Midterm Project: Clustering and Dimensionality Reduction Analysis

Advanced Machine Learning
Astana IT University (AITU)

Angsar Shaumen

January 2026

Contents

1 Introduction

Clustering and dimensionality reduction are two pillars of unsupervised machine learning. Clustering algorithms, such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [1], allow us to uncover hidden structures in unlabeled data by grouping similar instances based on density rather than simple distance to a centroid. Dimensionality reduction techniques, such as Principal Component Analysis (PCA) [3] and t-Distributed Stochastic Neighbor Embedding (t-SNE) [2], enable the visualization of high-dimensional data by projecting it into a lower-dimensional space.

The objective of this study is twofold:

1. **Clustering:** To implement and evaluate the DBSCAN algorithm on the "Mall Customers" dataset, exploring the impact of its hyperparameters (*eps* and *min_samples*) and comparing it against the traditional K-Means algorithm.

2. **Dimensionality Reduction:** To apply and compare PCA (linear) and t-SNE (non-linear) techniques on the MNIST written digits dataset to understand their efficacy in preserving global versus local data structures.

2 Theoretical Background

2.1 DBSCAN

DBSCAN groups points that are closely packed together. It distinguishes between three types of points:

- **Core Points:** Have at least *min_samples* points within distance *eps*.
- **Border Points:** Reachable from a core point but have fewer than *min_samples* neighbors.
- **Noise Points:** Not reachable from any core point.

Unlike K-Means, DBSCAN does not require specifying the number of clusters *a priori*, can find arbitrarily shaped clusters, and is robust to outliers [1].

2.2 PCA vs. t-SNE

- **PCA:** A linear transformation that projects data onto orthogonal axes maximizing variance. It preserves global structure.
- **t-SNE:** A non-linear probabilistic technique that minimizes the Kullback-Leibler divergence between high-dimensional and low-dimensional distributions. It excels at preserving local neighborhoods [2].

3 Methodology

3.1 Data Description

Mall Customers Dataset: Used for clustering. Contains features: Customer ID, Gender, Age, Annual Income (k$), Spending Score (1-100). We selected *Annual Income* and *Spending Score* for analysis.

MNIST Dataset: Used for dimensionality reduction [4]. Contains 70,000 grayscale images of handwritten digits. We used a subset of 5,000 samples for computational efficiency.

3.2 Implementation

All algorithms were implemented in Python using `scikit-learn` [5].

- **Clustering:** We performed a parameter sweep for DBSCAN ($eps \in [0.1, 1.0]$, $min\_samples \in \{3, 5, 10, 20\}$).
- **Visualization:** Results were scaled using `StandardScaler` before processing.

4 Results and Discussion

4.1 Clustering Analysis (Mall Customers)

The optimal configuration for DBSCAN was found to be **eps=0.35** and **min_samples=3**, maximizing the Silhouette Score.

