数据

2021/9/13

数据是什么?

- 样本和属性的集合
- 属性（Attribute）是一个样本的特性或特征
  - 例如: 性别、年龄、气温
  - Attribute is also known as variable（变量）, field（域）, characteristic（特征）, or feature（特征）
- 一组属性用来描述一个样本（Object）
  - Object is also known as record（记录）, point（点）, case（事例）, sample（抽样）, entity（实体）, or instance（实例）

属性

样本

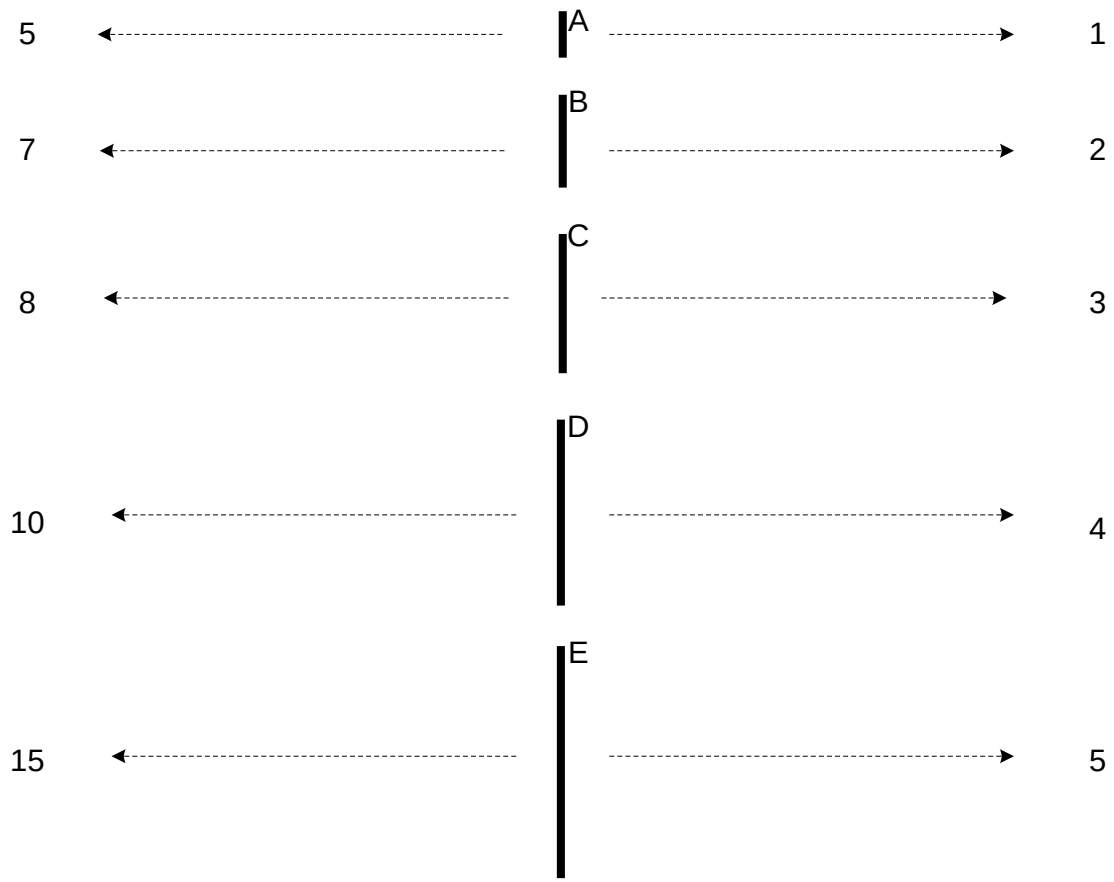
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

属性值

- 属性值是赋给一个属性的“数值（number）”或“符号（number）”
- 属性和属性值的区别
  - 同一个属性可以用不同的属性值来表示
    - 如：长度可以以米、里或公里为单位来表示
  - 不同的属性可以用相同的属性值来表示
    - 如: ID和年龄的属性值都是实数
    - 但属性的含义可能不一样
      - ID取值没有限制，但年龄取值得在一定范围
      - 对于ID的属性来说，“12”可能表示第12格；但对于年龄的属性，“12”则意味着12岁

同一个属性可以用不同的方式测量

- 属性可以用一种不描述属性全部性质的方式测量



两种不同的测量标度下的线段长度测量值

属性的类型

- 类别属性（定性的）
  - 标称 （值仅仅只是不同的名字，只用于区分对象）
    - 如: ID数值，性别，邮政编码
  - 序数（值提供足够的信息确定对象的序）
    - 如：成绩，年级，高度{高, 中, 短}
- 数值属性（定量的）
  - 区间（值之间的差是有意义的）
    - 如：日历日期，温度（摄氏度）
  - 比率（差和比率计算都是有意义的）
    - 如：绝对温度、年龄、质量

