3η Εργασία
Αναγνώριση Προτύπων
Μπούρχα Ιωάννα 58019

Επιβλέπων καθηγητής: Ηλίας Θεοδωρακόπουλος

Ακαδημαϊκό έτος: 2023 - 2024

ΠΕΡΙΕΧΟΜΕΝΑ

ΑΣΚΗΣΗ 1

Κατεβάστε το seeds Dataset από τη διεύθυνση https://archive.ics.uci.edu/ml/machine-learning-databases/00236/seeds_dataset.txt .

Το αρχείο περιέχει μορφολογικές μετρήσεις 210 σπόρων από τρεις ποικιλίες σιτηρών (ω1: Kama, ω2: Rosa, ω3: Canadian). Οι πρώτες 7 στήλες περιέχουν τις μετρηθείσες τιμές των μορφολογικών χαρακτηριστικών και η τελευταία στήλη περιλαμβάνει την ετικέτα της ποικιλίας στην οποία ανήκει κάθε σπόρος.

Προκειμένου να μην χρειάζεται κάθε φορά να φορτώνω χειροκίνητα το αρχείο που περιέχει τα δεδομένα, συνδέω το google colab με τον λογαριασμό μου στο google drive μέσω της βιβλιοθήκης `google.colab`. Με την εντολή `mount` φορτώνονται τα περιεχόμενα του Drive μου. Έπειτα, προσδιορίζω την διεύθυνση του αρχείου seeds_dataset.txt, το οποίο και διαβάζω με την βιβλιοθήκη `numpy`, ώστε να το αντιμετωπίζω σαν πίνακα, ο οποίος περιέχει δεδομένα τύπου float.

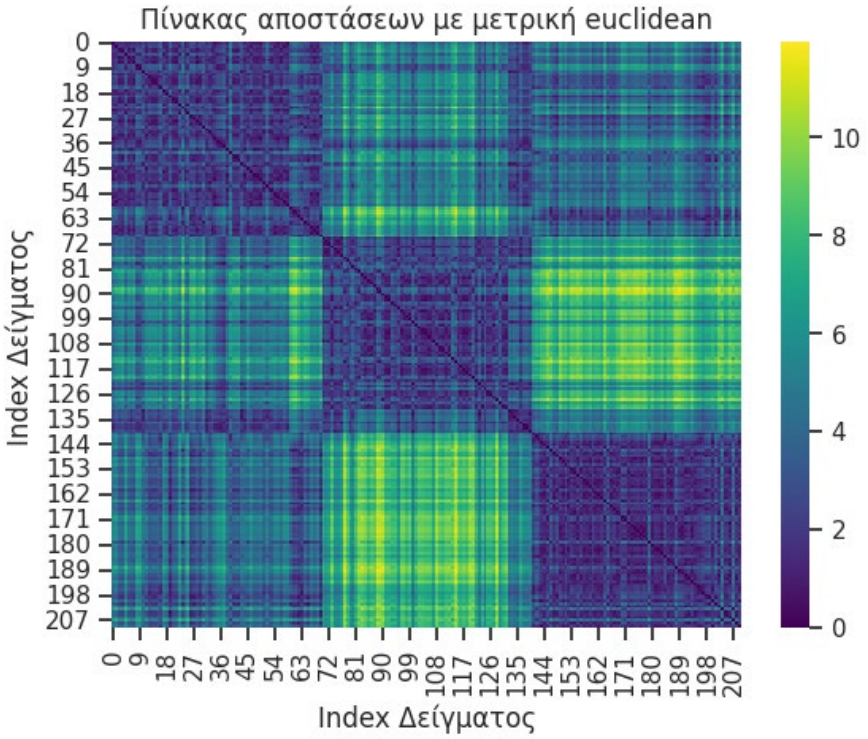
Εάν δεν θέλω να συνδεθώ στο Drive, αλλά να φορτώσω χειροκίνητα το αρχείο `seeds_dataset.txt`, χρησιμοποιώ την τελευταία εντολή του κελιού.

Αρχικά, διαχωρίζω τα δεδομένα από τις ετικέτες και αποθηκεύω τις τιμές στις μεταβλητές `data` και `labels` αντίστοιχα. Από τα περιεχόμενα τις δεύτερης, συμπεραίνω ότι κάθε μία από τις 3 κλάσεις αποτελείται από 70 δείγματα.

ΕΡΩΤΗΜΑ Α

Απεικονίστε τον πίνακα αποστάσεων των δεδομένων για Ευκλείδεια και cosine μετρικές. Ποιες κλάσεις πιστεύετε ότι είναι ευκολότερο να διαχωριστούν μεταξύ τους; Γιατί;

