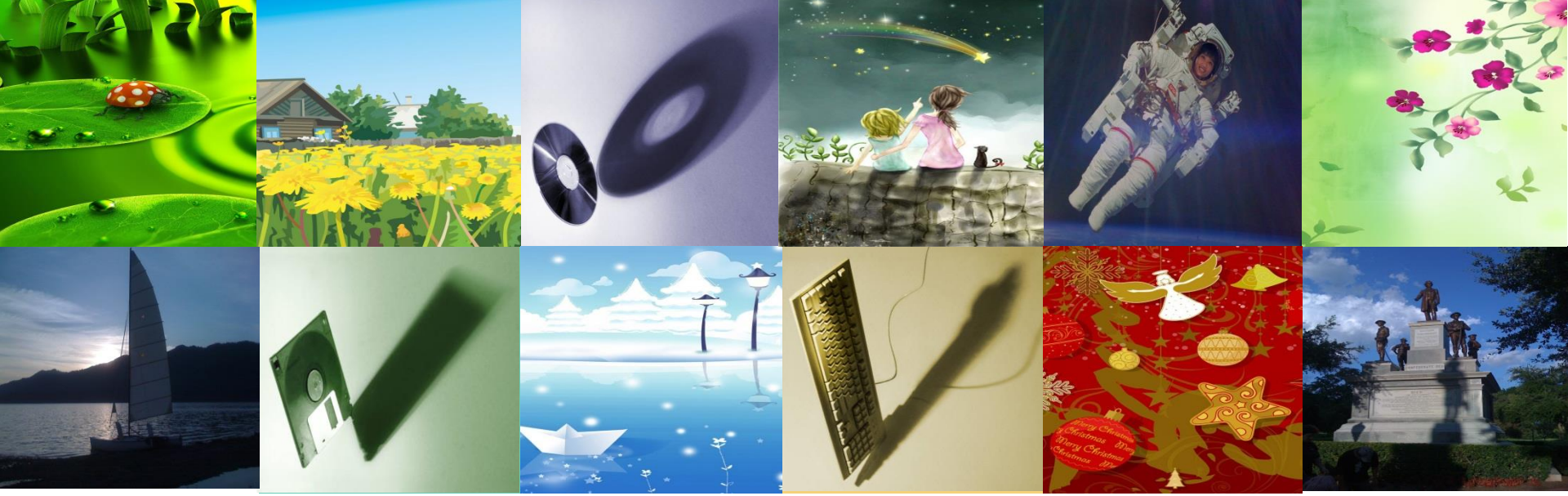


Data Warehouse and Data Mining
第2章 数据预处理

北京邮电大学
计算机学院
王小茹

Data Mining

Lecture Notes for Chapter 2

Introduction to Data **Pre-processing**

Something about Data

本章讨论一些与数据相关的问题，它们对于数据挖掘的成败至关重要。

数据类型 数据集的不同表现在多方面。例如，用来描述数据对象的属性可以具有不同的类型——定量的或定性的，并且数据集可能具有特定的性质，例如，某些数据集包含时间序列或彼此之间具有明显联系的对象。毫不奇怪，数据的类型决定我们应使用何种工具和技术来分析数据。此外，数据挖掘研究常常是为了适应新的应用领域和新的数据类型的需要而展开的。

数据的质量 数据通常远非完美。尽管大部分数据挖掘技术可以忍受某种程度的数据不完美，但是注重理解和提高数据质量将改进分析结果的质量。通常必须解决的数据质量问题包括存在噪声和离群点，数据遗漏、不一致或重复，数据有偏差或者不能代表它应该描述的现象或总体情况。

使数据适合挖掘的预处理步骤 通常，原始数据必须加以处理才能适合于分析。处理一方面是要提高数据的质量，另一方面要让数据更好地适应特定的数据挖掘技术或工具。例如，可能需要将连续值属性（如长度）转换成具有离散的分类值的属性（如短、中、长），以便应用特定的技术。又如，数据集属性的数目常常需要减少，因为属性较少时许多技术用起来更加有效。

根据数据联系分析数据 数据分析的一种方法是找出数据对象之间的联系，之后使用这些联系而不是数据对象本身来进行其余的分析。例如，我们可以计算对象之间的相似度或距离，然后根据这种相似度或距离进行分析——聚类、分类或异常检测。诸如此类的相似性或距离度量很多，要根据数据的类型和特定的应用做出正确的选择。

An example

- 如果只给你一组数据，但不告诉你数据的含义,.....

012	232	33.5	0	10.7
020	121	16.9	2	210.1
027	165	24.0	0	427.6
...				

每行包含一个病人的信息，由 5 个字段组成。

我们想使用前面 4 个字段预测最后一个字段。

2.1 What is Data?

- Collection of data objects and their attributes

- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object
  - Object is also known as **record**, **point**, **case**, **sample**, **entity**, or **instance**

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

2.1.1 Attribute Values 属性值

- **Attribute values** are numbers or symbols assigned to an attribute

- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different
      - ID has no limit but age has a maximum and minimum value

Measurement scale 测量标度

- **测量标度**是将数值或符号值与对象的属性相关联的规则（函数）
- 测量过程是使用测量标度将一个值（符号值或数值）与一个特定对象的特定属性相关联的过程。

Types of Attributes

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

Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness（相异性）： = ≠
  - Order（序）： < >
  - Addition（加法）： + -
  - Multiplication（乘法）： * /
- 属性类型：确定对应于属性基本性质的数值的性质
  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & addition
  - Ratio attribute: all 4 properties

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male*, *female*}	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {*good*, *better*, *best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

Attribute Level	Transformation	Comments
Nominal	Any permutation of values 任意一对一地变换	如果所有雇员的ID号重新复制，不会出现任何不同
Ordinal	An order preserving change of values, i.e., 保序变换 *new_value = f(old_value)* where *f* is a monotonic function.	等价地：good, better best可以用{1, 2, 3} or by { 0.5, 1, 10}替换.
Interval	*new_value =a * old_value + b* where a and b are constants	摄氏温度和华氏温度可以互相计算，只是零度的位置不同.
Ratio	*new_value = a * old_value*	程度可以用米 meters 或英尺 feet度量.

Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes

- Continuous Attribute
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight.
  - Practically, real values can only be measured and represented using a finite number of digits.
  - Continuous attributes are typically represented as floating-point variables.

Asymmetric attribute 非对称的属性

- 对于非对称的属性（asymmetric attribute），出现非零属性值才是最重要的。
- 例如：
  - 大学生选课记录数据集中，数据集的对象是每个学生，每个属性记录了学生是否选秀了大学的某个课程；
  - 对于每个学生，如果选修了对应某个属性的课程，则对应属性取值为1，否则为0；
  - 由于学生只选修所有可选课程中的很小一部分，则这种数据集的大部分值都是0.
  - 因此，关注非零值将更有意义，更有效。
  - 如果在不选修的课程上比较，则大部分学生都是相似的。
  - 只有非零值才重要的二元属性是非对称的二元属性。

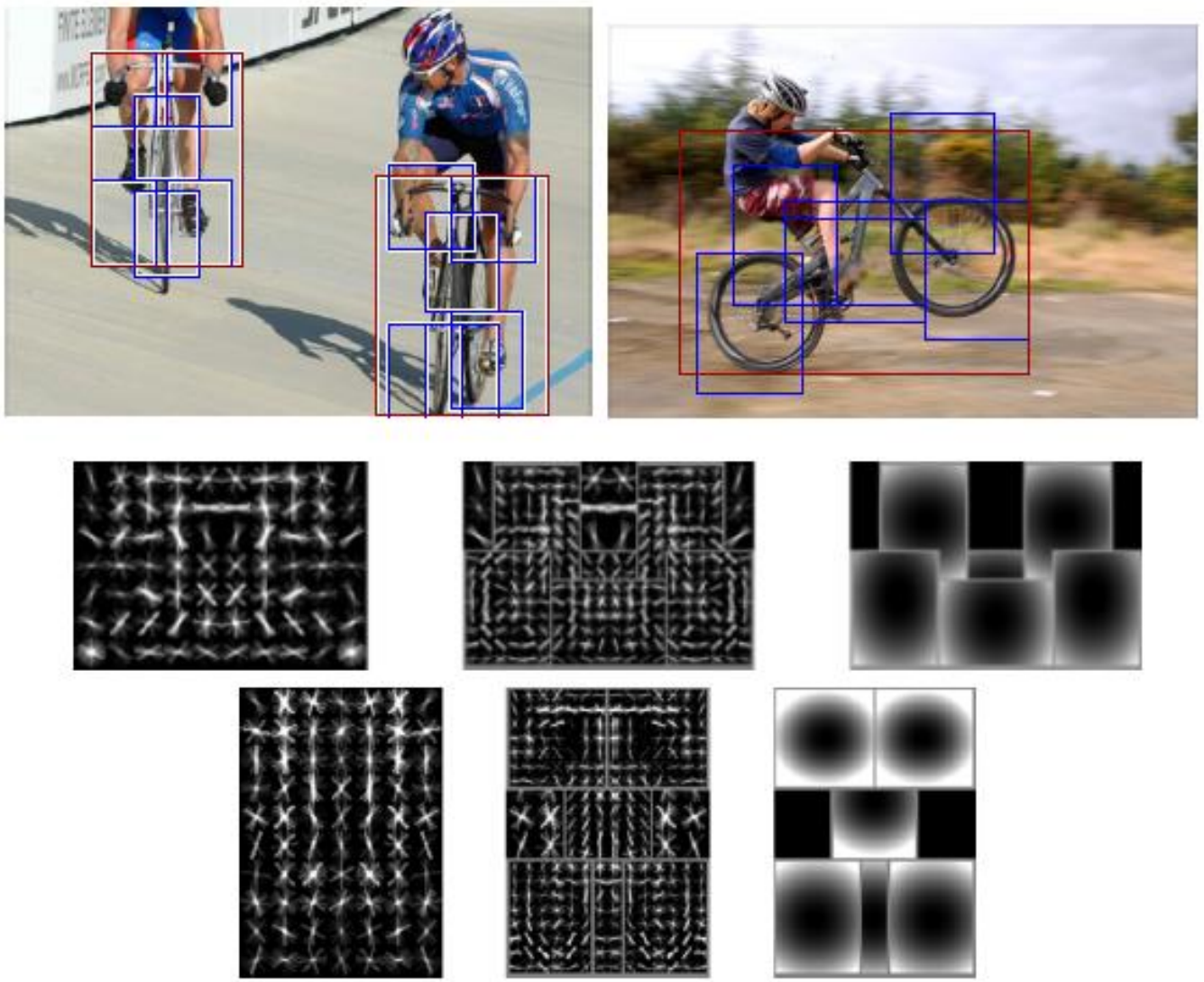
2.1.2 Types of data sets

- **Record** 记录数据
  - **Data Matrix**
  - **Document Data**
  - **Transaction Data**
- **Graph** 基于图形的数据
  - **World Wide Web**
  - **Molecular Structures**
- **Ordered** 有序的数据
  - **Spatial Data**
  - **Temporal Data**
  - **Sequential Data**
  - **Genetic Sequence Data**

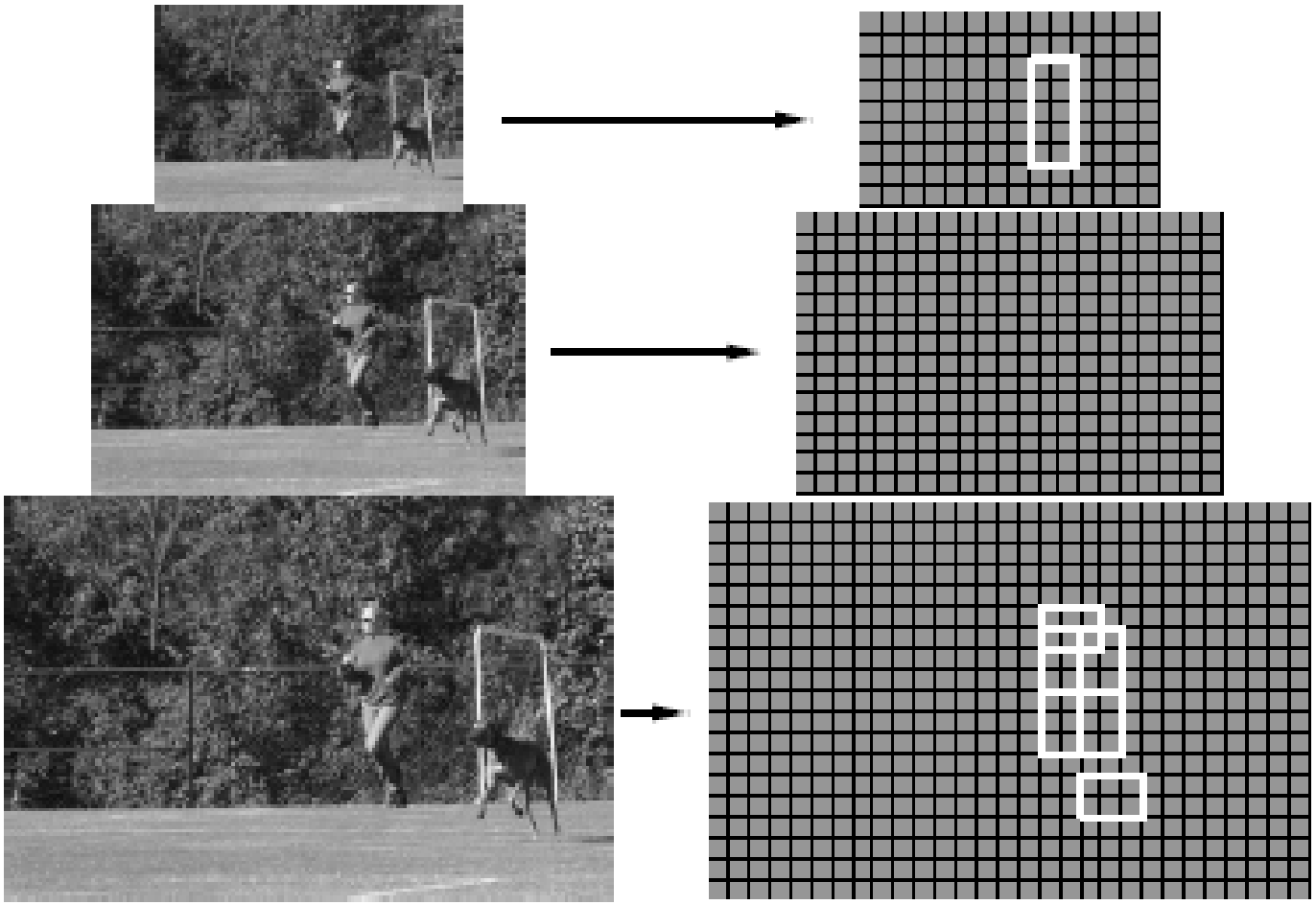
1. Important Characteristics of Structured Data

- **Dimensionality（维度）**
  - 是数据集中的对象具有的属性的数目
  - 低维度数据往往与中、高维度数据有质的不同
  - **Curse of Dimensionality（维度灾难）使得维度归约（dimensionality reduction）是必不可少的。**
- **Sparsity（稀疏性）**
  - 对于具有非对称特征的数据集，一个对象的大部分属性上的值都为0，如非零项少于1%。
  - 稀疏性对于存储和处理都是非常有利的
- **Resolution （分辨率）**
  - 可以在不同的分辨率下得到数据，并且在不同分辨率下数据的性质也不同。
  - 数据的模式依赖于分辨率，如果分辨率太高，模式可能看不出来，或者淹没在噪声中，如果分辨率太低，模式可能不出现。

An example



An example



2. Record Data 记录数据

- 数据集是记录records（数据对象）的汇集，每个记录包含固定的数据字段或属性集
  - 记录数据通常放在平展文件或关系数据库中。
  - 数据挖掘一般不在关系数据库上挖掘。

