



Exploring Data

September 26, 2017

Ming-Syan Chen

1



Tentative Class Agenda

- Class 1 – (9/12) Introduction to data science
- Class 2 – (9/19) Python basic
- Class 3 – (9/26) Introduction to data, HW#1
- Class 4 – (10/3) More on data, OLAP
No class on 10/10
- Class 5 – (10/17) Python ML library scikit-learn, HW#2 (using scikit-learn)
- Class 6 – (10/24) PCA/SVD, Text and Web data, project announcement
- Class 7 – (10/31) Introduction to data mining: classification,

2

Tentative Class Agenda (cont'd)

- Class 8 – (11/7) Association mining
- Class 9 – (11/14) Clustering, [brief on Keras and Tensorflow](#) **HW#3** (using Keras and Tensorflow)
- Class 10 – (11/21) Social network, (project abs. due on 11/19)
- Class 11 – (11/28) R-1
- Class 12 – (12/5) R-2, **HW#4** (data mining with R)
- Class 13 – (12/12) Exam (in class, closed book)
- Class 14 – (12/19) Cloud computing, [Introduction to GPU programming](#), **HW#5**
- Class 15 – (12/26) Project presentation I
- Class 16 – (1/2) Project presentation II

Happy New Year!

3

Class Announcement

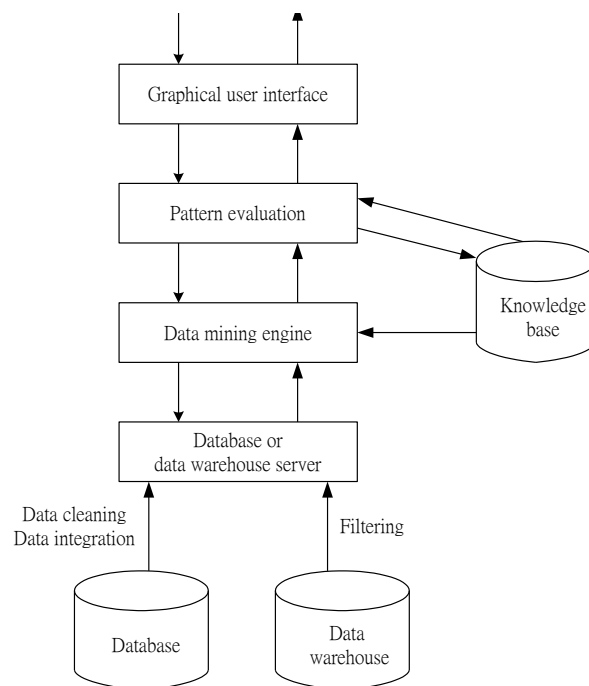
- A class Web page is ready
 - Ceiba
- TA: 謝有恒, 蔡仕竝; [吳齊軒](#) (for R)
- Grading (tentative):
 - HW 45%
 - 5 in total
 - Midterm 25% (in class, closed book)
 - Project 30% (including 5% on abstract, 5% on presentation)
 - Final grades will be adjusted based on overall scores.
 - Class performance +-...

4

Course Materials

- Introduction to data science
- Exploring Data
- Mining data
 - Classification, association, and clustering
- Data manipulation at Scale

5



6

Description on Data

(Chapter 2-3 in Vipin’s book)

7

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

<i>Tid</i>	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

Objects

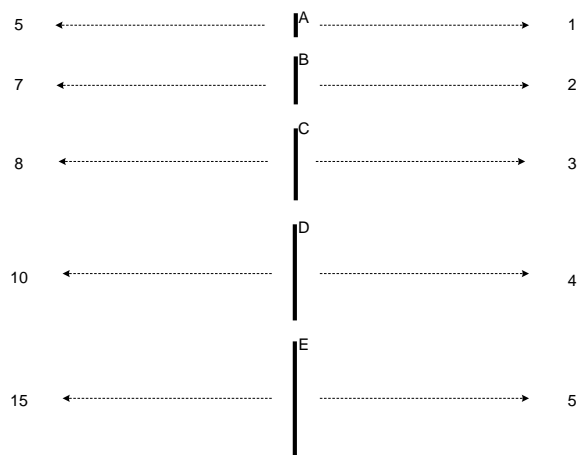
Attribute Values

- 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

9

Measurement of Length

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



10

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

11

Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
 - Distinctness: = \neq
 - Order: < >
 - Addition: + -
 - Multiplication: * /
 - Nominal attribute: distinctness
 - Ordinal attribute: distinctness & order
 - Interval attribute: distinctness, order & addition
 - Ratio attribute: all 4 properties

12



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. ($=$, \neq)	zip codes, employee ID numbers, eye color, sex: { <i>male</i> , <i>female</i> }	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to 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
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

13



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.

