Data

Huiping Cao

Outline

- Types of data
- Data quality
- Measurement of proximity
- Data preprocess

What is Data?

Tid	Refund	Marital Status	Taxable In-come	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

- Collection of data objects and their attributes
- Attributes: a property or characteristic of an object
- Objects: a collection of attributes
- Attribute values: numbers or symbols assigned to an attribute

Types of Attributes

- Nominal
  - Examples: ID numbers, eye colors, 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

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

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

Discrete and Continuous 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
  - Data matrix
  - Document data
  - Transaction data
- Graph
  - World Wide Web
  - Molecular structures
- Ordered
  - Spatial data
  - Temporal data
  - Sequential data
  - Genetic sequence data

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 In-come	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

- Data object: point in a multi-dimensional space where each dimension represents a distinct attribute
- Represented by an $m \times n$ matrix, where there are m rows, one for each object, and n columns, one for each attribute

x	y	temperature	humidity	soil moisture
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	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.

Document Data

- A short/long document:
  The Cable News Network is an American basic cable and satellite television channel that is owned by the Turner Broadcasting System division of Time Warner
- Remove common words (or stopwords). These words are referred to as stopwords. They include
  - Articles (a, an, the,)
  - Prepositions (in, on, of,)
  - Conjunctions (and, or, but, if,)
  - Pronouns (I, you, them, it)
  - Possibly some verbs, nouns, adverbs, adjectives (make, thing, similar, etc.)

  Cable News Network American basic cable satellite television channel owned Turner Broadcasting System division Time Warner

Document Data

- Stemming
  - Cable News Network American basic cable satellite television channel owned Turner Broadcasting System division Time Warner
  - Stemming: Replace all the *variants* of a word with the single stem of the word.
  - Variants include plurals, gerund forms (ing-form), third person suffixes, past tense suffixes, etc.
  - Example: connect, connects, connected, connecting, connection, etc.
  - cable news network american basic cable satellite television channel own turner broadcast system division time warner
- Form a matrix

Document Data

- Example documents
  - doc1: cable news network american basic cable satellite television channel own turner broadcast system division time warner
  - doc2: Comcast Corporation, formerly registered as Comcast Holdings, is a U.S.-based multinational mass media company and is the largest broadcasting and largest cable company in the world by revenue
  - ...
- Form a matrix

term	doc1	doc2	...
broadcast	1	0	...
cable	1	1	...
comcast	0	1	...
mass	0	1	...
news	1	0	...
...	...	...	...

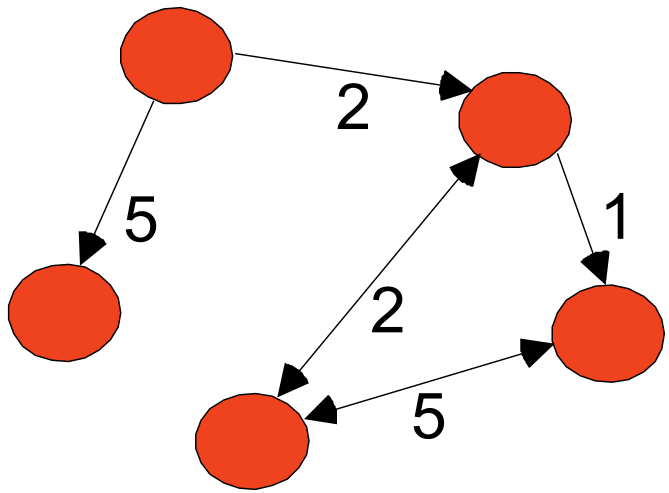
Transaction Data

- A special type of record data, where
  - Each record (transaction) involves a set of items.
  - Items: individual products that were purchased
  - A transaction: the set of products purchased by a customer during one shopping trip

