Assignment 4: PCA and EM Algorithm

Name: Md Raihan Sobhan
Student ID: 1905095
Course: CSE472 (Machine Learning Sessional)

November 22, 2024

Introduction

Principal component analysis (PCA) and the expectation-maximization (EM) algorithm are two of the most widely used unsupervised methods in machine learning. In this assignment, we had to implement the following things.

- Principal Component Analysis (PCA) for dimensionality reduction.
- UMAP and t-SNE visualizations for analyzing high-dimensional data.
- Expectation-Maximization (EM) algorithm for estimating parameters of a Poisson mixture model.

How to Run the Code

- Place the dataset files (`pca_data.txt` and `em_data.txt`) in the same directory as the script.
- Run the Jupyter Notebook `1905095.ipynb`. Install necessary libraries before running.
- Generated plots and results will be saved in the working directory.

Principal Component Analysis (PCA)

The PCA algorithm was implemented without using library functions for the PCA process. Only basic matrix operations like eigendecomposition were utilized.

PCA Scatter Plot

The data was projected onto the two principal components with the highest eigenvalues, producing the following scatter plot:

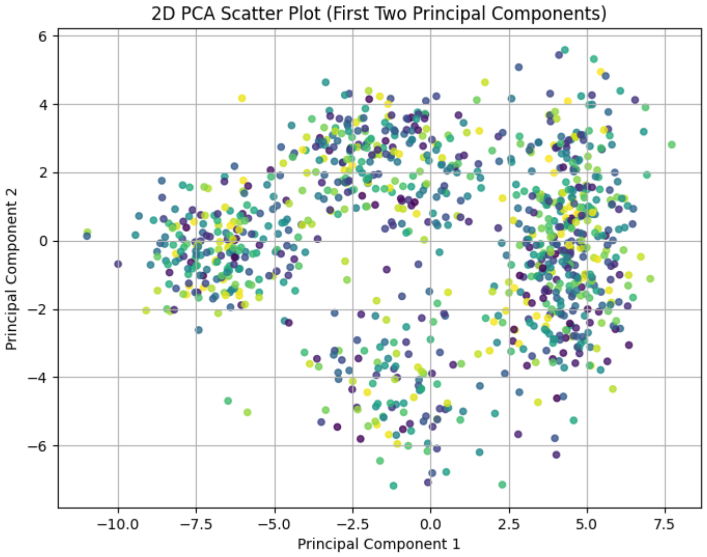

Figure 1: PCA Scatter Plot of First Two Principal Components

UMAP and t-SNE Visualizations

The UMAP and t-SNE algorithms were applied to the original dataset using library functions to generate 2D visualizations. These techniques complement PCA by preserving the local and global structure of the data.

UMAP Plot

The UMAP plot of the dataset is shown below:

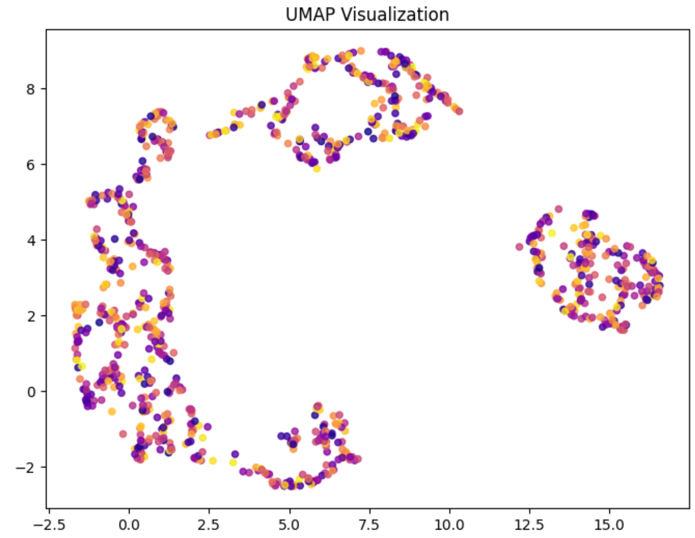

Figure 2: UMAP Visualization

t-SNE Plot

The t-SNE plot of the dataset is shown below:

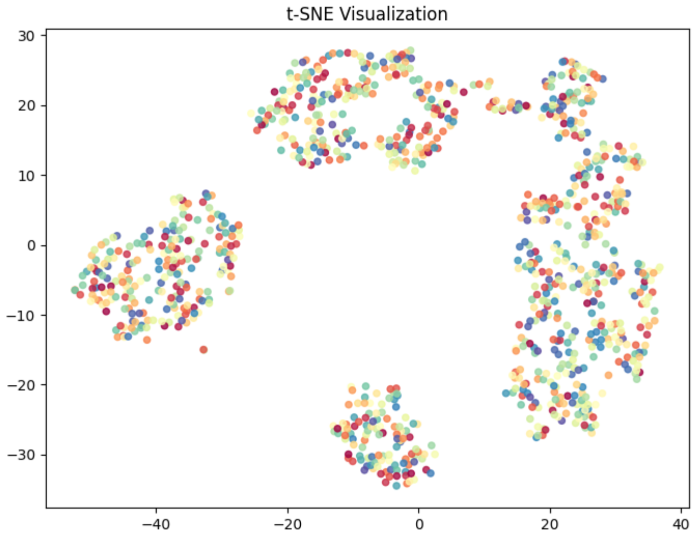

Figure 3: t-SNE Visualization

Expectation-Maximization (EM) Algorithm

The EM algorithm was implemented to estimate the parameters of a Poisson mixture model based on the number of children in families.

Estimated Parameters

The algorithm estimated the following parameters:

- **Proportion of families with family planning (π):** `0.3562`
- **Proportion of families with family planning (π):** `0.6438`
- **Mean number of children (with family planning, λ_1):** `1.7851`
- **Mean number of children (without family planning, λ_2):** `4.9113`

Log-Likelihood Convergence

The log-likelihood values during EM algorithm iterations are shown in the figure below:

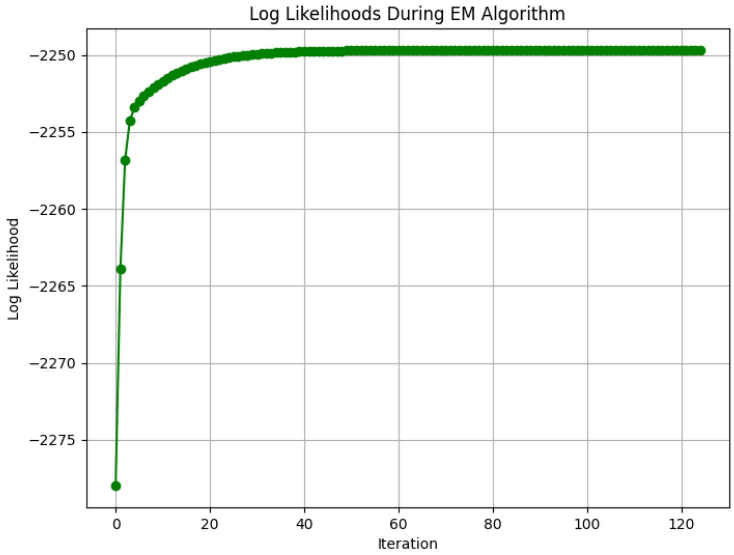

Figure 4: Log-Likelihood Convergence During EM Algorithm

Frequency Histogram with Estimated Distributions

The histogram of the dataset, overlaid with the estimated Poisson distributions, is shown below:

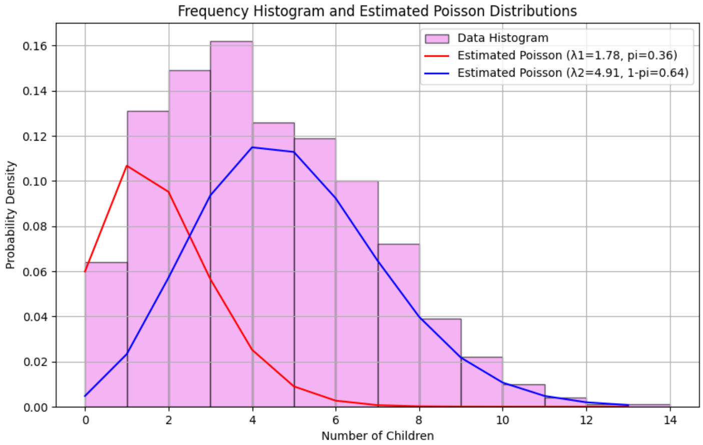

Figure 5: Frequency Histogram with Estimated Poisson Distributions

Conclusion

We implemented the PCA and the EM algorithm successfully and drew necessary plots. The results demonstrate the efficacy of these techniques for dimensionality reduction, visualization, and parameter estimation.