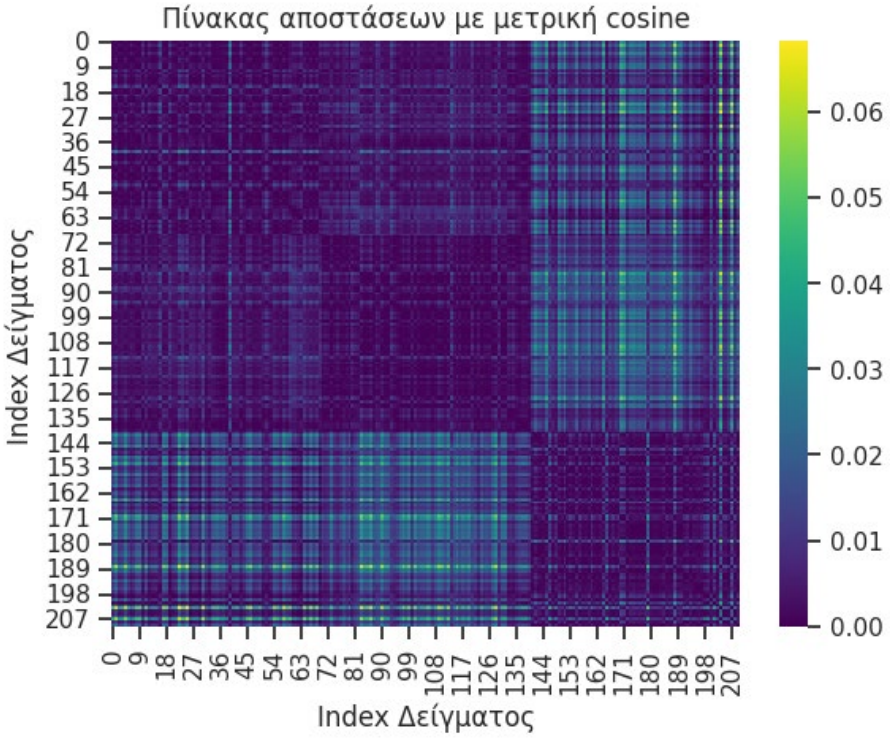
Για τον υπολογισμό των αποστάσεων με τις αντίστοιχες μετρικές χρησιμοποιώ την βιβλιοθήκη `cdist`, η οποία γενικά χρησιμοποιείται για τον υπολογισμό αποστάσεων μεταξύ δύο σημείων. Μικρή τιμή απόστασης μεταξύ δύο δειγμάτων σημαίνει ότι αυτά είναι παρόμοια, ενώ μεγάλη σημαίνει ότι είναι ανόμοια. Για την απεικόνιση του πίνακα χρησιμοποιώ τις βιβλιοθήκες `matplotlib` και `seaborn`.



Εικόνα 1: Πίνακας αποστάσεως για ευκλείδεια απόσταση.

Η διαγώνιος με το σκούρο μπλε δηλώνει ότι το κάθε δείγμα είναι πάρα πολύ κοντά με τον εαυτό του, όπως και αναμενόταν. Όσο πιο φωτεινό είναι το τετράγωνο τόσο μεγαλύτερη είναι η απόσταση μεταξύ των δύο δειγμάτων. Για την πρώτη κλάση (πρώτα 70 δείγματα) παρατηρώ σχεδόν μηδενική απόσταση μεταξύ των δειγμάτων της αλλά παρόμοιες αποστάσεις για τα δείγματα των κλάσεων 2 και 3. Η κλάση 2 (επόμενα 70

δείγματα) σημειώνει σχεδόν μηδενική απόσταση μεταξύ των δειγμάτων της, μέτρια με τα δείγματα της κλάσης 1 και μεγάλη με τα δείγματα της κλάσης 3. Η κλάση 3 (τελευταία 70 δείγματα) σημειώνει σχεδόν μηδενική απόσταση μεταξύ των δειγμάτων της, μικρή για τα δείγματα της κλάσης 1 και μεγάλη για τα δείγματα της κλάσης 2. Συνεπώς, οι κλάσεις 2 και 3 είναι εύκολα διαχωρίσιμες.



Εικόνα 2: Πίνακας αποστάσεως για cosine απόσταση.

Όπως και προηγουμένως, όσο πιο σκούρο το χρώμα ενός τετραγώνου τόσο πιο κοντά βρίσκονται τα δύο δείγματα. Οι αποστάσεις των δειγμάτων της κλάσης 1 είναι μικρές τόσο για τα ίδια τα δεδομένα της κλάσης όσο και για εκείνα της κλάσης 2. Αντίθετα, η απόστασή τους από τα δεδομένα της κλάσης 3 είναι αισθητά μεγαλύτερη. Ομοίως για την κλάση 3. Συνεπώς, για αυτήν την περίπτωση, μονάχα η κλάση 3 είναι εύκολα διαχωρίσιμη από τις υπόλοιπες.

ΕΡΩΤΗΜΑ Β

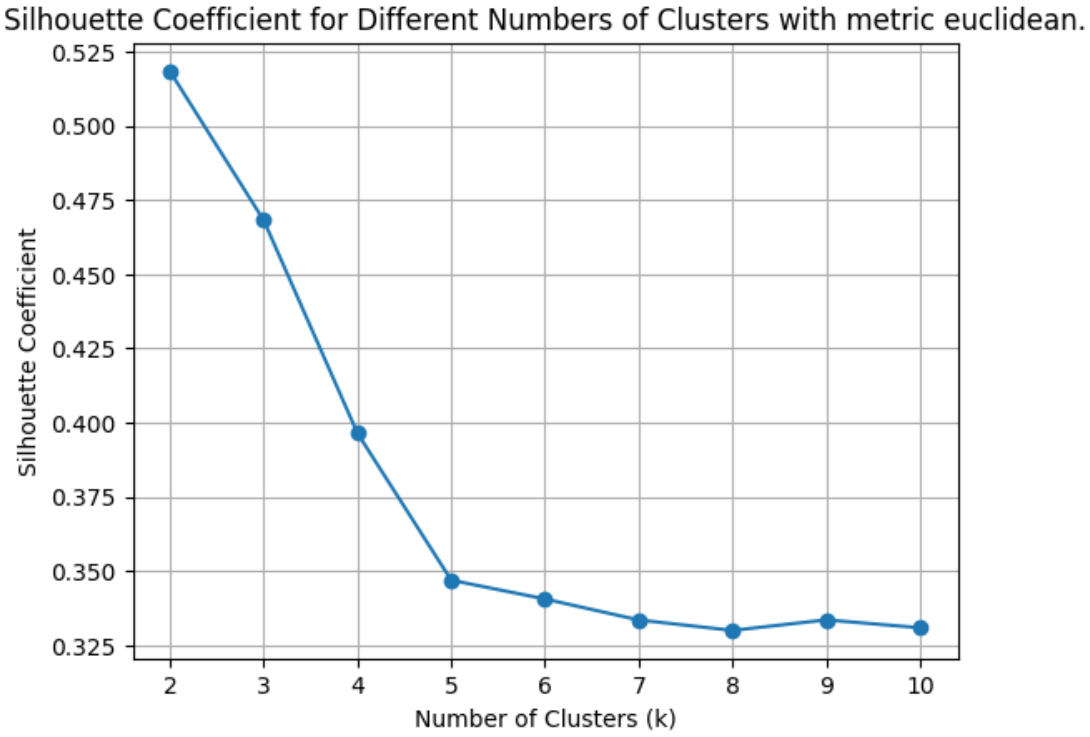
Να υπολογίσετε το Silhouette Coefficient για την ομαδοποίηση των 7-διάστατων δεδομένων σε k=2, 3, ... 10 κλάσεις με τη μέθοδο k-means και Ευκλείδεια ή squared Euclidean μετρική. Απεικονίστε το διάγραμμα του Silhouette και σχολιάστε ποιος είναι ο βέλτιστος αριθμός κλάσεων σύμφωνα με το κριτήριο αυτό.

