Clustering Report

MNIST & CIFAR10

Περιγραφή: 3rd_Project.ipynb - Colab

Σκοπός της εργασίας είναι αφού μειώσουμε τις διαστάσεις των δύο dataset (Mnist & Cifar10) χρησιμοποιώντας Isomap, να κάνουμε ομαδοποίηση (clustering) για διαφορετικό αριθμό ομάδων (clusters), να αξιολογήσουμε την ικανότητα του αλγορίθμου να ομαδοποιεί τα δεδομένα της ίδιας κλάσης σε διακριτές συστάδες και να χρησιμοποιήσουμε τα αποτελέσματα της ομαδοποίησης (clustering) για την ταξινόμηση, ώστε να αξιολογήσουμε την ακρίβεια της μεθόδου.

Datasets:

Τα dataset που θα χρησιμοποιήσουμε είναι το Mnist dataset με τα χειρόγραφα ψηφία από 0 έως 9 και το Cifar10, που περιλαμβάνει 10 διαφορετικές κλάσεις αντικειμένων (σκύλος, γάτα, αυτοκίνητο,..κλπ).

Import Libraries:

Εισάγουμε τις απαραίτητες βιβλιοθήκες που θα χρησιμοποιήσουμε.

Dimensionality Reduction «Isomap»:

Εφαρμόζουμε τη μέθοδο Isomap για μείωση σε 2 διαστάσεις. Η συνάρτηση με το όνομα f«dimentionality_reduction_train», παίρνει ως όρισμα 2 μεταβλητές: α) ένα σύνολο δεδομένων X (train set: Mnist & Cifar10), β) τον αριθμό των n_components, τον οποίο ορίζουμε ίσο με 2

Μέσα στη συνάρτηση εκτελείται η μέθοδο Isomap και έπειτα η μέθοδος fit_transfrom() εφαρμόζεται στα δεδομένα X train και επιστρέφει τα δεδομένα σε 2 διαστάσεις.

Αντίστοιχα δημιουργούμε μία ξεχωριστή μέθοδο για τα δεδομένα του test set, μόνο που σε αυτή την περίπτωση εφαρμόζεται η μέθοδος transform.

Apply Spectral Clustering:

Η συνάρτηση «perform_clustering» εκτελεί spectral clustering σε ένα σύνολο δεδομένων X train με βάση τον αριθμό των n_clusters, και τον αριθμό των γειτόνων n_neighbors, οι οποίοι καθορίζονται από το χρήστη. Χρησιμοποιούμε το «SpectralClustering» από την scikit-learn.

Η παράμετρος affinity καθορίζει τον τρόπο με τον οποίο υπολογίζεται η ομοιότητα μεταξύ των δεδομένων και ορίζεται ως ”nearest_neighbors”, δηλαδή ο υπολογισμός της ομοιότητας γίνεται με βάση τις κοντινές γειτονιές κάθε σημείου.

Η παράμετρος assign_labels καθορίζει τη στρατηγική που θα ακολουθήσει ο αλγόριθμος για να αναθέσει τις ετικέτες.

Η μέθοδος fit_predict() εφαρμόζεται στα δεδομένα X train και επιστρέφει τις ετικέτες που δηλώνουν σε ποιο cluster ανήκει κάθε δείγμα.

Visualization of Clusters

Η μέθοδος plot_side_by_side εμφανίζει 2 διαγράμματα το ένα δίπλα στο άλλο. Το αριστερό διάγραμμα απεικονίζει τα δεδομένα (X train set) με τις πραγματικές τους ετικέτες (labels), ενώ το δεξί διάγραμμα απεικονίζει τα δεδομένα (X train set) με τις ετικέτες (labels) που έχουν προκύψει από τη διαδικασία του clustering. Τα δεδομένα απεικονίζονται σε δύο διαστάσεις με διαφορετικά χρώματα για κάθε ομάδα.

