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

Lecture 1c: Unsupervised learning and 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 approaches we will explore: k-means clustering, and self-organizing maps. We will use these algorithms to identify hidden patterns and structures in data without explicit guidance. The key concepts covered in this lecture include:

- **Unsupervised learning** is a type of machine learning that involves training algorithms on unlabeled data. The goal of unsupervised learning is to identify patterns and structures in data without explicit guidance. Unsupervised learning is particularly useful when dealing with large volumes of unstructured data or when the desired outcomes are unknown.
- **Clustering** is a common unsupervised learning technique that involves dividing a dataset into distinct groups, or clusters, based on the similarity of data points. Clustering algorithms aim to group data points that are more similar to each other than to those in other clusters.
- **K-means clustering** is a popular and straightforward clustering algorithm that partitions a dataset into k clusters. The algorithm iteratively assigns data points to the nearest cluster center and updates the cluster centers based on the mean of the assigned points.

2 K-means clustering

The K-means algorithm, originally developed by Lloyd in the 1950s but not published until much later in 1982 (1), is an example of an `unsupervised learning`. 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.

K-means is a popular unsupervised machine learning algorithm used for clustering data points into K distinct groups based on their similarity. In this approach, the algorithm partitions the dataset into K (specified by you) clusters, with each cluster represented by a centroid (the mean of the data points in the cluster). Then the algorithm iteratively assigns data points to the nearest cluster centroid and updates the centroids based on the mean of the assigned points. Puesdo code for the k-means algorithm is shown in Algorithm 1.

Algorithm 1 Unsupervised naive k-means clustering (Lloyd's algorithm)

Input: $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^m\}$, number of clusters K and initial centroids $\{\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^m\}$
Output: Cluster assignments $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$ and updated cluster centroids $\{\mu_1, \mu_2, \ldots, \mu_K\}$
flag $\leftarrow$ `false` ▷ flag to indicate convergence: `true` for convergence and `false` otherwise
while flag is `false` **do**
  for $\mathbf{x} \in \mathcal{D}$ **do** ▷ Iterate over all data points in $\mathcal{D}$
    $c_i \leftarrow \arg\min_j ||\mathbf{x} - \mu_j||^2$ ▷ Assign data point $\mathbf{x}$ to the closest cluster centroid (Euclidean distance)
  end for

  $\hat{\mu} \leftarrow \mu$ ▷ Store the current best cluster centroids
  for $j = 1$ to K **do** ▷ Iterate over all clusters
    $\mu_j \leftarrow \frac{1}{|c_j|} \cdot \sum_{\mathbf{x} \in c_j} \mathbf{x}$ ▷ Update cluster centroid μ_j where $|c_j|$ is the number of data points in cluster c_j
  end for

  if $||\mu - \hat{\mu}|| < \epsilon$ **then** ▷ Check for convergence: based on the change in cluster centroids
    flag $\leftarrow$ `true` ▷ Set flag to `true` to terminate the algorithm
  end if
end while

3 Estimating the number of clusters

The k-means algorithm is simple and intuitive, but it has some limitations. One of the main drawbacks of k-means is that it requires the number of clusters K to be specified in advance, which can be challenging when the number of clusters is unknown. There are several methods to estimate the number of clusters, including the elbow method, the silhouette method, or performance metrics such as the Davies-Bouldin index, the Dunn index or the Calinski-Harabasz index.

3.1 Calinski-Harabasz index

4 Summary and Conclusion

In this lecture, we introduced the concept of unsupervised learning and discussed two common unsupervised learning algorithms: k-means clustering and self-organizing maps. Unsupervised learning is a type of machine learning that involves training algorithms on *unlabeled data* to identify patterns and structures within data without explicit guidance. Clustering is a common unsupervised learning technique that involves dividing a dataset into distinct groups, or clusters, based on the similarity of data points. We explored k-means clustering, which partitions a dataset into k clusters and identify patterns that may not be immediately apparent.

References

1. Lloyd S. Least squares quantization in PCM. IEEE Transactions on Information Theory. 1982;28(2):129–137. doi:10.1109/TIT.1982.1056489.