CS7641: Unsupervised Learning and Dimensionality Reduction

Nayeem Aquib
naquib3@gatech.edu

***Abstract*—** This report presents an extensive exploration of clustering and dimensionality reduction techniques on two datasets. A comparative analysis of the clustering algorithms and dimensionality reduction algorithms is provided. Additionally, the report evaluates the performance of a neural network learner when incorporating dimensionality reduction techniques and cluster labels as features. Overall, the report provides a comprehensive analysis of the interplay between clustering, dimensionality reduction, and neural network performance.

1 Dataset Description

Two datasets were used for this report. One is a dataset of video game sales across different platforms and different regions with 16598 rows and 11 columns. The information of this dataset can be potentially used for understanding market trends, consumer preferences, and the impact of various factors on video game success. Another is a dataset of college applicants with different metrics from their applications including the decision of the college with 8358 rows and 7 columns. This dataset can be used to explore correlations between demographic, economic, and academic factors and admission outcomes.

2 Hypotheses

Sales figures could vary by genre, suggesting trends in video game preferences over time. For instance, certain genres may have been more popular during specific decades.

Higher academic performance (GPA and ACT scores) correlates with a higher likelihood of admission, indicating the weight of academic credentials in admission decisions. Also, Applicants from higher household income brackets

have a higher rate of admission, potentially reflecting the influence of socio-economic status on educational opportunities.

3 Clustering Algorithms

The chosen clustering algorithms are: Gaussian Mixture Models (**GMM**) variation of **EM** and **KMeans**.

GMM(Time complexity O(n*k*d^2) for n samples, k components, and d dimensions) was chosen as it can potentially identify the clusters with different shapes and densities, which may correspond to various game genres and their sales patterns. For the college admissions dataset, GMM can account for the varied distribution of applicants' profiles.

KMeans(Time complexity O(n*k*i*d) for n samples, k clusters, i iterations, and d dimensions) was selected because it will group games based on their sales figures, assuming Euclidean distance is a good measure of similarity between sales profiles. In college admissions data, KMeans may cluster applicants into distinct categories based on their academic and socioeconomic status, testing the hypothesis that certain profiles are more likely to gain admission.

For the video game sales dataset, I focused on the sales figures in different regions as features, while for the college admission dataset, I used GPA, ACT scores, and household income as the features. After applying the clustering algorithms to the two datasets, I received the following clusters:

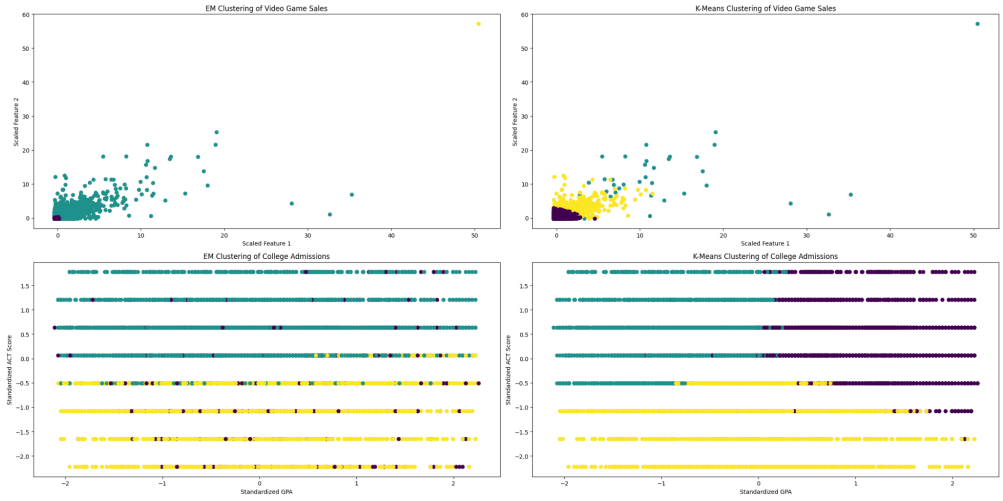

Figure 1: Clustering Algorithms

For the **video game sales data**, the **GMM** algorithm created clusters that are diffuse, which indicate that the sales data does not follow a strictly Gaussian distribution. There appears to be a small cluster of high-sales games far from the main group, which **makes sense** because there are blockbuster titles with significantly higher sales than the average game. The **K-Means** algorithm has produced clusters that suggest a gradation in sales figures, with a more distinct grouping of lower-sales games and some separation among higher-sales games. The presence of outliers, particularly in the **GMM** clustering, raises questions about the influence of exceptionally successful titles on the overall sales landscape. For the **college admissions data**, the **GMM** algorithm seems to have identified clusters primarily based on the ACT score, suggesting that the ACT score is a significant factor in differentiating clusters. This result supports the hypothesis that academic performance is a strong determinant in the admissions process. **K-Means** has generated more dispersed clusters compared to GMM. The clusters obtained **make sense** within the context of college admissions, where standardized test scores and GPAs are typically key factors in decision-making. I then compared the clusters against the "admitted" labels to see if there is a match, which provided a 82% match for GMM and 74% match for K-Means. I wonder if including essays, recommendations, or extracurricular activities as part of the dataset would change the clustering dynamics.

Performance-wise, for **GMM**, I would adjust the covariance type (full, tied, diag, spherical) to better fit the data distribution. For the **K-Means** cluster, I would experiment with the number of clusters (k) using methods like the elbow method or silhouette score first before moving forward.

4 Dimensionality Reduction Algorithms

The chosen algorithms are **PCA(O(n*d^2)+O(d^3)** for n samples and d dimensions**)**, **ICA(O(n*d^3)** for n samples and d dimensions**)**, **Randomized Projections(O(n*d*k)** for projecting n samples from d dimensions to k dimensions**)**, and Multidimensional Scaling or **MDS(O(n^3)** due to eigenvalue decomposition, distance computations, and iterations for stress minimization**)**. **MDS** was chosen because it can reveal the non-linear relationships that may exist between different genres and their sales over time.

After applying the dimensionality reduction algorithms to the two datasets, I received the following plots:

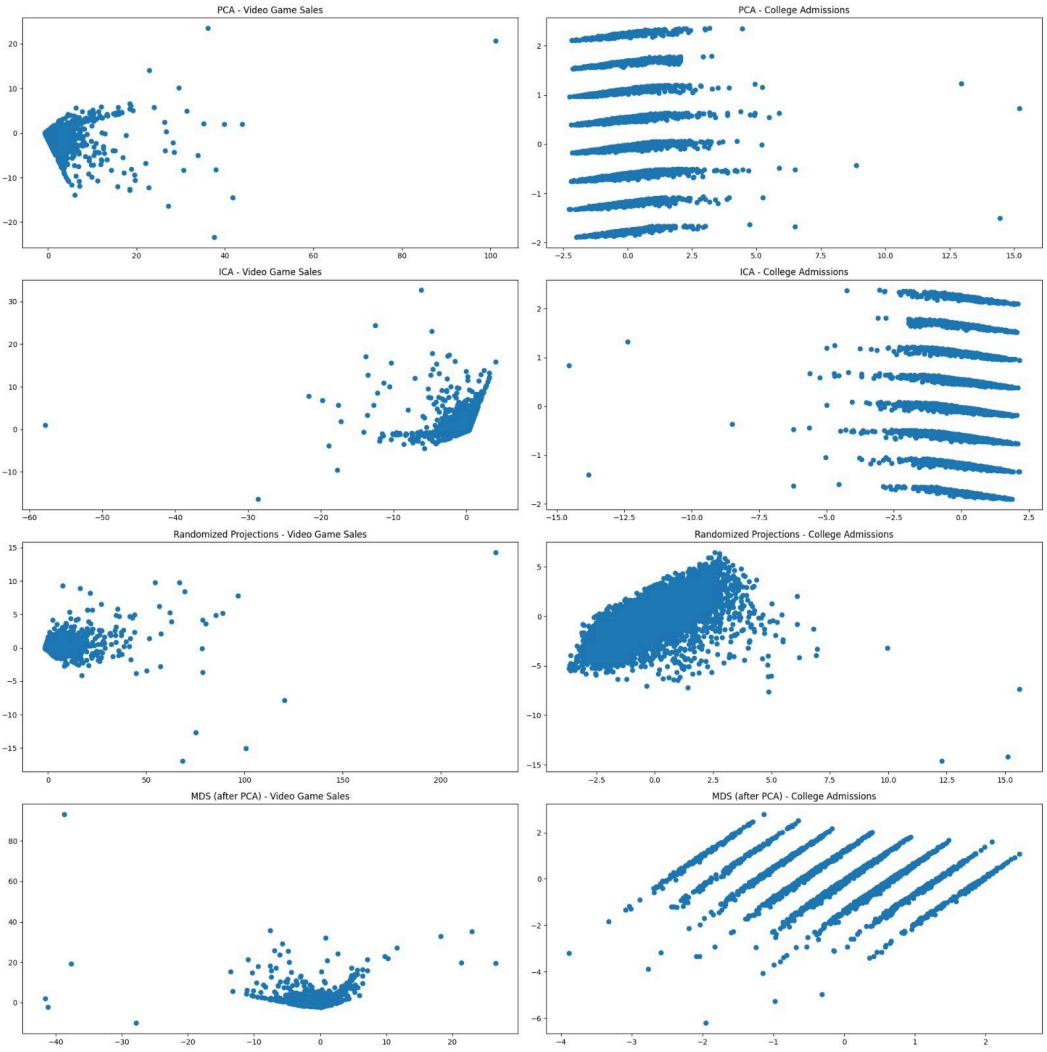

Figure 2: Dimensionality Reduction Algorithms

For **video games sales data**, the **PCA** plot shows a clear gradient or "elbow" shape, reflecting the disparity in global sales, with a few games achieving very high sales and many games with much lower sales. From the **eigenvalues, [3.67e+00 7.55e-01 3.57e-01 2.10e-01 7.85e-06]**, it seems that the first principal component captures significantly more variance than the others. The rapid drop-off after the first principal component suggests that the video game sales data can be largely represented by the first component. This probably means that a single factor, like the global popularity of a game, could be the main driver of the variance in sales figures. **ICA** appears to have separated out sources of variation in the data, potentially uncovering underlying factors affecting video game sales. The **kurtosis value**, **[792.61 189.14],** suggests that a few blockbuster

games have sales far beyond the average. I sorted the dataset based on sales and confirmed the suggestions of Kurtosis. For **randomized projections,** I calculated how well data is reconstructed by randomized projections and found an **Average Reconstruction Error of 7.087e-30** and a **Standard Deviation of Reconstruction Error of 1.61e-29,** which are very close to zero, suggesting that the reconstruction of the data from the randomized projections is extremely accurate. When I introduced noise, the **Average Reconstruction Error with Noise** was **2.10e-29** and the **Standard Deviation of Reconstruction Error with Noise was 1.01e-28,** which also suggests the same. The **variance** across 100 runs was **2.04,** suggesting that the transformation is reproducibly capturing its structure, despite the inherent randomness of the method. **MDS** seems to reveal a non-linear manifold, showing clusters that could correspond to different tiers of sales performance, with some outliers that likely represent blockbuster hits. Disregarding the final eigenvalue which is close to 0, the **rank of the video game sales data is 4**. To check for **collinearity**, I calculated the VIF values, which came out as **[1.00, 24462.44, 9369.69, 3509.05, 1305.57, 88689.90]**, suggesting that there is substantial multicollinearity among the variables in the video game sales dataset. So one or more of these variables can be predicted from the others with a high degree of accuracy. However, due to this, small changes in the data can lead to large changes in the model coefficients.

For **college admissions data**, the **PCA** plot shows distinct horizontal bands, likely corresponding to discrete ACT score ranges. From the **Eigenvalues, [1.0929668 1.00006178 0.907425],** it seems that the variance in the college admissions dataset is quite evenly distributed across the three principal components. There doesn't seem to be a single dimension that dominates. In the case of the **ICA** plot, similar to PCA, there are distinct bands, but they are not as cleanly separated, indicating that ICA components are not primarily driven by the ACT score alone and may be capturing more subtle independent factors. The **kurtosis** value, **[17.85589559 -0.35983463],** suggests that the dataset has both **leptokurtic component** and **platykurtic component.** So there are some candidates with unique backgrounds in the dataset. After checking the dataset out, the primary component that stuck out was the number of native Hawaiian or other pacific islander were significantly low. In case of **randomized projections**, the visible layers suggest that the projection still preserves some structure related to ACT scores or GPA levels. The **Average Reconstruction Error** was **3.88e-29** and the **Standard Deviation of Reconstruction Error** was **2.22e-28,** which are

