

Data Mining: Data

Lecture Notes for Chapter 2

Introduction to Data Mining , 2nd Edition
by
Tan, Steinbach, Kumar

Outline

- Attributes and Objects
- Types of Data
- Data Quality
- Similarity and Distance

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, dimension, or feature
- A collection of attributes describe an **object**
 - Object is also known as record, point, case, sample, entity, or instance

Attributes				
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

A More Complete View of Data

- Data may have parts
- Attributes (objects) may have relationships with other attributes (objects)
- More generally, data may have structure
- Data can be incomplete
- We will discuss this in more detail later

Attribute Values

- **Attribute values** are numbers or symbols assigned to an attribute for a particular object
- 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

Measurement of Length

- The way you measure an attribute may not match the attributes properties.



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 from 1-10), grades, height {tall, medium, short}
 - **Interval**
 - ◆ Examples: calendar dates, temperatures in Celsius or Fahrenheit.
 - **Ratio**
 - ◆ Examples: temperature in Kelvin, length, counts, elapsed time (e.g., time to run a race)

Properties of Attribute Values

- The type of an attribute depends on which of the following properties/operations it possesses:
 - Distinctness: $= \neq$
 - Order: $< >$
 - Differences are meaningful : $+ -$
 - Ratios are meaningful $* /$
 - Nominal attribute: distinctness
 - Ordinal attribute: distinctness & order
 - Interval attribute: distinctness, order & meaningful differences
 - Ratio attribute: all 4 properties/operations

Difference Between Ratio and Interval

- Is it physically meaningful to say that a temperature of 10° is twice that of 5° on
 - the Celsius scale?
 - the Fahrenheit scale?
 - the Kelvin scale?

- Consider measuring the height above average
 - If Bill's height is three inches above average and Bob's height is six inches above average, then would we say that Bob is twice as tall as Bill?
 - Is this situation analogous to that of temperature?

		Attribute Type	Description	Examples	Operations
Categorical Qualitative		Nominal	Nominal attribute values only distinguish. (=, ≠)	zip codes, employee ID numbers, eye color, sex: { <i>male</i> , <i>female</i> }	mode, entropy, contingency correlation, χ^2 test
		Ordinal	Ordinal attribute values also order objects. (<, >)	hardness of minerals, { <i>good</i> , <i>better</i> , <i>best</i> }, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric Quantitative		Interval	For interval attributes, differences between values are meaningful. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, <i>t</i> and <i>F</i> tests
		Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, current	geometric mean, harmonic mean, percent variation

This categorization of attributes is due to S. S. Stevens

		Attribute Type	Transformation	Comments
Categorical Qualitative		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}.
Numeric Quantitative		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.

This categorization of attributes is due to S. S. Stevens

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 Attributes

- Only presence (a non-zero attribute value) is regarded as important
 - ◆ Words present in documents
 - ◆ Items present in customer transactions
- If we met a friend in the grocery store would we ever say the following?

“I see our purchases are very similar since we didn’t buy most of the same things.”
- We need two asymmetric binary attributes to represent one ordinary binary attribute
 - Association analysis uses asymmetric attributes
- Asymmetric attributes typically arise from objects that are sets

Key Messages for Attribute Types

- The types of operations you choose should be “meaningful” for the type of data you have
 - Distinctness, order, meaningful intervals, and meaningful ratios are only four properties of data
 - The data type you see – often numbers or strings – may not capture all the properties or may suggest properties that are not present
 - Analysis may depend on these other properties of the data
 - ◆ Many statistical analyses depend only on the distribution
 - Many times what is meaningful is measured by statistical significance
 - But in the end, what is meaningful is measured by the domain

Types of data sets

- Record
 - Data Matrix
 - Document Data
 - Transaction Data
- Graph
 - World Wide Web
 - Molecular Structures
- Ordered
 - Spatial Data
 - Temporal Data
 - Sequential Data
 - Genetic Sequence Data

Important Characteristics of Data

- Dimensionality (number of attributes)
 - ◆ High dimensional data brings a number of challenges
- Sparsity
 - ◆ Only presence counts
- Resolution
 - ◆ Patterns depend on the scale
- Size
 - ◆ Type of analysis may depend on size of data

Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

<i>Tid</i>	<i>Refund</i>	<i>Marital Status</i>	<i>Taxable Income</i>	<i>Cheat</i>
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 a data set can be represented by an m by 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

Document Data

- 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

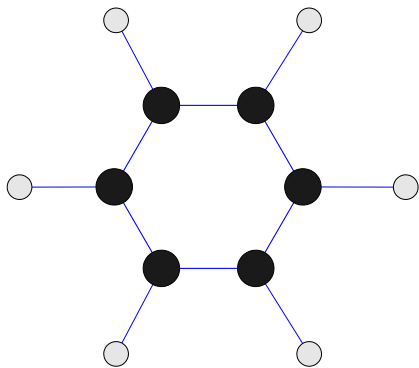
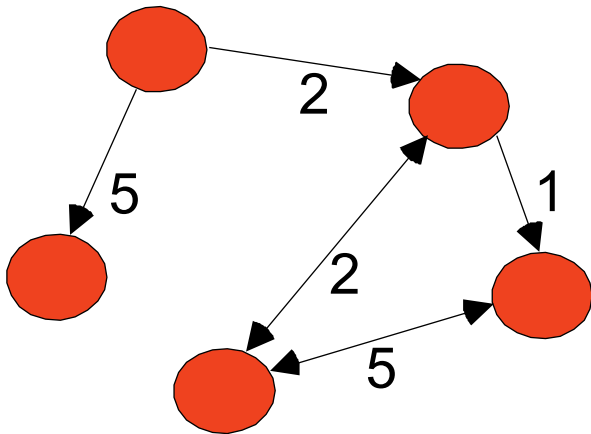
Transaction Data

- A special type of data, where
 - Each transaction involves a set of items.
 - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.
 - Can represent transaction data as record data

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Graph Data

- Examples: Generic graph, a molecule, and webpages



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.

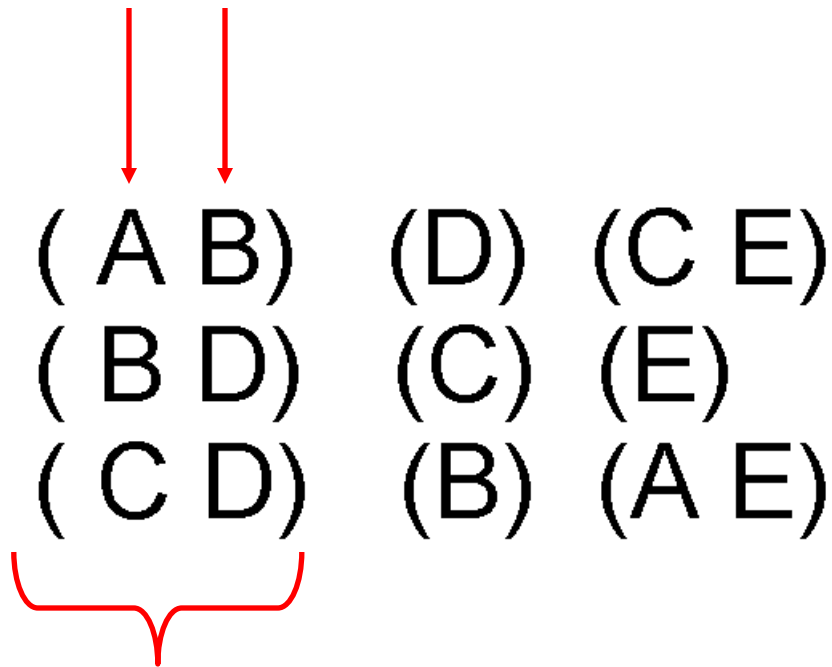
Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

