Knowledge Discovery & Data Mining

– Introduction to KDD–

Instructor: Yong Zhuang

yong.zhuang@gvsu.edu

Outline

- What is Data Mining?
- What is Data?
- Attribute

What is Data Mining?

Definition 1:

- The aim of data mining is to make sense of large amounts of data in the context of some domain.

What is Data Mining?

Definition 1:

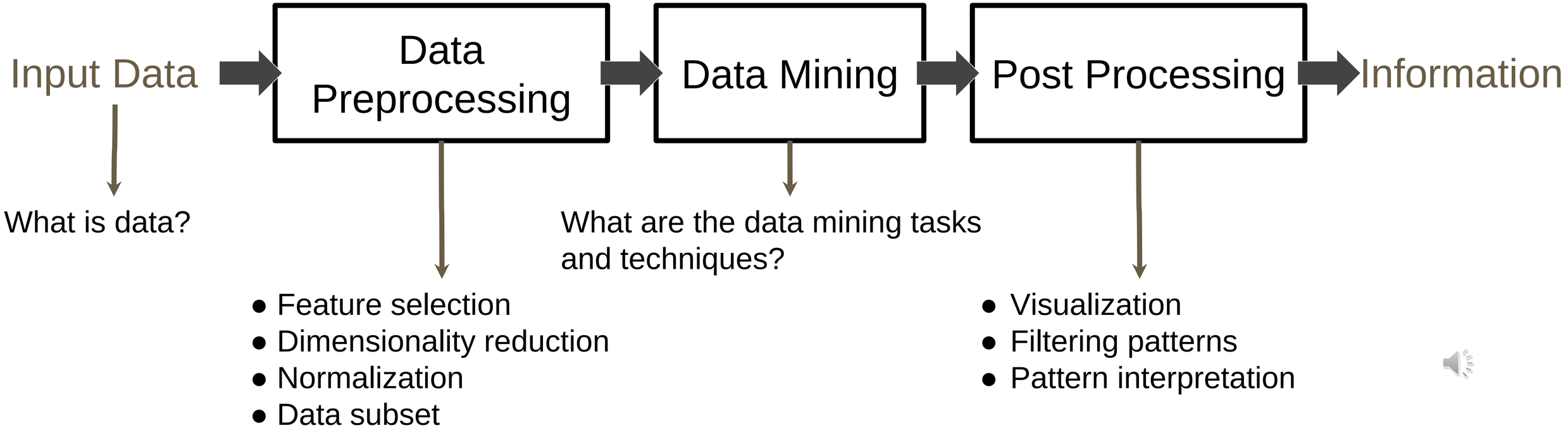
- The aim of data mining is to **make sense** of **large amounts of data** in the context of some **domain**.

- **Make sense** – gain insights that are understandable, valid, novel, and useful.
- **Large amounts of data** – data mining techniques are meant to be applied to large datasets that cannot be analyzed manually
- **In the context of some domain** – successful data mining projects depend heavily on access to domain knowledge (i.e., it's important to know where your data came from, why it's important, what each attribute means, etc.)!

What is Data Mining?

Definition 2:

- The process of automatically discovering useful information from large data repositories



Outline

- What is Data Mining?
- What is Data?
- Attribute

What is data?

Data is a collection of objects and their attributes.

Attribute is also be known as a feature, variable, field, or characteristic.

An **object** is also known as a record, data point, sample, entity, or instance.

Attributes

Objects

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2

An **attribute** is a property or characteristic of an object.

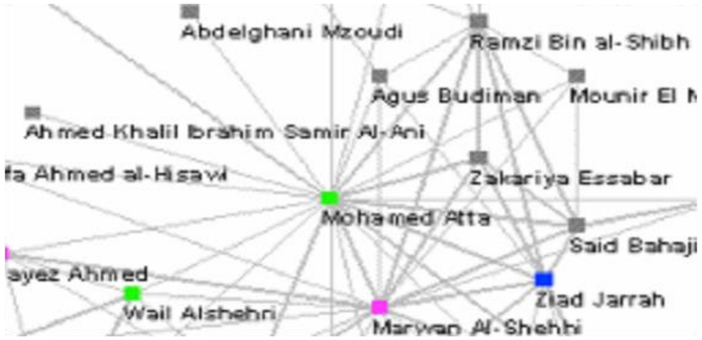
There are many types of data

Records

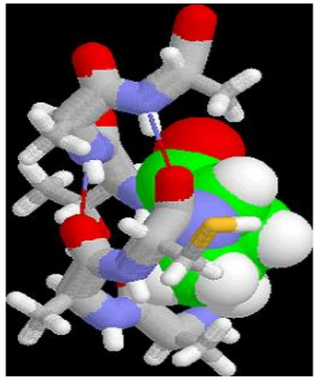
Tid	Refund	Marital Status	Taxable Income	Class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

