UGANDA CHRISTIAN UNIVERSITY
A Centre of Excellence in the Heart of Africa

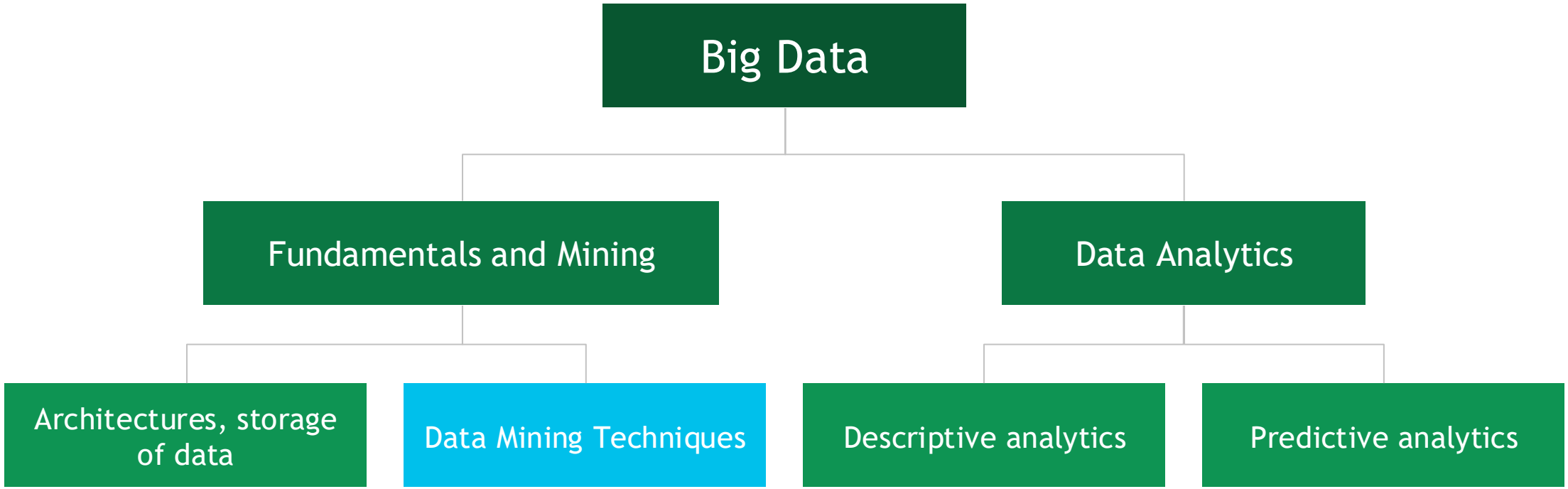
DSC3108: Big Data Mining and Analytics

Lecture 03 (BSCS, BSDS_3:1)

Topic: *Other Data Mining Techniques-Association Mining*

Dr. Daphne Nyachaki Bitalo
Department of Computing & Technology
Faculty of Engineering, Design & Technology

COURSE OVERVIEW



Lecture Objectives and Learning outcomes

The Objectives of this lecture are :

- Understand association data mining techniques that can be used for predictive analyses

By the end of this lecture, students should be able to:

- Get practical experience working with association mining algorithms and unsupervised learning.

Lecture Overview

- Data Mining Techniques
- Association mining parameters
- Association mining algorithms
- Applications of association mining
- Hands-on practical

Data Mining Techniques

1. Data mining process: CRISP-DM methodology
2. Data exploration and visualization
3. Association rule mining
4. Classification algorithms (Decision trees, Naïve Bayes, Support Vector Machines)
5. Clustering algorithms (K-means, hierarchical clustering)
6. Outlier detection

CRISP-DM Method

CRISP-DM (Cross-Industry Standard Process for Data Mining) is a widely used methodology that provides a framework for conducting data mining projects. It outlines a series of phases that guide the process from business understanding to deployment.

Documentation: Maintain clear documentation throughout the project.

CRISP-DM Method

Phases of CRISP-DM:

1. **Business Understanding:**
   - ❑ Define the business objectives and goals.
   - ❑ Identify the relevant data sources.
   - ❑ Create a project plan.
2. **Data Understanding:**
   - ❑ Collect and gather the necessary data.
   - ❑ Explore the data to understand its characteristics, quality, and completeness.
   - ❑ Identify potential data quality issues.

CRISP-DM Method

3. Data Preparation:

- ❑ Clean and preprocess the data to address any quality issues.
- ❑ Transform the data into a suitable format for analysis.
- ❑ Create features or attributes that are relevant to the problem.

4. Modeling:

- ❑ Select appropriate data mining techniques based on the business objectives.
- ❑ Build and train models using the prepared data.
- ❑ Evaluate the performance of the models.

CRISP-DM Method

5. **Evaluation:**

- ❑ Assess the quality and reliability of the models.
- ❑ Compare the performance of different models.
- ❑ Validate the models using unseen data.