属性值的特点

- 属性的类型由其所支持的操作类型决定：
  - 相异: = ≠
  - 排序: < >
  - 相加: + -
  - 相乘: * /

  - 标称属性：相异
  - 序数属性：相异、排序
  - 区间属性：相异、排序、相加
  - 比率属性：相异、排序、相加、相乘

属性类型	变换	注释
标称	任何的一对一变换	如果所有雇员的ID重新赋值，会出现什么异常吗？
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
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

离散属性和连续属性

- 离散属性
  - 只包含有限个或无限可数个值
  - 如：邮政编码，计数，文档中的词
  - 一般表达为实数变量
  - 注：二值属性是离散属性的特例

- 连续属性
  - 实数作为属性值
  - 如：温度，高度，重量
  - 实际上，实值只能使用有限数字来测量和表示
  - 连续属性通常表示为浮点变量

数据集类型

- 记录
  - 数据矩阵
  - 文档数据
  - 事务数据
- 基于图形的数据
  - 万维网
  - 分子结构
- 有序数据
  - 空间数据
  - 时间数据
  - 时序数据
  - 基因组序列数据

结构化数据的重要特征

- 维度
  - 维灾难
- 稀疏性
  - 只计数非零值
- 分辨率
  - 数据模依赖于分辨率

记录数据

- 数据集是记录的汇集，每个记录包含固定的数据字段集。

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

数据矩阵

- 如果一个数据集族中的所有数据对象都具有相同的数值属性集，则数据对象可以看做多维空间中的点，其中每个维代表对象的一个不同属性。
- 这样的数据对象集可以用一个m×n的矩阵表示，其中m行，一个对象一行； n列，一个属性一列。

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

- 文档可用词向量表示
  - 每个词是向量的一个分量
  - 每个分量的值是对应词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

- 一种特殊类型的记录数据, 其中
  - 每个记录涉及一系列的项
  - 考虑一个杂货店。顾客一次购物所购买的商品的集合就构成一个事务，而购买的商品是项。

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

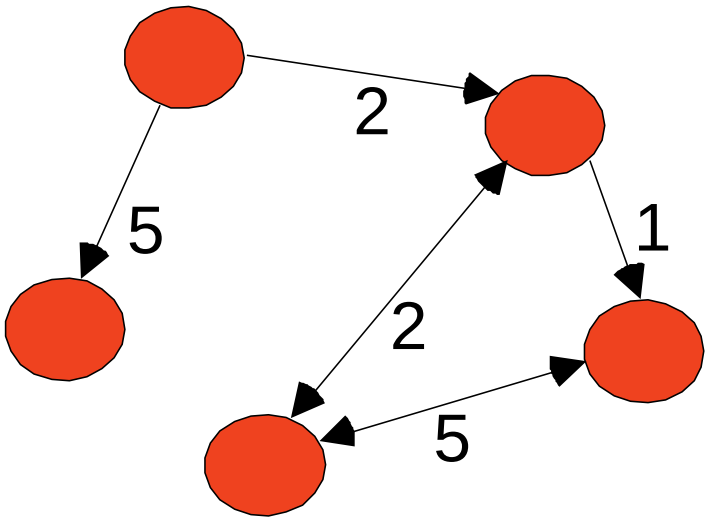
网站日志数据

- 网站运行写下的数据, 其中
  - 每个记录表示一次访问记录
  - 每行记录有5部分组成：访问者IP、访问时间、访问资源、访问状态（HTTP状态码）、本次访问流量

```
27.19.74.143 - - [30/May/2013:17:38:20 +0800] "GET /static/image/common/faq.gif HTTP/1.1" 200 1127
110.52.250.126 - - [30/May/2013:17:38:20 +0800] "GET /data/cache/style_1_widthauto.css?y7a HTTP/1.1" 200 1292
27.19.74.143 - - [30/May/2013:17:38:20 +0800] "GET /static/image/common/hot_1.gif HTTP/1.1" 200 680
27.19.74.143 - - [30/May/2013:17:38:20 +0800] "GET /static/image/common/hot_2.gif HTTP/1.1" 200 682
27.19.74.143 - - [30/May/2013:17:38:20 +0800] "GET /static/image/filetype/common.gif HTTP/1.1" 200 90
110.52.250.126 - - [30/May/2013:17:38:20 +0800] "GET /source/plugin/wsh_wx/img/wsh_zk.css HTTP/1.1" 200 1482
110.52.250.126 - - [30/May/2013:17:38:20 +0800] "GET /data/cache/style_1_forum_index.css?y7a HTTP/1.1" 200 2331
110.52.250.126 - - [30/May/2013:17:38:20 +0800] "GET /source/plugin/wsh_wx/img/wx_jqr.gif HTTP/1.1" 200 1770
27.19.74.143 - - [30/May/2013:17:38:20 +0800] "GET /static/image/common/recommend_1.gif HTTP/1.1" 200 1030
```

