第二章　数据预处理

- 关于数据
- 为什么要预处理数据？
- 描述性数据汇总
- 数据清理
- 数据集成和变换
- 数据规约
- 小结

什么是数据?

- 数据对象的集合及其属性
- 属性是对象的性质或者特征
  - ✓ 例如：人眼睛的颜色，温度等
  - ✓ 属性也可以理解为变量，领域，特征或者特点
- 描述一个对象的属性集合
  - ✓ 对象也可以理解为记录，观点，案例，样本，实体或者实例

Attributes

Objects

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

离散和连续属性

离散属性

- ✓ 只有一个有限集和可数无限集
- ✓ 例如：邮政编码，计数，或者是一个文档集合的词集
- ✓ 通常表示为整数变量
- ✓ 注：二进制属性是离散属性的特殊情况

连续属性

- ✓ 实数作为属性值
- ✓ 例如：温度，高度，或者重量.
- ✓ 特别的，实际值只能用有限位数的数字测量和表示
- ✓ 连续性属性通常用浮点变量表示

数据集的类型

- 记录
  - ✓ 数据矩阵
  - ✓ 文本数据
  - ✓ 交易数据
- 图表
  - ✓ 互联网
  - ✓ 分子结构
- 顺序的
  - ✓ 空间的数据
  - ✓ 时间的数据
  - ✓ 连续的数据
  - ✓ 基因序列数据

记录数据

- 由记录集合组成的数据，每一个记录又由一个固定的属性集组成

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

- 如果数据对象具有相同的一套固定的数值属性，那么数据对象可以被认为是一个多维空间中的点，其中每个维度代表了不同的属性

- 这样的数据集可以用**m * n**的矩阵表示，**m**行，每行代表一个对象，**n**列，每列代表一个属性

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

文本数据

- 每个文档都成为一个“**term**”向量
  - ✓ 每个**term**都是向量的一个分量,
  - ✓ 每个分量的值就是对应的**term**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

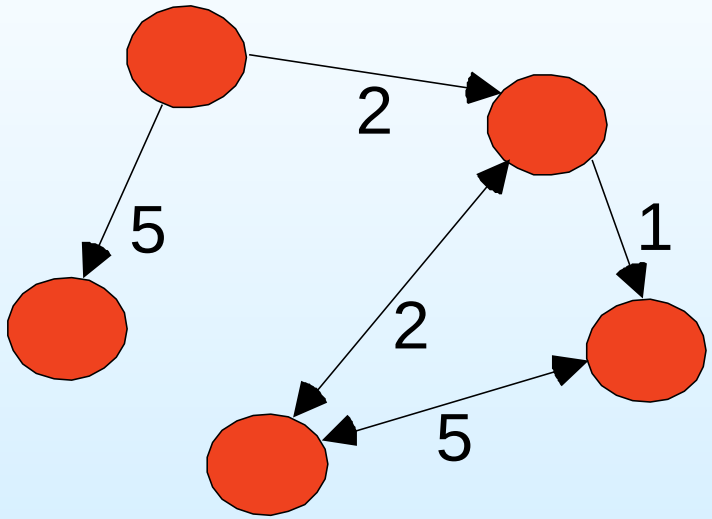
交易数据

- 一组特殊类型的记录数据
  - ✓ 每个交易记录都涉及一组项目
  - ✓ 例如：考虑一个杂货店，一个顾客一次购物所买的一组商品就构成一次交易，这些购买的商品就是项目