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

Data Matrix 数据矩阵

- 如果数据集中所有的数据对象具有相同的数值型属性集，则数据对象可以看作多维空间中的点（向量），其中每维代表对象的一个不同的属性。

- 数据对象集可以用一个m×n的矩阵表示，一行一个对象，一列一个属性。
- 可以使用标准的矩阵操作对数据进行变换和处理

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

Document Data 文档数据

- 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

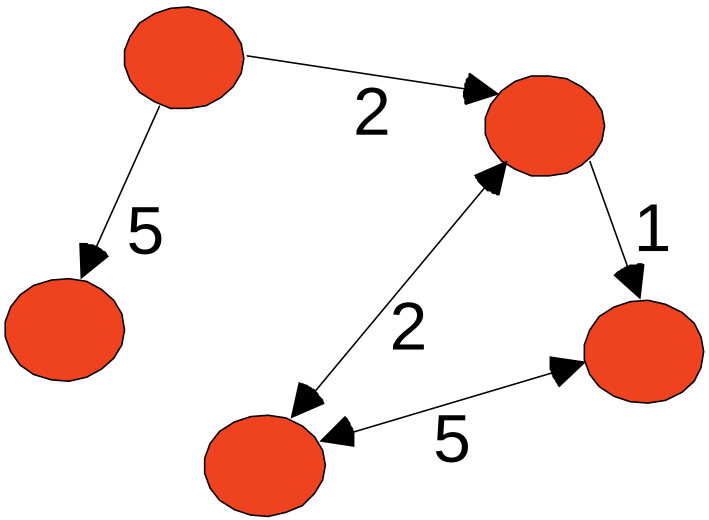
Transaction Data 事务数据

- 一种特殊类型的记录数据，其中：
  - 每个记录 (transaction)涉及一系列项（items）.
  - 例如，购物篮数据（market basket data），顾客一次购物所购买的商品的集合就构成一个事务，购买的商品是项（item）
  - 记录的字段是非对称的属性，常常也是二元的，指出商品是否已买，可以是离散 的或连续的。

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

3. Graph Data 基于图形的数据

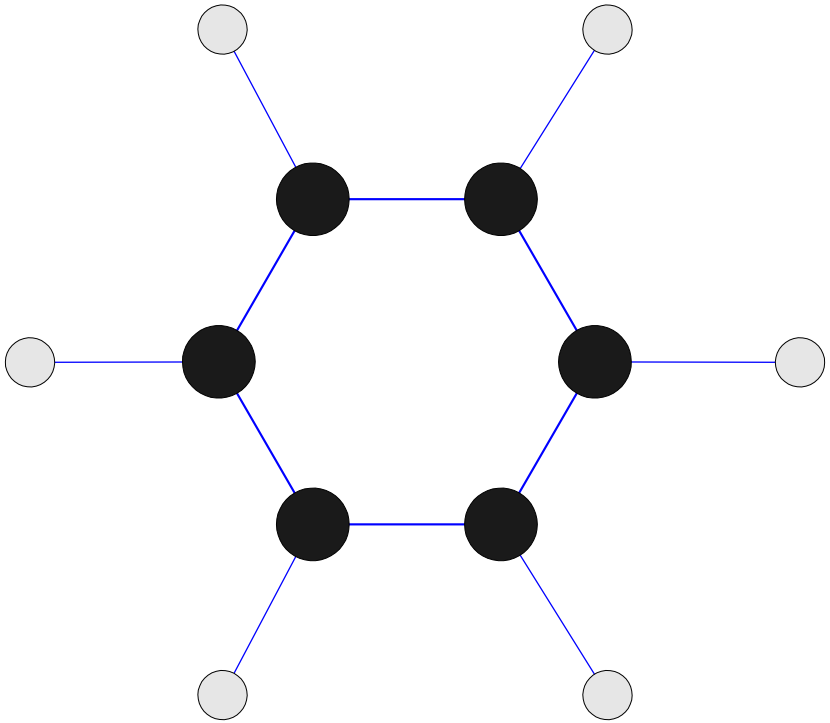
- Examples: Generic graph and HTML Links




```
<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
```

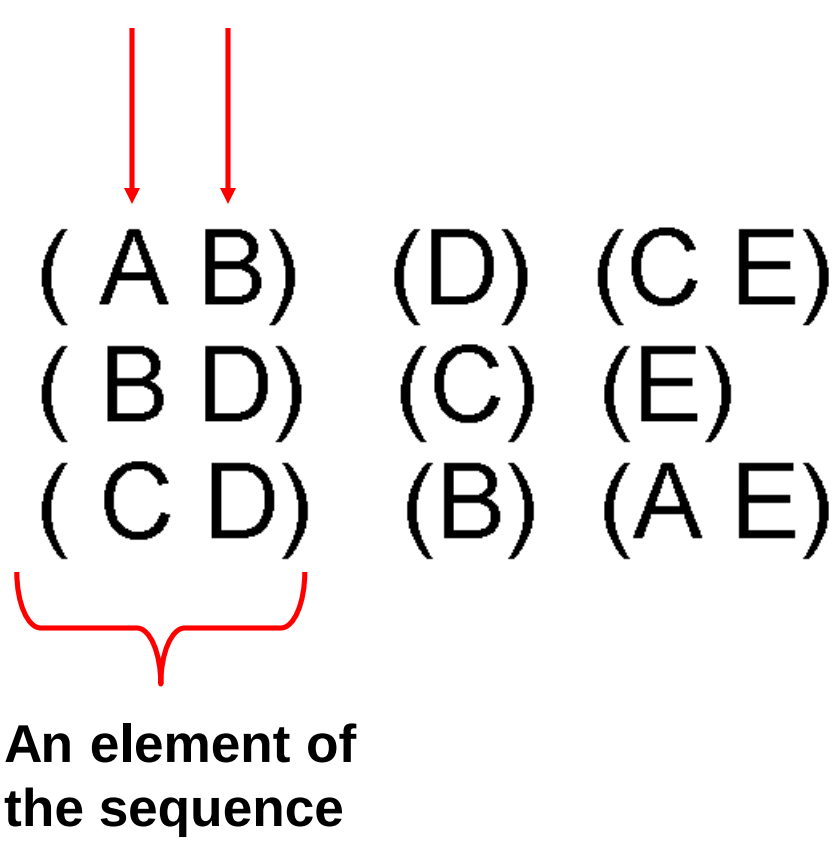
Chemical Data

- Benzene Molecule苯分子: C_6H_6



4. Ordered Data 有序数据

- 某些数据类型，属性具有涉及时间或空间序的联系。
- **Sequences of transactions 时序数据，时间数据（temporal data）**



Ordered Data

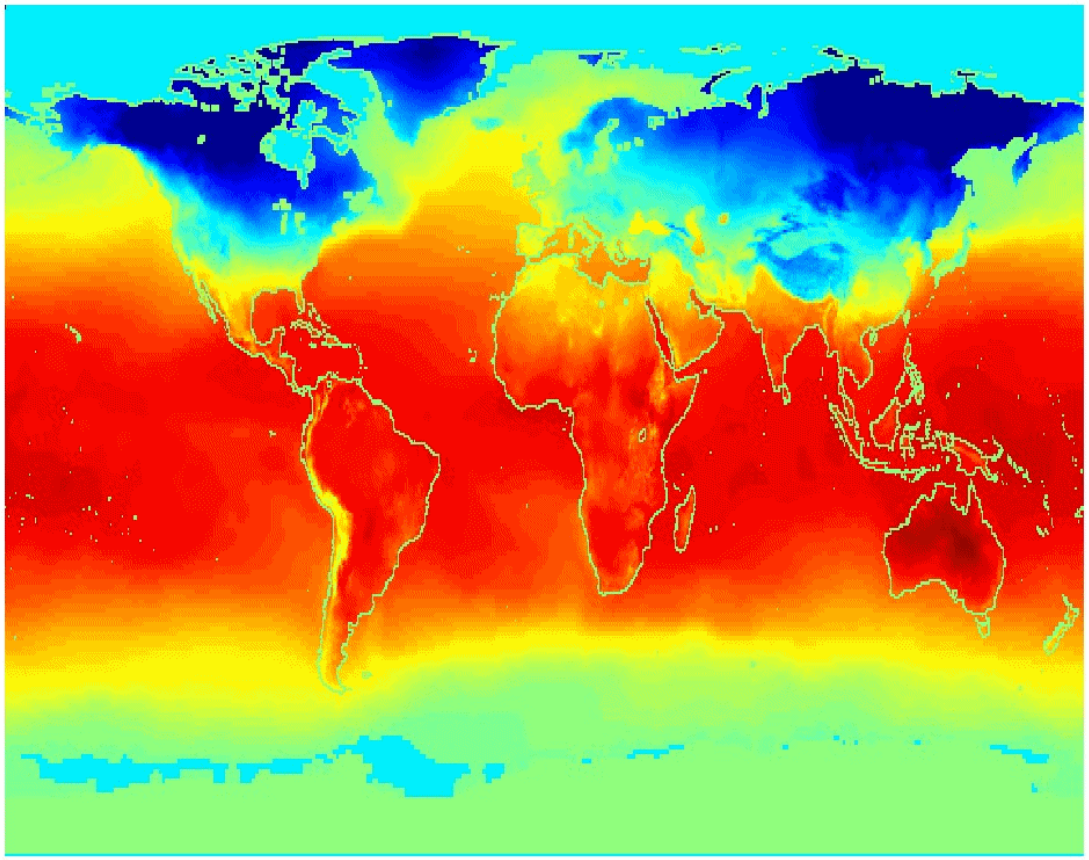
- 数据集是各个实体的序列，除了没有时间戳之外，与时序数据类似，只是有序序列考虑项的位置。
  - Genomic sequence data 基因组序列数据

GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG

Ordered Data

- 空间数据，具有空间属性，如位置或区域
  - **Spatio-Temporal Data** 空间温度数据

Average Monthly Temperature of land and ocean



2.2 Data Quality

数据挖掘使用的数据常常是为其他用途收集的，或者在收集时未明确其目的。因此，数据挖掘常常不能“在数据源头控制质量”。相比之下，统计学的实验设计或调查往往其数据质量都达到了一定的要求。

由于无法避免数据质量问题，因此数据挖掘着眼于两个方面：

（1）数据质量问题的检测和纠正，通常称作**数据清理**（data cleaning）。

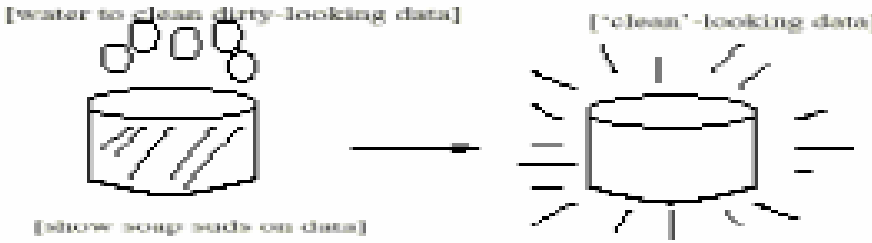
（2）使用可以容忍低质量数据的算法。

数据预处理的流程

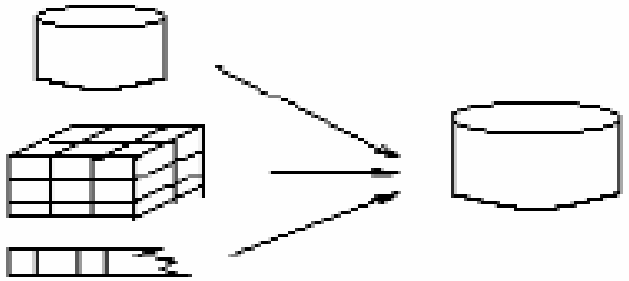
- 数据清理
  - 填充缺失值, 识别/去除离群点, 光滑噪音, 并纠正数据中的不一致
- 数据集成
  - 多个数据库, 数据立方体, 或文件的集成
- 数据变换
  - 规范化和聚集
- 数据归约
  - 得到数据的归约表示, 它小得多, 但产生相同或类似的分析结果：维度规约、数值规约、数据压缩
- 数据离散化和概念分层

数据预处理的流程

Data Cleaning



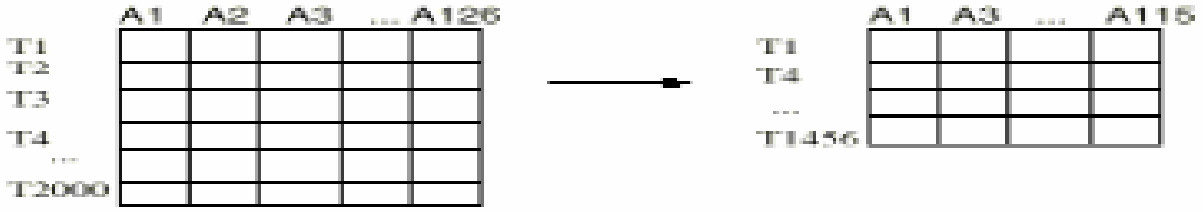
Data Integration



Data Transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



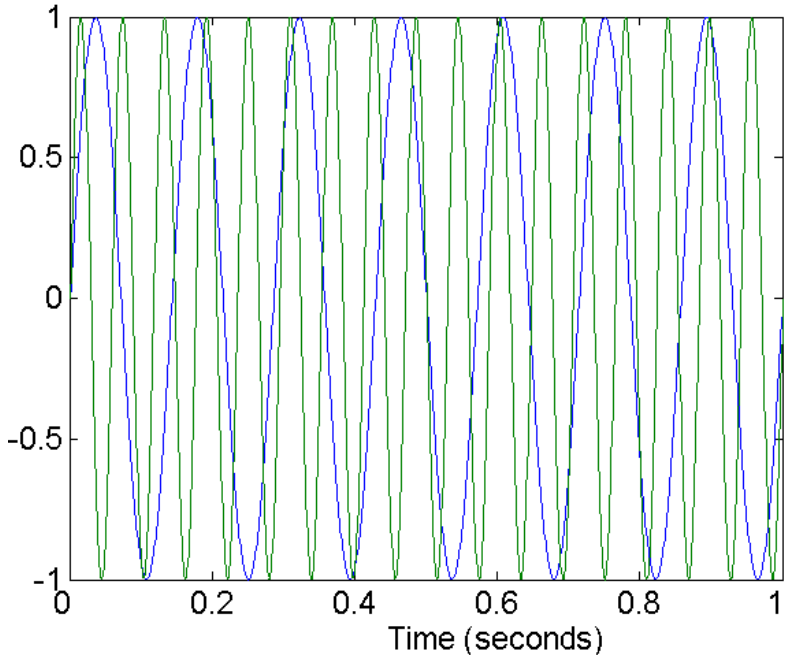
Data Quality

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

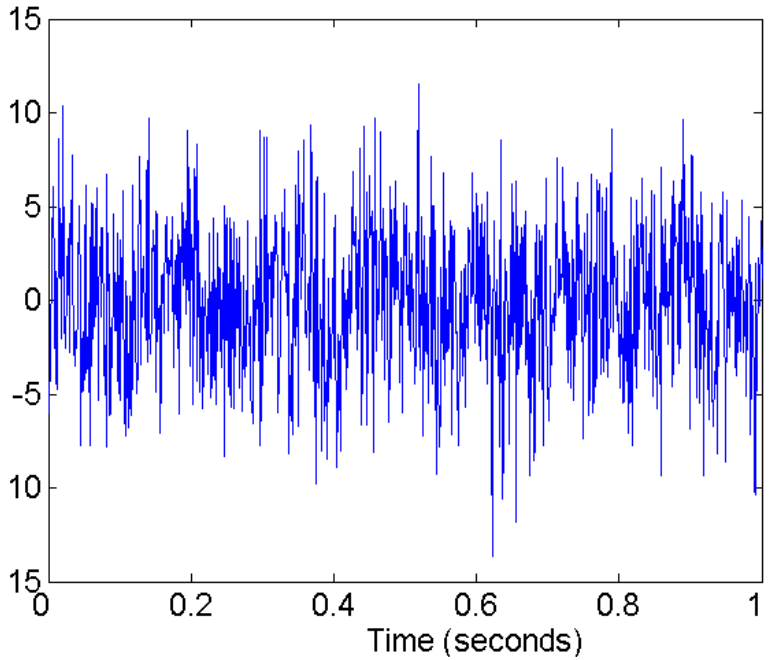
- Examples of data quality problems:
  - Measurement error 测量误差
  - Noise and outliers 噪声和离群点
  - missing values 缺失值
  - duplicate data 重复数据