基于图形的数据

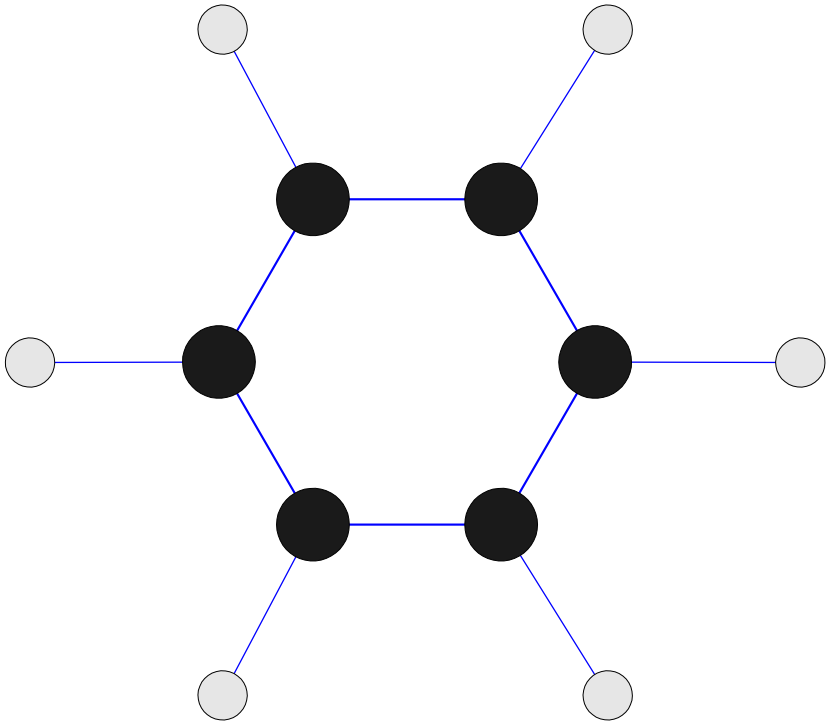
- 如: 通用图和HTML链接




```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

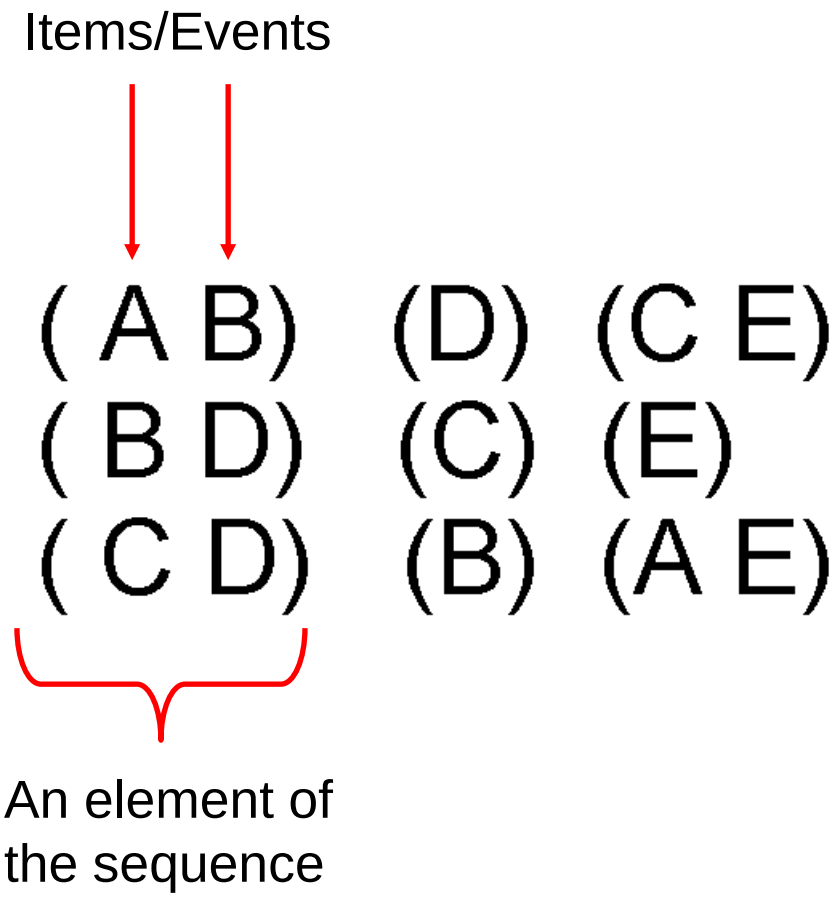
化学数据

- 苯分子：C_6H_6



有序数据

- 时序事务数据



有序数据

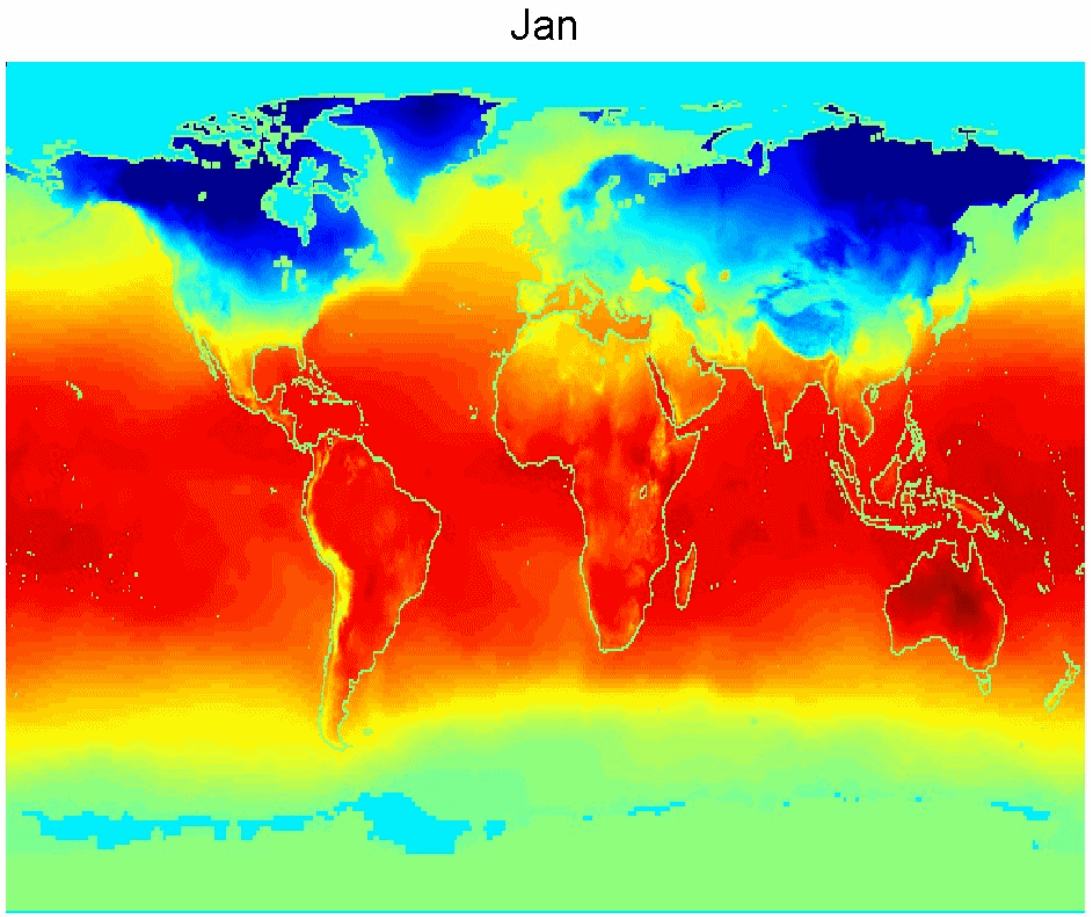
- 基因组序列数据

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

有序数据

- 空间温度数据

陆地和海洋的平均月温度



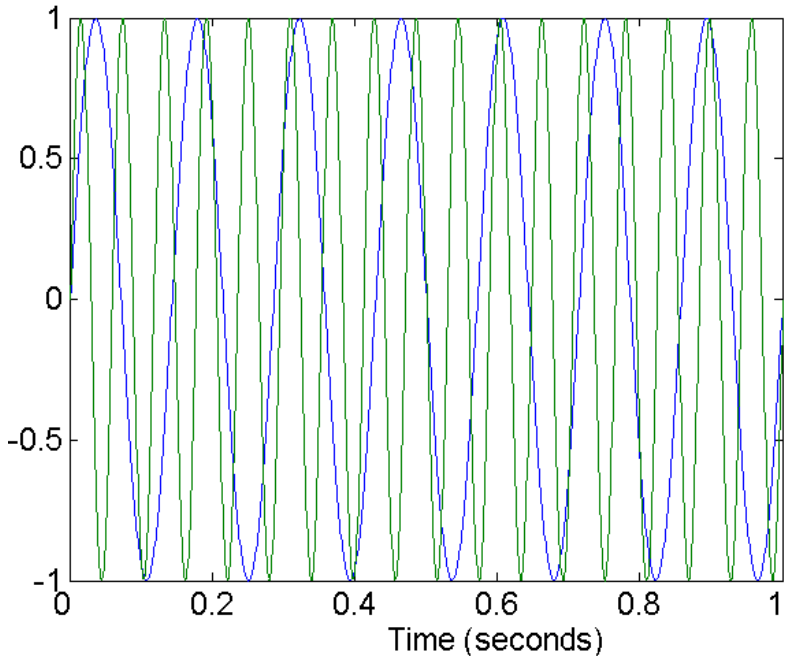
数据质量

- 什么样的数据质量问题？
- 我们如何检测数据的问题？
- 我们能做些什么来解决这些问题？

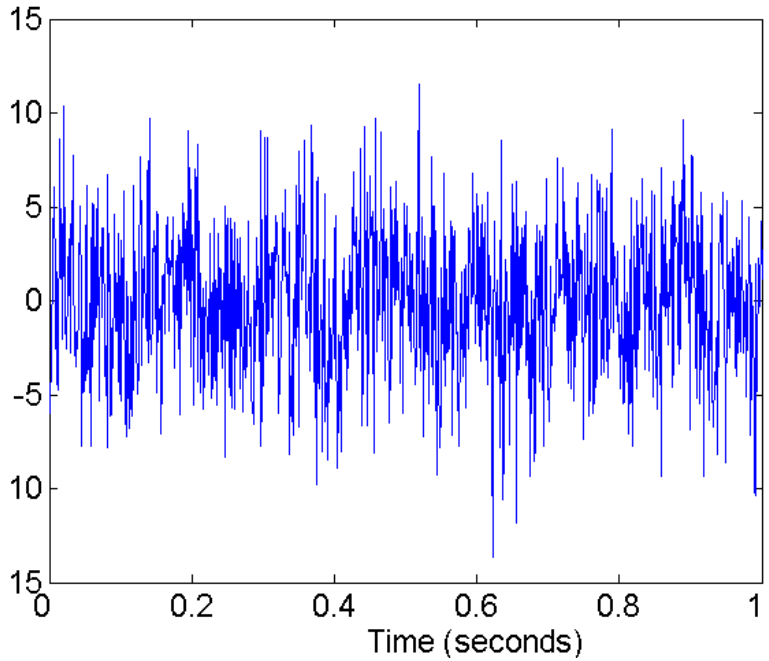
- 数据质量问题举例:
  - 噪声和离群点
  - 遗漏值
  - 重复数据

噪声

- 噪声指原始值的修改
  - 如：在破旧的电话上谈话时，人的声音失真，电视屏幕上出现“雪花”



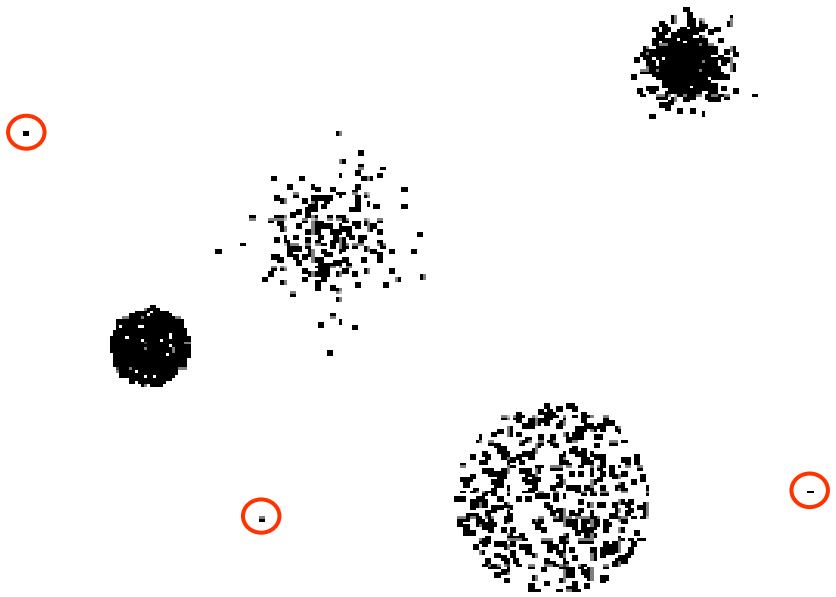
Two Sine Waves



Two Sine Waves + Noise

离群点

- 离群点是具有不同于数据集中其他大部分数据对象的特征的数据对象。



遗漏值

- 遗漏值的原因
  - 没有收集到信息
    (如：人们拒绝透漏他们的年龄和体重)
  - 属性不适用于所有情况
    (如：年收入不适用于儿童)

- 处理遗漏值
  - 删除数据对象
  - 估计遗漏值
  - 在分析期间忽略遗漏值
  - 替换为所有可能的值（以其概率加权）

重复数据