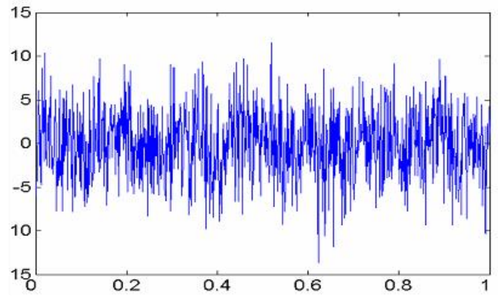
Image

Network (Link)

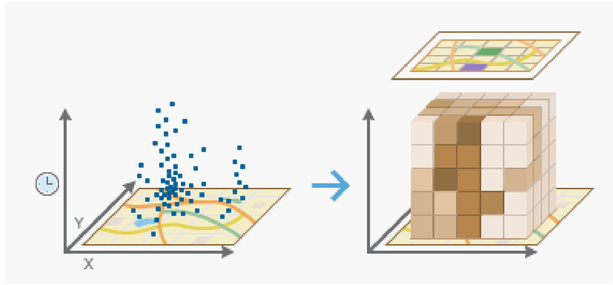


Graphs





Time series



Spatio-temporal

Where does your data come from?

- Primary data analysis
  - Data is generated with a particular question in mind through careful design of experiments
  - Data is analyzed to prove or disprove a hypothesis

- Secondary data analysis
  - Data is collected without specific question in mind
  - Data is analyzed to model its underlying structure and find consistent and replicable patterns

Data mining is mostly concerned with secondary data analysis

Outline

- What is Data Mining?
- What is Data?
- Attribute

Attribute

- An **attribute** is a property or characteristic of an object. The word “attribute”, “dimension”, “feature”, and “variable” are often used interchangeably in the literature.
  - **Dimension**: is commonly used in data warehousing.
  - **Feature**: is commonly used in machine learning.
  - **Variable**: is commonly used in statistics.
  - **Attribute**: is commonly used by data mining and database professionals

Attribute(Vocabularies)

- Observed values for a given attribute are known as observations.
- A set of attributes used to describe a given object is called an attribute vector(or feature vector)
- The distribution of data involving one attribute (or variable) is called univariate.

Attribute Values

- Numeric values – expressed as numbers (e.g., real numbers, integers, etc.)
- Symbolic values – describe qualitative concepts or categories

iata	name	city	state	country	latitude	longitude
00M	Thigpen	Bay Springs	MS	USA	31.95376472	-89.23450472
00R	Livingston Municipal	Livingston	TX	USA	30.68586111	-95.01792778
00V	Meadow Lake	Colorado Springs	CO	USA	38.94574889	-104.5698933
01G	Perry-Warsaw	Perry	NY	USA	42.74134667	-78.05208056
01J	Hilliard Airpark	Hilliard	FL	USA	30.6880125	-81.90594389
01M	Tishomingo County	Belmont	MS	USA	34.49166667	-88.20111111
02A	Gragg-Wade	Clanton	AL	USA	32.85048667	-86.61145333
02C	Capitol	Brookfield	WI	USA	43.08751	-88.17786917
02G	Columbiana County	East Liverpool	OH	USA	40.67331278	-80.64140639
03D	Memphis Memorial	Memphis	MO	USA	40.44725889	-92.22696056

Nominal Attribute Type (Symbolic)

Nominal:

- Do not have any meaningful order and are not quantitative
- Examples:
  - *Hair_color* = {*black, red, brown, grey*},
  - occupation, ID numbers, zip codes, …

Nominal Attribute Type (Symbolic)

Nominal:

- Do not have any meaningful order and are not quantitative
- Examples:
  - *Hair_color* = {*black, red, brown, grey*},
  - occupation, ID numbers, zip codes, …

If Hair_color = {0,2,5,8}
where 0: black, 2: red, 5: brown, and 8: grey
What is the type of Hair_color?

Nominal Attribute Type (Symbolic)

Nominal:

- Do not have any meaningful order and are not quantitative
- Examples:
  - *Hair_color* = {*black, red, brown, grey*},
  - occupation, ID numbers, zip codes, …

If Hair_color = {0,2,5,8}
where 0: black, 2: blond, 5: brown, and 8: grey
What is the type of Hair_color?

Nominal !

