Customer Segmentation using Clustering: An Approach using Autoencoders

Md Sadique, K Mohan Madhav, Shivang Gupta
2022A7PS0156P, 2022A2PS1657P, 2022A1PS0643P

CS F415 - DATA MINING

Problem Statement

Abstract

Customer segmentation is an essential step for companies looking to personalize marketing and improve customer relationship management. In this project, we introduce an innovative method that involves using autoencoders to perform non-linear dimensionality reduction before clustering. We use K-Means, Gaussian Mixture Models (GMM), and DBSCAN on the reduced representations to reveal insightful customer segments. This approach utilizes the UK Online Retail Dataset and aims to enhance cluster quality and interpretability by identifying intricate patterns in customer behavior.

Dataset

UK Online Retail Dataset (from UCI Machine Learning Repository): 541,909 customer transactions with features used in RFM (Recency, Frequency, Monetary) analysis.

Dataset Features

- **InvoiceNo**: Unique ID for each transaction. Transactions with a 'C' prefix indicate cancellations.
- **StockCode**: Unique product code.
- **Description**: Name of the product.
- **Quantity**: Number of units per transaction.
- **InvoiceDate**: Timestamp of the transaction.
- **UnitPrice**: Price per unit in sterling.

- **CustomerID**: Unique identifier for each customer.
- **Country**: Country where the customer resides.

Implementation Roadmap

1. Data Preprocessing

- **Feature Engineering**: Computed RFM scores to summarize customer purchasing behavior.
- **Data Cleaning**: Removed incomplete records and handled transaction cancellations.
- **Feature Scaling**: Standardized RFM features to ensure balanced input to the autoencoder and clustering algorithms.

2. Non-Linear Dimensionality Reduction with Autoencoders

Classic linear methods such as PCA are unable to embed intricate, non-linear patterns in customer data. To overcome this, autoencoders—based on deep learning—were utilized to learn compressed, informative representations even under the presence of high dimensions and noise. The model was trained with mean squared error loss for reconstruction, and early stopping was used to avoid overfitting. The resulting encoded representations were next utilized as input for the clustering algorithms.

Encoder (non-linear transformation):

$$\mathbf{z} = f_\phi(\mathbf{x}) = \mathrm{ReLU}(\mathbf{W}_e\mathbf{x} + \mathbf{b}_e)$$

Decoder (Reconstruction):

$$\hat{\mathbf{x}} = g_\theta(\mathbf{z}) = \sigma(\mathbf{W}_d\mathbf{z} + \mathbf{b}_d)$$

Where-

$$\mathbf{x} \in \mathbb{R}^d \quad \text{(input)}, \quad \mathbf{z} \in \mathbb{R}^k, \quad k \ll d$$

Loss Function (Reconstruction Error):

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \|x_i - \hat{x}_i\|^2$$

3. Clustering Techniques

- **K-Means**: Applied to the encoded features to partition customers into distinct, non-overlapping groups based on purchasing patterns. The optimal number of clusters was determined using the elbow method and silhouette score.

  Objective Function:

  $$\arg\min_{\{C_k\}} \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

  Centroid Update:

  $$\mu_k^{(t+1)} = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i$$

- **Gaussian Mixture Model (GMM)**: Utilized for probabilistic clustering, capturing overlapping customer segments and complex cluster shapes in the latent space. **Density Function:**

  $$p(x) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

  E-Step:

  $$\gamma(z_{nk}) = \frac{\alpha_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \alpha_j \mathcal{N}(x_n|\mu_j, \Sigma_j)}$$

  M-Step:

  $$\mu_k = \frac{\sum_n \gamma(z_{nk}) x_n}{N_k}, \quad \Sigma_k = \frac{\sum_n \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T}{N_k}, \quad N_k = \sum_n \gamma(z_{nk})$$

- **DBSCAN**: Employed to identify dense regions and outliers, particularly effective for discovering irregular clusters and noise in compressed feature space.

Epsilon Neighborhood:

$$N_\varepsilon(p) = \{q \in D \mid \text{dist}(p, q) \leq \varepsilon\}$$

Core Point Condition:

$$|N_\varepsilon(p)| \geq \text{MinPts}$$

4. Performance Evaluation Metrics

- **Silhouette Score**: Evaluated the quality of clusters; higher scores indicate better-defined groupings.