Hyper-Parameter Tuning

Η συνάρτηση Tuning_Clustering πραγματοποιεί την αναζήτηση υπερπαραμέτρων για τη μέθοδο Spectral Clustering. Η συνάρτηση παίρνει ως παραμέτρους τα δεδομένα X train, τα true_labels που είναι οι πραγματικές ετικέτες των δεδομένων, τα n_clusters, που ορίζονται κάθε φορά από εμάς και τα n_neighbors_values, που είναι μία λίστα με διαφορετικούς αριθμούς γειτόνων που χρησιμοποιούνται για το spectral clustering.

Με τη χρήση της συνάρτησης Tuning_Clustering θέλουμε να βρούμε την καλύτερη τιμή του n_neighbors για το spectral clustering.

Αρχικοποιούμε το λεξικό best_neighbor, το οποίο περιλαμβάνει τη λίστα n_neighbor_values με όλες τις τιμές των γειτόνων που έχουμε ορίσει και δημιουργούμε ένα πίνακα results=[], που θα κρατάει τα αποτελέσματα για κάθε τιμή του n_neighbors.

Με μία for loop() διατρέχουμε όλες τις τιμές των γειτόνων που περιέχει η λίστα n_neighbor_values. Στη συνέχεια καλούμε τη συνάρτηση spectral_clustering για να εφαρμόσουμε τον αλγόριθμο spectral clustering στα δεδομένα X με τον προκαθορισμένο αριθμό clusters και την τρέχουσα τιμή του n_neighbors.

Υπολογίζουμε τις μετρικές ARI (Adjusted Rand Index) και NMI (Normalized Mutual Info Score), οι οποίες αξιολογούν την ποιότητα του clustering.

1. Adjusted Rand Index (ARI): Μετράει την ομοιότητα των ετικετών που προκύπτουν από το clustering με τις πραγματικές ετικέτες (True Labels) - (αποκλείει την πιθανότητα να υπάρχει συμφωνία μεταξύ των ετικετών λόγω τυχαιότητας).
2. Normalized Mutual Information: Μετράει την αμοιβαία πληροφορία μεταξύ των ετικετών που προκύπτουν από το clustering και των πραγματικών ετικετών.

Για την αξιολόγηση της καλύτερης παραμέτρους (n_neighbors) χρησιμοποιούμε τη μετρική ARI. Επομένως, ο γείτονας που θα δώσει το καλύτερο ARI score θα επιλεγεί ως η καλύτερη παράμετρος.

Top 3 Labels in every Cluster

Η συνάρτηση find_top_labels παίρνει ως παραμέτρους τα δεδομένα X train, τα labels που έχουν προκύψει από τα cluster, τα πραγματικά labels και μία μεταβλητή top_n, η οποία δηλώνει τον αριθμό των κορυφαίων labels.

Με το unique_labels = np.unique(labels) βρίσκουμε τα μοναδικά clusters, που έχουν προκύψει από το clustering. Στη συνέχεια δημιουργούμε μία κενή λίστα top_labels_summary=[], η οποία θα περιέχει τα κορυφαία labels και τα ποσοστά τους.

Με μία for loop() διατρέχουμε τον πίνακα με τα unique labels και βρίσκουμε τις θέσεις των δειγμάτων που ανήκουν στο ίδιο cluster. Στη συνέχεια εξάγουμε τα πραγματικά labels αυτών των δειγμάτων.

Με την bincount() υπολογίζουμε τη συχνότητα των πραγματικών ετικετών μέσα σε ένα cluster .

Τέλος ταξινομούμε τα labels κατά φθίνουσα σειρά, παίρνουμε τα top_n και υπολογίζουμε το ποσοστό των top labels στο cluster.

Classify Test Data based on Cluster Centroid

Η συνάρτηση classify_with_clusters κατηγοριοποιεί τα δεδομένα ελέγχου X_test με βάση τα κέντρα (centroids) των clusters, που προκύπτουν από τη διαδικασία του clustering υπολογίζονται από τα δεδομένα εκπαίδευσης.

