Getting to Know Your Data

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

School Of Computer Engineering

Datamining and Data warehousing (CS 2004)

Dr. Amiya Ranjan Panda
Assistant Professor [II]
School of Computer Engineering,
Kalinga Institute of Industrial Technology (KIIT),
Deemed to be University,Odisha

3 Credit

Lecture Note 02

Acknoledgement

A Special

Thanks to

J. Han and M. Kamber.

&

Tan, Steinbach, Kumar

for their slides and books, which I have used for preparation of these slides.

Chapter Contents

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- DataVisualization
- Measuring Data Similarity and Dissimilarity
- Summary

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**

Data Objects & Attribute

- Data sets are made up of data objects.
- A **data object** represents an entity.
- Examples:
  - ✓ sales database: customers, store items, sales
  - ✓ medical database: patients, treatments
  - ✓ university database: students, professors, courses
- Also called *samples* , *examples, instances, data points, objects, tuples*.
- Data objects are described by **attributes**.
- Database rows -> data objects; columns ->attributes.
- **Attribute** (or **dimensions, features, variables**): a data field, representing a characteristic or feature of a data object.

  E.g., customer _ID, name, address
- Types:
  - ➢ Nominal
  - ➢ Binary
  - ➢ Numeric: quantitative
    - ✓ Interval-scaled, Ratio-scaled

Attribute Values

- ❑ 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

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 in {tall, medium, short}
  - ➢ Interval
    - ✓ Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - ➢ Ratio
    - ✓ Examples: temperature in Kelvin, length, time, counts

Types of Attributes...

- ❑ **Nominal :** Nominal means relating to names. The values of Nominal attributes are symbols or names of things. For example:
  - ✓ *Hair_color = {auburn, black, blond, brown, grey, red, white}*
  - ✓ marital status, occupation, ID numbers, zip codes
  - ✓ The values don't have any meaningful order about them.
- ❑ **Binary:** Nominal attribute woth only two states (0 and 1), where 0 typically means the attribute is absent, and 1 means that it is present. Binary attributes are reffered to as Boolean if the two states correspond to true and false.
  - ➢ **Symmetric binary:** both outcomes equally important
    - ✓ e.g., gender
  - ➢ **Asymmetric binary:** outcomes not equally important.
    - ✓ e.g., medical test (positive vs. negative)
    - ✓ Convention: assign 1 to most important outcome (e.g., HIV positive)

Types of Attributes...

- ❑ **Ordinal**
  - ✓ Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - ✓ Size = {small, medium, large}, grades, army rankings
  - ✓ Other examples of ordinal attributes include Grade (e.g., A+, A, A−, B+, and so on) and Professional rank. Professional ranks can be enumerated in a sequential order, such as assistant, associate, and full for professors,
- ❑ The central tendency of an ordinal attribute can be represented by its mode and its median (the middle value in an ordered sequence), but the mean cannot be defined.
- ❑ Qualitative attributes are describes a feature of an object, without giving an actual size or quantity. The values of such qualitative attributes are typically words representing categories.

Types of Attributes...

- ❑ Quantity (integer or real-valued)
- ❑ **Interval**
  - ➢ Measured on a scale of **equal-sized units**
  - ➢ Values have order
    - ✓ E.g., *temperature in C°or F°, calendar dates*
  - ➢ No true zero-point
- ❑ **Ratio**
  - ➢ Inherent **zero-point**
  - ➢ We can speak of values as being an order of magnitude larger than the unit of measurement (10 K° is twice as high as 5 K°).
    - ✓ e.g., *temperature in Kelvin, length, counts, monetary quantities*

Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:

Distinctness:	= ≠
Order:	< >
Addition:	+ -
Multiplication:	* /

- Nominal attribute: distinctness
- Ordinal attribute: distinctness & order
- Interval attribute: distinctness, order & addition
- Ratio attribute: all 4 properties

Properties of Attribute Values...

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, $\chi 2$ 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

Properties of Attribute Values...

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.
  - 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.

Types of data sets

- ❑ Record
  - ➢ Relational records
  - ➢ Data matrix, e.g., numerical matrix, crosstabs
  - ➢ Document data: text documents: term-frequency vector
  - ➢ Transaction data
- ❑ Graph
  - ➢ World Wide Web
  - ➢ Social and infomation network
  - ➢ Molecular Structures
- ❑ Ordered
  - ➢ Video data: sequence of images
  - ➢ Temporal data: time-series
  - ➢ Sequential Data: transaction sequences
  - ➢ Genetic sequence data
- ❑ Spatial, image and multimedia
  - ➢ Spatial data: maps
  - ➢ Image data

Important Characteristics of Structured Data

- ❑ Dimensionality
  - ➢ Curse of Dimensionality
- ❑ Sparsity
  - ➢ Only presence counts
- ❑ Resolution
  - ➢ Patterns depend on the scale
- ❑ Distribution
  - ➢ Centrality and dispersion

Record Data

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

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 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

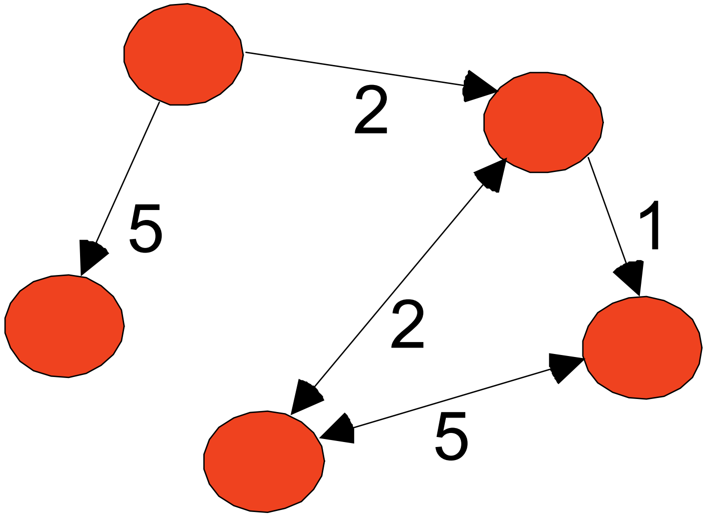
Transaction Data

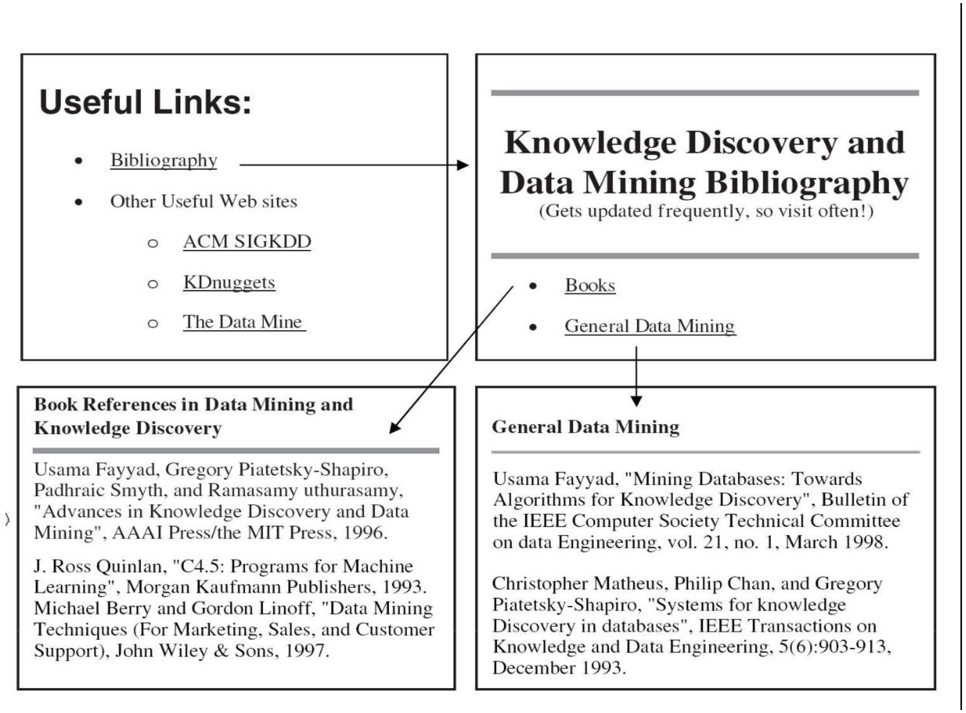
- A special type of record data, where
  - each record (transaction) involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

TID	*Items*
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 and HTML Links





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.

