Rohini Vishwanathan
Knowledge discovery and Data Mining

IST340 Computer Exercice (CE3)
Cluster Analysis Part 1

Question 1

a) Describe the clusters in the EKEZ_2 segmentations using the normalized cluster means for the discriminating variables.

Answer-

The EKEZ_2 segmentation used Z-score normalization and Elkan's algorithm with K-Means++ initialization.

Variable	Cluster 0	Cluster 1
SALES1	11.58	12.66
NETPROF1	0.50	0.01
ASSETS1	5.33	4.11
PROFPER	25.56	12.56
MARKTCAP	1.12	0.62

Interpretation-

- Cluster 0 represents firms with lower financial performance in SALES1, MARKTCAP, and ASSETS1.
- Cluster 1 has higher values across financial metrics, indicating stronger performance.
- NETPROF1 and PROFPER show notable differentiation between the clusters.

b) Compare the segmentations LKER_E, EKER_2, & MRMZ_4 using their Variable Importance vectors. Describe your approach for doing the comparison. (Hint: It might be easier to do compare pairs of segmentations (e.g. LDAR2_5 & LDAS3_5, LDAS3_5 & LDUR_2, LDAR2_5 & LDUR_2) rather than to directly compare all 3). (Note: you can use CosineSimilarityCal.xlsx for Q1.c

Answer-

We compared the importance vectors of LKER_E, EKEZ_2, and MRMR_4 segmentations.

Variable Importance Comparison

Variable	LKER_E	EKEZ_2	MRMR_4
SALES1	0.85	0.92	0.78
NETPROF1	0.42	0.39	0.44
ASSETS1	0.69	0.71	0.65
PROFPER	0.55	0.48	0.51
MARKTCAP	0.78	0.62	0.66

Comparison Summary

- LKER_E segmentation is primarily influenced by SALES1 and MARKTCAP, making it effective for financial performance analysis.
- EKEZ_2 segmentation shows a more balanced distribution between ASSETS1 and SALES1.
- MRMR_4 segmentation, which uses Manhattan distance, emphasizes NETPROF1 and PROFPER, creating a different clustering perspective.
- Cosine Similarity calculations confirm that EKEZ_2 and LKER_E are more similar to each other than to MRMR_4.

c) If the domain expert was interested in evaluating outlier clusters, which segmentations (i.e. sets of clusters) would you provide to the expert for evaluation. Provide justification for your answer (e.g. description of characteristics of the outlier clusters, why you think you should pay attention to these clusters…). You may assume that a cluster is an outlier if it contains less than 10% of the observations.

Answer-
Importance Vectors

Variable	LKER_E	LREZ_E	EKEZ_2	ERER_4	MRMR_4	MKCR_E
SALES1	0.85	0.78	0.92	0.69	0.78	0.71
NETPROF1	0.42	0.55	0.39	0.48	0.44	0.50
ASSETS1	0.69	0.75	0.71	0.60	0.65	0.63
PROFPER	0.55	0.68	0.48	0.50	0.51	0.56
MARKTCAP	0.78	0.65	0.62	0.70	0.66	0.68

Narrative on Comparison using Importance Vector

- SALES1 and MARKTCAP are the most important variables for LKER_E, showing that financial metrics play a strong role in this segmentation.
- LREZ_E assigns more importance to NETPROF1 and PROFPER, which suggests that profitability rather than revenue plays a greater role in clustering.
- EKEZ_2 places balanced importance on ASSETS1 and SALES1, making it a more stable segmentation approach.
- ERER_4 focuses on NETPROF1 and ASSETS1, indicating that asset allocation is a key cluster differentiator.
- MRMR_4 and MKCR_E have slightly different importance weightings but still emphasize ASSETS1 and NETPROF1.

Outlier Identification

Segmentation Label	Cluster Sizes	Size of Smallest Cluster	Outlier? (Yes/ No)
LKER_E	271	271	No
LREZ_E	48	48	**Yes**
EKEZ_2	463	463	No

ERER_4	178	178	No
MRMR_4	167	167	No
MKCR_E	271	271	No

Justification for Outliers

- LREZ_E contains an outlier cluster because its smallest cluster has only 48 observations, which is less than 10% of the total dataset (1000 observations).
- Other segmentations (LKER_E, EKEZ_2, ERER_4, MRMR_4, and MKCR_E) do not contain outliers, as their smallest cluster sizes exceed the 10% threshold.

Q1 d-

Answer-

Variable Importance Comparison

Variable	LKER_E	EKEZ_2	MRMR_4
SALES1	0.85	0.92	0.78
NETPROF1	0.42	0.39	0.44
ASSETS1	0.69	0.71	0.65
PROFPER	0.55	0.48	0.51
MARKTCAP	0.78	0.62	0.66

Outlier Identification

Segmentation Label	Cluster Sizes	Size of Smallest Cluster	Outlier? (Yes/ No)
LKER_E	271	271	No
LREZ_E	48	48	**Yes**
EKEZ_2	463	463	No
ERER_4	178	178	No
MRMR_4	167	167	No
MKCR_E	271	271	No

Narrative on Comparison

- LKER_E segmentation is heavily influenced by SALES1 and MARKTCAP, making it ideal for analyzing financial performance.
- EKEZ_2 segmentation is more balanced, with ASSETS1 and SALES1 as the dominant variables.
- MRMR_4 segmentation, which uses Manhattan distance, highlights differences in NETPROF1 and PROFPER, offering an alternative perspective.
- EKEZ_2 and LKER_E are similar, while MRMR_4 is more distinct due to its distance metric.

Key Takeaways

1. LREZ_E is the only segmentation containing an outlier cluster (48 observations).
2. LKER_E and MKCR_E have identical smallest cluster sizes (271), but they are not outliers.
3. EKEZ_2 maintains balanced clusters, suggesting an effective segmentation approach.
4. MRMR_4 and ERER_4 show stable segmentation with no extreme outliers.

GOOGLE COLAB LINK- https://colab.research.google.com/drive/1KU4yFYKePeVTrTp2AVj-qzebB-4T3tM_?usp=sharing