- 数据集可能包含重复或几乎重复的数据对象。
  - 从异构源合并数据时的主要问题
- 如：
  - 一个人有多个电子邮件地址
- 去重复
  - 处理重复数据问题的过程

相似性和相异性

- 相似度
  - 两个对象相似程度的数值度量.
  - 两个对象越相似，相似度越高
  - 在区间[0,1]取值
- 相异度
  - 两个对象差异程度的数值度量
  - 相异度的最小值通常是0
  - 上限不同
- 接近度是指相似度或相异度

简单属性之间的相似度和相异度

假设p、q为两个样本。

属性类型	相异度	相似度
标称	$d = \begin{cases} 0 & if\ p = q \\ 1 & if\ p \neq q \end{cases}$	$s = \begin{cases} 1 & if\ p = q \\ 0 & if\ p \neq q \end{cases}$
序数（将值映射到0,..,m-1）	$d = \frac{\lvert p - q \rvert}{m - 1}$	$s = 1 - \frac{\lvert p - q \rvert}{m - 1}$
区间、比率	$d = \lvert p - q \rvert$	$s \propto \exp(-d)$, $s = \frac{1}{1 + d}$ $s = 1 - \frac{d - d_{min}}{d_{max} - d_{min}}$

欧几里得距离

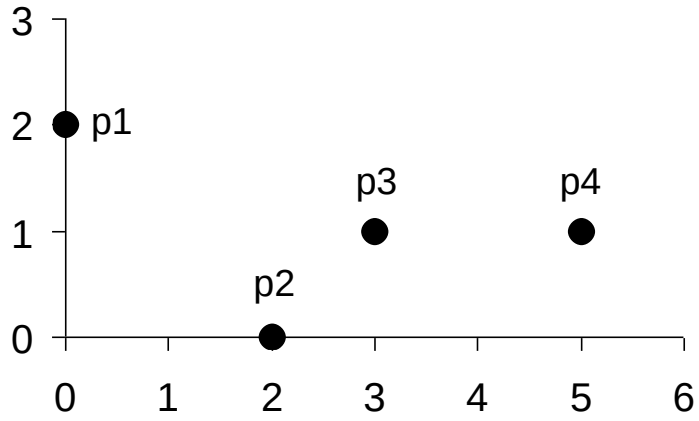
距离是特殊的相异度

- 欧几里得距离

$$dist = \sqrt{\sum_{j=1}^{m}(p_j - q_j)^2}$$

- 如果范围不同，则需要标准化.

欧几里得距离



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

闵可夫斯基距离

- 闵可夫斯基距离是欧几里得距离的一个泛化：

