PENNSTATE

1855

Chapter 2: Data

Presentation extended from the slides of the textbook, Introduction to Data Mining by Tan et al. and supplementary material

Overview

- What is data?
  - Data Types
  - Data Quality
- Data Preprocessing
- Data Similarity and Dissimilarity

What is Data?

- Data captures things, phenomena, etc, in forms of collection of *data objects* and their *attributes*
- An *attribute* is a property or characteristic of an object
  - E.g., eye color of a person, temperature, etc.
  - Attribute is also known as variable, feature, field or characteristic
- An *object* is described by a set of attributes
  - 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

Attribute Values

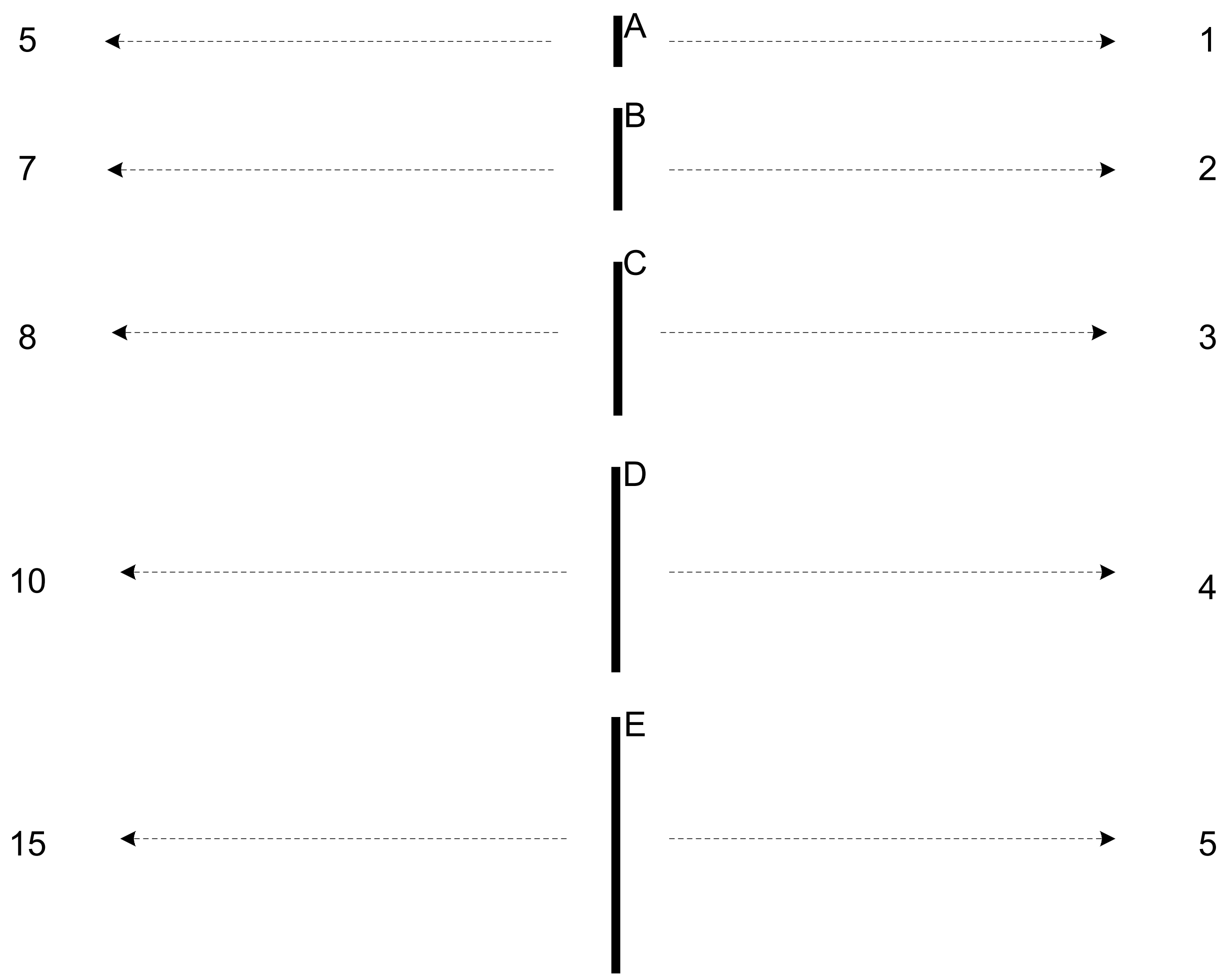
- Attribute values are *numbers* or *symbols* assigned to an attribute
- Distinction between attributes and attribute values
  - *Attribute is the semantic notation while the attribute value is the numeric measure or symbolic representation.*
  - Same attribute can be mapped to different attribute values
    - 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 both integers
    - But properties of attribute values can be different
      - ID has no limit but age has a maximum and minimum value

Process of Measurement

- The process of measurement is the application of a measurement scale to associate a value with a particular attribute of an object.
- The properties of an attribute may not be the same as the properties of the values used to measure the attribute
  - *Choose a measure carefully!*
  - Integers can be used to represent Employee attributes such as Age and ID Number, but not all integer operations can be meaningfully applied to them.

Measurement of Length

Different measurements can be used to capture the desired properties attributes, e.g., length, based on application requirements.



Types of Attributes

- There are different types of attributes
  - Nominal
    - Examples: ID numbers, eye color, zip codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale of 1-10), grades, height in {tall, medium, short}
  - Interval
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Ratio
    - 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: $= \neq$
  - Order: $< >$
  - Addition: $+ -$
  - Multiplication: $* /$

  相异性
  - 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	If all employee ID numbers were reassigned, would it make any difference?
Ordinal	An order preserving change of values, i.e., *new_value = f(old_value)* where *f* is a monotonic function.	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Interval	*new_value =a * old_value + b* where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
Ratio	*new_value = a * old_value*	Length can be measured in meters or feet.

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.
  - *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.

Data Sets

- Many types of data sets
  - Record Data, e.g., transaction, document, etc.
  - Graph Data, e.g., World Wide Web, molecular structures, social networks, etc.
  - Ordered Data, e.g., temporal data, sequential data, genetic sequence, spatial data, etc.

- Important characteristics of data Sets
  - Dimensionality: curse of dimensionality
  - Sparsity: only presence (with non-null values) counts
  - Resolution: patterns depend on the scale.

Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes
每个记录包含固定的字段集。

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**

Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as *points in a multi-dimensional space*, where each dimension represents a distinct attribute
- Such data set can be represented by an $m \times n$ matrix, where there are m rows, one for each object, and n columns, one for each attribute

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

Transaction Data

- Special type of record data & sparse data matrix each record (transaction) involves a set of items.
  - Consider a grocery store. The set of products (items) purchased by a customer during one shopping trip constitute a transaction.

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

  - Transaction dataset are usually represented as the above instead of sparse data matrix.

Document Data

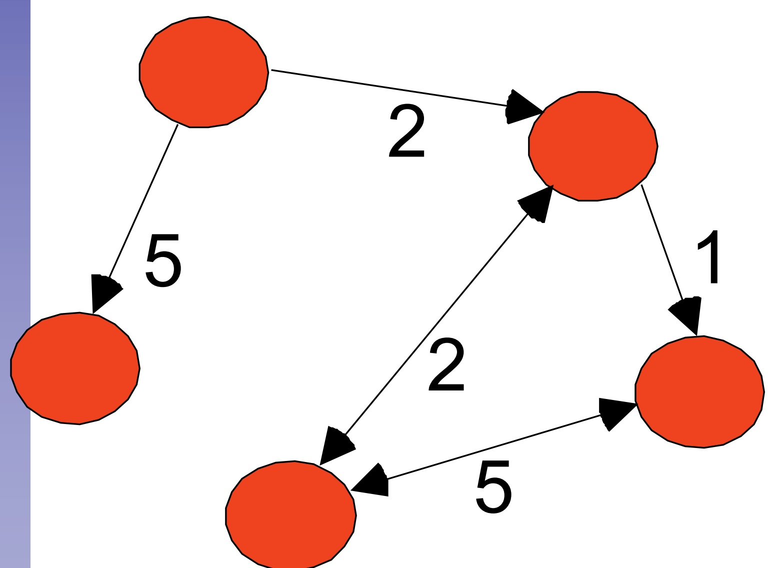
检索词向量

- Each document is 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	pla y	ball	score	game	wi n	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

Graph Data

- A graph is a powerful representation of data
  - captures *relationship* among objects
  - captures *complex structure* of objects
- Examples: Generic graph and HTML Links



Data Mining

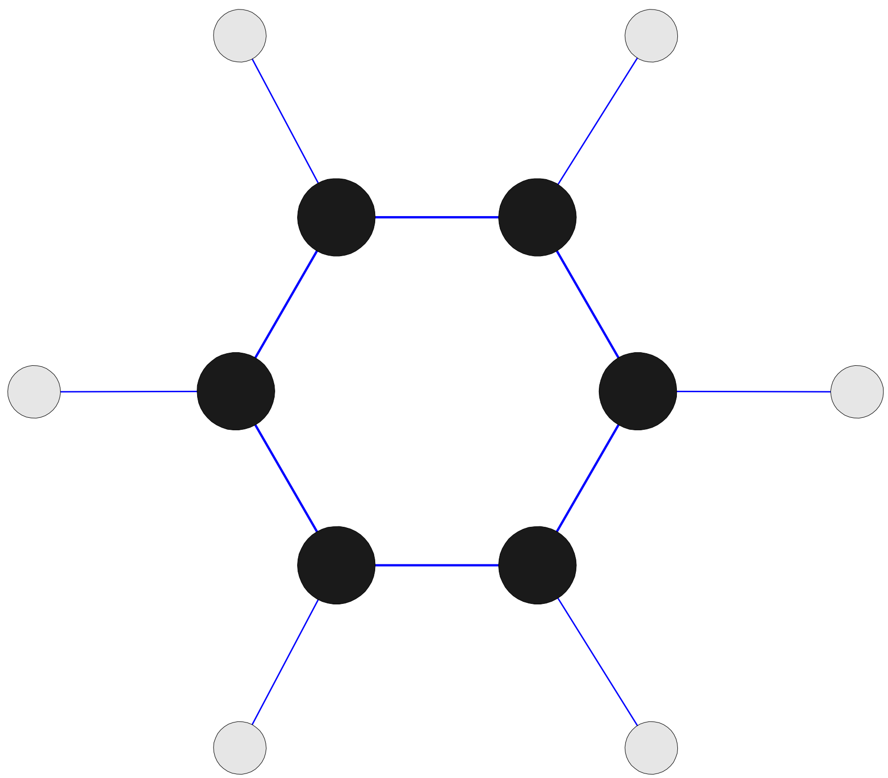
Graph Partitioning

Parallel Solution of Sparse Linear System of Equations

<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers

Chemical Data

- Benzene Molecule: C_6H_6



Ordered Data

- For some data sets, the attributes involve *order* in time and space.
- E.g., sequences of transactions.

