Lab Report for CW3 Data Clustering

[redacted]

***Abstract*—This report focuses on processing the final grades data from the 2023 INT104 course, employing various clustering methods to identify potential relationships between the clustering results and the 'Programme' distribution. The experiments applied Gaussian Mixture Model (GMM), K-means and Hierarchical Clustering with 4 and 7 clusters respectively. Results indicate a positive correlation between Silhouette Score and accuracy for GMM, while K-means and Hierarchical Clustering display an inverse conclusion and indicate negative correlation.**

***Index Terms*—Data Clustering, GMM Clustering, Hierarchical Clustering, K-means Clustering**

I. Introduction

THIS report aims to analyze the grades data from the INT104 final exam in 2023. The dataset in 'xlsx' format contains 11 columns: 'Index', 'Gender', 'Programme', 'Grade', 'Total', 'MCQ', 'Q1', 'Q2', 'Q3', 'Q4', and 'Q5', with a total of 619 samples. As an unsupervised learning method, clustering results will be compared with the original 'Programme' distribution to explore potential relationships with the 'Programme' column.

II. Experiment Preparation

A. Data Pre-processing

The data is split into a 70% training set and a 30% test set. Initially, the 'Index' and 'Programme' columns were removed, retaining all other columns. To enhance visualization and evaluation, PCA was used for dimensionality reduction. Consistent with the CW1 experiment, both sets were standardized before PCA, and the top two principal components explained almost 95% of the variance. This minimal information loss facilitates representation and adjustment of clustering results. Subsequent experiments performed clustering analysis on both 2D and 3D reduced data.

B. Performance Measure

To evaluate clustering performance, the Silhouette Coefficient was used to represent intra-cluster cohesion and inter-cluster separation. For assessing the 'Programme' representation capability, each cluster was assigned the label of the most frequent 'Programme' within it, and overall accuracy was calculated. Confusion matrices were also utilized to display the results.

C. Cluster Number Determination

Determining the optimal number of clusters requires considering both clustering performance and accuracy. According to Fig. 1., the line graph levels off after the number of clusters reaches 4 and 7, suggesting these as optimal candidates for best clustering performance.

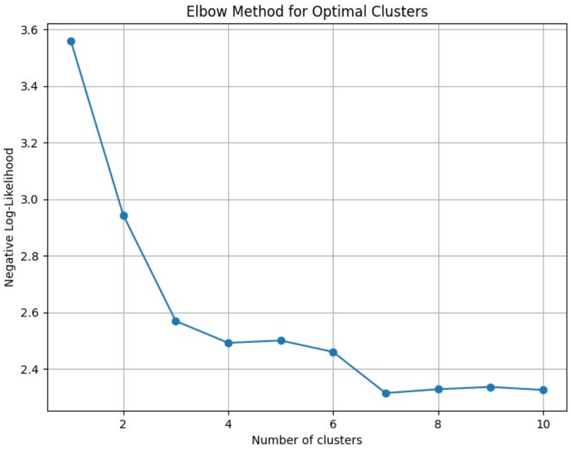

Fig. 1. Elbow Method for Optimal Clusters

D. Feature Extraction

To extract appropriate feature sets, Spearman's Rank Correlation Coefficient was applied to observe correlations among different features. As shown in Fig. 2., 'Total' and 'MCQ' exhibited a strong correlation, allowing for the removal of one. After retaining either 'Total' or 'MCQ', other strongly correlated features were also removed, such as 'Q2' and 'Q4' for 'Total', and 'Q2' and 'Q5' for 'MCQ'.

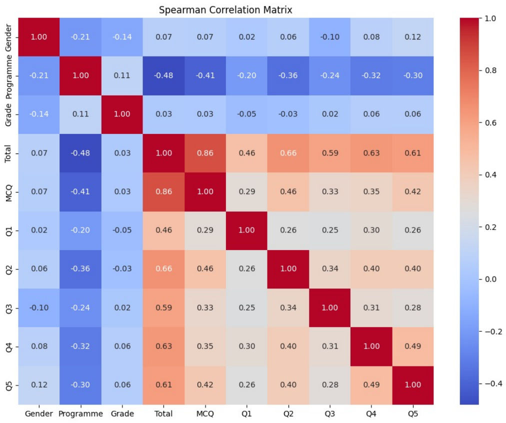

Fig. 2. Spearman Correlation Matrix

III. 4 Clusters

A. GMM Clustering

GMM was performed on both 2D and 3D training datasets, with evaluation on test datasets. The results for 2D and 3D showed a certain degree of similarity. For example, in the 2D results, the 'Programme' distribution in each cluster is shown in Fig. 3.. In the first cluster, there are many samples from three different 'Programmes'. However, this cluster was assigned to 'Programme 1', and other samples were considered misclassified. This indicates relatively low accuracy for this clustering, as evidenced by the experimental results. The 2D GMM clustering achieved a Silhouette score of 0.14 and accuracy of 0.49, while the 3D GMM clustering achieved a Silhouette score of 0.26 and accuracy of 0.49. The improvement from 0.14 to 0.26 is likely due to the increased distances brought by the additional third dimension, as shown in Fig. 4..

