Data Issues

Dr. Sanjay Ranka
Professor
Computer and Information Science and Engineering
University of Florida, Gainesville

ranka@cise.ufl.edu

What Is a Data Set ?

- Attributes (describe objects)
  - Variable, field, characteristic, feature or observation
- Objects (have attributes)
  - Record, point, case, sample, entity or item
- Data Set
  - Collection of objects

Type of an Attribute

- The type of an attribute depends on the following properties:
  - Distinctness: =, ≠
  - Order: <, >
  - Addition: +, -
  - Multiplication: *, /

Types of Attributes

Attribute Type	Description	Examples	Operations
Nominal	Each value represents a label. (Typical comparisons between two values are limited to "equal" or "no equal")	Flower color, gender, zip code	Mode, entropy, contingency correlation, χ^2 test
Ordinal	The values can be ordered. (Typical comparisons between two values are "equal" or "greater" or "less")	Hardness of minerals, {good, better, best}, grades, street numbers, rank, age	Median, percentiles, rank correlation, run tests, sign tests
Interval	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	Differences and ratios are meaningful. (*, /)	Monetary quantities, counts, age, mass, length, electrical current	Geometric mean, harmonic mean, percent variation

Transformations for different types

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	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 attributes
  - A discrete attribute has only a finite or countably infinite set of values. Eg. Zip codes, counts, or the set of words in a document
  - Discrete attributes are often represented as integer variables
- Binary attributes are a special case of discrete attributes and assume only two values
  - E.g. Yes/no, true/false, male/female
  - Binary attributes are often represented as Boolean variables, or as integer variables that take on the values 0 or 1
- Continuous attributes
  - A continuous attribute has real number values. E.g. Temperature, height or weight (Practically real values can only be measured and represented to a finite number of digits)
  - Continuous attributes are typically represented as floating point variables

Structured Data Sets

- Common types
  - Record
  - Graph
  - Ordered
- Three important characteristics
  - Dimensionality
  - Sparsity
  - Resolution

Record Data

- Most of the existing data mining work is focused around data sets that consist of a collection of records (data objects), each of which consists of fixed set of data fields (attributes)

Source: Data Mining – Introductory and Advanced topics by Margaret Dunham

Name	Gender	Height	Output
Kristina	F	1.6 m	Medium
Jim	M	2 m	Medium
Maggie	F	1.9 m	Tall
Martha	F	1.88 m	Tall
Stephanie	F	1.7 m	Medium
Bob	M	1.85 m	Medium
Kathy	F	1.6 m	Medium
Dave	M	1.7 m	Medium
Worth	M	2.2 m	Tall
Steven	M	2.1 m	Tall
Debbie	F	1.8 m	Medium
Todd	M	1.95 m	Medium
Kim	F	1.9 m	Tall
Amy	F	1.8 m	Medium
Lynette	F	1.75 m	Medium

Data Matrix

- If all objects in data set have the same set of numeric attributes, then each object represents a point (vector) in multi-dimensional space
- Each attribute of the object corresponds to a dimension

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, where each term is a component (attribute) of the vector, and where the value of each component of the vector 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

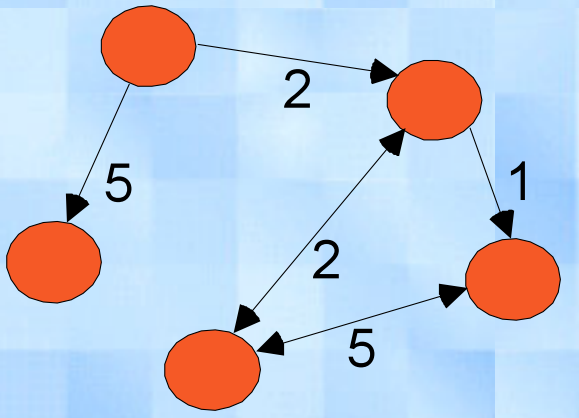
- Transaction data is a special type of record data, where each record (transaction) involves a set of items. For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are items

Source: Data Mining – Introductory and Advanced topics by Margaret Dunham

Transaction	Items
T1	Bread, Jelly, Peanut Butter
T2	Bread, Peanut Butter
T3	Bread, Milk, Peanut Butter
T4	Beer, Bread
T5	Beer, Milk

Graph Data

HTML Document




```
<tr>
<td>
<b>Publications:<br>
          </b>
<a href="books.htm">Books</a><br>
     
<a href="journal.htm">Journals</a><br>
     
 <a href="refconf.htm">Conferences</a><br>
         
<a href="workshop.htm">Workshops</a>
</td>
</tr>
```

