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.
- **Self-organizing maps (SOMs)** are another type of unsupervised learning algorithm that uses a neural network to map high-dimensional data onto a lower-dimensional grid.

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: Data points $\mathcal{D} = \{x_1, x_2, \ldots, x_n \in \mathbb{R}^m\}$, number of clusters K
Output: Cluster assignments $C = \{c_1, c_2, \ldots, c_n\}$ and cluster centroids $\{\mu_1, \mu_2, \ldots, \mu_K\}$
Randomly initialize K cluster centroids $\{\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^m\}$
flag $\leftarrow$ `true` ▷ flag to indicate convergence
while flag **do**
    for $x_i \in \mathcal{D}$ **do**
        $c_i \leftarrow \arg\min_j ||x_i - \mu_j||^2$
    end for
    $\hat{\mu} \leftarrow \mu$
    for $j = 1$ to K **do**
        $\mu_j \leftarrow \frac{1}{|C_j|} \cdot \sum_{x_i \in C_j} x_i$
    end for
    if $||\mu - \hat{\mu}|| < \epsilon$ **then**
        flag $\leftarrow$ `false`
    end if
end while
return cluster assignments C, cluster centroids $\{\mu_1, \mu_2, \ldots, \mu_K\}$

3 Self-organizing maps (SOMs)

Self-organizing maps (SOMs), originally developed by Kohonen (2), are another type of unsupervised learning algorithm that uses a graph-like structure to map high-dimensional data onto a lower-dimensional grid. In the literature, you may also see these referred to as Kohonen maps or topographic maps, or described as a type of artificial neural network (although they are distinct, and much different in several important ways from traditional neural networks). SOMs can be used for clustering, visualization, and dimensionality reduction. They differ from traditional neural networks in that they use a `competitive learning` approach to map input data to a lower-dimensional grid.

A self-organizing map consists of a rectangular (or potentially hexagonal) grid of nodes organized in a two-dimensional lattice. Each node is associated with a weight vector $\mathbf{w}_j \in \mathbb{R}^n$ which has the same dimension as the input data $\mathbf{x} \in \mathbb{R}^n$. The training of SOMs, which determines the weight vector $\mathbf{w}$, involves two main phases: competition and cooperation.

- **Phase I: Competition**: For each input vector, the node with the weight vector most similar to the input (usually determined by Euclidean distance $|| \cdot ||_2$) is identified as the Best Matching Unit (BMU). This process encourages nodes to compete for representing specific regions of the input space.
- **Phase II: Cooperation**: Once the BMU is identified, the weights of the neighboring nodes are updated to become more like the input vector. This is done using a neighborhood function $h : \mathbb{R} \to \mathbb{R}$ that defines how much influence the BMU has on its neighbors based on their distance from it on the grid.

The weights of the BMU and its neighbors are adjusted at iteration t according to the following expression:

$$w_{ij}(t+1) = w_{ij}(t) + \alpha(t) h_{ij}(t)(x_i - w_{ij}(t)) \tag{1}$$

where $w_{ij}(t)$ is the weight of the node at position (i, j) at iteration t, $\alpha(t)$ is the learning rate at iteration t, $h_{ij}(t)$ is the neighborhood function at iteration t, and $\mathbf{x}_i$ is the input vector. The neighborhood function

$h_{ij}(t)$ is typically a Gaussian function that decreases with the distance from the BMU, i.e., something like:

$$h_{ij}(t) = \exp\left(-\frac{d_{ij}^2}{2\sigma^2(t)}\right) \tag{2}$$

where d_{ij} is the distance between the BMU and the node at position (i, j), and $\sigma(t)$ is the neighborhood radius at iteration t. The learning rate $\alpha(t)$ and the neighborhood radius $\sigma(t)$ are typically annealed over time to allow the network to converge to a stable state.

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 two clustering algorithms: k-means clustering, which partitions a dataset into k clusters, and self-organizing maps, which use a neural (like) network to map high-dimensional data onto a lower-dimensional grid. These algorithms can be used to uncover hidden structures in data, visualize complex datasets, 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.

2. Kohonen T. Self-organized formation of topologically correct feature maps. Biological Cybernetics. 1982;43(1):59–69. doi:10.1007/BF00337288.