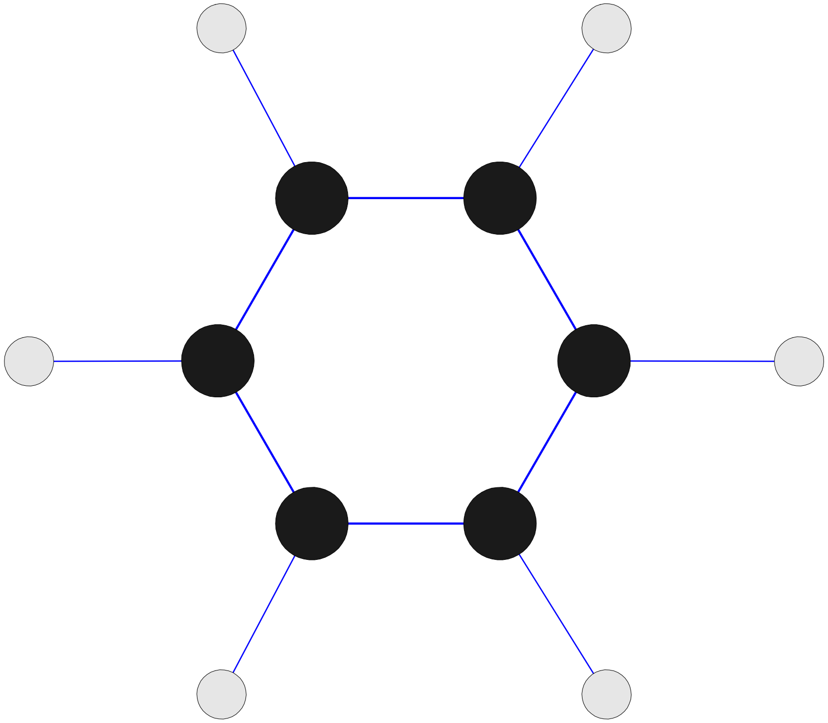
General Data Mining

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.

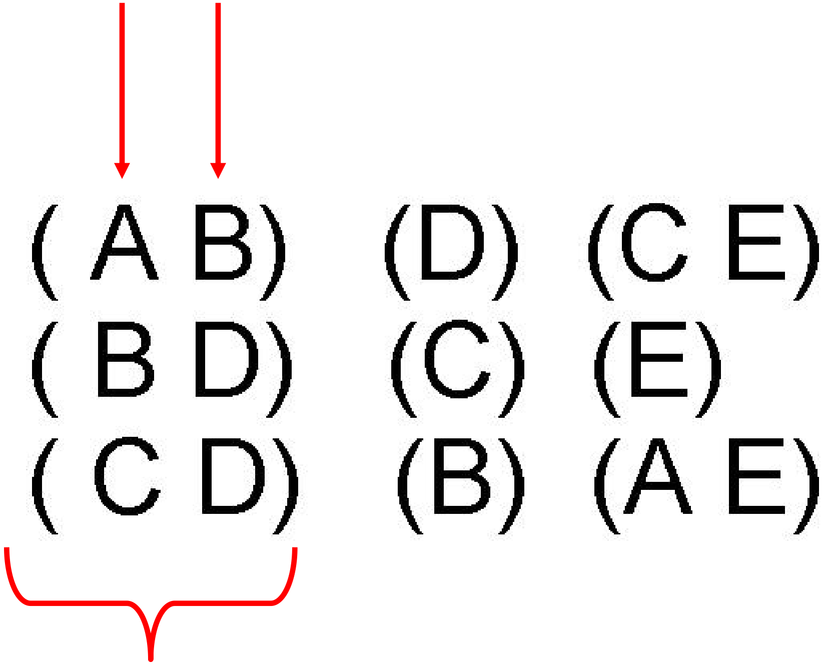
Chemical Data

- ❑ Benzene Molecule: C_6H_6



Ordered Data

Sequences of transactions



Items/Events

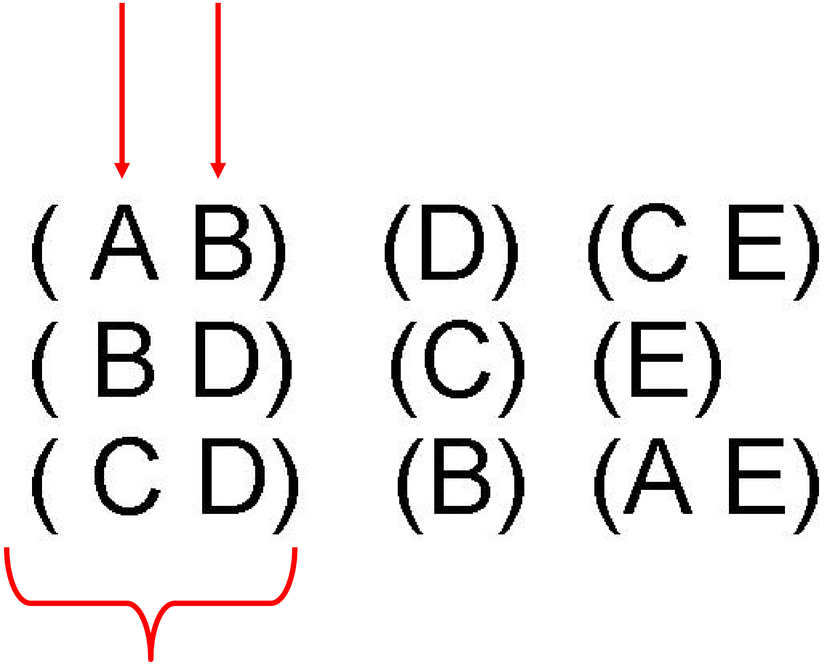
(A B) (D) (C E)
(B D) (C) (E)
(C D) (B) (A E)

