12/01/2023

MACHINE LEARNING
3rd Task

Leisure:

Applied Informatics

Department:

Introduction to Computer Science and Technology

Theme:"Unsupervised learning – Clustering»

Club:

1. Michael Zervas, ics20015
2. Nestoras Sotiriou, ics20012
3. Christos Drikos, ics20043

HELLENIC
REPUBLIC
UNIVERSITY
OF MACEDONIA

Academic year 2022-2023

Table of Contents

Charts Table of Contents

Introduction

Based on the fashion-mnist dataset, we try to find the best clustering technique on the corresponding raw and pca projected data. Specifically, 3 different clustering techniques are used: Mini-Batch K-Means, Hierarchical clustering (Agglomerative Clustering) and Gaussian Mixture. For each of them, 4 performance indices are calculated for cluster numbers from 3 to 11: Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index and Adjusted Mutual Information, both in raw and pca projected data. Finally based on these, we find the best number for the clusters and at the end there is the best combination of data (raw or pca projected) with clustering technique.

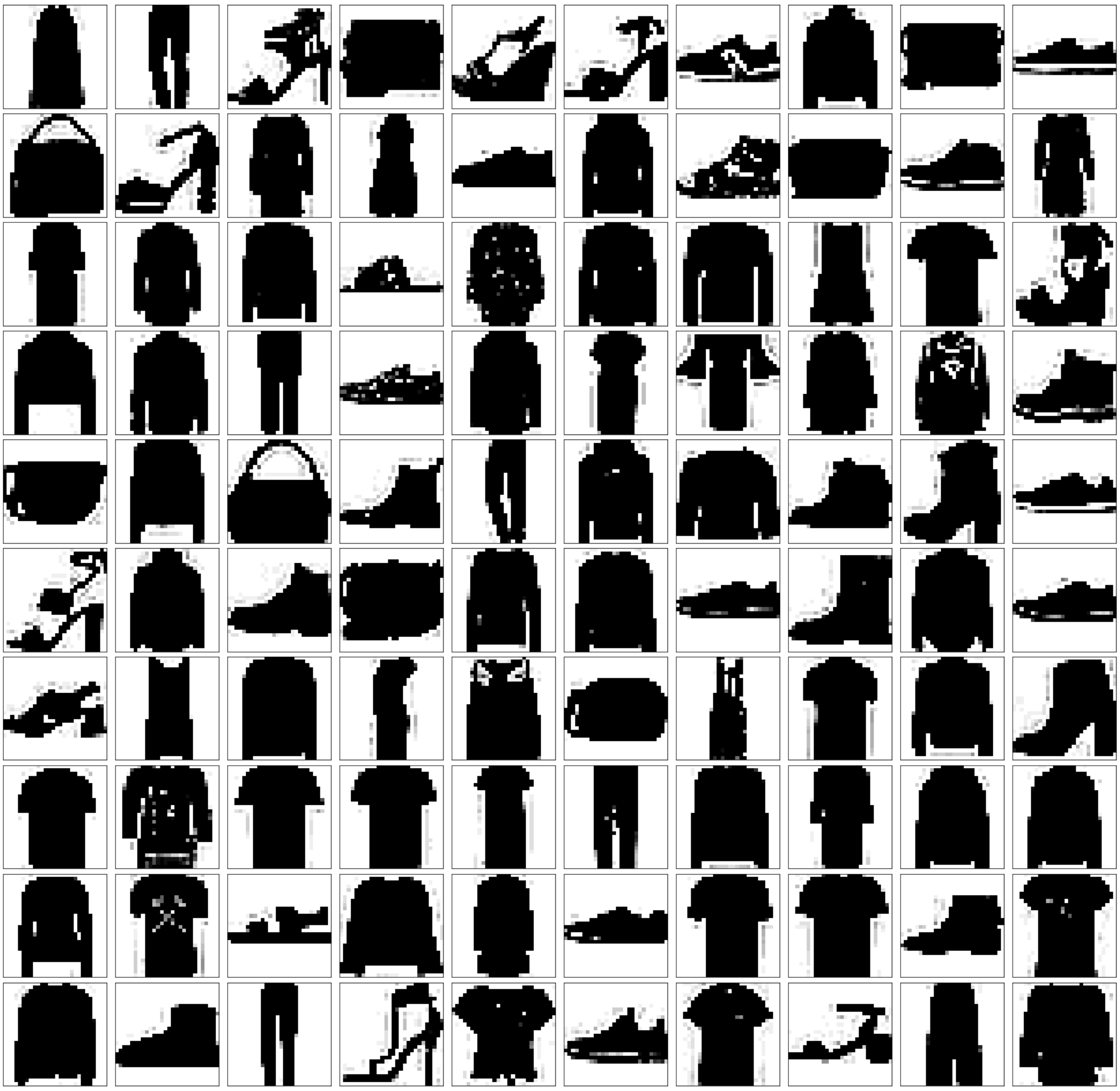
Methods applied

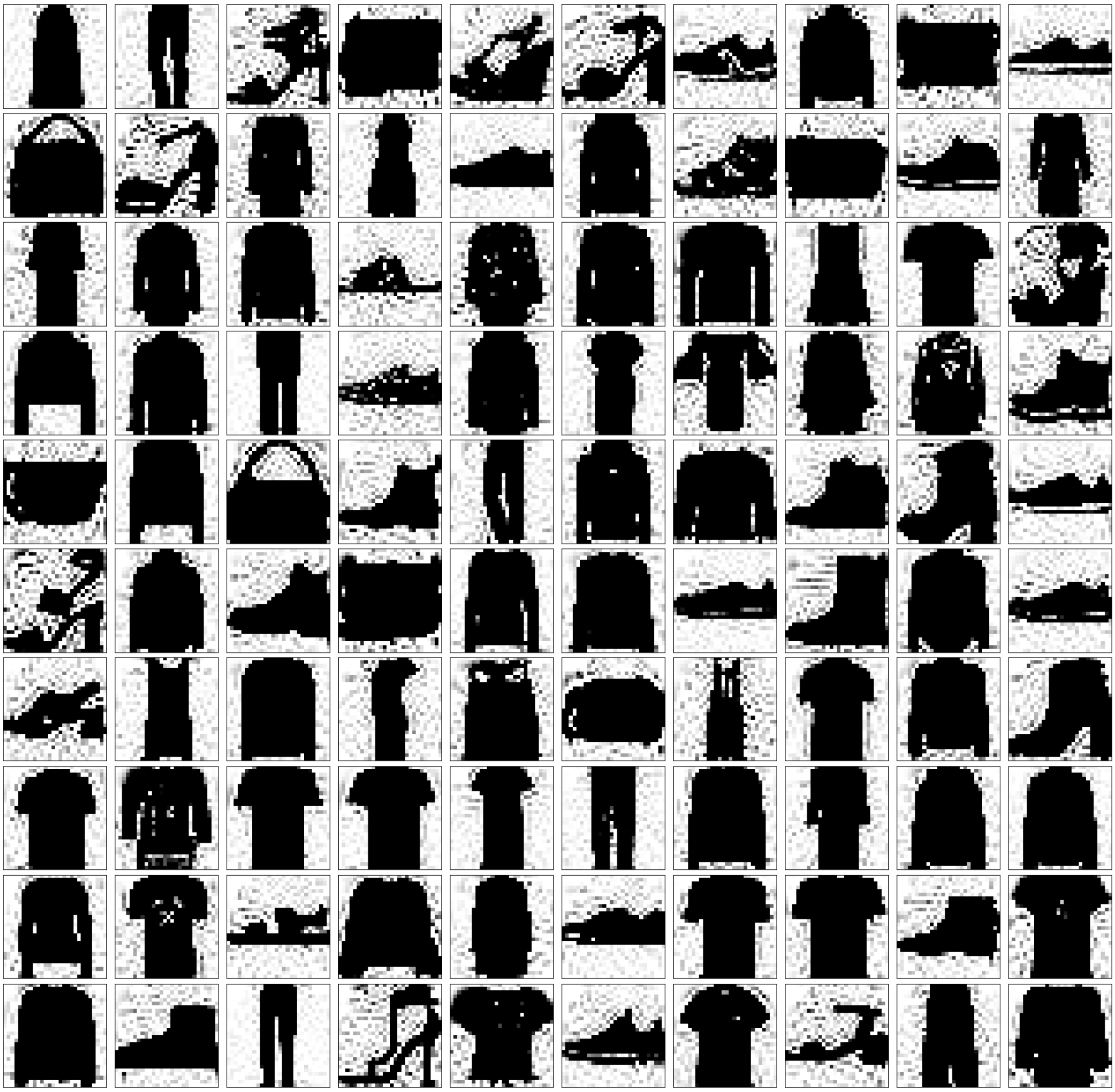
Language used: Python

1Theand 2Thequestion)
The dataset was loaded through the keras library and the separation into train and test was done through the ready function load_data() and then for validation set, a part (10%) of the train data was used.

3Thequestion)
A 99% variance was chosen for the principal components analysis.

5Thequestion)