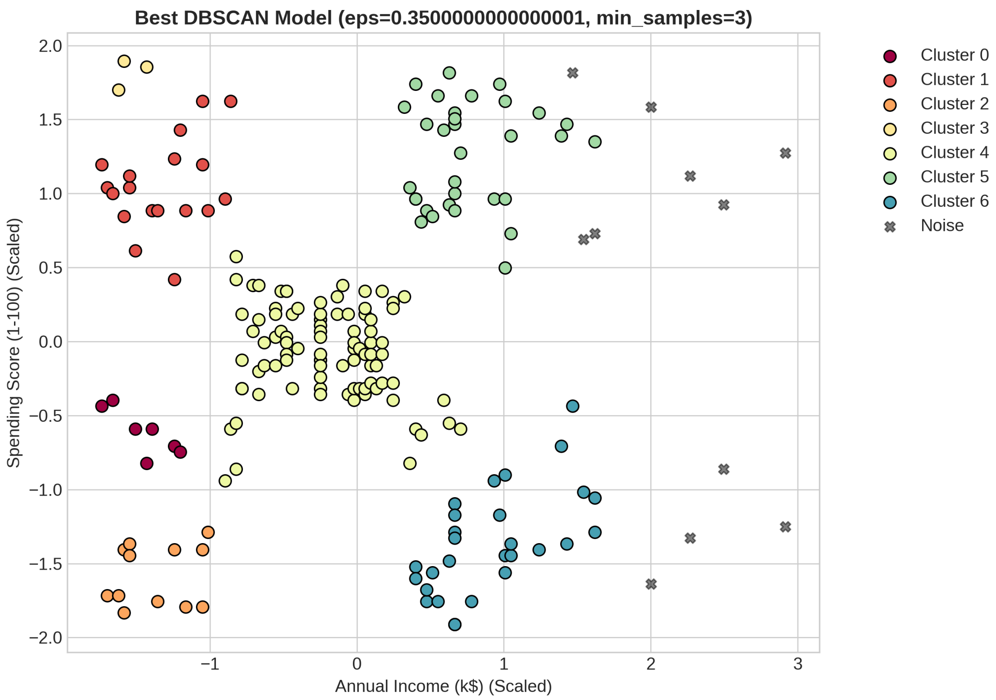

Figure 1: Best DBSCAN Clustering Result. The algorithm successfully isolates distinct spending behaviors and identifies outliers (black crosses).

4.1.1 Sensitivity Analysis

Increasing *eps* rapidly reduces the number of clusters. As shown in Figure 2, the stable region is between 0.3 and 0.5.

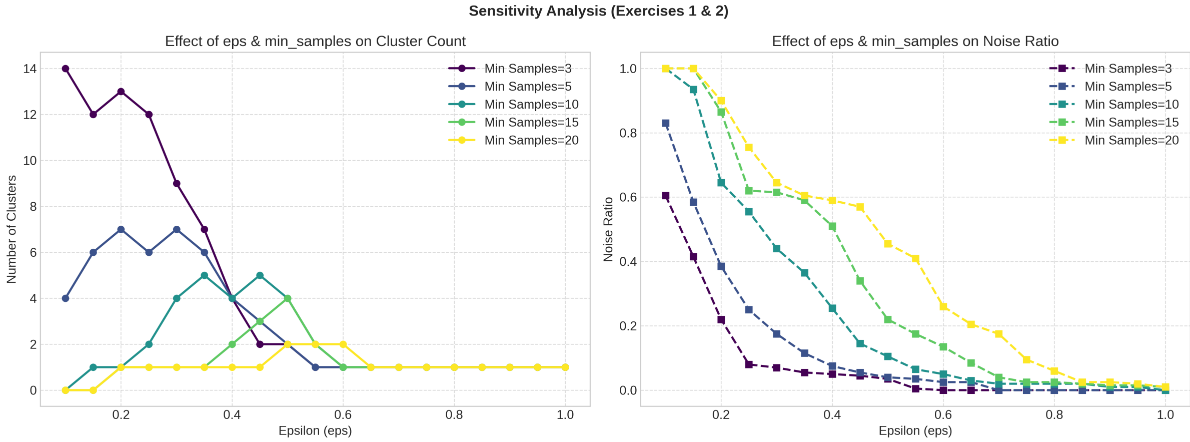

Figure 2: Sensitivity Analysis: Effect of *eps* and *min_samples* on Clusters and Noise. Note how higher *min_samples* (lighter colors) requires higher *eps* to form clusters.

4.1.2 Comparison with K-Means

K-Means forces all points, including outliers, into spherical clusters (Figure 3). DBSCAN provides cleaner segmentation by explicitly marking outliers as noise.