Οι παράμετροι της συνάρτησης είναι: a) τα δεδομένα εκπαίδευσης X train, b) Οι ετικέτες που αντιστοιχούν στα clusters για τα δεδομένα εκπαίδευσης, c) τα δεδομένα του test set που θέλουμε να ταξινομήσουμε και d) οι πραγματικές ετικέτες για τα δεδομένα δοκιμής (test set).

Εξάγουμε τα unique labels από την λίστα με τα cluster_labels, τα οποία αντιπροσωπεύουν τα διαφορετικά clusters.

Υπολογίζουμε τα κέντρα των clusters με μία for loop() με την εξής διαδικασία:

Διατρέχει όλα τα unique labels που έχουμε εξάγει και κάθε φορά και στη μεταβλητή group_points κρατάμε τα δείγματα του X train set, τα οποία ανήκουν όλα στο ίδιο cluster. Έπειτα υπολογίζουμε το μέσο όρο των δεδομένων κατά στήλη για κάθε διάσταση του dataset(group_points.mean(axis=0)) και το προσθέτουμε στον προδημιουργημένο πίνακας (centroids=[]). Στη συνέχεια ταξινομούμε τα δεδομένα δοκιμής, υπολογίζοντας την ευκλείδεια απόσταση κάθε σημείου από κάθε κέντρο cluster (np.linalg.norm). Με το np.argmin() επιστρέφουμε το index του κεντροειδούς που βρίσκεται πιο κοντά στο δεδομένο δοκιμής . Τέλος αποθηκεύουμε το label (index του πλησιέστερου centroid) στη λίστα labels_test και υπολογίζουμε την ακρίβεια της κατηγοριοποίησης.

Εφαρμογή

Δοκιμάζουμε να κάνουμε ομαδοποίηση των δεδομένων των δύο dataset Mnist & Cifar10 με διαφορετικό αριθμό cluster = 5, 10, 15. Προφανώς και γνωρίζουμε ότι τα πραγματικά labels των dataset είναι 10, ωστόσο μας ενδιαφέρει να δούμε με την μέθοδο του clustering πως πραγματοποιείται ο διαχωρισμός. Και για τα δύο dataset χρησιμοποιούμε ένα υποσύνολο του συνολικού dataset που περιλαμβάνει τα πρώτα 25.000 δείγματα.

Για Cluster = 5:

Το Mnist dataset (best n_neighbors= 200) (metrics: ARI= 0.1785, NMI= 0.312).

Top Labels για κάθε Cluster: Classification Accuracy = 0.29

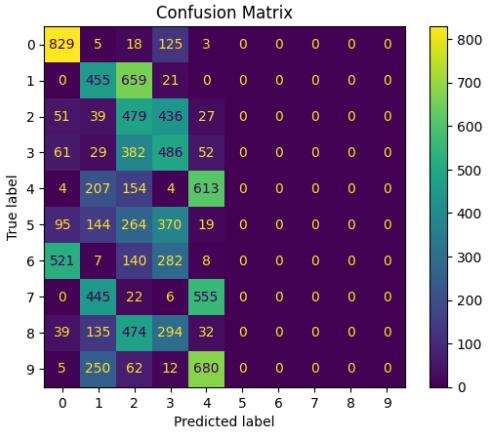
Top labels for each cluster in MNIST:

	Cluster	Top Label	Percentage
0	0	0	0.532643
1	0	6	0.286890
2	0	5	0.074310
3	1	7	0.256085
4	1	1	0.252548
5	1	9	0.170377
6	2	1	0.264940
7	2	2	0.193955
8	2	8	0.153368
9	3	3	0.227736
10	3	6	0.180659
11	3	2	0.180150
12	4	9	0.321827
13	4	4	0.310461
14	4	7	0.278257

