

Foundation of Artificial Intelligence

Assignement 3

Davide Vitagliano

Academic Year 2023/2024

Sommario

[1. Requirements 1](#_Toc157442012)

[2. Introduction 1](#_Toc157442013)

[3. Designing the solution 2](#_Toc157442014)

[3.1. Gaussian Mixture 3](#_Toc157442015)

[3.1.1. Formalization 3](#_Toc157442016)

[3.1.2. Code explanation 4](#_Toc157442017)

[3.1.3. Conclusion 4](#_Toc157442018)

[3.2. Normalized Cut 4](#_Toc157442019)

[3.2.1. Formalization 4](#_Toc157442020)

[3.2.2. Code explanation 4](#_Toc157442021)

[3.2.3. Conclusion 5](#_Toc157442022)

[3.3. Mean Shift 5](#_Toc157442023)

[3.3.1. Formalization 5](#_Toc157442024)

[3.3.2. Code Explanation 6](#_Toc157442025)

[3.3.3. Conclusion 6](#_Toc157442026)

[4. Comparison 6](#_Toc157442027)

# Requirements

Perform clustering of the MNIST dataset using:

* Gaussian Mixture with diagonal covariance;
* Spectral Clustering with Normalized Cut;
* Mean Shift.

The clustering is performed at varying levels of dimensionality reduction through PCA from 2 to 72 with a step length of 10. PCA is used to assess the effect of the dimensionality in learning time and accuracy.

We also vary the extracted clusters as the number of clusters ‘k’ varies from 10 to 15, for the Gaussian Mixture and Normalized Cut, while for Mean Shift we vary the kernel width.

For this assignment we used GPU acceleration to increase the performance of the training process using NVIDIA CUDA Toolkit.

# Introduction

The first thing to do during the training process is applying feature selection for dimensionality reduction. This is a process of selecting a subset of the original features that are most relevant and informative for the classification. The main benefits of feature selection are:

* Reducing computational cost: fewer features means less time and space required for training and testing models;
* Improving model performance: removing irrelevant or redundant features can reduce noise and overfitting, and increase accuracy and generalization.

In our case we use Principal Components Analysis to reduce the dimensionality of the dataset by maximizing the variance of each dimension. Then we choose which model to train to classify the dataset.

Mixture of Gaussians clustering is a probabilistic model that is used to group data points into clusters based on their similarity. Each cluster is modeled as a Gaussian distribution with its own mean and covariance matrix. The algorithm estimates the parameters of these Gaussians by iteratively computing the probability of each data point belonging to each cluster and updating the parameters of the Gaussians based on these probabilities. This algorithm is useful when the data points are not clearly separable and belong to multiple clusters with different means and variances. It's commonly used in image segmentation, where the goal is to partition an image into regions that correspond to different objects or backgrounds.

Normalized Cut clustering is a technique used in multivariate statistics to perform dimensionality reduction before clustering in fewer dimensions. It’s a type of spectral clustering that makes use of the spectrum, eigenvalues, of the similarity matrix of the data to perform clustering in fewer dimensions. The similarity matrix is provided as an input and consists of a quantitative assessment of the relative similarity of each pair of points in the dataset. It’s commonly used for image segmentation and partitions points into two sets based on the eigenvector corresponding to the second-smallest eigenvalue of the symmetric normalized Laplacian matrix.

Mean Shift clustering is a clustering algorithm that can identify clusters in a dataset with arbitrary shapes and sizes. It works by updating candidates for centroids to be the mean of the points within a given region. These candidates are then filtered in a post-processing stage to eliminate near-duplicates to form the final set of centroids. The algorithm aims to discover “blobs” in a smooth density of samples. It’s a centroid-based algorithm that uses a flat kernel. Seeding, which is choosing the initial cenroids, is performed using a binning technique for scalability. Mean Shift clustering is widely used in computer vision, image processing, and bioinformatics.

In opposition of supervised learning, where we use accuracy as an indicator of performance of the classifier, in clustering we use the rand index to evaluate such performance.

For each value of ‘k’ we provide the value of the Rand index as:

Where

* *n:* it’s the number of samples in the dataset;
* *a:* it’s the number of pairs of samples that represent the same digit and that are clustered together;
* *b:* it’s the number of pairs of samples that represent different digits and are placed in different clusters.

