BDA Final Project

HuaHsuan Chen
CSIE, NTU
b11902157@ntu.edu.tw

https://github.com/ShuttonChen/BDA-final-project

June 11, 2025

1 Methodology

This work implements an end-to-end clustering pipeline on two datasets (public and private) comprising numerical features. The overall workflow is:

1. **Data Preprocessing**: Standardize each feature to zero mean and unit variance to eliminate scale effects.
2. **Dimensionality Reduction**: Apply Principal Component Analysis (PCA) to project the data into a lower-dimensional subspace that retains at least 75% of the original variance.
3. **Clustering**: Perform K-Means clustering on the PCA-transformed data with the number of clusters set to

$$k = 4n - 1,$$

where n is the original feature dimension.

This methodology is applied identically to both the public (4-dimensional) and private (6-dimensional) datasets, enabling fair comparison of clustering performance.

2 Data Preprocessing

Before clustering, we apply the following steps to ensure feature comparability and improve clustering performance:

Standardization: Transform each feature to zero mean and unit variance:

$$X_{\text{scaled}} = \frac{X - \boldsymbol{\mu}}{\boldsymbol{\sigma}},$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are the per-feature mean and standard deviation.

3 Principal Component Analysis (PCA)

We reduce the data to a lower-dimensional subspace that retains 75% of the total variance:

- Compute the covariance matrix:

$$\Sigma = \frac{1}{m-1} X_{\text{scaled}}^T X_{\text{scaled}}.$$

- Obtain eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and corresponding eigenvectors.
- Select the smallest integer d such that

$$\frac{\sum_{i=1}^{d} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \geq 0.75.$$

- Project the data onto the first d principal components:

$$X_{\text{reduced}} = X_{\text{scaled}}\, W_d,$$

  where $W_d \in R^{n \times d}$ contains the top d eigenvectors.

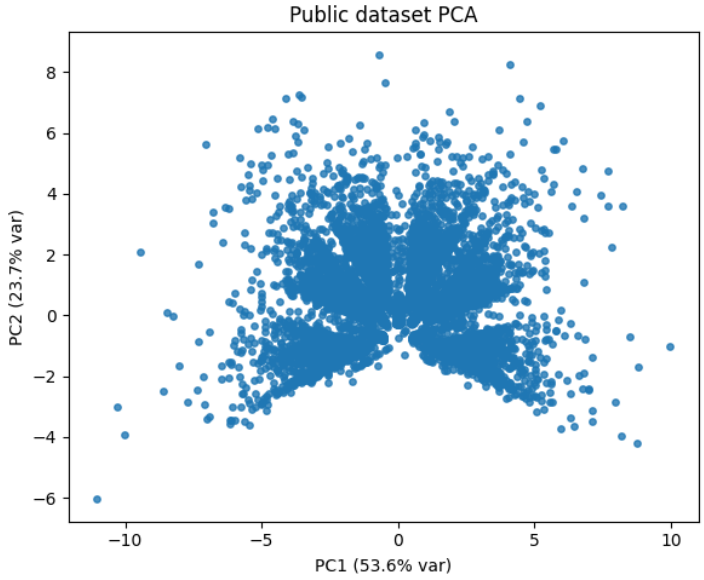

Figure 1: 2D PCA projection of the public dataset onto the first two principal components, which explain 89.5% and 8.0% of the total variance, respectively.

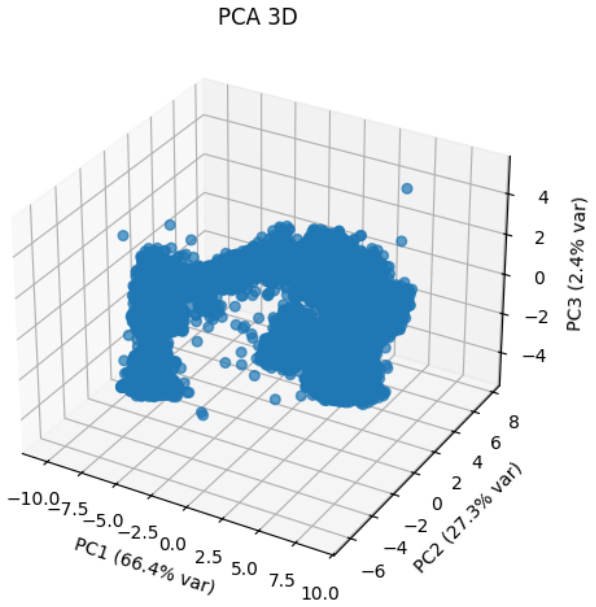

Figure 2: Visualization of PCA-based dimensionality reduction. (a) Public dataset in 2D showing clear separation along PC1 and PC2. (b) Private dataset in 3D illustrating the data structure across PC1, PC2 and PC3.