Ο δείκτης Silhouette Coefficient αξιολογεί έναν αλγόριθμο ομαδοποίησης συγκρίνοντας πόσο κοντά είναι τα δεδομένα μέσα στην εκάστοτε ομάδα σε σχέση με τις υπόλοιπες. Παίρνει τιμές από -1 μέχρι 1. Η τιμή 1 δηλώνει ότι τα δείγματα είναι πολύ καλά απομονωμένα μεταξύ των ομάδων και ότι βρίσκονται πολύ κοντά στο κέντρο της ομάδας τους. Η τιμή 0 δηλώνει ότι τα δείγματα βρίσκονται κοντά στο σύνορο απόφασης γειτονικών ομάδων. Η τιμή -1 δηλώνει ότι το δείγμα έχει τοποθετεί σε λάθος ομάδα. Επομένως, όσο μεγαλύτερη η τιμή του τόσο καλύτερο το αποτέλεσμα.

Για την υλοποίηση του k-means και τον υπολογισμό του Silhouette Coefficient χρησιμοποιώ τις έτοιμες συναρτήσεις της βιβλιοθήκης `sklearn`. Η συνάρτηση για τον k-means χρησιμοποιεί εξ' ορισμού την τετραγωνική ευκλείδεια απόσταση. Η διαδικασία υλοποιείται με την συνάρτηση `calcilate_silhoutte_coefficient`.

Προκειμένου σε κάθε εκτέλεση του κώδικα να παίρνω το ίδιο αποτέλεσμα καθόρισα την τιμή της παραμέτρου `n_init` της συνάρτησης `KMeans` σε 1. Η παράμετρος αυτή καθορίζει τον αριθμό των φορών που εκτελείται ο αλγόριθμος k-means με διαφορετική αρχικοποίηση των κέντρων των κλάσεων. Για την τιμή που έδωσα, ο αλγόριθμος θα εκτελεστεί μία φορά με ψευδοτυχαία αρχικοποίηση κέντρων η οποία καθορίζεται από το την παράμετρο `random_state`.

Όπως φαίνεται και από την εικόνα 3, η βέλτιστη τιμή των clusters για τα τρέχοντα δεδομένα είναι 2.

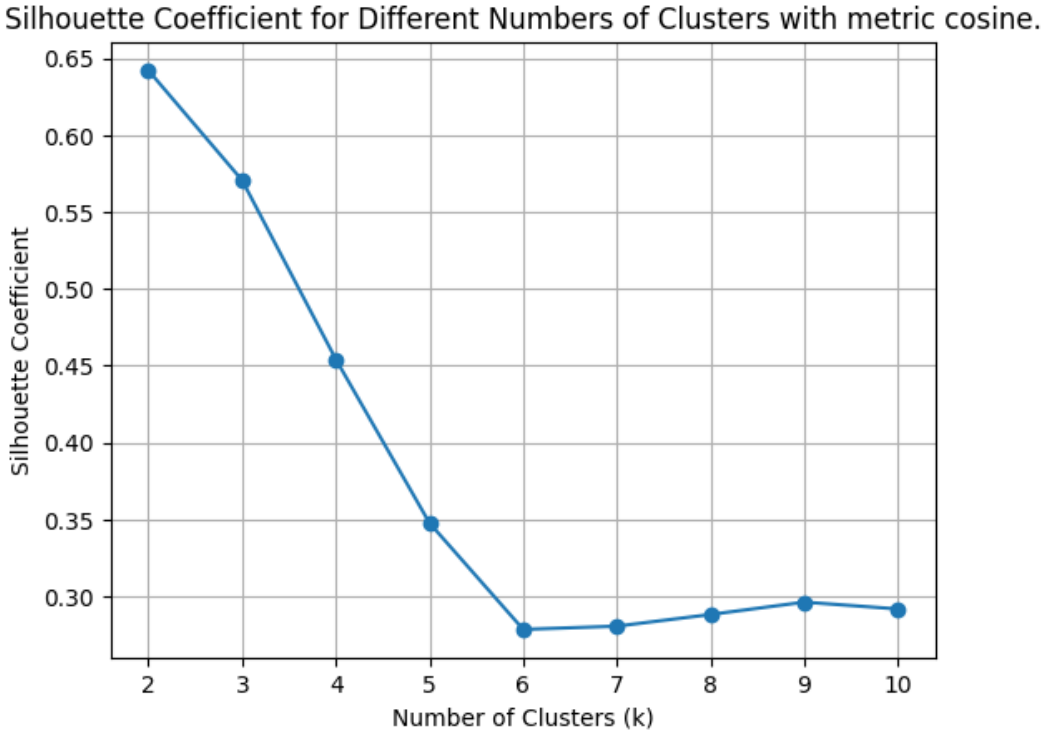


Εικόνα 3: Τιμές δείκτη Silhouette Coefficient για ευκλείδεια απόσταση για k-means

ΕΡΩΤΗΜΑ Γ

Να κανονικοποιηθούν τα δεδομένα ώστε κάθε χαρακτηριστικό να έχει μηδενική τιμή και μοναδιαία variance. Υπολογίσετε εκ νέου το Silhouette Coefficient στα κανονικοποιημένα δεδομένα για cosine μετρική. Απεικονίστε το νέο διάγραμμα του Silhouette. Τι παρατηρείτε?

