Unsupervised Learning Final Project

- Customer Segmentation of Mall Customers
-
-

Problem Statement

- Businesses want to understand customers to design targeted marketing strategies.
- Key question: Can we group customers with similar characteristics without labels?
- This is a perfect case for unsupervised machine learning.

Dataset

- Mall Customers dataset: 200 customers.
- Features: Age, Annual Income, Spending Score, Gender.
- No missing values; numerical features standardized for clustering.

Exploratory Data Analysis (EDA)

- From histograms and scatterplots:
- • Younger customers often have higher Spending Scores.
- • Income and Spending Score are not strongly correlated.
- • Dataset covers a wide range of ages and income levels.

Methods

- Applied three clustering methods:
- 1) KMeans — popular and simple.
- 2) Gaussian Mixture Models (GMM) — probabilistic, supports elliptical clusters.
- 3) DBSCAN — density-based; detects noise and outliers.

KMeans Results

- Model selection via Elbow method and Silhouette score: k = 5.
- Clear separation into five groups in PCA space.
- Interpretation examples:
- • Low income & low spending — cost-conscious customers.
- • High income & high spending — VIP customers.
- • Young with moderate income & high spending — brand-conscious youth.
- • Two additional balanced groups with moderate behavior.

—

Gaussian Mixture Models (GMM) Results

- Selected number of clusters using BIC: k = 4 (best).
- GMM captures overlapping elliptical clusters better than KMeans in visualization.
- Competitive evaluation; in some cases outperforming KMeans.

DBSCAN Results

- Chose eps using the k-distance plot.
- Detected a few clusters and identified noise points.
- Clusters less balanced than KMeans/GMM — more suitable for anomaly detection.

Model Comparison

- KMeans: Good interpretability and intuitive results; assumes roughly spherical clusters.
- GMM: Best trade-off between statistical validation (BIC) and interpretability.
- DBSCAN: Useful for noise detection, less ideal for structured segmentation here.

Conclusion

- Best model: GMM with 4 clusters (chosen by BIC).
- KMeans with 5 clusters: strong baseline with intuitive business insights.
- DBSCAN highlights noise/anomalies, offering an additional perspective.

Future Work

- Include Gender and additional demographics in clustering or post-hoc analysis.
- Use t-SNE or UMAP for enhanced visualization of cluster structure.
- Validate clusters with business metrics: campaign response, customer lifetime value (LTV).

Closing

- Unsupervised learning enables practical customer segmentation.
- Comparing multiple models reveals complementary insights.
- Thank you for watching.