

Data Mining: Data

Lecture Notes for Chapter 2

Introduction to Data Mining

Fang Zhou

Data-related issues

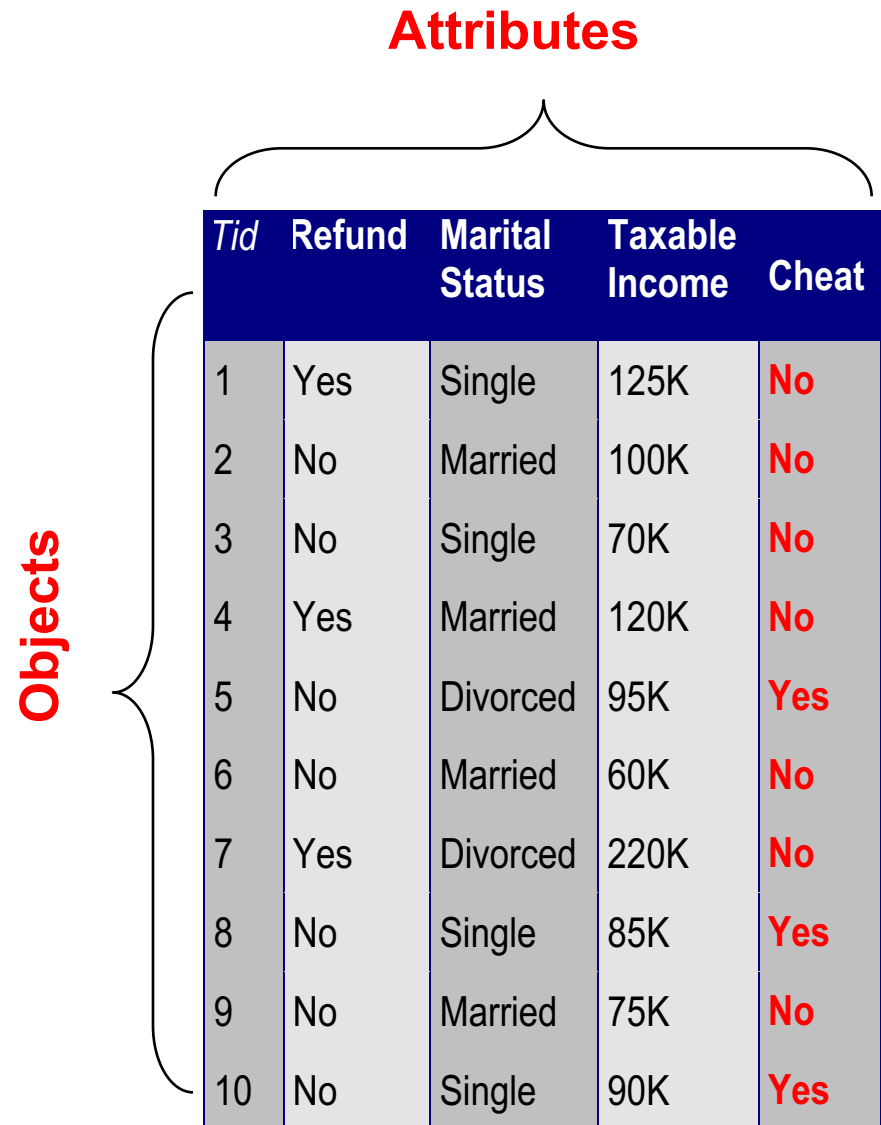


Data-related issues

- The type of data
- The quality of the data
 - Noise and outliers
 - Missing, inconsistent, duplicate data
 - Data that is biased
- Preprocessing steps

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, dimension, or feature
- A collection of attributes describe an **object**
 - Object is also known as record, point, case, sample, entity, or instance



The diagram illustrates the relationship between data objects and attributes. A table is shown with five columns: Tid, Refund, Marital Status, Taxable Income, and Cheat. A bracket labeled 'Attributes' spans the top of these columns. A bracket labeled 'Objects' spans the left side of the rows, which are indexed from 1 to 10.