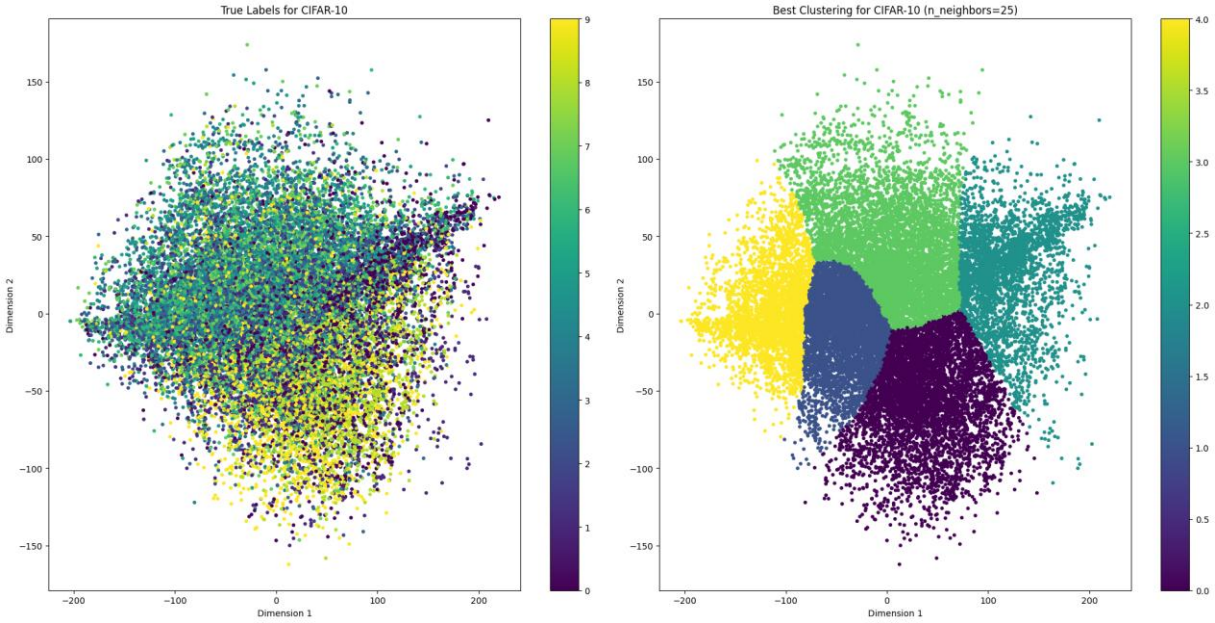
Classification Report:

	precision	recall	f1-score	support
0	0.52	0.85	0.64	980
1	0.27	0.40	0.32	1135
2	0.18	0.46	0.26	1032
3	0.24	0.48	0.32	1010
4	0.31	0.62	0.41	982
5	0.00	0.00	0.00	892
6	0.00	0.00	0.00	958
7	0.00	0.00	0.00	1028
8	0.00	0.00	0.00	974
9	0.00	0.00	0.00	1009
accuracy			0.29	10000
macro avg	0.15	0.28	0.20	10000
weighted avg	0.15	0.29	0.20	10000

Παρατηρούμε ότι το 0 καταφέρνει να το ομαδοποιήσει αρκετά ικανοποιητικά σε σχέση με τα υπόλοιπα labels, αφού έχει πετύχει ποσοστό 53% (δηλαδή το cluster 0 αποτελείται 53% από μηδενικά).



To Cifar10 dataset (best n_neighbors= 25) (metrics: ARI= 0.0399, NMI= 0.0658).



