CHEME 5820: Machine Learning for Engineers **Spring 2025**

Lecture 1c: Unsupervised learning and k-means clustering

Lecturer: Jeffrey Varner

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

1 Introduction

This lecture introduces the first unsupervised learning approach we will explore: k-means clustering. The primary objective of clustering is to partition a dataset into distinct groups or clusters such that the data points within each cluster exhibit a higher degree of similarity to one another than to those in other clusters. The Lloyd-Forgy algorithm [1] published in 1982 (although the approach was developed much earlier and published in part by Forgy earlier) is a straightforward and widely employed approach for clustering. The algorithm is easy to understand and implement, and it often produces clusters that are useful in practice.

The k-means algorithm is an example of an `unsupervised learning algorithm`. Unsupervised learning focuses on discovering patterns and structures in data without the guidance of labeled examples or explicit feedback. Unlike supervised learning (which we will explore in future lectures), where algorithms are trained on labeled datasets, unsupervised learning algorithms operate with raw, unlabeled data to identify inherent groupings, anomalies, or relationships. This approach is particularly valuable when dealing with large volumes of unstructured data or when the desired outcomes may be unknown. Typical applications of unsupervised learning include clustering (which we are discussing today), dimensionality reduction, and anomaly detection. Unsupervised learning can provide valuable insights and facilitate data by uncovering hidden structures in data.

2 K-means Clustering

Fill me in

3 Summary and Conclusions

This lecture introduced unsupervised learning, K-means clustering, and its implementation through the Lloyd-Forgy algorithm. Unsupervised learning algorithms autonomously identify patterns and structures without predefined categories, making them valuable in customer segmentation, image processing, and anomaly detection applications. K-means clustering organizes data into distinct groups based on similarity. The Lloyd-Forgy algorithm enhances the efficiency and accuracy of this clustering process by iteratively refining the cluster centroids. As the demand for data-driven decision-making continues to grow, the capability of unsupervised learning techniques to uncover hidden relationships and optimize data interpretation becomes increasingly essential for both businesses and researchers.

References

[1] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.