Noise 噪声

- 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



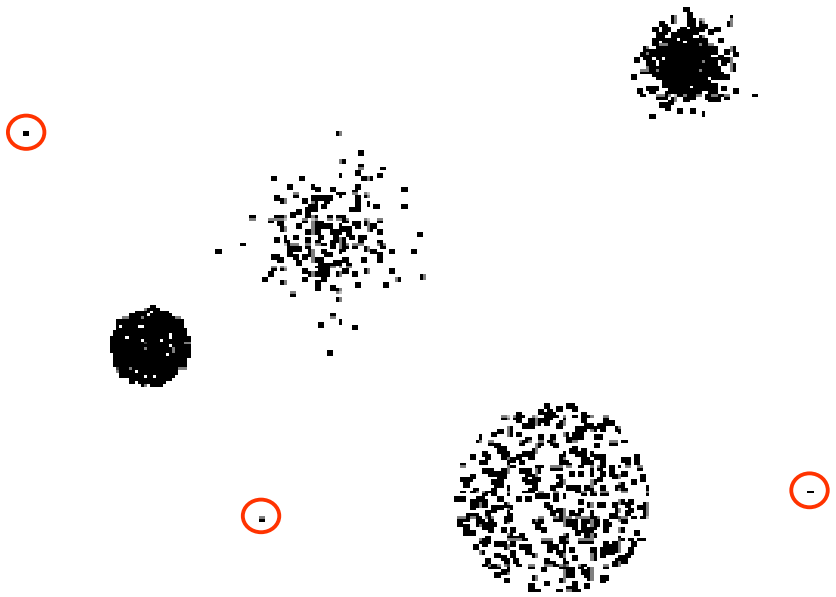
Two Sine Waves



Two Sine Waves + Noise

Outliers 离群点

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set



去噪技术

- 分箱Binning method:
  - 排序数据，分布到等频/等宽的箱/桶中
  - 箱均值光滑、箱中位数光滑、箱边界光滑, etc.
- 聚类Clustering
  - 检测和去除 离群点/孤立点 outliers
- 回归 Regression
  - 回归函数拟合数据
- 计算机和人工检查相结合
  - 人工检查可疑值 (e.g., deal with possible outliers)

分箱：简单的离散化方法

- 分箱：通过考察数据的近邻（即周围的值）来光滑有序数据的值。由于考察近邻的值，因此是一种局部光滑技术。

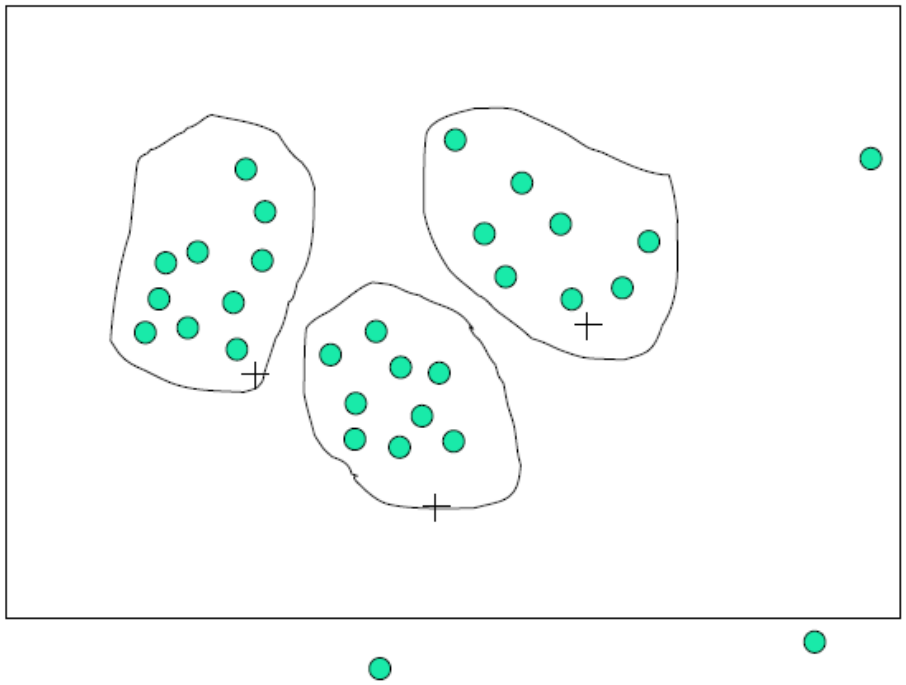
- 等宽度Equal-width (distance) 剖分:
  - 分成大小相等的n个区间: 均匀网格 uniform grid
  - 若*A*和*B*是 属性的最低和最高取值, 区间宽度为: $W = (B-A)/N$.
  - 孤立点可能占据重要影响 may dominate presentation
  - 倾斜的数据处理不好.
- 等频剖分 (frequency) /等深equi-depth :
  - 分成n个区间, 每一个含近似相同数目的样本
  - Good data scaling
  - 类别属性可能会非常棘手.

Binning Methods for Data Smoothing

* Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
* Partition into (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

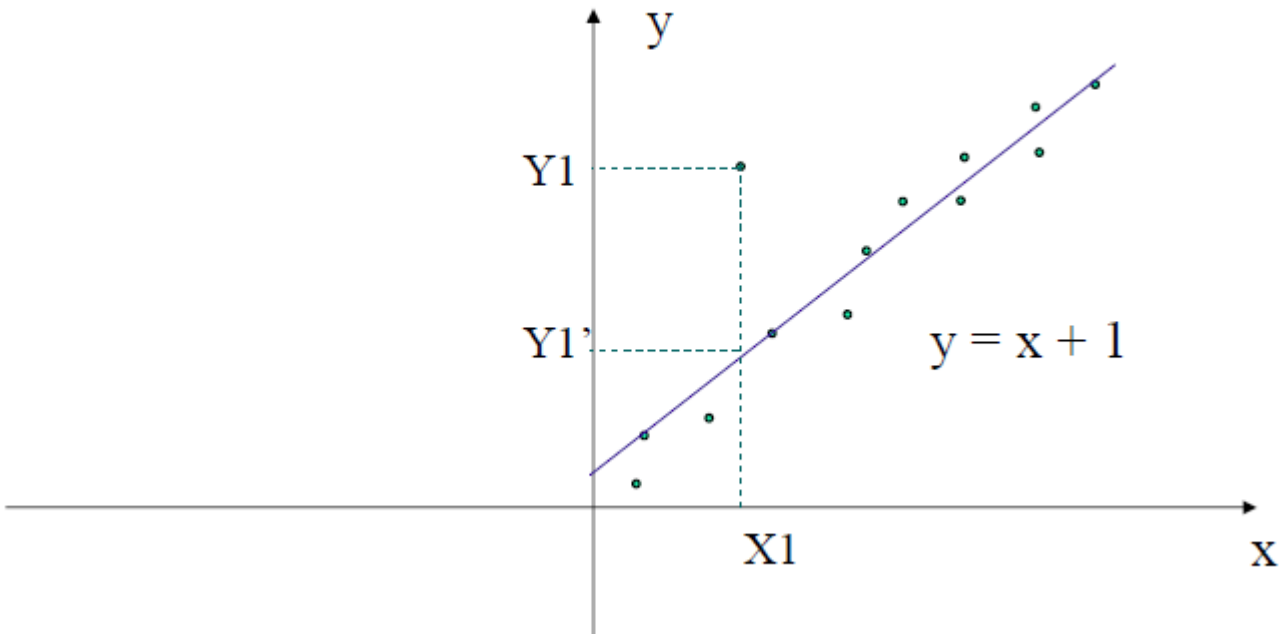
聚类: clustering

- 通过聚类检测离群点，将类似的值组织成群或簇。直观地，落在簇集合之外的值视为离群点。



回归: Regression

- 用一个函数（如回归函数）拟合数据来光滑数据。
- 分为线性回归和多元回归。



Missing Values 缺失值

- Reasons for missing values
  - Information is not collected
    (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases
    (e.g., annual income is not applicable to children)

- Handling missing values
  - Eliminate Data Objects
  - Estimate Missing Values
  - Ignore the Missing Value During Analysis
  - Replace with all possible values (weighted by their probabilities)

如何处理缺失数据?

- 为属性填上丢失的值:
  - 忽略元组
  - 人工填写缺失值
  - 使用一个全局常量填充缺失值
  - 使用属性的均值填充缺失值
  - 使用与给定元组属同一类的所有样本你属性均值
  - 使用最可能的值填充缺失值

如何处理缺失数据?

- 忽略元组
  - 当缺少类标号时，通常这样做(假定挖掘任务涉及分类)。除非元组有多个属性缺少值，否则该方法不是很有效；
- 人工填写缺失值
  - 乏味+费时+不可行

如何处理缺失数据?

- 自动填充
  - 使用一个全局常量填充缺失值
    - 将缺失的属性值用同一个常数，如unknown,或-∞。
    - 可能会形成一个新的class，这个新class具有相同的填充常数值；
    - 虽然简单，却不可靠；
  - 使用属性的均值填充缺失值
    - 例如用平均薪水值填写 某个元组缺失的salary属性值；
  - 使用与给定元组属同一类的所有样本的属性均值
    - 先给元组分类，用不同类别的均值填充缺失的本类属性值；——妙极了！
  - 使用最可能的值填充缺失值
    - 基于推理的方法，如Bayes，回归，决策树等推理预测；

Duplicate Data 重复数据

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeous sources
- Examples:
  - Same person with multiple email addresses
- Data cleaning
  - Process of dealing with duplicate data issues

重复数据处理：模式集成

- 来自多个信息源的现实世界的等价实体如何匹配？
  - 实体识别问题 Entiry identification problem:
    - A数据库中的customer_id与B数据库中的cust_number是相同的属性？
    - Bill Clinton = William Clinton?
  - 集成不同来源的元数据
  - 冲突数据值的检测和解决
    - 对真实世界的实体，其不同来源的属性值可能不同
    - 原因：不同的表示，不同的尺度，如公制 vs英制

冗余数据的处理

- 冗余数据Redundant data （集成多个数据库时出现）
  - 目标识别：同一个属性在不同的数据库中有不同的名称
  - 衍生数据：一个属性值可由其他表的属性推导出, e.g., 年收入

  - 小心的集成多个来源的数据可以帮助降低和避免结果数据集中的冗余和不一致，提高数据挖掘的速度和质量

检测冗余数据：相关分析

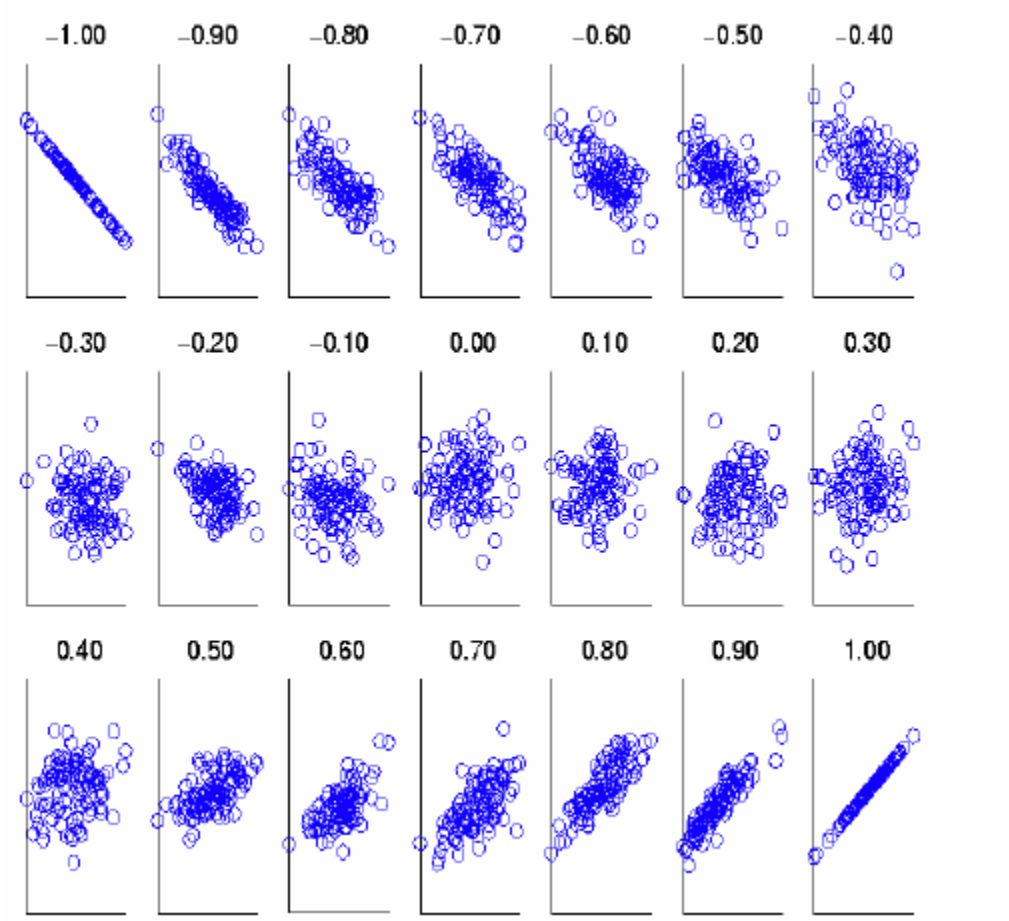
- 相关分析 *correlation analysis* /协方差分析 *covariance analysis*
  - 可用于检测冗余数据
  - **Correlation coefficient (also called Pearson's product moment coefficient)**
  - 相关系数（皮尔逊相关系数）

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \overline{A})(b_i - \overline{B})}{(n-1)\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\overline{A}\,\overline{B}}{(n-1)\sigma_A\sigma_B}$$

n元组个数, $\overline{A}$ 和 $\overline{B}$ 属性A和B上的平均值, σ_A and σ_B分别为各自标准差， $\Sigma(a_i b_i)$ is the AB叉积 cross-product之和.

  - **If $r_{A,B} > 0$, A and B 整相关 (A's values increase as B's).** 值越大相关程度越高.
  - $r_{A,B} = 0$: 不相关; $r_{AB} < 0$: 负相关

相关性的视觉评价



Scatter plots showing the similarity from –1 to 1.

相关 (线形关系)

- 相关测量的是对象间的线性关系
- **To compute correlation, we standardize data objects, A and B, and then take their dot product**

$$a'_k = (a_k - mean(A)) / std(A)$$

$$b'_k = (b_k - mean(B)) / std(B)$$

$$correlation(A, B) = A' \bullet B'$$

协方差Covariance (Numeric Data)

- **Covariance is similar to correlation**

$$Cov(A,B) = E((A-\bar{A})(B-\bar{B})) = \frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n}$$

Correlation coefficient:

$$r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$$

n元组个数, $\overline{A}$ 和 $\overline{B}$ 属性A和B上的平均值, σ A and σ B分别为各自标准差.

