

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.

Key concepts

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 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.

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. Pseudo code for the k-means algorithm is shown in Algorithm 1.

Algorithm 1 Unsupervised K-Means Clustering

```

1: Input: Data points  $\mathcal{D} = \{x_1, x_2, \dots, x_n \in \mathbb{R}^m\}$ , number of clusters  $K$ 
2: Output: Cluster assignments  $C = \{c_1, c_2, \dots, c_n\}$  and cluster centroids  $\{\mu_1, \mu_2, \dots, \mu_K\}$ 
3: Randomly initialize  $K$  cluster centroids  $\{\mu_1, \mu_2, \dots, \mu_K \in \mathbb{R}^m\}$ 
4: flag  $\leftarrow$  true ▷ flag to indicate convergence
5: while flag do
6:   for  $x_i \in \mathcal{D}$  do
7:      $c_i \leftarrow \arg \min_j \|x_i - \mu_j\|^2$ 
8:   end for
9:    $\hat{\mu} \leftarrow \mu$ 
10:  for  $j = 1$  to  $K$  do
11:     $\mu_j \leftarrow \frac{1}{|C_j|} \cdot \sum_{x_i \in C_j} x_i$ 
12:  end for
13:  if  $\|\mu - \hat{\mu}\| < \epsilon$  then
14:    flag  $\leftarrow$  false
15:  end if
16: end while
17: return cluster assignments  $C$ , cluster centroids  $\{\mu_1, \mu_2, \dots, \mu_K\}$ 

```

3 Self-organizing maps (SOMs)

Self-organizing maps (SOMs) are another type of unsupervised learning algorithm that uses a Graph structure to map high-dimensional data onto a lower-dimensional grid. SOMs are a type of artificial neural network that can be used for clustering, visualization, and dimensionality reduction. However, they differ from traditional neural networks in that they use a **competitive learning** approach to map input data to a lower-dimensional grid.

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