very close to zero, suggesting that the reconstruction of the data from the randomized projections is extremely accurate.With **noise**, the **Average Reconstruction Error** and the **Standard Deviation of Reconstruction Error** was **4.83e-29** and **2.96e-28** respectively, suggesting the same. The **variance** across 100 runs was **1.48,** suggesting that the transformation is reproducibly capturing its structure, despite the inherent randomness of the method. **MDS** has produced a plot with diagonal striping, which could indicate that the admission decision-making process is complex and influenced by multiple, possibly non-linearly interacting factors. Based on the eigenvalues, the **rank of the college admission data is 3**. To check for collinearity, I calculated the VIF values, which came out as **[1.0, 1.0086, 1.00, 1.0086],** suggesting there is virtually no multicollinearity among the variables in the college admission dataset. This is great for regression modeling, making the model more reliable for inference or prediction regarding the impact of each independent variable on the outcome variable.

Performance-wise, for **PCA** and **ICA**, I would select the appropriate number of components. For **Random Projections**, I would use sparse random projections, and for **MDS**, I would consider using a variant like SMACOF(Scaling by Majorizing a Complex Function) that scales better.

4 Reapply Clustering Algorithms on Dimensionally Reduced Datasets

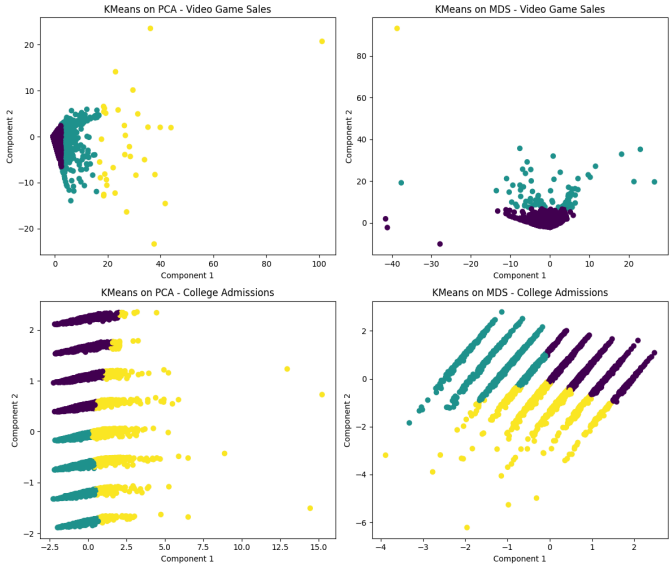

Fig 3: Clustering Algorithms on Reduced Datasets

The **PCA** plot for video game sales shows distinct clusters, potentially representing different genres. **MDS**, however, doesn't show as clear of a differentiation by clusters. In the **PCA** plot for college admissions, the stratification probably indicates that higher academic achievements do indeed play a significant role in admissions, aligning with the hypothesis. The **MDS** plot shows a diagonal striping pattern, suggesting that admissions decisions may involve a combination of factors, including academic performance and potentially other attributes.

In both datasets, the clusters have a very similar pattern. This is most likely due to 1. The outlier blockbuster games stay as outliers. However, the other genre specific games have a normal sales pattern. 2. Higher academic achievements do indeed play a significant role in admissions