- 正covariance: **If $Cov_{A,B} > 0$, 则A 和B 同时倾向于大于期望值.**
- 负covariance: **If $Cov_{A,B} < 0$，则如果 A〉其期望值, B is likely to be smaller than its expected value.**
- Independence: **$Cov_{A,B} = 0$ but the converse is not true:**
  - **Some pairs of random variables may have a covariance of 0 but are not independent. Only under some additional assumptions (e.g., the data follow multivariate normal distributions) does a covariance of 0 imply independence**

Co-Variance: An Example

$$Cov(A,B) = E((A-\bar{A})(B-\bar{B})) = \frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n}$$

- **It can be simplified in computation as**

$$Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}$$

- 设两个股票 **A** 和**B** 一周内值如下 **(2, 5), (3, 8), (5, 10), (4, 11), (6, 14).**
- 问：如果股票是由同行业趋势的影响，它们的价格将一起上升或下降？
  - **E(A) = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4**
  - **E(B) = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6**
  - **Cov(A,B) = (2×5+3×8+5×10+4×11+6×14)/5 − 4 × 9.6 = 4**
- **Thus, A and B rise together since Cov(A, B) > 0.**

相关分析 (名义数据Nominal Data)

- X2 (chi-square) test 开方检验
  - σ ij是(a_i,b_j)的观测频度（实际计数）
  - eij是(a_i,b_j)的期望频度
  - N数据元组的个数

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(\sigma_{ij} - e_{ij})^2}{e_{ij}}$$

属性 \ A	a_1	a_2	i→	a_c
b_1				
b_2				
j↓				
b_r				

B

(A=ai,B=bj)

$$e_{ij} = \frac{count(A = a_i) * count(B = b_j)}{N}$$

- X2 值越大,相关的可能越大
- 对 X2 值贡献最大的项，其实际值与期望值相差最大的相
- 相关不意味着因果关系

Chi-Square 卡方值计算: 例子

	Play chess	Not play chess	Sum (row)
看小说	250(90)	200(360)	450
不看小说	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

$$e_{11} = \frac{count(\text{看小说}) * count(\text{下棋})}{N} = \frac{450 * 300}{1500} = 90$$

- **X^2 (chi-square) 计算(括号中的值为期望计值，由两个类别的分布数据计算得到)**

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- **结果表明like_fiction 和play_chess 关联**

2.3 Data Preprocessing 数据预处理

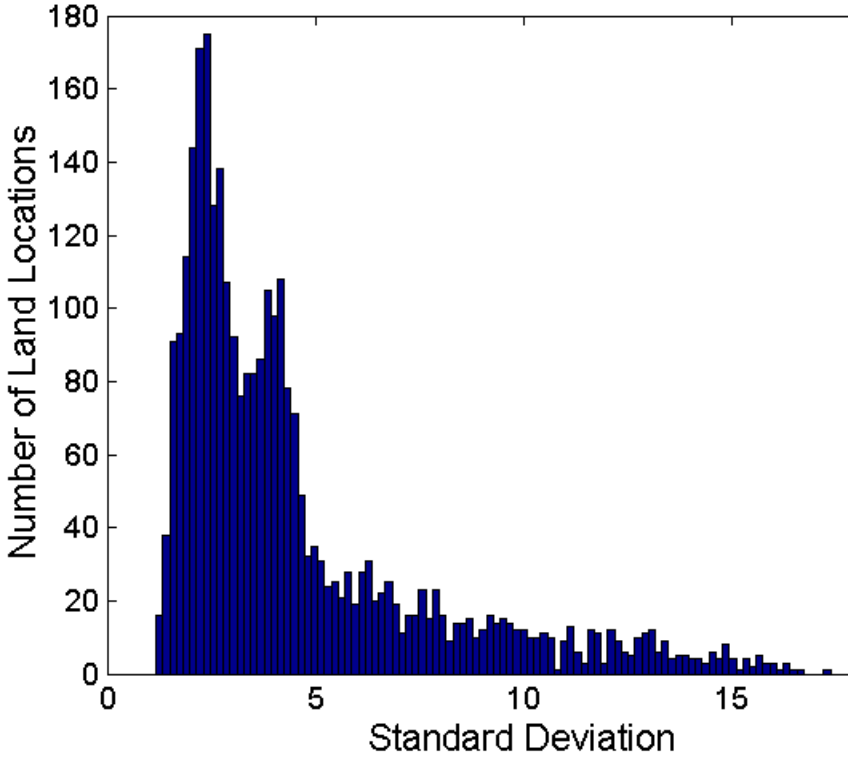
- Aggregation 聚集
- Sampling 抽样
- Dimensionality Reduction 维度归约
- Feature subset selection 特征子集选择
- Feature creation 特征创建
- Discretization and Binarization 离散和二元化
- Attribute/variable Transformation 变量变换

2.3.1 Aggregation 聚集

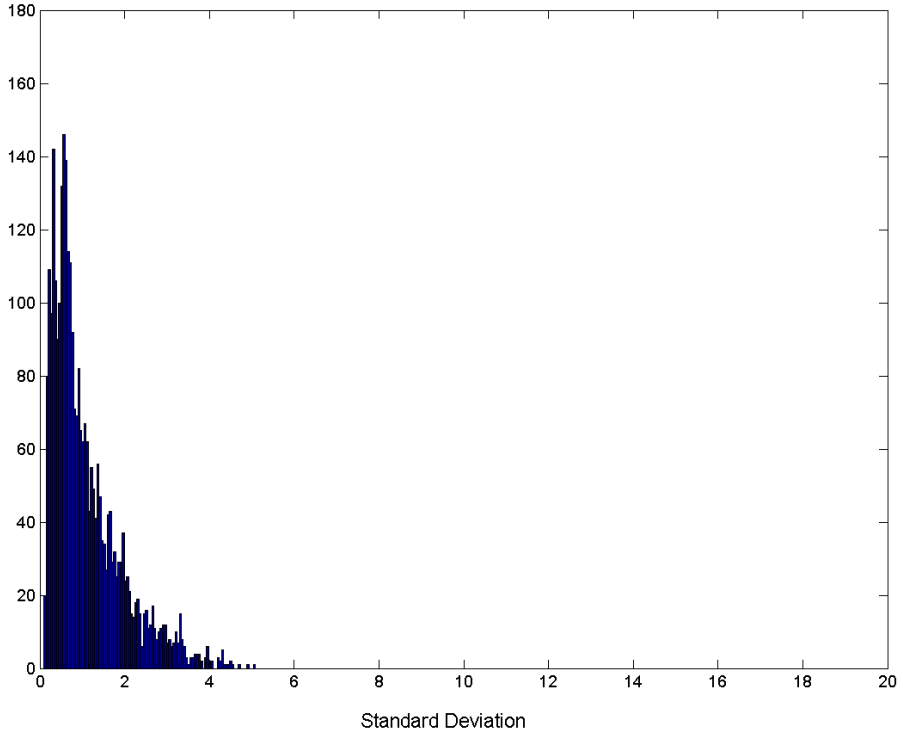
- 少就是多，聚集将两个或多个对象合并成单个对象
- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc
  - More “stable” data
    - Aggregated data tends to have less variability
- Drawback
  - Missing valuale information

Aggregation

Variation of Precipitation in Australia



Standard Deviation of Average Monthly Precipitation



Standard Deviation of Average Yearly Precipitation

2.3.2 Sampling 抽样

- **Sampling 是一种选择数据对象子集进行分析的常用方法.**
- **Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming.**
- **Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.**

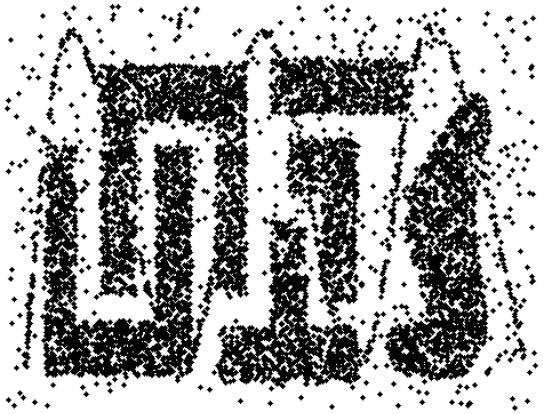
Sampling …

- The key principle for effective sampling is the following:
  - **如果样本具有代表性，则使用样本与使用整个数据集的效果几乎相同.**
  - 样本具有代表性，前提是它近似地具有与原有数据集相同的（感兴趣）的性质。

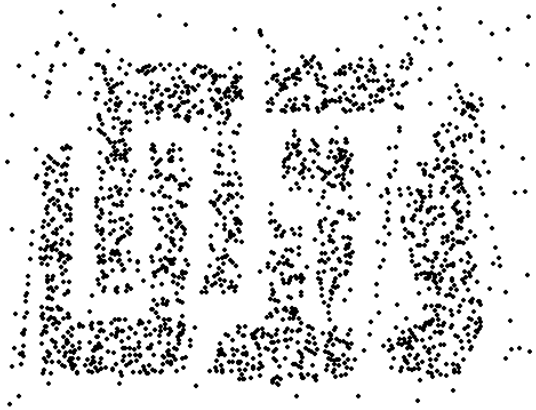
1. Types of Sampling

- Simple Random Sampling 简单随机抽样
  - 选取任何特定项的概率相同
- Sampling without replacement 无放回抽样
  - As each item is selected, it is removed from the population
- Sampling with replacement 有放回抽样
  - Objects are not removed from the population as they are selected for the sample.
    - In sampling with replacement, the same object can be picked up more than once
- Stratified sampling 分层抽样
  - 当总体由不同的对象组成，每种类型的对象数量差别很大时，简单随机抽样不能充分代表不太频繁出现的对象类型。
  - Split the data into several partitions; then draw random samples from each partition

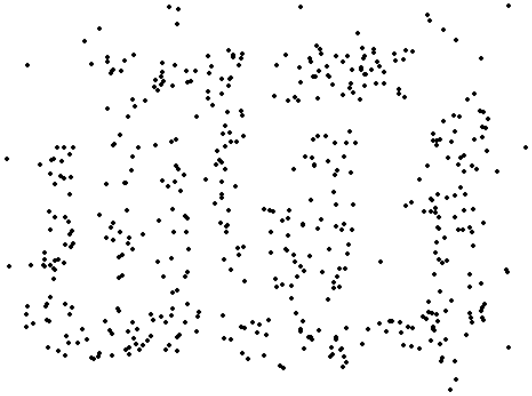
Sample Size and loss of information



8000 points



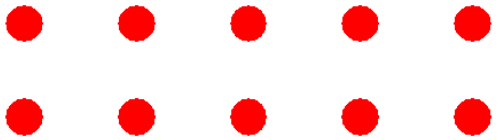
2000 Points

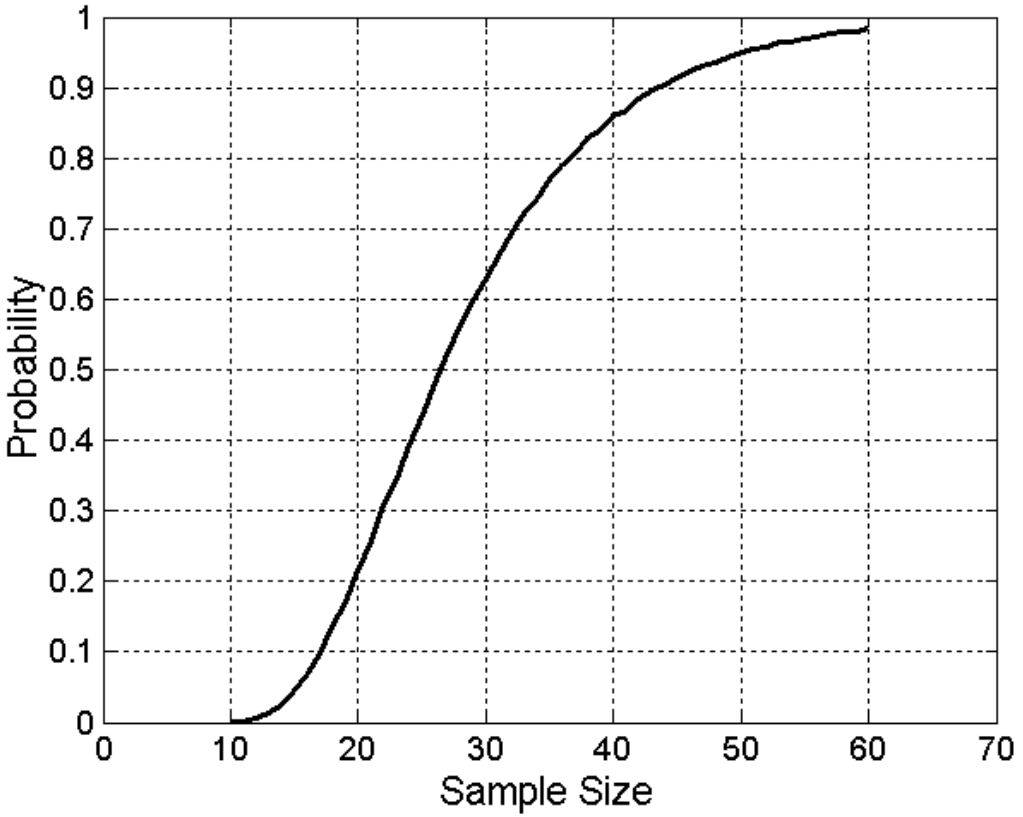


500 Points

Sample Size

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





维度归约可以删除不相关的特征，并降低噪声

避免维度灾难

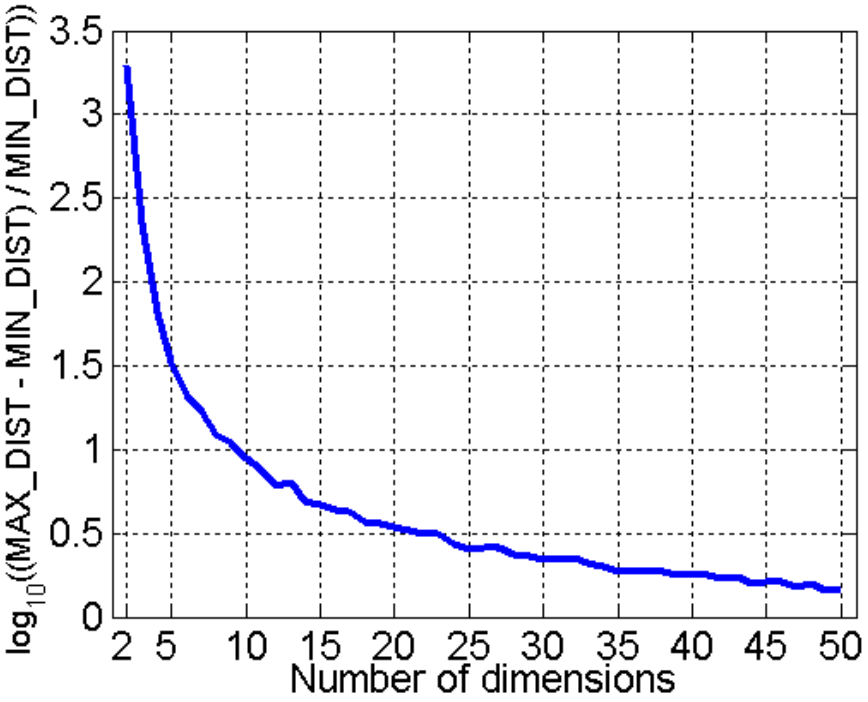
使得模型更加易于理解

更加可视化

2.3.3 DIMENSIONALITY REDUCTION 维度归约

1. Curse of Dimensionality 维度灾难

- 当数据维度的增加时，数据在所占据的空间中越来越稀疏；
  - 对于分类，则可能意味着没有足够的数据对象来创建差异性模型；
  - 对于聚类，点之间的密度和距离（对于聚类非常重要）都将失去意义。