Benzene Molecule: C₆H₆

Ordered Data

□ Sequences of transactions

Items/Events



**An element of
the sequence**

Ordered Data

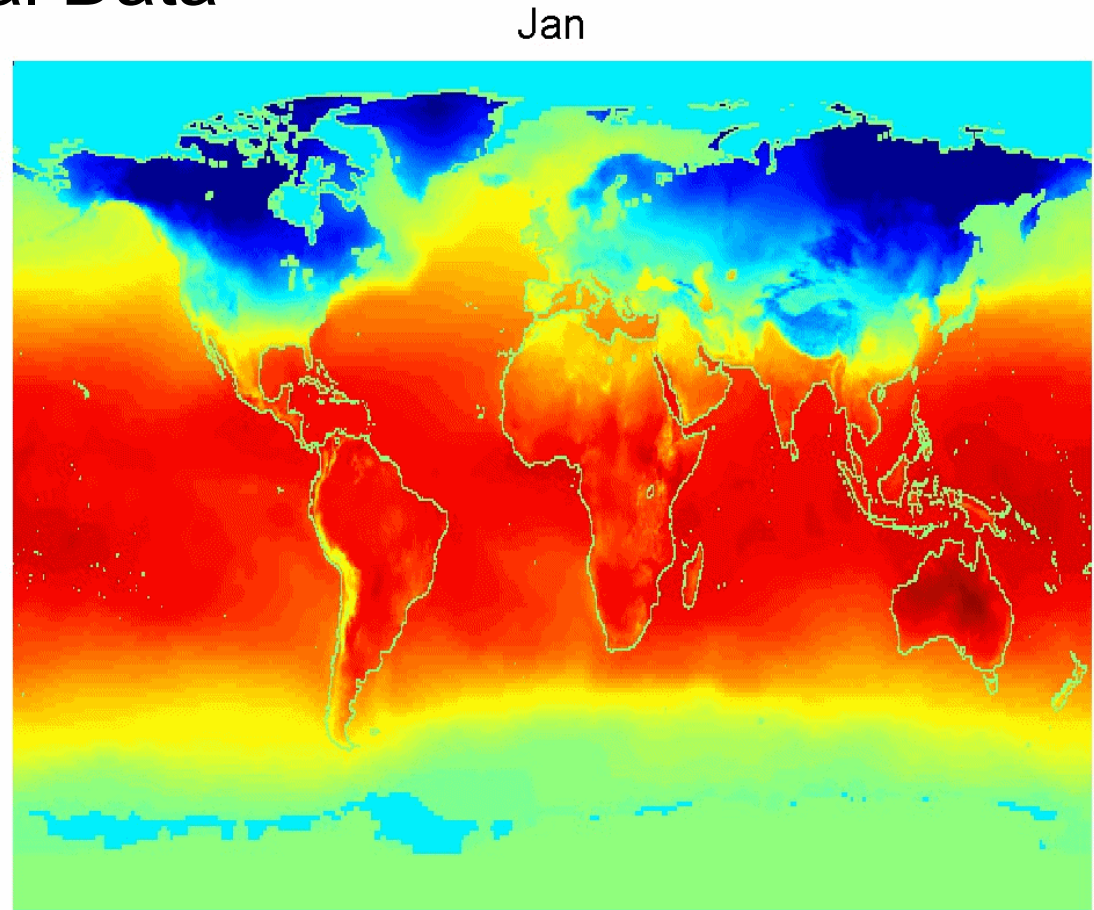
□ Genomic sequence data

**GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCCGCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG**

Ordered Data

□ Spatio-Temporal Data

**Average Monthly
Temperature of
land and ocean**



Data Quality

- Poor data quality negatively affects many data processing efforts

“The most important point is that poor data quality is an unfolding disaster.

- Poor data quality costs the typical company at least ten percent (10%) of revenue; twenty percent (20%) is probably a better estimate.”

Thomas C. Redman, DM Review, August 2004

- Data mining example: a classification model for detecting people who are loan risks is built using poor data
 - Some credit-worthy candidates are denied loans
 - More loans are given to individuals that default

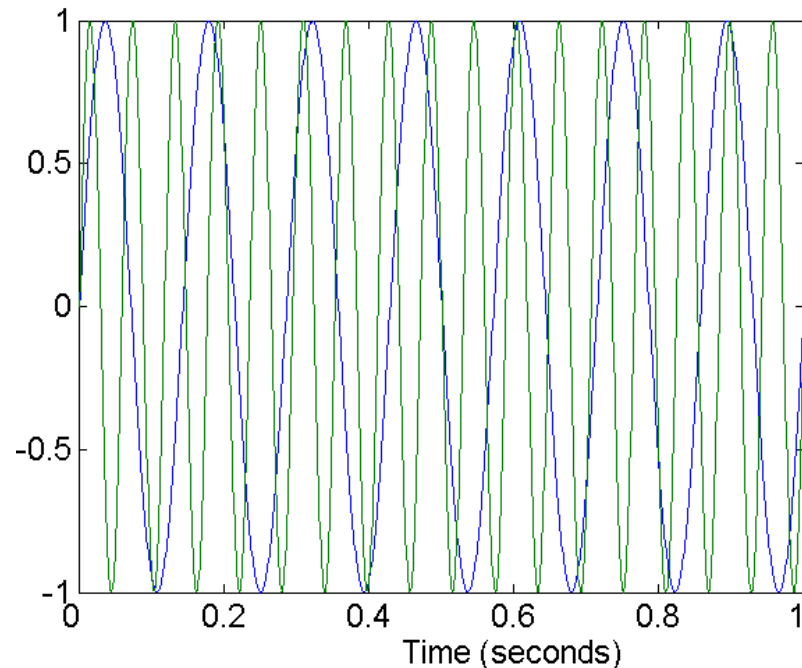
Data Quality ...