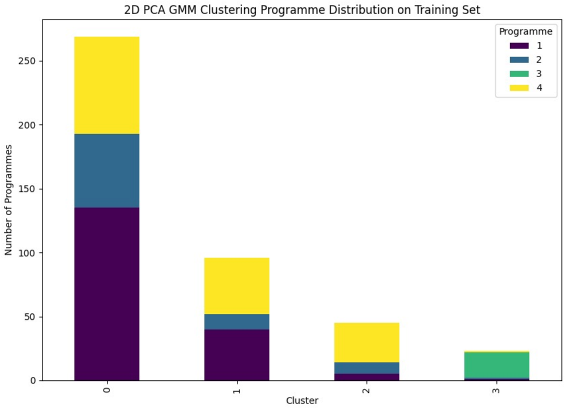

Fig. 3. 'Programme' Distribution for 2D GMM Clustering

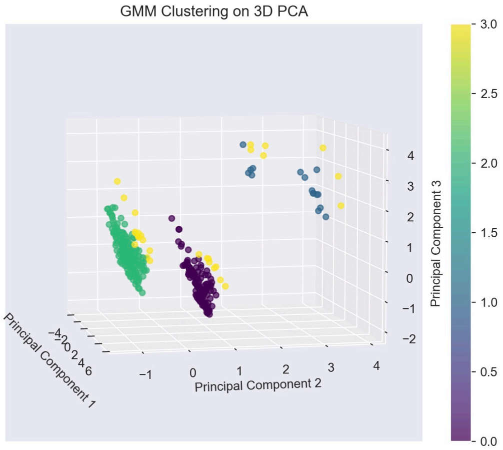

Fig. 4. GMM Clustering on 3D PCA

The results of GMM are not ideal, but they are predictable since the features do not strictly follow a Gaussian distribution as experimented in CW2. Further attempts on different feature sets will be discussed.

B. K-means Clustering

K-means clustering was also performed on both 2D and 3D datasets. The confusion matrices for 2D and 3D clustering results reveal an interesting point: 2D clustering fails to identify any test samples from 'Programme 3', while 3D clustering almost perfectly identifies all samples. This difference is likely due to the algorithm used for assigning labels. Even with 4 clusters, there is no guarantee of a one-to-one correspondence between clusters and the 'Programme'. To maximize accuracy, the algorithm always chooses the most frequently occurring 'Programme' as the label for the cluster. In this case, the algorithm assigns 'Programme 1' to clusters 0 and 1, and 'Programme 4' to clusters 2 and 3, as shown in Fig. 5.. Consequently, samples from 'Programme 2 and 3' cannot be correctly classified, which is confirmed by the resulting confusion matrix.

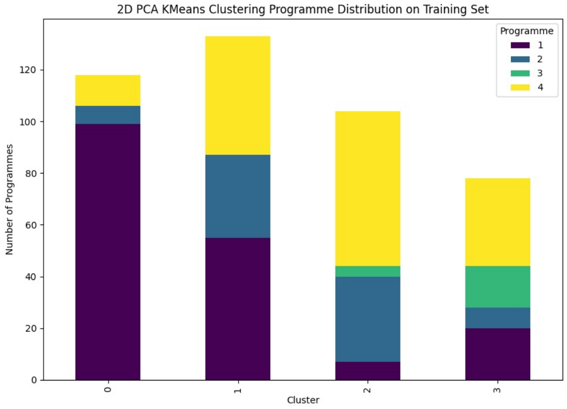

Fig. 5. 'Programme' Distribution for 2D K-means Clustering

The purity of cluster 0 is very high, which is the ideal result, while others are relatively low. The 2D K-means clustering has got the Silhouette score of 0.40 and accuracy of 0.51, and the 3D K-means clustering has got the Silhouette score of 0.36 and accuracy of 0.61. Compared to GMM clustering, both Silhouette score and accuracy have enhanced.

C. Hierarchical Clustering

Agglomerative Hierarchical Clustering was applied to 2D and 3D training datasets in this experiment. Ward's method was used for merging, as it aims to minimize variance and typically produces compact and uniform clusters, effectively handling noise and outliers. The resulting 2D clusters faced a similar dilemma as K-means clustering, sacrificing 'Programme 2 and 3' for better prediction of 'Programme 1 and 4'. Meanwhile, the 3D results only sacrificed 'Programme 2', indicating that the additional dimension 'PC3' contains crucial information related to 'Programme 3'.