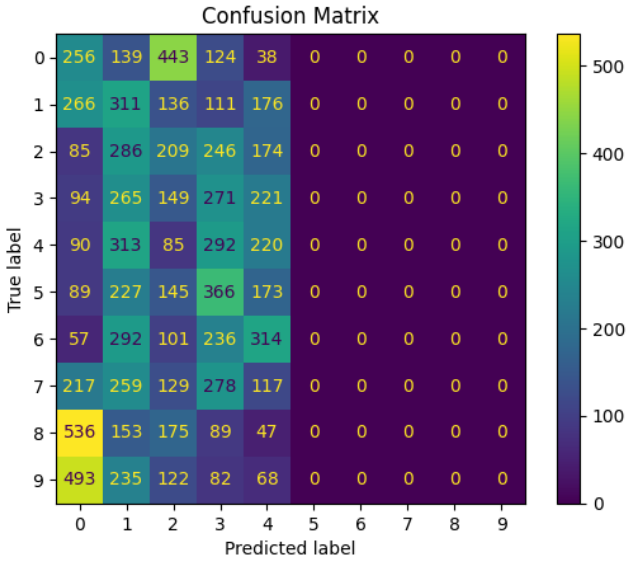
Top Labels για κάθε Cluster: Classification Accuracy = 0.13

Top labels for each cluster in CIFAR-10:

	Cluster	Top Label	Percentage
0	0	9	0.229903
1	0	8	0.229376
2	0	1	0.131926
3	1	6	0.131148
4	1	1	0.126036
5	1	2	0.125507
6	2	0	0.250650
7	2	2	0.127204
8	2	8	0.119688
9	3	5	0.150784
10	3	4	0.138641
11	3	3	0.137104
12	4	6	0.208866
13	4	3	0.145499
14	4	4	0.126462

Classification Report:

	precision	recall	f1-score	support
0	0.12	0.26	0.16	1000
1	0.13	0.31	0.18	1000
2	0.12	0.21	0.16	1000
3	0.13	0.27	0.18	1000
4	0.14	0.22	0.17	1000
5	0.00	0.00	0.00	1000
6	0.00	0.00	0.00	1000
7	0.00	0.00	0.00	1000
8	0.00	0.00	0.00	1000
9	0.00	0.00	0.00	1000
accuracy			0.13	10000
macro avg	0.06	0.13	0.08	10000
weighted avg	0.06	0.13	0.08	10000



Για Cluster=10:

To Mnist dataset (best n_neighbors= 25) (metrics: ARI= 0.191, NMI= 0.338).

Top Labels για κάθε Cluster: Classification Accuracy = 0.05

Top labels for each cluster in MNIST:

	Cluster	Top Label	Percentage
0	0	6	0.223932
1	0	3	0.205812
2	0	2	0.177436
3	1	2	0.300584
4	1	3	0.229364
5	1	6	0.165814
6	2	9	0.328080
7	2	7	0.309415
8	2	4	0.288262
9	3	7	0.509890
10	3	9	0.268864
11	3	4	0.159341
12	4	1	0.513901
13	4	8	0.154260
14	4	2	0.129596
15	5	0	0.325241
16	5	6	0.280539
17	5	3	0.197303
18	6	4	0.389633
19	6	9	0.304654
20	6	7	0.132228
21	7	0	0.654428
22	7	6	0.260799
23	7	5	0.046436
24	8	8	0.277289
25	8	5	0.243838
26	8	1	0.195863
27	9	1	0.484861
28	9	8	0.192036
29	9	5	0.147242

Classification Report:

	precision	recall	f1-score	support
0	0.14	0.16	0.15	980
1	0.00	0.00	0.00	1135
2	0.01	0.01	0.01	1032
3	0.02	0.01	0.01	1010
4	0.06	0.08	0.07	982
5	0.05	0.05	0.05	892
6	0.00	0.00	0.00	958
7	0.00	0.00	0.00	1028
8	0.24	0.17	0.20	974
9	0.01	0.01	0.01	1009
accuracy			0.05	10000
macro avg	0.05	0.05	0.05	10000
weighted avg	0.05	0.05	0.05	10000

Παρατηρούμε ότι στο cluster 3 έχουμε 50.9% από το label 7, στο cluster 4 51.3% το label 1 και στο cluster 7 65,4% το label 0.

Το Cifar10 dataset (best n_neighbors= 200) (metrics: ARI= 0.0358, NMI= 0.07).