# Designing the solution

The first portion of code is the same for every classifier, we define the parent class *Classifier*:

* *\_\_init\_\_* method:
  + It loads the MNIST dataset using the *fetch\_openml* function, which fetches datasets from the OpenML repository. The data is then normalized by dividing each feature by 255;
  + An *ATOMClassifier* object is created with the loaded data. This object acts as a wrapper to divide the dataset in training and test set.
* *\_fit\_predict\_* method:
  + The training data is transformed using PCA and then used to fit the model;
  + The test data is also transformed using PCA and then used to make predictions.
* *\_compute\_rand\_index* method:
  + Pairwise equality is calculated for the test labels and the predicted labels;
  + The Rand index is calculated using these pairwise equalities;
  + If the calculated Rand index is greater than the current best Rand index, the best parameters are updated with the current number of clusters, Rand index, and fitting and prediction times.

## Gaussian Mixture

### Formalization

Gaussian Mixture Clustering is a method of clustering that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. Here’s a detailed explanation:

* **Gaussian Mixture Model**: Suppose there are k clusters, and the mean and variance are estimated for each k. But since there are k such clusters and the probability density is defined as a linear function of densities of all these k distributions, the parameters cannot be estimated in closed form;
* **Expectation-Maximization algorithm**: The EM algorithm is an iterative way to find maximum-likelihood estimates for model parameters when the data is incomplete or has some missing data points or has some hidden variables. The algorithm chooses some random values for the missing data points and estimates a new set of data. These new values are then recursively used to estimate a better first date, by filling up missing points, until the values get fixed;
* **Estimation Step and Maximization Step**: These are the two most important steps that are iteratively performed to update the model parameters until the model convergence;
* **Clustering**: Once the parameters of the Gaussian distributions are learned, we can then cluster the samples and our error will be nearly as accurate as the Bayes optimal classifier.

### Code explanation

The class *Mixture* is a subclass of *Classifier* and implements a new method:

* *train* method:
  + it iterates over different numbers of PCA components;
  + for each PCA execution it iterates over different number of clusters;
  + for each ‘k’ clusters it calls the *\_fit\_predict* method of *Classifier* with a Gaussian Mixture with diagonal covariance and ‘k’ clusters as the model parameter.

### Conclusion

Algorithm task: multiclass classification.

Parallel processing with 12 cores.

Parallelization backend: loky

Dataset stats ==================== >>

Shape: (70000, 785)

Train set size: 60000

Test set size: 10000

-------------------------------------

Memory: 439.60 MB

Scaled: False

Outlier values: 415414 (0.9%)

| **Components** | **Clusters** | **Fit Time** | **Predict Time** | **Rand Index** |
| --- | --- | --- | --- | --- |
| 2 | 15 | 2.1631 | 0.0450 | 0.8701 |
| 12 | 14 | 3.0613 | 0.0394 | 0.8946 |
| 22 | 15 | 7.2375 | 0.0337 | 0.8960 |
| 32 | 15 | 4.0653 | 0.0411 | 0.8933 |
| 42 | 14 | 5.0723 | 0.0290 | 0.8936 |
| 52 | 12 | 4.7651 | 0.0312 | 0.8860 |
| 62 | 14 | 8.6649 | 0.0442 | 0.8749 |
| 72 | 15 | 10.4653 | 0.0294 | 0.8556 |
| 82 | 14 | 9.8508 | 0.0445 | 0.8647 |
| 92 | 15 | 7.3651 | 0.0383 | 0.8531 |
| 102 | 13 | 6.4071 | 0.0352 | 0.8504 |
| 112 | 15 | 14.0154 | 0.0388 | 0.8504 |
| 122 | 13 | 11.8794 | 0.0399 | 0.8526 |
| 132 | 13 | 9.6332 | 0.0393 | 0.8481 |
| 142 | 14 | 10.0557 | 0.0461 | 0.8577 |
| 152 | 11 | 10.6206 | 0.0374 | 0.8452 |
| 162 | 11 | 8.2323 | 0.0377 | 0.8385 |
| 172 | 14 | 16.3227 | 0.0419 | 0.8425 |
| 182 | 13 | 11.3045 | 0.1135 | 0.8491 |
| 192 | 14 | 12.7288 | 0.0477 | 0.8233 |
| 202 | 15 | 12.6883 | 0.0434 | 0.8334 |

From the table we notice that greater values of the rand index are obtained by higher number of clusters with lower value of PCA components.