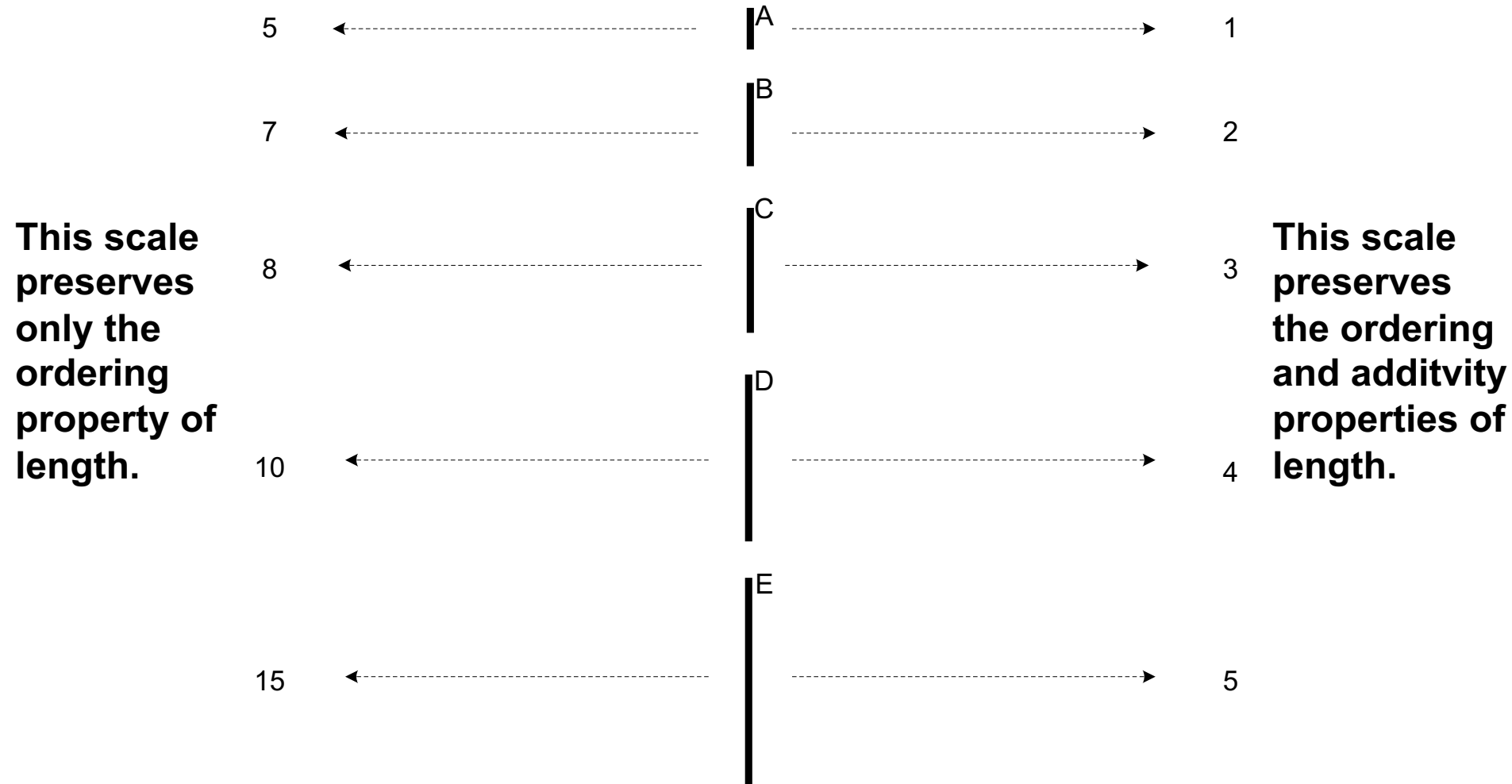
Attributes				
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

