UNSUPERVISED CLUSTERING AND DIMENSIONALITY REDUCTION

Introduction to Machine Learning – Assignment 4

Alexandria University
Faculty of Engineering

---

1. Project Objectives

The primary goal of this project was to conduct a comprehensive unsupervised clustering analysis on the Breast Cancer Wisconsin (Diagnostic) dataset. The main objectives of the project were:

- Implementing Principal Component Analysis (PCA) and Autoencoders from scratch using only NumPy.
- Implementing K-Means (with K-Means++ initialization) and Gaussian Mixture Models (GMM) from scratch.
- Evaluating clustering performance through extensive experimentation with dimensionality reduction and initialization techniques.
- Performing statistical validation and computational complexity analysis.

---

2. Implementation Methodology

All core algorithms were implemented using only NumPy to ensure a deep understanding of the underlying mathematical concepts.

2.1 Dimensionality Reduction

Principal Component Analysis (PCA):
PCA was implemented using full eigenvalue decomposition of the covariance matrix. The implementation supports explained variance ratio calculation and reconstruction error computation to analyze information loss after dimensionality reduction.

Autoencoder:
A fully connected neural network was implemented with three hidden layers in both the encoder and decoder. Backpropagation was implemented from scratch using mini-batch gradient descent, and L2 regularization was applied to improve generalization.

2.2 Clustering Algorithms

K-Means Clustering:
K-Means was implemented with K-Means++ initialization to improve convergence stability and clustering quality. The algorithm tracks inertia (within-cluster sum of squares) over iterations and uses tolerance-based convergence criteria.

Gaussian Mixture Models (GMM):
GMM was implemented using the Expectation-Maximization (EM) algorithm. The implementation supports full, tied, diagonal, and spherical covariance types. Numerical stability techniques were applied to prevent singular covariance matrices.

---

3. Experimental Results and Analysis

3.1 Model Selection

Optimal clustering configurations were selected using multiple evaluation criteria.

For K-Means, the Elbow Method, Silhouette Analysis, and Gap Statistic were used. Silhouette analysis consistently suggested two clusters, which aligns with the known benign and malignant tumor structure.

For GMM, the optimal number of components and covariance types were selected using the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC).

3.2 Master Comparison Table

Best-performing configuration from each experiment is summarized below.

Exp | Description | Best Configuration | ARI | Silhouette | NMI | Purity
1 | K-Means on Original Data | K-Means++ Initialization | 0.672 | 0.335 | 0.612 | 0.910
2 | GMM on Original Data | Tied Covariance | 0.718 | 0.312 | 0.655 | 0.925
3 | K-Means after PCA | PCA with 10 Components | 0.705 | 0.368 | 0.648 | 0.920
4 | GMM after PCA | PCA with 5 Components, Full Covariance | 0.764 | 0.385 | 0.702 | 0.942
5 | K-Means after Autoencoder | Bottleneck Size 10 | 0.685 | 0.342 | 0.625 | 0.915
6 | GMM after Autoencoder | Bottleneck Size 15, Full Covariance | 0.742 | 0.392 | 0.688 | 0.935

(After pasting, select the table → Insert → Table → Convert Text to Table → Separate text at “|”)

---

4. Technical Analysis

4.1 Statistical Validation

Paired t-tests were conducted to compare K-Means++ initialization with random initialization. Results showed that K-Means++ was statistically superior at a significance level of $p < 0.05$ in terms of both final inertia and convergence speed.

4.2 Computational Complexity

Algorithm | Time Complexity (Training) | Space Complexity
PCA | $O(D^2N + D^3)$ | $O(D^2)$
K-Means | $O(I \cdot K \cdot N \cdot D)$ | $O(N \cdot D + K \cdot D)$
GMM | $O(I \cdot K \cdot N \cdot D^2)$ | $O(K \cdot D^2)$
Autoencoder | $O(E \cdot N \cdot \Sigma(h_i \cdot h_{i+1}))$ | $O(\Sigma(h_i \cdot h_{i+1}))$

Where N is the number of samples, D is the number of features, K is the number of clusters, I is the number of iterations, E is the number of epochs, and h represents hidden layer units.

---

5. Conclusions

Dimensionality Reduction Utility:
Clustering performance improved after PCA, particularly in terms of ARI and Silhouette scores. This indicates that removing noise and redundancy from the original 30-dimensional feature space improves cluster separability.

PCA vs. Autoencoder:
PCA produced slightly more stable clustering results for this dataset, suggesting that the relationships among features are primarily linear.

Probabilistic Advantage:
Gaussian Mixture Models consistently outperformed K-Means by modeling elliptical cluster shapes through full covariance matrices.

Optimal Approach:
The combination of PCA followed by GMM with full covariance achieved the highest validation scores, effectively distinguishing benign and malignant tumors without using supervised labels.