12/01/2023

# ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

## 3η Εργασία

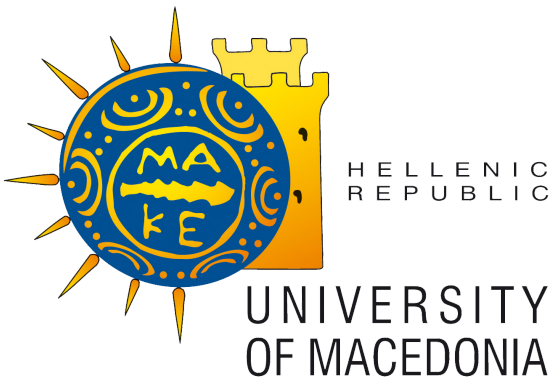

**Σχολή**:
Εφαρμοσμένης Πληροφορικής

**Τμήμα**:
Εισαγωγή στην Επιστήμη και Τεχνολογία Υπολογιστών


**Θέμα:** «Μη επιβλεπόμενη μάθηση – Συσταδοποίηση»


**Ομάδα**:

1. Μιχαήλ Ζέρβας , ics20015
2. Νέστορας Σωτηρίου , ics20012
3. Χρήστος Δρίκος , ics20043


Ακαδημαϊκό έτος 2022-2023

# Πίνακας περιεχομένων



# Πίνακας περιεχομένων γραφημάτων

# Εισαγωγή

Βάση του dataset fashion-mnist, προσπαθούμε να βρούμε την καλύτερη τεχνική συσταδοποίησης στα αντίστοιχα raw και στα pca projected data. Συγκεκριμένα γίνεται χρήση 3 διαφορετικών τεχνικών συσταδοποίησης: Mini-Batch K-Means, Hierarchical clustering (Agglomerative Clustering) και Gaussian Mixture. Για κάθε μία από αυτές υπολογίζονται και 4 δείκτες απόδοσης για αριθμούς cluster από 3 έως 11: Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index και Adjusted Mutual Information, τόσο στα raw όσο και στα pca projected data. Τελικά βάση αυτών, βρίσκουμε τον καλύτερο αριθμό για τα cluster και στο τέλος βρίσκεται ο καλύτερος συνδυασμός data (raw ή pca projected) με τεχνική συσταδοποίησης.