Ordered Data: Transaction Data

(AB) (D) (CE)

(BD) (C) (E)

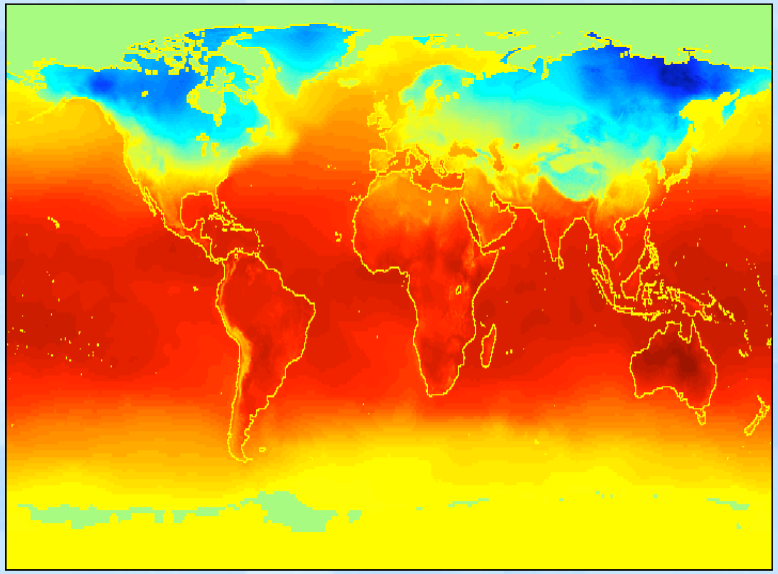
(CD) (B) (AE)

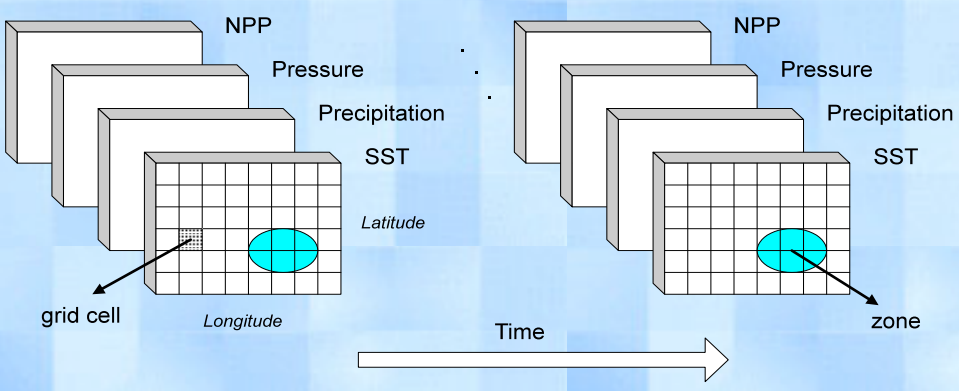
Ordered Data: Genomic Sequence Data

GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG

Ordered Data: Spatio-temporal Data

Ocean and Land Temperature (Jan 1982)





- Find global climate patterns of interest to earth scientists
- Global snapshots of values for a number of variables on land surfaces and water surfaces
- Monthly over a range of 10 to 50 years

Data Quality

- The following are some well known issues
  - Noise and outliers
  - Missing values
  - Duplicate data
  - Inconsistent values

Noise and Outliers

- Noise –Modification of original value
  - random
  - non-random (artifact of measurement)
- Noise can be
  - temporal
  - spatial
- Signal processing can reduce (generally not eliminate) noise
- Outliers – Small number of points with characteristics different from rest of the data

Missing Values: Eliminate Data Objects with missing values

- A simple and effective strategy is to eliminate those records which have missing values. A related strategy is to eliminate attribute that have missing values
- Drawback: you may end up removing a large number of objects

Missing Values: Estimating Them

- Price of the IBM stock changes in a reasonably smooth fashion. The missing values can be estimated by interpolation
- For a data set that has many similar data points, a nearest neighbor approach can be used to estimate the missing value. If the attribute is continuous, then the average attribute value of the nearest neighbors can be used. While if the attribute is categorical, then the most commonly occurring attribute value can be taken

Missing Values: Using the Missing Value As Another Value

- Many data mining approaches can be modified to operate by ignoring missing values
- E.g. Clustering - Similarity between pairs of data objects needs to be calculated. If one or both objects of a pair have missing values for some attributes, then the similarity can be calculated by using only the other attributes.