Items/Events

(A B) (D) (C E)
(B D) (C) (E)
(C D) (B) (A E)

An element of the sequence

Ordered Data

- Genomic sequence data

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

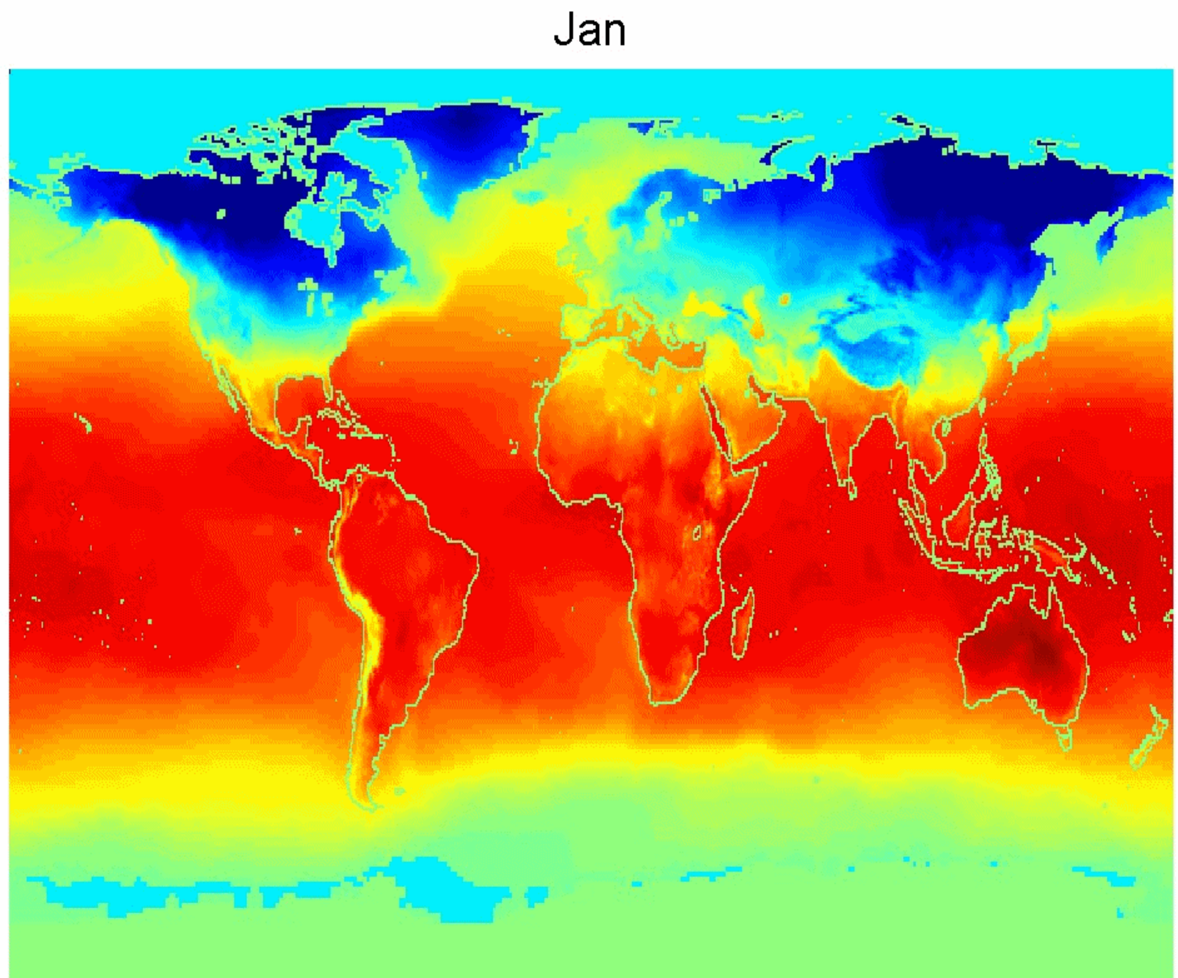
Ordered Data

- Time series data

Ordered Data

- Spatio-Temporal Data

Average Monthly Temperature of land and ocean



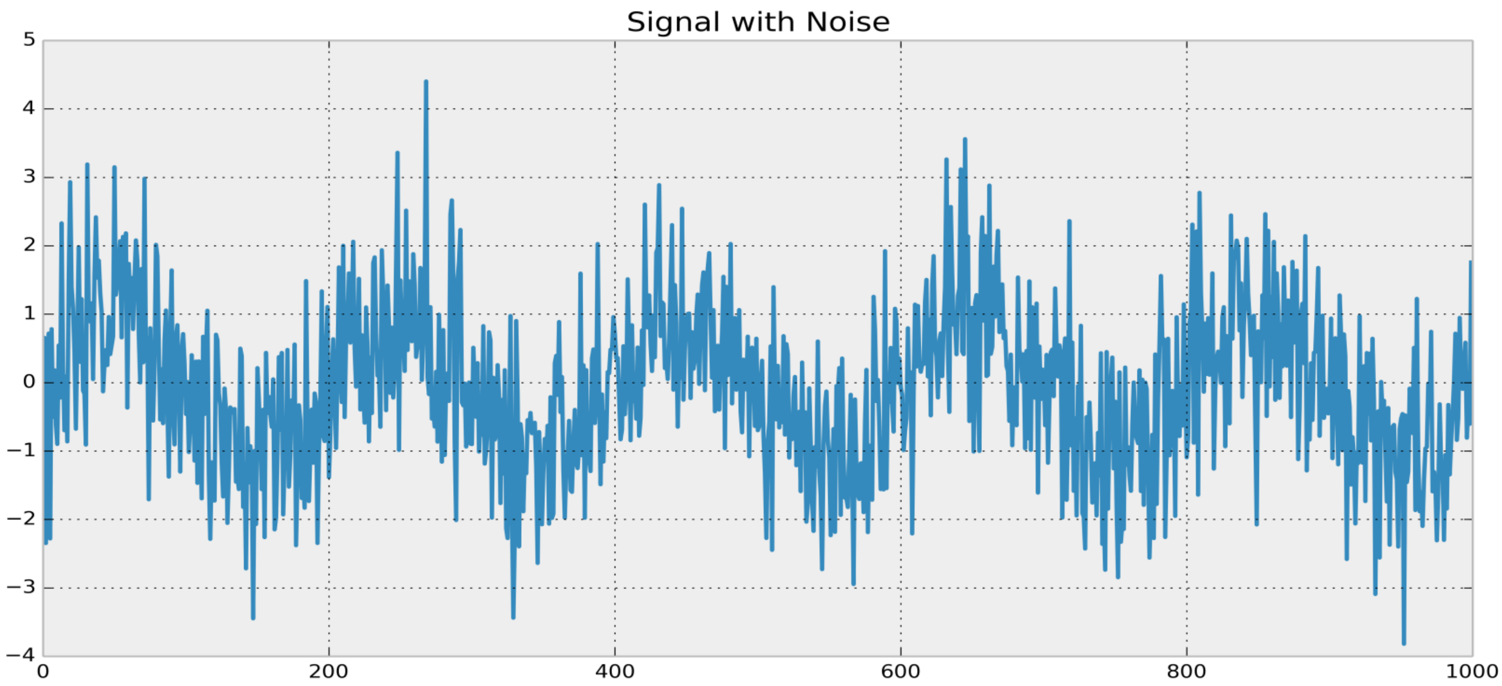
Data Quality

- Data mining applications are often using data collected for other (or future) applications, and thus facing serious data quality issues.
  - What kinds of data quality problems?
  - How can we detect problems with the data?
  - What can we do about these problems?
- Many data quality issues are related to *measurement* and *data collection*. For examples:
  - Noise and outliers
  - missing values, inconsistent values
  - duplicate data

Noise

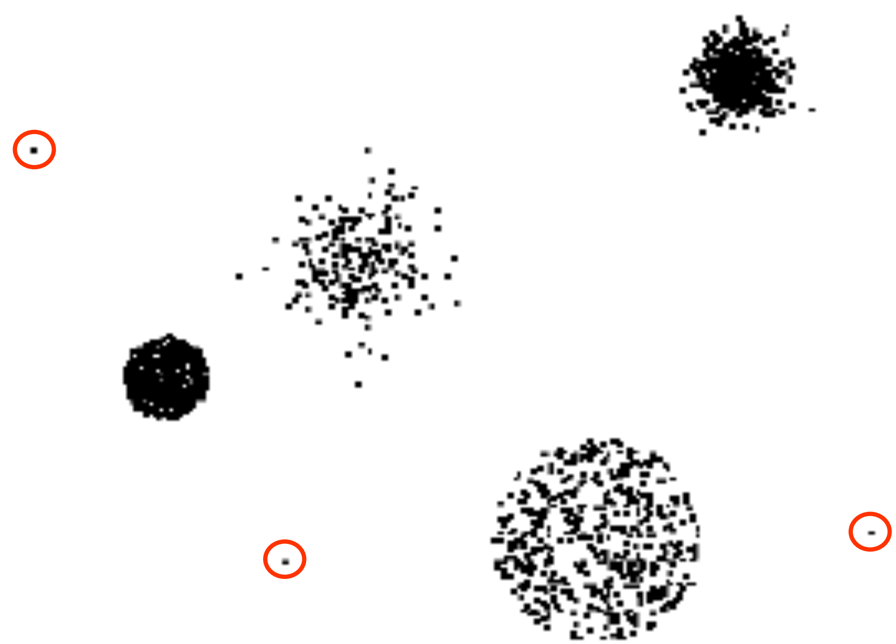
是指

- Noise refers to deviation from the original values
  - E.g., distortion of a person's voice when talking on a poor phone and "snow" on television screen
  - Measurement error



Outliers

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



Missing Values

- 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 (or attributes w/ caution)
  - Estimate Missing Values
  - Ignore the Missing Value during Analysis
  - Replace with all possible values (weighted by their probabilities)

不一致

Inconsistent Data

- Inconsistent data needs to be detected/corrected, if all possible.

