2025

CS 4407
Data Mining & Machine Learning

LEARNING JOURNAL UNIT 5

SANA UR REHMAN

INSTRUCTOR: NIRMAL ADHIKARI

INTRODUCTION

This week's exploration of **decision tree algorithms** provided a deeper understanding of supervised learning and model interpretability in data mining. Through readings, self-quiz, discussion post, and programming assignment, I examined how entropy and information gain drive feature selection and how the ID3 and C4.5 algorithms generate effective classification models. Using R, I built and evaluated a decision tree on the *Ionosphere* dataset, applying theoretical concepts to practical implementation.

WHAT I DID AND HOW I DID IT

I began the week by reviewing key topics such as **entropy**, **information gain**, and the operational steps of **ID3** and **C4.5** algorithms. To reinforce the theory, I implemented a classification model using R's `rpart` library on the *Ionosphere Radar* dataset. The process included data loading, splitting into training and test subsets, generating the tree, and computing the model's accuracy.

The code generated a tree that split features like *V5* and *V27* based on threshold values, predicting whether radar returns were "good" or "bad." I used the `predict()` function to obtain test predictions and created a confusion matrix with the `table()` function to calculate accuracy. The resulting accuracy of **87.4%** validated that the model generalized well to unseen data. I also plotted the tree structure using `plot()` and `text()` functions, which visually demonstrated how recursive partitioning captures data relationships.

In the discussion post, I extended this learning by analyzing how **decision trees** apply to real-world context specifically in **medical diagnosis** and **customer churn prediction**. These

examples emphasized the interpretability advantage of decision trees, making them valuable in domains where transparent decisions matter.

Reactions and Feedback

Initially, I found entropy and information gain calculations conceptually dense. Understanding how each feature's split contributes to reduced uncertainty requires careful review. However, seeing the R implementation translate those ideas into a working model was rewarding. My peers commented that my discussion effectively explained interpretability in medical and business applications, reinforcing my understanding of how theory connects to real-world problem-solving.

Feelings and Attitudes

At first, I felt uncertain about whether I fully grasped the mathematical intuition behind entropy. As I progressed through the exercises, I became more confident in applying formulas and interpreting results. I appreciated how the **visual clarity** of decision trees aligns with human reasoning each branch represents a clear "if–then" rule, making the model easier to explain than many black-box algorithms. This improved my comfort level with statistical modeling and coding integration.

What I Learned

This unit taught me how **information gain** quantifies the usefulness of attributes during tree construction. I learned to compute entropy manually and then validate it programmatically, confirming my calculations. I also gained practical experience in **splitting data**, **training models**, and **evaluating classification accuracy** in R.

The readings reinforced that decision trees are **non-parametric models**, making no assumptions about data distribution—an advantage in complex datasets (Han, Pei, & Tong, 2022). Moreover, I realized that while decision trees are interpretable, they can easily be **overfit** without pruning, highlighting the importance of balancing model depth and generalization (Sarker, 2021).

Reflection and Application

What surprised me most was how easily decision trees can visualize data-driven decisions. The most challenging part was tracing how entropy decreases after each split. Once I saw the numeric values correspond to intuitive data separations, the logic became clearer. I recognize that I am gaining analytical skills to interpret model metrics and programming confidence in R.

This week helped me realize that I learn best by **combining conceptual study with active experimentation**. Implementing algorithms myself deepens retention and builds intuition. I plan to apply these techniques in future projects involving classifications such as predicting student performance or detecting anomalies in sensor data by using decision trees as baseline models.

The most important takeaway is that **interpretability is power**. Decision trees remind me that transparency in model logic is as critical as accuracy, especially when decisions affect real-world outcomes.

REFERENCES

Han, J., Pei, J., & Tong, H. (2022). *Data mining: Concepts and techniques* (4th ed.). Morgan Kaufmann.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science, 2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Word count: *639*