$$dist = \left(\sum_{j=1}^{m} |p_j - q_j|^r \right)^{\frac{1}{r}}$$

- r = 1. City block (Manhattan, taxicab, L_1 norm) distance.
  - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors
- r = 2. Euclidean distance
- $r \to \infty$. "supremum" (L_{max} norm, L_∞ norm) distance.
  - This is the maximum difference between any component of the vectors

闵可夫斯基距离

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

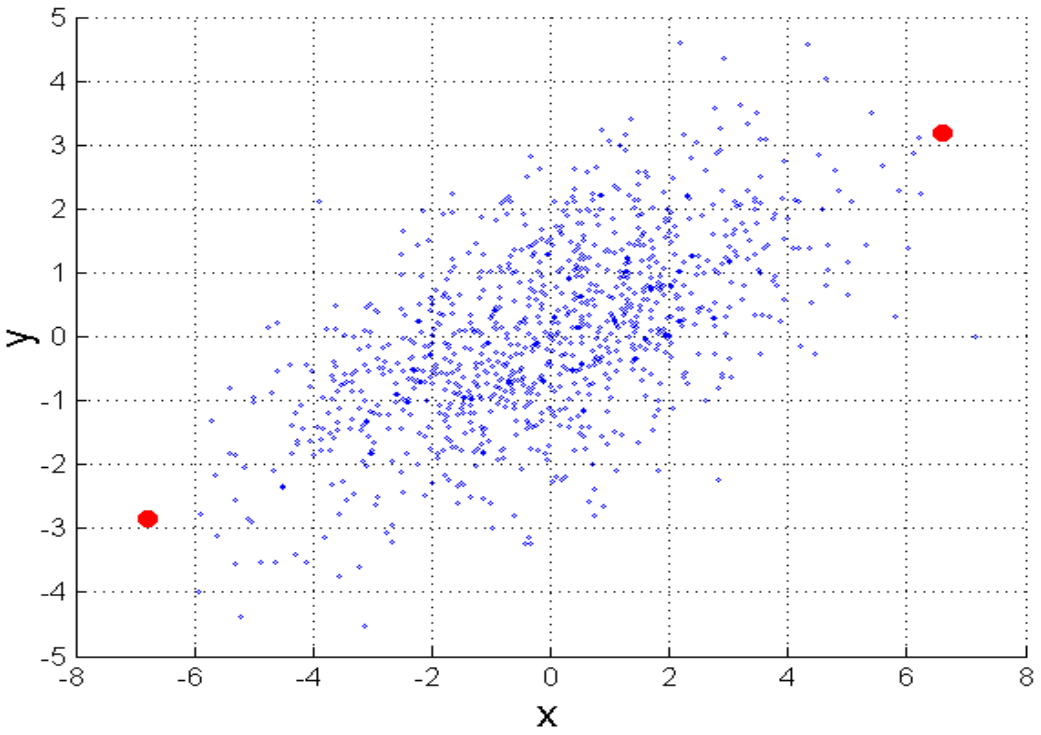
L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

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

Distance Matrix

马哈拉诺比斯距离

$$mahalanobis(p,q) = (p-q)\Sigma^{-1}(p-q)^T$$

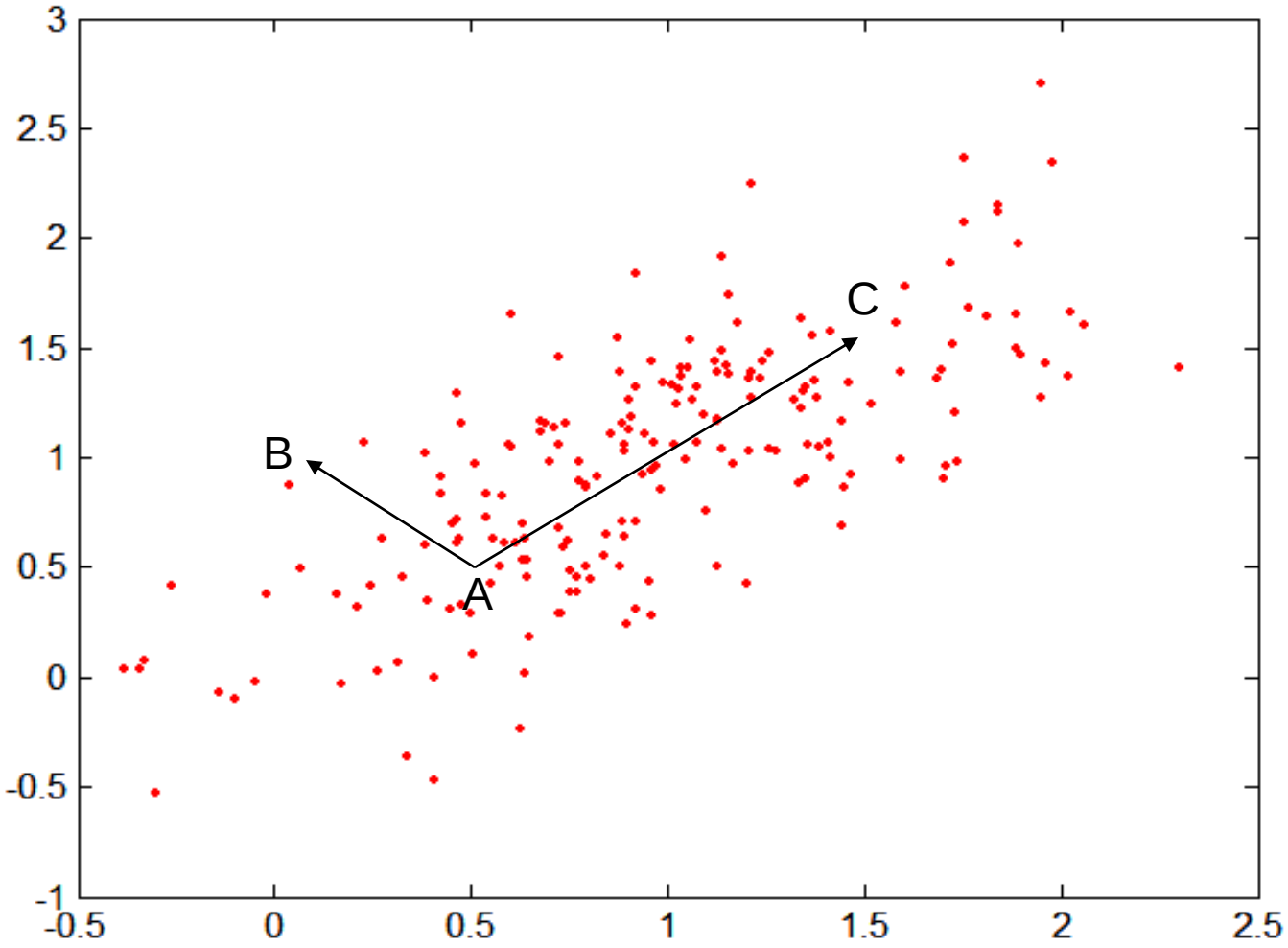


Σ 是输入数据X的协方差矩阵

$$\Sigma_{j,k} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)(X_{ik} - \overline{X}_k)$$

对两个红点来说, 欧几里得距离是 14.7,马哈拉诺比斯距离是 6.

马哈拉诺比斯距离



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

