Understanding of Structured Dataset

BY:

DR. VIPIN KUMAR

DEPARTMENT OF COMPUTER SCIENCE &IT

MAHATMA GANDHI CENTRAL UNIVERSITY

MOTIHARI, BIHAR

Outlines…

- Defining data
- Types of data
  - Structured data
  - Semi-structured data
  - Unstructured data
- Representation of structured data
- Description of keys related to data description
- Types of attributes
  - Categorical attribute
  - Continuous attribute
- Example of Iris dataset with description

Types of Data

There are three types of dataset:

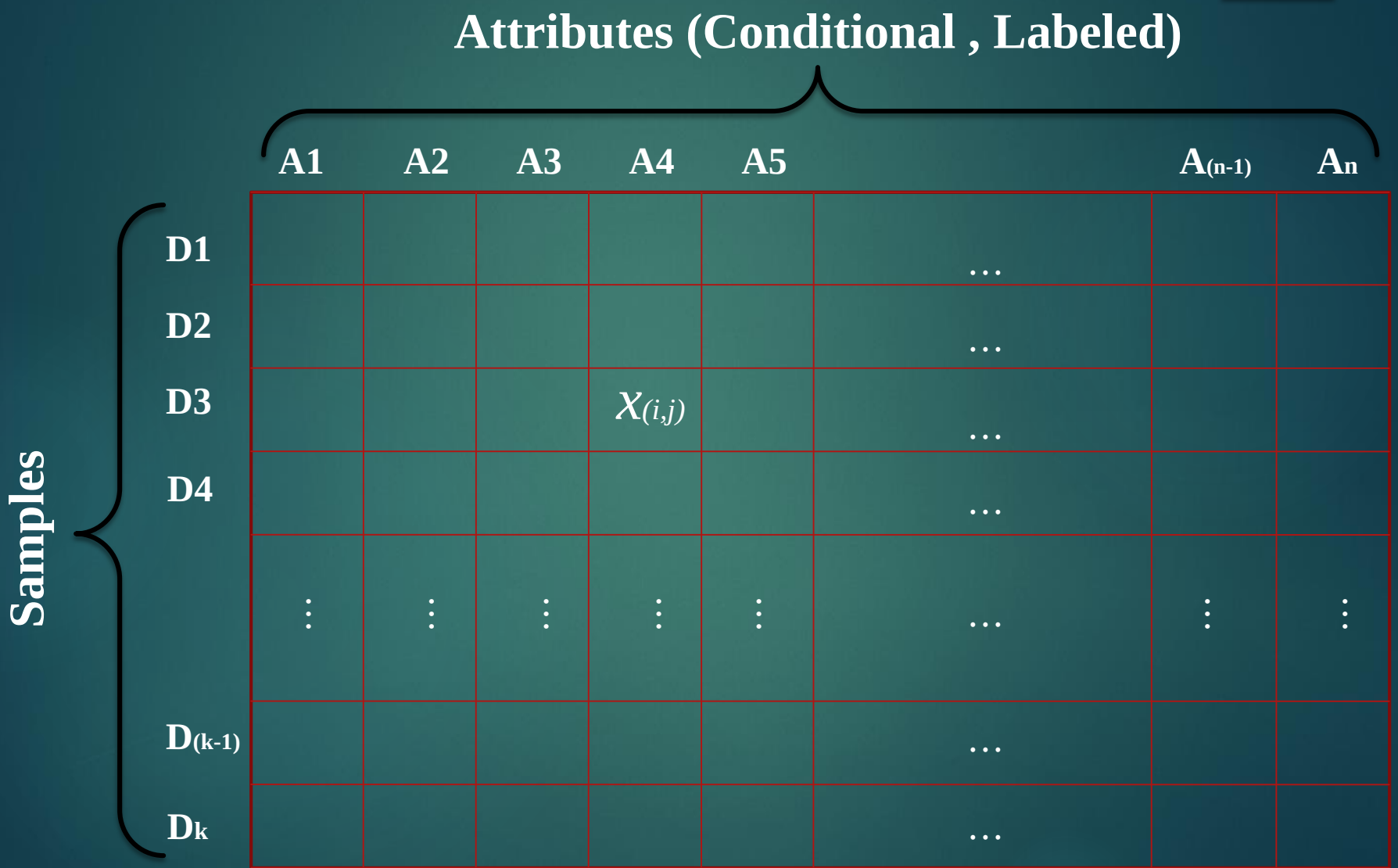
1. Structured data:

   - This type of data comprises with rows and columns, where generally rows and columns implies to instance and their corresponding features respectively.
   - For instance: Tabular data, SQL database etc.

Types of Data

2. Semi-structure data:
   - It has structure but not as rows and columns format.
   - For instance: HTML, XML, JSON, etc.
3. Unstructured data:
   - Data can stored in any format instead of structured and semi-structured format.
   - It may have audios videos, images, signals reading, text corpus, etc.

Structured Data Representation



Structured Data Representation

- **Rows** in structured data are generally called as instances, samples, experiences, entities.
- **Columns** in structured data are commonly called as Attributes or Features of instances.
- **Instance or Samples** are the independent entities which have recorded separately while data collection.
- **Features or Attributes** is a set of characteristics of individual samples which are independently recorded. There are types of attribute namely Conditional attribute and Decision attribute.

Structured Data Representation

- Types of attributes:
  - Broadly, it can be divided into two categories: Continues and categorical attributes
  - Continuous attribute: it has integer or floating values corresponding to all samples.
    - e.g. salary of employee, age of people, temperature, device signals, etc.
  - Categorical attribute: It has discreet values of corresponding to all samples.
    - e.g. [Male, Female], [low medium, high], [government, private] etc.

Weather Dataset

S.N.	Outlook	Temperature	Humidity	Windy	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rainy	Mild	High	Weak	Yes
D5	Rainy	Cool	Normal	Weak	Yes
D6	Rainy	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rainy	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rainy	Mild	high	Strong	No

Iris dataset
(Donor: Michael Marshall, 1988)

- Number of Samples= 150 (50 each classes)
- Number of Attribute= 04
- Number of classes= 03
  - Iris-setosa
  - Iris-versicolour
  - Iris-virginica

Conti...

S.N.	Sepetal Length (cm)	Sepetal Width (cm)	Petal Length (cm)	Petal Width (cm)	Classes
D1	5.4	3.4	5.0	0.4	Iris-Setosa
D2	5.2	4.1	5.0	0.1	Iris-Setosa
D3	5.5	4.2	4.0	0.2	Iris-Setosa
D4	7.0	3.2	4.7	1.4	Iris-versicolour
D5	6.4	3.2	4.5	1.5	Iris-versicolour
D6	6.9	3.1	4.9	1.5	Iris-versicolour
D7	6.3	3.3	6.0	2.5	Iris-verginica
D8	5.8	2.7	5.1	19	Iris-verginica
D9	7.1	3.0	5.9	2.1	Iris-verginica

Bibliography

- Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Pearson Education India, 2016.
- Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.
- Mitchell, Tom M. "Machine learning." (1997).
- Alpaydin, Ethem. *Introduction to machine learning*. MIT press, 2020.
- Murphy, Kevin P. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- Burkov, Andriy. *The hundred-page machine learning book*. Quebec City, Can.: Andriy Burkov, 2019.
- Burkov, Andriy. *The hundred-page machine learning book*. Quebec City, Can.: Andriy Burkov, 2019.

Thank You