4 K-Means Clustering Algorithm

Algorithm 1 K-Means Clustering

Require: Reduced data $X_{\text{reduced}} \in R^{m \times d}$, number of clusters k
Ensure: Cluster labels $\{\ell_i\}_{i=1}^{m}$
1: Initialize k centroids $\{c_j\}_{j=1}^{k}$ using k-means++
2: **repeat**
3: **for** $i = 1, \ldots, m$ **do**
4: Assign sample x_i to the nearest centroid:

$$\ell_i = \arg\min_{1 \le j \le k} \|x_i - c_j\|_2^2.$$

5: **end for**
6: **for** $j = 1, \ldots, k$ **do**
7: Update centroid:

$$c_j = \frac{1}{|C_j|} \sum_{i:\ell_i = j} x_i,$$

where $C_j = \{i : \ell_i = j\}$.
8: **end for**
9: **until** Assignments do not change or reach maximum iterations

5 Clustering Evaluation

5.1 Silhouette Score

For each sample i, compute:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

where

- $a(i)$ is the average distance from sample i to all other points in its cluster.
- $b(i)$ is the minimum average distance from sample i to points in any other cluster.

The overall Silhouette Score is the mean of $s(i)$ over all samples, with range $[-1, 1]$.

5.2 Davies–Bouldin Index

Defined as:

$$DB = \frac{1}{k} \sum_{j=1}^{k} \max_{h \neq j} \frac{S_j + S_h}{M_{jh}},$$

where

- S_j is the average distance of all samples in cluster j to its centroid.
- M_{jh} is the distance between centroids of clusters j and h.

A lower DB index indicates better clustering (low intra-cluster variance and high inter-cluster separation).

6 Results and Discussion

Based on the clustering runs, we obtained the following metrics:

- Public dataset (4 dimensions):
  - PCA reduced from 4 to 2 dimensions.
  - Silhouette Score: 0.6445 (indicating good intra-cluster cohesion and inter-cluster separation).
  - Davies–Bouldin Index: 0.7217 (relatively low, confirming compact and well-separated clusters).
- Private dataset (6 dimensions):
  - PCA reduced from 6 to 3 dimensions.
  - Silhouette Score: 0.5666 (indicating moderate to good clustering quality).
  - Davies–Bouldin Index: 0.7461 (maintaining decent compactness and separation).

Overall, both datasets achieved satisfactory clustering performance. Therefore, we have reason to believe that the classification performance based on this algorithm should be good.

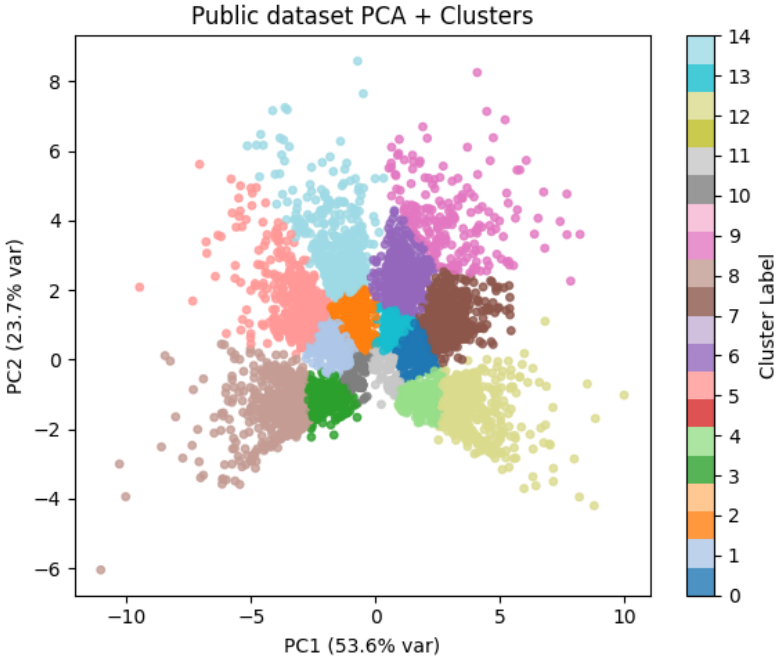

Figure 3: Public data result Visualization

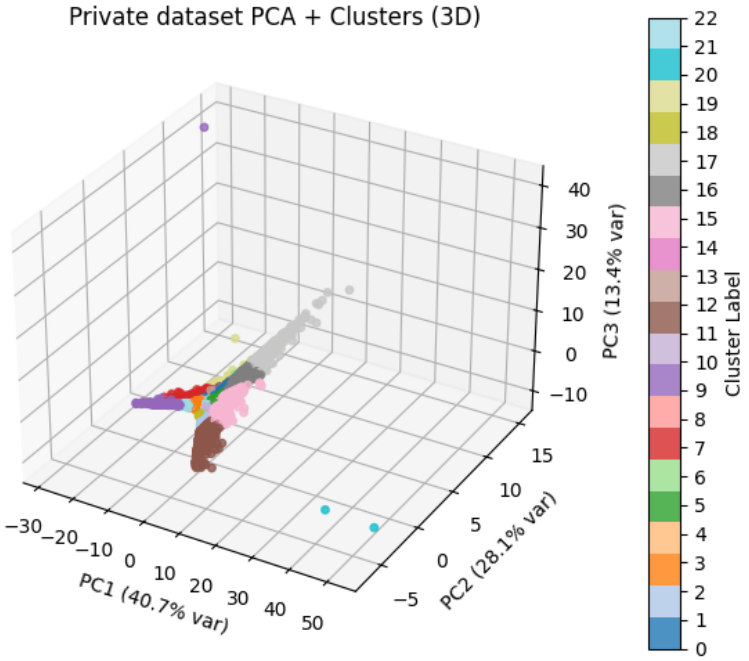

Figure 4: Private data result Visualization.