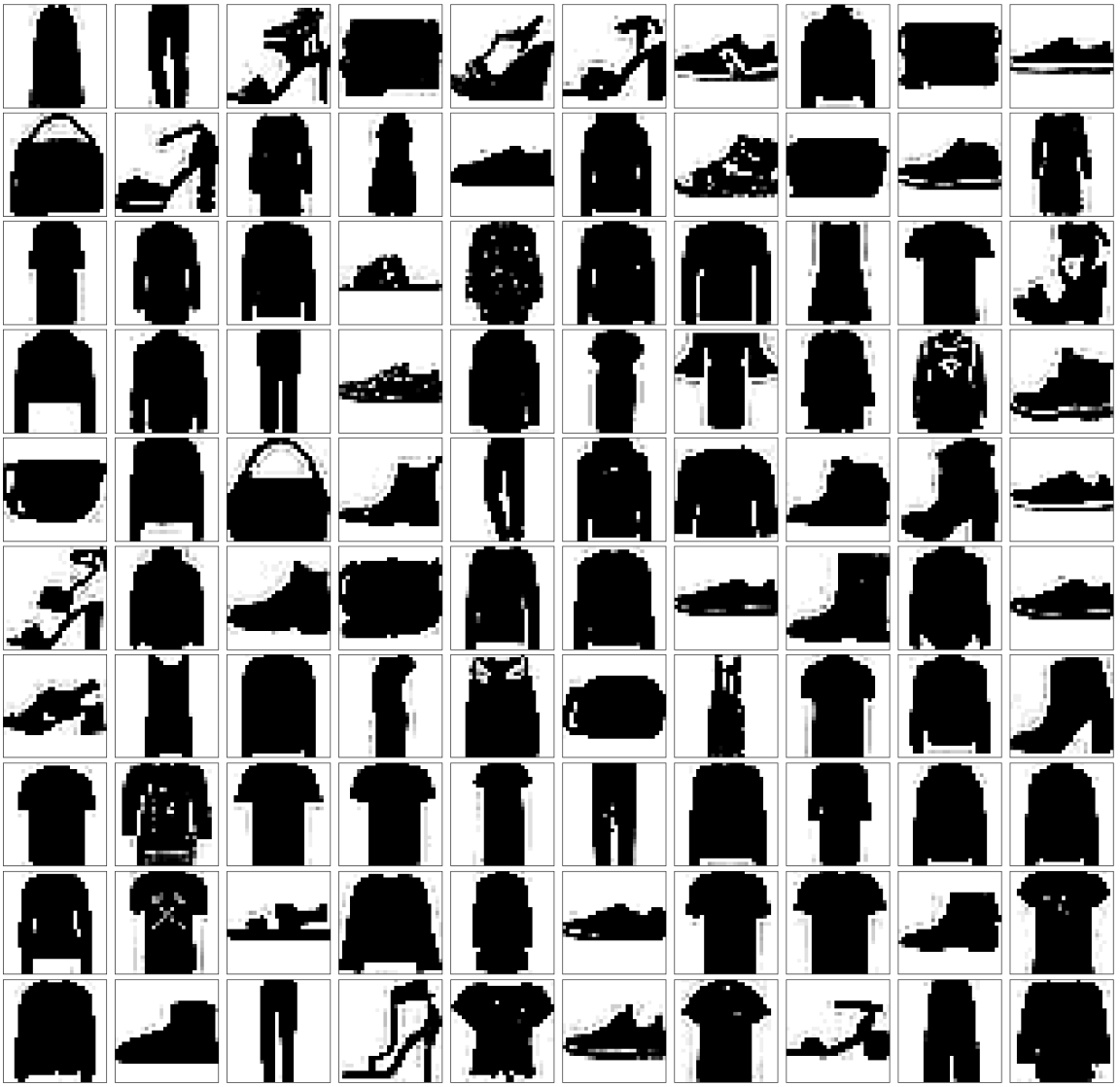
# Μέθοδοι που εφαρμόστηκαν

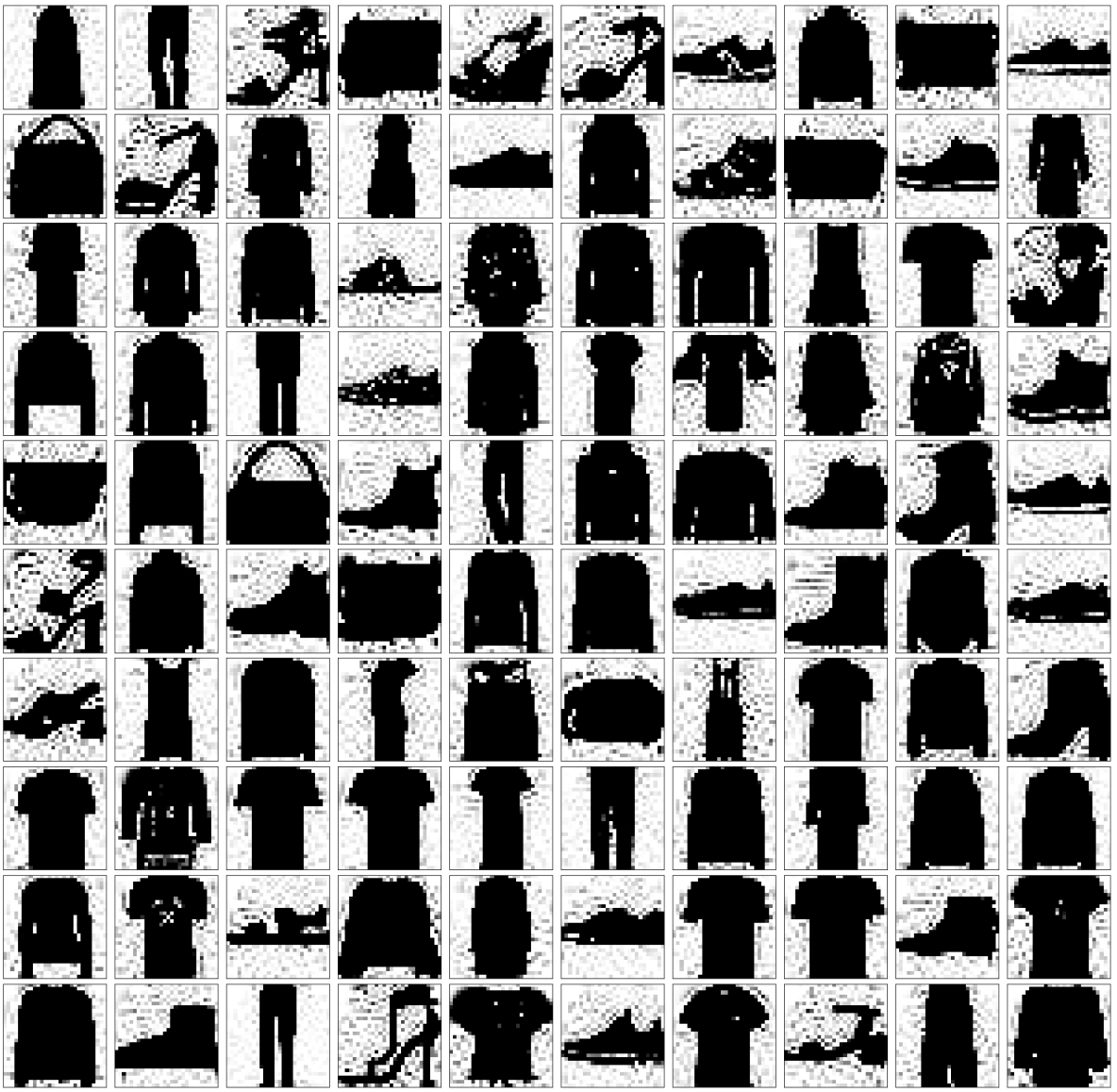
Γλώσσα που χρησιμοποιήθηκε: Python

1ο και 2ο ερώτημα)
Το dataset φορτώθηκε μέσα από την βιβλιοθήκη keras και ο διαχωρισμός σε train και test έγινε έγινε μέσω της έτοιμης συνάρτησης load_data() και στη συνέχεια για validation set, χρησιμοποιήθηκε ένα μέρος (10%) του train data.

3ο ερώτημα)
Επιλέχθηκε 99% ποσοστό διασποράς για την ανάλυση σε κύριες συνιστώσες.

5ο ερώτημα)
Τροποποιήθηκε η plot_digits() ώστε να εξασφαλίζει την εμφάνιση τουλάχιστον 1 στοιχείου από κάθε κλάση. Παρακάτω φαίνονται τα αποτελέσματα (πρώτα τα raw validation data και μετά τα pca projected validation data):





Παρατηρούμε πως λόγω της μείωσης των συνολικών συνιστωσών, έπεσε η ανάλυση των εικόνων στα pca projected data.

7ο ερώτημα)

- **Mini-Batch K-Means:**
  Ο αλγόριθμος Mini-Batch K-Means είναι μια παραλλαγή του αλγορίθμου K-Means και έχει σχεδιαστεί για να είναι πιο αποδοτικός στη μνήμη και πιο γρήγορος όταν εργάζεται με μεγάλα σύνολα δεδομένων. Το σύνολο δεδομένων χωρίζεται σε μικρότερα batches (υποσύνολα) και ο αλγόριθμος k-means εφαρμόζεται σε κάθε mini-batch ξεχωριστά.