Προκειμένου να κανονικοποιηθούν τα δεδομένα αξιοποιώ την συνάρτηση `StandardScaler` της βιβλιοθήκης `sklearn`. Στην συνέχεια, καλώ την συνάρτηση `calcilate_silhoutte_coefficient` με ορίσματα τα κανονονικοποιημένα δεδομένα και την μετρική ‘`cosine`’. Τα αποτελέσματα φαίνονται στην εικόνα 4.



Εικόνα 4: Τιμές δείκτη Silhouette Coefficient για cosine απόσταση για k-means

Συγκρίνοντας τα αποτελέσματα των ερωτημάτων Β και Γ παρατηρώ ότι και για τα δύο παρατηρείται η βέλτιστη τιμή των clusters για τα τρέχοντα δεδομένα είναι 2, αλλά στην δεύτερη περίπτωση σημειώνεται αύξηση της τιμής του δείκτη. Συνεπώς, στην δεύτερη περίπτωση βελτιώνεται σημαντικά η ομοιογένεια εντός των ομάδων καθιστώντας πιο εύκολο τον διαχωρισμό τους.

ΕΡΩΤΗΜΑ Δ

Ομαδοποιείστε τα δεδομένα σε 3 κλάσεις με τη μέθοδο k-means και squared Euclidean μετρική. Υπολογίστε το Rand Index για τη σύγκριση της παραγόμενης ομαδοποίησης με τις ετικέτες του dataset. Επαναλάβετε 5 φορές (με τυχαία αρχικοποίηση κέντρων), και υπολογίστε την μέση τιμή και το variance του Rand Index.

Αναφορικά με τον αλγόριθμο k-means ακολουθείται η προαναφερόμενη προσέγγιση.

Ο δείκτης Rand Index είναι ένα μέτρο αξιολόγησης της ομοιότητας των clusters. Κατά κύριο λόγο χρησιμοποιείται για να συγκρίνουμε το αποτέλεσμα της ομαδοποίσης διαφορετικών αλγόριθμων. Οι υψηλές τιμές δηλώνουν ότι τα αποτελέσματα των δύο αλγορίθμων συμφωνούν σε μεγάλο βαθμό. Ο υπολογισμός του γίνεται με την συνάρτηση `adjusted_rand_score` της βιβλιοθήκης `sklearn`. Προτιμήθηκε αυτή η συνάρτηση έναντι της `rand_score`, διότι η τροποποιημένη είναι πιο αξιόπιστος. Ο απλός Rand Index δεν προσαρμόζεται καλά στην τυχαιότητα, με αποτέλεσμα να δίνει σχετικά υψηλό σκορ ακόμα και σε περιπτώσεις που δεν υπάρχει πραγματική συσχέτιση μεταξύ των πραγματικών ετικετών και εκείνων που παρήχθησαν από την ομαδοποίηση. Οι τιμές κάθε επανάληψης αποθηκεύονται σε μία λίστα και στο τέλος αυτών υπολογίζεται ο μέσος όρος και η διακύμανση του δείκτη.

```
-- Squared Euclideian metric
Μέση τιμή Rand Index:  0.7153642264424888     Διακύμανση          Rand          Index:
6.306419691952044e-06
```

ΕΡΩΤΗΜΑ Ε

Επαναλάβετε το προηγούμενο ερώτημα για cosine μετρική. Σε περίπτωση που η βιβλιοθήκη που χρησιμοποιείτε δεν υποστηρίζει k-means με cosine μετρική, μπορείτε να χρησιμοποιήσετε την προσέγγιση της κανονικοποίησης των διανυσμάτων σε μοναδιαίο μήκος ακολουθούμενο από k-means με Ευκλείδεια μετρική.

Η βιβλιοθήκη που χρησιμοποιήθηκε προηγουμένως δεν καθιστά εφικτή την αλλαγή της μετρικής στον υπολογισμό της απόστασης. Επομένως, ακολουθώντας την υπόδειξη της εκφώνησης και το εργαστήριο του μαθήματος, κανονικοποιώ τα δεδομένα και στην συνέχεια τα διαιρώ με το μήκος.

```
-- Cosine metric
Μέση τιμή Rand Index:  0.6889693498587033     Διακύμανση Rand Index: 0.0
```

Ποια μετρική επιτυγχάνει ομαδοποίηση πιο πιστή στις πραγματικές ομάδες των δεδομένων;

Η απάντηση καθορίζεται από την μέση τιμή και την διακύμανση του rand index της κάθε υλοποίησης. Η μέση τιμή προσδιορίζει την απόδοση του αλγορίθμου. Υψηλή τιμή δηλώνει ότι τα αποτελέσματα συνάδουν με τις πραγματικές ετικέτες των δεδομένων. Η διακύμανση προσδιορίζει πόσο σταθερά είναι τα αποτελέσματα του αλγορίθμου. Υψηλή τιμή δηλώνει αστάθεια, δηλαδή τα αποτελέσματα διαφέρουν σημαντικά σε διαδοχικές εκτελέσεις. Συνεπώς, ο ιδανικός συνδυασμός είναι υψηλή μέση τιμή και χαμηλή διακύμανση, προκειμένου ο αλγόριθμος να είναι ακριβής και αξιόπιστος. Από τα παραπάνω είναι φανερό ότι η βέλτιστη υλοποίηση είναι η πρώτη (ευκλείδεια) καθώς έχει την μεγαλύτερη μέση τιμή και την μικρότερη διακύμανση του Rand Index.

BONUS

Έστω ότι σχεδιάζετε ένα σύστημα που χρειάζεται να εκτελεί ταξινόμηση με τη μέθοδο Nearest Neighbor χρησιμοποιώντας ένα πολύ μεγάλο σύνολο δεδομένων αναφοράς (εκπαίδευσης). Μπορείτε να σκεφτείτε έναν τρόπο να μειωθεί το υπολογιστικό κόστος κάθε νέας ταξινόμησης, αξιοποιώντας τεχνικές ομαδοποίησης? Περιγράψτε το σκεπτικό σας και τα βήματα του αλγορίθμου.

