2025

CS 4407
Data Mining & Machine Learning

LEARNING JOURNAL UNIT 8

SANA UR REHMAN

INSTRUCTOR: NIRMAL ADHIKARI

UNIT 8 LEARNING JOURNAL: DISCOVERING PATTERNS IN DATA

WEEKLY ACTIVITIES AND INITIAL REACTIONS

This week marked a significant shift in perspective as we moved from supervised learning to unsupervised learning, specifically focusing on clustering. My learning process involved a deep dive into the provided reading materials to understand the fundamental difference: operating without predefined labels. I explored the mechanics of both K-means and hierarchical clustering. The major activity of my week was researching and constructing the discussion forum post, for which I detailed a comprehensive implementation plan for agglomerative hierarchical clustering. To consolidate this knowledge, I also completed the self-quiz and the review quiz for the final exam.

My initial reaction was a mix of curiosity and slight unease. Supervised learning feels comfortable because you have a "ground truth" to measure against. Unsupervised learning, as I quickly learned, is far more of an exploratory art. There is no single "correct" answer, only interpretations that are more or less useful. This ambiguity was my biggest initial hurdle.

CHALLENGES AND DISCOVERIES

The most challenging part of the week was moving from the simple theory of K-means (which seems intuitive) to the more complex, choice-driven process of hierarchical clustering. This became the focus of my discussion post. Researching for the post was the key learning activity. I had to articulate the precise step-by-step implementation, which forced me to grapple with the critical decisions a data scientist must make: choosing a distance metric (like Euclidean) and, more

importantly, a linkage criterion (like Ward or complete). I realized that these choices are not trivial and can completely change the resulting clusters.

What surprised me most was the central importance of the dendrogram. I had previously seen dendrograms and considered them a mere visualization, but my research revealed they are the *primary output* of the algorithm. The dendrogram isn't just a picture; it *is* the full, nested hierarchy of clusters. As noted by Espinoza (2011), this visualization is a powerful tool for interpreting complex data, such as in gene expression, allowing the analyst to visually determine a natural "cut point" for forming distinct groups. This was a "lightbulb" moment—the algorithm doesn't give you a final answer, it gives you a map, and it's your job to interpret it.

GAINED SKILLS AND PERSONAL REALIZATIONS

This week, I gained a foundational skill in exploratory data analysis. I am learning how to find inherent structures in data without being told what to look for. This requires a different set of muscles: critical thinking, justification, and a comfort with ambiguity. The quizzes served as my main feedback, confirming I had the terminology of linkage and dendrograms correct, but the discussion post was the true test of my understanding.

As a learner, this unit revealed that I value clear metrics and "right answers." Unsupervised learning is challenging my reliance on this. It's forcing me to become a better "detective," using clues from the data rather than a pre-defined answer key. I can immediately see how to apply this. The concept of clustering is the engine behind customer segmentation, anomaly detection (e.g., finding fraudulent transactions), and even organizing document libraries.

The most important thing I am thinking about now is the trade-off between K-means and hierarchical clustering. K-means is fast and scalable, but you must guess "K" beforehand. Hierarchical clustering is computationally expensive (O(n^2) or worse), but it gives you the complete picture and doesn't require pre-specifying the number of clusters. This highlights that there is no "best" algorithm, only the most appropriate one for the problem at hand, which is a core concept in machine learning (Singh et al., 2007).

References

Espinoza, F. A. (2011). Using hierarchical clustering and dendrograms to interpret gene expression data. *PLoS Computational Biology, 7*(6), e1001116. https://stmc.unm.edu/research/pdf/espinozabullmathbiol11.pdf

Singh, Y., Bhatia, P. K., & Sangwan, O. (2007). A review of studies on machine learning techniques. *International Journal of Computer Science and Security, 1*(1), 70-84. https://www.researchgate.net/profile/Pradeep-Bhatia-2/publication/41845861_A_REVIEW_OF_STUDIES_ON_MACHINE_LEARNING_TECHNIQUES/links/55489c350cf26a7bf4dadba4/A-REVIEW-OF-STUDIES-ON-MACHINE-LEARNING-TECHNIQUES.pdf?origin=scientificContributions

Wordcount: 600