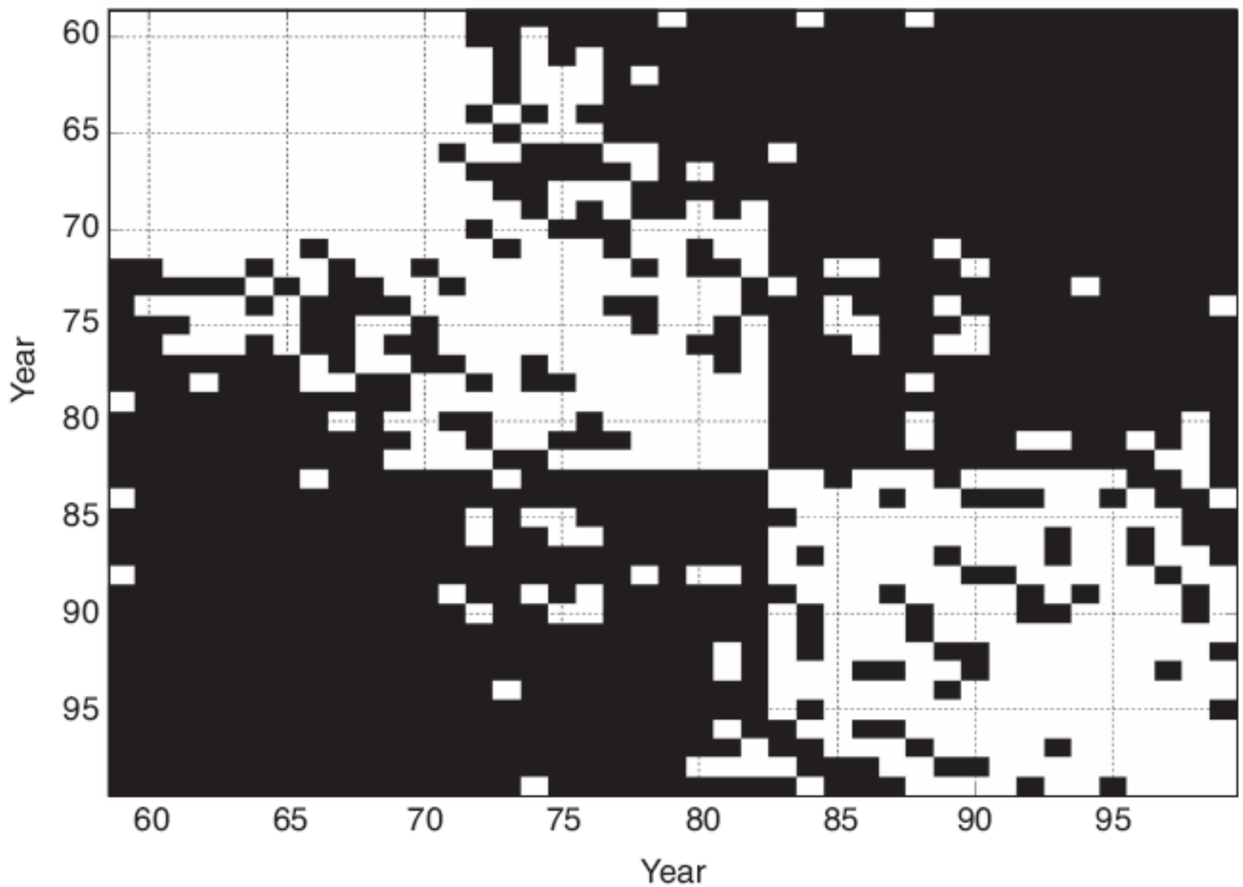


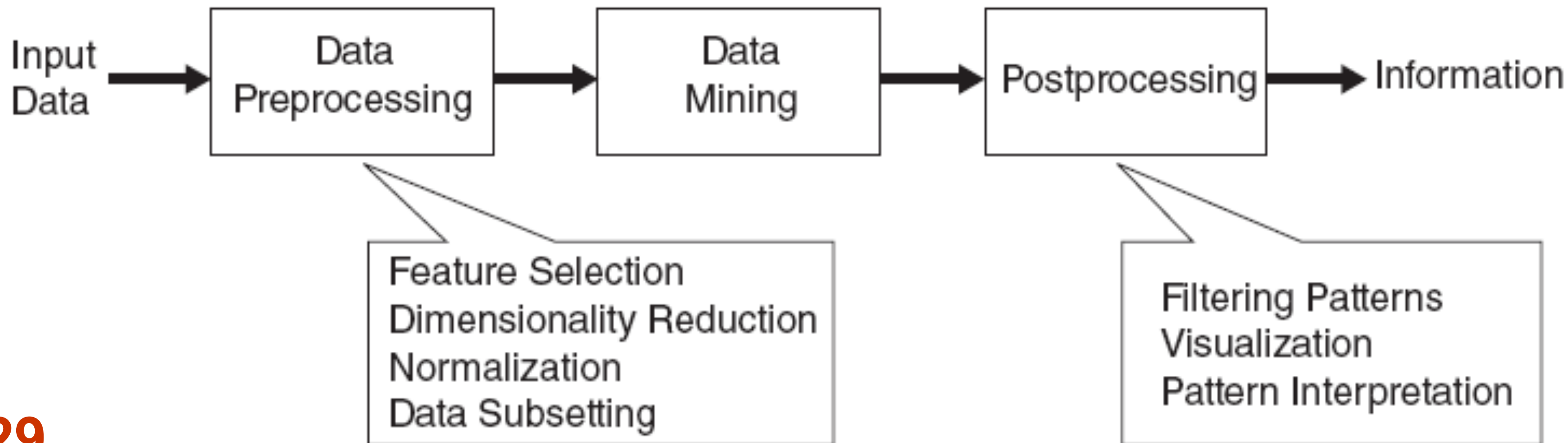
Figure 2.7. Correlation of SST data between pairs of years. White areas indicate positive correlation. Black areas indicate negative correlation.

Duplicate Data

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

Data Preprocessing

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



Aggregation

- Combining two or more attributes (or objects) into a single attribute (or object)

- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale 改变范围或标度
    - Cities aggregated into regions, states, countries, etc
  - More “stable” data
    - Aggregated data tends to have less variability

Aggregation: Sales Data

- Change the scope/scale of data to provide a higher-level view.

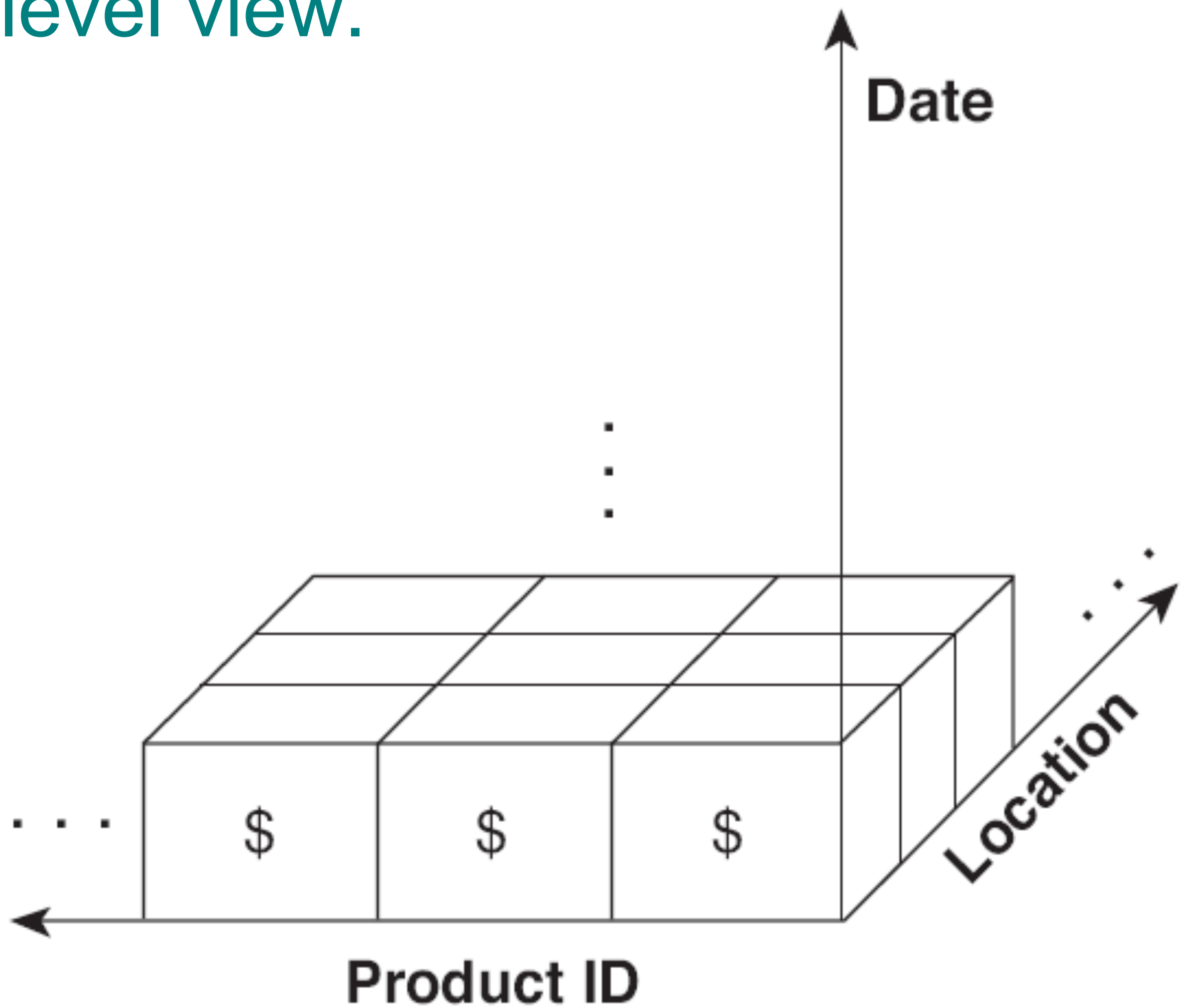
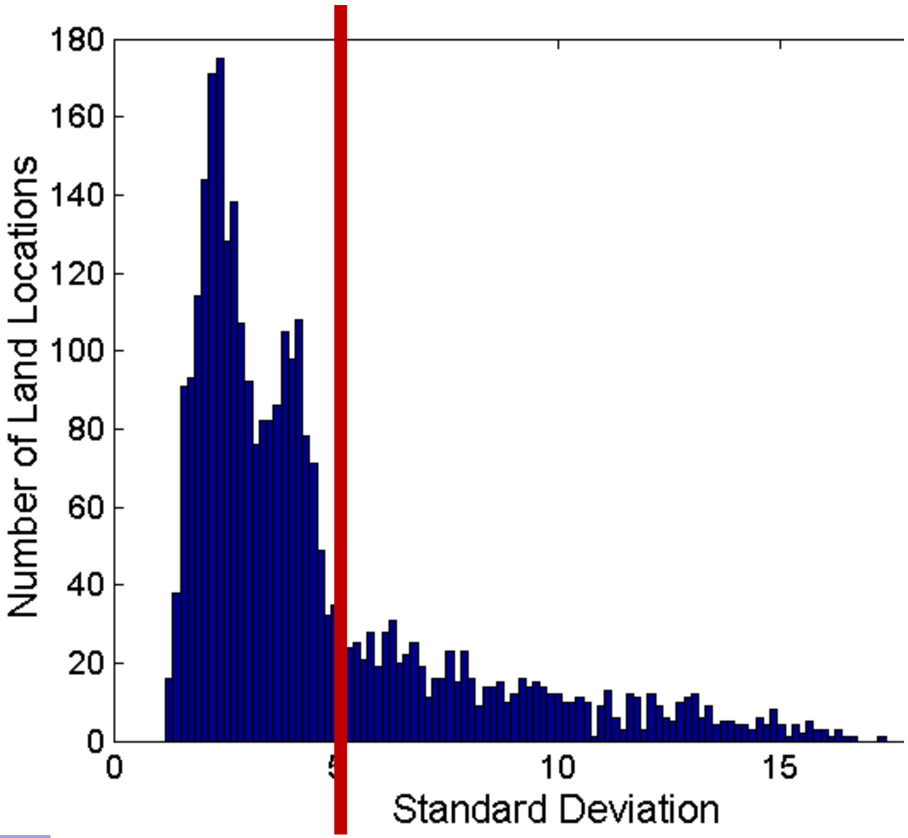