An element of the sequence

Ordered Data

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

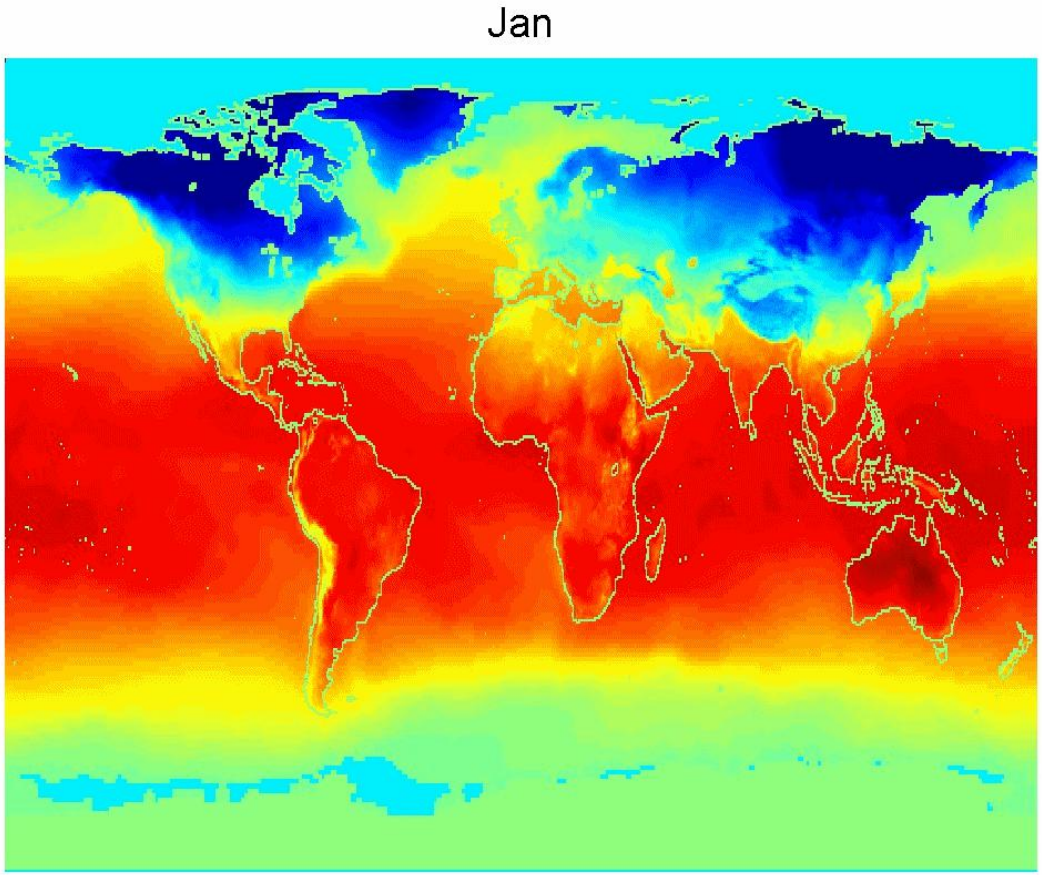
- ❑ Genomic sequence data

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

Ordered Data

- ❑ Spatio-Temporal Data

Average Monthly Temperature of land and ocean



Data Quality

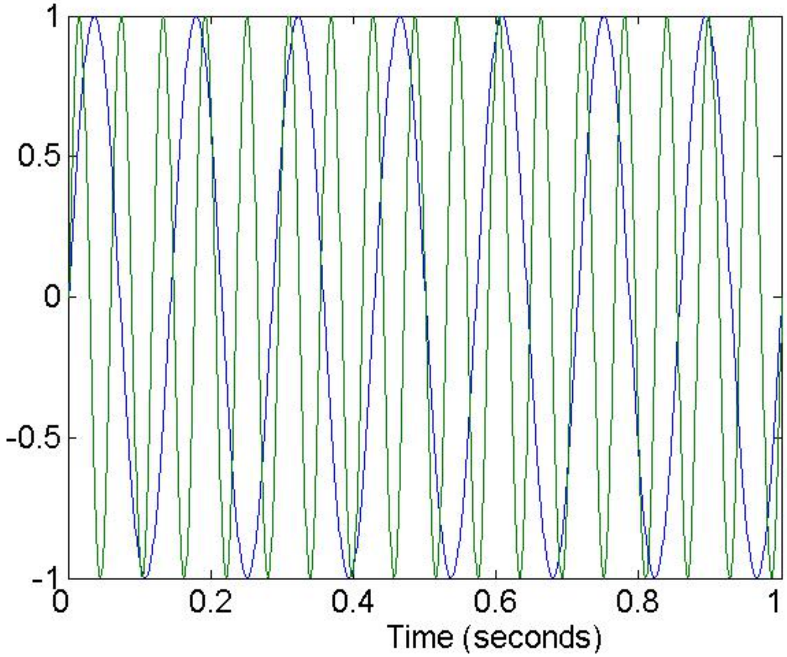
- ❑ What kinds of data quality problems?
- ❑ How can we detect problems with the data?
- ❑ What can we do about these problems?

Data Preparation = Cleaning the Data

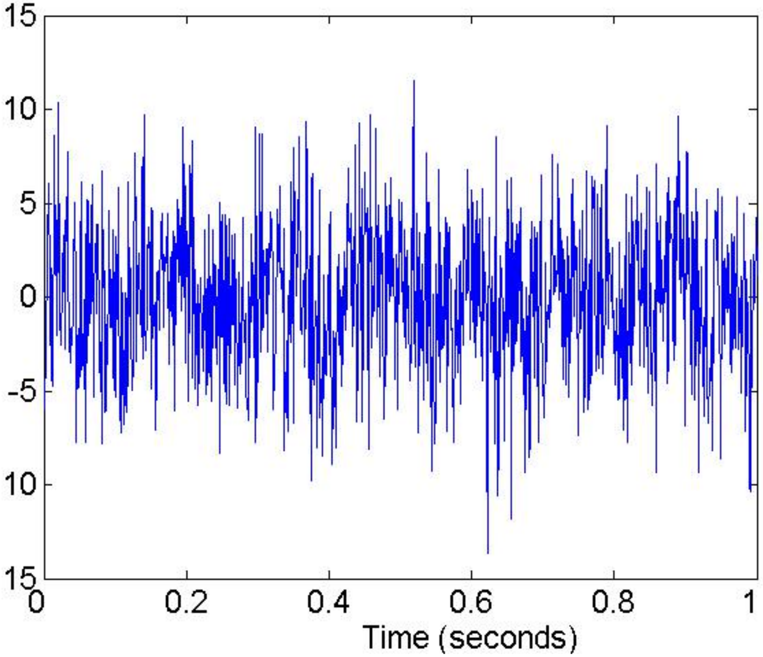
- ❑ Data Preparation can take **40-80%** (or more) of the effort in a data mining project
  - ✓ Dealing with NULL (missing) values
  - ✓ Dealing with errors
  - ✓ Dealing with noise
  - ✓ Dealing with outliers (unless that is your science!)
  - ✓ Transformations: units, scale, projections
  - ✓ Data normalization
  - ✓ Relevance analysis: Feature Selection
  - ✓ Remove redundant attributes
  - ✓ Dimensionality Reduction

Noise

- ❑ For objects, noise is an extraneous object
- ❑ For attibutes, 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



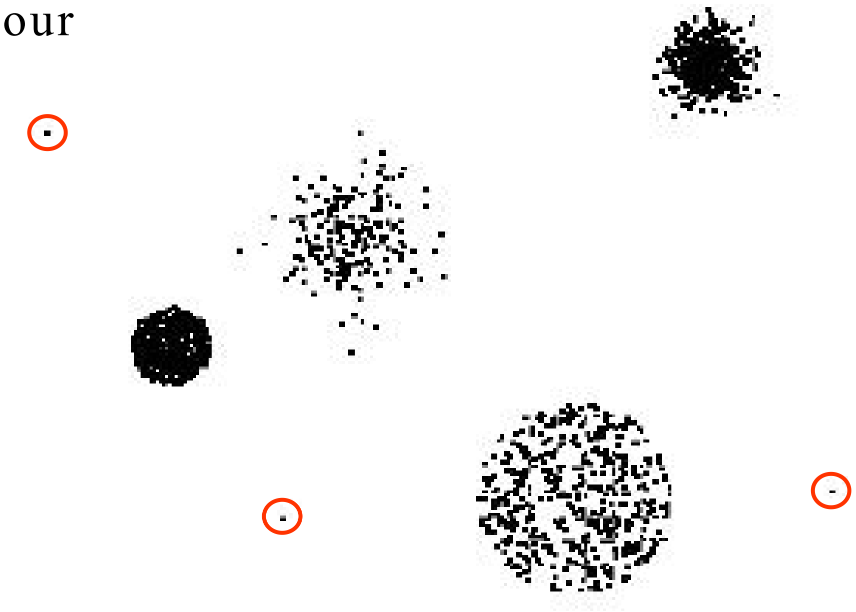
Two Sine Waves



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



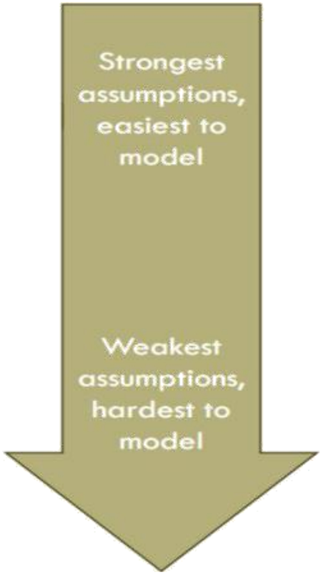
Missing Values

- ❑ 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
  - ➢ Estimate Missing Values
  - ➢ Ignore the Missing Value During Analysis
  - ➢ Replace with all possible values (weighted by their probabilities)

Types of Missing Values

- □ Some definitions are based on representation: Missing data is the lack of a recorded answer for a particular field.
  - ➢ Missing completely at random (MCAR)
  - ➢ Missing at Random (MAR)
  - ➢ Missing Not at Random (MNAR)



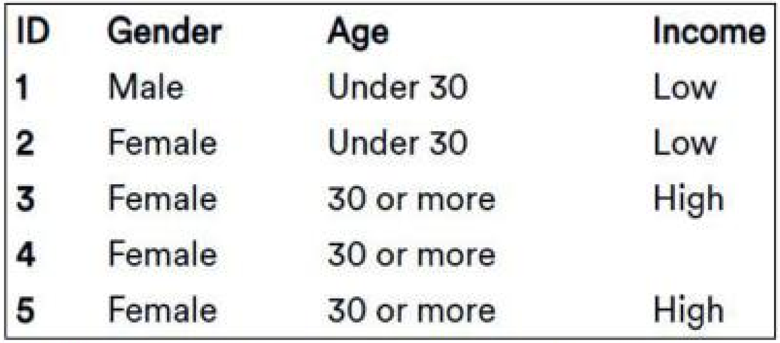
Strongest assumptions, easiest to model

Weakest assumptions, hardest to model

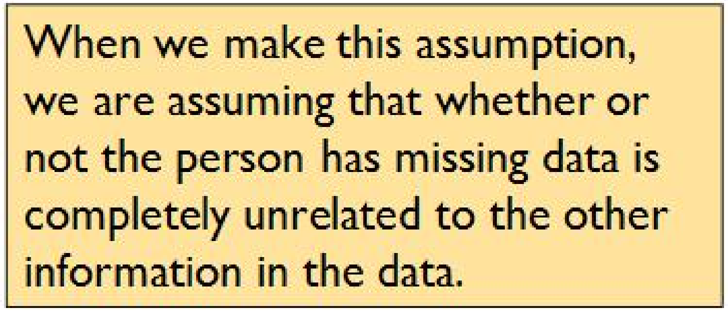
> "All I know is that you throw out missing data or fill it in and make an informative note of it."