14

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

15

Important Characteristics of Structured Data

- **Dimensionality**
 - **Curse of Dimensionality**
- **Sparsity**
 - **Only presence counts**
- **Resolution**
 - **Patterns depend on the scale**

16

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

17

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

18

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

19

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.

<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



20

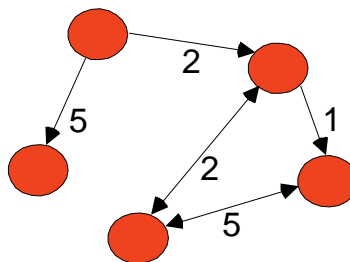
Types of data sets

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

21

Graph Data

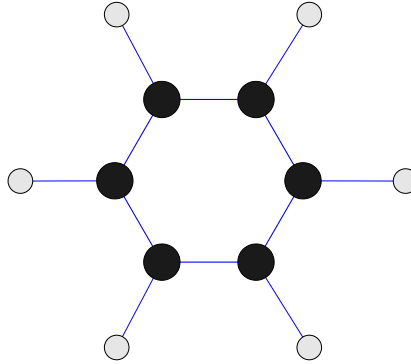
- Many data could be represented by a graph structure
 - Examples: Web graph and HTML Links
- A directed graph with links labelled



22

Chemical Data

- Benzene Molecule: C_6H_6



23

Types of data sets

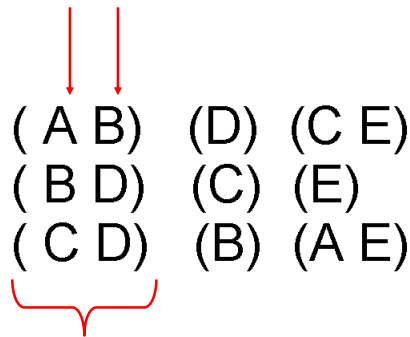
- **Record**
 - Data Matrix
 - Document Data
 - Transaction Data
- **Graph**
 - World Wide Web
 - Molecular Structures
- **Ordered**
 - Spatial Data
 - Temporal Data
 - Sequential Data
 - Genetic Sequence Data

24

Ordered Data

- Sequences of transactions

Items/Events



25

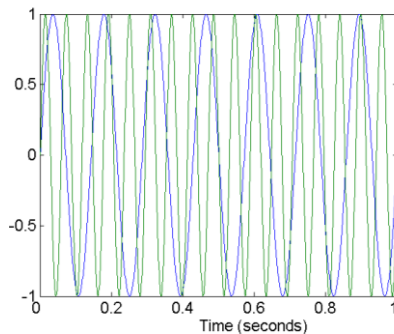
Data Quality

- 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

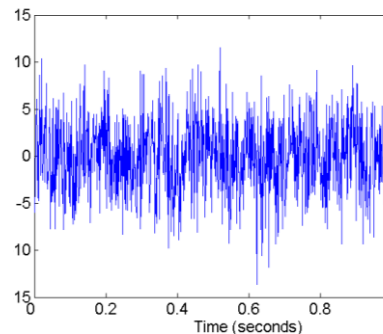
26

Noise

- 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

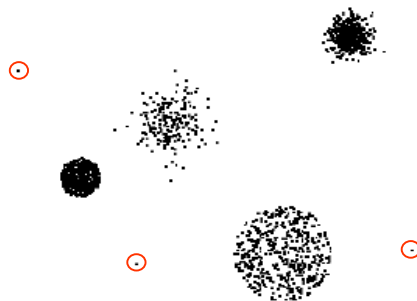


Two Sine Waves + Noise

27

Outliers

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



28



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)

29



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

30



Data Preprocessing

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

31



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

32

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.