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

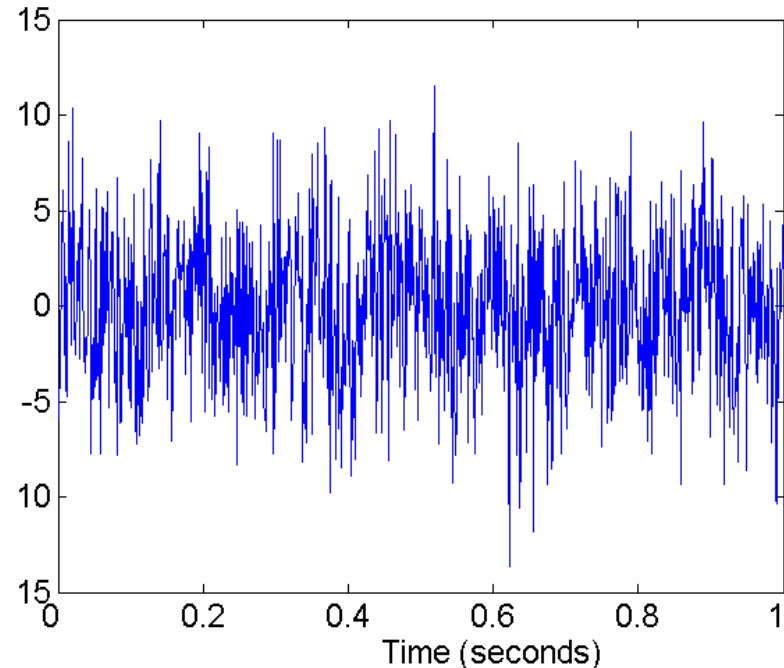
- Examples of data quality problems:
 - Noise and outliers
 - Missing values
 - Duplicate data
 - Wrong data
 - Fake data

Noise

- For objects, noise is an extraneous object
- For attributes, 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



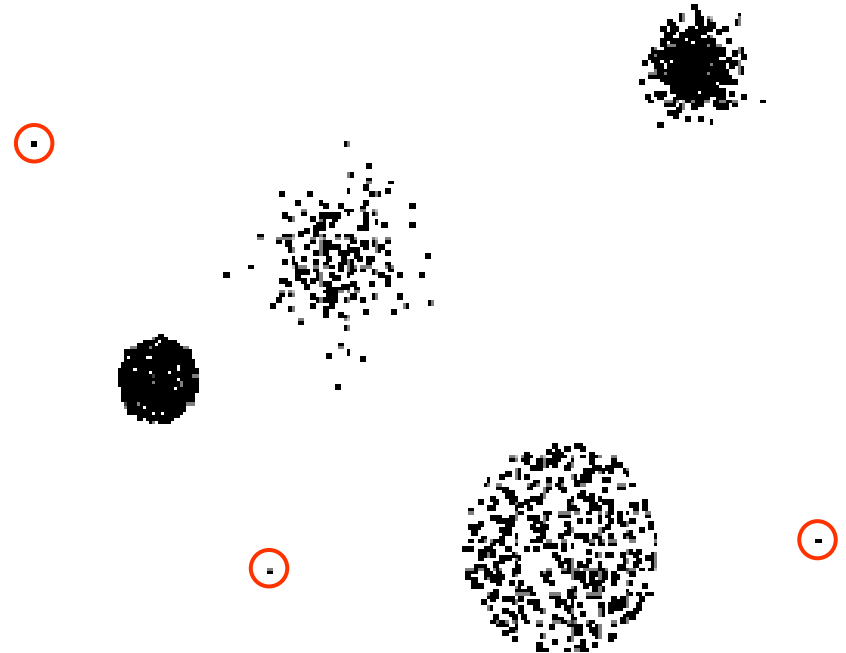
Two Sine Waves



Two Sine Waves + Noise

Outliers

- **Outliers** are data objects with characteristics that are considerably different than most of the other data objects in the data set
 - **Case 1:** Outliers are noise that interferes with data analysis
 - **Case 2:** Outliers are the goal of our analysis
 - ◆ Credit card fraud
 - ◆ Intrusion detection



□ Causes?

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 variables
- Estimate missing values
 - ◆ Example: time series of temperature
 - ◆ Example: census results
- Ignore the missing value during analysis

Missing Values ...

- Missing completely at random (MCAR)
 - Missingness of a value is independent of attributes
 - Fill in values based on the attribute
 - Analysis may be unbiased overall
- Missing at Random (MAR)
 - Missingness is related to other variables
 - Fill in values based other values
 - Almost always produces a bias in the analysis
- Missing Not at Random (MNAR)
 - Missingness is related to unobserved measurements
 - Informative or non-ignorable missingness
- Not possible to know the situation from the data

Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
 - Major issue when merging data from heterogeneous sources
- Examples:
 - Same person with multiple email addresses
- Data cleaning
 - Process of dealing with duplicate data issues
- When should duplicate data not be removed?