Missing Completely at Random (MCAR)

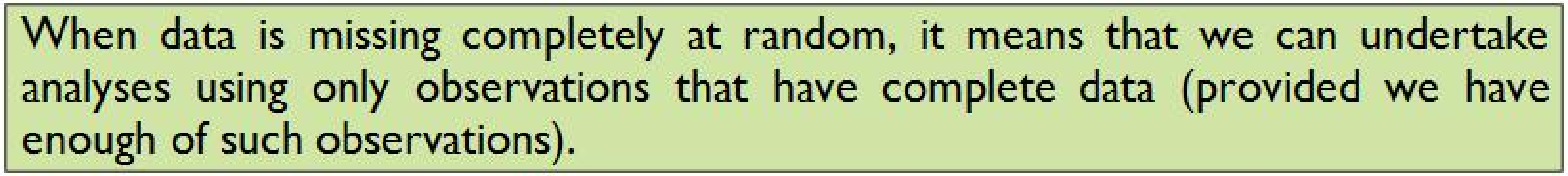
- ❑ Missingness of a value is independent of attributes
  - ✓ Fill in values based on the attribute
  - ✓ Analysis may be unbiased overall
- ❑ The missingness on the variable is completely unsystematic.
- ❑ Example when we take a random sample of a population, where each member has the same chance of being included in the sample.



ID	Gender	Age	Income
1	Male	Under 30	Low
2	Female	Under 30	Low
3	Female	30 or more	High
4	Female	30 or more	
5	Female	30 or more	High



When we make this assumption, we are assuming that whether or not the person has missing data is completely unrelated to the other information in the data.



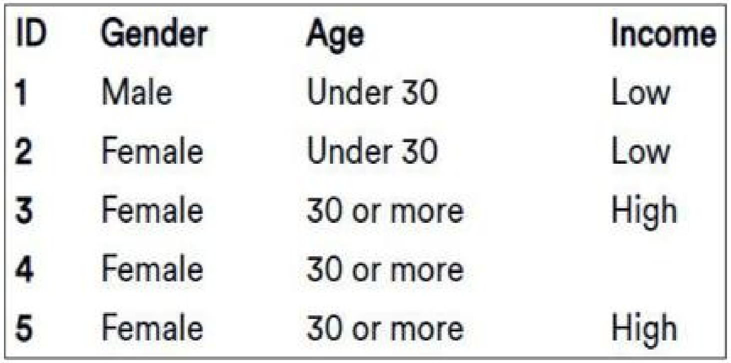
When data is missing completely at random, it means that we can undertake analyses using only observations that have complete data (provided we have enough of such observations).

Missing at Random (MAR)

- ❑ Missingness is related to other variables
- ❑ Fill in values based other values
- ❑ Almost always produces a bias in the analysis

Example of MAR is when we take a sample from a population, where the probability to be included depends on some known property.

A simple predictive model is that income can be predicted based on gender and age. Looking at the table, we note that our missing value is for a Female aged 30 or more, and observations says the other females aged 30 or more have a High income. As a result, we can predict that the missing value should be High.



ID	Gender	Age	Income
1	Male	Under 30	Low
2	Female	Under 30	Low
3	Female	30 or more	High
4	Female	30 or more	
5	Female	30 or more	High

There is a systematic relationship between the inclination of missing values and the observed data, but not the missing data. All that is required is a probabilistic relationship

Missing not at Random (MNAR) - Nonignorable

- ❑ Missingness is related to unobserved measurements
- ❑ When the missing values on a variable are related to the values of that variable itself, even after controlling for other variables.

MNAR means that the probability of being missing varies for reasons that are unknown to us.

Data was obtained from 31 women, of whom 14 were located six months later. Of these, three had exited from homelessness, so the estimated proportion to have exited homelessness is 3/14 = 21%. As there is no data for the 17 women who could not be contacted, it is possible that none, some, or all of these 17 may have exited from homelessness. This means that potentially the proportion to have exited from homelessness in the sample is between 3/31 = 10% and 20/31 = 65%. As a result, reporting 21% as being the correct result is misleading. In this example the missing data is nonignorable.

Strategies to handle MNAR are to find more data about the causes for the missingness, or to perform what-if analyses to see how sensitive the results are under various scenarios.

Formalize the definitions

- ❑ Let's X represent a matrix of the data we “expect” to have; X = {Xo, Xm} where Xo is the observed data and Xm the missing data.

1. **MCAR: P(R| Xo , Xm) = P(R)**
2. **MAR: P(R| Xo , Xm) = P(R| Xo)**

- ❑ Let’s define R as a matrix with the same dimensions as X where Ri,j = 1 if the datum is missing, and 0 otherwise.

3. **MNAR: No simplification.**

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

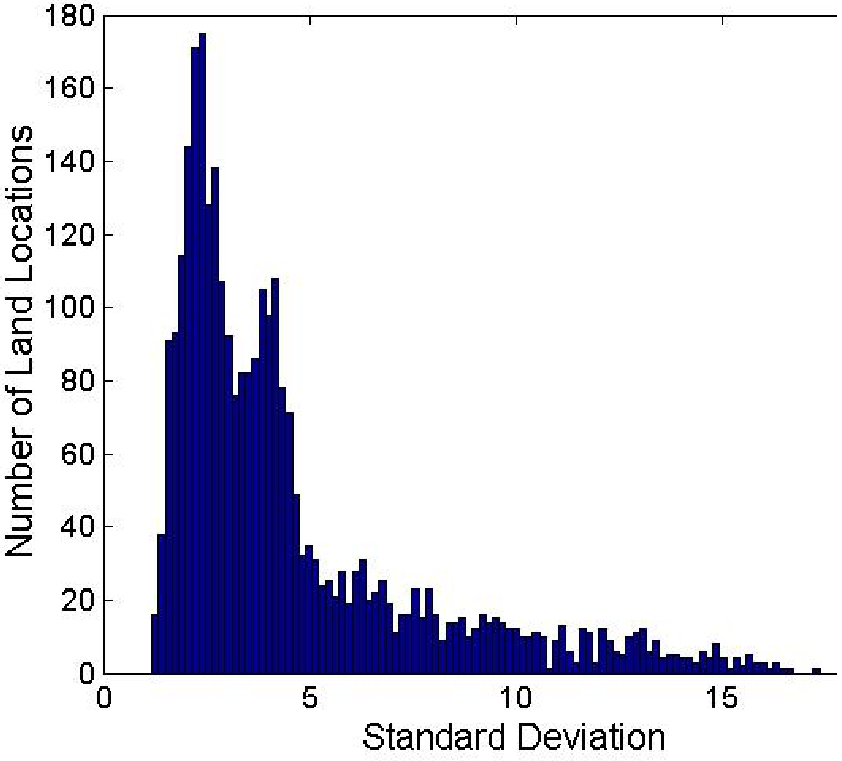
- ❑ Aggregation
- ❑ Sampling
- ❑ Dimensionality Reduction
- ❑ Feature subset selection
- ❑ Feature creation
- ❑ Discretization and Binarization
- ❑ Attribute Transformation

Aggregation

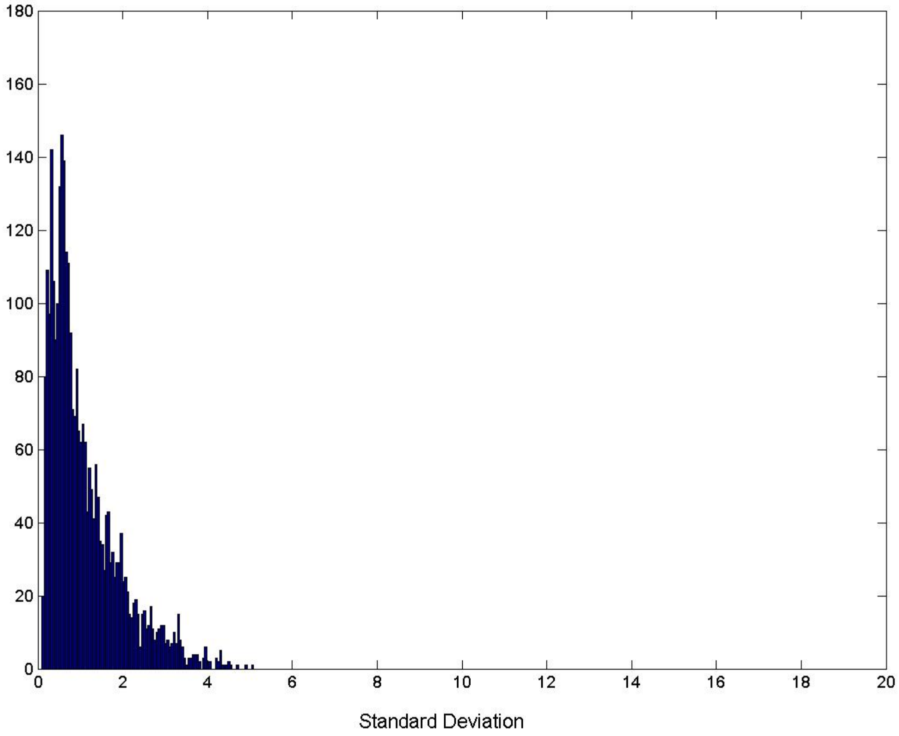
- Combining two or more attributes (or objects) into a single attribute (or object)
- Purpose
  - Data reduction
    - ✓ Reduce the number of attributes or objects
  - Change of scale
    - ✓ Cities aggregated into regions, states, countries, etc
  - More “stable” data
    - ✓ Aggregated data tends to have less variability

Aggregation...