- **Cluster Visualization**: Used to qualitatively assess the separation and structure of customer segments in the reduced feature space.

Results

Before AutoEncoding

- **K-Means** - Silhouette Score: 0.64

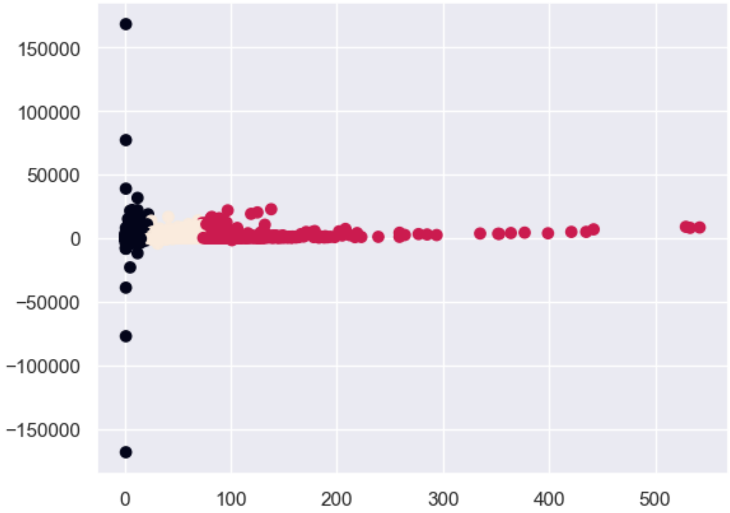

Figure 1: Clusters Formed by KMeans Before Autoencoding

- **GMM** - Silhouette Score: 0.76

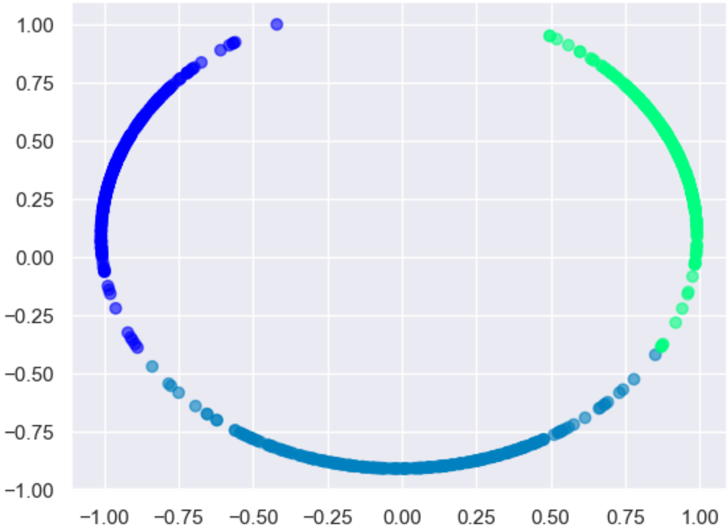

Figure 2: Clusters Formed by GMM Before Autoencoding

- **DBSCAN** - Silhouette Score: 0.626

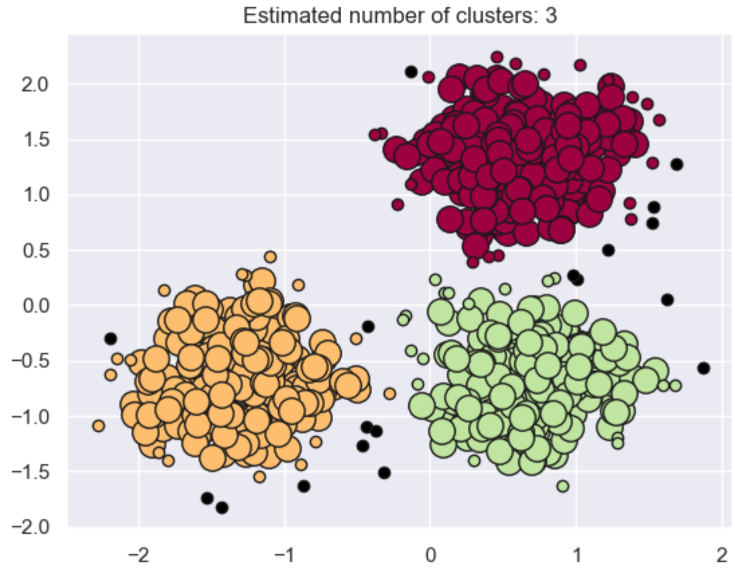

Figure 3: Clusters Formed by DBSCAN Before Autoencoding

After Encoding

- **K-Means** - Silhouette Score: 0.65