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

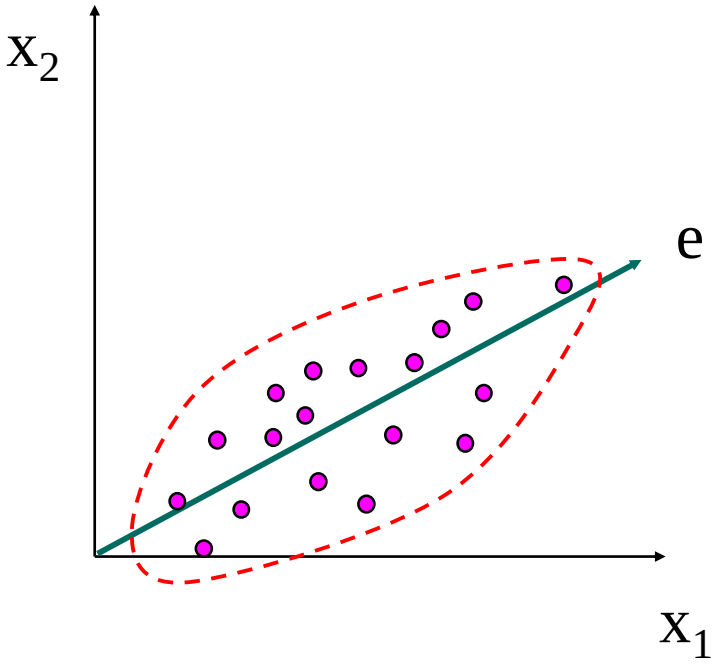
Dimensionality Reduction

- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise

- Techniques
  - Principle Component Analysis 主成成分析
  - Singular Value Decomposition 奇异值分解
  - Others: supervised and non-linear techniques

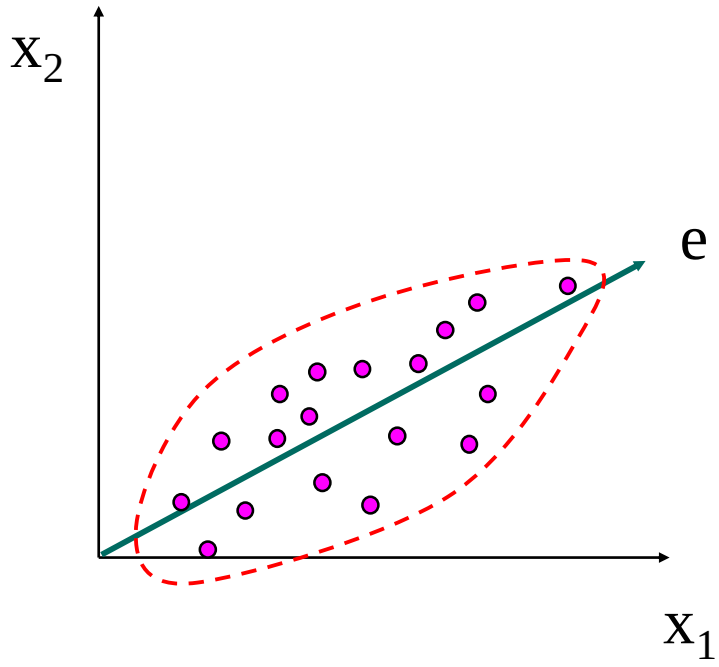
Dimensionality Reduction: PCA 主成成分分析

- 适用于连续属性的线性代数技术，它找出新的属性（主成分），这些属性是原属性的线性组合，是相互正交的，并且捕获了数据的最大变异。



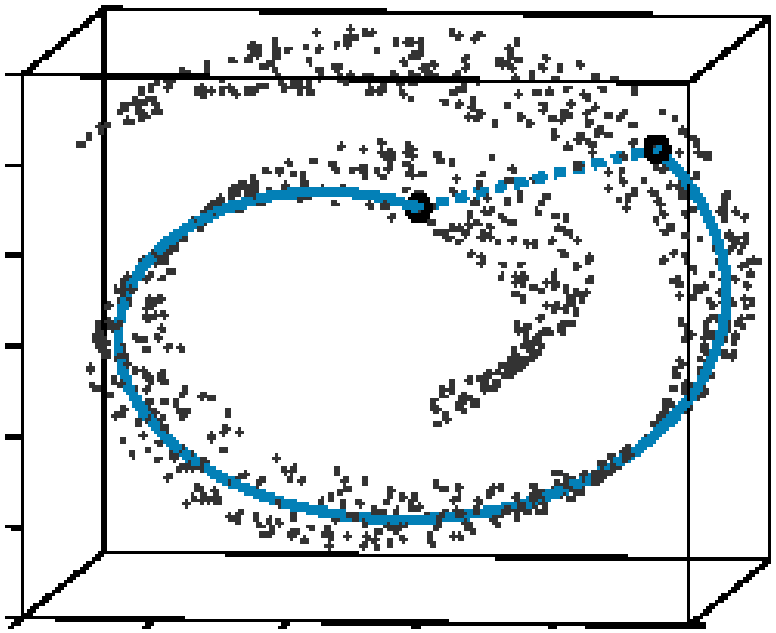
Dimensionality Reduction: PCA

- Find the eigenvectors of the covariance matrix
- The eigenvectors define the new space



Dimensionality Reduction: ISOMAP

By: Tenenbaum, de Silva, Langford (2000)



- Construct a neighbourhood graph
- For each pair of points in the graph, compute the shortest path distances – geodesic distances

Dimensionality Reduction: PCA

Dimensions = 206

2.3.4 Feature Subset Selection 特征子集选择

- Another way to reduce dimensionality of data
- Redundant features 冗余特征
  - duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid
- Irrelevant features 不相关特征
  - contain no information that is useful for the data mining task at hand
  - Example: students' ID is often irrelevant to the task of predicting students' GPA

Feature Subset Selection

- Techniques:
  - Brute-force approch: 蛮力法
    - Try all possible feature subsets as input to data mining algorithm
  - Embedded approaches: 嵌入方法
    - Feature selection occurs naturally as part of the data mining algorithm
  - Filter approaches: 过滤方法
    - Features are selected before data mining algorithm is run
  - Wrapper approaches: 包装方法
    - Use the data mining algorithm as a black box to find best subset of attributes

Feature Subset Selection：特征加权

- 保留或删除特征的方法

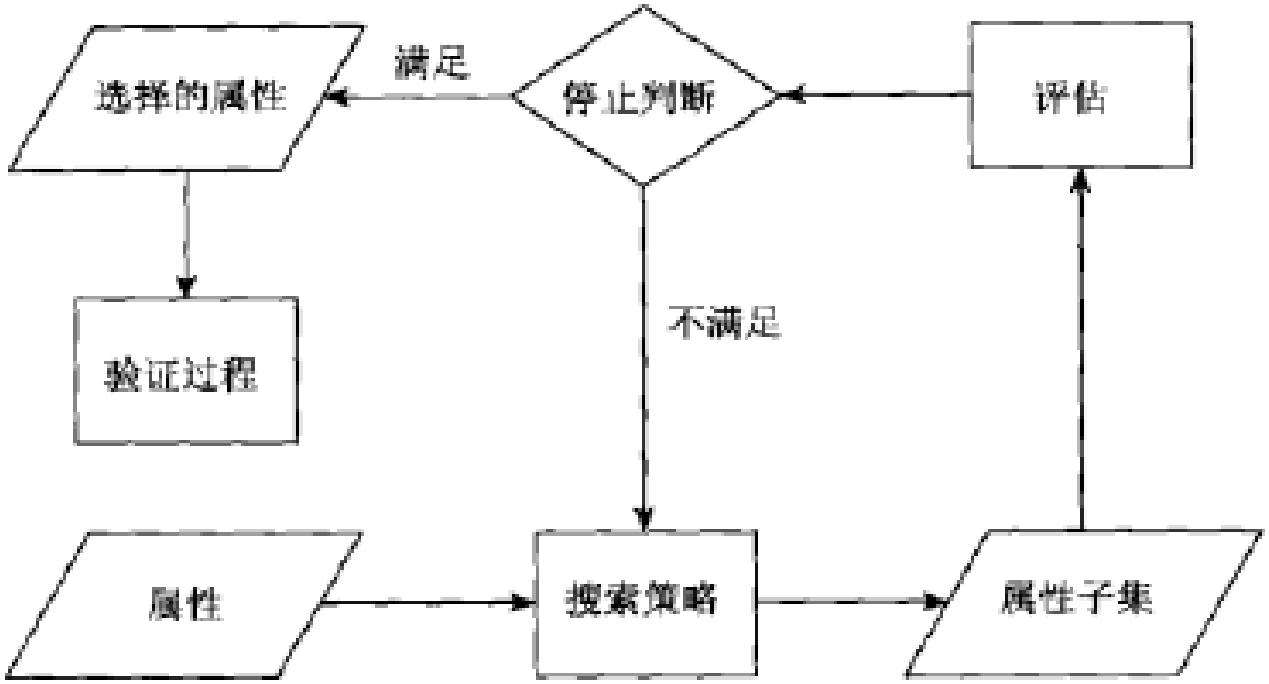


图 2-11　特征子集选择过程流程图

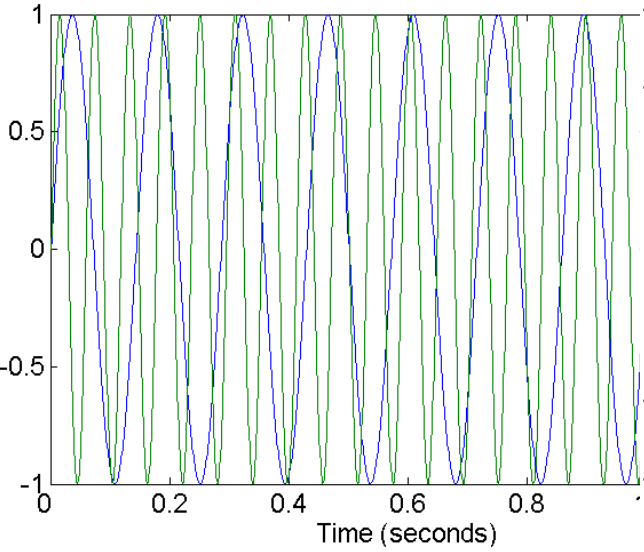
2.3.5 Feature Creation 特征创建

- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes

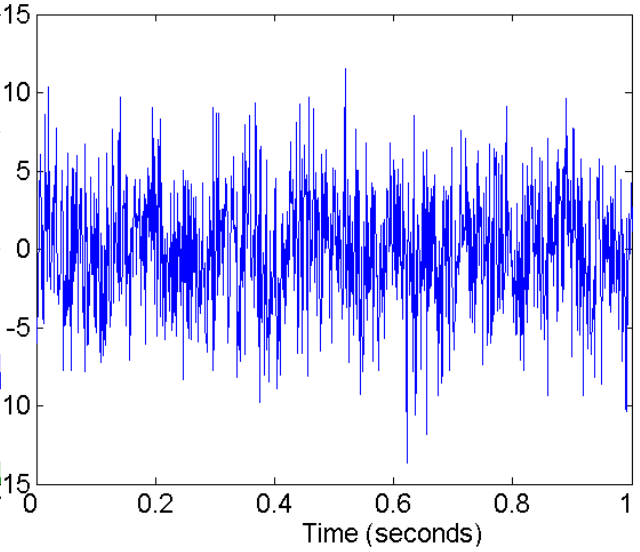
- Three general methodologies:
  - Feature Extraction 特征提取
    - domain-specific
  - Mapping Data to New Space 映射数据到新的空间
  - Feature Construction 特征构造
    - combining features