## Normalized Cut

### Formalization

Normalized Cut clustering is a graph-based approach for unsupervised data analysis. Here’s a detailed explanation:

* **Graph Representation**: In this method, data points are represented as nodes of a graph. The edges between nodes represent the similarity between data points;
* **Graph Partitioning**: The goal is to partition the graph into disjoint clusters such that the similarity within a group is high and the similarity across groups is low;
* **Normalized Cut**: The normalized cut criterion measures both the total dissimilarity between the different groups as well as the total similarity within the groups. It is defined as the total weight of the edges that need to be removed to achieve the partition, normalized by the sum of the weights of the edges touching each group;
* **Optimization Problem**: The problem of finding the partition that minimizes the normalized cut is a difficult optimization problem. It is often solved within an eigenvector based framework. However, in some cases, the calculation of eigenvectors can be computationally expensive, especially when dealing with high dimensional data;
* **Relaxation and Approximation**: To make the problem tractable, it is often relaxed to a continuous domain, where it can be solved using techniques from linear algebra. The continuous solution is then discretized to obtain the final clustering5.

### Code explanation

The class *Cut* is a subclass of *Classifier* and implements two new method:

* *\_predict* method:
  + Compute the similarity between the test and training data using the Radial basis function (RBF) kernel.
  + For each test sample, find the indices of the k most similar training samples.
  + Get the labels of these k nearest neighbors.
  + For each test sample, predict the label as the most common label among its k nearest neighbors.
* *train* method:
  + it iterates over different numbers of PCA components;
  + for each PCA execution it iterates over different number of clusters;
  + for each ‘k’ clusters it fits the model to the PCA transformed training data and it calls the *\_predict* method.

### Conclusion

Algorithm task: multiclass classification.

Parallel processing with 12 cores.

Parallelization backend: loky

Dataset stats ==================== >>

Shape: (20000, 785)

Train set size: 15000

Test set size: 5000

-------------------------------------

Memory: 125.60 MB

Scaled: False

| **Components** | **Clusters** | **Fit Time** | **Predict Time** | **Rand Index** |
| --- | --- | --- | --- | --- |
| 2 | 13 | 7.1343 | 6.7567 | 0.4070 |
| 12 | 15 | 41.8378 | 6.7142 | 0.6549 |
| 22 | 13 | 22.3378 | 6.9242 | 0.6457 |
| 32 | 14 | 35.3986 | 6.6503 | 0.6673 |
| 42 | 15 | 31.3453 | 7.1892 | 0.6647 |
| 52 | 15 | 33.2318 | 6.6997 | 0.6163 |
| 62 | 15 | 29.6366 | 6.5566 | 0.6383 |
| 72 | 14 | 29.9100 | 6.7236 | 0.6456 |
| 82 | 14 | 31.6500 | 6.6321 | 0.6447 |
| 92 | 14 | 35.6481 | 6.7835 | 0.6393 |
| 102 | 14 | 32.8069 | 6.7474 | 0.6235 |
| 112 | 14 | 29.1009 | 6.6096 | 0.6384 |
| 122 | 15 | 40.8752 | 6.7151 | 0.6373 |
| 132 | 15 | 33.0893 | 6.8115 | 0.6395 |
| 142 | 15 | 40.2319 | 6.9212 | 0.6356 |
| 152 | 15 | 38.3191 | 6.7580 | 0.6379 |
| 162 | 14 | 41.7602 | 6.7023 | 0.6406 |
| 172 | 15 | 46.6879 | 6.7731 | 0.6396 |
| 182 | 15 | 38.3690 | 6.8668 | 0.6408 |
| 192 | 15 | 38.1069 | 6.9284 | 0.6375 |
| 202 | 15 | 39.8118 | 7.0768 | 0.6447 |

From the table we notice that the values of the rand index are more or less the same for every value of PCA components, but higher number of clusters.

## Mean Shift

### Formalization

Mean Shift clustering is a centroid based algorithm that aims to discover “blobs” in a smooth density of samples. Here’s a detailed explanation:

* **Initialization**: Each data point is initialized as a cluster centroid;
* **Shifting**: The algorithm works by updating candidates for centroids to be the mean of the points within a given region, also called bandwidth. Each of the sliding windows is shifted towards higher density regions by shifting their centroid to the data-points’ mean within the sliding window. This step will be repeated until no shift yields a higher density;
* **Selection of Sliding Windows**: When multiple sliding windows overlap, the window containing the most points is preserved, and the others are deleted;
* **Assigning Data Points**: The data points are assigned to the sliding window in which they reside;
* **Kernel Density Estimation**: Mean Shift is shifting the windows to a higher density region by shifting their centroid to the mean of the data-points inside the sliding window. Higher density regions correspond to regions with more samples, and lower density regions correspond to regions with fewer points;
* **Choosing the Right Bandwidth**: Depending on the bandwidth, the resulting clusters can look quite different. Therefore instead of selecting the bandwidth by hand, we can estimate it using our data.

