

Data preprocessing



Elena Baralis
Politecnico di Torino




Data set types

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



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006


2



Tabular Data


- A collection of records
 - Each record is characterized by 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



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006


3



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



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006


4



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



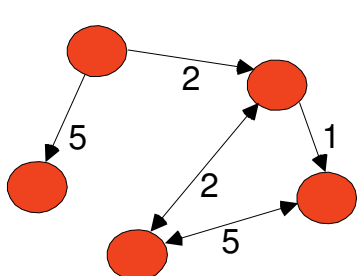
From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

5




Graph Data

- Examples: Generic graph and HTML Links




```

<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
          
```



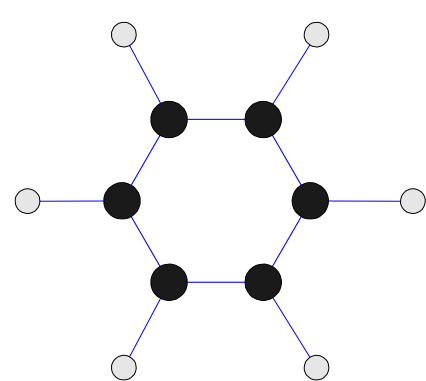
From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006


6



Chemical Data


- Benzene Molecule: C_6H_6





From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

7



Ordered Data

- Sequences of transactions

Items/Events

↓ ↓


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

(D)
(C)
(B)

(C E)
(E)
(A E)