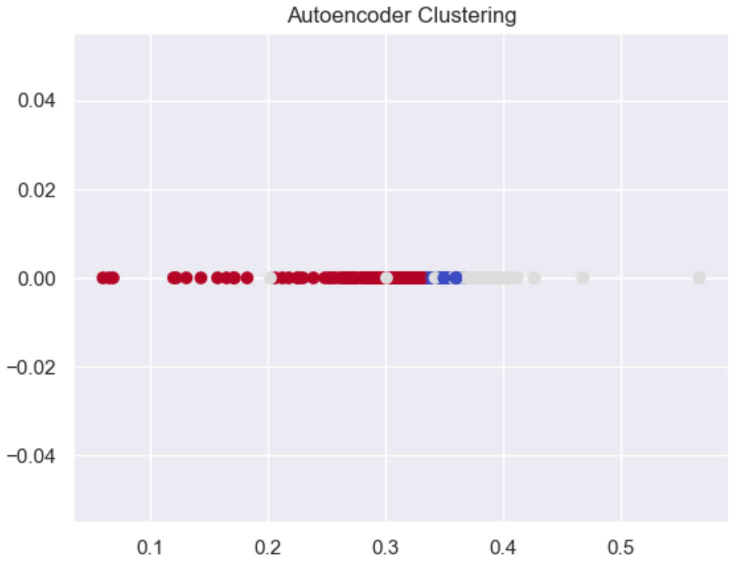

Figure 4: Clusters Formed by KMeans After Autoencoding

- **GMM** - Silhouette Score: 0.64
- **DBSCAN** - Silhouette Score: 0.79

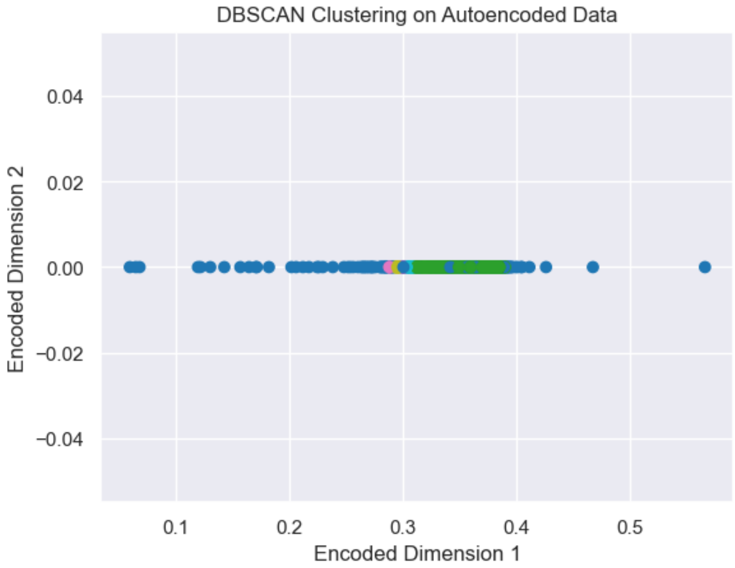

Figure 5: Clusters Formed by DBSCAN After Autoencoding

Summary

Algorithm	Before Encoding	After Encoding
K-Means	0.64	0.65
GMM	0.76	0.64
DBSCAN	0.62	0.79

References

- John, Jeen & Shobayo, Olamilekan & Ogunleye, Bayode. (2023). An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market. *Analytics*, 2, 809–823. `https://doi.org/10.3390/analytics2040042`

- D. Chen. ”Online Retail,” UCI Machine Learning Repository, 2015. `https://doi.org/10.24432/C5BW33`