CS 4407-01 Data Mining and Machine Learning

Instructor: Professor Shabia Shabir

Name: Ryohei Hayashi

Learning Journal 5

**Overview of the Week**

This week, we focused on the basics of decision trees, including their structure and operation, the calculation of entropy and information gain, and the detailed mechanics of the ID3 and ID4.5 algorithms. Additionally, we used R programming to create decision trees and build classification models based on datasets. Through practical programming assignments, we developed hands-on skills and gained a deeper understanding of these concepts. In the discussion assignment, we researched the practical applications of decision trees and examined two specific use cases for their use in solving classification problems.

**Personal Reflections**

Learning about decision trees introduced many new terms and concepts, which initially felt overwhelming, but overall, the content was highly engaging. Calculating entropy and information gain manually helped me intuitively grasp the structure of the data and the mechanisms behind decision tree splits. Although the calculations were complex, I gained a concrete understanding of "why a specific node was selected" and "how information gain optimizes the decision tree."

In the programming assignment, using R's `rpart` package to create decision trees provided immediate feedback through visual outputs, which enhanced my motivation to learn. At the same time, working with a complex dataset prompted me to reflect on the trade-offs between computation speed and accuracy, an important consideration for applying algorithms to real-world problems. These experiences highlighted challenges I aim to explore further in the future.

**Topics Studied in Depth**

The most impactful learning this week was understanding the process of optimizing data splits through entropy and information gain calculations. For example, entropy quantifies "uncertainty in data," while information gain measures "how much uncertainty is reduced." Using this understanding, I practiced calculating entropy for various attributes in a dataset and selecting the attribute with the highest information gain.

In the programming assignment, we worked with the Ionosphere dataset collected in Goose Bay. The dataset, consisting of 34 continuous attributes and one categorical attribute, posed a binary classification problem to differentiate between "good" and "bad" radar signals. I began by splitting the dataset into training and testing subsets and created a model using the `rpart` method. Then, using the `predict` method, I calculated predicted values and evaluated the model's accuracy. This practical exercise provided a comprehensive understanding of the end-to-end process of constructing and evaluating decision tree models on real-world data.

Additionally, the discussion assignment allowed me to investigate practical use cases for decision trees. For example, examining how decision trees are used in medical diagnosis and product recommendation systems helped deepen my understanding of their practical applicability.

**Future Challenges and Goals for Next Week**

Next week, the focus will shift to artificial neurons and the perceptron algorithm. I aim to deepen my understanding of the mathematical and logical skills foundational to artificial intelligence. Additionally, I plan to clarify the differences between data mining and machine learning to create a solid base for applying these concepts to practical and research contexts.

Reflecting on this week's work, I feel there were areas where my consideration of algorithmic efficiency and practical limitations could have been more thorough. I intend to address these aspects in the coming weeks to enhance my understanding of how to apply decision trees and other algorithms effectively.

Word Count: 531

References

1. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R.* New York, NY: Springer.