Η μέθοδος Nearest Neighbor (NN) αφορά την ταξινόμηση των δεδομένων με βάση την κατηγορία στην οποία ανήκουν ο πλησιέστερος γείτονάς του. Το πρόβλημα της μεθόδου αυτής είναι ότι ενδέχεται να γίνει εξαιρετικά απαιτητική υπολογιστικά για μεγάλα σύνολα δεδομένων.

Το υπολογιστικό αυτό κόστος οφείλεται στις συγκρίσεις του τρέχοντος δείγματος με κάθε δείγμα του συνόλου δεδομένων. Ένας τρόπος, λοιπόν, να μειώσουμε το υπολογιστικό κόστος είναι να προ-ομαδοποιήσουμε το μεγάλο σύνολο δεδομένων σε μικρότερα. Με άλλα λόγια, αντί να συγκρίνουμε το νέο δείγμα με όλο το μεγάλο σύνολο το συγκρίνουμε με εκπροσώπους του, τα κέντρα των ομάδων στις οποίες έχει διασπαστεί. Με αυτόν τον τρόπο περιορίζεται η αναζήτηση στην ομάδα που ενδέχεται να περιέχει το δείγμα.

Ψευδοαλγόριθμος:

1. Προεπεξεργασία - Ομαδοποίηση:
Χρησιμοποιώ κάποια μέθοδο ομαδοποίησης, π.χ. k-means, για να χωρίσω το μεγάλο σύνολο δεδομένων σε μικρότερες ομάδες. Για κάθε ομάδα υπολογίζω το κέντρο τους.

2. Εφαρμογή του Nearest Neighbor:
Συγκρίνω κάθε νέο δείγμα με τα κέντρα που υπολόγισα προηγουμένως.

3. Περιορισμένη Αναζήτηση:
Για την ομάδα της οποίας το κέντρο είναι πιο κοντά στο νέο δείγμα, συγκρίνω κάθε δείγμα της με το νέο ώστε να βρω αυτά που βρίσκονται πιο κοντά.

4. Ταξινόμηση:
Τελικά, το δείγμα ταξινομείται στην κλάση της οποίας η ετικέτα εμφανίζεται τις περισσότερες φορές στο πλήθος των κοντινότερων γειτόνων.

ΑΣΚΗΣΗ 2

ΕΡΩΤΗΜΑ Α

Επιλέξτε έναν αλγόριθμο ιεραρχικής ομαδοποίησης της αρεσκείας σας (agglomerative ή divisive) και εφαρμόστε τον στο ίδιο σύνολο δεδομένων με τα προηγούμενα. Περιγράψτε συνοπτικά την τεχνική που επιλέξατε και τις παραμέτρους που χρησιμοποιήσατε (π.χ. μετρική, linkage method κλπ.). σε μία παράγραφο.

Ο αλγόριθμος agglomerative ακολουθεί bottom-up προσέγγιση. Κάθε σημείο των δεδομένων θεωρείται ως μία ανεξάρτητη ομάδα και σταδιακά οι κοντινότερες συνενώνονται σε μεγαλύτερες έως ότου επιτευχθεί ο επιθυμητός αριθμός ομάδων. Η απλότητα του και εύκολη υλοποίηση του τον έχει καταστήσει σε έναν από τους δημοφιλούς αλγόριθμους ομαδοποίησης. Σημειώνεται ότι για μεγάλα σύνολα δεδομένων έχει μεγάλο υπολογιστικό κόστος.

Ο αλγόριθμος devise ακολουθεί top-down προσέγγιση. Τα δεδομένα θεωρούνται μία μεγάλη ομάδα η οποία σταδιακά διαιρείται σε μικρότερες έως ότου φτάσουμε στον επιθυμητό αριθμό ομάδων. Δεν είναι τόσο δημοφιλής όσο ο agglomerative καθώς είναι πιο περίπλοκος στην υλοποίηση. Επίσης, παρουσιάζει και αυτός υψηλό υπολογιστικό κόστος για μεγάλα σύνολα δεδομένων.