Binary Attribute Type (Symbolic)

Binary: Nominal attribute with only 2 states (0 and 1)

- **Symmetric** binary: both outcomes equally important
  - e.g., gender
- **Asymmetric** binary: outcomes not equally important.
  - e.g., medical test (positive vs. negative)
  - Convention: we code the most important outcome, which is usually the rarer one, by 1 (e.g., HIV positive) and the other by 0 (e.g., HIV negative).

Ordinal Attribute Type (Symbolic)

Ordinal: Values have a meaningful order (ranking) but magnitude between successive values is not known.

- *drink_size = {small, medium, large},*
- *grades = {A+, A, A−, B+, …}*
- *customer satisfaction = {1: very dissatisfied, 2: dissatisfied, 3: neutral, 4: satisfied, and 5: very satisfied}*

Interval-scaled Attribute (Numeric)

Interval-scaled attributes:

- Measured on a scale of **equal-sized units**
- can be positive, 0, or negative.
- Values have order
  - E.g., *temperature, calendar dates*
- No true zero-point
  - The year 0 does not correspond to the beginning of time, so we cannot say year 2000 is twice as year 1000

Ratio-scaled Attribute (Numeric)

Ratio-scaled attributes:

- Inherent **zero-point**
- e.g., years_of_experience (the objects are employees), *number_of_words (the objects are documents), weight, height, and speed…*
  - *You are 100 times richer with $100 than with $1.*

Discrete vs. Continuous Attributes

There are many ways to organize attribute types. The field of machine learning typically treats attributes as either discrete or continuous.

Discrete Attribute

- Has only a finite or countably infinite set of values
  - Finite number of values: *hair_color, smoker, medical_test, and drink_size*
  - Integer variables: the values 0 to 110 for the attribute *age*
  - Countably infinite set of values: *customer_ID, Zip codes*
- Note: Binary attributes are a special case of discrete attributes

Continuous Attribute:

- Has real numbers as attribute values
  - E.g., temperature, height, or weight
- In practice, Continuous attributes are typically represented as floating-point variables.

Example

What are the types of attributes in the following dataset?

Pause

iata	name	city	state	country	latitude	longitude
00M	Thigpen	Bay Springs	MS	USA	31.95376472	-89.23450472
00R	Livingston Municipal	Livingston	TX	USA	30.68586111	-95.01792778
00V	Meadow Lake	Colorado Springs	CO	USA	38.94574889	-104.5698933
01G	Perry-Warsaw	Perry	NY	USA	42.74134667	-78.05208056
01J	Hilliard Airpark	Hilliard	FL	USA	30.6880125	-81.90594389
01M	Tishomingo County	Belmont	MS	USA	34.49166667	-88.20111111
02A	Gragg-Wade	Clanton	AL	USA	32.85048667	-86.61145333
02C	Capitol	Brookfield	WI	USA	43.08751	-88.17786917
02G	Columbiana County	East Liverpool	OH	USA	40.67331278	-80.64140639
03D	Memphis Memorial	Memphis	MO	USA	40.44725889	-92.22696056

Example

What are the types of attributes in the following dataset?

Pause

iata	name	city	state	country	latitude	longitude
00M	Thigpen	Bay Springs	MS	USA	31.95376472	-89.23450472
00R	Livingston Municipal	Livingston	TX	USA	30.68586111	-95.01792778
00V	Meadow Lake	Colorado Springs	CO	USA	38.94574889	-104.5698933
01G	Perry-Warsaw	Perry	NY	USA	42.74134667	-78.05208056
01J	Hilliard Airpark	Hilliard	FL	USA	30.6880125	-81.90594389
01M	Tishomingo County	Belmont	MS	USA	34.49166667	-88.20111111
02A	Gragg-Wade	Clanton	AL	USA	32.85048667	-86.61145333
02C	Capitol	Brookfield	WI	USA	43.08751	-88.17786917
02G	Columbiana County	East Liverpool	OH	USA	40.67331278	-80.64140639
03D	Memphis Memorial	Memphis	MO	USA	40.44725889	-92.22696056

- **name(**symbolic): discrete, nominal
- **city(**symbolic): discrete, nominal
- **state(**symbolic): discrete, nominal
- **country(**symbolic): discrete, nominal
- **latitude(**numeric): continuous, Interval-scaled
- **longitude(**numeric): continuous, Interval-scaled

Summary

- What is Data Mining?
- What is Data?
- Attribute
  - Symbolic: Nominal, Binary, Ordinal
  - Numeric: Interval-scaled, Ratio-scaled,
  - Discrete vs. Continuous