Assignment 4 Report on
Wholesale Customers Dataset

Maitreyee Das Urmi (501218269)

Toronto Metropolitan University

CPS803 - Machine Learning

Dr. Elodie Lugez

Nov 26, 2024

The wholesale customer dataset reflects customers' annual spending on different product categories. This dataset appealed to me as interesting because it deals with customer segmentation through clustering an unlabeled dataset. I was curious to see how effectively it could divide all the instances into separate clusters based on numerical spending habits. The dataset contains eight features in total (Channel, Region, Fresh, Milk, Grocery, Frozen, Detergents_Paper, and Delicassen). Each feature, except for Channel and Region, refers to a product category, and the values represent annual spending in monetary units. A machine learning model trained on this dataset would reflect a real-world example of customer segmentation for market analysis, effectively distinguishing groups like high spenders, bulk buyers, etc.

For my project, I trained a machine learning model using the KMeans clustering algorithm. I excluded the first two features (Channel and Region) since they are categorical and not product categories, making them irrelevant for this clustering process. Using Scikit-learn's StandardScaler() method, I scaled my dataset before applying it to train the model because clustering algorithms rely heavily on distances between data points. The KMeans algorithm uses Euclidean distance to assign data points to clusters. If features have different scales, the distances will be biased toward features with larger magnitudes. Standardization transforms data to have a mean of 0 and a standard deviation of 1, ensuring equal contribution from all features.

Standardization Formula:

$$\mathbf{X}\text{scaled} = (\mathbf{X} - \mu) / \sigma$$

Where**,**

- **X**scaled is the standardized value
- **X** is the original value
- **μ** is the mean of the feature
- **σ** is the standard deviation of the feature

Next, I used a for loop to iterate through values in the range of 2 to 10 (excluding 10) to compute and plot a graph with silhouette scores against the number of clusters. Silhouette scores help identify the optimal value of k (the number of clusters) for the KMeans algorithm. Looking at Fig. 1, the silhouette score is highest (>0.45) when k=3. Therefore, I trained the KMeans algorithm with n_clusters=3, as it had the highest silhouette score from the range, indicating well-separated clusters compared to other values of k.

To visualize the clusters in 2D, I used Principal Component Analysis (PCA) to reduce the dataset to two dimensions (principal components). PCA helps retain the maximum variance in the data while simplifying visualization. I then plotted the clusters and their centroids, with the first principal component on the x-axis and the second principal component on the y-axis. Fig. 2 shows the final result, displaying three distinct clusters of customers.

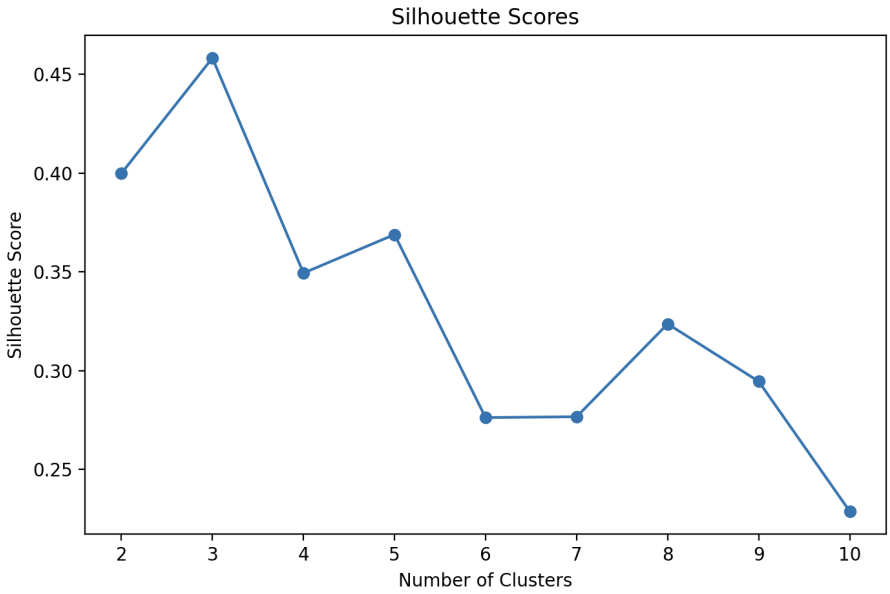

Fig 1: Illustrates the relationship between the number of clusters and their corresponding silhouette scores.

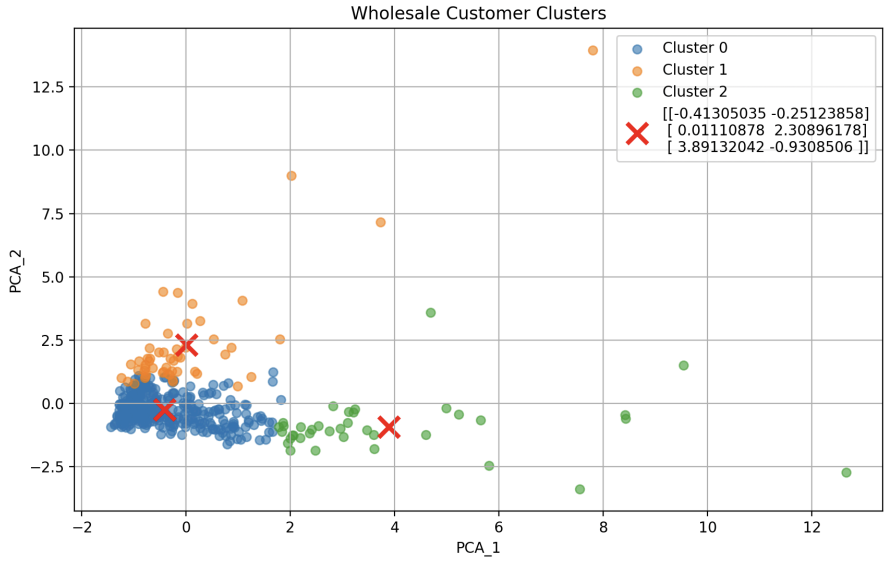

Fig 2: Illustrates three distinct clusters, each representing a different segment of wholesale customers.

The machine learning model trained using KMeans algorithm effectively identifies three distinct groups in the dataset. In Fig 2, it can be seen that cluster 0 is more compact implying that customers in that segment have relatively similar spending habits, whereas cluster 1 and 2 reflects customers with diverse spending patterns.

References:

UCI Machine Learning Repository. (n.d.). Wholesale Customers Dataset. Retrieved from https://archive.ics.uci.edu/dataset/292/wholesale+customers