AutoEncoders & Clustering Project Report

Hemn Sheikholeslami
610399214

June 10, 2024

Abstract

This report details a project focused on utilizing AutoEncoders for dimensionality reduction of the MNIST dataset followed by clustering using multiple methods to identify the most effective technique.

1 Introduction

This project explores various clustering techniques applied to the MNIST dataset of handwritten digits post-dimensionality reduction using an AutoEncoder. The goal is to determine which clustering method segments the data into clusters that align well with actual digit labels.

2 Preprocessing

2.1 Data Loading and Preprocessing

The MNIST dataset consists of 70,000 grayscale images of digits, normalized and flattened to prepare for the AutoEncoder. This involves scaling pixel values to a range of 0 to 1 and reshaping the images from 28x28 to a vector of 784 pixels.

2.2 AutoEncoder Design and Training

An AutoEncoder with multiple layers was designed and trained to minimize reconstruction errors such as Mean Squared Error (MSE). The training aimed to optimize parameters including the number of neurons, learning rate, and batch size.

3 Model Implementation

3.1 Feature Reduction

The trained AutoEncoder effectively reduced the dimensionality of the dataset, which was crucial for the subsequent clustering processes. The efficiency of this reduction was verified by visual inspection of the reconstructed images.

4 Method Comparison

4.1 Clustering Techniques

We applied K-Means, DBSCAN, HDBSCAN, and Gaussian Mixture Model (GMM). Each method was evaluated based on silhouette scores and Adjusted Rand Index (ARI).

4.2 Evaluation of Clustering Outcomes

Among the methods, GMM provided the best results, showing high silhouette scores and ARI, indicating effective clustering according to the digit categories. The results were substantiated by visual inspection using t-SNE plots.

5 Image Reconstruction

Post-clustering, the extracted features for each cluster were used to reconstruct images using the decoder part of the AutoEncoder. This step was crucial for visually assessing the effectiveness of each clustering method.

6 Conclusion

The Gaussian Mixture Model proved most effective, suggesting that its flexible approach to cluster shape and density is particularly well-suited for complex datasets like MNIST. Future work may explore combining AutoEncoders with more advanced clustering techniques or applying the model to other datasets.

7 Report Preparation

The report is structured to include sections on Introduction, Preprocessing, Model Implementation, Method Comparison, and Conclusion. It includes comprehensive descriptions and visual evidence to support the assessments and findings.