The dendrogram for 2D clustering is shown in Fig. 6.. The 2D

TABLE I
2D Clustering Results of Different Feature Sets

	GMM Clustering		K-means Clustering		Hierarchical Clustering	
Removing	Silhouette Score	Accuracy	Silhouette Score	Accuracy	Silhouette Score	Accuracy
'Index', 'Programme'	0.14	0.49	0.40	0.51	0.40	0.51
'Index', 'Programme', 'MCQ'	0.17	0.49	0.40	0.53	0.39	0.51
'Index', 'Programme', 'MCQ', 'Q2'	0.28	0.49	0.39	0.54	0.42	0.51
'Index', 'Programme', 'MCQ', 'Q5'	0.24	0.52	0.43	0.51	0.39	0.51
'Index', 'Programme', 'MCQ', 'Q2', 'Q5'	0.31	0.53	0.45	0.48	0.32	0.56
'Index', 'Programme', 'Total'	0.21	0.54	0.40	0.52	0.37	0.52
'Index', 'Programme', 'Total', 'Q2'	0.31	0.49	0.44	0.50	0.36	0.52
'Index', 'Programme', 'Total', 'Q4'	0.21	0.49	0.41	0.49	0.41	0.47
'Index', 'Programme', 'Total', 'Q2', 'Q4'	0.28	0.49	0.41	0.51	0.45	0.49

TABLE II
3D Clustering Results of Different Feature Sets

	GMM Clustering		K-means Clustering		Hierarchical Clustering	
Removing	Silhouette Score	Accuracy	Silhouette Score	Accuracy	Silhouette Score	Accuracy
'Index', 'Programme'	0.26	0.49	0.36	0.61	0.29	0.59
'Index', 'Programme', 'MCQ'	0.28	0.60	0.40	0.59	0.28	0.57
'Index', 'Programme', 'MCQ', 'Q2'	0.26	0.53	0.41	0.57	0.32	0.59
'Index', 'Programme', 'MCQ', 'Q5'	0.35	0.49	0.44	0.57	0.39	0.58
'Index', 'Programme', 'MCQ', 'Q2', 'Q5'	0.38	0.60	0.47	0.59	0.39	0.61
'Index', 'Programme', 'Total'	0.30	0.49	0.40	0.60	0.29	0.58
'Index', 'Programme', 'Total', 'Q2'	0.31	0.58	0.40	0.57	0.38	0.58
'Index', 'Programme', 'Total', 'Q4'	0.23	0.51	0.39	0.61	0.38	0.59
'Index', 'Programme', 'Total', 'Q2', 'Q4'	0.34	0.56	0.42	0.57	0.39	0.59

hierarchical clustering achieved a Silhouette score of 0.40 and accuracy of 0.51, while the 3D hierarchical clustering achieved a Silhouette score of 0.29 and accuracy of 0.59.

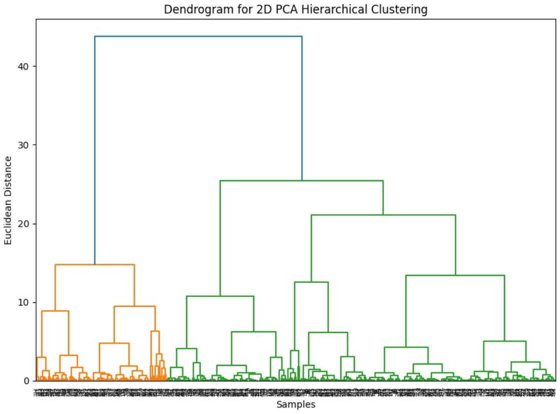

Fig. 6. Dendrogram for 2D Hierarchical Clustering

D. *Different Feature Sets*

As stated in 'Feature Exaction' section, attempts have been made to try various feature sets on each clustering method, and the results of 2D and 3D have been shown in TABLE I and TABLE II respectively.

IV. 7 Clusters

The number of clusters was set to 7, and all the previously mentioned actions were performed. The results show a significant accuracy increase in 2D, while remaining steady in 3D. Due to the denser distribution of samples in 2D images, increasing the number of clusters may lead to overfitting. The Silhouette Score decreases, especially in GMM, as points from different clusters are more likely to stick together.

V. Result

By observation and comparison, here are several findings and explanations:

- GMM clustering results in generally higher Silhouette score in 3D than 2D, while K-means and Hierarchical clustering seem not. One potential explanation is that the probability density functions that GMM applies is more effective in high-dimensional spaces since it takes the covariance structure of the data into account.
- GMM clustering results show a positive correlation between Silhouette Score and accuracy in both 2D and 3D. This indicates that removing features that do not follow a Gaussian distribution significantly improves the performance of GMM.
- K-means and hierarchical clustering results show a negative correlation between Silhouette Score and accuracy in both 2D and 3D, especially for K-means. This indicates that the 'Programme' feature itself does not exhibit distinct clustering characteristics. Therefore, optimizing clustering performance does not necessarily enhance the classification accuracy of the 'Programme' feature.

VI. Recommendation

Further experiments can apply cross-validation for optimization. Moreover, attempts with different choice of cluster amount can be made. Last but not least, 'Programme 2' consistently fails to be correctly predicted. Further research should be done to provide a more in-depth explanation.