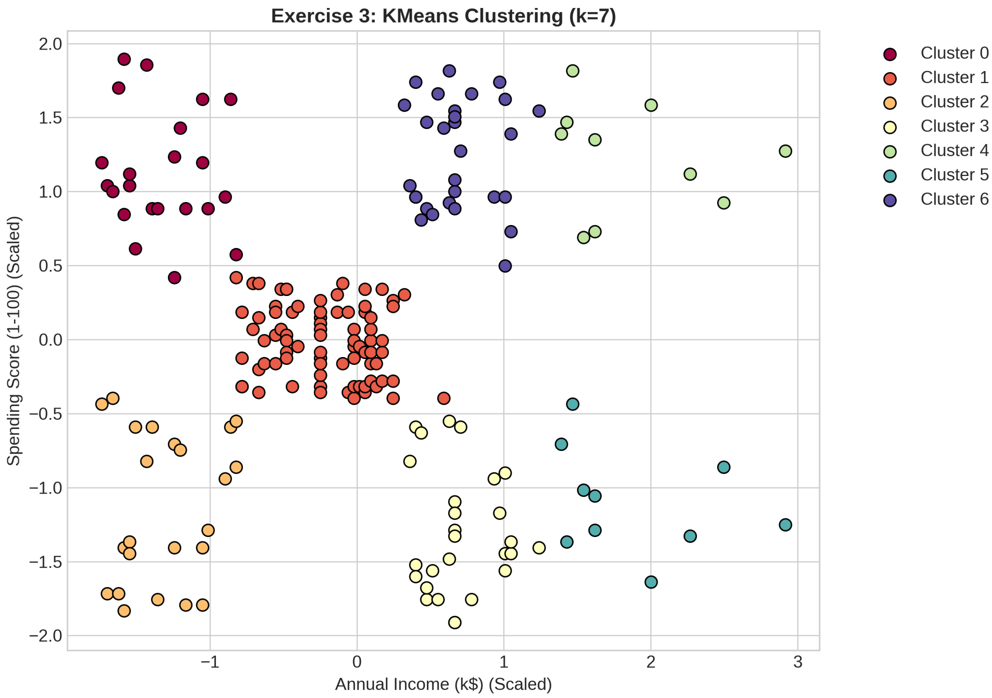

Figure 3: K-Means Clustering Comparison. Note how outliers are forced into clusters.

4.2 Dimensionality Reduction (MNIST)

4.2.1 PCA (Global Structure)

PCA captures global variance but overlaps complex digits (Figure 4). It fails to clearly separate the non-linear manifold of the digits.

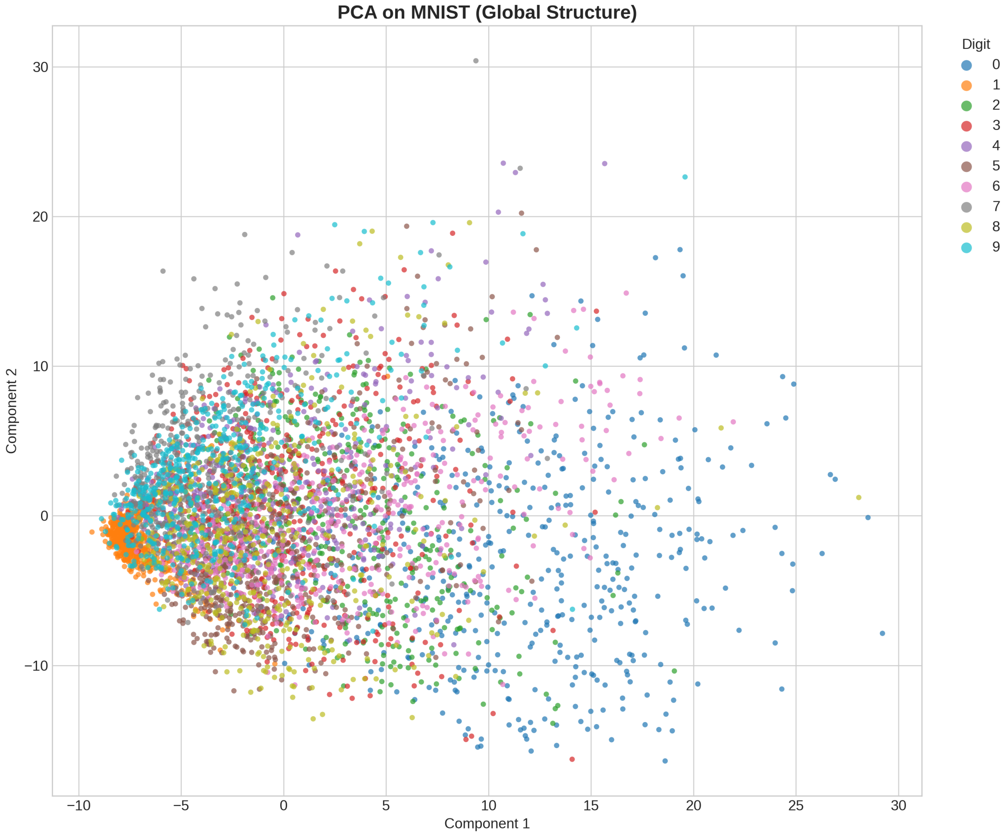

Figure 4: PCA Projection of MNIST. Digits overlap significantly.

4.2.2 t-SNE (Local Structure)

t-SNE successfully unrolls the manifold, creating distinct, well-separated islands for each digit (Figure 5).