}

An element of the sequence



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

8




Ordered Data

- Genomic sequence data


```

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



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

9

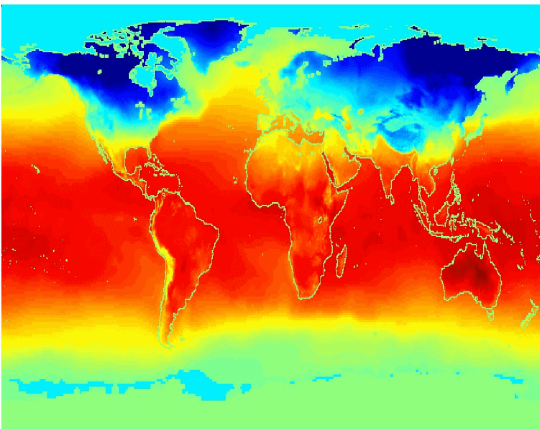



Ordered Data

- Spatio-Temporal Data


Average Monthly
Temperature of
land and ocean

Jan






From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006




Attribute types

- 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



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006


11



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




From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

12




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.




From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

13




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



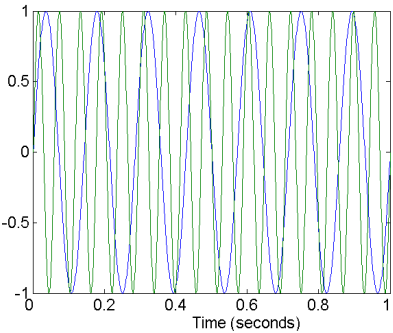
From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

14

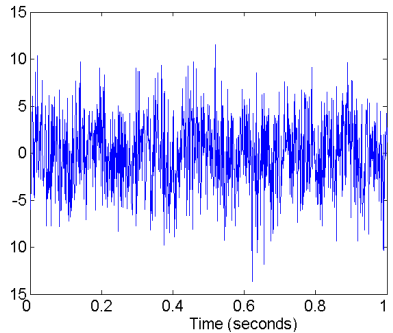


Noise


- 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



A line graph showing two sine waves, one in green and one in blue, plotted against time in seconds from 0 to 1. The y-axis ranges from -1 to 1. The waves are periodic and overlap significantly.



A line graph showing the same two sine waves as the previous plot, but with significant high-frequency noise added. The y-axis ranges from -15 to 15. The noise is represented by a dense, irregular blue line.




Two Sine Waves

Two Sine Waves + Noise

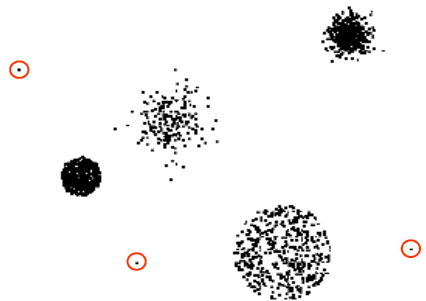
From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

15




Outliers

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




A scatter plot showing several clusters of black data points. There are four distinct clusters. Three of these clusters have a single red circle around them, indicating they are outliers. The red circles are located at approximately (0.1, 0.8), (0.3, -0.8), and (0.8, -0.8) in a coordinate system where the top-left cluster is at (0.5, 0.8).




From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

16




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)




From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

17




Important Characteristics of Structured Data

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




From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

18




Data Preprocessing

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




From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

19



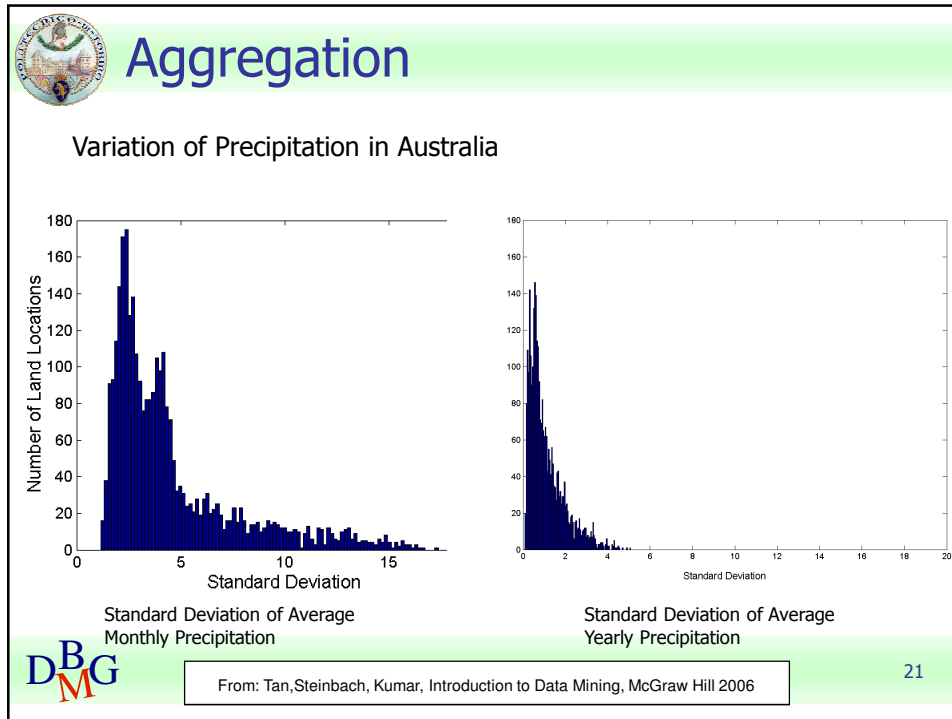
Aggregation

- 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



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

20




Data reduction

- It generates a reduced representation of the dataset. This representation is smaller in volume, but it can provide similar analytical results
 - sampling
 - It reduces the cardinality of the set
 - feature selection
 - It reduces the number of attributes
 - discretization
 - It reduces the cardinality of the attribute domain


From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

22




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.




From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

23




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




From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

24




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



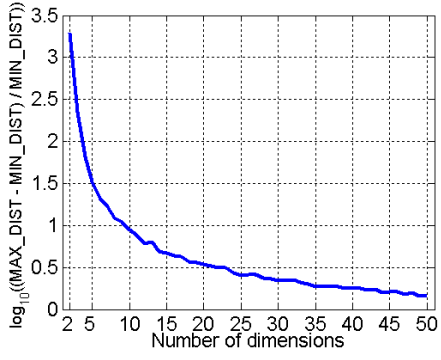
From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

25




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




From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

26




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



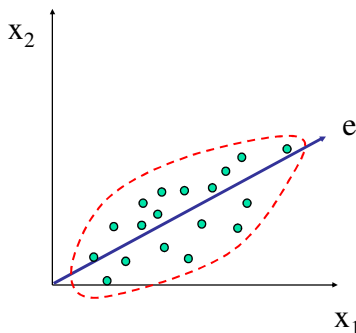
From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006


27



Dimensionality Reduction: PCA


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






From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

28




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




From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

29




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




From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

30




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



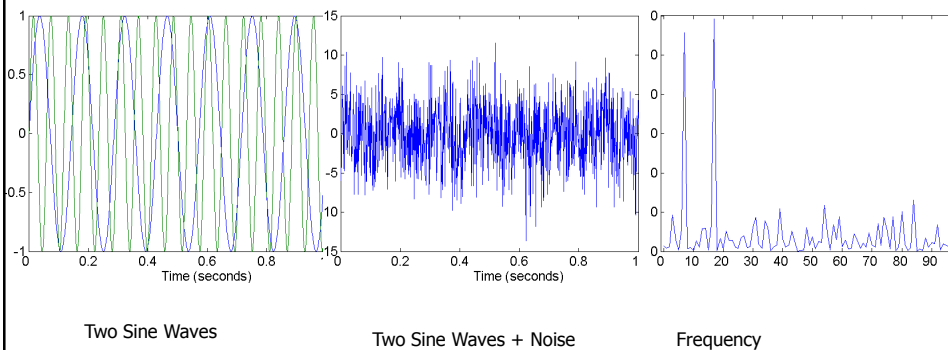
From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

31




Mapping Data to a New Space

- Fourier transform
- Wavelet transform




Two Sine Waves Two Sine Waves + Noise Frequency




From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

32




Discretization

- It splits the domain of a continuous attribute in a set of intervals
 - It reduces the cardinality of the attribute domain
- Techniques