Figure 3.31. Multidimensional data representation for sales data.

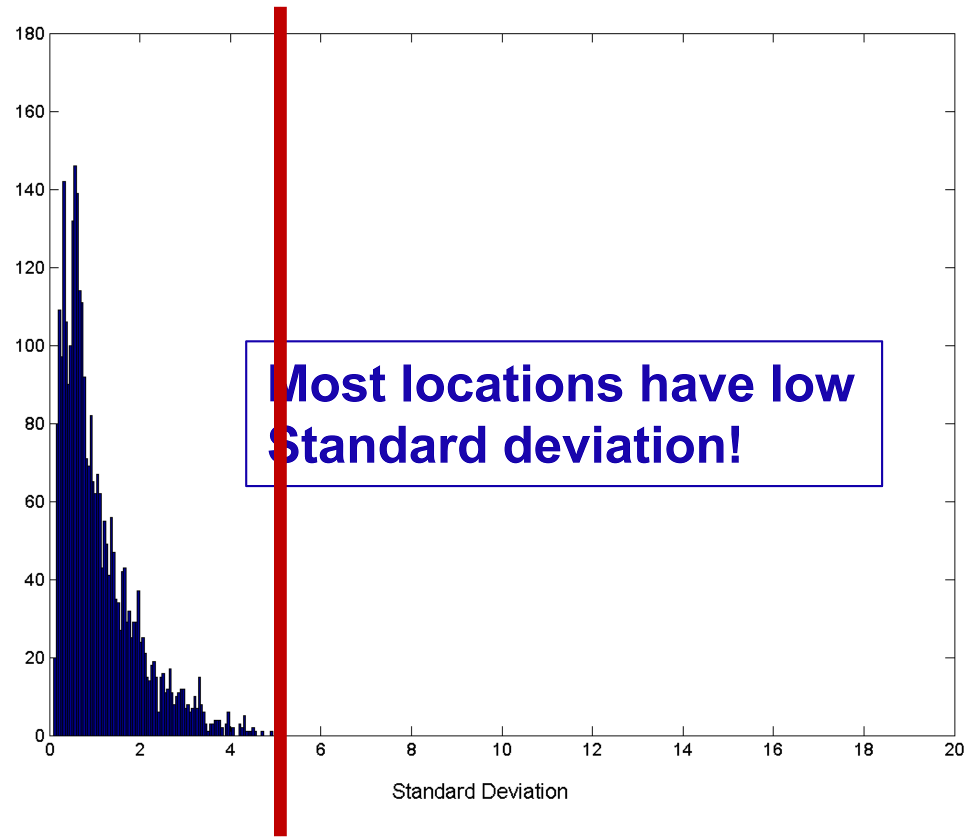
降水量

Aggregation – Precipitation Data

- The behavior of group is usually more stable than individuals.



Standard Deviation of Average Monthly Precipitation



Standard Deviation of Average Yearly Precipitation

Sampling

- Sampling is the main technique employed for data selection.
  - 事先调查
  - It is often used for both the preliminary investigation of the data and the final data analysis.
- 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.
  - Big data computing frameworks aim to address the processing needs. Recent advances help but sampling is still needed sometimes.

原理

Sampling Principle

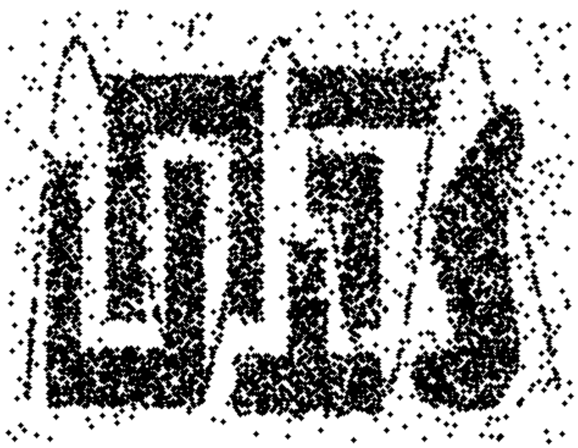
- The key principle for effective sampling is the following:
  - *Using a sample will work almost as well as using the entire data sets, if the sample is representative*
  - *A sample is representative if it has approximately the same property (of interest) as the original set of data*

Sampling Approaches

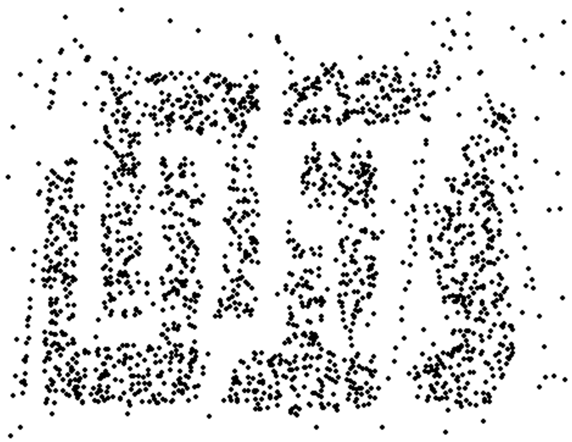
- Simple Random Sampling
  - There is an equal probability of selecting any item
- Sampling without replacement
  - An item is removed from the population after being selected
- Sampling with replacement
  - Objects are not removed from the population as they are selected for the sample.
    - the same object can be picked up more than once
    - Easy to analyze as probability not affected by sampling.
- Stratified sampling
  - Split the data into pre-specified groups; then draw random samples from each group

Sample Size

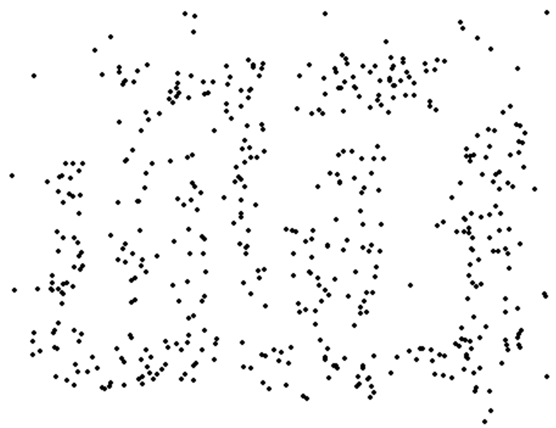
- Large sample is representative but loose the advantage of sampling.
- Small sample may result in missing or erroneous patterns