Mapping Data to a New Space

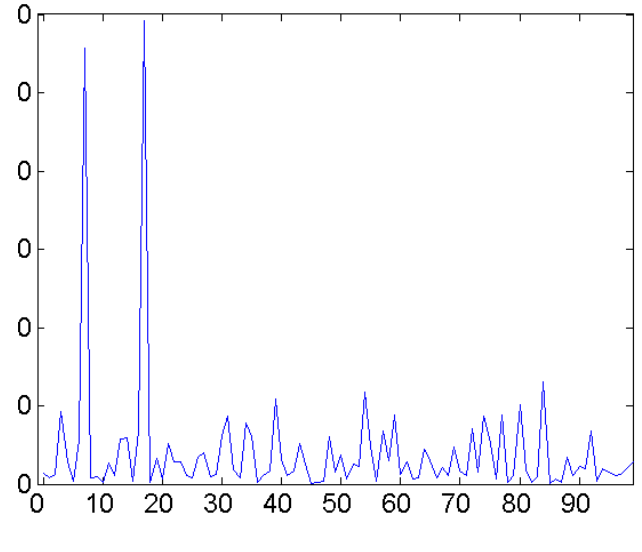
- Fourier transform
- Wavelet transform



Two Sine Waves



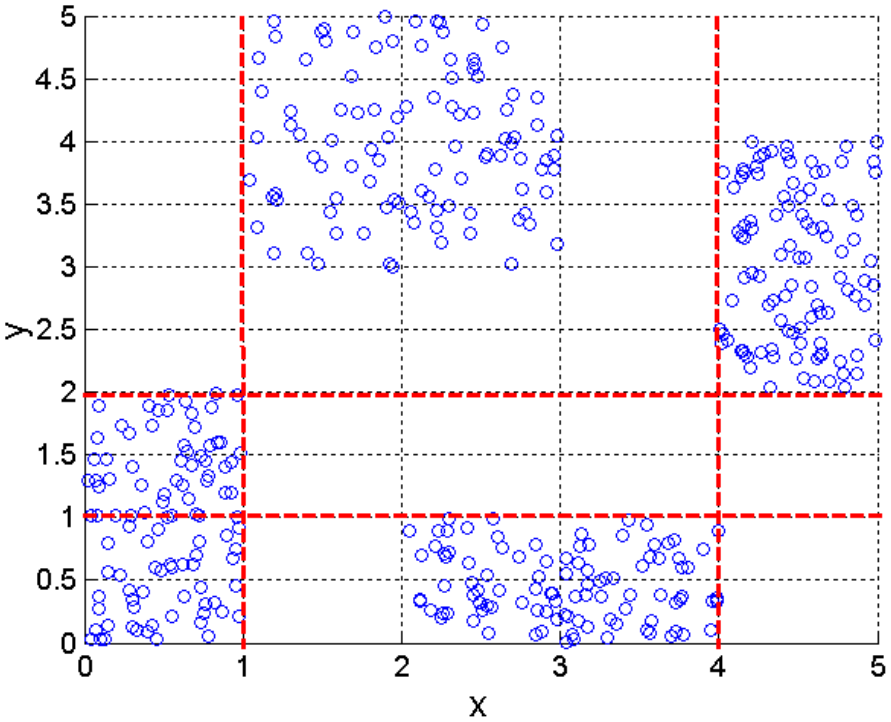
Two Sine Waves + Noise



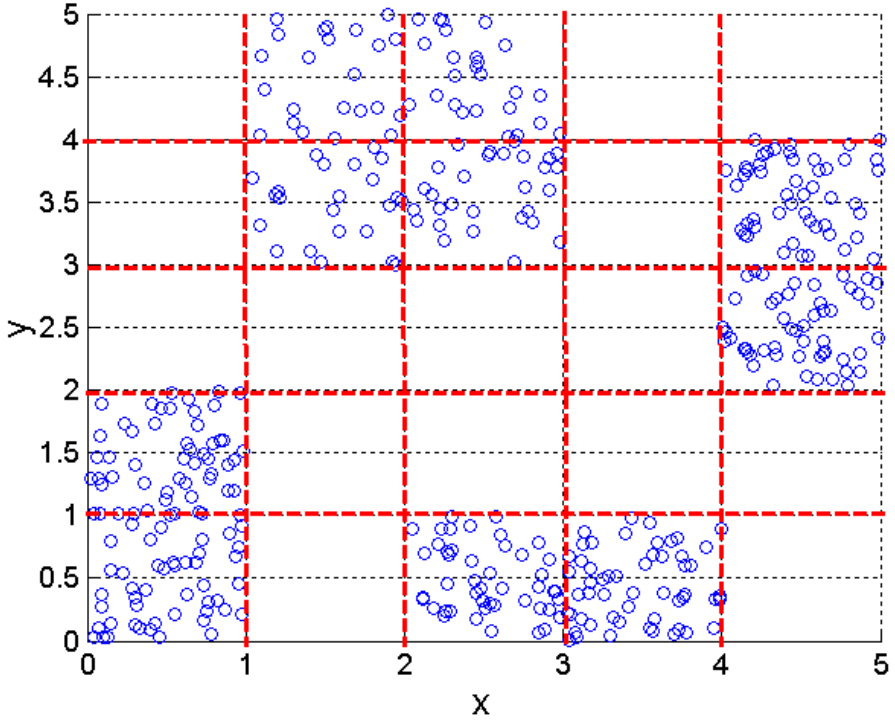
Frequency

Discretization Using Class Labels

Entropy based approach

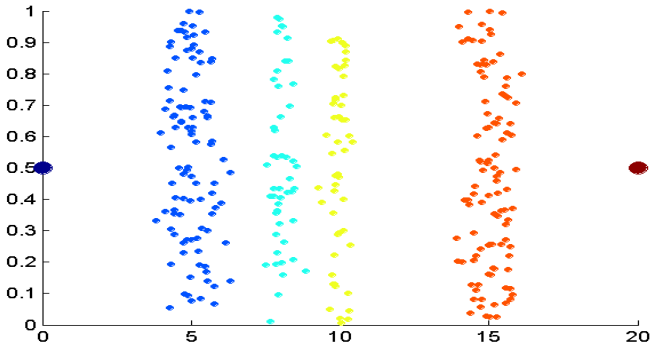


3 categories for both x and y

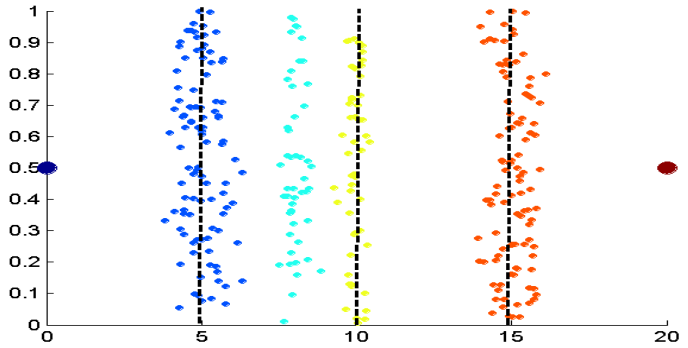


5 categories for both x and y

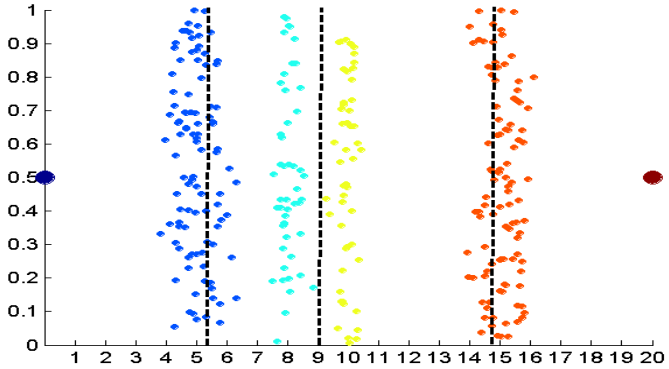
Discretization Without Using Class Labels



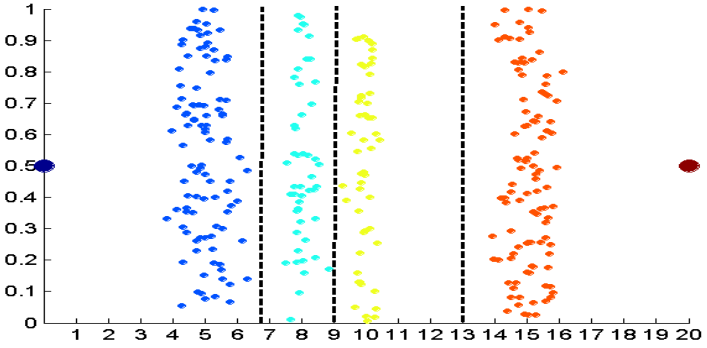
Data



Equal interval width



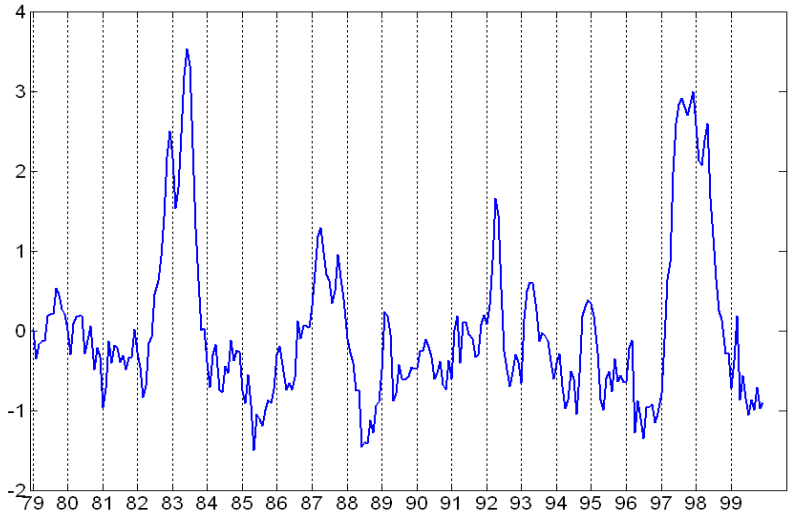
Equal frequency



K-means

Attribute Transformation 变量变换

- A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values
  - Simple functions: x^k, log(x), e^x, |x|
  - Standardization and Normalization



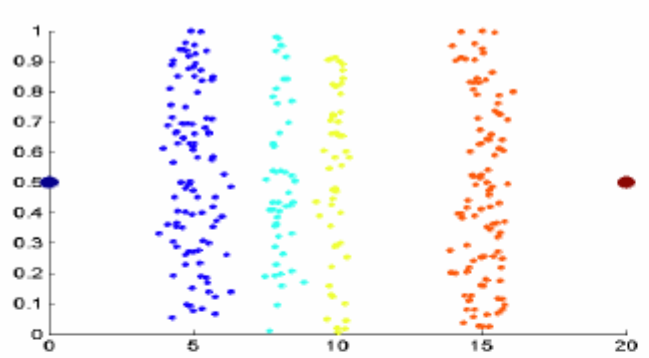
2.3.6 离散化和概念分层

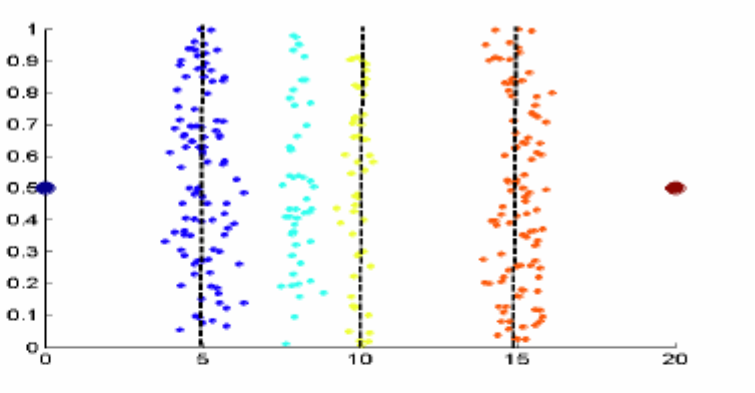
- 三种类型属性:
  - 名义 — values from an unordered set， color, profession
  - 顺序数 — values from an ordered set , e.g., military or academic rank
  - 连续 — real numbers
- 离散化 Discretization: 把连续属性的区域分成区间
  - 区间标号可以代替实际数据值
  - 利用离散化减少数据量
  - 有监督 vs. 无监督：是否使用类的信息
  - 某个属性上可以递归离散化
  - 分裂 Split (top-down) vs. 合并merge (bottom-up)
    - 自顶向下：由一个/几个点开始递归划分整个属性区间
- 递归离散化属性，产生属性值分层/多分辨率划分：概念分层

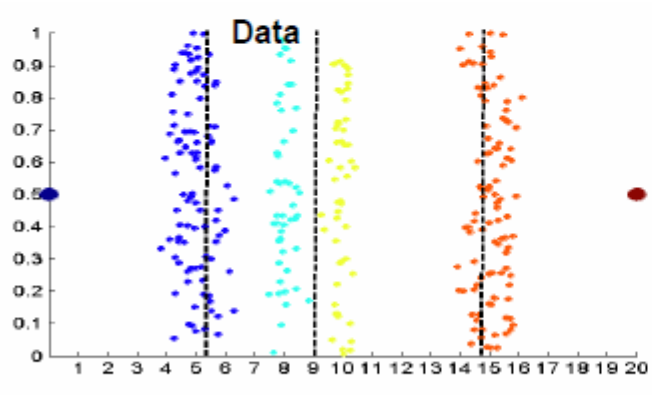
1.数值数据离散化/概念分层

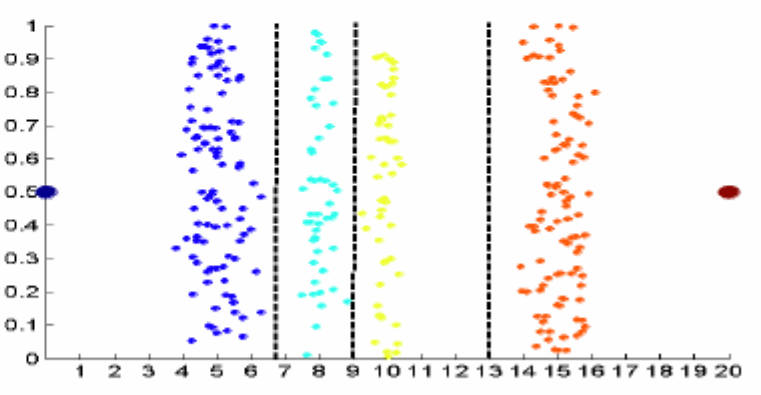
- 分箱 Binning(Top-down split, unsupervised)
- 直方图（Top-down split, unsupervised）
- 聚类 (unsupervised, top-down split or bottom-up merge)
- 基于χ^2 分析的区间合并(unsupervised, bottom-up merge)
- 基于熵 Entropy-based discretization
- 根据自然划分

不用类别(Binning vs. Clustering)









Equal frequency (binning)

K-means clustering leads to better results

基于熵Entropy的离散化

给定一个数据元组的集合 S，基于熵对 A 离散化的方法如下：

1. A 的每个值可以认为是一个潜在的区间边界
2. 选择的阈值 T 使其后划分得到的信息增益最大

$$I(S,T)=\frac{|S_1|}{|S|}Ent(S_1)+\frac{|S_2|}{|S|}Ent(S_2)$$

其中，S_1 和 S_2 分别对应于 S 中满足条件 $A<T$ 和 $A\geq T$ 的样本

$$Ent(S_1)=-\sum_{i=1}^{m}p_i\log_2(p_i)$$

其中，p_i 是类 i 在 S_1 中的概率，

等于 S_1 中类 i 的样本数除以 S_1 中的样本总数。

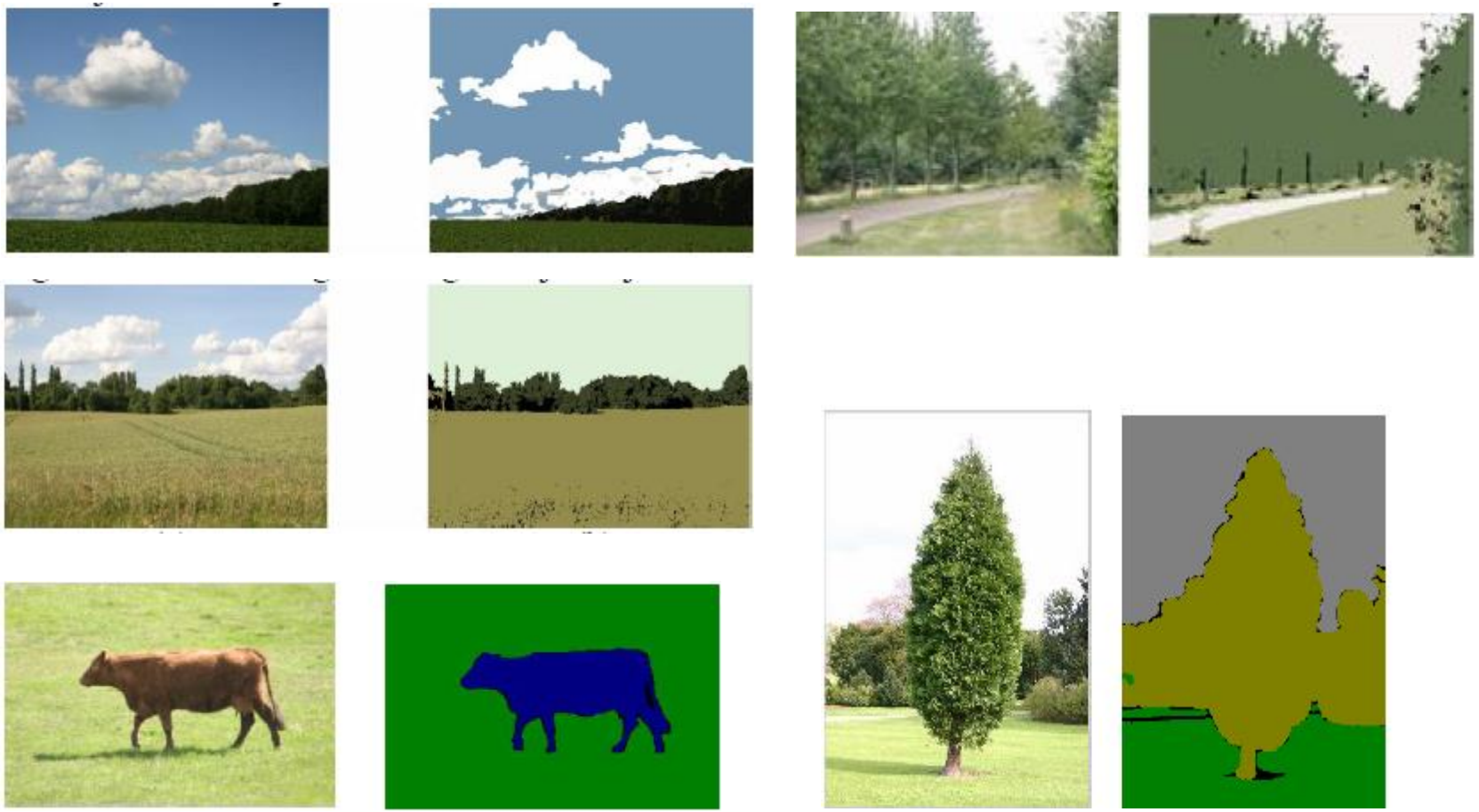
3. 直到满足某个终止条件 $Ent(S)-I(S,T)>\delta$

Chi-merge离散化

- **Chi-merge: χ^2-based discretization**
  - **有监督: use class information**
  - **自低向上: find the best neighboring intervals (具有相似的类别分布, i.e., low χ^2 values) to merge**
  - **递归地合并, until a predefined stopping condition**

