾难 Curse of Dimensionality

当维数增加时，数据在其占用的空间中变得越来越稀疏

对聚类和离群值检测至关重要的点之间的密度和距离的定义变得没有意义



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

维度规约 Dimensionality Reduction

目的：

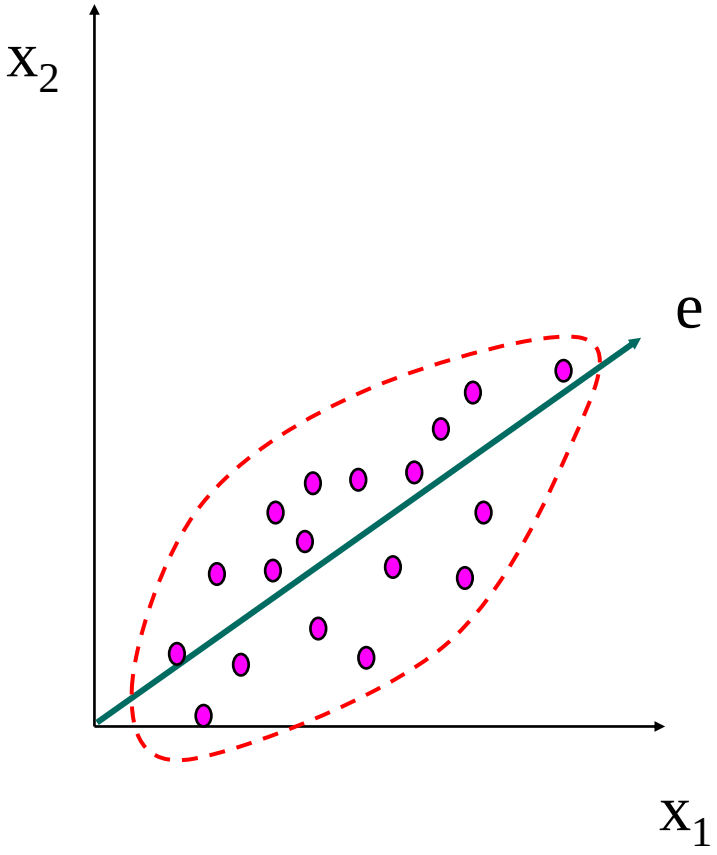
- 避免维度诅咒/灾难
- 减少数据挖掘算法所需的时间和内存
- 使数据更容易可视化
- 可能有助于消除不相关的功能或减少噪音

技术

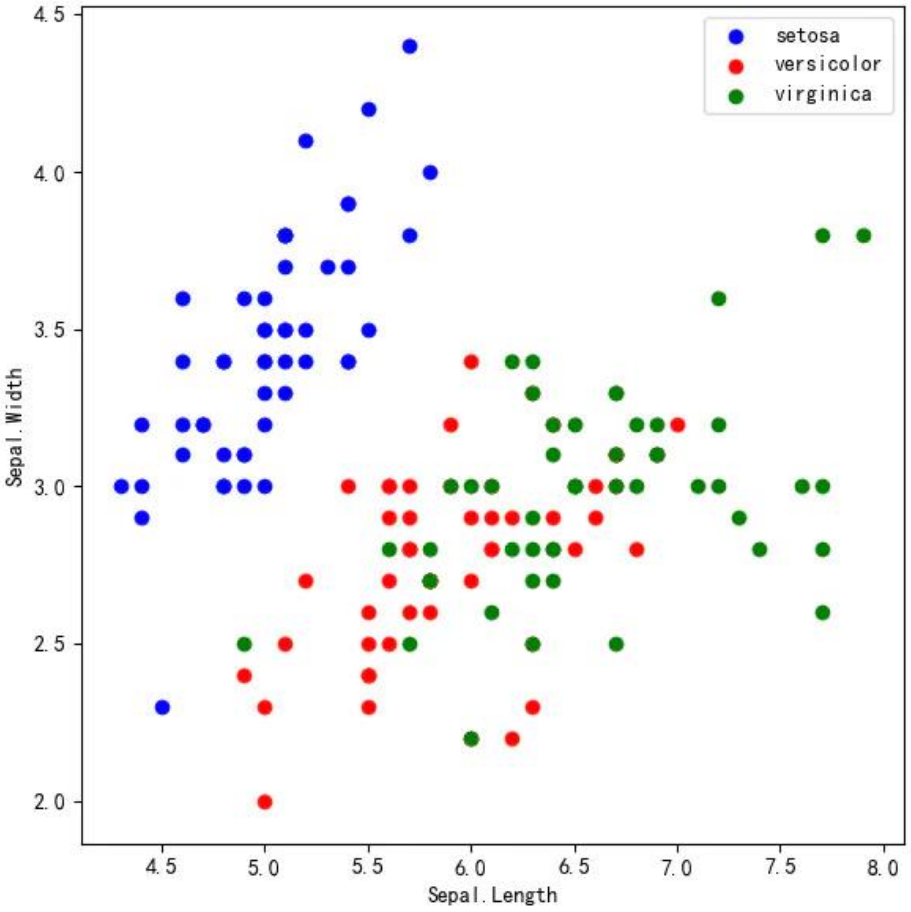
- 主成分分析（Principal Components Analysis，PCA）
- 奇异值分解（Singular Value Decomposition）
- 其他：有监督（supervised）和非线性技术