8000 points



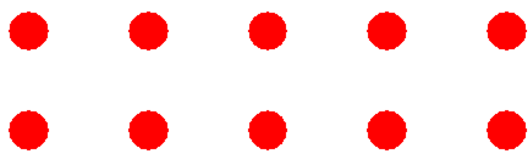
2000 Points

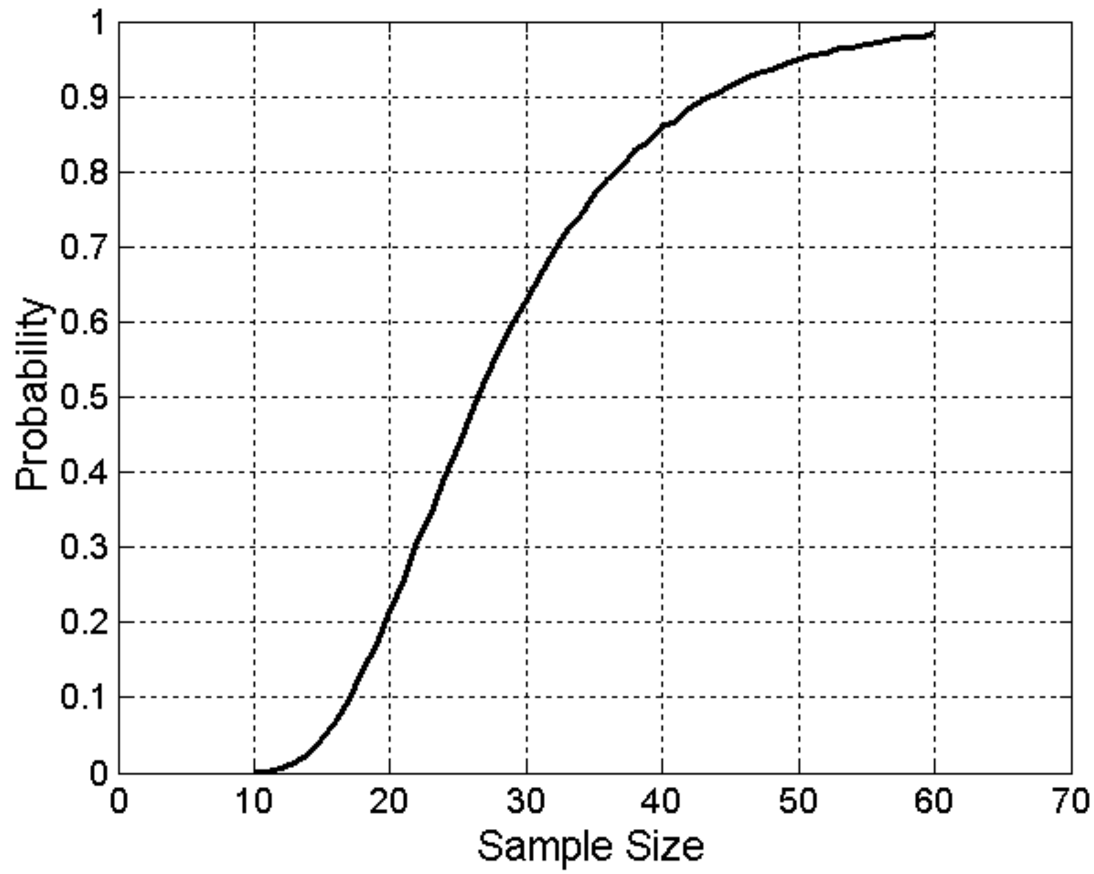


500 Points

Determine Sample Size

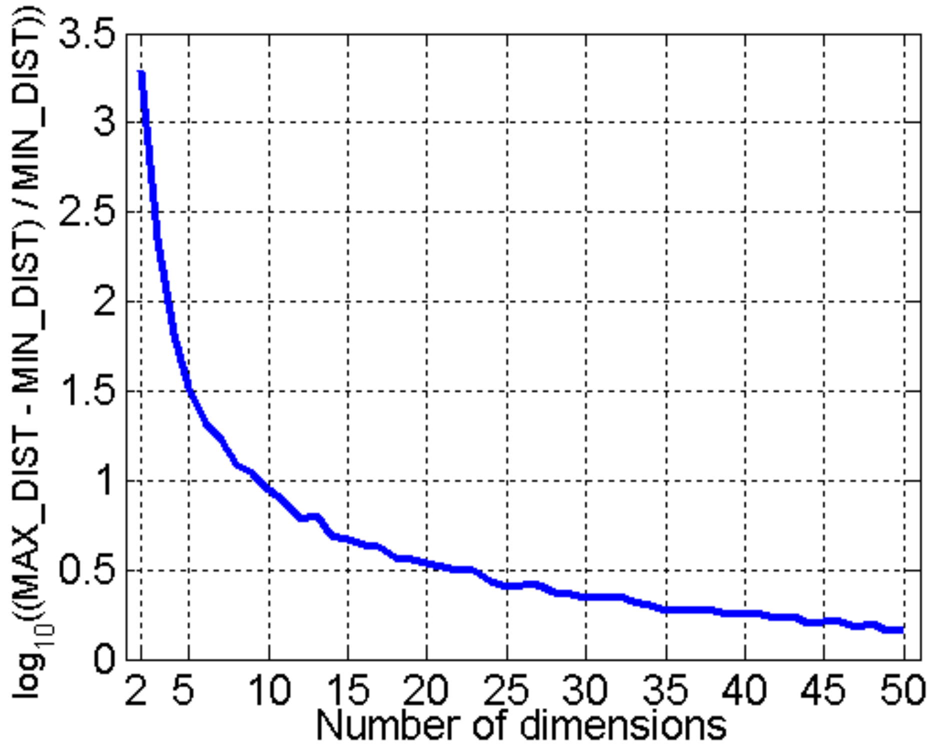
- What sample size is necessary to get at least one object from each of 10 groups?





Curse of Dimensionality

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



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

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
  - Linear Algebra Techniques
    - Principle Component Analysis
    - Singular Value Decomposition
  - Others: supervised and non-linear techniques, e.g., neural networks.

主成分分析

Principle Component Analysis

查找捕获数据中最大变化量的正交投影（即原始尺寸的线性组合）

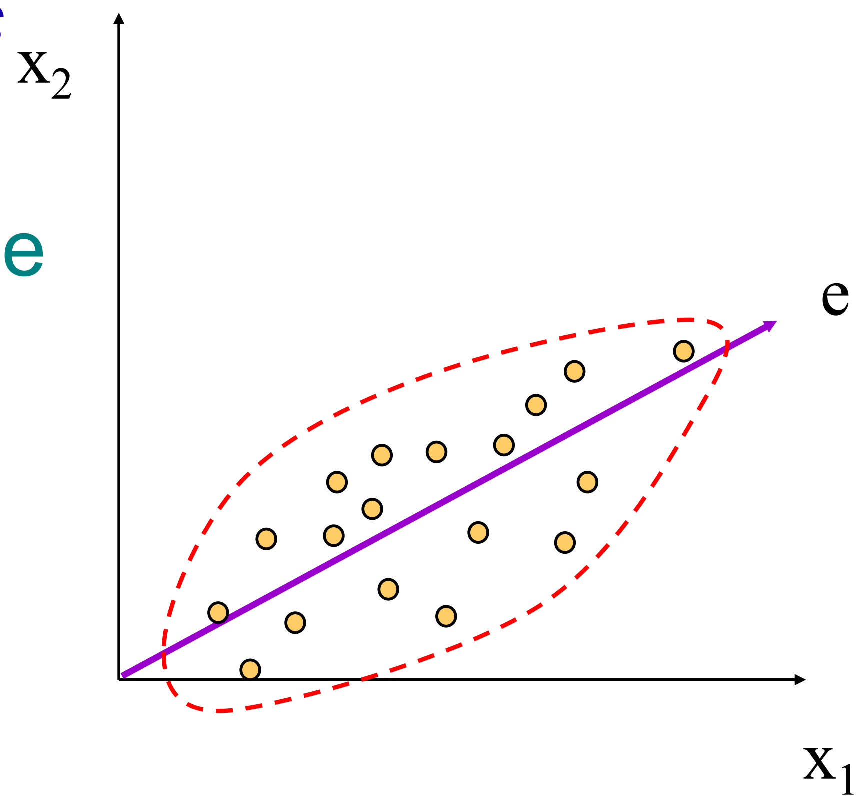
- Find orthogonal projections (i.e., linear combinations of original dimensions) that captures the largest amount of variation in data

在线性代数中，这是为了找到协方差矩阵的特征向量

- In Linear Algebra, this is to find the *eigenvectors* of the covariance matrix

特征向量（主成分）定义了新的维度减小的空间

- The eigenvectors (principle components) defines the new dimension-reduced space



Feature (Subset) Selection

- Select some features, eliminate others.
  - *Will this cause information loss?*

- Redundant features
  - 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
  - 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 Selection

- Brute-force (ideal) approach:
  - Try *all possible feature subsets* as input to data mining algorithm
- Embedded approaches:
  - Feature selection occurs naturally as part of the data mining algorithm
  - Decision tree operates in this manner
- Filter approaches:
  - Features are selected *before* data mining is performed
  - E.g., select attributes with low pair-wise correlation
- Wrapper approaches:
  - Use the data mining algorithm as a black box to find the best subset of attributes

Steps for Feature Subset Selection

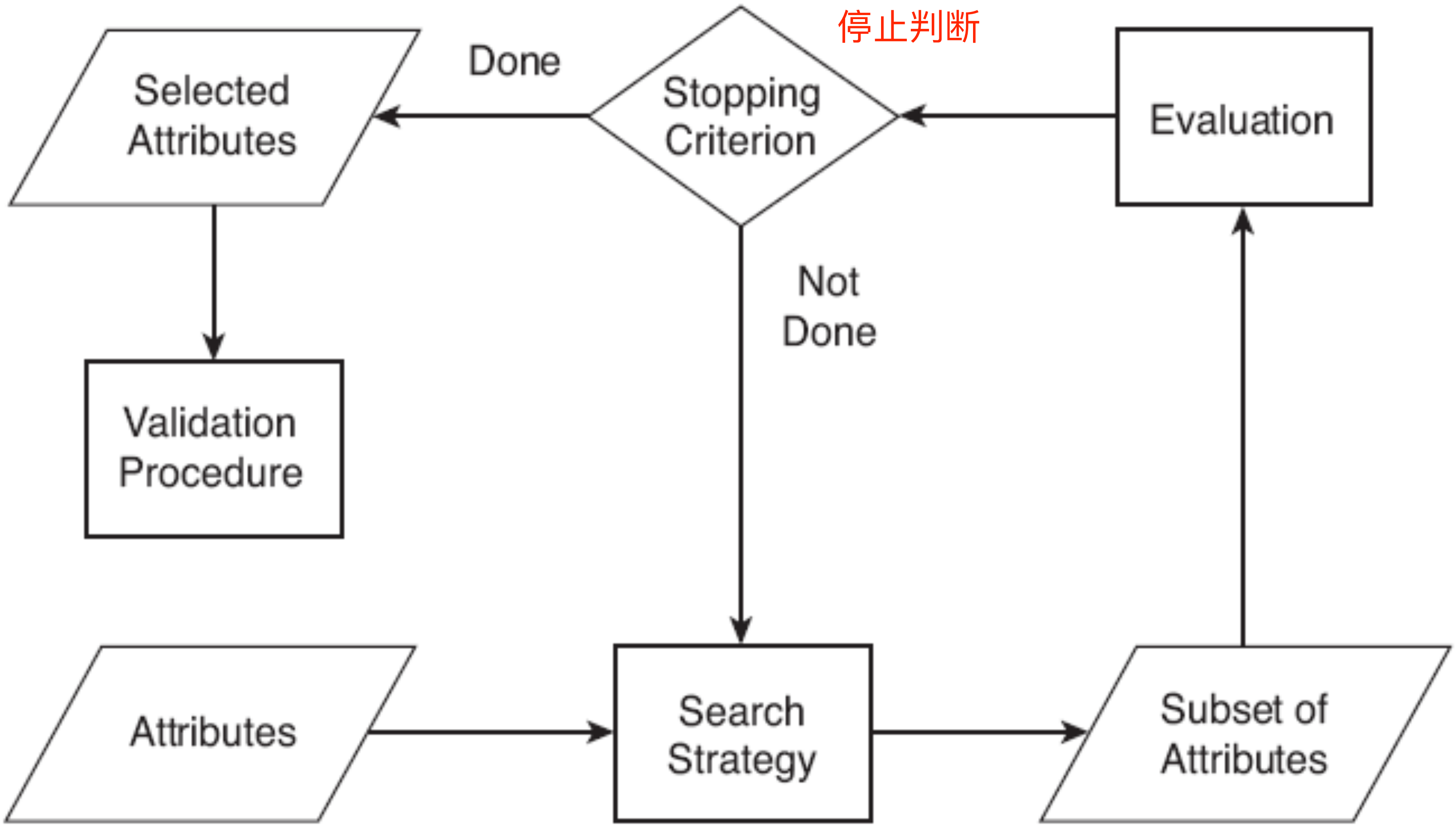


Figure 2.11. Flowchart of a feature subset selection process.

