Learning Guide Unit 3

Site: University of the People
Course: CS 3440-01 Big Data - AY2025-T1
Book: Learning Guide Unit 3

Printed by: Mejbaul Mubin
Date: Thursday, 5 September 2024, 2:30 PM

Description

Learning Guide Unit 3

Table of contents

Overview

Introduction

Reading Assignment

Discussion Assignment

Learning Journal

Self-Quiz

Graded Quiz

Checklist

Overview

UNIT 3: Analytical Theories and Methods

Topics

- Clustering principles
- Data analysis and big data
- Importance of data analysis

Learning Objectives

By the end of this Unit, you will be able to:

1. Explain the principles behind clustering.
2. Describe the techniques used in data analysis.

Tasks

- Peer assess Unit 2 Written Assignment
- Read the Learning Guide and Reading Assignments
- Participate in the Discussion Assignment (post, comment, and rate in the Discussion Forum)
- Complete and submit the Written Assignment
- Make entries to the Learning Journal
- Take the Self-Quiz
- Take the Graded Quiz

Introduction

Clustering is the practice of essentially grouping data points into similar groups for comprehensive data analysis and reporting. It is one of the main tasks used in the process of statistical analysis, pattern recognition, data compression, and computer graphics. Clustering is used by data scientists to gain important insights into the data they are examining by conducting this review in groups (clusters). By definition, unsupervised learning is a type of machine learning that searches for patterns in a data set with no pre-existing labels and a minimum of human intervention. Clustering can also be used for anomaly detection to find data points that are not part of any cluster, or outliers" (nvida.com, 2021).

Along with the benefits of clustering comes improved data analysis outcomes and techniques. Having an organized data set, organized by categories or similar data points makes the data analysis process infinitely better to organize and process.

The following points throw light on why clustering is required in data mining:

- Scalability − Large databases need to ensure that the clustering algorithms are scalable.
- Ability to deal with different kinds of attributes − Different types of data such as interval-based, categorical and binary need to have different algorithms applied to that data.
- Discovery of clusters with attribute shape − Clustering of any arbitrary shapes should be easily identifiable by the clustering algorithm.
- High dimensionality − The clustering algorithm should handle both low and high-dimensional data.
- Ability to deal with noisy data −Some databases contain data that is noisy, missing, or erroneous, and the algorithms should be sensitive to such data.
- Interpretability − Results of the clustering should be interpretable, comprehensible, and usable by data scientists (tutorialspoint.com, 2022).

Clustering involves grouping the data points into several groups so that the data points within each group are more similar to those within other groups. In this unit, we focus on data clustering as it pertains to big data.

The importance of data analysis is shown by its detailed approach to recording, analyzing, and presenting data in ways that businesses can use to help interpret market trends and understand better the operations of their business. Once you have your data categorized and ready for analysis you can then gain the insights needed to run your business more effectively. Data analysis can also help you understand your competitor's business by analyzing their trends and seeing how you compare in the marketplace.

Some of the most important approaches or methods used in data analysis include – regression analysis, simple linear regression analysis, hypothesis analysis, null hypothesis, content analysis, discourse analysis, grounded theory, and cross-tabulation. Some of these methods are specific to either quantitative or qualitative data analysis, which a few methods having the ability to be used in either instance.

In this unit, you will explore the principles of data clustering, how it helps improve upon data analysis, and why data analysis is an important concept for any business.

References

nvida.com. (2021, July 2). *Cluster analysis.* https://www.nvidia.com/en-us/glossary/data-science/clustering/

tutorialspoint.com. (2022). Data mining - cluster analysis. https://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm

Reading Assignment

Read through the following to better understand data clustering techniques, big data analysis and the importance of data analysis in the big data realm.

Arena, F., & Pau, G. (2020). An overview of big data analysis. *Bulletin of Electrical Engineering and Informatics*, *9*(4), 1646-1653. https://www.beei.org/index.php/EEI/article/view/2359/1532 licensed under CC BY-SA

- This article provides the reader with an overview of data analysis and how it impacts big data. The reader will get feel for what big data analysis is and how it applies to the realm of big data.

Chen, Z. L. (2022). Research and application of clustering algorithm for text big data. *Computational Intelligence and Neuroscience, 2022*, 8 pages. https://doi.org/10.1155/2022/7042778 licensed under CC BY 4.0

- This article discussed data clustering algorithms for big data text, and how dealing with big data text is very challenging given current technological challenges. The reader will be introduced to big data cluster analysis and the different methods involved. It provides clarity on what clustering involves.

Grant, A. (2020, January 3). *What is data analysis and why is it important?* MUO. https://www.makeuseof.com/tag/what-is-data-analysis/

- This article discussed with the increase in data collection, the world has become more data-driven. The article describes why data analysis is important and the role it plays in data mining.

Kaminsky, A. (2015). Chapter 14 - Massively Parallel. *Big CPU, big data: Solving the world's toughest computational problems with parallel computing*. licensed under CC 3.0

- Read Chapter 14 - Massively Parallel. This chapter goes into the process of describing what a massively parallel program is and what makes it possible. This section of the book shows the reader one area of big data clustering and how it relates to the big data realm.

