CS 4407-01 Data Mining and Machine Learning

Instructor: Professor Shabia Shabir

Name: Ryohei Hayashi

Learning Journal 1

**Summary of the Week**

This week, I studied data mining and machine learning. Specifically, I explored the differences between data mining and machine learning, the distinctions between supervised and unsupervised learning, various types of machine learning algorithms, and their applications. Additionally, I reviewed practical examples of data mining and machine learning, learned the basics of R programming, and implemented a simple program.

**Personal Reflections**

Through studying data mining and machine learning, I was able to reaffirm their practical significance. In particular, understanding the differences between supervised and unsupervised learning allowed me to see their potential applications in daily data analysis and business scenarios. Supervised learning, which involves training models on labeled data, is highly effective for tasks with clear objectives, such as classification and prediction. On the other hand, unsupervised learning, which identifies patterns or hidden structures in data, is better suited for exploratory analysis. I believe this knowledge will serve as a valuable foundation for selecting the appropriate methods in future data analysis tasks.

My first experience with R programming exercises posed some challenges, particularly in mastering basic operations and syntax. For example, understanding the precise structure of functions and managing data frames were initially confusing. However, repeated practice helped me become more comfortable with these processes. I found the ability to handle data visually to be highly engaging, as it allowed me to grasp not just numerical values but also the trends behind the data. I feel I am gradually developing a perspective that enables me to "utilize data effectively."

One particularly helpful resource was the explanation by James et al. (2013) regarding the fundamental differences between supervised and unsupervised learning. The text highlights how supervised learning is based on clearly labeled data, whereas unsupervised learning focuses on discovering patterns within data. This understanding was further solidified through hands-on exercises with R.

**Topics Studied in Depth**

The most impactful topic I studied this week was the comparison between supervised and unsupervised learning and their applications. In supervised learning, algorithms such as regression analysis, support vector machines (SVM), and random forests are widely used. For instance, predicting purchasing trends based on customer data effectively demonstrated the utility of these methods. Conversely, unsupervised learning, which employs algorithms like k-means clustering and principal component analysis (PCA), emphasized the importance of grouping data and reducing dimensions. This helped me understand the value of gaining new insights from previously unexplored data. Learning these differences with concrete examples provided a sense of how theoretical knowledge connects to real-world applications.

In terms of R programming, I experienced the full cycle from loading datasets to visualization. Notably, using the `summary()` function for basic statistical summaries and the `plot()` function for visualization proved to be highly practical and effective as the first steps in data analysis. Additionally, referring to Jackson (2002) for foundational knowledge on data mining helped reinforce my theoretical understanding of how tools like R can assist in data analysis.

Moreover, Edelstein (1998) offered valuable insights into the definitions and practical applications of data mining. This resource helped me concretely understand how the data analysis process is structured and utilized, enriching my grasp of the subject.

**Challenges and Goals for Next Week**

One challenge I aim to address moving forward is improving my ability to apply theoretical knowledge to real-world data. Next week, I plan to deepen my understanding of data storage and management and gain insights into data analysis tools. Specifically, I will study the characteristics of relational databases and NoSQL databases, the relationship between big data and machine learning, and the use of statistical packages such as SPSS and R. Through these studies, I aim to strengthen my understanding of foundational technologies in data mining and machine learning.

Word Count: 615

References

1. Edelstein, H. A. (1998). *Introduction to data mining and knowledge discovery.* Available from: http://www.twocrows.com/intro-dm.pdf
2. Jackson, J. (2002). *Data mining: A conceptual overview. Communications of the Association for Information Systems,* 8(2002), 267–296. Available from: https://aisel.aisnet.org/cais/vol8/iss1/19/
3. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R.* New York, NY: Springer.