维度规约：PCA

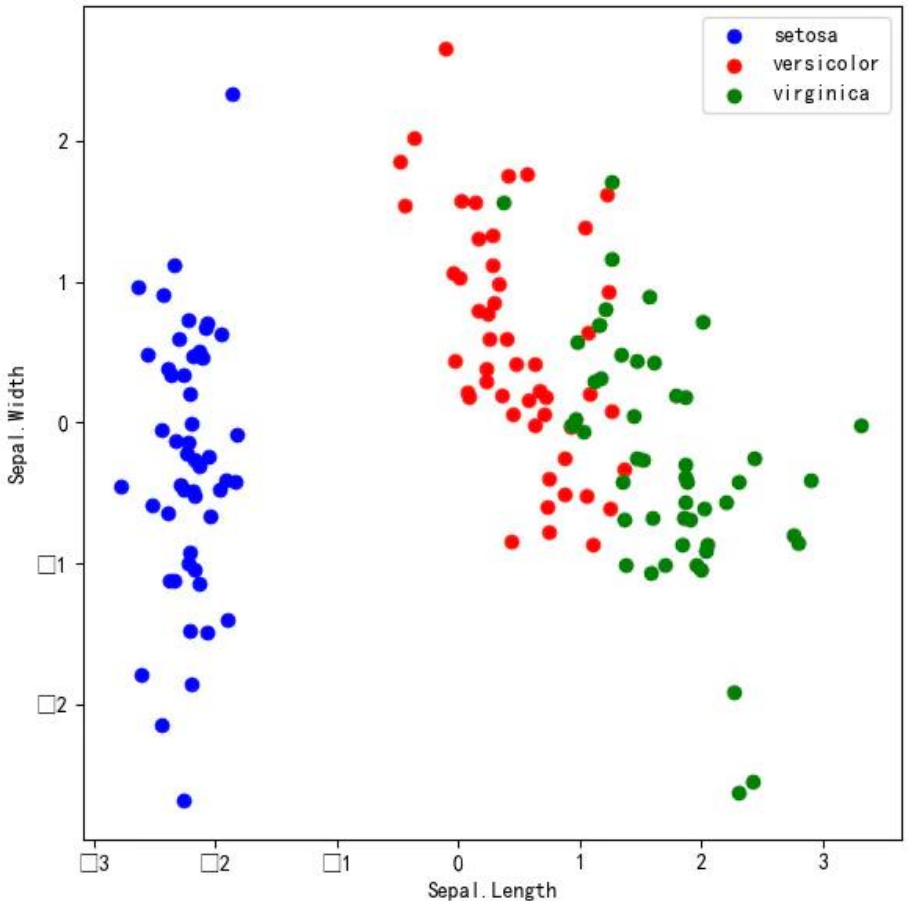
目标是找到一个可以捕获最大量数据变化的投影（projection）



维度规约：PCA



降维前



降维后

2.3.4 特征子集选择：Feature Subset Selection

减少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付的营业税额

不相关的功能

- 不包含对当前数据挖掘任务有用的信息
- 示例：学生ID通常与预测学生的GPA任务无关

技术：

- 嵌入方法
- 过滤方法
- 包装方法

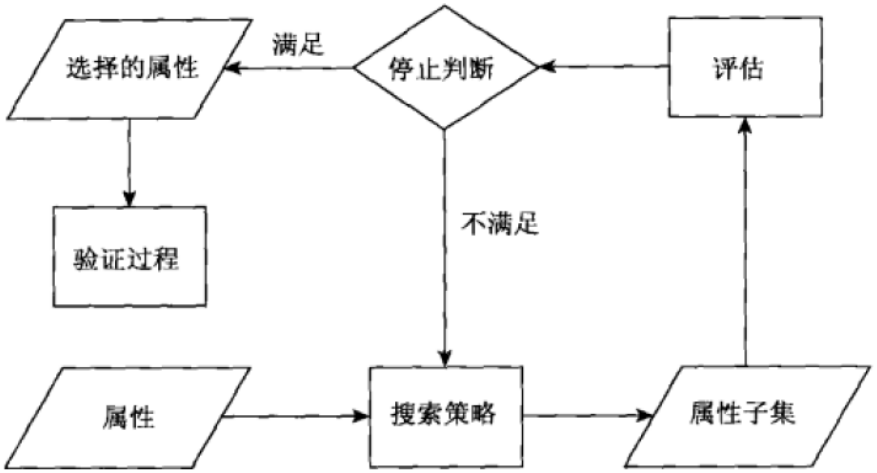


图 2-11　特征子集选择过程流程图

2.3.5 特征创建 Feature Creation

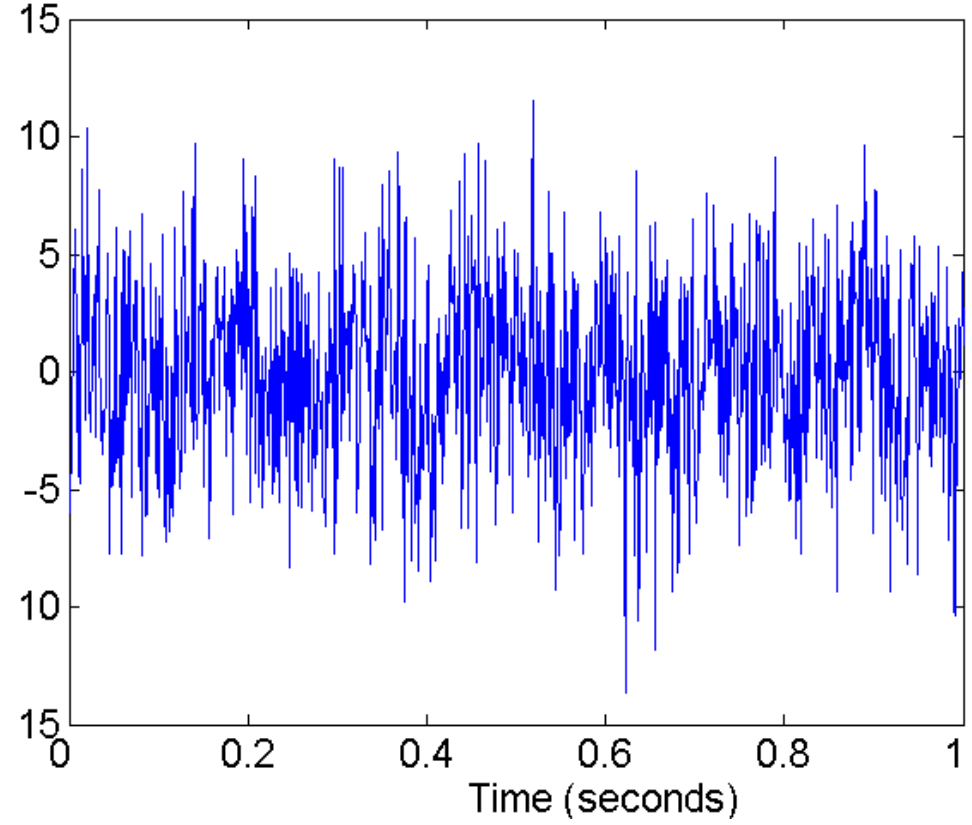
创建可以比原始属性更有效地捕获数据集中重要信息的新属性

三种通用方法：

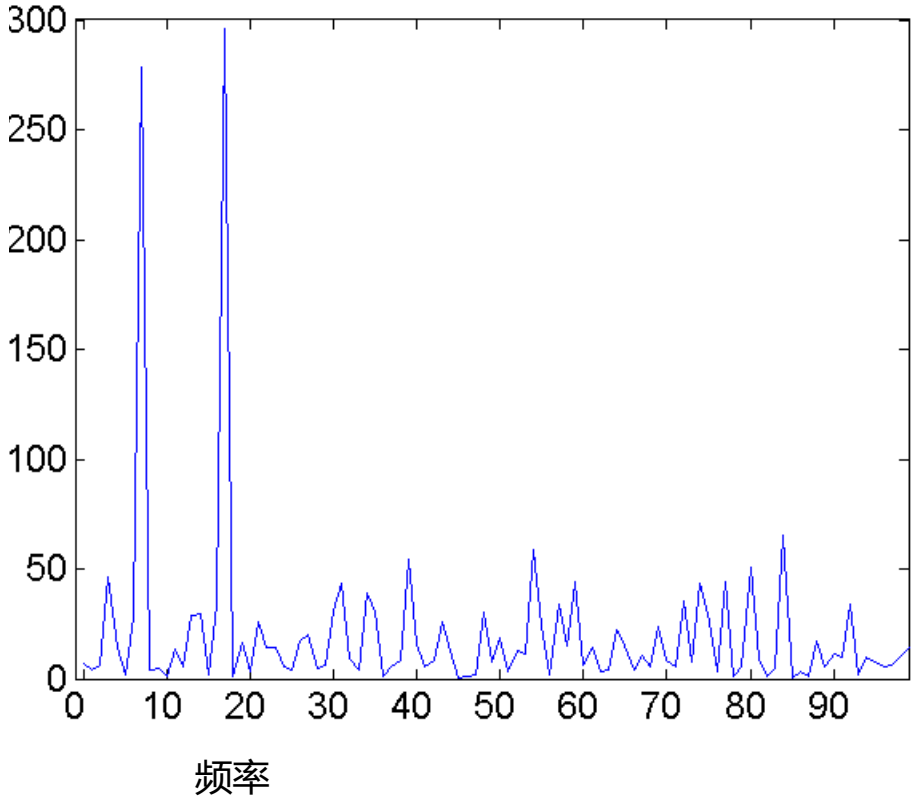
- 特征提取
  - 示例：从图像中提取边缘
- 特征构造
  - 示例：将质量除以体积以获得密度
- 将数据映射到新空间
  - 示例：傅立叶和小波分析

Mapping Data to a New Space

傅里叶（Fourier）和小波（wavelet）变换



Two Sine Waves + Noise



Frequency

2.5.6 二元化 Binarization

二值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类别属性，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进制属性
- 示例：眼睛的颜色深度和高度的测量值为{低，中，高}

2.5.7 离散化 Discretization

离散化是将连续属性转换为序数属性的过程

- 潜在的无限数量的值被映射到少数类别
- 离散化通常用于分类
- 如果自变量和因变量都只有几个离散值，那么许多分类算法效果都会提升
- 我们使用鸢尾花（Iris）数据集说明离散化的有用性

