Data Mining: Introduction

Lecture Notes for Chapter 2

Exploratory Data Analysis

Outline

Unit 2: Exploratory Data Analysis [6 Hrs.]

- Sources and types of data
- Non-graphical and graphical methods for exploring univariate, bivariate, multivariate data
- Visualization techniques

What are Data Source Types?

1. **Databases:** Structured data stored in relational databases like SQL, NoSQL databases, or data warehouses.
2. **APIs:** Data fetched from web services or applications via API calls.
3. **Flat Files:** Data from CSVs, Excel sheets, text files, or XML/JSON formats.
4. **Streaming Data:** Real-time data from IoT devices, sensors, or live feeds.
5. **Cloud Services:** Data stored in cloud platforms like AWS, Google Cloud, or Azure.
6. **Manual Input:** Data entered manually by users or operators into systems.
7. **Other Sources:** Data from alternative sources like RSS feeds, social media, or web scraping tools, often providing unstructured or semi-structured data that adds real-time insights to your analysis.

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

Attributes

Objects

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 for a particular object

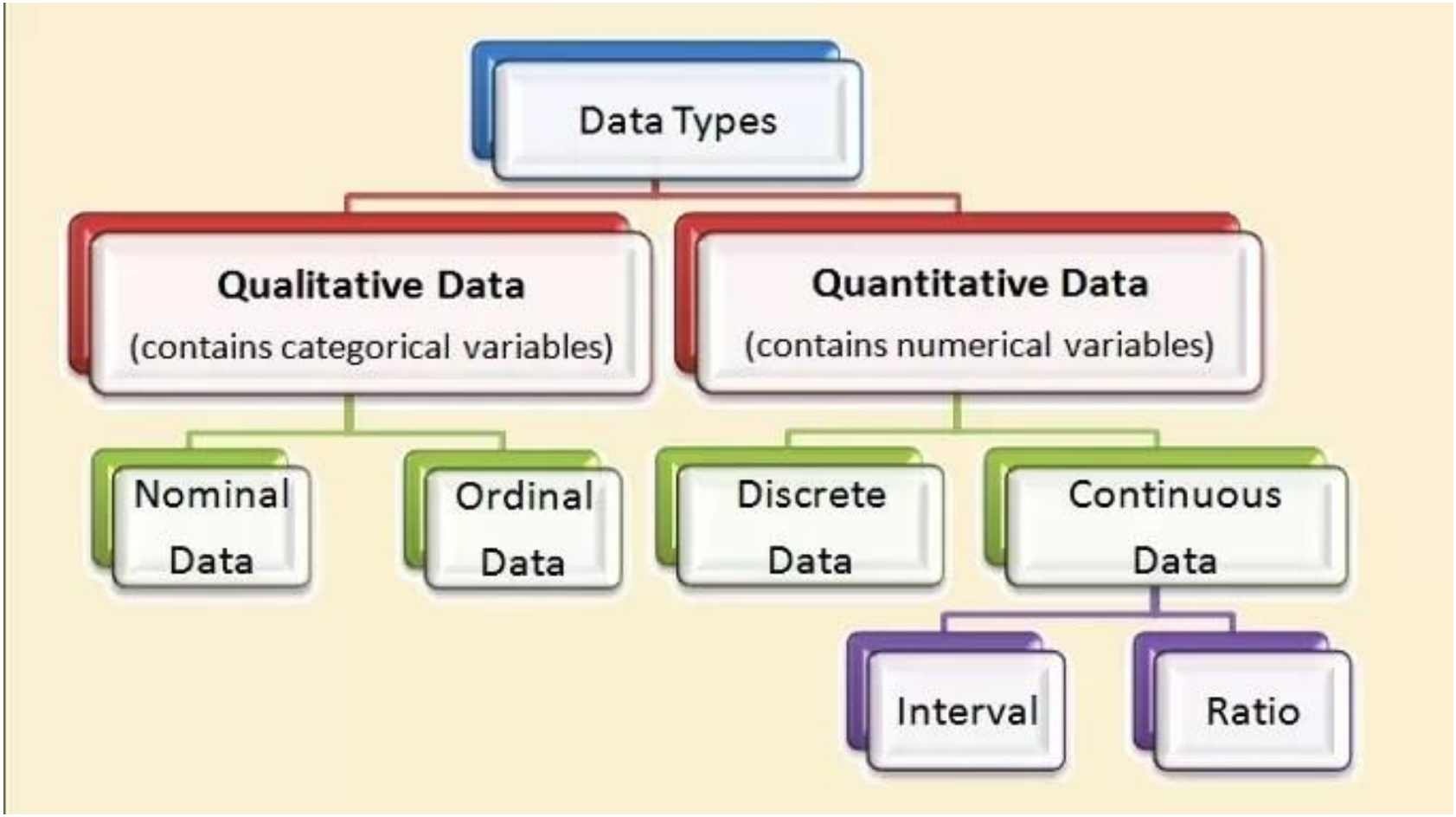
- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
  - But properties of attribute can be different than the properties of the values used to represent the attribute