Data Preprocessing

Data Preprocessing

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

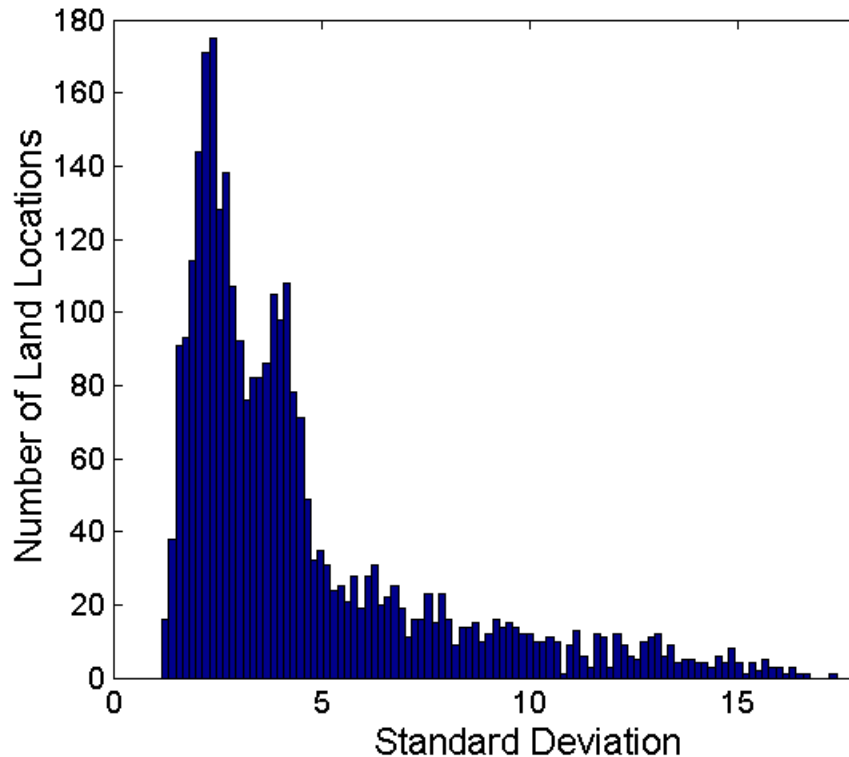
Aggregation

- Combining two or more objects into a single object

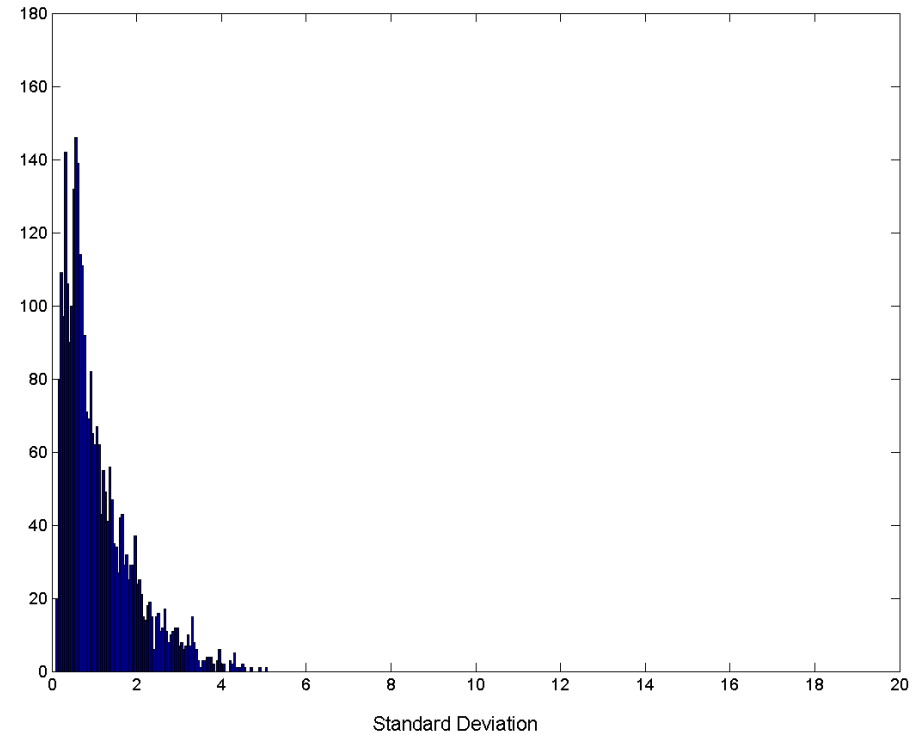
- Purpose
 - Data reduction
 - ◆ Reduce the number of attributes or objects
 - Change of scale
 - ◆ Cities aggregated into regions, states, countries, etc
 - More “stable” data
 - ◆ Aggregated data tends to have less variability

Aggregation

Variation of Precipitation in Australia



**Standard Deviation of Average
Monthly Precipitation**



**Standard Deviation of Average
Yearly Precipitation**

Sampling

- Sampling is the main technique employed for data selection.
 - 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.

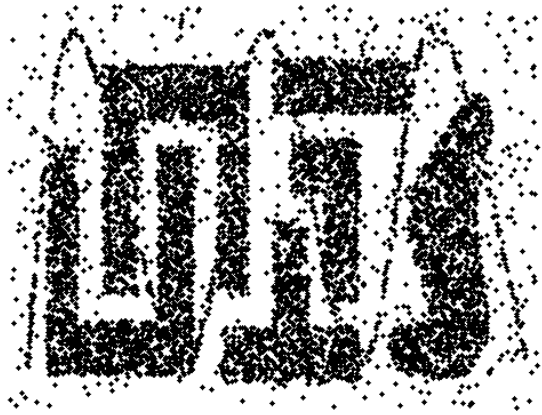
Sampling ...

- 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

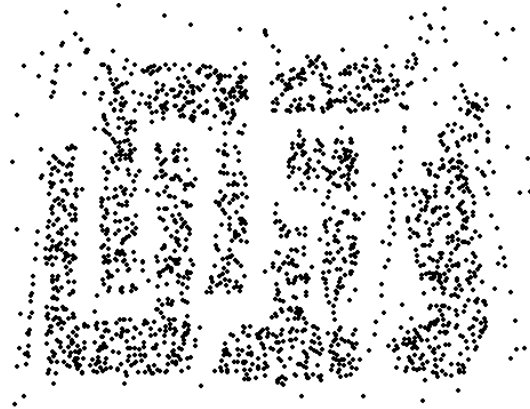
Types of Sampling

- Simple Random Sampling
 - There is an equal probability of selecting any particular item
- Sampling without replacement
 - As each item is selected, it is removed from the population
- Sampling with replacement
 - 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
 - **Split the data into several partitions**; then draw random samples from each partition

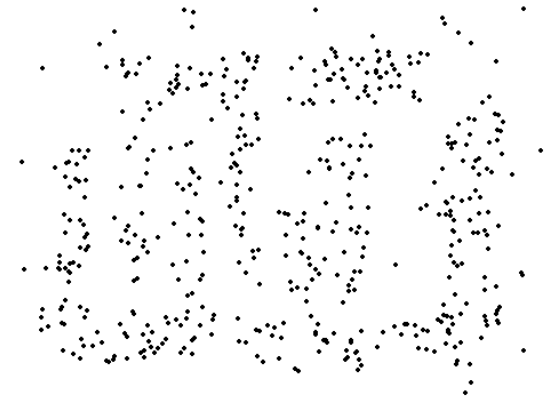
Sample Size



8000 points



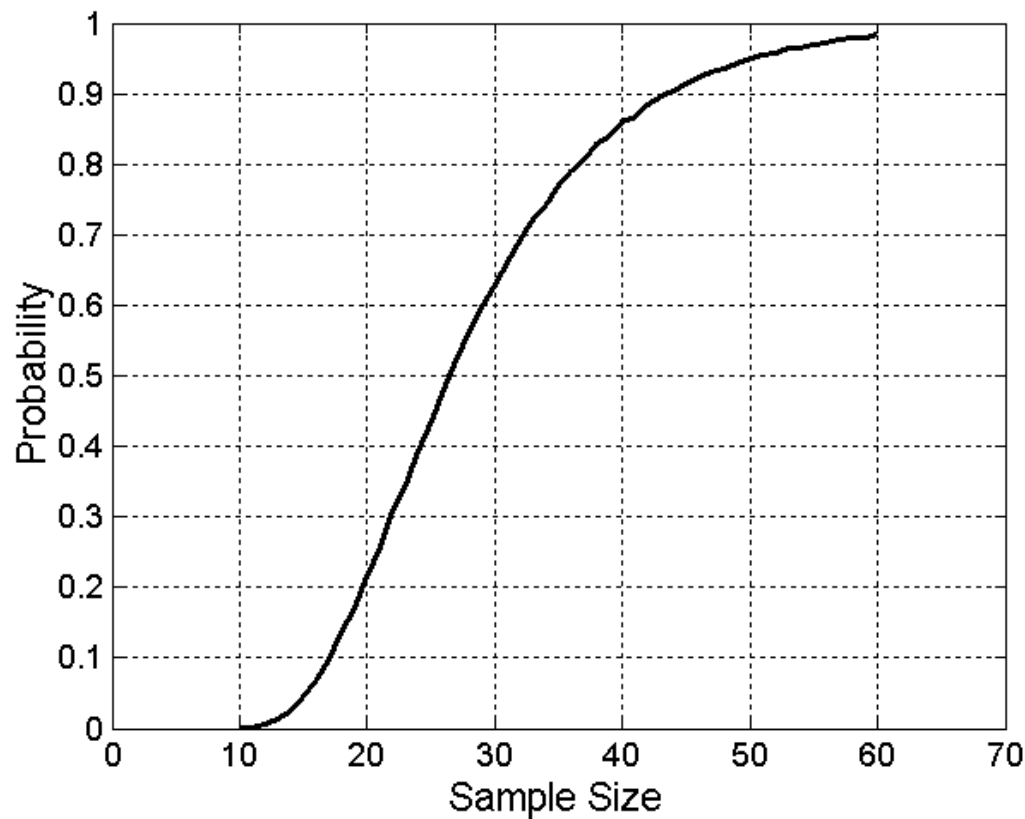
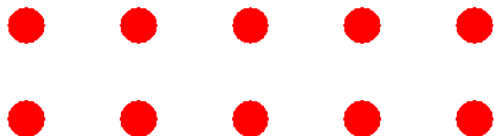
2000 Points