Feature Creation

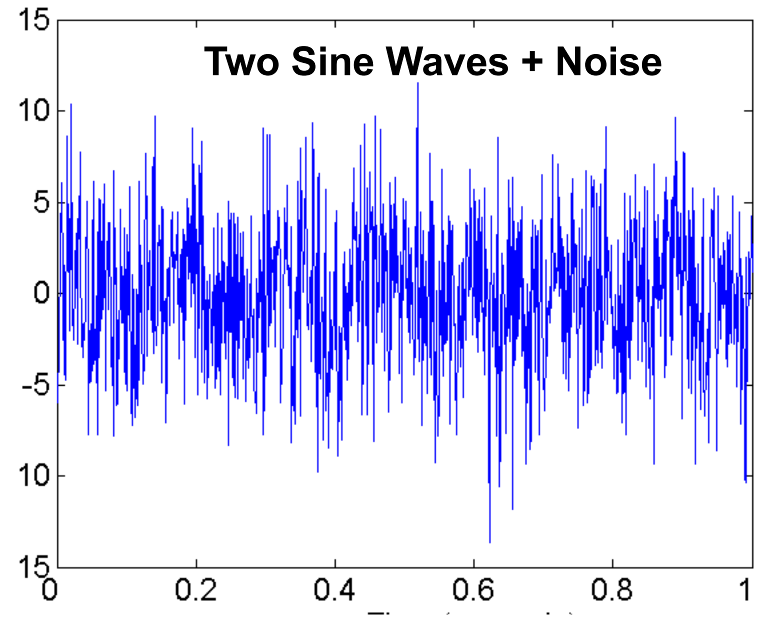
- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes

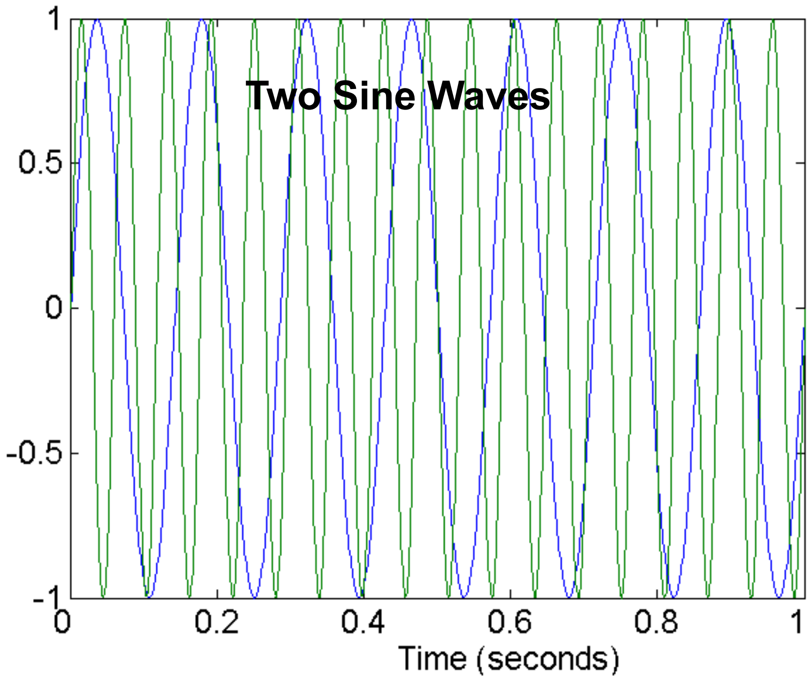
- Three general methodologies:
  - Feature Extraction
    - domain-specific
  - Feature Transformation
    - mapping data to new space
  - Feature Construction
    - combining features
    - E.g., derive speed from distance and interval from a vehicle terajectory dataset

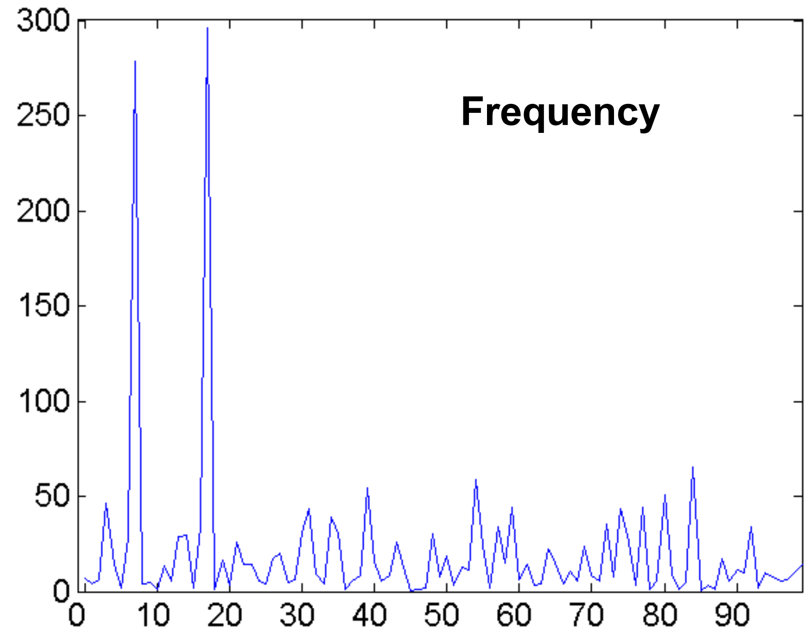
Mapping Data to a New Space

傅里叶变换：识别时间序列数据中的基本频率

- Fourier transform
- Wavelet transform







Discretization and Binarization

将连续属性变换成分类属性。

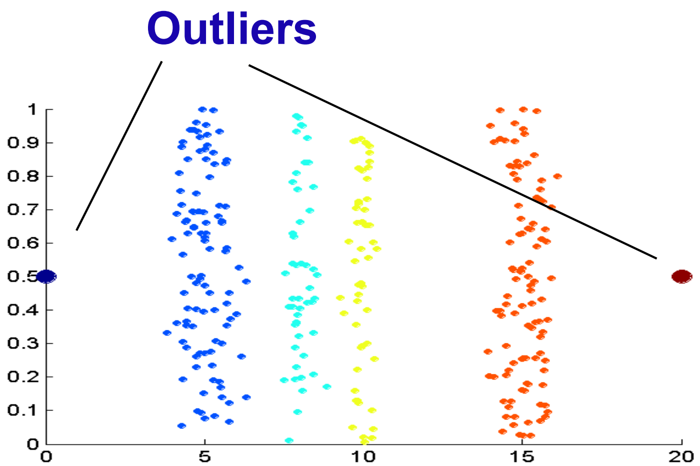
- *Discretization:* transform a continuous attribute into a categorical one.
  - Some data mining algorithms, e.g., classification, require categorical attributes.
- *Binarization*: transform continuous and categorical attributes into binary one
  - Association pattern discovery may require binary attributes.

连续属性离散化

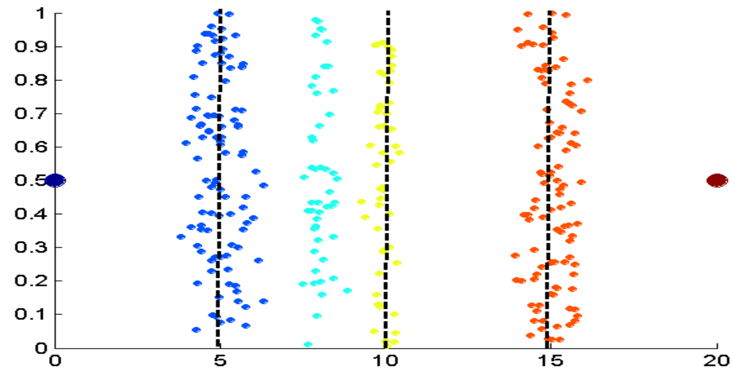
Discretization of Continuous Attributes

- Sort the attribute values and divide them into *n* intervals (by 指定 specifying *n-1* split points).
  - The key issues are how many split points to choose and where to put them.
- Depending on class information (i.e., labels) are used or not, discretization methods can be classified as follows:
  - Unsupervised discretization
    - Equal Width, Equal Depth/Frequency, Clustering-based
  - Supervised discretization
    - Class information is useful, as unsupervised methods usually result in intervals of objects with mixed labels.

