CentraleSupélec

Clustering and Modeling Customer Preferences

Xuanchong CHEN
Haonan ZHU
Yasmine Ben Cheikh

CentraleSupélec – Decision Making Project

February 6, 2025

CentraleSupélec

Introduction: Background and Goals

- **Supermarket Challenge**: Selecting optimal products based on customer preferences.
- **Traditional Approach**: Category managers make manual decisions.
- **Big Data Opportunity**: Using purchase history for clustering and preference modeling.
- **Goal**: Cluster customers and model decision functions.

Problem Definition

Methods:

- **Mixed-Integer Programming (MIP)**:
  - Precise modeling using **Utility Additive (UTA) model**.
  - High computational complexity.
- **Heuristic Approach**:
  - Scalable for large datasets.
  - Iterative optimization for preference-based clustering.

Evaluation Metrics:

- Explained Pairs Percentage.
- Clustering Intersection.

Problem Formulation:

- **Inputs**:
  - K: Number of clusters.
  - n: Number of criteria.
  - L: UTA segments.
  - P: Preference pairs.
- **Decision Variables**:
  - $u_i^k(x_i^l)$: Utility margin variables representing the marginal utility of criterion i at breakpoint l for cluster k.
  - z_j^k: Cluster assignment variables, indicating which cluster a preference pair $(x(j), y(j))$ belongs to.
  - σ_j^k: Error variables, accounting for deviations in preference modeling.
- **Objective Function**:
  - Minimize the total error across all preference pairs and clusters:

$$\min \sum_{j=1}^{P} \sum_{k=1}^{K} \sigma_j^k$$

Constraints:

- **Cluster Assignment**: Each preference pair j must be assigned to exactly one cluster:

$$\sum_{k=1}^{K} z_j^k = 1, \quad \forall j = 1, \ldots, P$$

- **Preference Satisfaction**: The utility of $x(j)$ must be greater than $y(j)$ in the assigned cluster k:

$$u^k(x(j)) \geq u^k(y(j)) + \varepsilon - \sigma_j^k - M(1 - z_j^k)$$

where ε is a small positive value ensuring strict preference.

Additional Constraints:

- **Monotonicity of Utility Functions**: The marginal utility functions must be monotonically increasing:

$$u_i^k(x_i^{l+1}) \geq u_i^k(x_i^l), \quad \forall k, \forall i, \forall l$$

- **Normalization of Utility Functions**: The total utility for each cluster is normalized:

$$\sum_{i=1}^{n} u_i^k(x_i^L) = 1, \quad \forall k$$

- **Boundary Conditions**: Utility values at the boundaries must be set:

$$u_i^k(x_i^0) = 0, \quad u_i^k(x_i^L) = 1, \quad \forall k, \forall i$$

MIP Results: 50s vs 300s Optimization

- **After 50s:** Best Objective = 133.8, Optimality Gap = 98.37%.
- **After 300s:** Best Objective = 106.06, Optimality Gap = 95.48%.
- **Feature Weights Analysis:** Cluster 1 and Cluster 2 have distinct feature importance.

```
Feature weights (p):
Feature 1, Cluster 1: 0.0
Feature 1, Cluster 2: 0.4200839646370762
Feature 2, Cluster 1: 0.3391943977014524З
Feature 2, Cluster 2: 0.0
Feature 3, Cluster 1: 0.660805602298547б
Feature 3, Cluster 2: 0.05631161996719469
Feature 4, Cluster 1: 0.0
Feature 4, Cluster 2: 0.523604415395729l
```

Discussion on ε, M, and MIP Gap

Impact of ε and M on Model Performance:

- Small ε $(0.001 - 0.01)$: Lower preference explanation and clustering accuracy.
- Optimal ε $(0.05 - 0.1)$: Highest accuracy (92.75% explained preferences, 95.50% clustering intersection).
- The parameter M is crucial: a small M (1.1) lowers clustering accuracy, while a large M (10) improves clustering consistency via stronger assignment penalties.

M	ε	**Percentage of explained preferences**	**Cluster intersection for all samples**	**Gap**
1.1	0.05	0.8775	0.9352	92.36%
1.1	0.1	0.8875	0.9014	99.13%
1.1	0.2	0.9065	0.8586	97.75%
10	0.05	0.901	0.9005	100%
10	0.1	0.9275	0.9550	98.26%
10	0.2	0.881	0.7210	99.14%

Table: Influence of ε and M on Model Performance

Challenges in Reducing MIP Gap:

- **Gap remained high despite optimizations.**
- **Tried methods:**
  - Adjusting ε and M values.
  - Extending solver runtime.
  - Improved constraint relaxation.
- **Conclusion:** More advanced methods (e.g., branch-and-bound improvements) needed.