- Variation of Precipitation in Australia



Standard Deviation of Average Monthly Precipitation



Standard Deviation of Average Yearly Precipitation

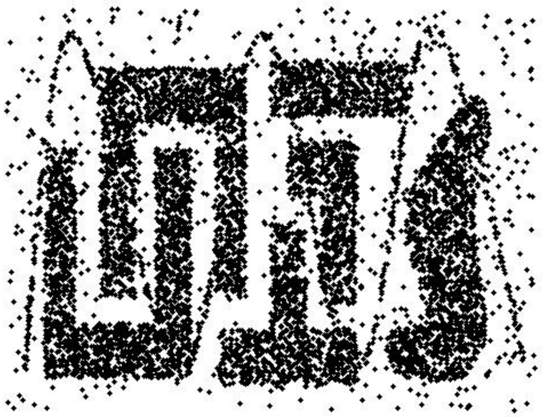
Sampling

- ❑ 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.
- ❑ 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

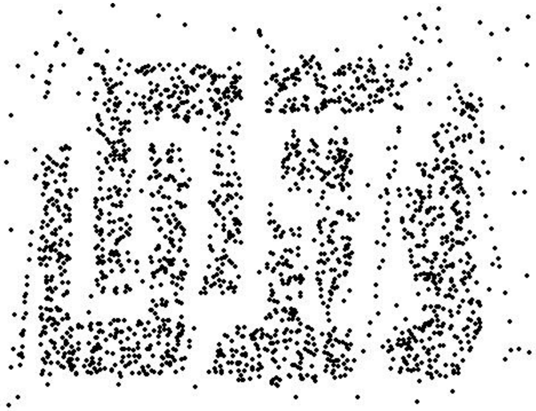
Types of Sampling

- ❑ 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

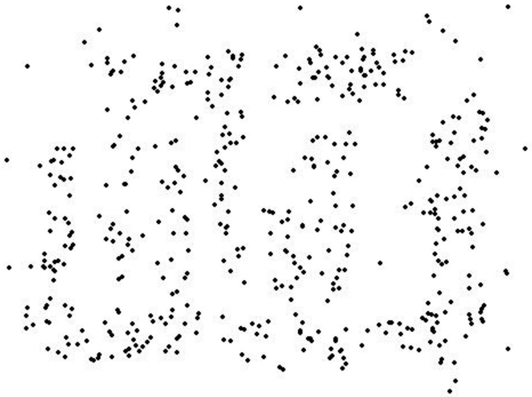
Sample Size



8000 points



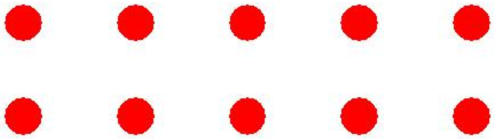
2000 Points

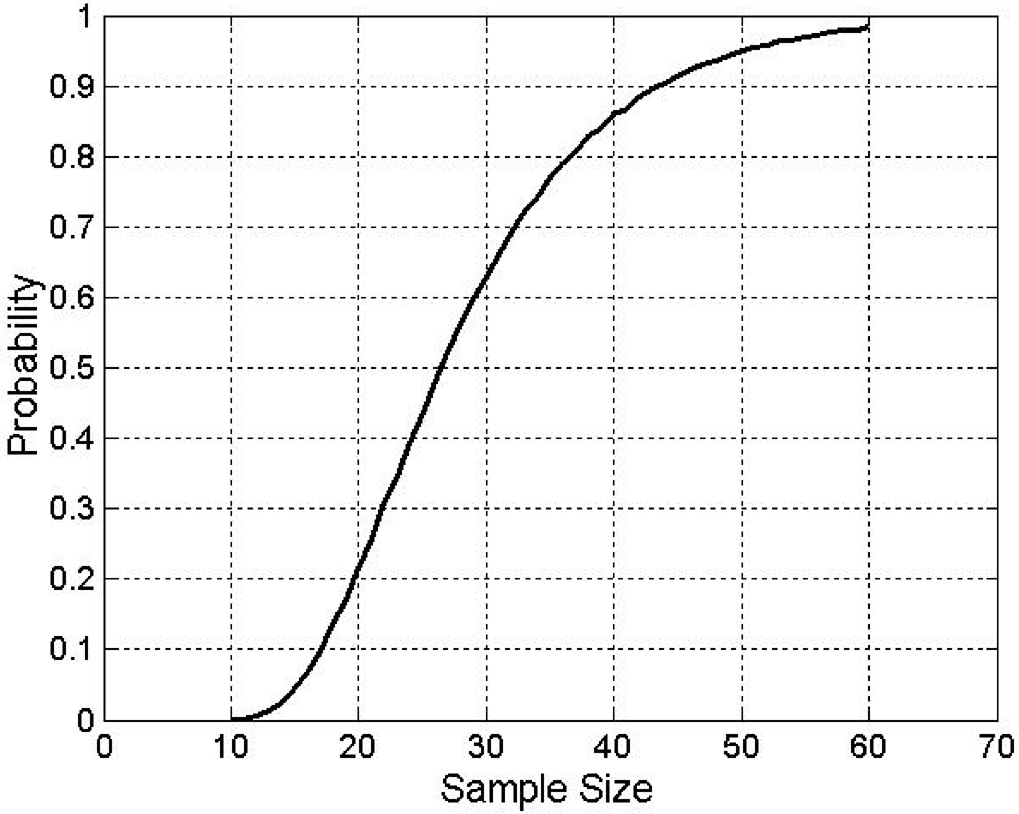


500 Points

Sample Size

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





Curse of Dimensionality

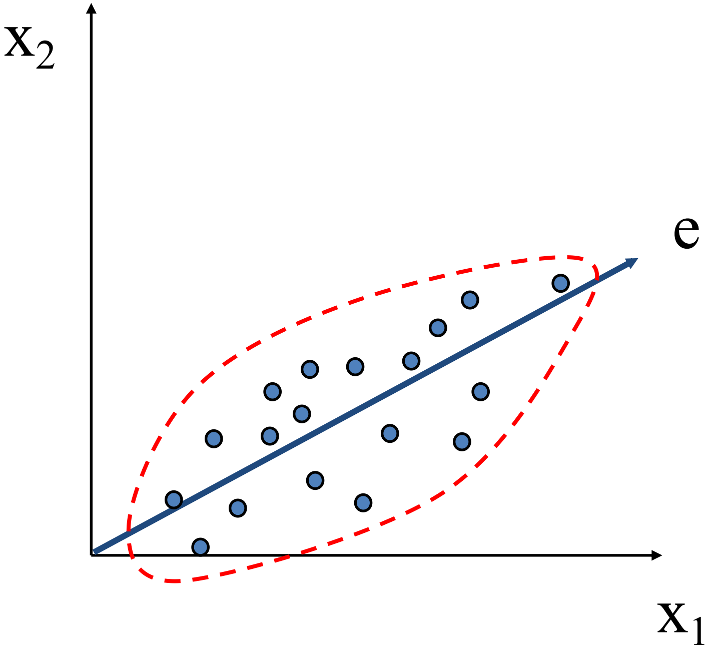
- ❑ When dimensionality increases, data becomes increasingly sparse in the space that it occupies
- ❑ Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful