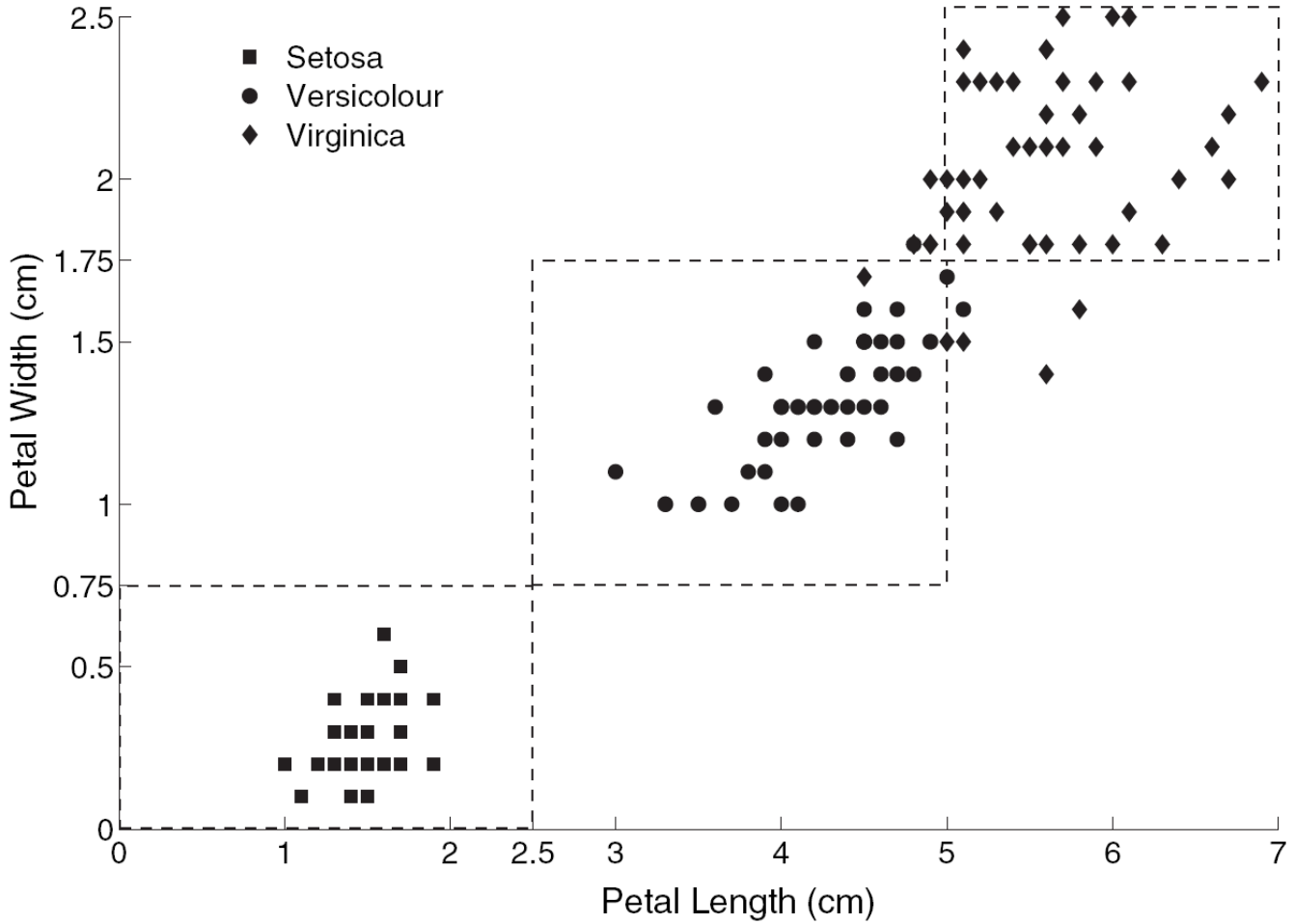
Iris Sample Data Set

鸢尾花植物数据集

- Can be obtained from the UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html
- From the statistician Douglas Fisher
- 三种花型（类）：
  - Setosa
  - Versicolour
  - Virginica
- 四个（非类别）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 Example