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

Measurement of Length

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



Types of Attributes

- There are different types of attributes
 - Nominal (标称)
 - ◆ Examples: ID numbers, eye color, zip codes
 - Ordinal (序数)
 - ◆ Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
 - Interval (区间)
 - ◆ Examples: calendar dates, temperatures in Celsius or Fahrenheit.
 - Ratio (比例)
 - ◆ 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

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: $\{male, female\}$	mode, entropy, contingency correlation, χ^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

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.
- Binary attributes are a special case of discrete attributes

- Continuous Attribute

- Has real numbers as attribute values
- **Examples:** temperature, height, or weight.
- Continuous attributes are typically represented as floating-point variables.

Asymmetric Attributes

- Only presence (a non-zero attribute value) is regarded as important
 - ◆ Words present in documents
 - ◆ Items present in customer transactions
- If we met a friend in the grocery store would we ever say the following?

“I see our purchases are very similar since we didn’t buy most of the same things.”

Quiz

- Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio).
 - 1) Brightness as measured by a light meter.
 - 2) Brightness as measured by people's judgments.
 - 3) ISBN numbers for books.
 - 4) Angles as measured in degrees between 0° and 360° .
 - 5) Number of patients in a hospital.
 - 6) Distance from the center of campus.

Question

- Can you think of a situation in which identification numbers would be useful for prediction?

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

<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

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

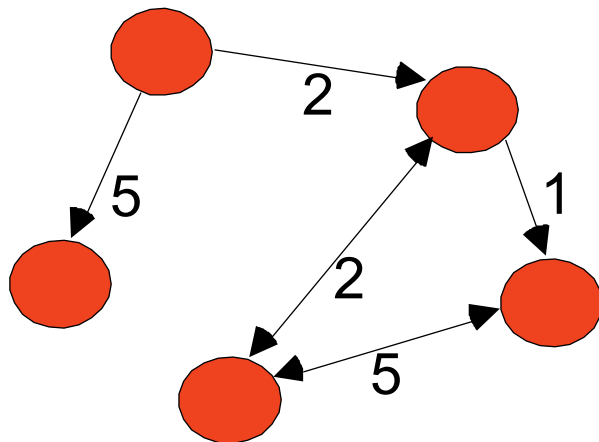
Transaction Data

- 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.

<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

Graph Data

- Examples: Generic graph and webpages



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.

