CS 4407-01 Data Mining and Machine Learning

Instructor: Professor Shabia Shabir

Name: Ryohei Hayashi

Learning Journal 8

**​Overview of the Week**

This week, I studied unsupervised learning, which is used when data is not labeled. Unsupervised learning has unique characteristics compared to supervised learning. The focus was on clustering methods, particularly the K-means clustering algorithm and hierarchical clustering. I also aimed to understand the role of dendrograms, which are tools used to visualize the results of hierarchical clustering.

Additionally, I explored how the choice of the number of clusters (K) in K-means clustering affects the results. By referring to the paper by Pham et al. (2005), I considered effective methods for determining the optimal number of clusters. Through these topics, I gained a deeper understanding of selecting and implementing appropriate analytical methods based on data characteristics.

**Personal Reflections**

Through this week’s learning, I realized the potential value of unsupervised learning. In particular, the process of discovering hidden patterns in unlabeled data was fascinating. While K-means clustering is a relatively simple algorithm, I found it challenging to determine the initial cluster centers and the optimal number of clusters (K), as these significantly affect the results. Pham et al. (2005) emphasized the importance of understanding data characteristics to make informed choices, which resonated with me as a critical takeaway.

On the other hand, hierarchical clustering stood out for its intuitive approach to grouping and dividing data, aided by dendrograms. These visualizations make it easier to grasp the appropriate levels for dividing clusters, which I found particularly useful. I could imagine its practical applications, such as analyzing customer segmentation in marketing or processing medical data where hierarchical structures are significant.

Additionally, I began to consider the challenges of evaluating the results of unsupervised learning. Since there is no ground truth, assessing the quality of the outcomes is not straightforward. This aspect raised my curiosity, and I felt that working with actual datasets would help deepen my insights.

**Topics Studied in Depth**

**Unsupervised Learning and Clustering**

Unsupervised learning focuses on uncovering patterns and structures in unlabeled datasets, with clustering being one of the most representative approaches. The K-means clustering algorithm assigns data points to clusters iteratively while optimizing the cluster centroids. This iterative process ensures each cluster becomes as homogeneous as possible.

I learned the strengths and limitations of K-means clustering in depth. Its simplicity and efficiency are notable advantages, but the algorithm heavily depends on the initial selection of cluster centroids, which can lead to unstable results. Regarding the choice of K, I studied methods like the elbow method and gap statistics to enhance accuracy.

**Hierarchical Clustering and Dendrograms**

Hierarchical clustering organizes data into hierarchical groups, and dendrograms serve as a tool for visualizing the clustering process. Dendrograms graphically represent how data points are grouped or divided, helping to determine the most suitable cluster divisions intuitively. For example, they are highly effective in fields like marketing, where analyzing hierarchical customer segmentation is crucial.

The analysis of dendrograms involves selecting a "cut-off point" to determine the number of clusters. This visualization makes it easier to comprehend the complexity of the data structure, which I found particularly advantageous. I also studied the flexibility of hierarchical clustering, noting that it can use either an agglomerative approach (merging smaller clusters) or a divisive approach (splitting larger clusters), depending on the data's nature.

Furthermore, I discussed the high computational cost of hierarchical clustering, especially when applied to large datasets. This highlighted the need for optimizations in similarity calculations and the development of faster algorithms to address efficiency challenges.

Word Count: 573

References

1. Sherrie W. (2019).­­ *Introduction to Machine Learning*.
2. Jonathan R.S. *Unsupervised Learning and Principal Components Analysis*
3. Pham, D. T., Dimov, S. S., & Nguyen, C. D. (2005). Selection of K in K-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 219(1), 103-119*. Available from http://www.ee.columbia.edu/~dpwe/papers/PhamDN05-kmeans.pdf