EXAMPLES

1 图像的颜色数值归约



颜色的量化

一幅图像的颜色一般非常多,尤其是真彩色图像,因此直方图矢量的维数会非常多。如果对 HSV 空间进行适当的量化后再进行计算直方图,则计算量要少得多。在 HSV 空间中,H从 0 到 360 变化时,色调依次呈现为我们熟悉的红、橙、黄、绿、青、蓝、紫,而且每一种色调对应的 H分量的区域并不均匀。因此根据视觉对颜色的心理感觉,我们将 H、S、V三个分量按照人的颜色感知进行非等间隔的量化,将 H分量分为不等间隔的 7份。当 V很小时,视觉感觉为黑色,可以忽略 H的影响。最后量化结果为:

$$
H = \begin{cases} 0 & \text{if } h \in (330, 22) \\ 1 & \text{if } h \in [22, 45] \\ 2 & \text{if } h \in (45, 70) \\ 3 & \text{if } h \in [70, 155] \\ 4 & \text{if } h \in (155, 180) \\ 5 & \text{if } h \in [180, 272] \\ 6 & \text{if } h \in (272, 330] \end{cases}
$$

$$
S = \begin{cases} 0 & \text{if } s \in (0.1, 0.6) \\ 1 & \text{if } s \in [0.6, 1] \end{cases}
$$

$$
V = 0 \quad \text{if } (0.18, 1)
$$

颜色直方图

按照以上的量化级,把 3个颜色分量合成为一维特征矢量:

l=HQsQv+SQv+V。

式中:Qs和 Qv分别是分量 S、V的量

取 Qs=2,Qv=1,可得

$$l=\begin{cases} 2H+S+V & \text{if } s\in(0\ 1,1] \text{ and } v\in(0 \\ 14 & \text{if } v\in[0,0\ 18] \\ 15 & \text{if } s\in[0,0\ 1] \text{ and } v\in(0\ 18,0\ 65] \\ 16 & \text{if } s\in[0,0\ 1] \text{ and } v\in[0\ 65,0\ 85] \\ 17 & \text{if } s\in[0,0\ 1] \text{ and } v\in[0\ 85,1] \end{cases}$$

这样,就把 H, S, V 3个分量在一维矢量上分布开来。根据表达式可知,l的取值范围为 [0, 17]所以计算 l可以获得 18柄的一维直方图,这样量化可以有效地减少图像受光照的影响。

颜色直方图

彩色直方图 H定义为:

H = {(h[d1], h[d2], ……, h[dk], ……h[dn]) | $\sum_{k=1}^{n} h[dk]=1,\ 0\le h[dk]\le 1$。

其中 h[dk]表示第 k中色彩的像素在图像中出现的次数。

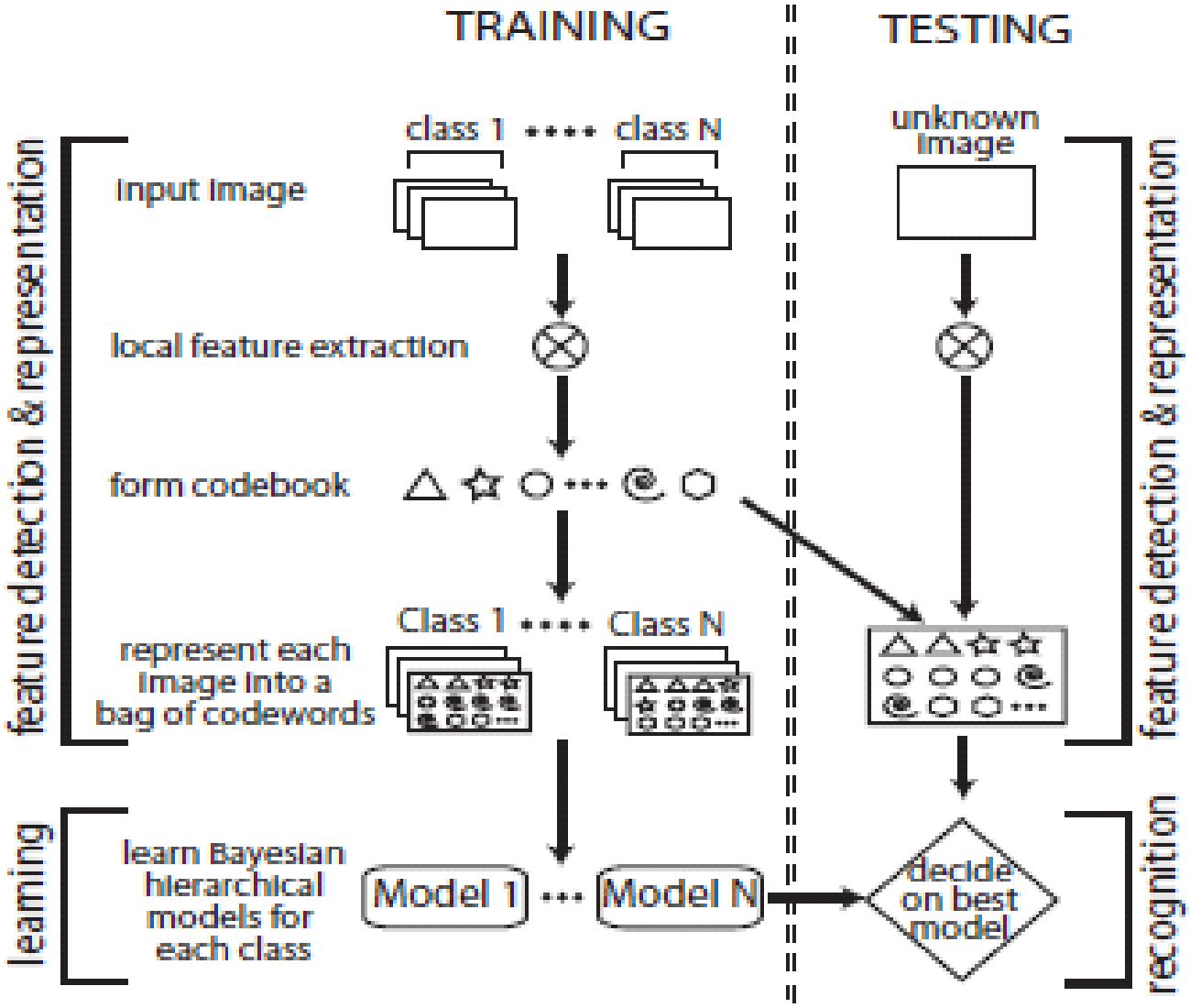
```
%*************************************************************************
%                    图像检索——提取颜色特征
%HSV空间颜色直方图(将RGB空间转化为HSV空间并进行非等间隔量化，
%将三个颜色分量表示成一维矢量，再计算其直方图作为颜色特征
%function : Hist = ColorHistogram(Image)
%Image    : 输入图像数据
%Hist     : 返回颜色直方图特征向量256维
%*************************************************************************
function Hist = ColorHistogram(Image)
% Image = imread('E:\\2\\5.jpg');
[M,N,Q] = size(Image);
[h,s,v] = rgb2hsv(Image);
H = h; S = s; V = v;
h = h*360;
```

```
%将hsv空间非等间隔量化：
%  h量化成16级；
%  s量化成4级；
%  v量化成4级；
for i = 1:M
    for j = 1:N
        if h(i,j)<=15||h(i,j)>345
            H(i,j) = 0;
        end
        if h(i,j)<=25&&h(i,j)>15
            H(i,j) = 1;
        end
        if h(i,j)<=45&&h(i,j)>25
            H(i,j) = 2;
        end
        if h(i,j)<=55&&h(i,j)>45
            H(i,j) = 3;
        end
        if h(i,j)<=80&&h(i,j)>55
            H(i,j) = 4;
        end
        if h(i,j)<=108&&h(i,j)>80
            H(i,j) = 5;
        end
        if h(i,j)<=140&&h(i,j)>108
            H(i,j) = 6;
        end
        if h(i,j)<=165&&h(i,j)>140
            H(i,j) = 7;
        end
```

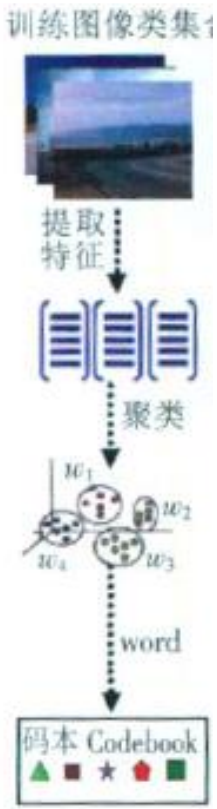
```
%将三个颜色分量合成为一维特征向量：L = H*Qs*Qv+S*Qv+v；Qs,Qv分别是S和V的量化级数， L取值范围[0,255]
%取Qs = 4; Qv = 4
for  i = 1:M
    for j = 1:N
        L(i,j) = H(i,j)*16+S(i,j)*4+V(i,j);
    end
end
%计算L的直方图
for i = 0:255
    Hist(i+1) = size(find(L==i),1);
end
Hist = Hist/sum(Hist);
```

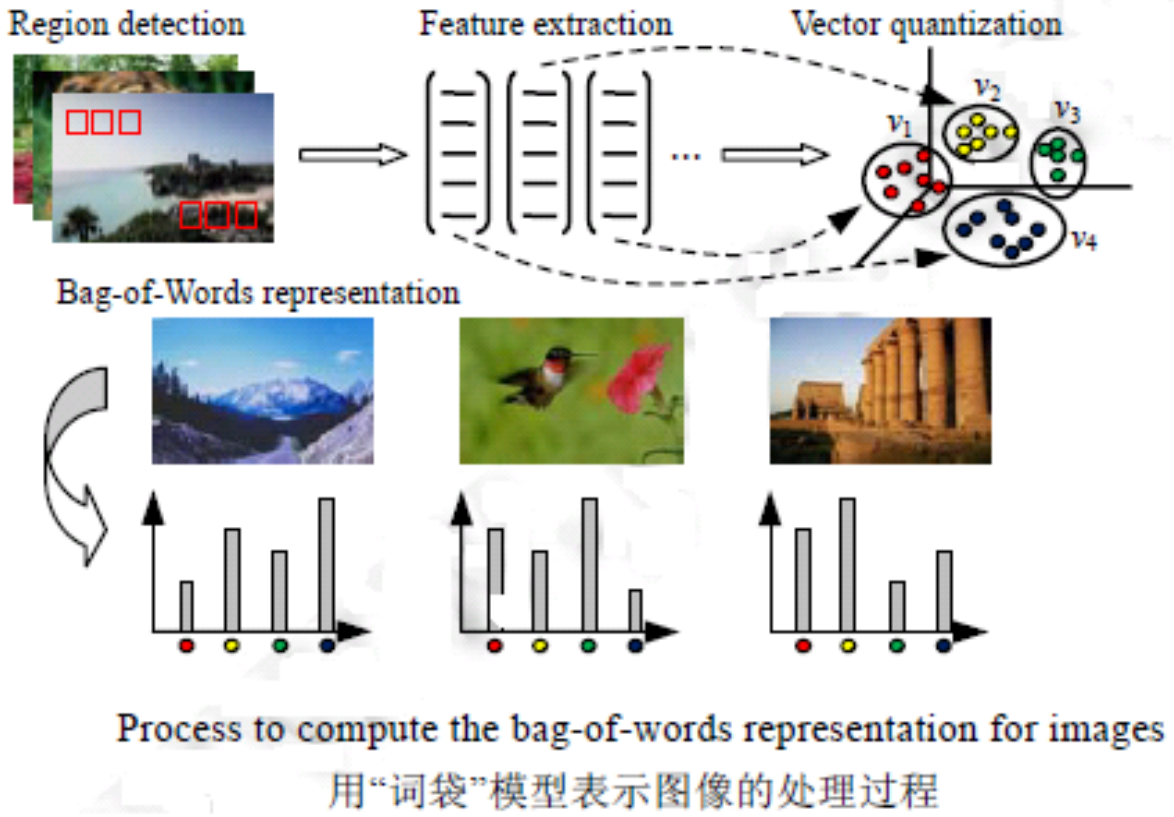
图像特征的离散化

算法流程


TRAINING
TESTING
class 1
class N
unknown image
Input Image
local feature extraction
form codebook
Class 1
Class N
represent each image into a bag of codewords
feature detection & representation
learning
learn Bayesian hierarchical models for each class
Model 1
Model N
decide on best model
recognition