500 Points

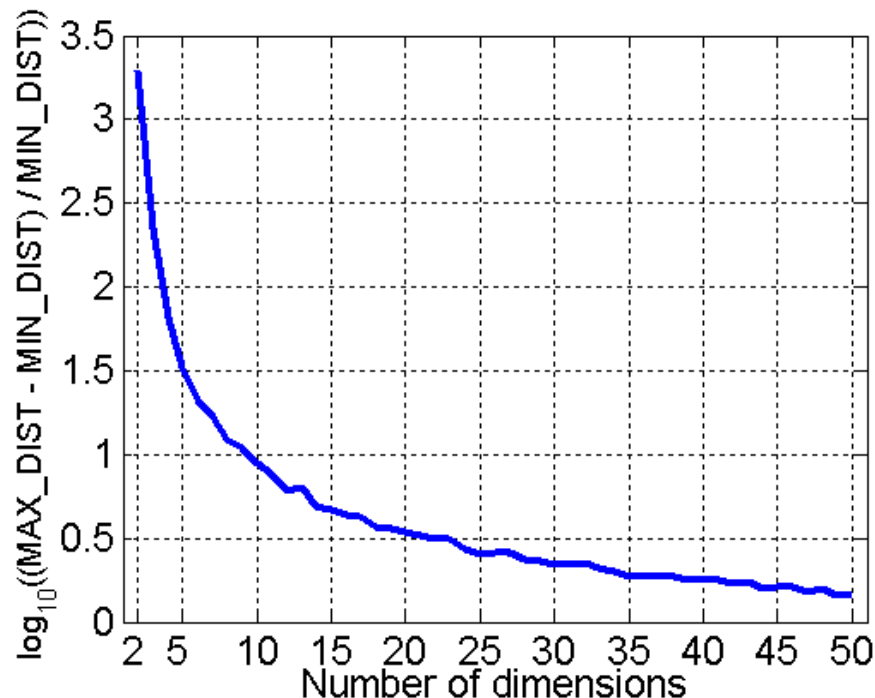
Sample Size

- 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

As the number of dimensions increases, the time of most algorithms increase exponentially in data mining.

Dimensionality Reduction

□ Purpose:

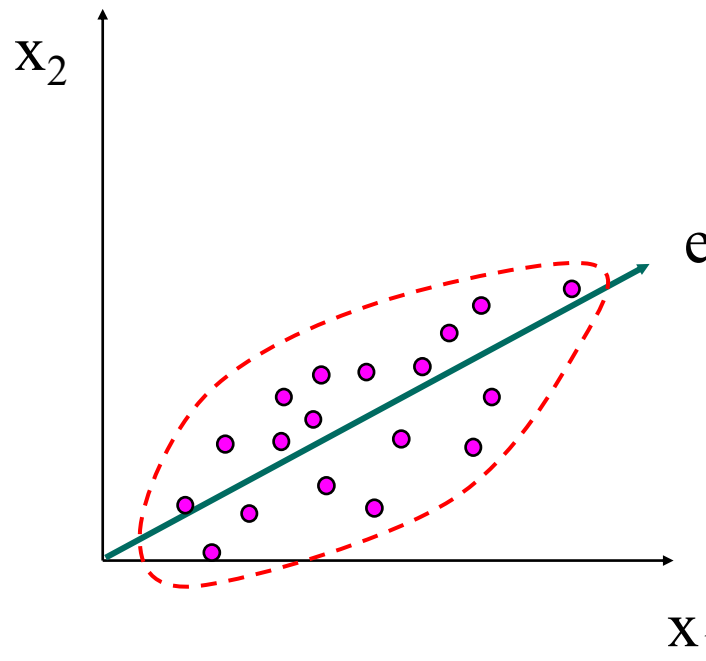
- Avoid curse of dimensionality
- Reduce amount of time and memory required by data mining algorithms
- Allow data to be more easily visualized
- May help to eliminate irrelevant features or reduce noise

□ Techniques

- Principle Component Analysis
- Singular Value Decomposition
- Others: supervised and non-linear techniques

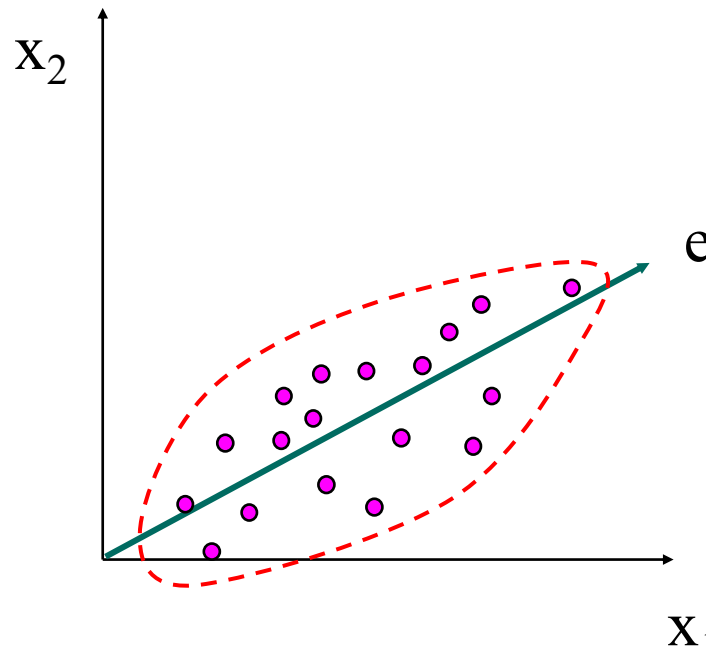
Dimensionality Reduction: PCA

- Goal is to find a projection that captures the largest amount of variation in data