Ordered Data

- Genomic sequence data

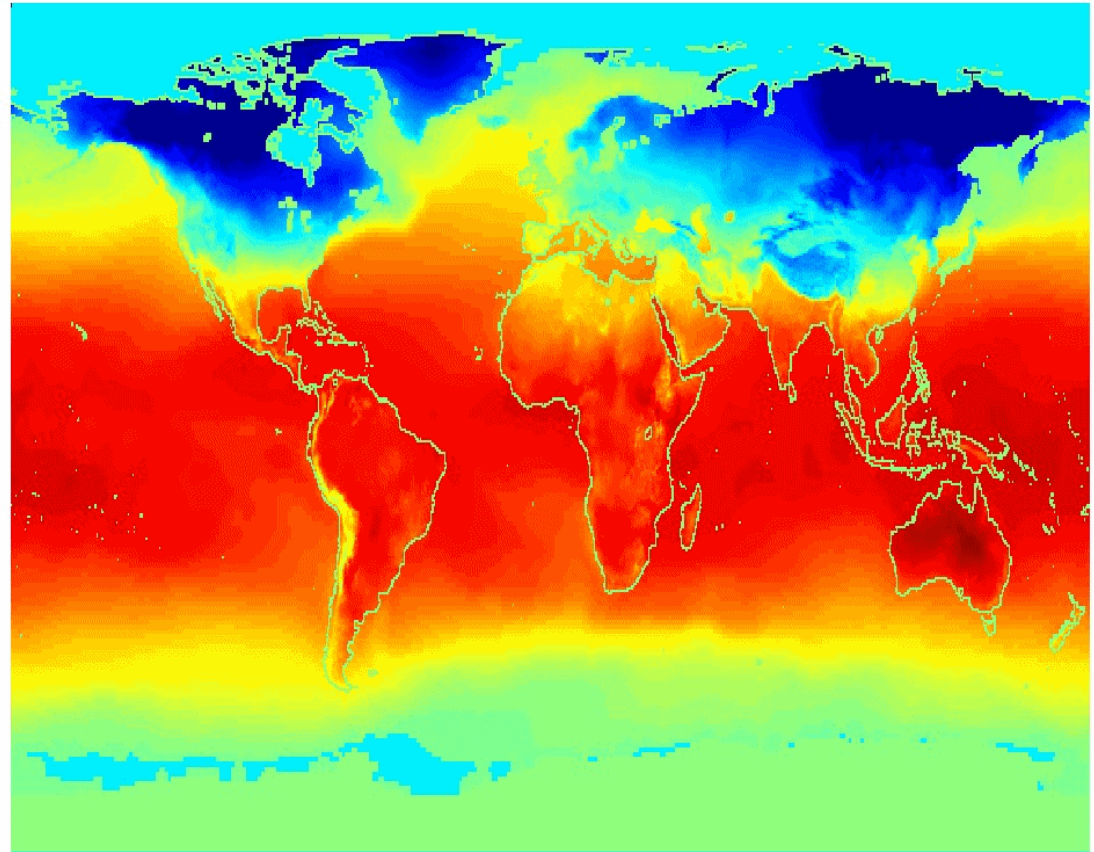
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG

Ordered Data

- Spatio-Temporal Data

**Average Monthly
Temperature of
land and ocean**

Jan



Data-related issues

- The type of data
- **The quality of the data**
- Preprocessing steps

Data Quality

- Data quality is a major concern in Data Mining and Knowledge Discovery tasks.
- Why: At most all Data Mining algorithms induce knowledge strictly from data.
- The **quality of knowledge** extracted highly depends on the **quality of data**.

Data Quality

- What problems should we worry about?
- How can we detect problems with the data?
- What can we do about these problems?

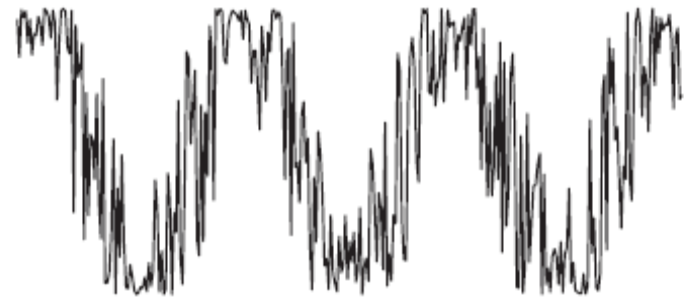
- Examples of data quality problems:
 - Noise and outliers
 - missing values
 - duplicate data

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) Time series.