Top Labels για κάθε Cluster: Classification Accuracy = 0.1

Top labels for each cluster in CIFAR-10:

	Cluster	Top Label	Percentage
0	0	5	0.161182
1	0	3	0.160060
2	0	7	0.148467
3	1	8	0.149931
4	1	0	0.132271
5	1	7	0.122230
6	2	2	0.152457
7	2	4	0.141208
8	2	6	0.134695
9	3	5	0.172067
10	3	4	0.170753
11	3	3	0.154553
12	4	9	0.260507
13	4	1	0.188260
14	4	8	0.152831
15	5	6	0.231434
16	5	3	0.153713
17	5	2	0.126655
18	6	0	0.221542
19	6	2	0.141983
20	6	5	0.103223
21	7	6	0.186541
22	7	4	0.133891
23	7	2	0.119247
24	8	0	0.312217
25	8	2	0.131222
26	8	3	0.107843
27	9	8	0.306561
28	9	9	0.228231
29	9	0	0.161829

Classification Report:

	precision	recall	f1-score	support
0	0.04	0.04	0.04	1000
1	0.08	0.09	0.09	1000
2	0.14	0.23	0.18	1000
3	0.16	0.12	0.14	1000
4	0.04	0.04	0.04	1000
5	0.09	0.06	0.07	1000
6	0.06	0.07	0.07	1000
7	0.09	0.10	0.10	1000
8	0.06	0.03	0.04	1000
9	0.23	0.21	0.22	1000
accuracy			0.10	10000
macro avg	0.10	0.10	0.10	10000
weighted avg	0.10	0.10	0.10	10000

Για Cluster=15:

To Mnist dataset (best n_neighbors= 200) (metrics: ARI= 0.186, NMI= 0.3525).

Top Labels για κάθε Cluster: Classification Accuracy = 0.06

Top labels for each cluster in MNIST:

	Cluster	Top Label	Percentage
0	0	2	0.288153
1	0	3	0.255020
2	0	8	0.172189
3	1	4	0.391875
4	1	9	0.390625
5	1	7	0.131250
6	2	4	0.350117
7	2	9	0.303864
8	2	7	0.233021
9	3	8	0.245972
10	3	2	0.241676
11	3	5	0.227712
12	4	3	0.341758
13	4	2	0.174038
14	4	6	0.168294
15	5	4	0.465079
16	5	9	0.315344
17	5	3	0.054497
18	6	6	0.363831
19	6	0	0.269128
20	6	5	0.144462
21	7	0	0.593704
22	7	6	0.244510
23	7	3	0.077599
24	8	1	0.564269
25	8	8	0.162736
26	8	5	0.093750
27	9	1	0.426220
28	9	5	0.236095
29	9	8	0.235528
30	10	0	0.685015
31	10	6	0.252294
32	10	5	0.031346
33	11	7	0.573211
34	11	9	0.257903
35	11	4	0.082363
36	12	2	0.261726
37	12	6	0.249566
38	12	3	0.231616
39	13	7	0.703427
40	13	9	0.216822
41	13	4	0.033022
42	14	1	0.624402
43	14	8	0.123206
44	14	2	0.087919

Classification Report:

	precision	recall	f1-score	support
0	0.01	0.01	0.01	980
1	0.00	0.00	0.00	1135
2	0.03	0.02	0.02	1032
3	0.19	0.15	0.17	1010
4	0.00	0.00	0.00	982
5	0.01	0.01	0.01	892
6	0.38	0.31	0.34	958
7	0.00	0.00	0.00	1028
8	0.16	0.12	0.14	974
9	0.00	0.00	0.00	1009
10	0.00	0.00	0.00	0
11	0.00	0.00	0.00	0
12	0.00	0.00	0.00	0
13	0.00	0.00	0.00	0
14	0.00	0.00	0.00	0
accuracy			0.06	10000
macro avg	0.05	0.04	0.05	10000
weighted avg	0.08	0.06	0.07	10000

Παρατηρούμε ότι τα label 0, 1, 7 έχουν συγκεντρωθεί σε αρκετά ικανοποιητικό βαθμό σε συγκεκριμένα clusters.