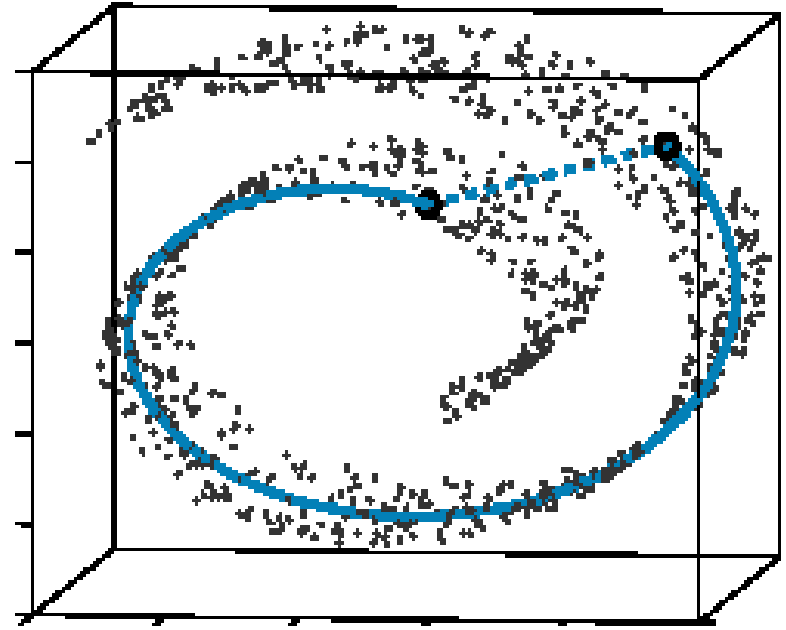
Dimensionality Reduction: PCA

- 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



Feature Subset Selection

- Another way to reduce dimensionality of data
- 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 Subset Selection

□ Techniques:

- Brute-force 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
- Filter approaches:
 - ◆ Features are selected before data mining algorithm is run
- Wrapper approaches:
 - ◆ Use the data mining algorithm as a black box to find best subset of attributes

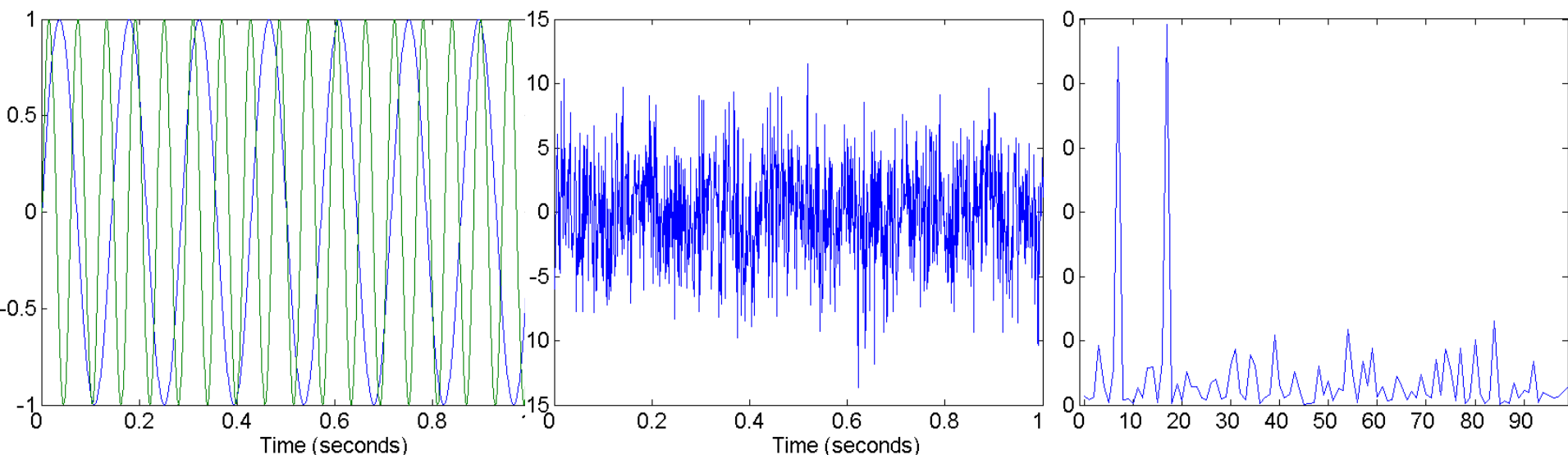
Feature Creation

- 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
 - Mapping Data to New Space
 - Feature Construction
 - ◆ combining features