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

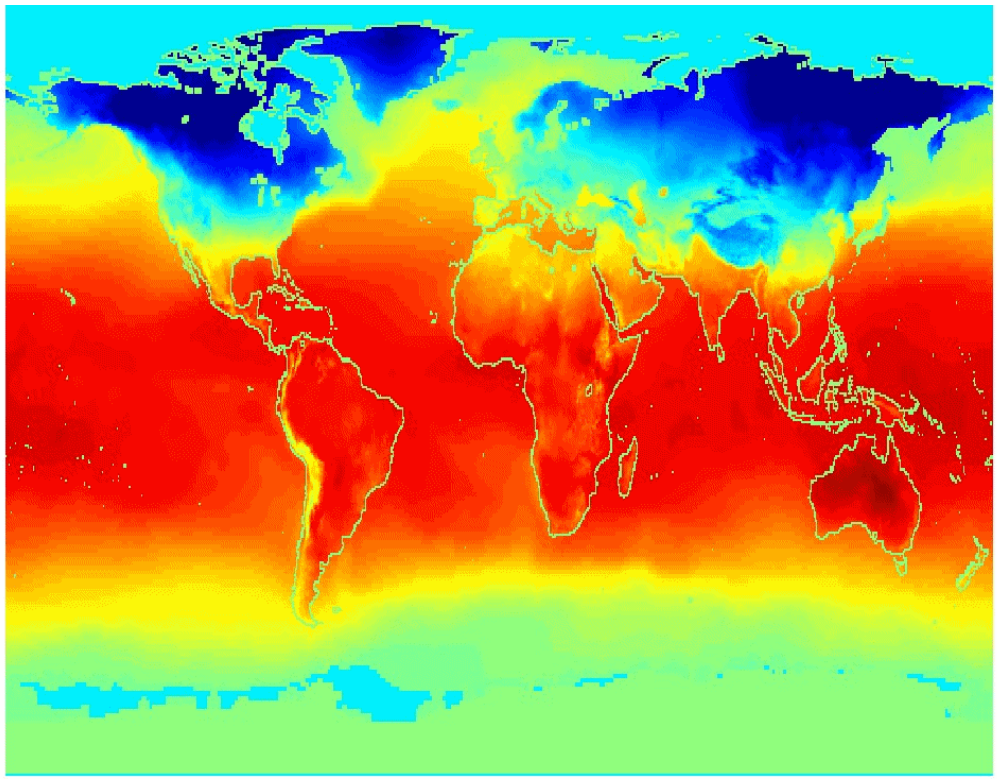


Ordered Data – Genomic Sequence Data

GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG

Ordered Data – Spatiotemporal Data

- Average monthly temperature of land and ocean

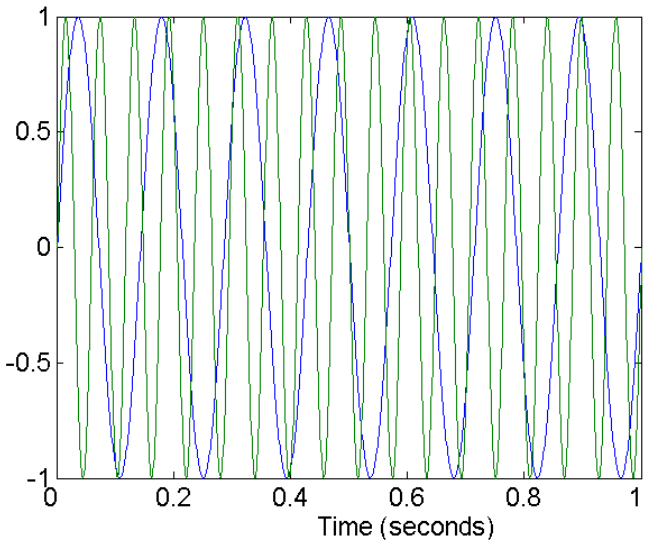


Data Quality

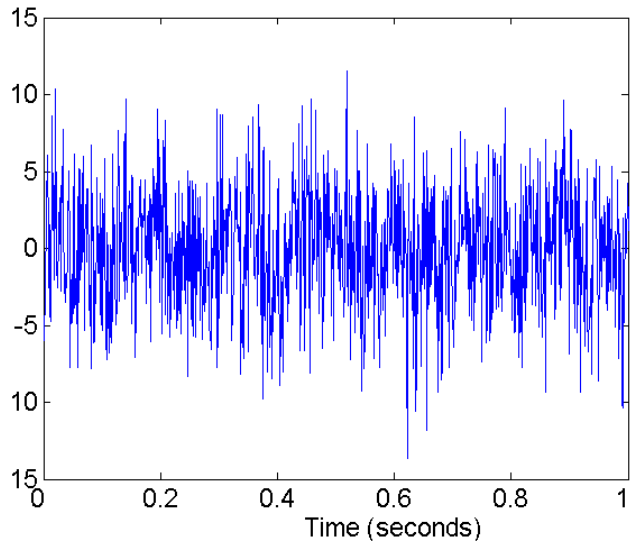
- Impossible to prevent data quality issues
- Data mining focus on
  - (1) Data cleaning: detection and correlation of data quality problems;
  - (2) The use of algorithms tolerating poor data quality
- Examples of data quality problems (related to data measurement and collection)
  - Noise and outliers
  - Missing values
  - Duplicate data

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
- Reduce noise
- Robust algorithms to tolerate noise



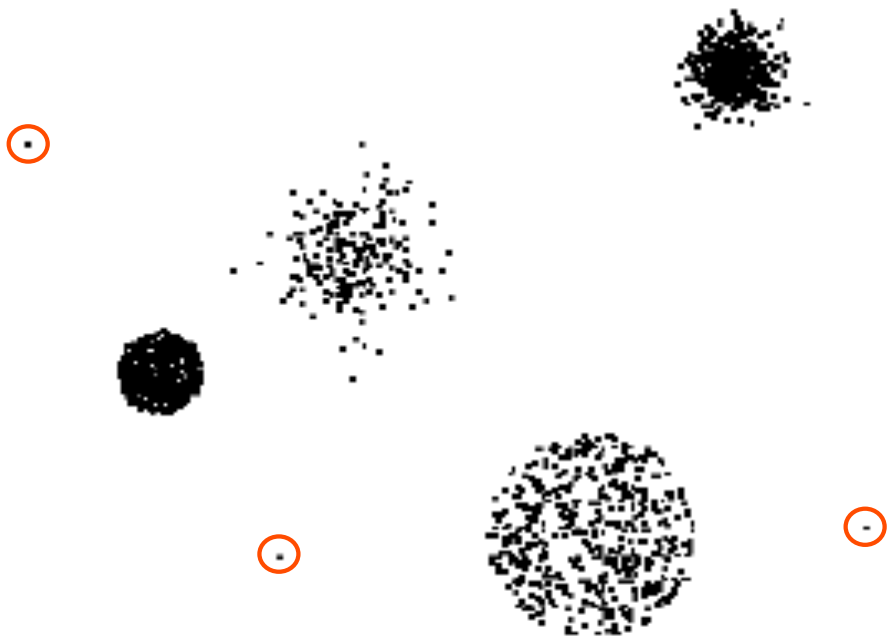
Two Sine Waves



Two Sine Waves + Noise

Outliers

- Data objects with characteristics that are considerably different than most of the other data objects in the data set
- vs. Noise
- Can be of users' interest



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: a few objects with missing values
  - Estimate missing values (interpolation)
  - Ignore the missing values during analysis

Duplicate Data

- Major issue when merging data from heterogeneous sources
- Examples
  - Same person with multiple email addresses
- Two situations
  - Two objects are intrinsically one
  - Similar objects, but still different objects