2025

CS 4407
Data Mining & Machine Learning

LEARNING JOURNAL UNIT 4

SANA UR REHMAN

INSTRUCTOR: NIRMAL ADHIKARI

What I Did

This week, I studied several approaches to classification in data mining and machine learning, including Bayes Theorem, nearest neighbor methods, decision trees, neural networks, and logistic regression. A significant portion of my learning was hands-on: in my programming assignment, I implemented the k-Nearest Neighbors (kNN) algorithm in R to classify points into two groups. By experimenting with different values of *k*, I observed how classification outcomes could change, such as when a test point switched from Class A to Class B depending on whether *k=1* or *k=3*. Alongside this, I prepared a discussion post that explored data normalization and standardization, connecting these preprocessing techniques to classification accuracy and model reliability.

My Reactions

I found the practical assignment particularly engaging, as it made abstract concepts more concrete. Visualizing training data clusters and classifying test points gave me a clear sense of how proximity influences predictions. At the same time, I realized how sensitive classification can be to the choice of parameters, which deepened my appreciation for model tuning. The reading and coding complemented each other well, as I saw theory applied directly to problems.

Feedback and Interactions

While I did not receive direct peer feedback this week, the structured instructions in the assignment and resources from the course served as valuable guidance. The discussion board also helped me reflect on preprocessing choices. Writing about normalization and standardization clarified for me why scaling matters before applying classifiers, especially distance-based methods like kNN (Han, Kamber, & Pei, 2011).

Feelings and Attitudes

Initially, I felt somewhat overwhelmed by the variety of classifiers introduced in this unit. However, completing the coding task boosted my confidence. I now feel more motivated to dive deeper into algorithms beyond kNN, particularly logistic regression and neural networks, which I recognize as foundational in machine learning applications.

What I Learned

This week solidified several key ideas. I learned that classifiers are at the core of supervised learning, where labeled data guides predictions. I also reinforced the principle that preprocessing, such as normalization, directly impacts algorithm performance (James, Witten, Hastie, & Tibshirani, 2021). Most importantly, I discovered the role of *k* in balancing sensitivity and stability in kNN classification. Too small a *k* risks overfitting, while larger values may oversimplify the results.

Challenges and Reflections

One of the most challenging aspects was reconciling different classification approaches conceptually. It was difficult at first to see how methods like Bayes or neural networks could all serve the same purpose but in very different ways. I overcame this by comparing them in terms of data requirements and assumptions. This process surprised me and made me wonder about the trade-offs in selecting classifiers for real-world problems.

Skills and Application

I recognize that I am building both technical coding skills in R and conceptual knowledge about machine learning algorithms. I am realizing that I learn best when I can connect theory

with practice through visualizations and examples. Already, I see ways to apply these insights: for example, in analyzing survey or customer data, I can use classifiers while ensuring proper preprocessing to avoid biased outcomes.

Final Thought

One important thing I am thinking about is how preprocessing, model choice, and parameter tuning form an interconnected chain. A classifier alone is not enough, its performance depends greatly on how data is prepared and how decisions like choosing *k* are made.

References

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Elsevier.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer.

Wordcount: 556