6. Deployment:

- ❑ Integrate the chosen model into the production environment.
- ❑ Monitor the model's performance and update it as needed.

Data exploration and visualisation

In data mining, exploration helps an analyst understand the data, identify patterns, and communicate findings effectively. Handles various aspects such as;

Summary Statistics: Calculate measures like mean, median, mode, standard deviation, and percentiles to get a basic understanding of the data distribution.

Data Profiling: Examine data types, missing values, outliers, and inconsistencies.

Univariate Analysis: Analyze individual variables to understand their distributions and characteristics.

Data exploration and visualisation

Handles various aspects such as;

Bivariate Analysis: Examine relationships between pairs of variables.

Multivariate Analysis: Explore relationships among multiple variables.

Data Storytelling: Use visualizations to create compelling narratives and communicate insights effectively.

Association rule mining

- This technique discovers interesting relationships between items in a large dataset. These relationships, typically represented as rules, can be used to make predictions or recommendations.
- The technique uses unsupervised machine learning algorithms to find the hidden rules (associations) in data
- Association rule mining assigns key statistical parameters that are different from correlations (i.e. Support, Lift, Confidence)

Association rule mining parameters

1. Support

Support is a measure of how frequent a item or an item set appears in a dataset. For example, what is the support for the item set { Milk + Cheese } everytime a customer goes to a shop?

Order	Item
1	Milk
2	**Milk, Cheese**
3	Milk, Water, Vegetables
4	**Milk, Cheese**, Yogurt, Water
5	Water, Vegetables

$$Support_{Milk+Cheese} = \frac{Occurances of Milk, Cheese}{Total} = \frac{2}{5} = 0.4$$

$$Support_{X+Y} = How\ Frequent\ is\ the\ combination - \{X + Y\}$$

Association rule mining parameters

2. Confidence

Confidence is a measure of how often this rule is found to be true. Or the probability that an event will occur given that another event has occurred. So what is the probability that a customer who purchases milk will also purchase cheese?

$$Confidence_{Milk->Cheese} = \frac{Support_{Milk+Cheese}}{Support_{Milk}} = \frac{0.4}{\frac{4}{5}} = \frac{0.4}{0.8} = \frac{1}{2} = 0.5$$

Association rule mining parameters

3. Lift

Lift defines that strength of the relationship or association of purchasing milk and cheese

$$Lift_{Milk->Cheese} = \frac{Support_{(X \bigcup Y)}}{Support_X \times Support_Y} = \frac{0.4}{0.8 \times 0.4} = 1.25$$

Association rule mining algorithms

1. Apriori algorithm

- The Apriori Algorithm searches through small datasets to identify different rules(associations). In this algorithm, there are product clusters that pass frequently, and then strong relationships between these products and other products are sought.

- It's designed to efficiently discover frequent itemsets and association rules in small datasets.

Association rule: A rule of the form $X \rightarrow Y$, where X and Y are itemsets. X is called the antecedent, and Y is called the consequent.

Support: The fraction of transactions that contain both X and Y.

Confidence: The probability that Y will occur given that X has occurred.

Association rule mining algorithms

2. FP-Growth algorithm

- The frequent pattern growth algorithm also uses unsupervised learning but is more efficient than apriori algorithm
- The algorithm uses a divide-and-conquer approach to generate frequent itemsets efficiently. The main idea of the FP-growth algorithm is to represent the dataset in a compact form called the frequent pattern tree
- Limited to categorical data only.

Association rule mining algorithms

3. ECLAT algorithm

- Equivalence Class Clustering and List Intersection offers a more efficient approach compared to Apriori, especially for dense datasets.

Key Concepts:

- **Equivalence class:** A group of items that always appear together in transactions.
- **Item prefix:** The initial part of an itemset.
- **Item list:** A list of transactions that contain a specific item.

Applications of association mining

Market basket analysis: Identifying products that are frequently purchased together.

Recommendation systems: Suggesting items to users based on their past purchases or preferences.

Web usage mining: Analyzing user behavior on a website to improve its design.

Medical diagnosis: Identifying patterns in patient data to assist in diagnosis.

In-class practical

Import the necessary libraries for association rule mining:

Libraries in python:

mlextend,

apyori,

Apriori

Use the datasets provided on Moodle

UGANDA CHRISTIAN UNIVERSITY
A Centre of Excellence in the Heart of Africa

Uganda Christian University
P.O. Box 4 Mukono, Uganda
Tel: 256-312-350800
https://ucu.ac.ug/ Email: info@ucu.ac.ug.
@ugandachristianuniversity @UCUniversity
@UgandaChristianUniversity

Department of Computing & Technology
FACULTY OF ENGINEERING, DESIGN AND TECHNOLOGY
Tel: +256 (0) 312 350 863 | WhatsApp: +256 (0) 708 114 300
@ucucomputeng @ucu_ComputEng
https://cse.ucu.ac.ug/ Email: dct-info@ucu.ac.ug