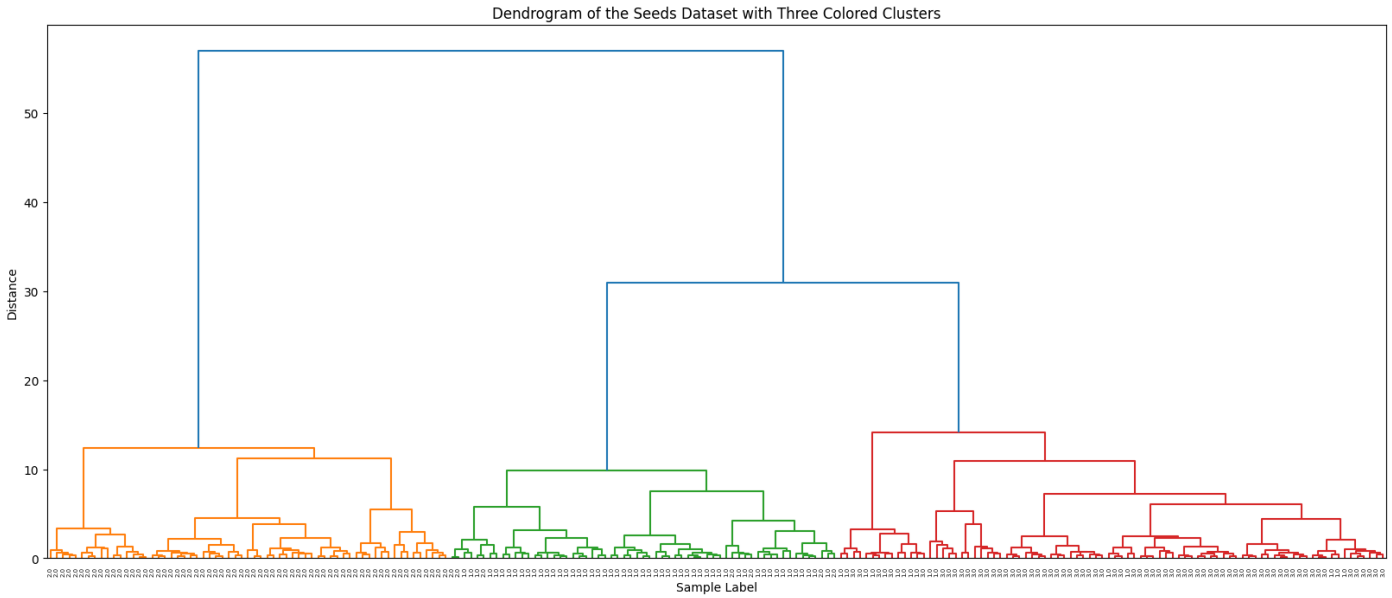
Με βάση τα παραπάνω επιλέγω τον agglomerative. Η υλοποίησή γίνεται με την βοήθεια της συνάρτησης `AgglomerativeClustering` βιβλιοθήκης `sklearn`. Ως μετρική απόστασης επιλέχθηκε η μέθοδος `'Ward'`, η οποία είναι αποτελεσματική στον εντοπισμό σφαιρικών συστάδων στοχεύοντας στην ελαχιστοποίηση της συνολικής διακύμανσης εντός των ομάδων κατά την συγχώνευση. Τα στοιχεία κάθε συστάδας είναι όσο το δυνατόν πιο όμοια μεταξύ τους. Η μετρική που ορίζεται αυτόματα για τον υπολογισμό της απόστασης είναι η ευκλείδεια.

ΕΡΩΤΗΜΑ Β

Κατασκευάστε το δενδρόγραμμα που προέκυψε από την ομαδοποίηση, και σχολιάστε πως σχετίζεται με τις ποικιλίες των σιτηρών (κλάσεις) των αντίστοιχων δεδομένων.

Το δενδρόδιαγραμμα υλοποιείται με την συνάρτηση `dedrodiagram` της βιβλιοθήκης. Χρησιμοποιώ την συνάρτηση `linkage` της `scipy` με τις ίδιες παραμέτρους που χρησιμοποίησα στον αλγόριθμο. Με την παράμετρο `orientation` καθορίζεται η κατεύθυνση του. Η παράμετρος `color_threshold` ορίζει το κατώφλι για τον χρωματισμό των διαφορετικών ομάδων στο δενδρόγραμμα. Οποιαδήποτε συγχώνευση ομάδων με απόσταση μεγαλύτερη από 15 θα εμφανίζεται σε διαφορετικό χρώμα.

Το δενδρόδιαγραμμα απεικονίζει την ομοιότητα των διαφορετικών σπόρων βάσει των μορφολογικών μετρήσεων που δόθηκαν στο dataset. Ο άξονας y αποτυπώνει την απόσταση, δηλαδή την μορφολογική διαφορά μεταξύ συγκρίνουμενων ομάδων και ο άξονας x την πραγματική ετικέτα των δειγμάτων.



Εικόνα 5: Δενδροδιάγραμμα άσκησης 2α.

Σύμφωνα και με το δενδροδιάγραμμα της εικόνας 5, εντοπίζονται τρεις διακριτές ομάδες, οι οποίες έχουν χρωματιστεί και με διαφορετικό χρώμα (πορτοκαλί, πράσινο, κόκκινο). Σύμφωνα με τις τιμές του άξονα x

παρατηρώ ότι οι τρεις κλάσεις ομαδοποιούνται σε ξεχωριστές ομάδες χωρίς πολλά σφάλματα. Συγκεκριμένα, η πορτοκαλί ομάδα αποτελείται μονάχα από δείγματα της κλάσης 2. Ελάχιστα δείγματα της κλάσης αυτής εντοπίζονται σε άλλες ομάδες. Η πράσινη ομάδα αποτελείται από δείγματα της κλάσης 1 με ελάχιστα δείγματα της κλάσης 2 και κανένα της κλάσης 3. Η κόκκινη ομάδα αποτελείται από δείγματα της κλάσης 3 και ελάχιστα δείγματα της κλάσης 1. Συνεπώς, κάθε μία από τις τρεις ομάδες ενδέχεται να αντιπροσωπεύει μία από τις τρεις διαφορετικές ποικιλίες σιτηρών.

ΕΡΩΤΗΜΑ Γ

Χρησιμοποιήστε κάποιο κριτήριο εξωτερικής επικύρωσης (π.χ. rand index, adjusted rand index, mutual information κλπ) και συγκρίνετε την ομαδοποίηση αυτή με την αντίστοιχη που επιτυγχάνει ο k-means σε σχέση με τις ετικέτες των δεδομένων. Ποια μετρική επιτυγχάνει ομαδοποίηση πιο πιστή στις πραγματικές ομάδες των δεδομένων;

Για να χρησιμοποιήσω ένα κριτήριο εξωτερικής επικύρωσης χρειάζομαι τις πραγματικές ετικέτες και τις ετικέτες που προκύπτουν από την ομαδοποίηση. Οι δεύτερες αποκτούνται με την συνάρτηση `fcluster` της βιβλιοθήκης `scipy`.

Όπως και προηγουμένως, επιλέχτηκε ο Adjusted Rand Index. Όσο μεγαλύτερη η τιμή του τόσο περισσότερο μοιάζουν οι πραγματικές ετικέτες με εκείνες της ομαδοποίησης. Για διάφορες μεθόδους δοκίμασα μετρικές απόστασης ευκλείδεια και συνημίτονου. Σημειώνω ότι οι δύο τελευταίες λειτουργούν αποκλειστικά με ευκλείδεια απόσταση.