 - N intervals with the same width $W=(v_{\max} - v_{\min})/N$
 - Easy to implement
 - It can be badly affected by outliers and sparse data
 - Incremental approach
 - N intervals with (approximately) the same cardinality
 - It better fits sparse data and outliers
 - Non incremental approach
 - clustering
 - It fits well sparse data and outliers

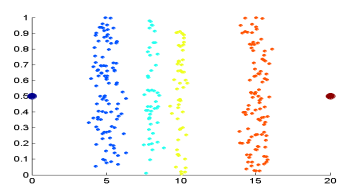


From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

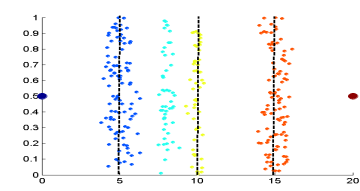
33



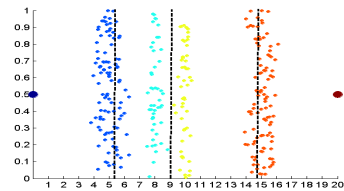
Discretization



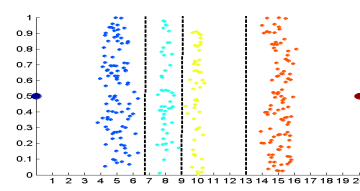
Data




Equal interval width



Equal frequency




K-means



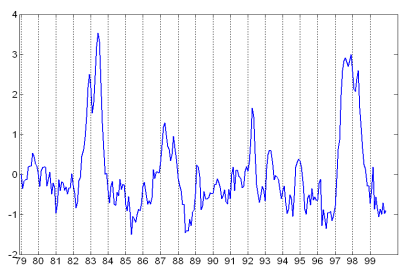
From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006


34



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





From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

35





Normalization

- It is a type of data transformation
 - The values of an attribute are scaled so as to fall within a small specified range, typically $[-1, +1]$ or $[0, +1]$
- Techniques
 - min-max normalization

$$v' = \frac{v - \min_k}{\max_k - \min_k} (\text{new_max}_k - \text{new_min}_k) + \text{new_min}_k$$
 - z-score normalization $v' = \frac{v - \text{mean}_k}{\text{stand_dev}_k}$
 - decimal scaling


$$v' = \frac{v}{10^j} \quad j \text{ is the smallest integer such that } \max(|v'|) < 1$$


36




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



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

37




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 $	$s = -d, s = \frac{1}{1+d} \text{ or } s = 1 - \frac{d - \min_d}{\max_d - \min_d}$

Table 5.1. Similarity and dissimilarity for simple attributes



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

38




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.

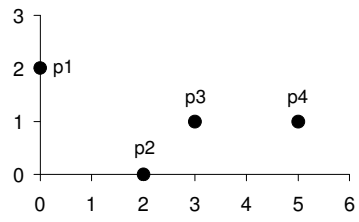


From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

39




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



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

40




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 k th attributes (components) of data objects p and q .




From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

41




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.



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

42




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

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006


43



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**



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006


44



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 .



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

45




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


$SMC = \text{number of matches} / \text{number of attributes}$
 $= (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

$J = \text{number of 11 matches} / \text{number of not-both-zero attributes values}$
 $= (M_{11}) / (M_{01} + M_{10} + M_{11})$



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

46




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)


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

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



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

47



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 norm of vector d .
- Example:

$d_1 = 3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0$
 $d_2 = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2$


$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$
 $||d_1|| = (3^2 + 2^2 + 0^2 + 5^2 + 0^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2)^{0.5} = (42)^{0.5} = 6.481$
 $||d_2|| = (1^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 1^2 + 0^2 + 2^2)^{0.5} = (6)^{0.5} = 2.245$

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



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

48




Combining Similarities

- Sometimes attributes are of many different types, but an overall similarity is needed.

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:


$$\delta_k = \begin{cases} 0 & \text{if the } k^{th} \text{ attribute is a binary asymmetric attribute and both objects have a value of 0, or if one of the objects has a missing values for the } k^{th} \text{ attribute} \\ 1 & \text{otherwise} \end{cases}$$
3. Compute the overall similarity between the two objects using the following formula:

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



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

49




Combining Weighted Similarities

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

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

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



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

50