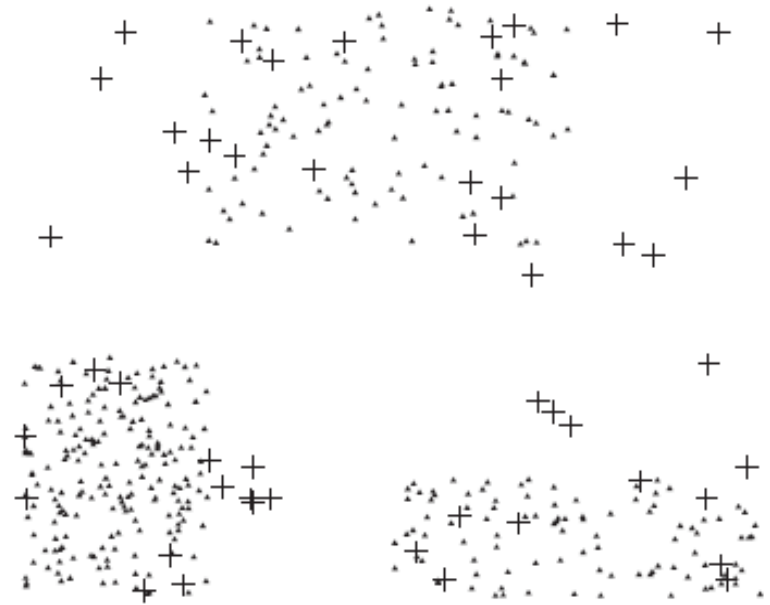
(b) Time series with noise.

Figure 2.5. Noise in a time series context.

Noise



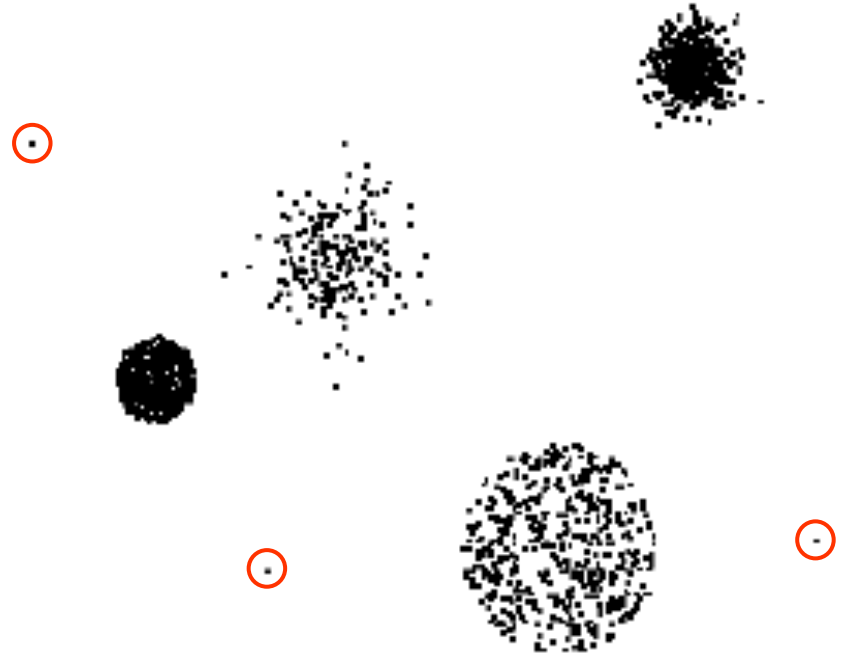
(a) Three groups of points.



(b) With noise points (+) added.

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



Quiz

- Distinguish between noise and outliers. Be sure to consider the following questions.
 - a) Is noise ever interesting or desirable?
 - b) Are outliers always noise objects?
 - c) Are noise objects always outliers?

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

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

Question

- Consider the problem of finding the K nearest neighbors of a data object.

- 1: for $i = 1$ to N do
- 2: Find the distances of the i th object to all other objects.
- 3: Sort these distances in decreasing order.
(Keep track of which object is associated with each distance.)
- 4: return the objects associated with the first K distances of the sorted list
- 5: end for

- a) Describe the potential problems with this algorithm if there are duplicate objects in the data set. Assume the distance function will only return a distance of 0 for objects that are the same.
- b) How would you fix this problem?