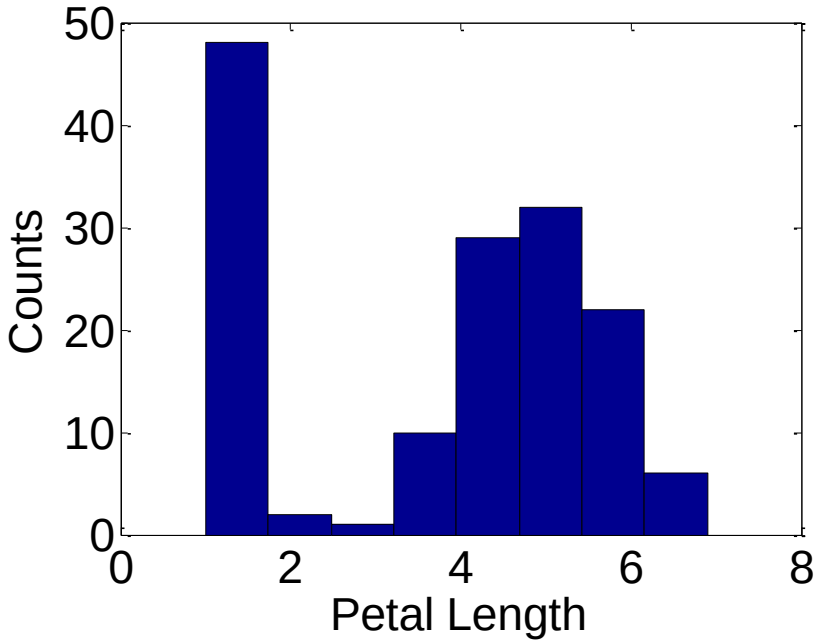


花瓣宽度低或花瓣长度低意味着 Setosa。
花瓣宽度中等或花瓣长度中等表示 Versicolour。
花瓣高或花瓣长高意味着 Virginica。

离散化: Iris Example

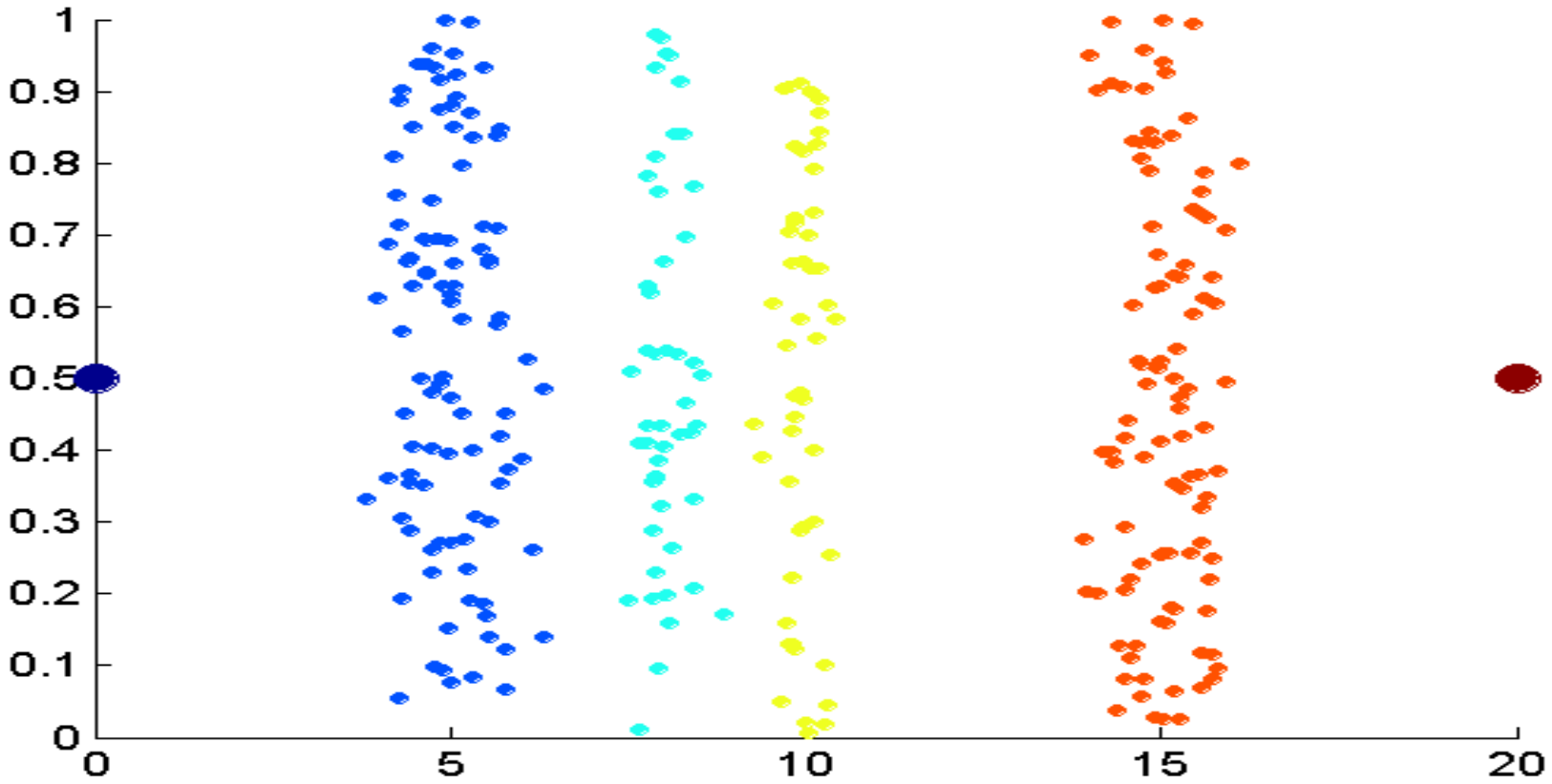
我们如何区分好的离散化是什么?