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

图片数据

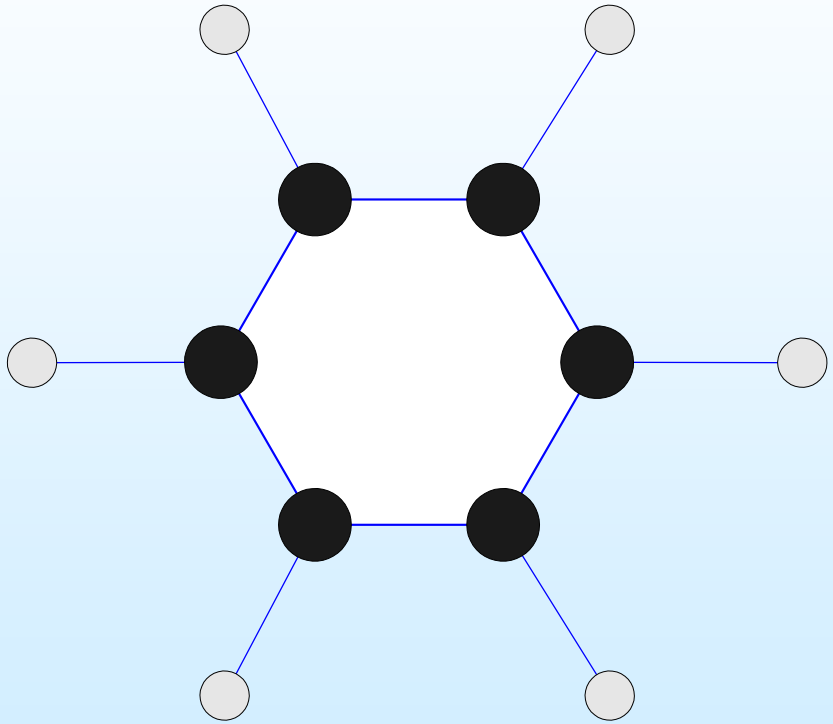
例如：一般的图片和网页链接




```
<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

- **Benzene Molecule (苯分子): C_6H_6**



顺序的数据

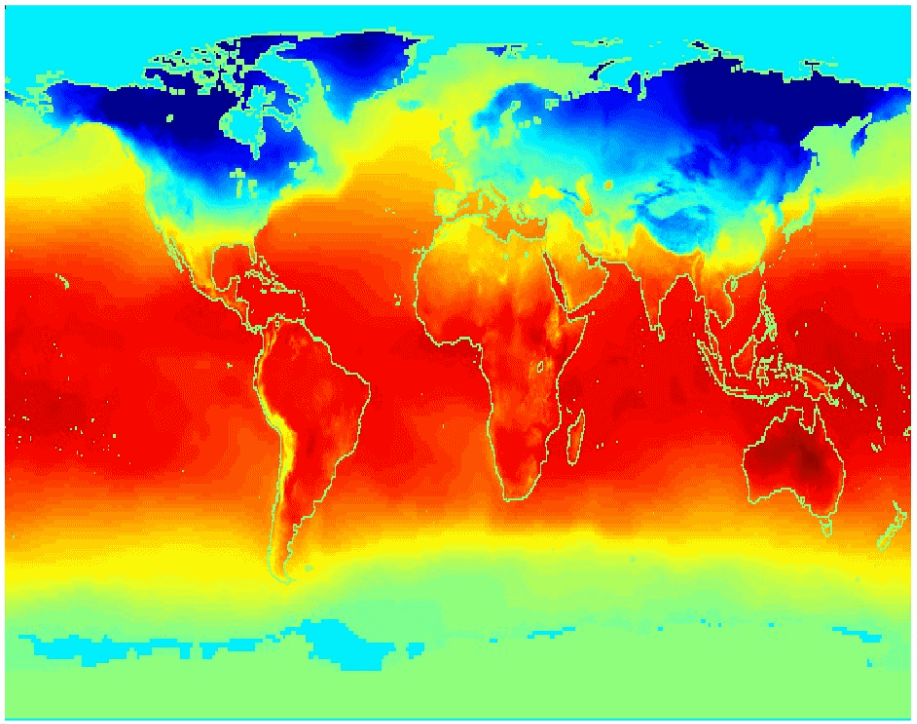
- 染色体序列数据

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

顺序的数据

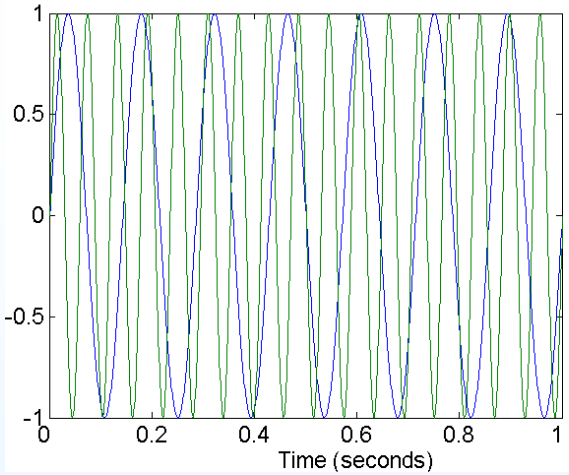
- 时空数据

陆地和海洋的月平均温度

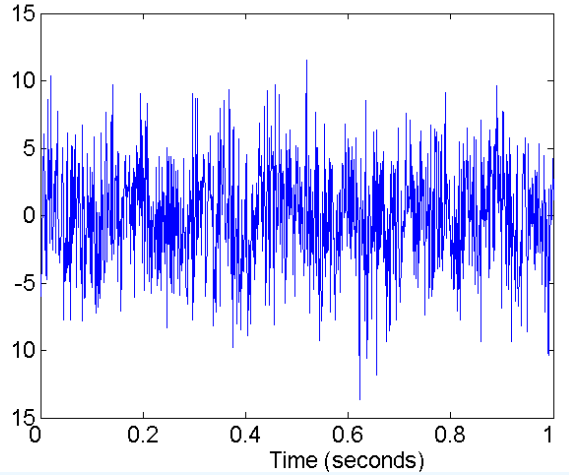


数据质量

- 数据质量问题的例子：
  - ✓ 噪音和离群点
  - ✓ 缺失值
  - ✓ 重复数据

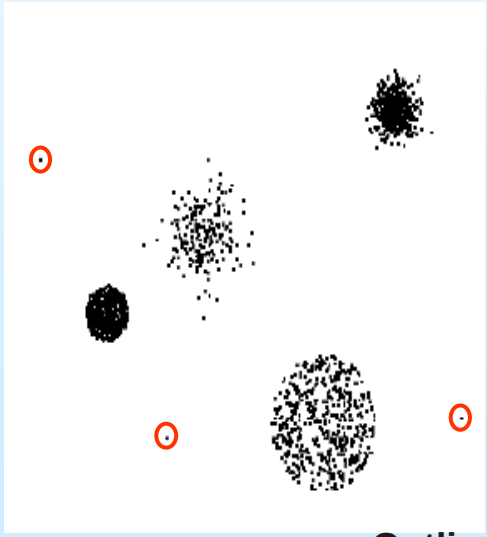


Two Sine Waves



Two Sine Waves + Noise

- 如何从数据中发现问题？
- 怎么处理这些问题？



Outliers