Step 1: K-Means Clustering

- Initial method: Mean difference ($X - Y$) as utility.
- Issues: Non-linearity, feature interaction ignored, sensitive to outliers.

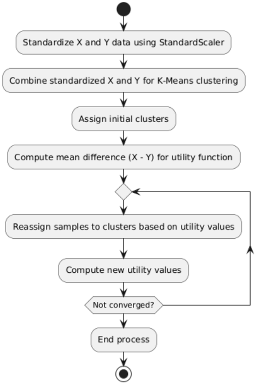


Figure: Initial Program Flowchart

Step 2: Gradient-based Optimization

- Iterative updates to utility function.
- Improved modeling accuracy.
- Limitations: Still unstable clustering results.

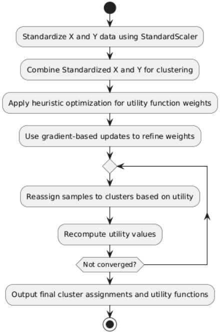


Figure: Gradient-based Utility Function

Step 3: Deep Learning for Utility Function

- Neural network to approximate utility.
- Used cross-entropy loss and Adam optimizer.
- Issues: Did not explicitly model preference constraints.

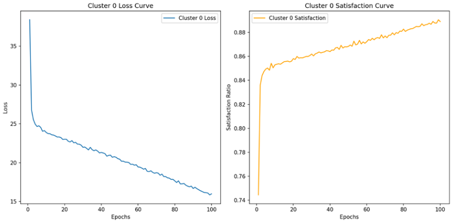


Figure: Correlation of NN Model Training Effect with Preference Satisfaction Ratio

Step 4: Soft-KMeans with Preference Constraints

- Allowed samples to belong to multiple clusters.
- Added explicit preference constraint for clustering.
- **Results:** Explained preferences: 99.7%, Clustering intersection: 80.22%.

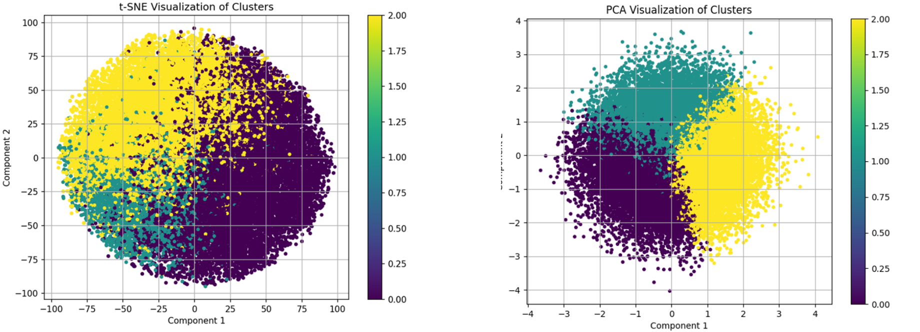
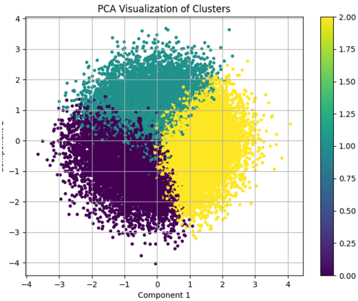


Figure: Soft-Kmeans and constrained preference-based Soft-KMeans cases

- **Goal**: Improve clustering accuracy without expensive MIP computations.
- **Key Improvements**:
  - Use **preference-aware clustering** to optimize group assignments.
  - Constraints ensure $U(X) > U(Y)$ holds in final clustering.
- **Results**:
  - Explained Preferences: 99.7%.
  - Clustering Intersection: 80.22%.



Future Prospects: Alternating Optimization

Potential Next Steps:

- **Iteratively refine** clustering and utility functions.
- Reduce preference violations at each step.
- Improve convergence by combining clustering and optimization.

Mathematical Formulation:

$$\min_{U,C} \sum_{i=1}^{N} L(U(X_i), C_i) + \lambda R(U)$$

Future Goal: Improve clustering quality by jointly optimizing preference constraints.

Conclusion & Future Work

- MIP offers **high precision** but **low scalability**.
- Heuristic methods balance **accuracy and efficiency.**
- **Future Directions**:
  - Further optimize Soft-KMeans constraints.
  - Use Reinforcement Learning to improve clustering strategies.
  - Combine Transformer models for time-series preference data.

The END

Thank you!