Mean Shift is a simple cluster method that works very well on spherical-shaped data. Furthermore, it automatically selects the number of clusters contrary to other clustering algorithms. Also, the output of Mean Shift is not dependent on the initialization since, at the start, each point is a cluster. The downside to Mean Shift is that it is computationally expensive O(n²). It can be sensitive to the choice of kernel and the radius of the kernel.

### Code Explanation

The class *Shift* is a subclass of *Classifier* and implements a new method:

* *train* method:
  + it iterates over different numbers of PCA components;
  + for each PCA execution it iterates over different number of kernel bandwidth;
  + for each ‘k’ kernel bandwidth it calls the *\_fit\_predict* method of *Classifier* with a Mean Shift with ‘k’ bandwidth as the model parameter.

### Conclusion

Algorithm task: multiclass classification.

Parallel processing with 12 cores.

Parallelization backend: loky

Dataset stats ==================== >>

Shape: (70000, 785)

Train set size: 60000

Test set size: 10000

-------------------------------------

Memory: 439.60 MB

Scaled: False

Outlier values: 415414 (0.9%)

| **Components** | **Clusters** | **Fit Time** | **Predict Time** | **Rand Index** |
| --- | --- | --- | --- | --- |
| 2 | 5 | 3.3333 | 0.3640 | 0.1002 |
| 12 | 5 | 9.0991 | 0.0247 | 0.1002 |
| 22 | 5 | 10.0237 | 0.0290 | 0.4375 |
| 32 | 5 | 8.1232 | 0.0339 | 0.4706 |
| 42 | 5 | 6.2824 | 0.0345 | 0.6504 |
| 52 | 5 | 7.4788 | 0.0374 | 0.6495 |
| 62 | 5 | 7.8912 | 0.0499 | 0.6514 |
| 72 | 5 | 8.6858 | 0.0419 | 0.7451 |
| 82 | 5 | 8.0007 | 0.0506 | 0.7782 |
| 92 | 5 | 9.2022 | 0.0427 | 0.8196 |
| 102 | 5 | 9.6179 | 0.0412 | 0.7981 |
| 112 | 5 | 10.7555 | 0.0494 | 0.8325 |
| 122 | 5 | 10.9896 | 0.0526 | 0.8305 |
| 132 | 5 | 11.9510 | 0.0428 | 0.8311 |
| 142 | 5 | 13.2086 | 0.0572 | 0.8306 |
| 152 | 5 | 14.7382 | 0.0393 | 0.8321 |
| 162 | 5 | 13.3471 | 0.0346 | 0.8317 |
| 172 | 5 | 16.1631 | 0.0478 | 0.8312 |
| 182 | 5 | 16.0770 | 0.0385 | 0.8304 |
| 192 | 5 | 15.5176 | 0.0385 | 0.8315 |
| 202 | 5 | 18.0257 | 0.0363 | 0.8465 |

From the table we notice that greater values of the rand index are obtained by lower number of clusters with higher value of PCA components.

# Comparison

| **Model** | **Components** | **Clusters** | **Fit Time** | **Predict Time** | **Rand Index** |
| --- | --- | --- | --- | --- | --- |
| Gaussian Mixture | 22 | 15 | 7.2375 | 0.0337 | 0.8960 |
| Normalized Cut | 32 | 14 | 35.3986 | 6.6503 | 0.6673 |
| Mean Shift | 202 | 5 | 18.0257 | 0.0363 | 0.8465 |

Before explaining the conclusions, we keep in mind that the Normalized Cut model uses a subset of the dataset, while the other two model use the entire dataset.

From the table we can make some observations:

* Gaussian Mixture and Mean Shift models have significantly higher rand index score, with the former to be the highest;
* The fit and predict time of Normalized cut are the highest between the three models, thus we can presume that with the entire dataset this model is impracticable;
* The Mean Shift model identifies more less clusters than the other two models.

In general the Gaussian Mixture model is the best choice in terms of training time and performance.