Data-related issues

- The type of data
- The quality of the data
- **Preprocessing steps**

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

Example: Precipitation in Australia

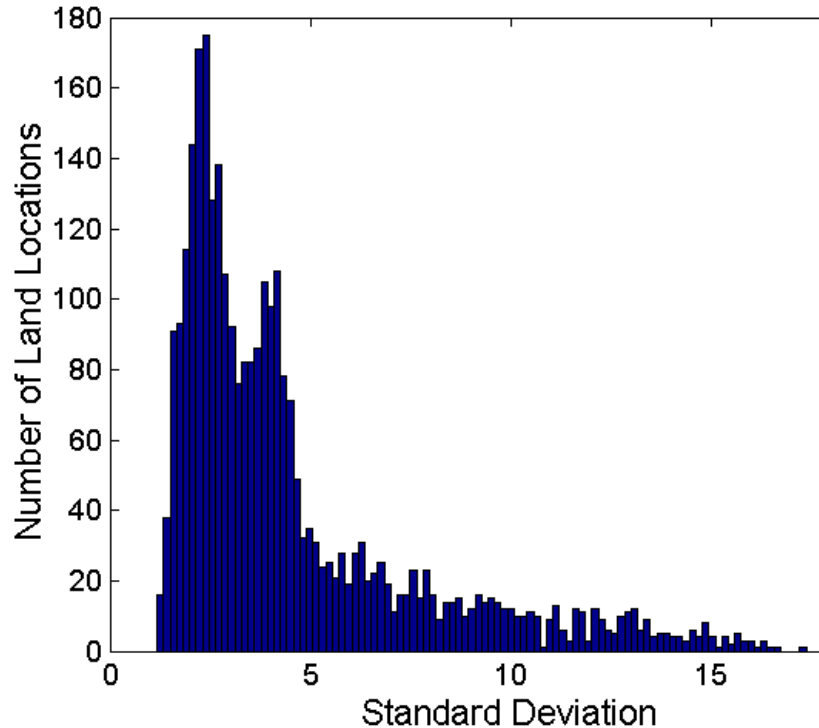
- This example is based on precipitation in Australia from the period 1982 to 1993.

The next slide shows

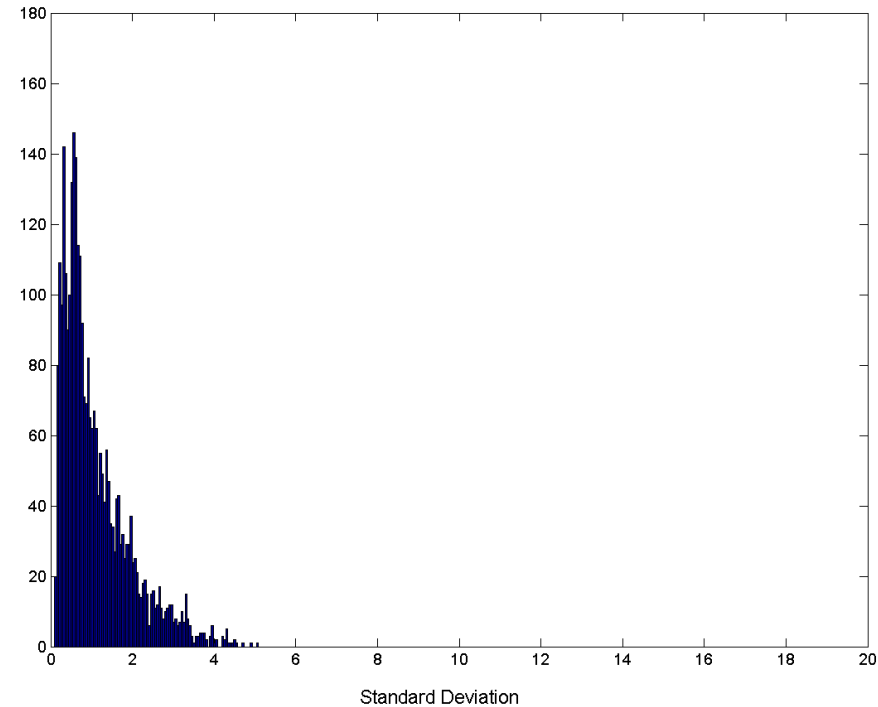
- A histogram for the standard deviation of average monthly precipitation for 3,030 0.5° by 0.5° grid cells in Australia, and
- A histogram for the standard deviation of the average yearly precipitation for the same locations.