Mapping Data to a New Space

- Fourier transform
- Wavelet transform



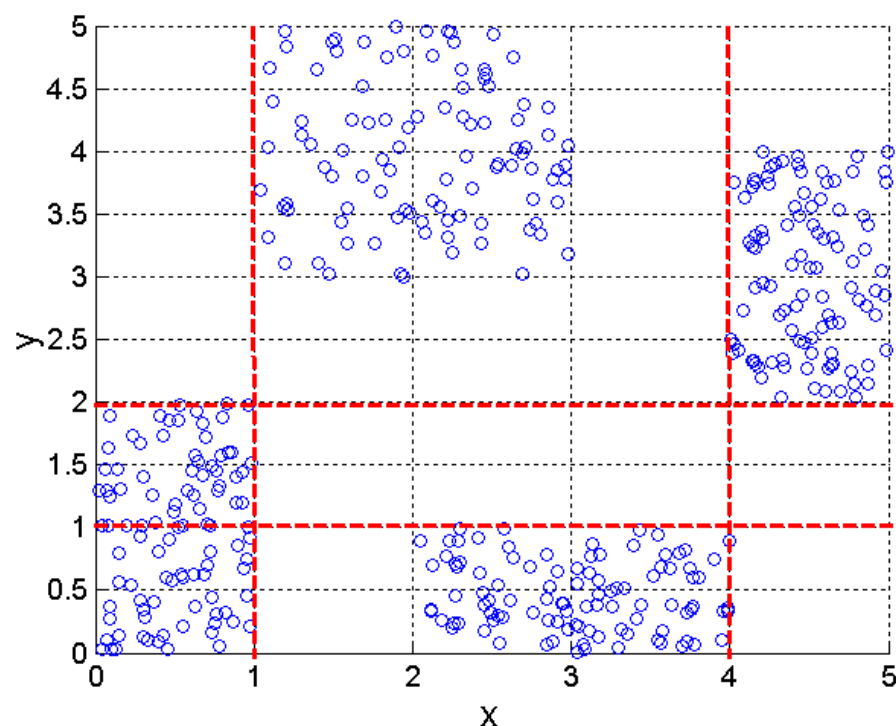
Two Sine Waves

Two Sine Waves + Noise

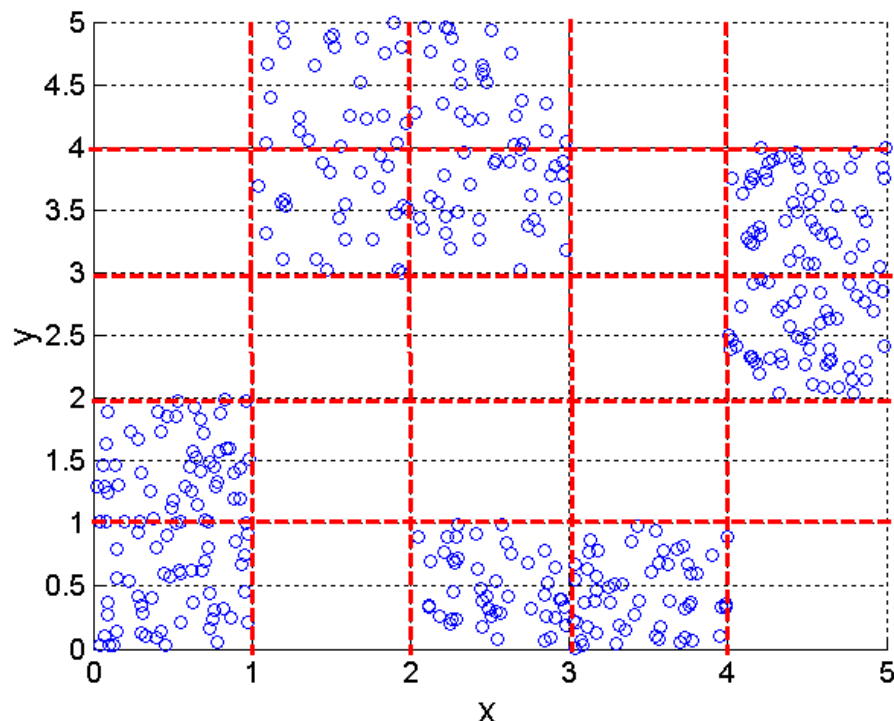
Frequency

Discretization Using Class Labels

□ Entropy based approach

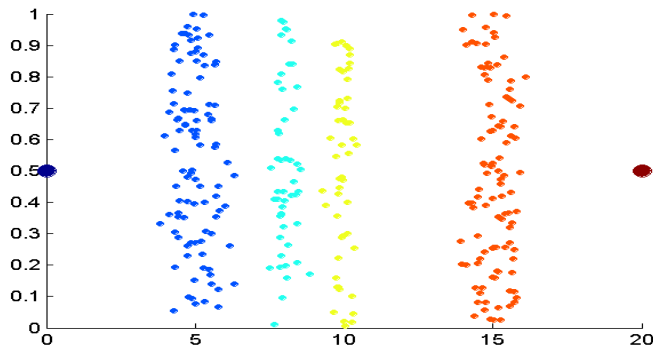


3 categories for both x and y

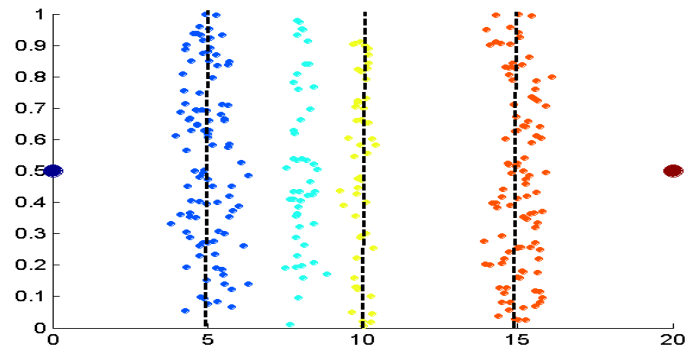


5 categories for both x and y

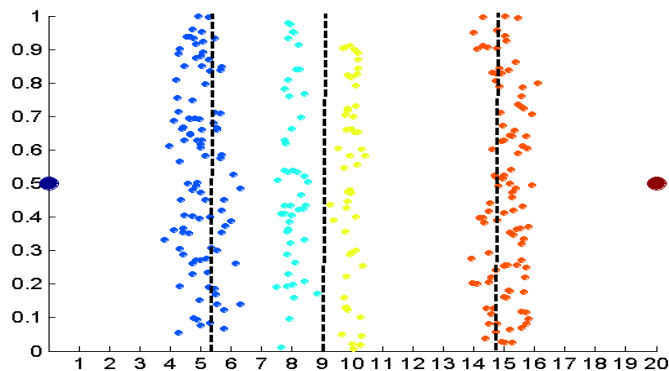
Discretization Without Using Class Labels



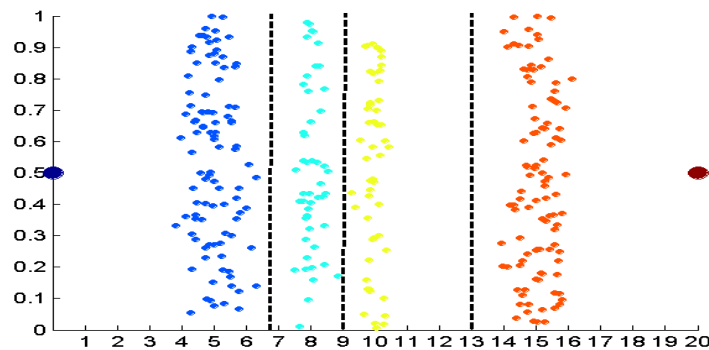
Data



Equal interval width



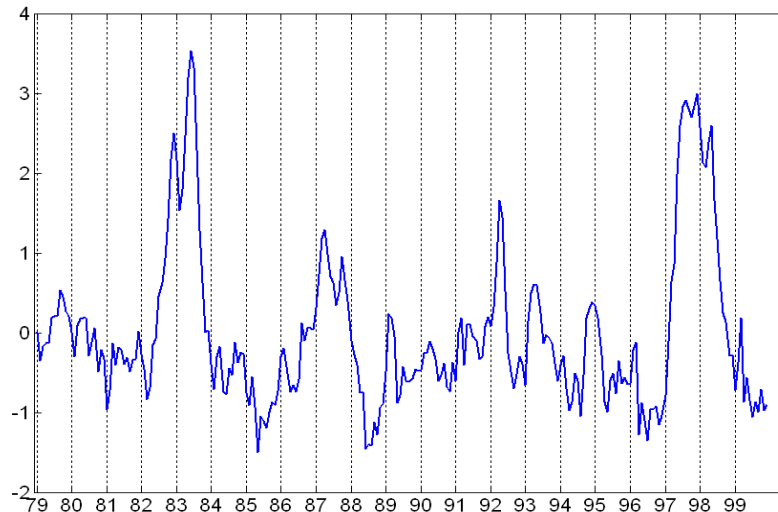
Equal frequency



K-means

Attribute Transformation

- 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 Measures

□ Similarity measure

- Numerical measure of how alike two data objects are.
- Is higher when objects are more alike.
- Often falls in the range $[0,1]$

□ Dissimilarity measure

- Numerical measure of how different two data objects are
- Lower when objects are more alike
- Minimum dissimilarity is often 0
- Upper limit varies

□ **Proximity** refers to a similarity or dissimilarity

Similarity/Dissimilarity for Simple Attributes

The following table shows the similarity and dissimilarity between two objects, x and y , with respect to a single, simple attribute.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = x - y / (n - 1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = x - y $	$s = -d, s = \frac{1}{1+d}, s = e^{-d},$ $s = 1 - \frac{d - \min_d}{\max_d - \min_d}$

Euclidean Distance

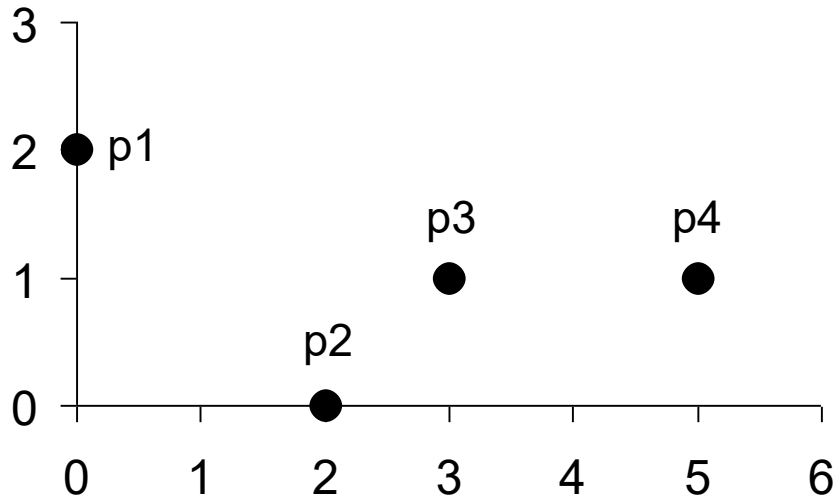
□ Euclidean Distance

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

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} .