距离的常用属性

- 距离，例如欧几里得距离，有一些常见属性：

  1. $d(p, q) \geq 0$ for all p and q and $d(p, q) = 0$ only if $p = q$. (Positive definiteness)
  2. $d(p, q) = d(q, p)$ for all p and q. (Symmetry)
  3. $d(p, r) \leq d(p, q) + d(q, r)$ for all points p, q, and r. (Triangle Inequality)

  其中，$d(p, q)$是数据点或数据对象p和q之间的距离或称相异度

- 满足这些属性的距离是度量

相似度的常见属性

- 相似度，也有一些众所周知的属性。

1. *s(p, q) = 1* (or maximum similarity) only if *p = q*.
2. *s(p, q) = s(q, p)* for all *p* and *q*. (Symmetry)

其中，*s(p, q)*是数据点或数据对象*p*和*q*之间的相似度

二元数据的相似性度量

- 常见的情况是对象p和q只有二进制属性
- 使用以下公式计算相似度

M_{01} = the number of attributes where p was 0 and q was 1
M_{10} = the number of attributes where p was 1 and q was 0
M_{00} = the number of attributes where p was 0 and q was 0
M_{11} = the number of attributes where p was 1 and q was 1

- 简单匹配和 Jaccard系数

SMC = number of matches / number of attributes