Types of Data

Types of attributes in DM

1. **Qualitative Attributes** (Nominal (N), Ordinal (O), Binary(B))
2. **Quantitative Attributes** (Numeric, Discrete, Continuous)
3. **Biological Sequences**
4. **Time Series**
5. **Images**
6. **Sound**
7. **Video**

Types of Data



Types of Attributes: Approach 1

Attribute Type		Description	Examples
Categorical (Qualitative)	Nominal	The values of a nominal attribute are just different names; i.e., nominal values provide only enough information to distinguish one object from another. $(=, \neq)$	zip codes, employee ID numbers, eye color, gender
	Ordinal	The values of an ordinal attribute provide enough information to order objects. $(<, >)$	hardness of minerals, $\{good, better, best\}$, grades, street numbers
Numeric (Quantitative)	Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. $(+, -)$	calendar dates, temperature in Celsius or Fahrenheit
	Ratio	For ratio variables, both differences and ratios are meaningful. $(*, /)$	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current

Approach 2: 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

Approach 3

- **Character**: values are represented in forms of character or set of characters (string).
- **Number**: values are represented in forms of number.
- Number may be in form of whole number, decimal number.

Types of data sets

1. **Record**
   - Data that consists of a collection of records, each of which consists of a fixed set of attributes.
     - **Data Matrix**
     - **Document Data**
     - **Transaction Data**
2. **Graph**
3. **Ordered**

Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute
- Such data set can be represented by an ***m*** by ***n*** matrix, where there are m rows, one for each object, and n columns, one for each attribute

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, 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	pla y	ball	score	game	wi n	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

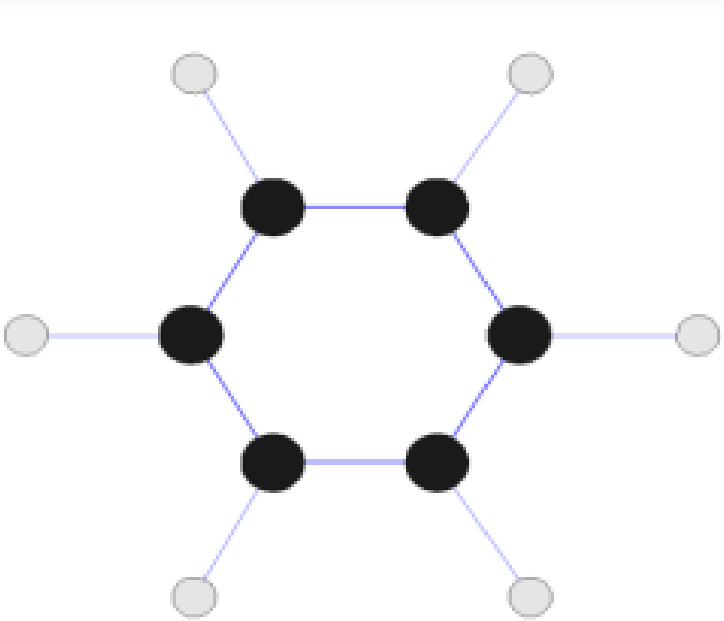
Transaction Data

- 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

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Graph

- Contains notes and connecting vertices.
- Eg: World Wide Web, Molecular Structures