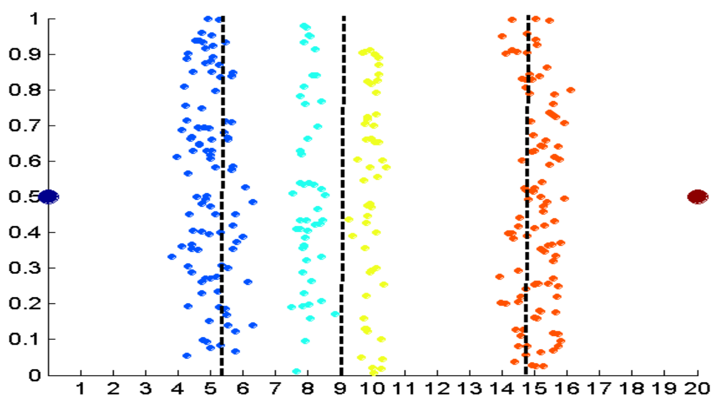
Discretization Without Using Class Labels



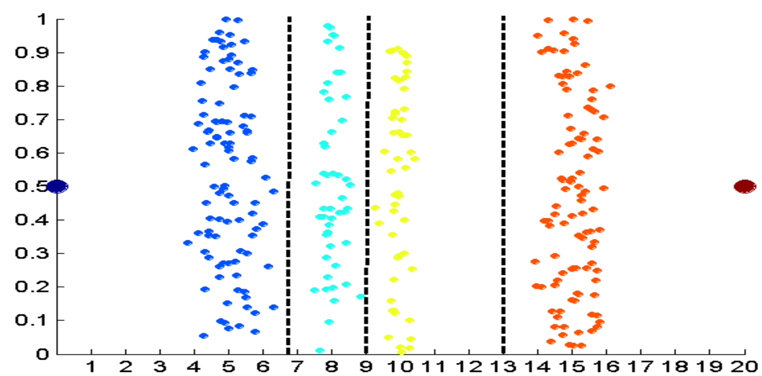
Data



Equal interval width



Equal frequency

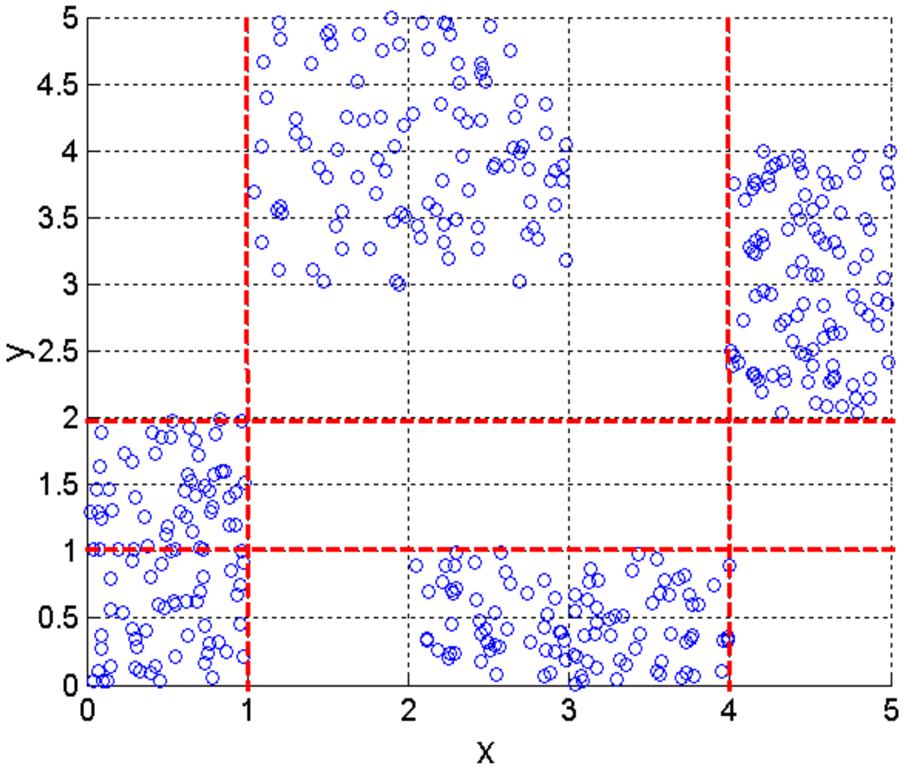


K-means

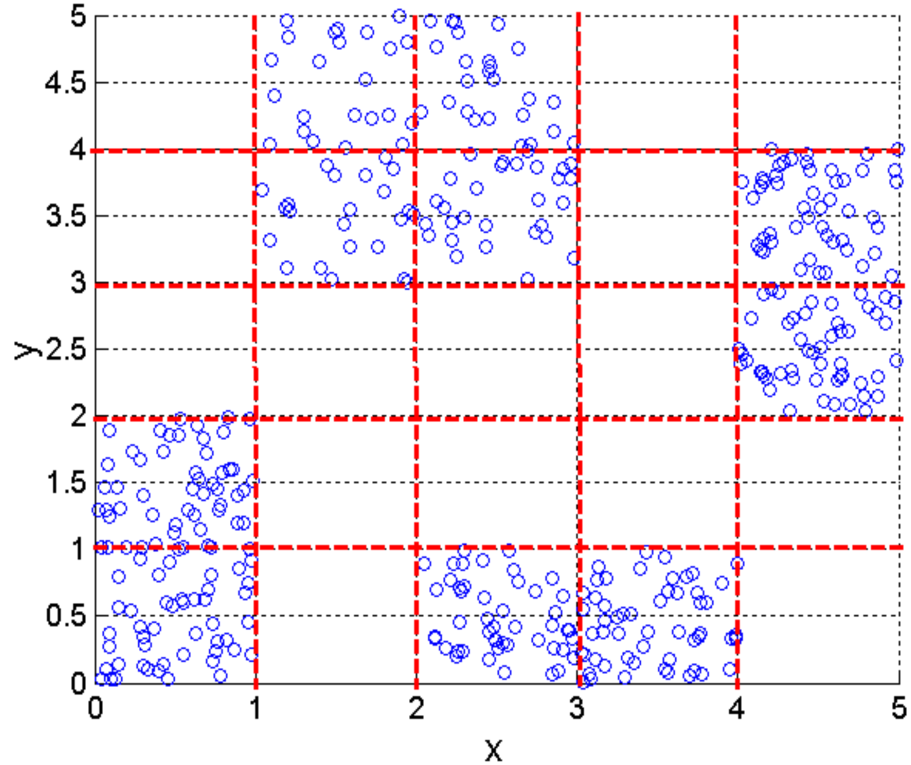
Split the points into 4 Intervals

Discretization Using Class Labels

- Entropy based approach: use entropy as a *purity* measure to divide the objects.



3 categories for both x and y



5 categories for both x and y

- In left figure, the separation in one dimension is not as good as two dimensions. In right figure, it's ok.

分类属性二元化

Binarization of Categorical Attributes

- Simple technique: *m* categories expressed in *log m* bits.
  - Sometimes create unintended relationship between bits
- One hot representation: *m* categories expressed in *m* bit. Only one bit is set to 1 and the rest bits are 0.
  - Bits are independent of each other

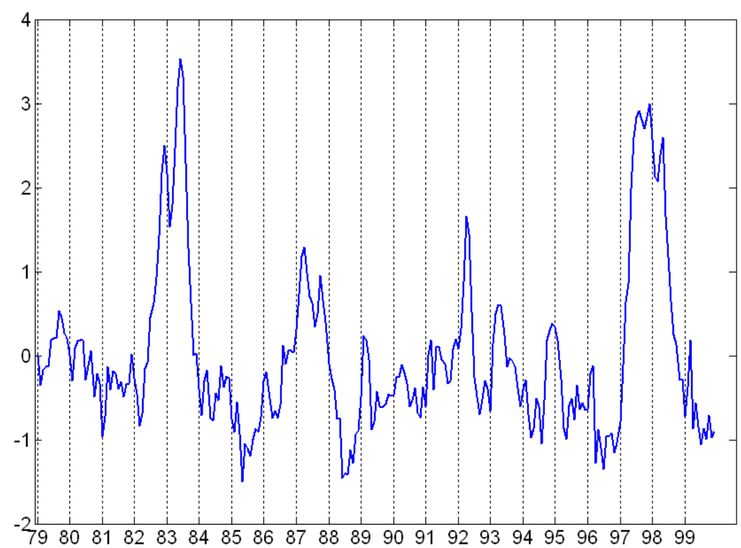
Categorical Attributes with Too Many Values

- *How to we handle it?*

- For ordinal attributes, *discretization* techniques could be used
- Categorical attributes with a lot of values can be combined, based on some relationship or taxonomy, to reduce the number of values
  - E.g., EE, CSE, IE all belong to College of Engineering.

Attribute Transformation

- Sometimes we need to transform attribute values into different form to amplify/smooth their effect in data mining algorithms.
  - E.g., salary and age are considered together by weighted sum. 加权和
- 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*



Similarity and Dissimilarity

- *Similarity* and *dissimilarity* are important for many data mining techniques, e.g., clustering.
- Similarity
  - Numerical measure of *how alike two data objects are*.
  - Is higher when objects are more alike.
  - Often falls in the range [0,1]
- Dissimilarity
  - Numerical measure of *how different are two objects*
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies

Similarity/Dissimilarity for Simple Attributes

p and *q* are the attribute values for two data objects.

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

Table 5.1. Similarity and dissimilarity for simple attributes

Euclidean Distance

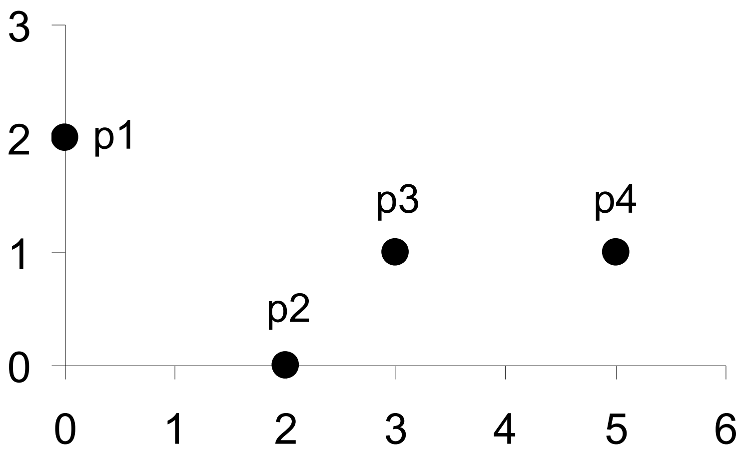
- Euclidean Distance

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

Where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k^th attributes (components) or data objects p and q.

- Standardization is necessary, if scales differ.

Euclidean Distance



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

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

Distance Matrix

Minkowski Distance

- *Minkowski Distance* is a generalization of Euclidean Distance

$$dist = \left(\sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}}$$

Where *r* is a parameter (*order*), *n* is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) of data objects *p* and *q*.