Το Cifar10 dataset (best n_neighbors= 100) (metrics: ARI= 0.029, NMI= 0.069).

Top Labels για κάθε Cluster: Classification Accuracy = 0.07

Top labels for each cluster in CIFAR-10:

	Cluster	Top Label	Percentage
0	0	6	0.156614
1	0	4	0.153439
2	0	2	0.150265
3	1	8	0.308517
4	1	0	0.183596
5	1	9	0.162145
6	2	2	0.147410
7	2	0	0.137948
8	2	7	0.114044
9	3	4	0.188482
10	3	5	0.181937
11	3	3	0.162304
12	4	6	0.232731
13	4	3	0.149841
14	4	2	0.141339
15	5	9	0.326021
16	5	8	0.242057
17	5	1	0.133888
18	6	9	0.250654
19	6	1	0.196232
20	6	8	0.152800
21	7	0	0.230962
22	7	2	0.156313
23	7	5	0.101703
24	8	8	0.230325
25	8	9	0.174187
26	8	0	0.122246
27	9	5	0.180838
28	9	3	0.172455
29	9	7	0.164671
30	10	0	0.342828
31	10	2	0.127162
32	10	3	0.112920
33	11	2	0.140792
34	11	4	0.116167
35	11	1	0.110814
36	12	6	0.179122
37	12	1	0.148280
38	12	3	0.113879
39	13	6	0.207524
40	13	4	0.152306
41	13	3	0.134709
42	14	5	0.157920
43	14	4	0.148688
44	14	6	0.144801

	precision	recall	f1-score	support
0	0.06	0.05	0.05	1000
1	0.14	0.08	0.10	1000
2	0.15	0.12	0.13	1000
3	0.15	0.07	0.10	1000
4	0.11	0.05	0.07	1000
5	0.01	0.01	0.01	1000
6	0.05	0.03	0.04	1000
7	0.09	0.09	0.09	1000
8	0.25	0.21	0.23	1000
9	0.03	0.01	0.02	1000
10	0.00	0.00	0.00	0
11	0.00	0.00	0.00	0
12	0.00	0.00	0.00	0
13	0.00	0.00	0.00	0
14	0.00	0.00	0.00	0
accuracy			0.07	10000
macro avg	0.07	0.05	0.06	10000
weighted avg	0.10	0.07	0.08	10000

Mnist Dataset

Number of Clusters	Best Neighbors	ARI	NMI	ACCURACY
5	200	0.1785	0.312	0.29
10	25	0.191	0.338	0.05
15	200	0.186	0.3525	0.06

Παρατηρώ ότι καθώς αυξάνουμε τα clusters οι μετρικές ARI και NMI φαίνεται να βελτιώνονται αλλά σε πολύ μικρό βαθμό. Ωστόσο, η ακρίβεια που παίρνουμε από τις προβλέψεις πέφτει δραστικά από τα 5 στα 10 clusters, άρα το clustering δεν καταφέρνει να κατανέμει τα δεδομένα σε ομάδες που ταιριάζουν μεταξύ τους.

Cifar10 Dataset

Number of Clusters	Best Neighbors	ARI	NMI	ACCURACY
5	25	0.0399	0.0658	0.13
10	200	0.0358	0.07	0.1
15	100	0.029	0.069	0.07

Το cifar10 είναι ένα πολύπλοκο dataset και οι τιμές των ARI και NMI είναι σχεδόν μηδενικές, γεγονός που δείχνει ότι τα clusters δεν αντανακλούν την πραγματική δομή των δεδομένων. Η ακρίβεια είναι επίσης χαμηλή με την καλύτερη τιμή να παρατηρείται για 5 clusters, ενώ μειώνεται περαιτέρω καθώς αυξάνουμε τον αριθμό των clusters.