= $(M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

J = number of 11 matches / number of not-both-zero attributes values
= $(M_{11}) / (M_{01} + M_{10} + M_{11})$

简单匹配系数

p = 1 0 0 0 0 0 0 0 0 0
q = 0 0 0 0 0 0 1 0 0 1

M_{01} = 2 (the number of attributes where p was 0 and q was 1)
M_{10} = 1 (the number of attributes where p was 1 and q was 0)
M_{00} = 7 (the number of attributes where p was 0 and q was 0)
M_{11} = 0 (the number of attributes where p was 1 and q was 1)

$$SMC = (M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00}) = (0+7) / (2+1+0+7) = 0.7$$

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

余弦相似度

- 如果d_1和 d_2 是两个文档向量, 那么

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / \|d_1\| \|d_2\|,$$

其中• 表示向量的点积， $\| d \|$ 是向量d的长度

- 如:

d_1 = **3 2 0 5 0 0 0 2 0 0**
d_2 = **1 0 0 0 0 0 0 1 0 2**

$d_1 \bullet d_2$= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5

$\|d_1\|$ = (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

$\|d_2\|$ = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2) $^{0.5}$ = (6) $^{0.5}$ = 2.245

$$\cos(d_1, d_2) = .3150$$

组合相似性的一般方法

- 有时，属性具有许多不同的类型，但是需要总体相似性。

1. For the k^{th} attribute, compute a similarity, s_k, in the range $[0, 1]$.
2. Define an indicator variable, δ_k, for the k_{th} attribute as follows:

$$\delta_k = \begin{cases} 0 & \text{if the } k^{th} \text{ attribute is a binary asymmetric attribute and both objects have} \\ & \text{a value of 0, or if one of the objects has a missing values for the } k^{th} \text{ attribute} \\ 1 & \text{otherwise} \end{cases}$$

3. Compute the overall similarity between the two objects using the following formula:

$$similarity(p, q) = \frac{\sum_{k=1}^{n} \delta_k s_k}{\sum_{k=1}^{n} \delta_k}$$

Thank You!