Contents

1. Data Structure
   - Data structure
   - Discrete & Continuous Attribute
   - Tabulation
   - Structured & unstructured data

2. Types of data
   - Record-based data
   - Graph-based data
   - Ordered data
Data Structure

Introduction to Big Data

01
Data Structure

What is dataset?

- What is dataset?

Attributes, Characteristics, Features, Variables, ..., etc.

<table>
<thead>
<tr>
<th>Tid</th>
<th>Refund</th>
<th>Marital Status</th>
<th>Taxable Income</th>
<th>Cheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Single</td>
<td>125K</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Married</td>
<td>100K</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Single</td>
<td>70K</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Married</td>
<td>120K</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Divorced</td>
<td>95K</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Married</td>
<td>60K</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Divorced</td>
<td>220K</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>Single</td>
<td>85K</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>Married</td>
<td>75K</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>No</td>
<td>Single</td>
<td>90K</td>
<td>Yes</td>
</tr>
</tbody>
</table>

- Set of objects consisting of their attributes.

Object

Instance

Individual

Sample

Subject

...
Data Structure
What is dataset?

➢ What is variables?
   Attributes, Characteristics, Features, ..., etc.

• Properties / features of specific object.

• i.e. Human
  • Eye size (mm unit)
  • Eye color
  • Skin color
  • Height (cm unit)
  • Wear glasses or not
  • Gender
  • Age
  • Length of finger (cm unit)
  • ...
Discrete & Continuous Attribute

Types of variable

➢ Continuous Attribute

• It has a real number as a property value

• A continuous variable is one which can take on infinitely many, uncountable values.

• Continuous attributes are usually represented as floating point variables
In Korean...?
Discrete & Continuous Attribute

Types of variable

➢ Discrete Attribute

• Finite or infinite set of countable values
• Usually expressed as an integer variable
• Binary attribute is a special form of discrete attribute
• Continuous variables can also be converted to discrete variables through binning
Tabulation
Importance of Tabulation

➢ It is a systematic and logical arrangement of classified data in rows and columns.

➢ Let’s think about how we can collect structured or unstructured data
Collection of the data
Two types of processes

Tabulation process

1. Experimental design
2. Determination of population
3. Set variables
4. Get values

➢ Usually, this is data-science process

Tabulation process

1. Experimental design
2. Determination of population
3. Get values
4. Prescreening

➢ Usually, we mentioned that ‘Data Mining’
Let’s try tabulation
Practice of tabulation

<table>
<thead>
<tr>
<th>AAAAA</th>
<th>AAAAA</th>
<th>AAAAA</th>
<th>AAAAA</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB</td>
<td>BB</td>
<td>BB</td>
<td>BB</td>
</tr>
<tr>
<td>CCC</td>
<td>CCC</td>
<td>CCC</td>
<td>CCC</td>
</tr>
<tr>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
</tr>
</tbody>
</table>

➢ Try to tabulize above unstructured data
Features of ‘unstructured’ data
Importance of tabulation

➢ It does not reside in traditional databases and data warehouses.

➢ It may have an internal structure, but does not fit a relational data model.

➢ It generated by both humans and machines.

➢ It will usually be textual and multimedia content, Machine-to-machine communication

➢ Examples include
  • Personal messaging – email, instant messages, tweets, chat
  • Business documents – business reports, presentations, survey responses
  • Web content – web pages, blogs, wikis, audio files, photos, videos
  • Sensor output – satellite imagery, geolocation data, scanner transactions
Current data
Zettabytes’ unstructured data

About 85% is unstructured data

1ZB = 1,126,000,000,000,000,000,000 bytes (approx)
Data sets of such size, complexity and volatility that their business value cannot be fully realised with existing data capture, storage, processing, analysis and management capabilities.
Types of Data
Introduction to Big Data
Types of Dataset

Data... data... data.......

- **Record-based data**
  - Data matrix
  - Document data
  - Transaction data
  - ...

- **Graph-based data**
  - World wide web
  - Molecular structure
  - Map data
  - ...

