Data Mining and Machine Learning Assignment 3

Ankan Kar (MCS202303), Sayan Bose(MCS202312)

April 21, 2024

1 Introduction

This study investigates the efficacy of semi-supervised learning techniques in conjunction with KMeans clustering for initializing the image classification process. Leveraging KMeans and MiniBatchKMeans algorithms, alongside Logistic Regression and Modified Logistic Regression with PCA dimensionality reduction, we conduct experiments on two distinct datasets: Fashion MNIST and Overhead MNIST. The aim is to identify a small subset of labeled images to bootstrap the classification task. Through a systematic evaluation, we analyze the performance of each approach in terms of classification accuracy providing insights into their suitability for semi-supervised image classification tasks.

2 Methodology

The study adopts a multifaceted methodological approach encompassing various components to ensure comprehensive analysis and equitable comparison across datasets. Maintaining consistency, the same methodology is uniformly applied to both datasets. By adhering to standardized procedures, the methodology enables meaningful interpretation of results and facilitates informed decision-making regarding the efficacy of the proposed techniques for semi-supervised learning in image classification tasks. The values of number of clusters (k) we took are 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600.

2.1 KMeans

This consists of iterating over various numbers of clusters (k) and training KMeans models on flattened training data. Representative data points for each cluster are identified, and labels are predicted for test data based on the nearest representatives. Accuracy is evaluated, and classification reports are generated to assess performance. This iterative process allows for comparison across different cluster configurations to determine the optimal setup for semi-supervised image classification.

2.2 MiniBathchKMeans

In this modified methodology utilizing MiniBatchKMeans, similar steps are followed as with KMeans. The key differences lie in the implementation of MiniBatchKMeans with specified batch size, aiming for computational efficiency. Additionally, instead of predicting labels based on the nearest representatives, the representative data points for each centroid are identified and utilized directly for

label prediction. This process remains consistent across different numbers of clusters (k), allowing for comparative analysis of performance.

2.3 Logistic Regression

This entails iterating over varying numbers of clusters (k) and sequentially performing cluster-based labeling on the training data. Subsequently, logistic regression models are initialized and trained on the flattened training data, incorporating the assigned cluster labels. Following training, predictions are made on the test data utilizing the trained logistic regression models. Finally, the performance of each model is evaluated through metrics such as accuracy score and classification report analysis, providing insights into the effectiveness of the semi-supervised learning approach.

2.4 Modified Logistic Regression

This follows an iterative process over various cluster sizes (k). Initially, cluster-based labeling assigns labels to the training data based on the identified clusters. Subsequently, logistic regression models specifically designed for multinomial classification are initialized, deviating from the binary logistic regression utilized in previous iterations. These models are then trained on the training data transformed using Principal Component Analysis (PCA), aiming to reduce dimensionality while preserving relevant information. Predictions are subsequently generated on the PCA-transformed test data. Finally, model performance is evaluated using standard metrics such as accuracy score and classification report analysis, ensuring consistent evaluation across different clustering techniques and dataset comparisons.

3 Results and Observations

The full classification report for each method for each dataset is present in the code provided. Some details are provided below:

3.1 Fashion MNIST

3.1.1 KMeans

The best accuracy we got is at k = 200 with accuracy score of 0.7528.

3.1.2 MiniBatchKMeans

The best accuracy we got is at k = 450 with accuracy score of 0.747.

3.1.3 Logistic Regression

The best accuracy we got is at k = 550 with accuracy score of 0.7751.

3.1.4 Modified Logistic Regression

The best accuracy we got is at k = 550 with accuracy score of 0.7743.

3.2 Overhead MNIST

3.2.1 KMeans

The best accuracy we got is at k = 600 with accuracy score of 0.3633.

3.2.2 MiniBatchKMeans

The best accuracy we got is at k = 600 with accuracy score of 0.3463.

3.2.3 Logistic Regression

The best accuracy we got is at k = 600 with accuracy score of 0.3227.

3.2.4 Modified Logistic Regression

The best accuracy we got is at k = 600 with accuracy score of 0.3634.

4 Conclusion

In conclusion, the comparative analysis of Fashion MNIST and Overhead MNIST datasets reveals notable differences in classification performance, predominantly influenced by intrinsic characteristics of the datasets such as image type and pixel separability. The Fashion MNIST dataset exhibits superior classification accuracy across moderately low values of k, suggesting a higher degree of separability and discernibility among image categories. Conversely, the Overhead MNIST dataset achieves its peak accuracy at a higher k value of 600, indicating potential for further improvement with higher k values. While the study's focus on low k values limits exploration of higher k values, the observed trends underscore the importance of dataset-specific considerations and parameter tuning in semi-supervised learning tasks. Thus, future investigations may benefit from exploring a broader range of k values to elucidate optimal clustering configurations for improved accuracy across diverse datasets.