Aggregation

Variation of Precipitation in Australia



Standard Deviation of Average
Monthly Precipitation



Standard Deviation of Average
Yearly Precipitation

The average **yearly** precipitation has less variability than the average **monthly** 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.

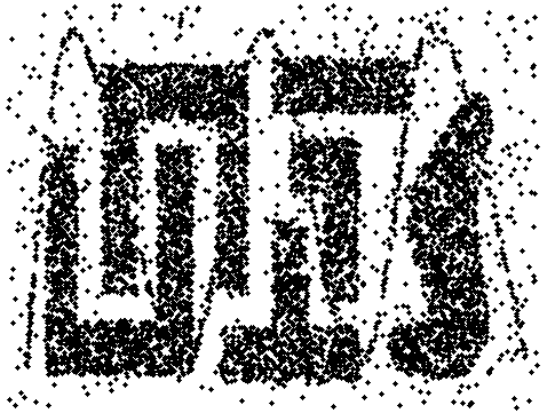
Sampling ...

- 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

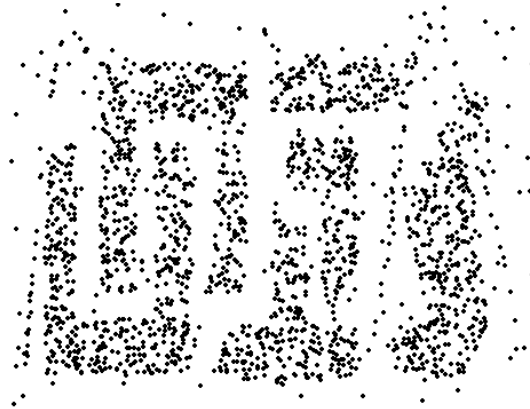
Types of Sampling

- 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

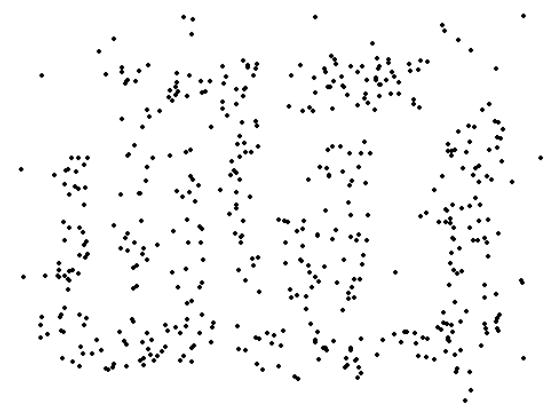
Sample Size



8000 points



2000 Points



500 Points

Quiz

- You are given a set of m objects that is divided into K groups, where the i^{th} group is of size m_i . If the goal is to obtain a sample of size $n < m$, what is the difference between the following two sampling schemes? (Assume sampling with replacement.)
 - a) We randomly select $n * m_i / m$ elements from each group.
 - b) We randomly select n elements from the data set, without regard for the group to which an object belongs.

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

Dimensionality Reduction

- Techniques

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

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

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

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
 - ◆ Example: dividing mass by volume to get density

Discretization

- **Discretization** is the process of converting a continuous attribute into an ordinal attribute
 - A potentially infinite number of values are mapped into a small number of categories
 - Discretization is commonly used in classification
 - Many classification algorithms work best if both the independent and dependent variables have only a few values

Binarization

- Binarization maps a continuous or categorical attribute into one or more binary variables
- Typically used for association analysis
- Often convert a continuous attribute to a categorical attribute and then convert a categorical attribute to a set of binary attributes
 - Association analysis needs asymmetric binary attributes
 - Examples: height measured as {low, medium, high}

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