Minkowski Distance: Examples

- *r* = 1. Manhattan distance (L_1 norm, City block, taxicab).
  - 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 (L_2 norm)
- $r \rightarrow \infty$. supremum distance (L_{max} norm, L_∞ norm)
  - This is the maximum difference between any component/attribute of the vectors
- Do not confuse *r* (order) with *n (dimension)*, i.e., all distances are defined for all numbers of dimensions.

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

Distance Matrix

Common Properties of a Distance

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

  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)

  where $d(p, q)$ is the distance (dissimilarity) between points (data objects), p and q.

- A distance that satisfies these properties is a *metric.*

Common Properties of a Similarity

- Similarities, also have some well known properties.

  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)

  where $s(p, q)$ is the similarity between points (data objects), p and q.

Similarity Between Binary Vectors

- A common situation is that objects, *p* and *q*, have only binary attributes
- Compute similarities with following quantities

  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

- *Simple Matching* and *Jaccard Coefficients*

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

SMC versus Jaccard: Example

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)

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

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

Cosine Similarity

- If d_1 and d_2 are two document vectors, then

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

where $\bullet$ indicates vector dot product and $||d||$ is the *length* of vector d.

- Example:

d_1 = **3 2 0 5 0 0 0 2 0 0**

d_2 = **1 0 0 0 0 0 0 1 0 2**

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

$||d_1|| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5}$

$= 6.481$

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

$= 2.245$

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

广义

Extended Jaccard Coefficient (Tanimoto)

- Jaccard Coefficient is mainly for measuring binary attributes
- Extended Jaccard (Tanimoto) Coefficient is a variation of Jaccard for continuous or count attributes
  - Reduces to Jaccard for binary attributes

$$T(p,q) = \frac{p \bullet q}{\|p\|^2 + \|q\|^2 - p \bullet q}$$

Correlation

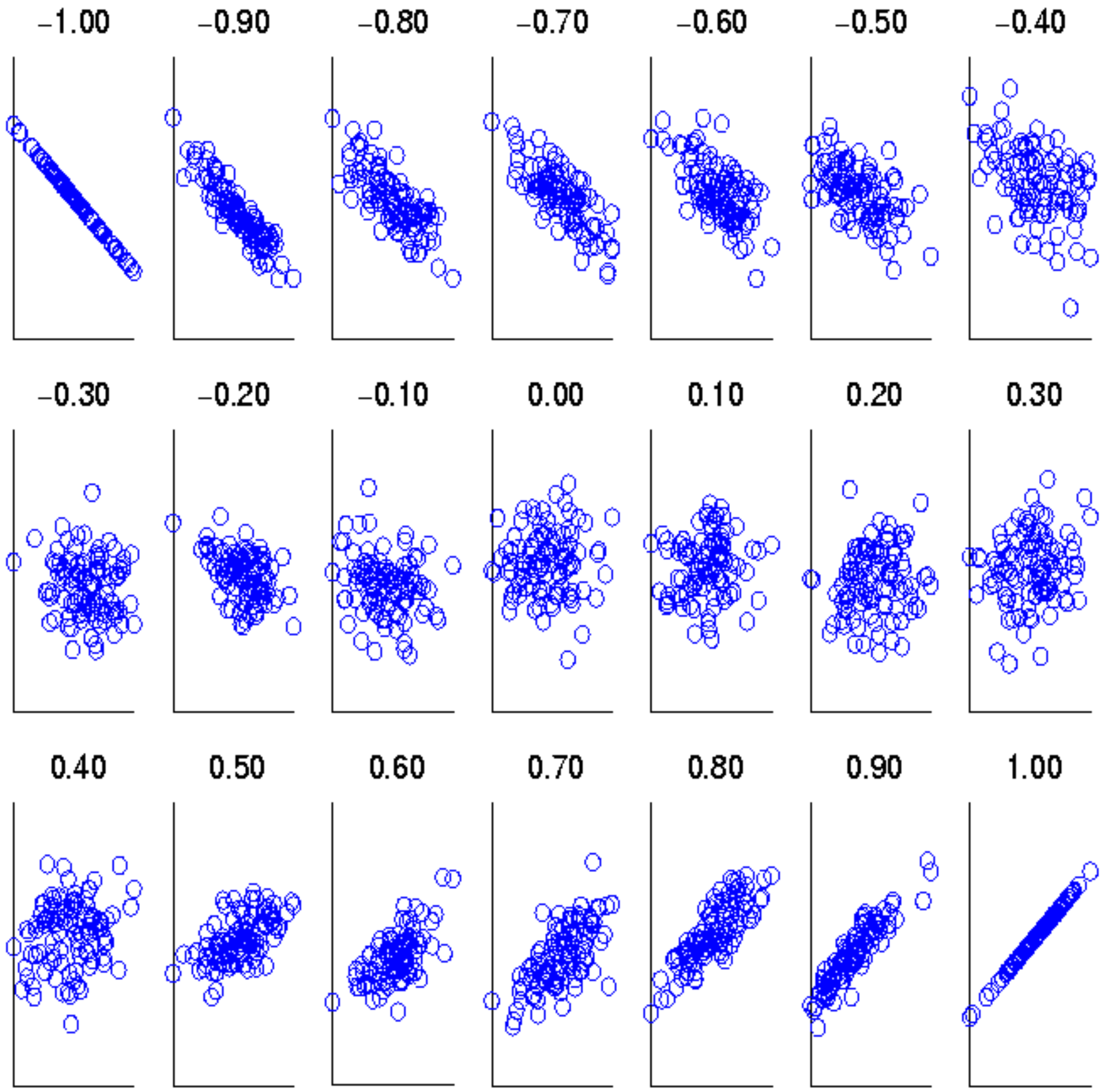
- Correlation measures the *linear relationship* between objects
- To compute correlation, we standardize data objects, *p* and *q*, and then take their dot product 用于二元变量和连续变量

$$p'_k = (p_k - mean(p)) / std(p)$$

$$q'_k = (q_k - mean(q)) / std(q)$$

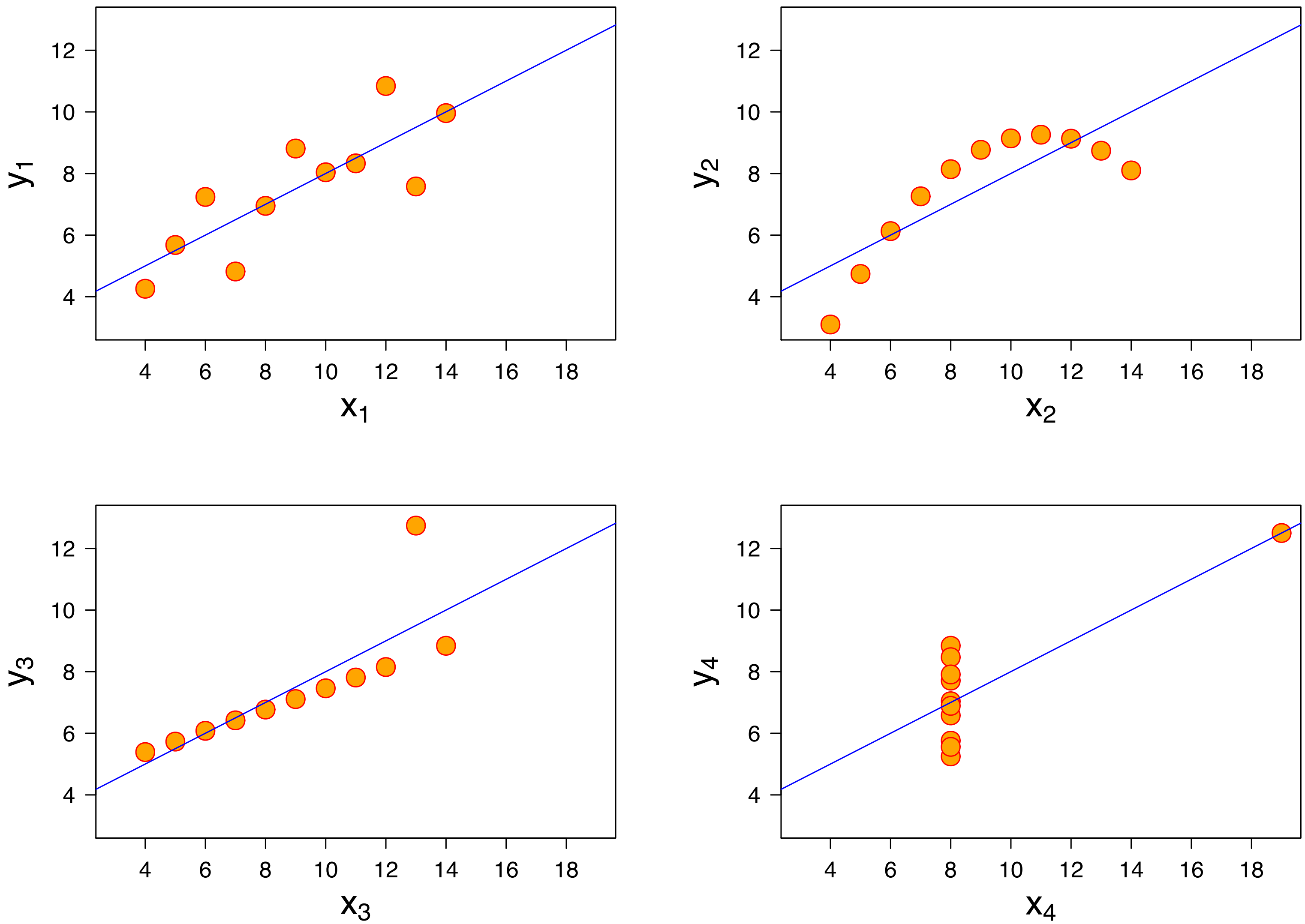
$$correlation(p, q) = p' \bullet q'$$

Visually Evaluating Correlation



Scatter plots showing the correlations from –1 to 1.

Anscombe's Quartet



General Approach for Combining Similarities

- Sometimes attributes are of many different types, but an overall similarity is needed.
- A simple strategy is to take *average* of effective attributes.

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:

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

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

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

Using Weights to Combine Similarities

- May not want to treat all attributes the same.
  - Use weights w_k which are between 0 and 1 and sum to 1.

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

$$distance(p,q) = \left(\sum_{k=1}^{n} w_k |p_k - q_k|^r \right)^{1/r}$$