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

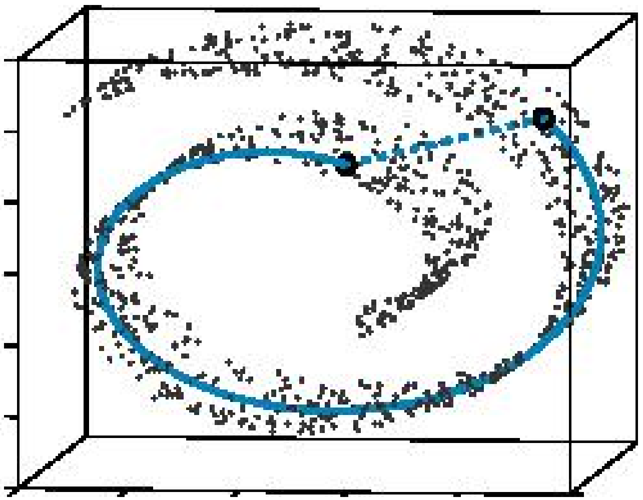
- ❑ Goal is to find a projection that captures the largest amount of variation in data



- ❑ 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

Feature Subset Selection

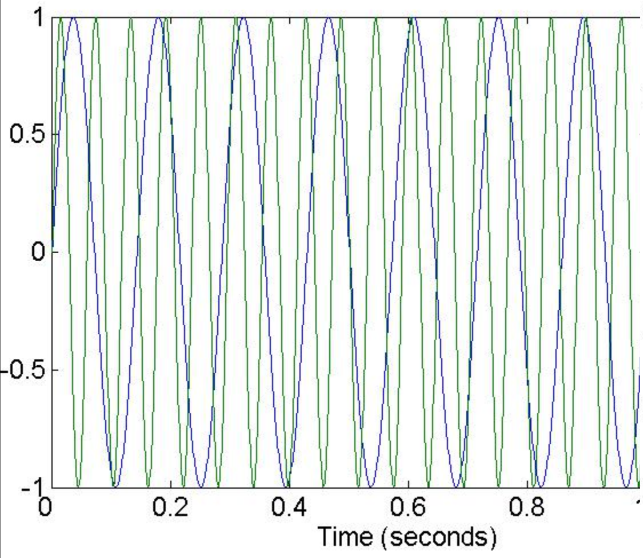
- Another way to reduce dimensionality of data
- Redundant features
  - duplicate much or all of the information contained in one or more other attributes
    - ✓ Example: purchase price of a product and the amount of sales tax paid
- Irrelevant features
  - contain no information that is useful for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

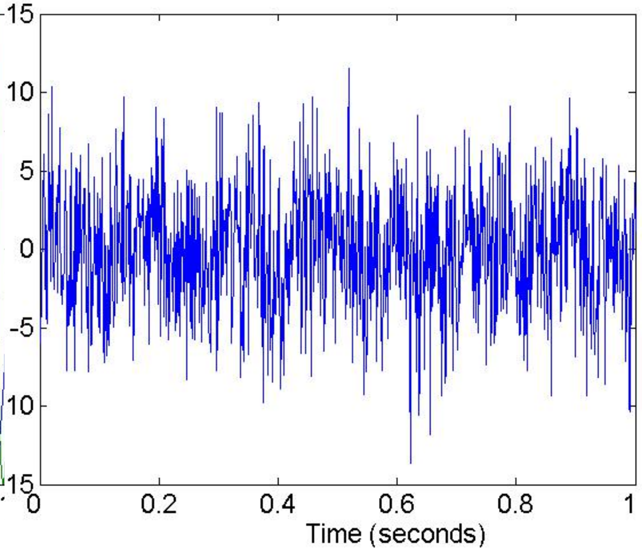
- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes
- Three general methodologies:
  - Feature Extraction
    - ✓ domain-specific
  - Mapping Data to New Space
  - Feature Construction
    - ✓ combining features

Mapping Data to a New Space

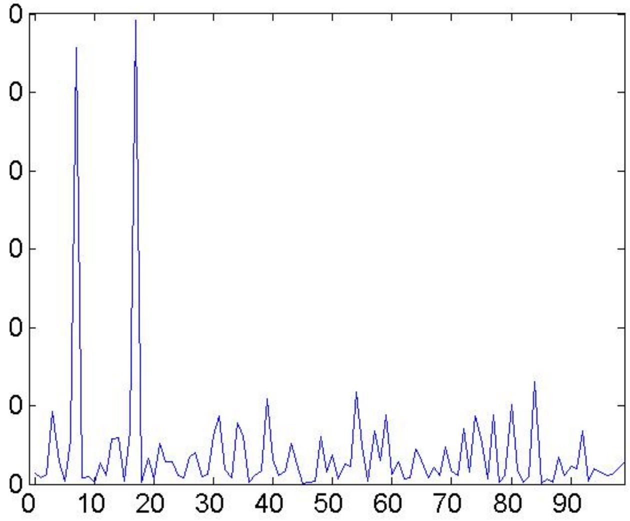
- ❑ Fourier transform
- ❑ Wavelet transform



