Identifying Cluster Structures in High-dimensional Data

Matthews Sejeso

School of Computer Science and Applied Mathematics,
University of the Witwatersrand, Johannesburg.
matthews.sejeso@wits.ac.za

SA Graduate Modelling Camp,
University of the Witwatersrand, Johannesburg.
15 - 18 January 2025.

Introduction

Data clustering involves grouping entities in a dataset into clusters based on their similarity.

Machine learning tasks are generally categorized into two types:

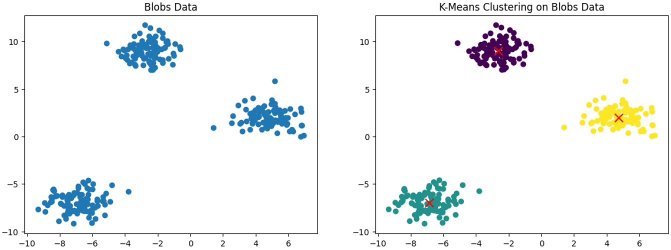
- Supervised learning, where data is paired with explicit labels.
- Unsupervised learning, where data does not have predefined labels.

Clustering is a key technique in unsupervised learning, used to uncover hidden patterns and structures in the data, especially in high-dimensional datasets.

It is a powerful tool for pattern recognition, offering valuable insights that might not be apparent from raw data alone.

Introduction

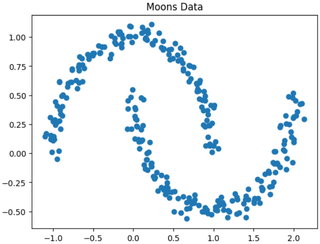
Clustering segments data into distinct groups based on similarity. Each cluster contains similar data points, while clusters are distinctly different.

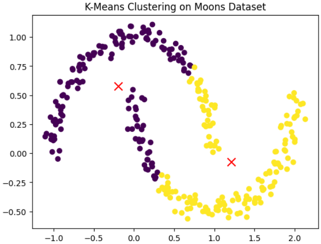


K-Means Clustering: Partition data into k clusters by minimizing the sum of squared distances within the cluster to the cluster's centre. Assume that clusters are spherical and separable in Euclidean space.

Introduction

Limitation of K-Means Clustering: Performs poorly with non-linear data.





High-Dimensional Data

High-dimensional data refers to datasets with many features compared to observations.

Characteristics of High-Dimensional Data:

1. Curse of Dimensionality - As the number of dimensions increases, the data points become sparse and distant from one another.

   - Many traditional machine learning algorithms fail due to the exponential increase in computational complexity and the reduced reliability of metrics like Euclidean distance.

2. Redundancy and Noise - High-dimensional datasets often contain redundant or irrelevant features that add noise, making it harder to identify meaningful patterns.
3. Manifold Hypothesis - Real-world high-dimensional data (e.g., images) often lie on or near low-dimensional manifolds within the high-dimensional space. For instance, images of an object taken under different conditions might vary along a lower-dimensional subspace.

Challenges of High-Dimensional Clustering

Applications of high-dimensional clustering include genomics (e.g., thousands of genes per sample), text processing (words as features), and image analysis (high-resolution pixel data) and more.

Key challenges in clustering high-dimensional data:

- The computational cost of processing high-dimensional data increases dramatically as dimensionality rises.
- Human intuition and conventional visualization techniques are limited to 2D or 3D spaces.
- As dimensionality increases, models become more complex and harder to interpret.

Understanding these challenges and their solutions is critical to ensure accurate and efficient analysis when working with high-dimensional data.

Subspace Clustering

Subspace clustering involves identifying and grouping data points that lie in distinct low-dimensional subspaces within a high-dimensional space.

Why It's Important:

- High-dimensional data often resides on lower-dimensional structures (manifolds or subspaces) rather than being uniformly distributed.
- Traditional clustering methods fail because they assume that the data form homogeneous clusters throughout the space.

Subspace clustering techniques:

- Algebraic: Assumes noise-free data; limited robustness.
- Iterative: Alternates between clustering and subspace estimation.
- Statistical: Uses generative models for structured data.
- Spectral: Constructs a similarity graph and leverages eigenstructure.

WITS UNIVERSITY

Spectral Clustering

Spectral clustering use graph theory and linear algebra to group data points. It then utilizes the eigenstructure of a similarity graph to identify clusters in high-dimensional or non-linear data spaces.

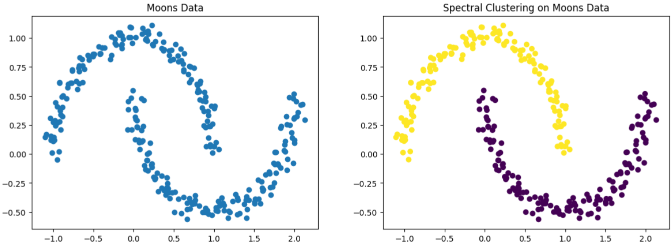
The steps of spectral clustering:

1. Construct a Similarity Graph -
   - Represent data points as a graph $G = (V, E)$, where: V is the set of nodes (data points). E is the set of edges (pairwise similarities).
   - Weight the edges using a similarity measure $w_{ij} = exp\left(\frac{\|x_i - x_j\|_2^2}{\sigma^2}\right)$ where σ controls the width of the neighborhood.
2. Construct the Graph Laplacian:
   - Unnormalized: $L = D - W$; or
   - Normalized: $L_{sym} = I - D^{-1/2} W D^{-1/2}$, where $D_{ii} = \sum_j^n w_{ij}$ and W Adjacency matrix which contains w_{ij}

WITS UNIVERSITY

Spectral Clustering

3. Compute Eigenvectors:
   - Find the first k eigenvectors (smallest eigenvalues) of the Laplacian.
   - Form a feature matrix U where columns correspond to the eigenvectors: $U = [u_1, u_2, \ldots, u_k]$
4. Apply K-Means Clustering:
   - Treat rows of U as data points and use K-Means to cluster these points.



Spectral Clustering

Challenges of spectral clustering:

- Choosing appropriate σ or the number of nearest neighbors.
- Avoiding noisy or spurious connections in high-dimensional data.
- Eigenvalue Decomposition is computationally expensive for large datasets.
- Requires experimentation and domain knowledge for optimal results.

Goal for the modeling camp

Explore robust approaches to spectral clustering in high-dimensional settings.

The core problem summarized:

Face clustering: given face images of multiple subjects, the goal is to find images that belong to the same subject.

References

1. R. Xu and D. C. Wunsch II, "Survey of clustering algorithms," IEEE Trans. Neural Networks, vol. 16, no. 3, pp. 645–678, 2005.
2. K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbour" meaningful?" in Int. Conf. Database Theory. Springer, 1999, pp. 217–235.
3. L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: a review," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 90–105, 2004.
4. V. S. Cherkassky and F. Mulier, Learning from Data: Concepts, Theory, and Methods, 1st ed. New York, NY, USA: John Wiley & Sons, Inc., 1998.
5. R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 25, no. 2, pp. 218–233, 2003.
6. I. Jolliffe, Principal component analysis (Springer Series in Statistics). Berlin, Germany: Springer, 2002.
7. R. Vidal, "Subspace clustering," IEEE Signal Processing Magazine, vol. 28, no. 2, pp. 52–68, 2011.

WITS UNIVERSITY