Machine Learning 2 - Assignment #1
Data Science Specialization
University of Antioquia

Professor: Hernán Felipe García Arias, PhD

March 5, 2025

Overview

In this assignment, you will explore clustering techniques by implementing the Gaussian Mixture Model (GMM) from scratch and applying it to two datasets. You will also implement the Elbow and Silhouette methods in combination with K-Means or GMM to evaluate the optimal number of clusters based on several performance metrics.

This assignment is divided into two tasks:

1. Implementing the Gaussian Mixture Model and evaluating clustering performance using the Rand Index.
2. Evaluating clustering quality and the optimal number of clusters using the Elbow method, Silhouette analysis, and clustering metrics.

Assignment 1: Implement Gaussian Mixture Model (GMM)

In this task, you are required to implement the GMM algorithm from scratch and apply it to the **FIFA 23 Players Dataset** and the **EastWestAirlines Dataset** for clustering. You should evaluate the performance of your clustering results using both qualitative and quantitative measures.

Datasets

- **FIFA 23 Players Dataset:** This dataset contains detailed attributes of professional soccer players. The objective is to cluster players based on their skills and playing styles. Relevant features for clustering include:
  - Age
  - Overall rating (general skill level)
  - Potential (maximum projected skill level)
  - Value (market price in €)

  - Wage (weekly salary)
  - Shooting, Passing, Dribbling (technical abilities)
  - Defending, Physicality (defensive capabilities)

  You can download the dataset from Kaggle: FIFA 23 Dataset - Kaggle.

- **EastWestAirlines Dataset:** This dataset contains information about airline customers and their behaviors. You should preprocess the dataset as necessary before applying the clustering algorithm. (Dataset available here.)

Tasks

1. Implement the Gaussian Mixture Model algorithm from scratch. Do not use libraries like `scikit-learn`'s GMM implementation for this part.
2. Fit the GMM to both datasets (FIFA 23 Players and EastWestAirlines) and perform clustering.
3. Evaluate your clustering results using the following methods:
   - **Rand Index:** Calculate the Rand Index to compare your clustering results against meaningful labels (e.g., player positions). Read more about the Rand Index here: Rand Index - Wikipedia.
   - **Qualitative Evaluation:** For both datasets, visualize and describe the resulting clusters (e.g., scatter plots, pair plots, or other visualizations that highlight the formed groups). For visualizations in Python, you may refer to this guide: Seaborn Visualization Library.

Hints and Useful Links

- For a detailed explanation on how Gaussian Mixture Models work and how to implement them, see this tutorial: Gaussian Mixture Model - scikit-learn documentation.
- To understand the mathematical background and principles behind GMM, this reference may help: Mixture Model - Wikipedia.

Assignment 2: Evaluating Clustering with Elbow and Silhouette Methods

In this task, you will evaluate the performance of clustering algorithms (K-Means or GMM) using the Elbow method and Silhouette analysis. You will also investigate the impact of clustering on four relevant metrics.

Tasks

1. Implement the Elbow method and Silhouette analysis to determine the optimal number of clusters for both the FIFA 23 and EastWestAirlines datasets.

2. Apply both K-Means and GMM clustering algorithms to each dataset and compare their performance in determining the optimal number of clusters.
3. Evaluate the clustering quality using the following four metrics:
    - **Silhouette Score:** Measures how similar a point is to its own cluster compared to other clusters. Learn more here: Silhouette Score - scikit-learn documentation.
    - **Inertia:** Sum of squared distances of points to their closest cluster center. Explanation can be found here: Inertia and Clustering Evaluation.
    - **Davies-Bouldin Index:** Measures the average similarity ratio of each cluster with its most similar cluster. Learn about it here: Davies-Bouldin Index - scikit-learn documentation.
4. Discuss how the choice of the number of clusters impacts each of the four metrics and provide a recommendation for the optimal number of clusters.

Deliverables

- Teams must submit an IPYNB notebook (Jupyter notebook) containing the complete solution to the task. This notebook should include:
    - An implementation of the Elbow method and Silhouette analysis for both datasets.
    - Clustering results using both K-Means and GMM for the FIFA 23 Players and EastWestAirlines datasets.
    - An evaluation of the clustering performance based on the four selected metrics.
    - A summary of results and analysis, including visualizations, interpretation of the results, and a discussion on how the choice of cluster number impacts clustering quality.

Grading Rubric

- **Correctness of Implementation (40%):** Accurate and complete implementation of the Gaussian Mixture Model and clustering evaluation methods.
- **Evaluation and Analysis (30%):** Quality and depth of the evaluation, including correct use of Rand Index, Silhouette, and other metrics.
- **Visualization and Reporting (20%):** Clear and informative visualizations, along with a well-structured report explaining your approach and findings.
- **Code Quality (10%):** Readability, documentation, and organization of your code.

Submission

Please submit your Jupyter notebooks in IPYNB format, including the complete code, analysis, and documentation in Markdown cells. The deadline for this assignment is 09/03/2025.