Modified plot_digits() to ensure at least 1 item from each class is displayed. Below are the results (first the raw validation data and then the pca projected validation data):





We notice that due to the reduction of the total components, the resolution of the images in the pca projected data fell.

7Thequestion)

- **Mini-Batch K-Means:**

  The Mini-Batch K-Means algorithm is a variant of the K-Means algorithm and is designed to be more memory efficient and faster when working with large datasets. The data set is divided into smaller batches (subsets) and the k-means algorithm is applied to each mini-batch separately.

- **Hierarchical clustering (Agglomerative Clustering):**

  It is a method of clustering data points based on their similarity. The algorithm starts by treating each data point as its own cluster and then iteratively combines the closest pair of clusters until a stopping criterion is met (such as when a maximum number of clusters is reached). There are two main types of hierarchical clustering: h

Agglomerative and Divisive. Agglomerative, which is the one used here, is a bottom-up approach, in which each point starts in its own cluster and pairs of clusters merge as they move up the hierarchy.

- **Gaussian Mixture:**
  The Gaussian Mixture Model (GMM) is a probabilistic model that assumes that all data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. GMM finds the parameters of these distributions that best fit the data. The final clusters are determined by the distribution to which each data point has the highest probability of belonging.

8Thequestion)

1.**Silhouette Coefficient**: This metric measures the similarity of each data point to its own cluster compared to other clusters. Its value ranges from -1 to 1, where higher values indicate that the data point matches its own cluster well.

2.**Calinski-Harabasz Index**: This metric measures the quality of a clustering. Higher values indicate a strong separation between clusters.

3.**Davies-Bouldin Index**: This metric measures the similarity between each cluster and its most similar cluster. Lower values indicate better separation between clusters.

4.**Adjusted Mutual Information**: This metric measures the similarity between the actual labels of the data points and the labels predicted by the clustering algorithm. A higher value indicates a better match between the actual tags and the predicted tags.

Below are summary tables of metrics for each cluster number and each clustering model. In each metric, 2 clusters with the best performance in the metric are highlighted.

Scoreboards:

Mini Batch K-Means / raw data:

Number of Clusters	Silhouette Coefficient Score	Calinski-Harabasz Index Score	Davies-Bouldin Index Score	Adjusted Mutual Information Score
3	**0.12**	**3280.46**	**1.79**	0.33
4	**0.07**	**3598.25**	**2.23**	0.39
5	0.00	2731.38	11.86	0.42
6	0.01	2486.66	9.04	0.41
7	- 0.05	2193.55	7.68	0.46
8	- 0.03	1993.77	6.88	0.51
9	0.02	2212.14	7.25	**0.52**
10	- 0.11	1512.44	12.25	0.49
11	- 0.03	1931.51	12.65	**0.52**

Mini Batch K-Means / pca projected data:

Number of Clusters	Silhouette Coefficient Score	Calinski-Harabasz Index Score	Davies-Bouldin Index Score	Adjusted Mutual Information Score
3	**- 0.21**	3.53	27.26	0.00
4	**- 0.30**	6.93	11.24	0.01
5	- 0.80	7.23	2.53	0.00
6	- 0.77	2.28	**2.30**	0.00

7	- 0.81	4.47	3.25	0.02
8	- 0.80	**33.81**	**2.05**	0.03
9	- 0.61	**146.51**	13.90	**0.15**
10	- 0.66	9.52	135.16	0.06
11	- 0.79	18.55	36.50	**0.09**

Hierarchical Clustering (Agglomerative Clustering) / raw data:

Number of Clusters	Silhouette Coefficient Score	Calinski-Harabasz Index Score	Davies-Bouldin Index Score	Adjusted Mutual Information Score
3	**0.12**	**3697.54**	**1.83**	0.38
4	**0.10**	**3818.39**	**2.86**	0.44
5	- 0.00	2869.12	7.53	0.46
6	- 0.01	2497.20	13.10	0.46
7	- 0.05	2332.04	6.40	0.48
8	- 0.14	2010.39	5.87	0.48
9	- 0.13	1764.30	5.73	0.50
10	- 0.11	2060.93	8.38	**0.53**
11	- 0.08	2116.79	9.40	**0.56**

Hierarchical Clustering (Agglomerative Clustering) / PCA projected data:

Number of Clusters	Silhouette Coefficient Score	Calinski-Harabasz Index Score	Davies-Bouldin Index Score	Adjusted Mutual Information Score
3	- 0.08	2119.75	**1.15**	0.38
4	- 0.10	2038.78	4.92	0.45
5	- 0.14	1716.83	**2.23**	0.47
6	- 0.06	**2302.17**	9.64	0.52
7	**- 0.01**	**2235.80**	10.34	**0.54**
8	**- 0.02**	1958.77	23.85	0.53
9	- 0.06	1738.83	21.92	**0.54**
10	- 0.08	1578.28	6.66	**0.54**
11	- 0.20	1445.45	9.28	**0.54**

Gaussian Mixture / raw data:

Number of Clusters	Silhouette Coefficient Score	Calinski-Harabasz Index Score	Davies-Bouldin Index Score	Adjusted Mutual Information Score
3	**0.17**	**3699.58**	**1.98**	0.36
4	**0.13**	**3990.26**	2.53	0.40
5	0.03	3198.98	**2.41**	0.44
6	- 0.04	2369.35	4.63	0.45
7	- 0.08	2045.28	23.58	0.47

8	- 0.09	1939.65	5.64	0.47
9	- 0.10	1780.32	5.32	0.49
10	- 0.08	1990.49	9.95	**0.51**
11	- 0.11	1827.42	13.75	**0.52**

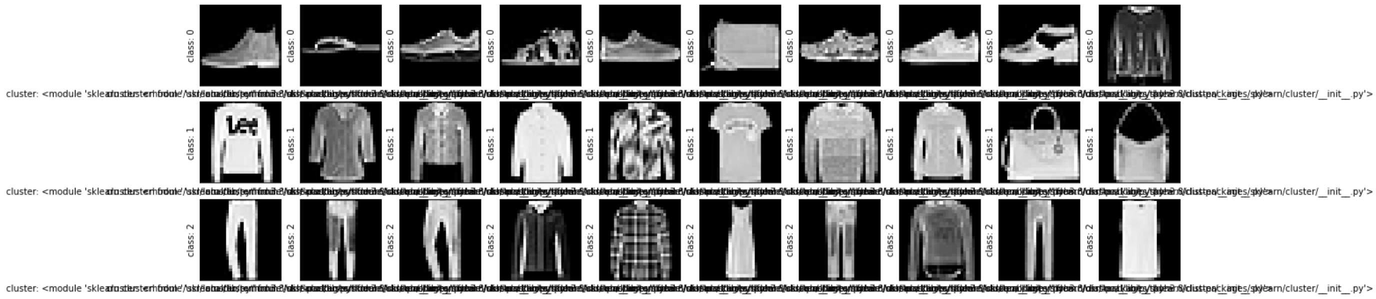
Gaussian Mixture / PCA projected data:

Number of Clusters	Silhouette Coefficient Score	Calinski-Harabasz Index Score	Davies-Bouldin Index Score	Adjusted Mutual Information Score
3	**0.09**	**3604.79**	**2.33**	0.31
4	- 0.11	1701.39	5.20	0.36
5	**- 0.10**	**1723.55**	10.79	0.43
6	- 0.26	1416.31	**4.01**	**0.44**
7	- 0.36	1629.76	117.88	0.43
8	- 0.18	1422.61	5.37	0.42
9	- 0.42	1186.93	8.41	0.42
10	- 0.27	623.63	7.97	0.42
11	- 0.35	930.36	21.35	0.46

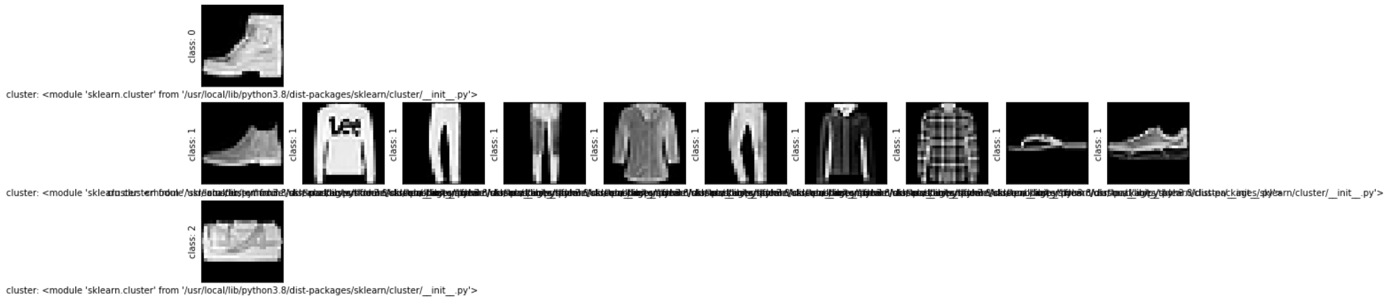
9Thequestion)

Sample clustering results for random images:

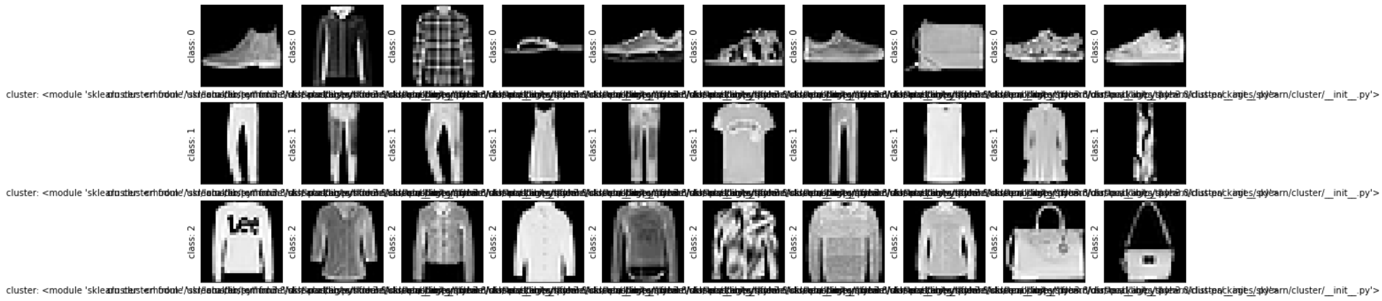
Mini Batch K-Means for raw data:



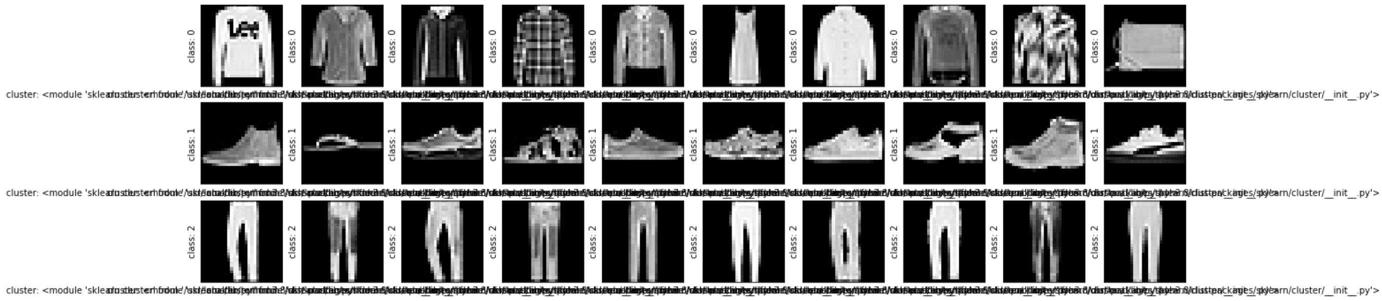
Mini Batch K-Means for PCA projected data:


class: 0
cluster: <module 'sklearn.cluster' from '/usr/local/lib/python3.8/dist-packages/sklearn/cluster/__init__.py'>
class: 1
class: 2
cluster: <module 'sklearn.cluster' from '/usr/local/lib/python3.8/dist-packages/sklearn/cluster/__init__.py'>

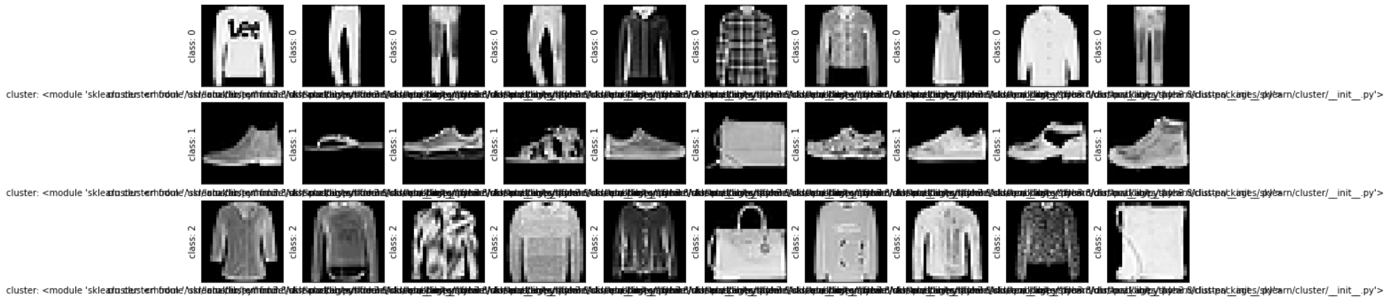
Hierarchical Clustering (Agglomerative Clustering) for raw data:



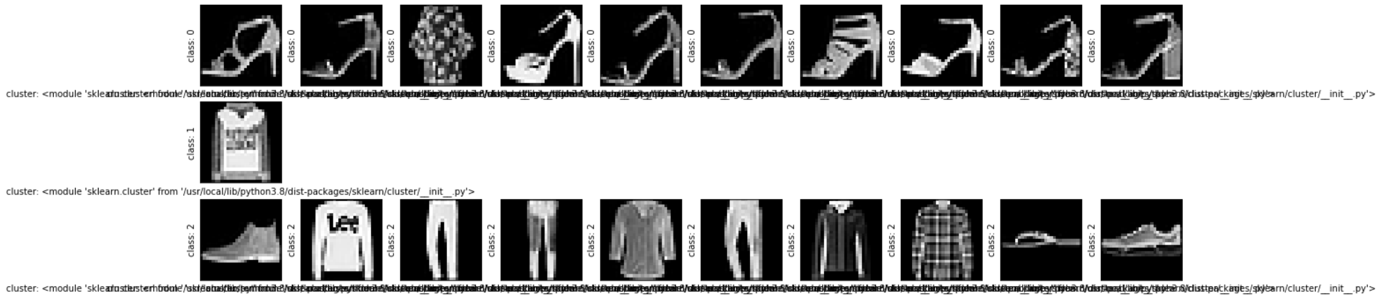
Hierarchical Clustering (Agglomerative Clustering) for PCA projected data:



Gaussian Mixture for raw data:


class: 0
class: 1
class: 2
cluster: <module 'sklearn.cluster' from '/usr/local/lib/python3.8/dist-packages/sklearn/cluster/__init__.py'>

Gaussian Mixture for PCA projected data:


class: 0
class: 1
cluster: <module 'sklearn.cluster' from '/usr/local/lib/python3.8/dist-packages/sklearn/cluster/__init__.py'>
class: 2

- ➢ *When do you have the best prices onclustering performance scores; when you have raw data or when you have pca projected data?*

  We notice that in general we have better clustering performance scores with the raw data than the PCA projected data. This is probably due to the fact that because the raw data have many dimensions (because they are images), the PCA technique may not be able to retain all the important information of the images. Also, having a total of 10 classes with clothes where each class has different features such as color, shape and others increases the complexity of the data and thus PCA may not be able to capture all the complexity present in the data, leading to poor clustering performance. Finally, the size of the images in the raw data is 28x28 pixels, which is quite small, so PCA may reduce the dimensions a lot and "lose" important information.

- ➢ *According toclustering performance scores, what is the ideal number of clusters for the dataset you have? Do you get the same result when you have raw and pca projected data?*

  Based on the clustering performance scores we find that the most ideal number of clusters for both raw and pca projected data ranges from 3 to 8. Specifically for raw data we find that the best number is 3 clusters, while for pca projected data it is 8 clusters. However, because in the pca projected data the 3 clusters have the second best performance after the 8, we keep for all the data the number 3 for the total clusters. From the tables presented previously we notice that for the same number of clusters we get different scores when we have raw data and different when we have pca projected data, with those in raw data being mainly better.

- ➢ *Whichclustering technique gives you the best results? In the end, which is the best combination of PCA or raw data / clustering approach?*

  The clustering technique with the best results, when we have 3 clusters, is the Gaussian Mixture. The best combination is the use of raw data and the Gaussian Mixture. This combination of data and clustering approach mainly shows better performance compared to the other combinations.

- ➢ *Present the clustering results of the best possible combination, based on the clustering metrics. Practically speaking, are the results acceptable? If not, what could you improve/do differently?*

The clustering results of the best possible combination, based on the cluster metrics, are shown below:

Number of Clusters	**Silhouette Coefficient Score**	**Calinski-Harabasz Index Score**	**Davies-Bouldin Index Score**	**Adjusted Mutual Information Score**
3	**0.17**	**3699.58**	**1.98**	0.36

In practice, these metrics show relatively satisfactory values. More specifically, the Silhouette Coefficient Score and the Adjusted Mutual Information Score present moderate scores, while the Calinski-Harabasz Index Score and Davies-Bouldin Index Score present a good score.
To improve the results, we could initially change the parameters with which the cluster model was initialized. Moreover, in general, we could use different clustering techniques or other numbers of clusters, so as to achieve better results.

Conclusions

In conclusion, for the fashion-MNIST dataset, the model that performs best, based on the tests performed and the performance of the corresponding metrics, is the Gaussian Mixture on the raw data and with a cluster number of 3. The results we get from this model combination and data are satisfactory but could be improved. To achieve better values in the metrics, we could either use different models, try additional possible numbers for the clusters, or initialize our models with different parameters (where possible).

Sources

https://www.geeksforgeeks.org/ml-mini-batch-k-means-clustering-algorithm/ https://www.geeksforgeeks.org/ml-hierarchical-clustering-agglomerative-and-divisive-clusteri ng/

https://scikit-learn.org/stable/modules/mixture.html