利用BOW词袋的方法生成词典





Process to compute the bag-of-words representation for images

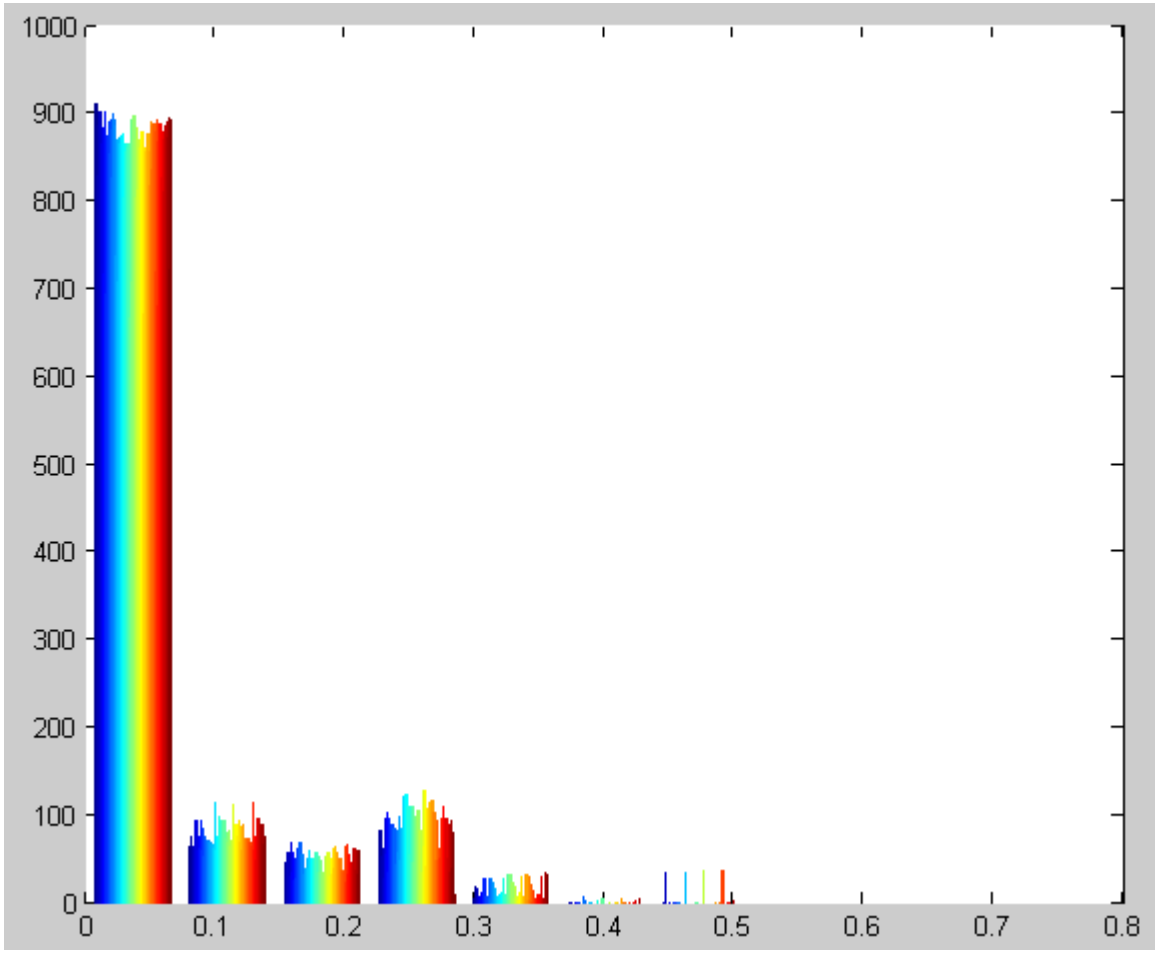
用“词袋”模型表示图像的处理过程

Dictionary

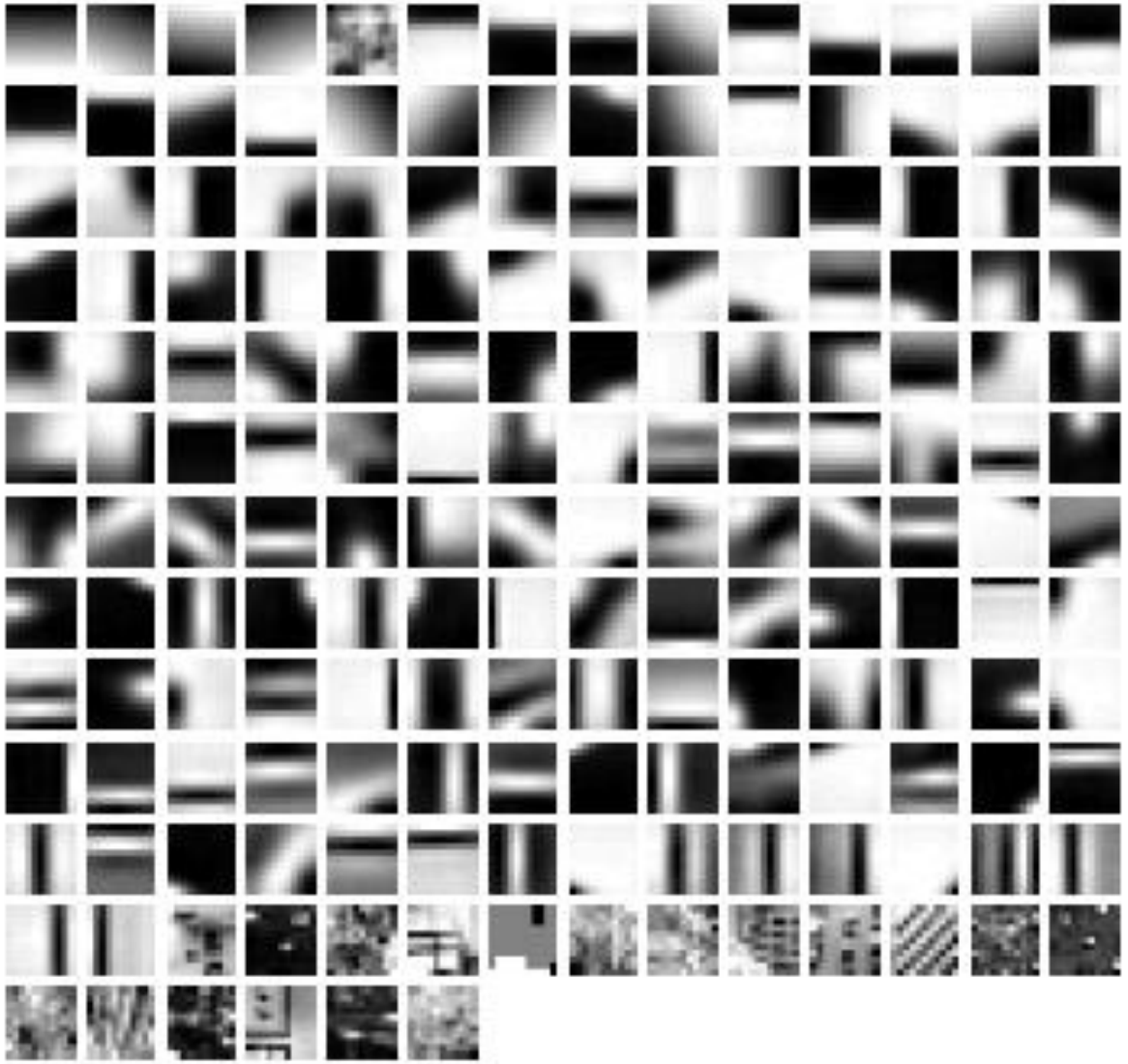
dictionary <200x128 double>

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0.0644	0.0648	0.0581	0.0476	0.0491	0.0567	0.0625	0.0635	0.0750	0.0851	0.0802	0.0597	0.0555	0.0594	0.0647	0.0629
2	0.0273	0.0156	0.0250	0.0285	0.0193	0.0149	0.0179	0.0303	0.0246	0.0192	0.0478	0.0777	0.0277	0.0120	0.0235	0.0323
3	0.0296	0.0456	0.1050	0.1464	0.0795	0.0242	0.0174	0.0289	0.0305	0.0504	0.1146	0.1591	0.1169	0.0314	0.0196	0.0165
4	0.0560	0.0477	0.0558	0.0617	0.0445	0.0364	0.0446	0.0547	0.0495	0.0475	0.0709	0.0915	0.0633	0.0477	0.0586	0.0635
5	0.0304	0.0153	0.0169	0.0123	0.0102	0.0281	0.0823	0.0561	0.0243	0.0095	0.0153	0.0087	0.0079	0.0479	0.2384	0.1022
6	0.0435	0.0493	0.0835	0.0532	0.0305	0.0345	0.0581	0.0514	0.0300	0.0299	0.0814	0.0616	0.0442	0.0641	0.1004	0.0592
7	0.0366	0.0418	0.0355	0.0348	0.0213	0.0342	0.0660	0.0659	0.0421	0.0411	0.0386	0.0397	0.0323	0.0368	0.0539	0.0531
8	0.0418	0.0492	0.0795	0.0647	0.0394	0.0365	0.0409	0.0427	0.0465	0.0382	0.0396	0.0390	0.0445	0.0682	0.0933	0.0781
9	0.0251	0.0462	0.1028	0.1056	0.0483	0.0208	0.0224	0.0137	0.0432	0.0479	0.0873	0.0899	0.0404	0.0330	0.0444	0.0396
10	0.0505	0.0267	0.0187	0.0257	0.0516	0.0173	0.0117	0.0254	0.0616	0.0241	0.0200	0.0223	0.0433	0.0144	0.0132	0.0269
11	0.0241	0.0286	0.0198	0.0182	0.0236	0.0128	0.0149	0.0195	0.0272	0.0244	0.0185	0.0204	0.0264	0.0196	0.0156	0.0184
12	0.0698	0.0209	0.0168	0.0692	0.1719	0.0492	0.0217	0.0402	0.0674	0.0227	0.0186	0.1068	0.2173	0.0500	0.0143	0.0330
13	0.0022	0.0113	0.0693	0.0090	0.0046	0.0078	0.0096	0.0042	0.0013	0.0423	0.2607	0.0258	0.0018	0.0258	0.1522	0.0081
14	0.0199	0.0141	0.0242	0.1201	0.1389	0.0372	0.0208	0.0205	0.0156	0.0142	0.0463	0.2378	0.2050	0.0220	0.0098	0.0110
15	0.0086	0.0713	0.2009	0.0726	0.0205	0.0174	0.0296	0.0152	0.0134	0.0378	0.0780	0.0354	0.0315	0.0980	0.1211	0.0335
16	0.0137	0.1129	0.2191	0.0692	0.0289	0.0238	0.0272	0.0107	0.0193	0.0925	0.1621	0.0386	0.0151	0.0723	0.1150	0.0347
17	0.0026	0.0209	0.2179	0.0333	0.0193	0.0149	0.0365	0.0055	0.0036	0.0246	0.1820	0.0246	0.0055	0.0374	0.2475	0.0257
18	0.0353	0.0319	0.0307	0.0309	0.0340	0.0633	0.1437	0.0875	0.0351	0.0471	0.0686	0.0669	0.0503	0.0510	0.0759	0.0509
19	0.0223	0.0804	0.1067	0.0394	0.0186	0.0167	0.0132	0.0093	0.0297	0.0691	0.0815	0.0387	0.0215	0.0171	0.0156	0.0168
20	0.0550	0.1210	0.1677	0.1023	0.0596	0.0424	0.0310	0.0255	0.0580	0.1376	0.1846	0.0884	0.0260	0.0151	0.0128	0.0176
21	0.0049	0.0055	0.0123	0.0269	0.0237	0.0100	0.0149	0.0021	0.0041	0.0031	0.0235	0.0704	0.0276	0.0106	0.0093	0.0029
22	0.0032	0.0097	0.1042	0.0369	0.0081	0.0373	0.2295	0.0722	0.0070	0.0275	0.2607	0.0815	0.0081	0.0119	0.0530	0.0191
23	0.0109	0.0158	0.0252	0.0546	0.1806	0.0653	0.0168	0.0080	0.0109	0.0089	0.0253	0.0895	0.2048	0.0744	0.0169	0.0057
24	2.5901e-04	3.0658e-04	3.6310e-04	3.2259e-04	0.0087	0.0592	0.0286	0.0031	1.9426e-04	1.6477e-04	2.6530e-04	0.0012	0.0143	0.0602	0.0253	0.0024
25	0.0302	0.0352	0.0282	0.0124	0.0180	0.0155	0.0218	0.0222	0.0226	0.0335	0.0277	0.0203	0.0207	0.0224	0.0253	0.0175
26	0.0712	0.0663	0.0289	0.0252	0.0379	0.0282	0.0209	0.0231	0.0678	0.0731	0.0261	0.0222	0.0545	0.0481	0.0218	0.0268
27	0.0135	0.0270	0.0169	0.0363	0.2874	0.4385	0.1261	0.0093	0.0145	0.0252	0.0358	0.0126	0.1028	0.3634	0.1523	0.0247
28	0.0225	0.0379	0.1024	0.0864	0.0405	0.0331	0.0353	0.0317	0.0298	0.0519	0.1397	0.1152	0.0351	0.0253	0.0514	0.0581

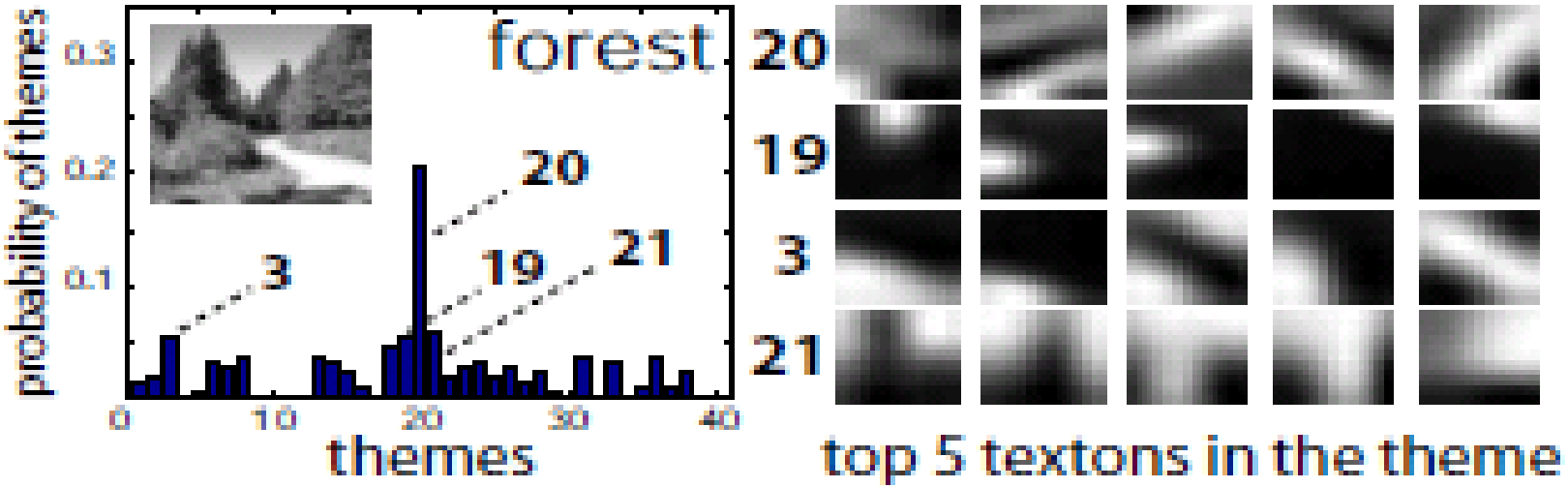
Hist



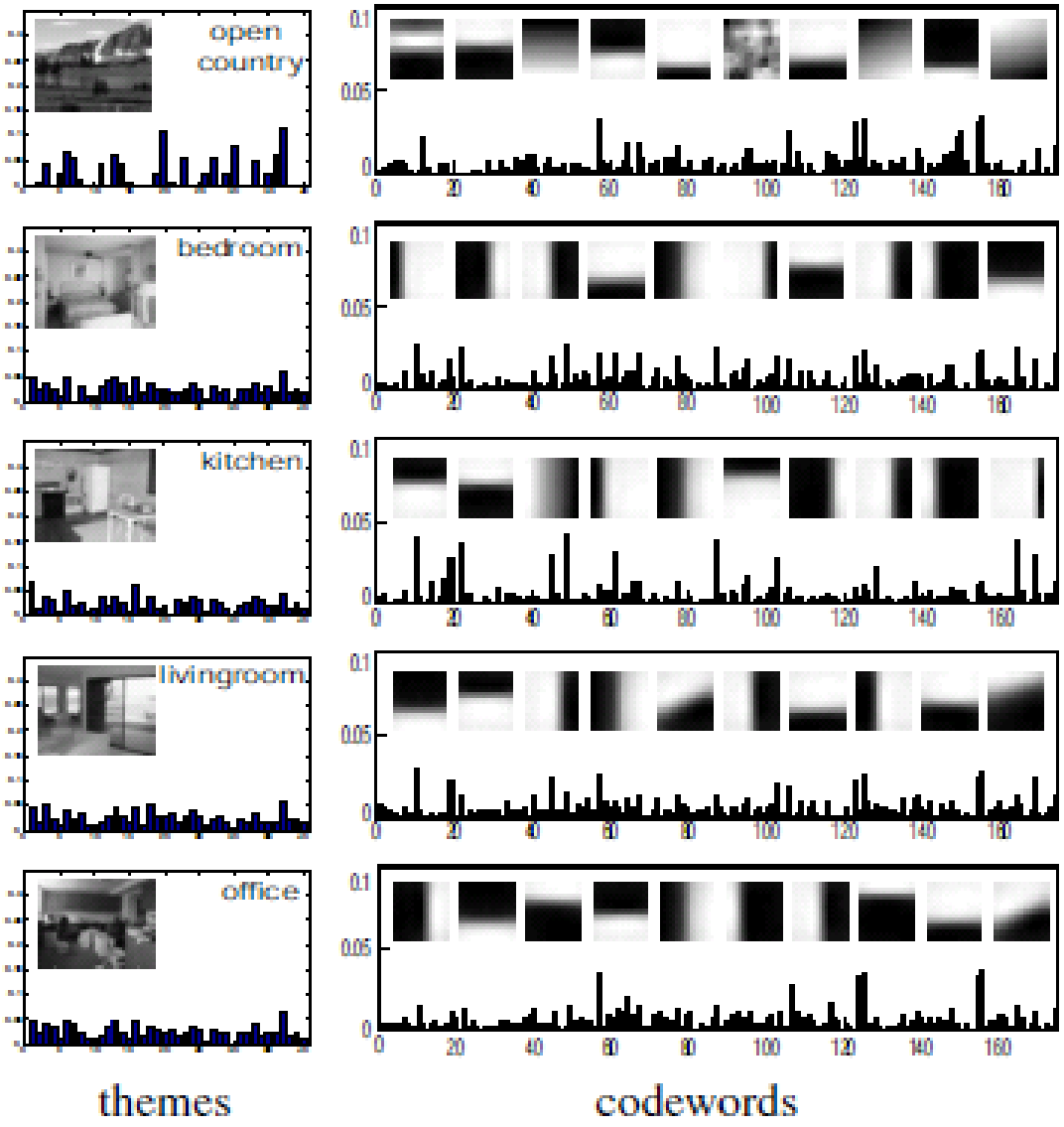
典型的视觉词典



主题和词典


probability of themes
0.3
0.2
0.1
forest
20
19
21
3
0
10
20
30
40
themes
20
19
3
21
top 5 textons in the theme

视觉词典表示的图例


open country
bedroom
kitchen
livingroom
office
themes
codewords