Method	Metric	ARI
single	euclidean	0.0024703116164171554
complete	euclidean	0.546135027762822
average	euclidean	0.7441752360248661
ward	euclidean	0.7131537289031059
centroid	euclidean	0.5664395805591734
single	cosine	9.202374344043254e-05
complete	cosine	0.7108638571583838
average	cosine	0.7889212187703276

Όπως φαίνεται, η μέγιστη τιμή του δείκτη προκύπτει για μέθοδο `average` και μετρική απόστασης `euclidean`.

Συγκριτικά με τον αλγόριθμο του k-means που υλοποιήθηκε στην προηγούμενη άσκηση παρατηρώ ότι αναφορικά με την ευκλείδεια μετρική απόστασης έχω παρόμοια ή χειρότερα αποτελέσματα ανάλογα με την μέθοδο, ενώ για την μετρική απόστασης cosine τα αποτελέσματα είναι κατά κύριο λόγο καλύτερα.

ΕΡΩΤΗΜΑ Δ

Σχολιάστε τα πλεονεκτήματα των ιεραρχικών τεχνικών ομαδοποίησης.

Μία από τις σημαντικότερες διαφορές ανάμεσα σε ιεραρχικές και μη ιεραρχικές τεχνικές ομαδοποίησης είναι ότι οι πρώτες δεν απαιτούν προκαθορισμένο αριθμό ομάδων. Αυτό τις καθιστά χρήσιμες σε περιπτώσεις όπου ο αριθμός των ομάδων δεν είναι γνωστός εκ των προτέρων. Παρουσιάζουν μεγάλη ευελιξία ως προς τις μετρικές αποστάσεων και μεθόδων σύνδεσης επιτρέποντας την προσαρμογή τους ανάλογα με τα δεδομένα του προβλήματος. Άρα, είναι ιδιαίτερα χρήσιμες σε πολύπλοκα δεδομένα, όπου οι σχέσεις μεταξύ των δειγμάτων μπορεί να μην είναι γραμμικές ή εύκολα διαχωρίσιμες. Επιπρόσθετα, οι ιεραρχικές τεχνικές δημιουργούν δενδροδιάγραμμα παρέχοντας οπτική αναπαράσταση της διαδικασίας ομαδοποίησης και των σχέσεων των ομάδων, με αποτέλεσμα να είναι πιο εύκολες στην ερμηνεία.

ΑΣΚΗΣΗ 3

Χρησιμοποιώντας το ίδιο dataset, να υλοποιηθούν και να απαντηθούν τα παρακάτω:

ΕΡΩΤΗΜΑ Α

Να εφαρμοστεί η μέθοδος PCA στα δεδομένα. Ποιες είναι οι ελάχιστες κύριες συνιστώσες που πρέπει να κρατήσετε ώστε να εξηγείται τουλάχιστον το 90% της variance του αρχικού dataset στη νέα απεικόνιση, και πόσες για το 99% αυτής;

Η Principal Component Analysis (PCA) είναι στατιστική τεχνική που χρησιμοποιείται για την μείωση των διαστάσεων των δεδομένων επιδιώκοντας την επίλυση του curse of dimensionalities εστιάζοντας στην αύξηση της διακύμανσης. Προκειμένου η μέθοδος να μην επηρεάζεται από τις διαφορετικές κλίμακες των μεταβλητών, τα δεδομένα κανονικοποιούνται ώστε να έχουν μέση τιμή μηδέν και τυπική απόκλιση ένα. Έπειτα, υπολογίζονται οι διακυμάνσεις των μεταβλητών και αποθηκεύονται σε έναν πίνακα. Στον πίνακα διακυμάνσεων η διαγώνιος περιέχει τις διακυμάνσεις των μεταβλητών. Τα ιδιοδιανύσματα του πίνακα αυτού αντιπροσωπεύουν τις κύριες συνιστώσες, ενώ οι ιδιοτιμές δείχνουν πόση διακύμανση έχει κάθε συνιστώσα. Τα ιδιοδιανύσματα ταξινομούνται σε φθίνουσα σειρά με βάση την ιδιοτιμή τους. Έτσι, το πρώτο ιδιοδιάνυσμα έχει την μέγιστη διακύμανση. Τα ιδιοδιανύσματα αυτά προσδιορίζουν τις συνιστώσες του νέου χώρου απεικόνισης. Ανάλογα το πόσες συνιστώσες θέλουμε, επιλέγουμε και τα αντίστοιχα πρώτα ταξινομημένα ιδιοδιανύσματα. Τα δεδομένα απεικονίζονται στον νέο αυτό χώρο διατηρώντας την μέγιστη δυνατή ποσότητα πληροφορίας καθώς έχουν την μέγιστη δυνατή διακύμανση.