- 无监督离散化：查找数据值中的中断
  - 示例：花瓣（Petal）长度



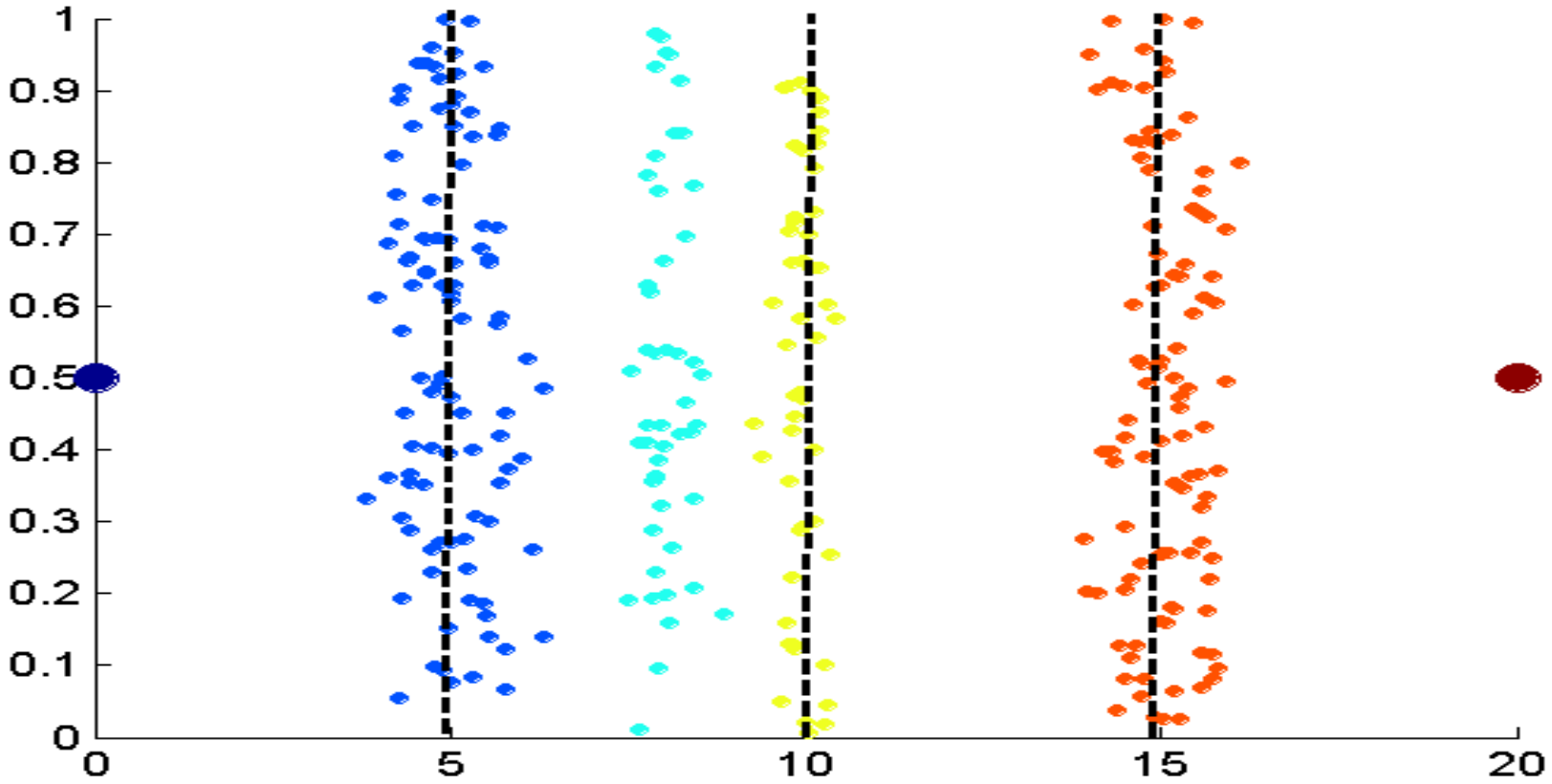
- 有监督的离散化：
  - 使用类标签查找中断

不使用标签/类别信息的离散化



数据由四组点和两个离群值组成。 数据是一维的，但添加了随机y分量以减少重叠。

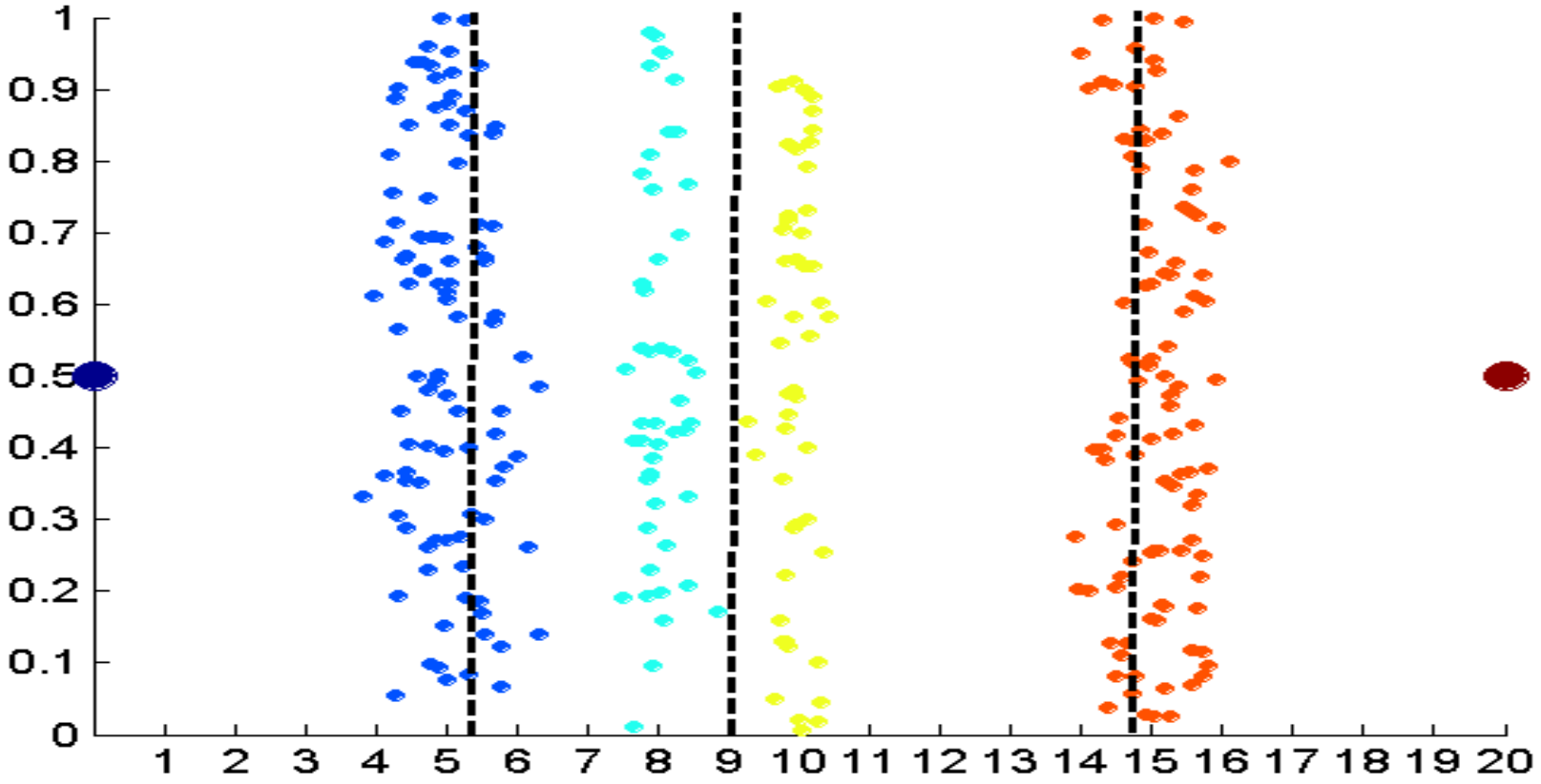
不使用标签/类别信息的离散化



等宽离散化

Equal interval width approach used to obtain 4 values.

