2025

CS 4407
Data Mining & Machine Learning

LEARNING JOURNAL UNIT 2

SANA UR REHMAN

INSTRUCTOR: NIRMAL ADHIKARI

Introduction

This week focused on tools and technologies for data mining and machine learning, including database systems, big data platforms, and statistical packages. I explored relational and analytical databases, NoSQL solutions such as MongoDB and Cassandra, and large-scale processing frameworks like Hadoop. I also studied visualization, reporting, and statistical analysis tools such as R, MATLAB, and open-source alternatives. My activities included completing the programming assignment, writing a discussion post on Hadoop, and taking the unit self-quiz.

Difficulties Faced

The most challenging part was understanding how analytical databases differ from traditional relational databases in terms of architecture and optimization. While relational databases were familiar, grasping columnar storage and its effect on query performance required extra reading (Elmasri & Navathe, 2020). I also found it demanding to connect the various database types with APIs and statistical tools, which required reviewing lecture notes and external references to see how these components work together.

Activities Performed

I completed a programming assignment comparing a traditional database, an analytical database, and a NoSQL database while demonstrating how MySQL, R, and Hadoop integrate within an analytics system. The assignment required synthesizing information from multiple sources and applying it to a cohesive analysis. For the discussion post, I researched Hadoop's architecture, including HDFS and the MapReduce model, and explained how it supports big data analytics (White, 2023; Shvachko et al., 2010). I also took the self-quiz, which reinforced key concepts such as the benefits of NoSQL databases and the use of APIs like WEKA and Orange.

Feedback and Interactions

Classmates commented on my discussion post, noting that the explanation of HDFS clarified how Hadoop ensures fault tolerance. This feedback confirmed that my explanation was accessible and thorough, which boosted my confidence in presenting technical topics. The instructor highlighted the clear link I made between Hadoop's distributed architecture and its practical applications, encouraging me to maintain this level of detail in future assignments.

Reflections and Reactions

I felt motivated as I worked through the programming assignment, especially when connecting MySQL, R, and Hadoop into an end-to-end analytics workflow. Initially, the scale and variety of technologies seemed overwhelming, but organizing my research and writing step by step helped me manage the complexity. I was surprised by how seamlessly Hadoop can integrate with statistical packages like R to process large datasets, which changed my view of big data as something accessible rather than intimidating.

Skills and Knowledge Gained

I strengthened my ability to compare database architectures and gained a deeper understanding of distributed computing. I also improved my research and technical writing skills, particularly in citing academic sources (Dean & Ghemawat, 2008). These skills will be valuable for future projects in data mining and machine learning, especially when designing scalable analytics systems.

Conclusion

This week expanded my understanding of how different database types, statistical packages, and big data frameworks work together to support analytics. I can now explain how traditional and analytical databases differ, describe the role of NoSQL solutions, and discuss how

Hadoop enables big data processing. Learning to integrate these tools has enhanced both my technical knowledge and my confidence in applying data mining concepts to real-world scenarios.

REFERENCES

Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM, 51*(1), 107–113. https://doi.org/10.1145/1327452.1327492

Elmasri, R., & Navathe, S. B. (2020). *Fundamentals of database systems* (7th ed.). Pearson.

Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010). The Hadoop Distributed File System. *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, 1–10. https://doi.org/10.1109/MSST.2010.5496972

White, T. (2023). *Hadoop: The definitive guide* (5th ed.). O'Reilly Media.

Word count: 519