Ordered

- Has Sequences of transactions
- Spatial Data
  - **Spatial data**, also known as geospatial data, is information about a physical object that can be represented by numerical values in a geographic coordinate system.
  - **Temporal Data** A temporal data denotes the evolution of an object characteristic over a period of time. Eg d=f(t).
  - **Sequential Data** Data arranged in sequence.

Continued…

- Non-graphical and graphical methods for exploring univariate, bivariate, multivariate data

...

- **EDA** is a way of exploring data through visual summaries and graphics, and there are several different types of EDA to choose from.
- **Univariate EDA** involves looking at a single variable at a time.
  - Univariate EDA can help you understand the data distribution and identify any outliers.

Univariate analysis

Non-Graphical Methods:

- **Summary statistics**:
  - Mean, median, mode
  - Variance, standard deviation
  - Range, interquartile range (IQR)
  - Skewness and kurtosis
- **Tables**:
  - Frequency tables
  - Percentile tables

...

Graphical Methods:

- **Histogram**: Distribution shape (normal, skewed, etc.)
- **Boxplot**: Spread, outliers, median
- **Stem-and-leaf plot**: Detailed distribution
- **Dot plot**: Simple view of distribution
- **Bar chart** (for categorical data)

Bivariate Analysis

Bivariate EDA involves looking at two variables at a time.

- can help you understand the relationship between two variables and identify any patterns that might exist.

Non-Graphical Methods:

Correlation coefficient (like Pearson's r): Measures linear relationship strength.

Covariance: Measures how two variables change together.

Contingency table (for categorical data): Joint frequencies.

Two-sample t-test (for comparing means).

...

Graphical Methods:

- **Scatter plot**: Visualizes relationship between two quantitative variables.
- **Side-by-side boxplots**: Compare distributions across categories.
- **Line graph**: Useful for trends over time.
- **Grouped bar chart**: For categorical data comparison.

...

- **Multivariate EDA** involves looking at three or more variables at a time.
- Multivariate EDA can help you understand the relationships between several variables and identify any complex patterns or outliers that might exist.

Multivariate Data (More than two variables)

Non-Graphical Methods:

- **Correlation matrix**: Pairwise correlations between variables.
- **Descriptive statistics table**: Means, variances, etc. for each variable.
- **Multivariate tests**:
  - MANOVA (Multivariate Analysis of Variance)
  - Multiple regression analysis

...

Graphical Methods:

- **Scatterplot matrix**: Multiple scatterplots between all variable pairs.
- **3D scatter plots**: For visualizing three variables together.
- **Heatmaps**: Show correlation matrices or other relationships.
- **Parallel coordinates plot**: Lines connecting multiple variables.
- **Bubble chart**: Adds a third variable using size.

Summary

Type	Non-Graphical Methods	Graphical Methods
Univariate	- Mean, Median, Mode - Variance, Std. Dev - Range, IQR - Frequency Table	- Histogram - Boxplot - Stem-and-Leaf Plot - Dot Plot - Bar Chart
Bivariate	- Correlation Coefficient - Covariance - Contingency Table - Two-sample t-test	- Scatter Plot - Side-by-Side Boxplots - Line Graph - Grouped Bar Chart
Multivariate	- Correlation Matrix - Descriptive Stats Table - MANOVA - Multiple Regression	- Scatterplot Matrix - 3D Scatter Plot - Heatmap - Parallel Coordinates Plot - Bubble Chart

The End