- **Hierarchical clustering (Agglomerative Clustering):**
  Είναι μια μέθοδος συσταδοποίησης σημείων δεδομένων με βάση την ομοιότητά τους. Ο αλγόριθμος ξεκινά αντιμετωπίζοντας κάθε σημείο δεδομένων ως το δικό του cluster και στη συνέχεια συνδυάζει επανειλημμένα το πλησιέστερο ζεύγος συστάδων μέχρι να ικανοποιηθεί ένα κριτήριο διακοπής (όπως όταν επιτυγχάνεται ένας μέγιστος αριθμός συστάδων). Υπάρχουν δύο κύριοι τύποι ιεραρχικής συσταδοποίησης: η

Agglomerative και η Divisive. Η Agglomerative, που είναι και αυτή που χρησιμοποιήθηκε εδώ, είναι μια προσέγγιση "από κάτω προς τα πάνω" (“bottom-up”), κατά την οποία κάθε σημείο ξεκινά στη δική του συστάδα και τα ζεύγη συστάδων συγχωνεύονται καθώς ανεβαίνουν στην ιεραρχία.

- **Gaussian Mixture:**
  Το Gaussian Mixture Model (GMM) είναι ένα πιθανοτικό μοντέλο που υποθέτει ότι όλα τα σημεία δεδομένων παράγονται από ένα μείγμα πεπερασμένου αριθμού κατανομών Gauss με άγνωστες παραμέτρους. Το GMM βρίσκει τις παραμέτρους αυτών των κατανομών που ταιριάζουν καλύτερα στα δεδομένα. Οι τελικές συστάδες καθορίζονται από την κατανομή στην οποία κάθε σημείο δεδομένων έχει τη μεγαλύτερη πιθανότητα να ανήκει.

$8^{ο}$ ερώτημα)

1. **Silhouette Coefficient**: Αυτή η μετρική μετρά την ομοιότητα κάθε σημείου δεδομένων με τη δική του συστάδα σε σύγκριση με άλλες συστάδες. Η τιμή της κυμαίνεται από -1 έως 1, όπου υψηλότερες τιμές υποδηλώνουν ότι το σημείο δεδομένων ταιριάζει καλά με τη δική του συστάδα.

2. **Calinski-Harabasz Index**: Αυτή η μετρική μετρά την ποιότητα μιας συσταδοποίησης. Υψηλότερες τιμές υποδηλώνουν έναν ισχυρό διαχωρισμό μεταξύ των συστάδων.

3. **Davies-Bouldin Index**: Αυτή η μετρική μετρά την ομοιότητα μεταξύ κάθε συστάδας και την πιο παρόμοιας σε αυτήν συστάδας. Χαμηλότερες τιμές υποδηλώνουν καλύτερο διαχωρισμό μεταξύ των συστάδων.

4. **Adjusted Mutual Information**: Αυτή η μετρική μετρά την ομοιότητα μεταξύ των πραγματικών ετικετών των σημείων δεδομένων και των ετικετών που προβλέφθηκαν από τον αλγόριθμο συσταδοποίησης. Υψηλότερη τιμή υποδηλώνει καλύτερη αντιστοιχία μεταξύ των πραγματικών ετικετών και των προβλεπόμενων ετικετών.

Παρακάτω παρατίθενται συγκεντρωτικοί πίνακες μετρικών για κάθε αριθμό cluster και για κάθε μοντέλο συσταδοποίησης. Σε κάθε μετρική υπογραμμίζονται 2 clusters με καλύτερη απόδοση στη μετρική.

## Πίνακες αποτελεσμάτων:

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** |

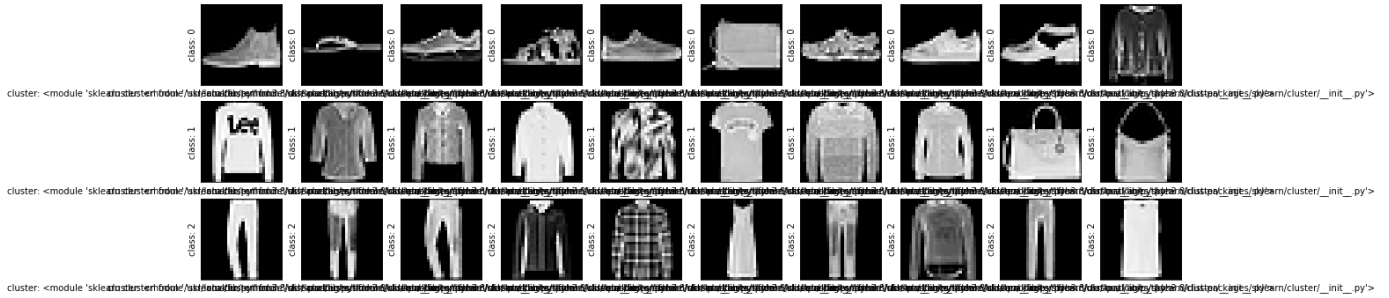
Gaussian Mixture / PCA projected data:

| 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 |

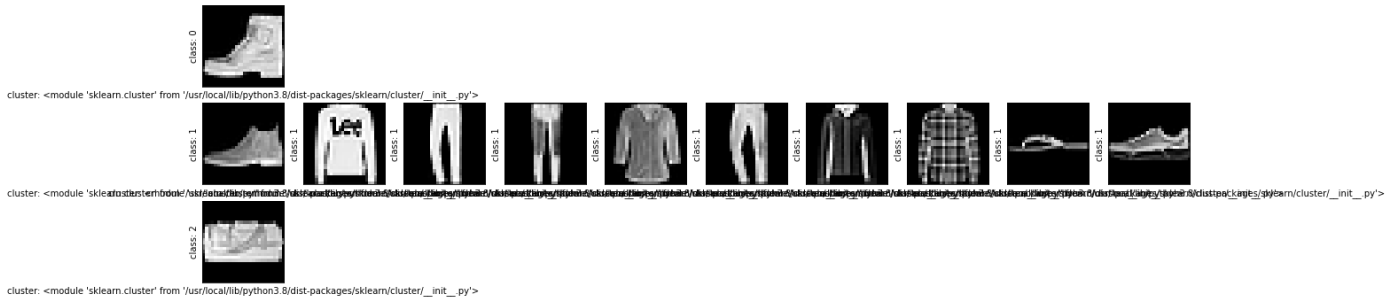
9ο ερώτημα)

Ενδεικτικά αποτελέσματα ομαδοποίησης για τυχαίες εικόνες:

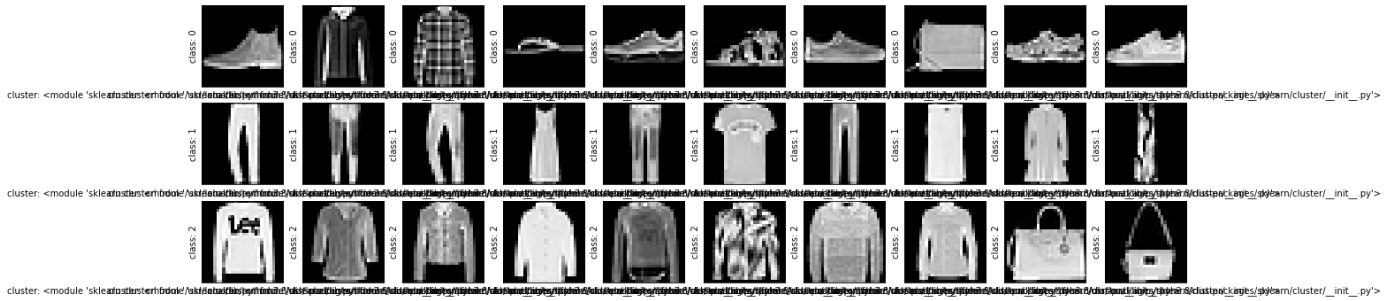
Mini Batch K-Means για raw data:



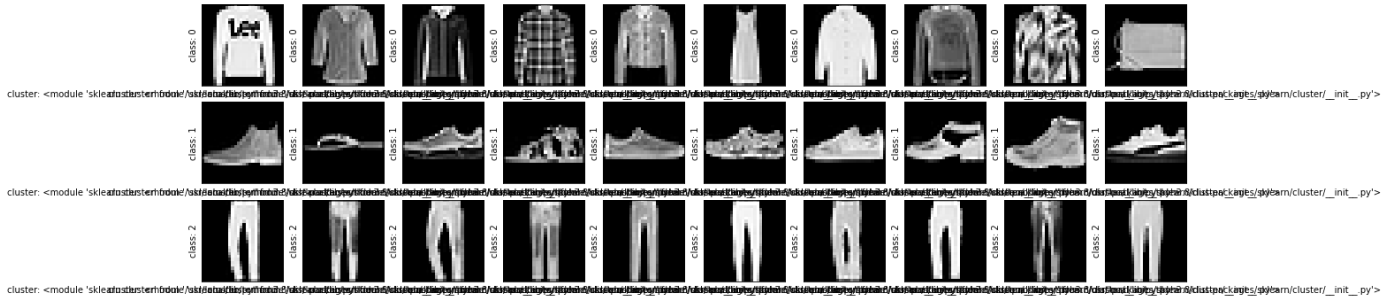
Mini Batch K-Means για PCA projected data:



Hierarchical Clustering(Agglomerative Clustering) για raw data:

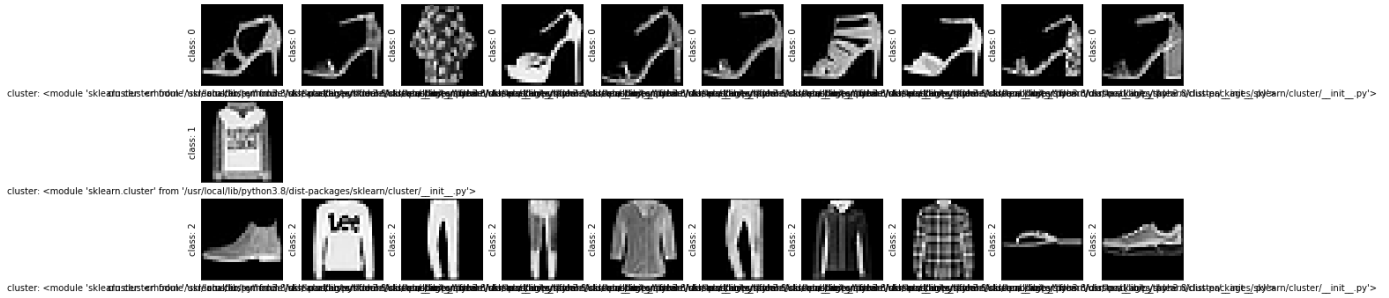


Hierarchical Clustering(Agglomerative Clustering) για PCA projected data:



Gaussian Mixture για raw data:

Gaussian Mixture για PCA projected data:



➢ *Πότε έχετε καλύτερες τιμές στα clustering performance scores; όταν έχετε raw data ή όταν έχετε pca projected data;*

Παρατηρούμε ότι γενικά έχουμε καλύτερα clustering performance scores με τα raw data από τα PCA projected data. Αυτό πιθανόν οφείλεται στο γεγονός ότι επειδή τα raw data έχουν πολλές διαστάσεις (επειδή είναι εικόνες), η τεχνική του PCA ίσως δεν είναι ικανή να διατηρήσει όλες τις σημαντικές πληροφορίες των εικόνων. Επίσης το ότι υπάρχουν συνολικά 10 κλάσεις με ρούχα όπου κάθε κλάση έχει διαφορετικά χαρακτηριστικά όπως χρώμα, σχήμα και άλλα, αυξάνει την πολυπλοκότητα των δεδομένων και έτσι το PCA ενδέχεται να μην είναι σε θέση να καταγράψει όλη την πολυπλοκότητα που υπάρχει στα δεδομένα, οδηγώντας σε κακή απόδοση συσταδοποίησης. Τέλος το μέγεθος των εικόνων στα raw data είναι 28x28 pixels το οποίο είναι αρκετά μικρό, οπότε το PCA ίσως μειώνει πολύ τις διαστάσεις και “χάνει” σημαντικές πληροφορίες.