□ 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

Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well known properties.
 1. $d(\mathbf{x}, \mathbf{y}) \geq 0$ for all \mathbf{x} and \mathbf{y} and $d(\mathbf{x}, \mathbf{y}) = 0$ only if $\mathbf{x} = \mathbf{y}$. (Positive definiteness)
 2. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ for all \mathbf{x} and \mathbf{y} . (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} . (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**

Common Properties of a Similarity

□ 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} .

Similarity Between Binary Vectors

- Common situation is that objects, \mathbf{x} and \mathbf{y} , have only binary attributes

- Compute similarities using the following quantities

f_{01} = the number of attributes where \mathbf{x} was 0 and \mathbf{y} was 1

f_{10} = the number of attributes where \mathbf{x} was 1 and \mathbf{y} was 0

f_{00} = the number of attributes where \mathbf{x} was 0 and \mathbf{y} was 0

f_{11} = the number of attributes where \mathbf{x} was 1 and \mathbf{y} was 1

- Simple Matching and 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

$$\mathbf{x} = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0$$

$$\mathbf{y} = 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1$$

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

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

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

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

$$\begin{aligned}\text{SMC} &= (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00}) \\ &= (0+7) / (2+1+0+7) = 0.7\end{aligned}$$

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

Cosine Similarity

□ If \mathbf{d}_1 and \mathbf{d}_2 are two document vectors, then

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

where $\langle \mathbf{d}_1, \mathbf{d}_2 \rangle$ indicates inner product or vector dot product of vectors, \mathbf{d}_1 and \mathbf{d}_2 , and $\|\mathbf{d}\|$ is the length of vector \mathbf{d} .

□ Example:

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

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

$$\langle \mathbf{d}_1, \mathbf{d}_2 \rangle = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

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

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

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

Correlation measures the linear relationship between objects

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

where we are using the following standard statistical notation and definitions

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

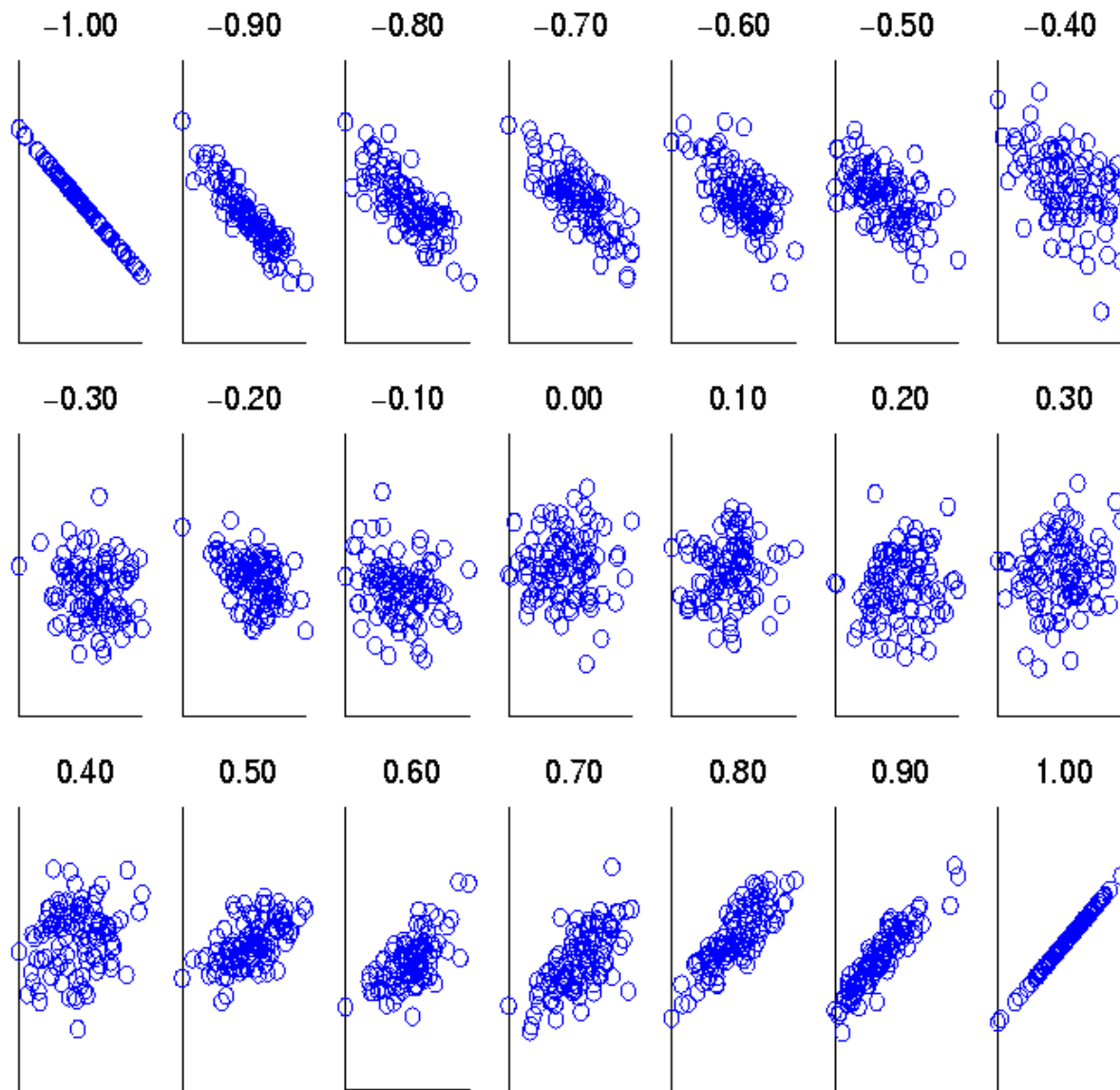
$$\text{standard_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2}$$

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

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

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

Visually Evaluating Correlation



**Scatter plots
showing the
similarity from
-1 to 1.**

Drawback of Correlation

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

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

$$y_i = x_i^2$$

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

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

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

Comparison of Proximity Measures

- Domain of application
 - Similarity measures tend to be specific to the type of attribute and data
 - Record data, images, graphs, sequences, 3D-protein structure, etc. tend to have different measures
- However, one can talk about various properties that you would like a proximity measure to have
 - Symmetry is a common one
 - Tolerance to noise and outliers is another
 - Ability to find more types of patterns?
 - Many others possible
- The measure must be applicable to the data and produce results that agree with domain knowledge