Khemka, T. (2021, December 15). *The importance of data analysis.* Business 360. https://b360nepal.com/the-importance-of-data-analysis/

- This article discussed how important data analysis is, and discussed four primary methods of data analysis.

Le, J. (2019, April 12). An introduction to big data: Clustering. Data notes. https://data-notes.co/an-introduction-to-big-data-clustering-1a911b83e590

- The blogger explains about clustering and its different models with examples.

Shaw, A. A (2020). *What is data analysis? Importance, types, process & methods.* Marketingtutor.net. https://www.marketingtutor.net/what-is-data-analysis/

- This article discussed how data analysis is used to analyze data and extract the most relevant information from that data.

Zerhari, B., Ait Lahcen, A., & Mouline, S. (2015). Big data clustering: Algorithms and challenges. https://www.researchgate.net/publication/276934256_Big_Data_Clustering_Algorithms_and_Challenges

- This article outlines algorithms and challenges faced with big data clustering.

Video Resources:

Computerphile. (2019, July 10). *Data analysis 0: Introduction to data analysis* [Video]. YouTube. https://www.youtube.com/watch?v=8GIbOJtUw8w&t=29s

- This video provides a basic introduction to what data analysis is.

Data science dojo. (2019, March 14). *Introduction to clustering* [Video]. YouTube. https://www.youtube.com/watch?v=4cxVDUybHrI

- This video will review the fundamental concepts of clustering and different types of clustering methods.

Quantra. (2021, February 24). *What is data analysis? | Why is it important? | How do you interpret and analyse data? | Quantra* [Video]. YouTube. https://www.youtube.com/watch?v=Lh6frjuGuZM

- This video provides an explanation of data analysis and its importance, and how to perform exploratory data analysis.

Discussion Assignment

- Given the importance of data analysis, discuss three possible data analysis techniques you can use. What does each technique bring to the analysis of data?
- Provide a discussion on the three different techniques selected, and include both the advantages and disadvantages of each.

Your Discussion should be a minimum of 200 words in length and not more than 300 words. Please include a word count. Following the APA standard, use references and in-text citations for the textbook and any other sources.

Use APA citations and references for the textbook and any other sources used; you should use at least 1 APA citation and reference, but you can use more if needed. Refer to the UoPeople APA Tutorials in the LRC for help with APA citations. You are required to post an initial response to the question/issue presented in the Forum and then respond to at least 3 of your classmates' initial posts. You should also respond to anyone who has responded to you. Don't forget to rate the postings of your classmates according to the Rating Guidelines. Review the Discussion Forum rating guidelines to see how your classmates will be rating your post.

After posting an appropriate, meaningful, and helpful response to your three classmates, you must rate their posts on a scale of 0 (unsatisfactory) to 10 (excellent).

10 (A) - Excellent, substantial, relevant, insightful, enriching, and stimulating contribution to the discussion. Also, uses external resources to support the position where required and/or applicable.
8 - 9 (B) - Good, quite substantial, and insightful, but missing minor details which would have otherwise characterized it as an excellent response.
6 - 7 (C) - Satisfactory insight and relevance, but required some more information and effort to have warranted a better rating.
4 - 5 (D) - Limited insight and relevance of the material; more effort and reflection needed to have warranted a satisfactory grading.
0 - 3 (F) - Unsatisfactory insight/relevance or failure to answer the question, reflecting a poor or limited understanding of the subject matter and/or the guidelines of the question.

The rating scores are anonymous; therefore, do NOT mention in your remarks the separate rating score you will give the peer. The instructor is the only person who knows which score matches the comment given to a peer. Some classmates may worry that some peers will not provide a fair rating, or be unable to provide accurate corrections for grammar or other errors. It is the instructor's responsibility to ensure fairness and accuracy.

Learning Journal

In this learning journal explain in detail three basic principles of data clustering,

The Learning Journal entry should be a minimum of 400 words and not more than 750 words. Use APA citations and references if you use ideas from the readings or other sources

The rubric detailing how you will be graded for this assignment can be found within the unit's assignment on the main course page.

Self-Quiz

The Self-Quiz gives you an opportunity to self-assess your knowledge of what you have learned so far.

The results of the Self-Quiz do not count towards your final grade. However,the quiz is an important part of the University's learning process and it is expected that you will take it to ensure understanding of the materials presented. Reviewing and analyzing your results will help you perform better on future Graded Quizzes and the Final Exam.

Please access the Self-Quiz on the main course homepage; it is listed inside the Unit.

Graded Quiz

The Graded Quiz will test your knowledge of all the materials learned thus far. The results of the quiz will count towards your final grade.

Please access the Graded Quiz on the main course homepage; it will be listed inside the Unit. After you click on it, the quiz's introduction will inform you of any time or attempt limits in place.

Good luck!

Checklist

- Peer assess Unit 2 Written Assignment
- Read the Learning Guide and Reading Assignments
- Participate in the Discussion Assignment (post, comment, and rate in the Discussion Forum)
- Make entries to the Learning Journal
- Take and submit the Self-Quiz
- Take and submit the Graded Quiz