➢ *Με βάση τα clustering performance scores, ποιος είναι ο ιδανικός αριθμός clusters για το dataset που έχετε; Παίρνετε το ίδιο αποτέλεσμα όταν έχετε raw και pca projected data;*

Βάση των clustering performance scores βρίσκουμε ότι οι ιδανικότεροι αριθμοί clusters τόσο για τα raw όσο και για τα pca projected data κυμαίνονται στο διάστημα 3 έως 8. Συγκεκριμένα για τα raw data βρίσκουμε ότι ο καλύτερος αριθμός είναι 3 clusters, ενώ για τα pca projected data είναι 8 clusters. Επειδή όμως στα pca projected data τα 3 clusters έχουν την δεύτερη καλύτερη επίδοση μετά τα 8, κρατάμε για όλα τα data τον αριθμό 3 για τα συνολικά clusters. Από τους πίνακες που παρουσιάστηκαν προηγουμένως παρατηρούμε ότι παίρνουμε, για τον ίδιο αριθμό clusters παίρνουμε διαφορετικά scores όταν έχουμε raw data και διαφορετικά όταν έχουμε pca projected data, με αυτά στα raw data να είναι κατά κύριο λόγο καλύτερα.

- ➢ *Ποια clustering τεχνική σας δίνει τα καλύτερα αποτελέσματα; Τελικά ποιος είναι ο καλύτερος συνδυασμός PCA ή raw data / clustering approach;*

  Η clustering τεχνική με τα καλύτερα αποτελέσματα,όταν έχουμε 3 clusters, είναι η Gaussian Mixture. Ο καλύτερος συνδυασμός είναι η χρήση των raw data και την Gaussian Mixture. Αυτός ο συνδυασμός data και clustering approach παρουσιάζει κατά κύριο λόγο καλύτερες αποδόσεις σε σχέσει με τους υπόλοιπους συνδυασμούς.

- ➢ *Παρουσιάστε τα αποτελέσματα ομαδοποίησης του καλύτερου δυνατού συνδυασμού, με βαση τα clustering metrics. Πρακτικά μιλώντας είναι τα αποτελέσματα αποδεκτά; Αν όχι, τι θα μπορούσατε να βελτιώσετε/κάνετε διαφορετικά;*

Τα αποτελέσματα ομαδοποίησης του καλύτερου δυνατού συνδυασμού, με βάση τα cluster metrics, φαίνονται παρακάτω:

| 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 |

Πρακτικά, οι μετρικές αυτές παρουσιάζουν σχετικά ικανοποιητικές τιμές. Πιο συγκεκριμένα το Silhouette Coefficient Score και το Adjusted Mutual Information Score παρουσιάζουν μέτριες βαθμολογίες, ενώ τα Calinski-Harabasz Index Score και Davies-Bouldin Index Score παρουσιάζουν καλή βαθμολογία.
Για την βελτίωση των αποτελεσμάτων θα μπορούσαμε αρχικά να αλλάζαμε τις παραμέτρους με τις οποίες αρχικοποιήθηκε το cluster model. Επιπλέον, γενικότερα, θα μπορούσαμε να χρησιμοποιήσουμε διαφορετικές τεχνικές συσταδοποίησης ή και άλλους αριθμούς clusters, έτσι ώστε να επιτύχουμε καλύτερα αποτελέσματα.

# Συμπεράσματα

Συμπερασματικά, για το fashion-MNIST dataset, το μοντέλο που αποδίδει καλύτερα, βάση των δοκιμών που έγιναν και των αποδόσεων των αντίστοιχων μετρικών, είναι το Gaussian Mixture στα raw data και με αριθμό cluster τα 3. Τα αποτελέσματα που λαμβάνουμε από αυτόν τον συνδυασμό μοντέλου και δεδομένων είναι ικανοποιητικά αλλά θα μπορούσαν να βελτιωθούν. Για την επίτευξη καλύτερων τιμών στις μετρικές, θα μπορούσαμε να χρησιμοποιήσουμε είτε διαφορετικά μοντέλα, είτε να δοκιμάσουμε επιπλέον πιθανούς αριθμούς για τα cluster, είτε να αρχικοποιήσουμε τα μοντέλα μας με διαφορετικές παραμέτρους (όπου αυτό είναι δυνατό).

# Πηγές

https://www.geeksforgeeks.org/ml-mini-batch-k-means-clustering-algorithm/
https://www.geeksforgeeks.org/ml-hierarchical-clustering-agglomerative-and-divisive-clustering/
https://scikit-learn.org/stable/modules/mixture.html