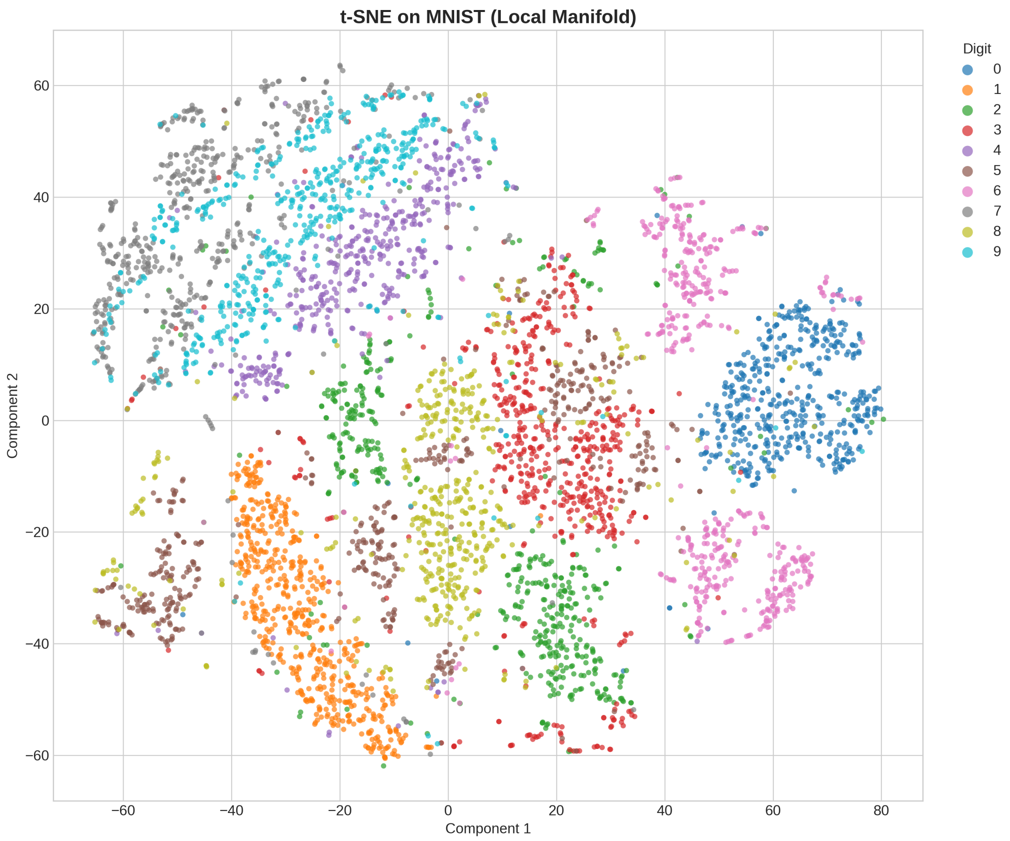

Figure 5: t-SNE Projection of MNIST. Clear separation of digit clusters.

5 Conclusion

This study demonstrated that DBSCAN is superior to K-Means for datasets with noise and non-spherical clusters, provided parameters are carefully tuned. For high-dimensional data like MNIST, t-SNE serves as a much more powerful visualization tool than PCA, effectively revealing class separability.

References

[1] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *KDD-96 Proceedings*, 226–231.

[2] Van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(11).

[3] Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer Series in Statistics.

[4] LeCun, Y., Cortes, C., & Burges, C. J. (1998). The MNIST Database of Handwritten Digits. Available at `http://yann.lecun.com/exdb/mnist/`

[5] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

A Control Questions Answers

A.1 DBSCAN

1. **What are eps and min_samples?** *eps* is the neighborhood radius; *min_samples* is the minimum neighbors required to form a dense region.
2. **Why can DBSCAN detect arbitrary shapes?** It uses density connectivity (chaining) rather than centroid distance.
3. **What does label -1 mean?** It represents Noise (Outliers).
4. **When is DBSCAN not suitable?** High-dimensional data (curse of dimensionality) or varying density clusters.
5. **How to select eps?** Using the k-distance graph (elbow method).

A.2 t-SNE

6. **Main objective?** Visualize high-dimensional data in low dimensions while preserving local structure.
7. **Similarity modeling?** High-dim: Gaussian; Low-dim: Student's t-distribution.
8. **Role of KL divergence?** It is the cost function to be minimized.
9. **Effect of perplexity?** Balances local vs. global attention (effective number of neighbors).
10. **Why different results?** Non-convex cost function and random initialization.

A.3 PCA

11. **Mathematical goal?** Maximize variance along orthogonal components.
12. **Difference from t-SNE?** PCA preserves global structure (linear); t-SNE preserves local (non-linear).
13. **Why faster?** Deterministic linear algebra vs. iterative optimization.
14. **When preferable?** Fore preprocessing, noise reduction, or needing global geometry.

A.4 Comparison

15. **Preservation?** Local: t-SNE; Global: PCA.
16. **Clearer for MNIST?** t-SNE, because digits lie on a non-linear manifold.