33

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

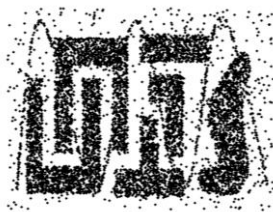
34

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

35

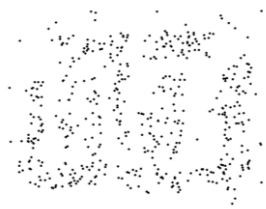
Sample Size



8000 points



2000 Points

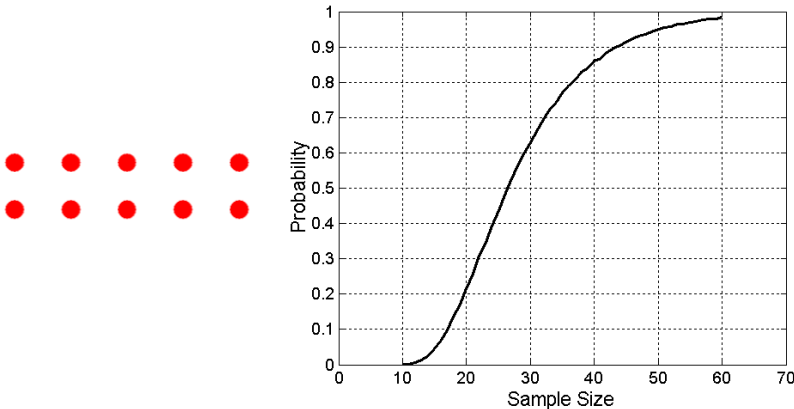


500 Points

36

Sample Size

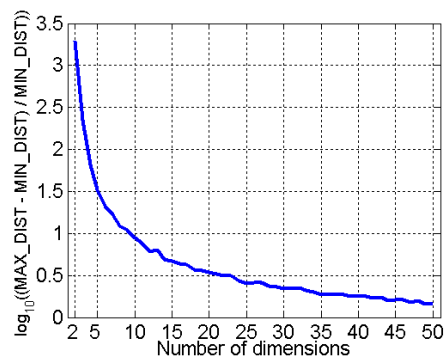
- What sample size is necessary to get at least one object from each of 10 groups.



37

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

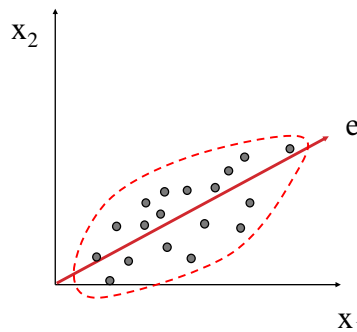
Dimensionality Reduction

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

39

Dimensionality Reduction: PCA

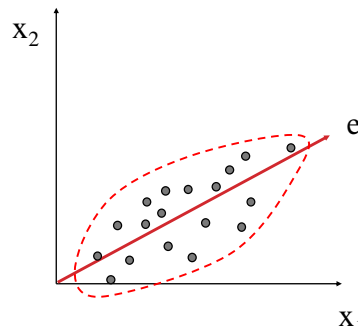
- Goal is to find a projection that captures the largest amount of variation in data



40

Dimensionality Reduction: PCA

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



41

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

42

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

43

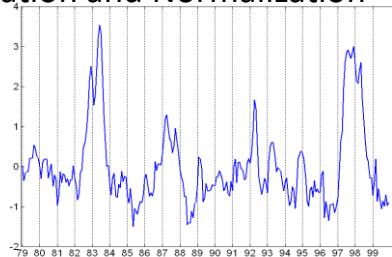
Feature Creation

- 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

44

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



45

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

46

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	$d = \frac{ p-q }{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{ p-q }{n-1}$
Interval or Ratio	$d = p - q $	$s = -d, s = \frac{1}{1+d}$ or $s = 1 - \frac{d - \min.d}{\max.d - \min.d}$

Table 5.1. Similarity and dissimilarity for simple attributes

47

Euclidean Distance

- Euclidean Distance

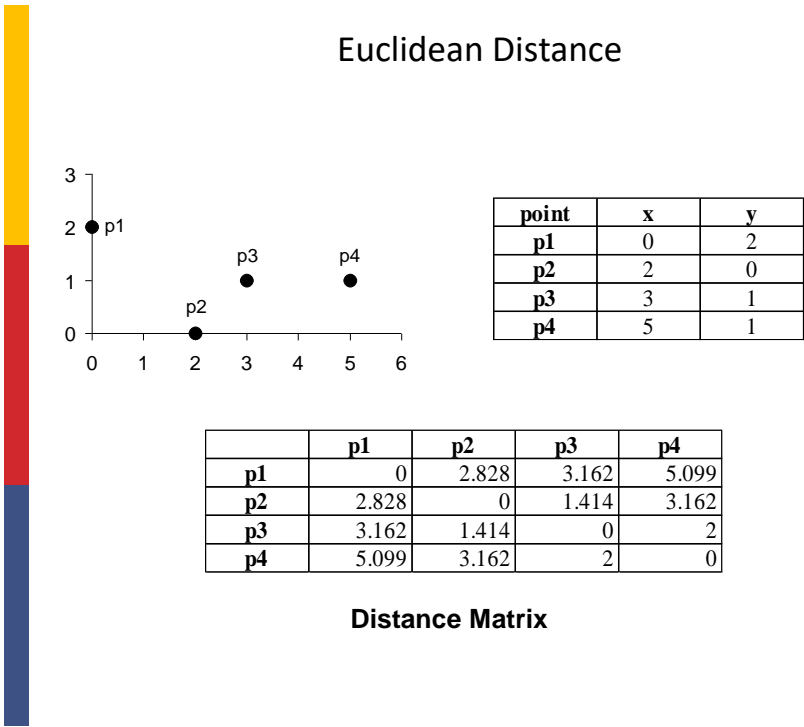
$$dist = \sqrt{\sum_{k=1}^n (p_k - q_k)^2}$$

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.

48

Euclidean Distance



49

Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

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

Where r is a parameter, 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 .

50

Minkowski Distance: Examples

- $r = 1$. 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.

51

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

52