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

Lecture 1c: Unsupervised learning and the k-means clustering algorithm

Lecturer: Jeffrey Varner

Introduction

In this lecture, we introduce one of the first unsupervised learning approaches we will explore: k-means clustering. The primary objective of clustering is to divide a dataset into distinct clusters, such that the data points within each cluster exhibit a higher degree of similarity than those in other clusters. The k-means algorithm is a straightforward and widely employed method for clustering. The algorithm is easy to understand and implement, and it often produces clusters that are useful in practice.

What is Unsupervised Learning?

The k-means algorithm is an example of an unsupervised learning algorithm. Unsupervised learning is a branch of machine learning that focuses on discovering patterns and structures in data without the guidance of labeled outputs 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.