5 Rerun Neural Network Learner on Reduced Datasets

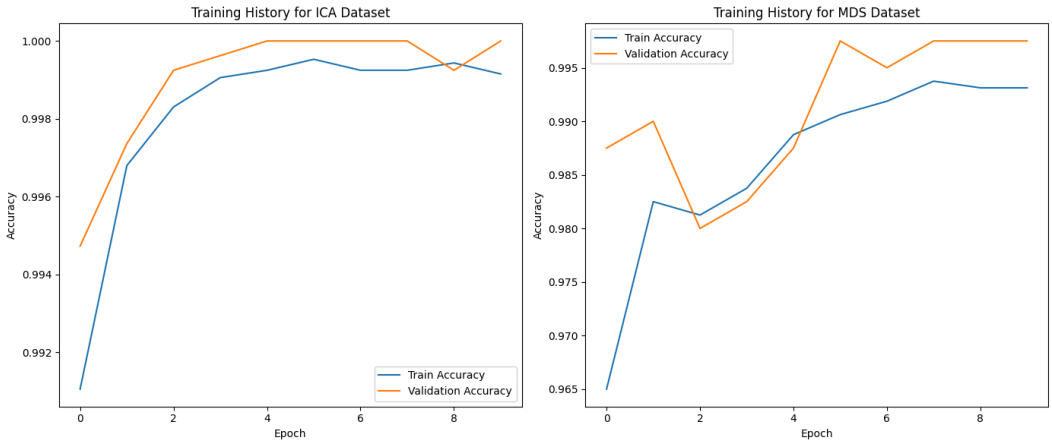

Fig 4: Rerun Neural Networks on Reduced Datasets

There is a visible gap between the training and validation accuracy in the **ICA** dataset, especially towards the end of training, which might indicate overfitting. The training accuracy for the **MDS** dataset is more volatile, with significant dips and recoveries, especially in the early epochs. This could suggest that the model parameters are not optimal for this data representation. The high accuracy achieved on the ICA dataset may indicate that the neural network can capture differences in sales figures that could be associated with game genres. The consistent high accuracy for the ICA dataset could imply that the genre has a strong correlation with sales, supporting the hypothesis.

6 Rerun Neural Network Learner on Cluster

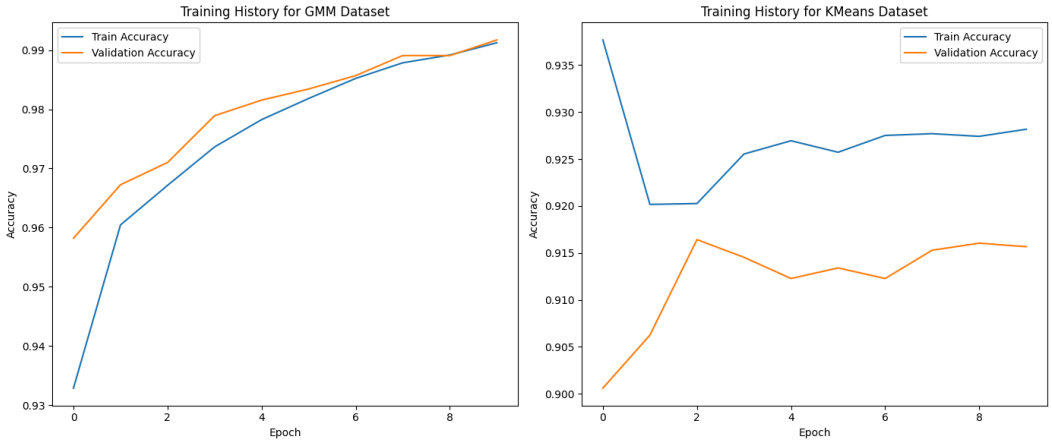

Fig 5: Return Neural Network Learner on Cluster

The accuracy for the GMM-labeled dataset increases rapidly and then plateaus, with the training and validation accuracy lines running close together. This indicates that the model is learning well and generalizing well to unseen data. Given the hypothesis about video game sales varying by genre, this plot suggests that the neural network might effectively discern trends or patterns related to these genres. The stable high accuracy would imply that genres, as captured by GMM clustering, are a strong predictor of sales figures. The training history for the KMeans-labeled dataset shows more fluctuation in validation accuracy, indicating potential overfitting or less stable learning. The training accuracy remains higher than the validation accuracy, which is typically a sign of the model learning specifics about the training data that do not generalize as well. Given the hypothesis, the fluctuations in accuracy could suggest that the KMeans clustering does not align as cleanly with genre-based trends in video game sales, or that the clustering is too simplistic to capture the nuances of sales trends over time.

The time complexity of a neural network is dependent on the number of layers, neurons, epochs, and data size. Training is usually O(n * e * l * m), with n samples, e epochs, l layers, and m neurons. However, when I ran my neural network learner on MDS, it was significantly faster (8x) then the previous run. This is probably due to Mini-Batch Gradient Descent and GPU acceleration.