Two Sine Waves



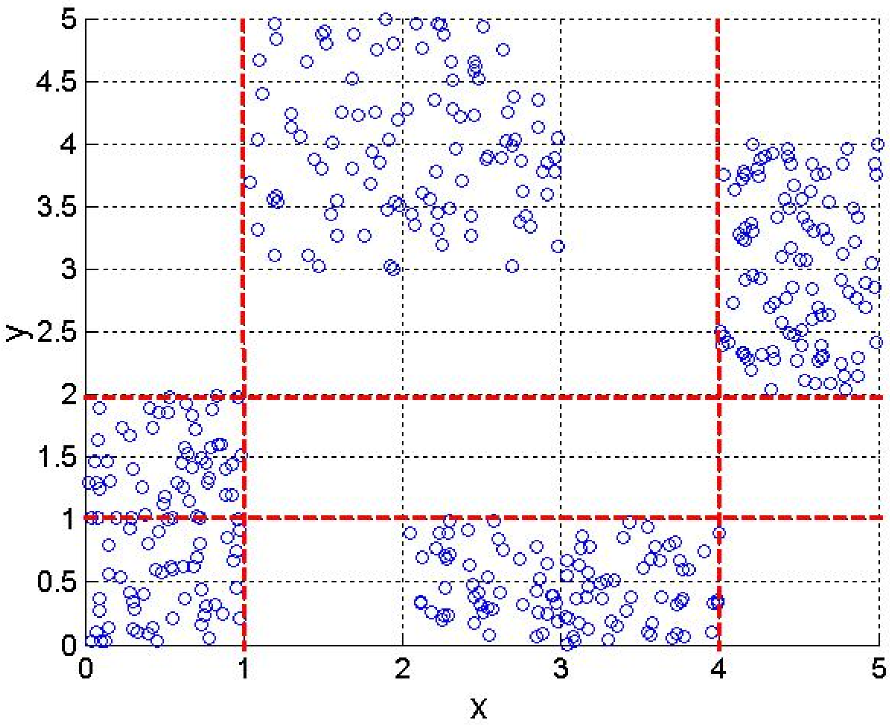
Two Sine Waves + Noise



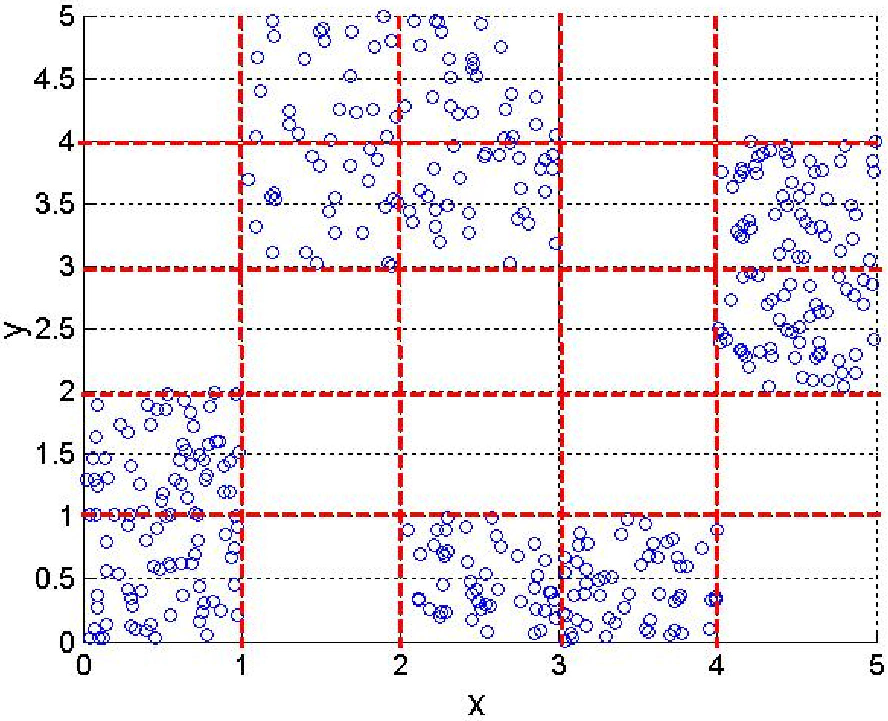
Frequency

Discretization Using Class Labels

- Entropy based approach

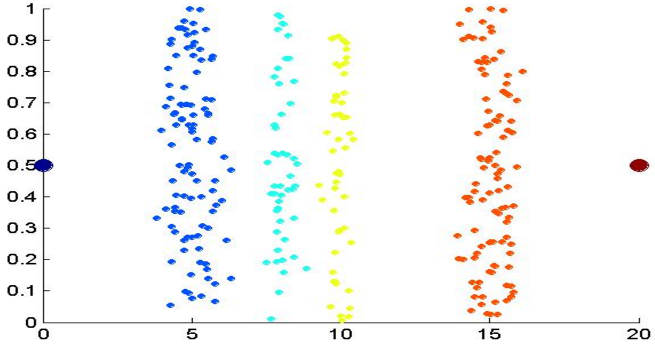


3 categories for both x and y

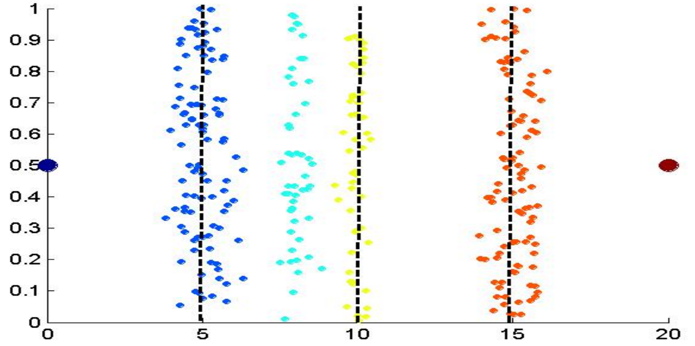


5 categories for both x and y

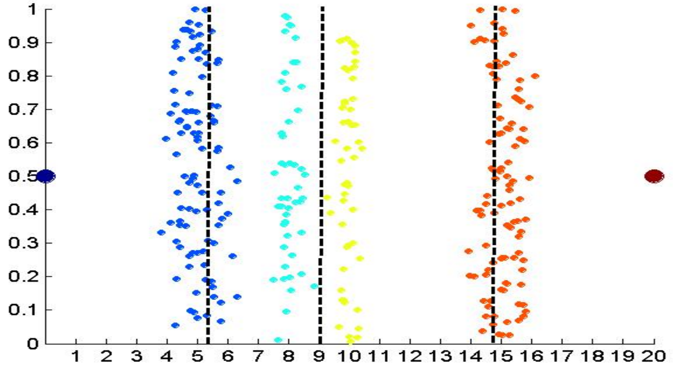
Discretization Without Using Class Labels



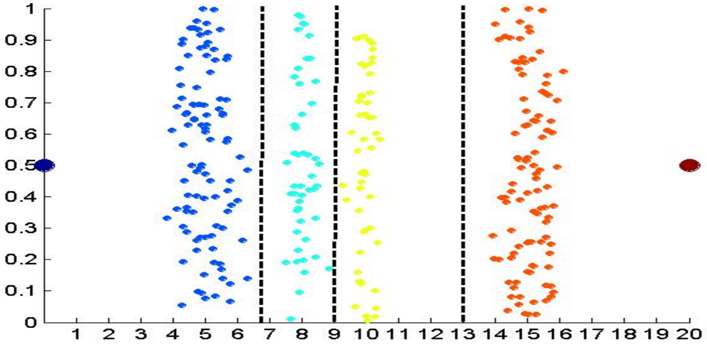
Data



Equal interval



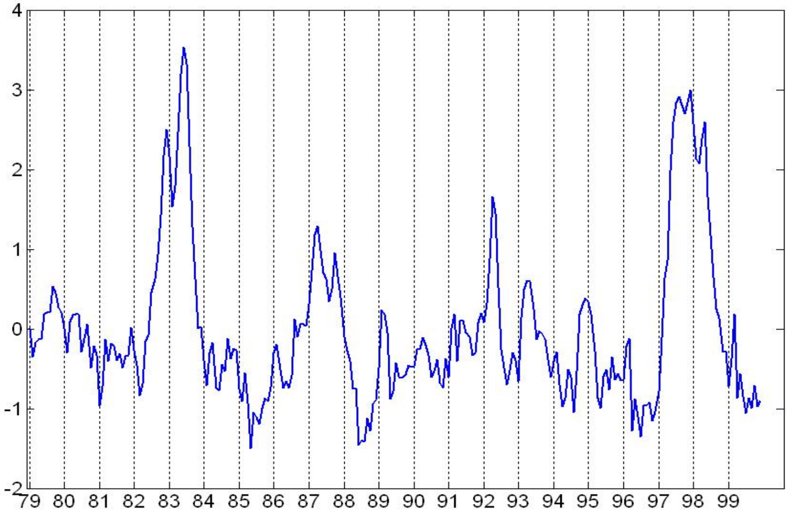
Equal frequency



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

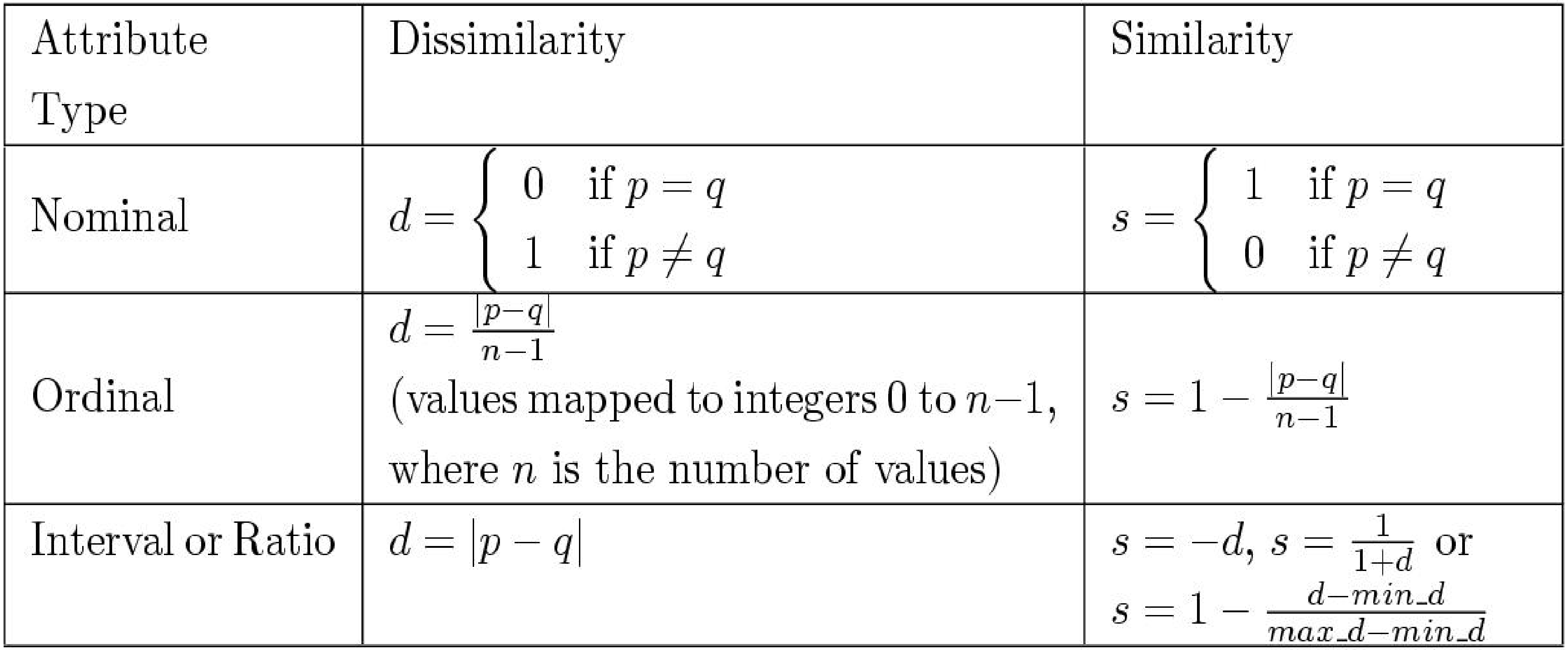


Similarity and Dissimilarity

- ❑ 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 data objects
  - ➢ 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

- ❑ 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

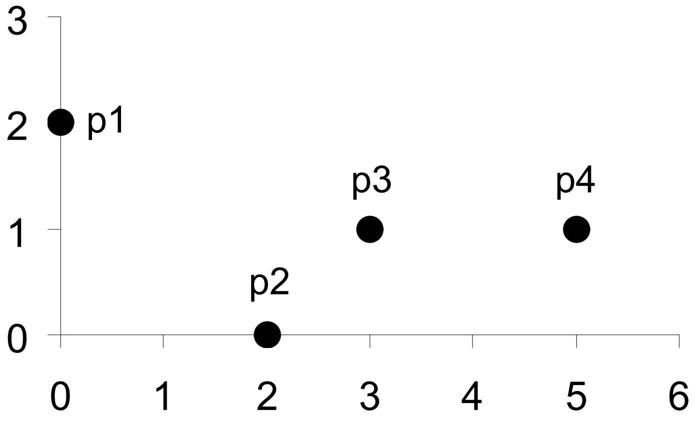
Euclidean Distance

- ❑ Euclidean Distance

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

- ❑ Where n is the number of dimensions (attributes) and pk and qk are, respectively, the kth 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, n is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) or data objects p and q.