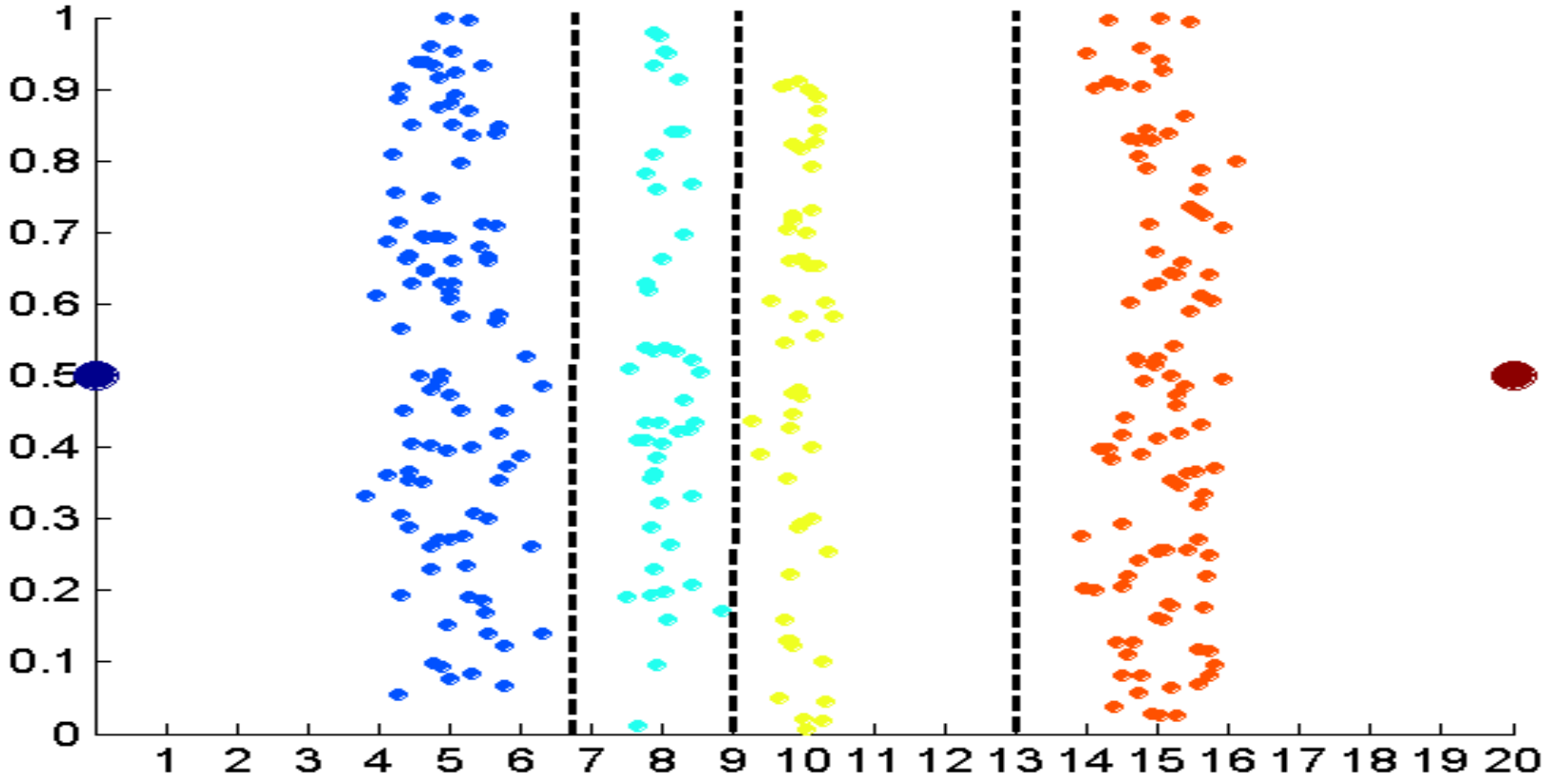
不使用标签/类别信息的离散化



等频率离散化

Equal frequency approach used to obtain 4 values.

不使用标签/类别信息的离散化



K均值离散化（聚类方法）

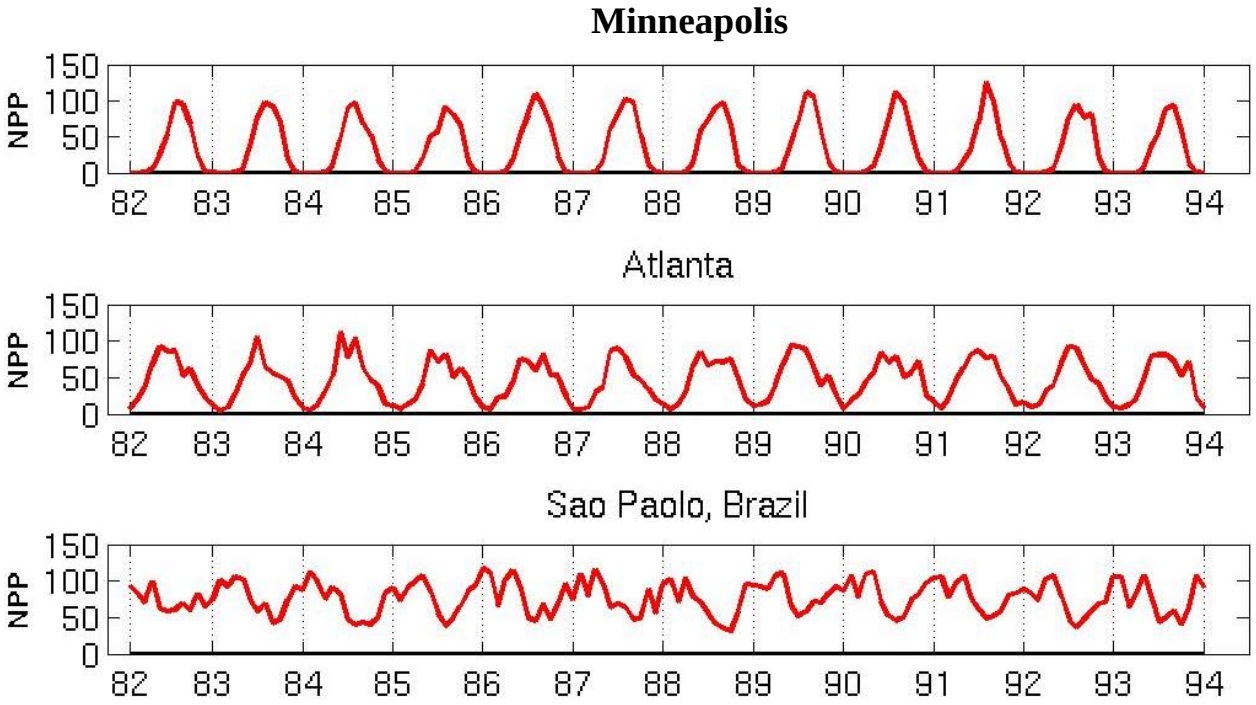
K-means approach to obtain 4 values.