- **Ordered data**
  - Spatial data
  - Temporal data
  - Sequential data
  - Genetic/Genomic sequence data

<table>
<thead>
<tr>
<th>Sale ID</th>
<th>Time</th>
<th>Customer</th>
<th>Product ID</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>S00001</td>
<td>12/1/2012 9:00:00 AM</td>
<td>C0001</td>
<td>P025</td>
<td>1</td>
</tr>
<tr>
<td>S00002</td>
<td>12/1/2012 9:05:58 AM</td>
<td>C0025</td>
<td>P025</td>
<td>3</td>
</tr>
<tr>
<td>S00003</td>
<td>12/1/2012 9:11:33 AM</td>
<td>C0010</td>
<td>P001</td>
<td>2</td>
</tr>
<tr>
<td>S00004</td>
<td>12/1/2012 9:17:16 AM</td>
<td>C0017</td>
<td>P023</td>
<td>4</td>
</tr>
<tr>
<td>S00005</td>
<td>12/1/2012 9:23:04 AM</td>
<td>C0018</td>
<td>P016</td>
<td>5</td>
</tr>
<tr>
<td>S00006</td>
<td>12/1/2012 9:28:43 AM</td>
<td>C0011</td>
<td>P018</td>
<td>4</td>
</tr>
<tr>
<td>S00007</td>
<td>12/1/2012 9:34:07 AM</td>
<td>C0015</td>
<td>P006</td>
<td>1</td>
</tr>
</tbody>
</table>
Types of Dataset

Record-based data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

<table>
<thead>
<tr>
<th>Tid</th>
<th>Refund</th>
<th>Marital Status</th>
<th>Taxable Income</th>
<th>Cheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Single</td>
<td>125K</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Married</td>
<td>100K</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Single</td>
<td>70K</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Married</td>
<td>120K</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Divorced</td>
<td>95K</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Married</td>
<td>60K</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Divorced</td>
<td>220K</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>Single</td>
<td>85K</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>Married</td>
<td>75K</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>No</td>
<td>Single</td>
<td>90K</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Types of Dataset

Data matrix

- This is a term derived from linear algebra.

- When composed of a fixed number of numerical attributes, an object (record) can be considered as a point in multidimensional space.

- Such data is represented by \( n \times p \) matrices, where \( n \) rows each represent an object and \( p \) columns each represent an attribute.

<table>
<thead>
<tr>
<th>Projection of x Load</th>
<th>Projection of y load</th>
<th>Distance</th>
<th>Load</th>
<th>Thickness</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.23</td>
<td>5.27</td>
<td>15.22</td>
<td>2.7</td>
<td>1.2</td>
</tr>
<tr>
<td>12.65</td>
<td>6.25</td>
<td>16.22</td>
<td>2.2</td>
<td>1.1</td>
</tr>
</tbody>
</table>
Types of Dataset

Document data

- Each document can be represented by a term vector.
- Each term corresponds to a component of the vector.
- Each value corresponds to the number of times the term appeared in the document.

<table>
<thead>
<tr>
<th></th>
<th>team</th>
<th>coach</th>
<th>play</th>
<th>ball</th>
<th>score</th>
<th>game</th>
<th>win</th>
<th>lost</th>
<th>timeout</th>
<th>season</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc 1</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Doc 2</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Doc 3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>
Types of Dataset
Transaction Data

- As a special type of record, each record (transaction) is a collection of items.

- It also known as ‘market basket data’.

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Coke, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Beer, Bread</td>
</tr>
<tr>
<td>3</td>
<td>Beer, Coke, Diaper, Milk</td>
</tr>
<tr>
<td>4</td>
<td>Beer, Bread, Diaper, Milk</td>
</tr>
<tr>
<td>5</td>
<td>Coke, Diaper, Milk</td>
</tr>
</tbody>
</table>
Types of Dataset
Graph Data

G = (V, E)
- V is set of vertices (nodes)
- E is set of edges (arcs or link)
Types of Dataset

Graph Data

➢ Set of HTML documents
Types of Dataset

Graph Data

- Chemical structure data

- Social network data
Types of Dataset
Ordered data

- Sequences of transactions

Items/Events

( A B) (D) (C E)
(B D) (C) (E)
(C D) (B) (A E)

An element of the sequence
Types of Dataset

Ordered data

- Genetic / Genomics sequence data
Types of Dataset

Ordered data

- Time-series data
Types of Dataset
Ordered data

➢ Spatio- Temporal data

Average Monthly Temperature of land and ocean
Topic of next class
Coming soon...

Thursday
- Introduction to R programming

Next Wednesday
- Data Preprocessing
- Quality control
- Similarity & dissimilarity
- Distance metric
- Fundamental statistics