Minkowski Distance:Example

- ❑ $r = 1$. City block (Manhattan, taxicab, L_1 norm) distance.

  A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors

- ❑ $r = 2$. Euclidean distance

- ❑ $r \rightarrow \infty$. "supremum" (L_{max} norm, L_∞ norm) distance.

  This is the maximum difference between any component of the vectors

- ❑ Do not confuse *r* with *n*, i.e., all these 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

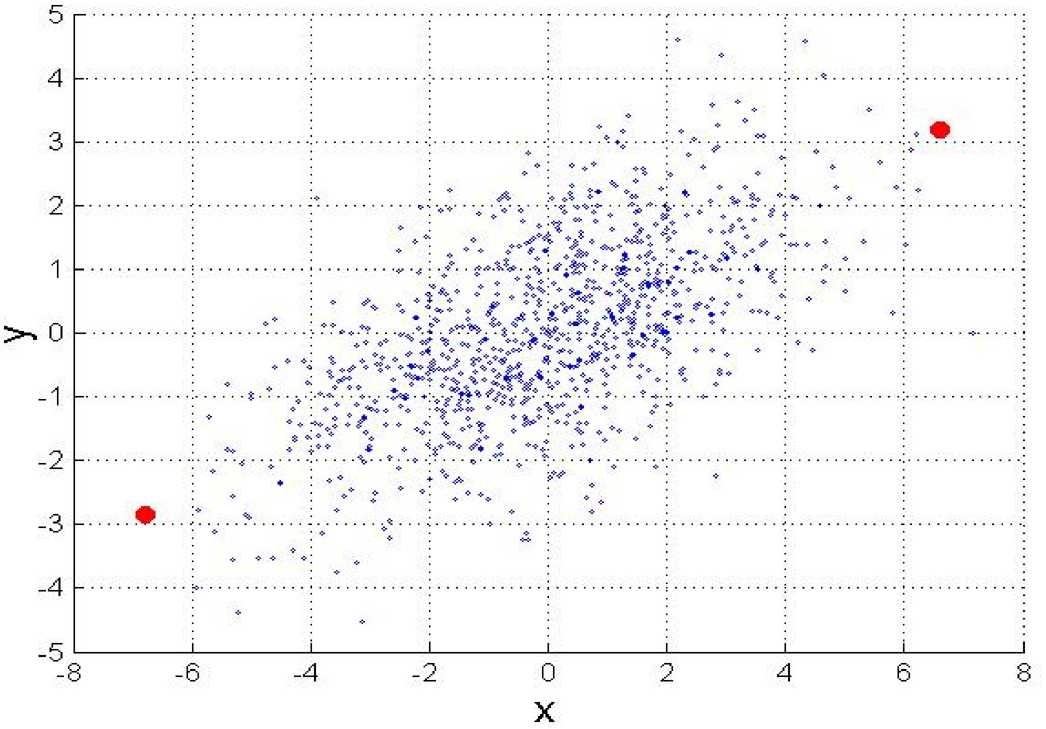
L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

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

Mahalanobis Distance

$$\boldsymbol{mahalanobi\ s(p,q) = (p-q)\Sigma^{-1}(p-q)^{T}}$$

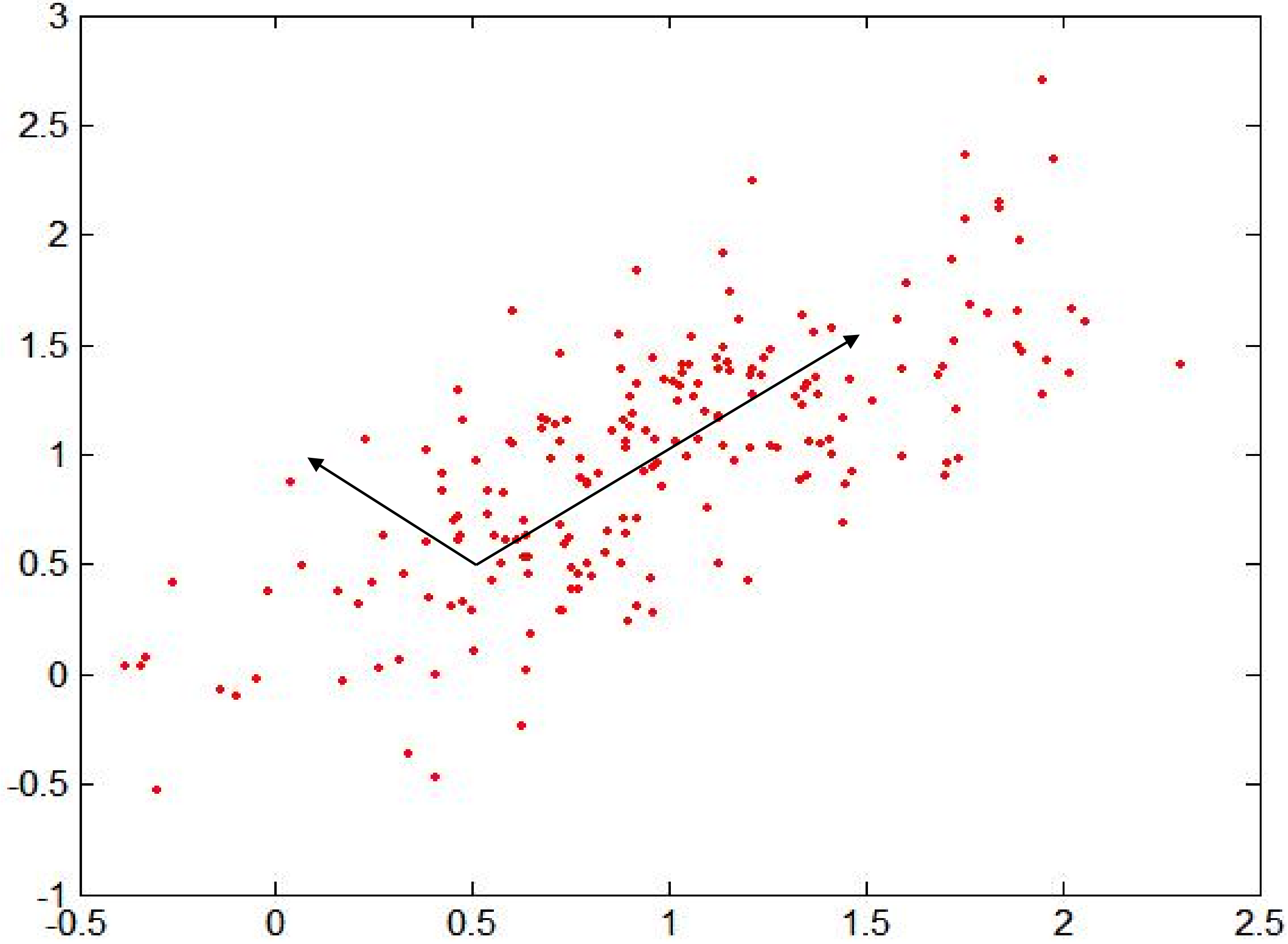


Σ is the covariance matrix of the input data *X*

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

For red points, the Euclidean distance is 14.7, Mahalanobis distance is 6.

Mahalanobis Distance



Covariance Matrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4

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

Similarity Between Binary Vectors

- ❑ Common situation is that objects, *p* and *q*, have only binary attributes

- ❑ Compute similarities using the 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)

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

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

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 • 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**

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

$\|d_1\| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{\mathbf{0.5}} = (42)^{\mathbf{0.5}} = 6.481$

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

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

Recommended Text and Reference Books

❑ **Text Book:**

- J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed., 2011

❑ **Reference Books:**

- H. Dunham. Data Mining: Introductory and Advanced Topics. Pearson Education. 2006.
- I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann. 2000.
- D. Hand, H. Mannila and P. Smyth. Principles of Data Mining.Prentice-Hall. 2001.

THANK YOU!