2.5.7 属性转换 Attribute Transformation

属性转换是一种将给定属性的整个值集映射到一组新的替换值的函数，以便可以使用一个新值来标识每个旧值

- 简单函数: x^k, log(x), e^x, |x|
- 归一化Normalization
  - 指用于根据出现频率，均值，方差，范围调整属性之间差异的方法
  - 去掉不需要的常见信号，例如季节性
- 在统计学中，标准化standardization是指减去均值并除以标准差

示例：植物生长的时间序列样本

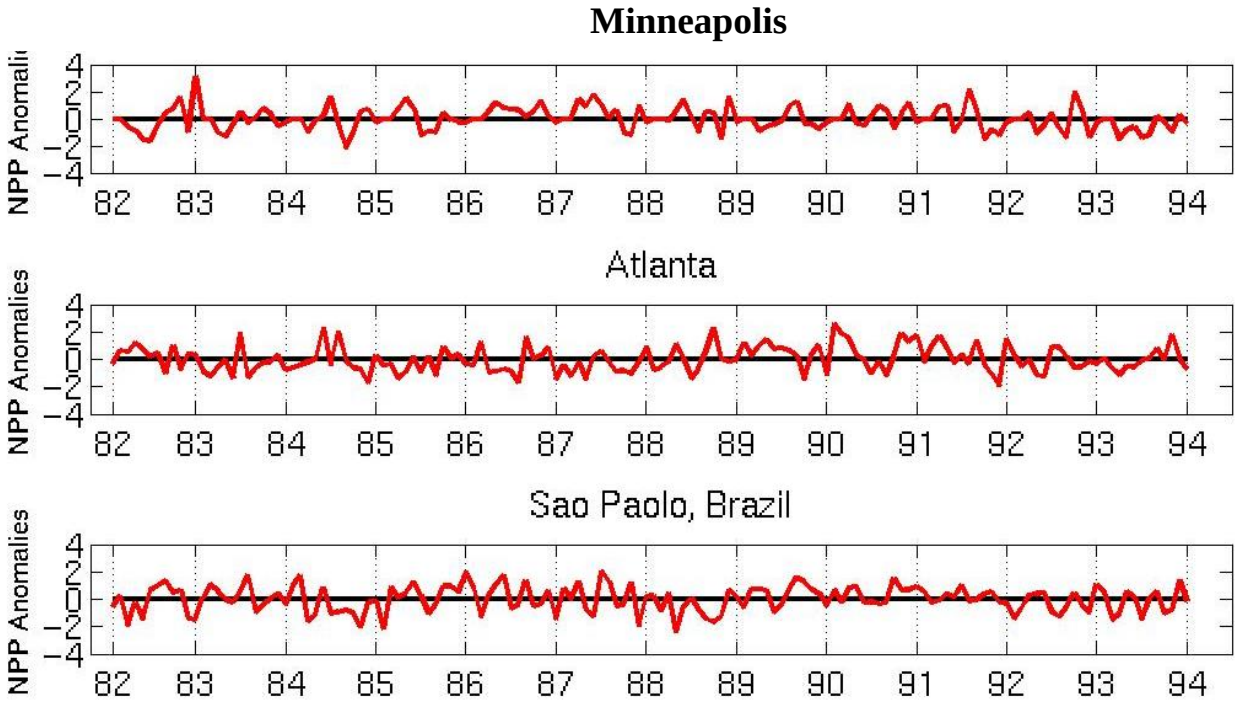


Net Primary Production (NPP) is a measure of plant growth used by ecosystem scientists.

Correlations between time series

	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

Seasonality Accounts for Much Correlation



Normalized using monthly Z Score:

Subtract off monthly mean and divide by monthly standard deviation

Correlations between time series

	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

邻近度（Proximity）比较

应用领域

- 相似性度量往往限定于属性和数据的类型
- 记录数据，图像，图形，序列，3D蛋白质结构等倾向于具有不同的度量方式

但是，邻近度(Proximity)度量应该具备一定的属性

- 对称(Symmetry)
- 容忍噪声和离群值/异常值
- 能够发觉更多类型的模式
- 等等

该度量策略应该适用于当前的数据，并且能够产生符合领域知识的结果

作业

杭电网络教学平台上

1. 注意截止时间
2. 需要提交（而非仅保存）

谢谢！

数据挖掘
教师：王东京
学院：计算机学院
邮箱：dongjing.wang@hdu.edu.cn