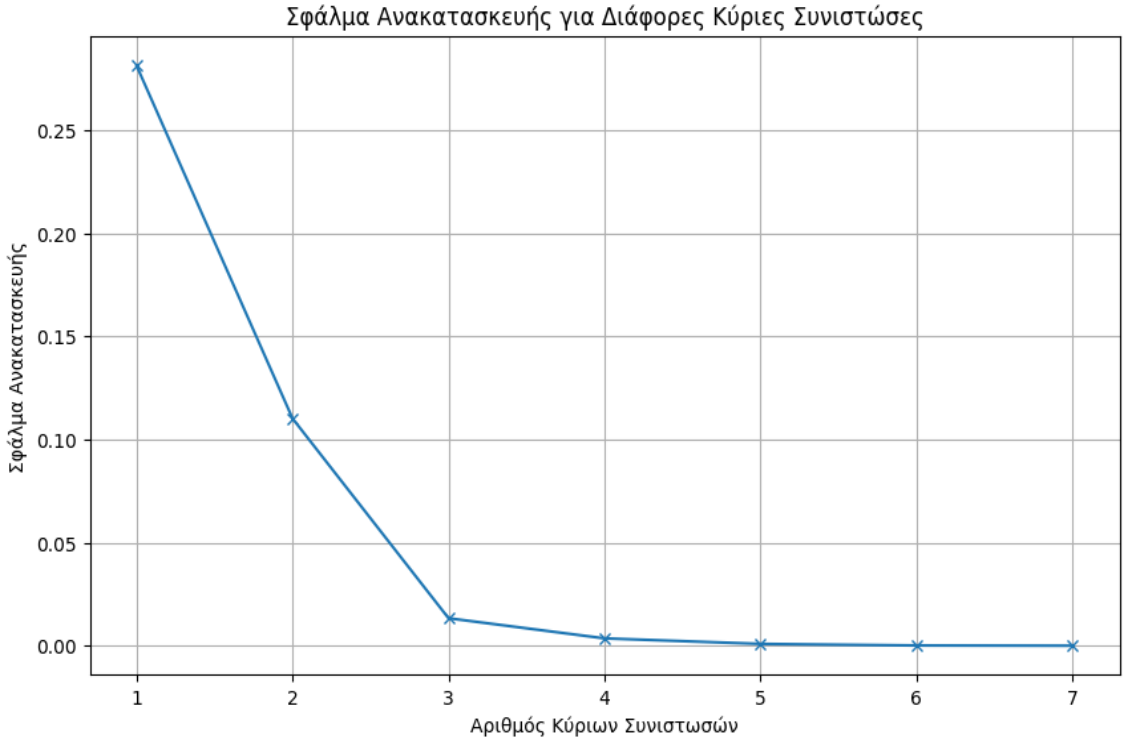
Για να εξηγηθεί τουλάχιστον το 90% της variance του αρχικού συνόλου δεδομένων, απαιτούνται 3 κύριες συνιστώσες, ενώ για το 99% της variance απαιτούνται 3 κύριες συνιστώσες. Αυτό δείχνει ότι τα δεδομένα μπορούν να περιγραφούν πολύ αποτελεσματικά με σχετικά λίγες διαστάσεις.

ΕΡΩΤΗΜΑ Β

Να υπολογισθεί το σφάλμα ανακατασκευής των δεδομένων χρησιμοποιώντας από 1 έως 7 κύριες συνιστώσες, και να αποτυπωθεί σε κατάλληλο διάγραμμα.

Η PCA μειώνει τις διαστάσεις των δεδομένων στον επιθυμητό αριθμό κύριων συνιστωσών διατηρώντας το μεγαλύτερο μέρος της διακύμανσης των δεδομένων. Κατά τον μετασχηματισμό των δεδομένων στον νέο χώρο των επιλεγμένων διαστάσεων διατηρούνται μονάχα οι πληροφορίες των συνιστωσών αυτών. Εάν επαναφέρουμε τα μετασχηματισμένα δεδομένα πίσω στον αρχικό χώρο, η ανακατασκευή δεν θα είναι τέλεια. Το σφάλμα ανακατασκευής υπολογίζεται ως η διαφορά των αρχικών και των ανακατασκευασμένων δεδομένων στον αρχικό χώρο. Στην παρούσα υλοποίηση χρησιμοποιήθηκε ο μέρος όρος του τετραγωνικού σφάλματος (MSE) για την καταγραφή του σφάλματος. Το σφάλμα αναμένεται να μειώνεται καθώς προστίθενται περισσότερες συνιστώσες, επειδή περισσότερες διαστάσεις σημαίνουν λιγότερη απώλεια πληροφοριών κατά τη διαδικασία ανακατασκευής.

Σύμφωνα με το διάγραμμα στην εικόνα 6, παρατηρούμε ότι το σφάλμα μειώνεται με την αύξηση των κύριων συνιστωσών που διατηρούμε. Είναι αξιοσημείωτο ότι η διατήρηση περισσότερων από 3 κύριες συνιστώσες βελτιώνει ελάχιστα την ανακατασκευή. Επομένως, τα δεδομένα μπορούν να αναπαρασταθούν με ελάχιστη περιττή πολυπλοκότητα και μικρό σφάλμα ανακατασκευής.

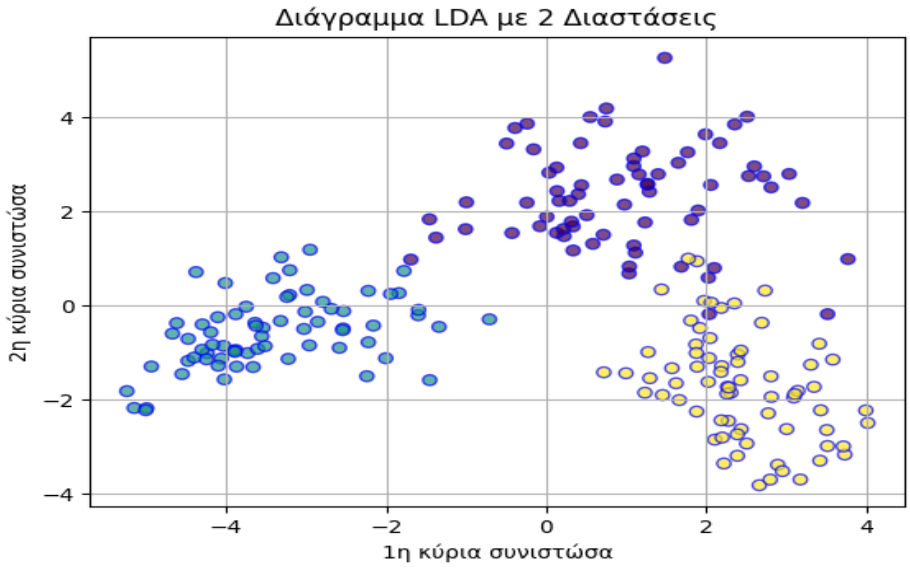


Εικόνα 6: Σφάλμα ανακατασκευής διατηρώντας διαφορετικό αριθμό κύριων συνιστωσών

ΕΡΩΤΗΜΑ Γ
Να εφαρμοστεί η μέθοδος LDA για την απεικόνιση του dataset σε 2 διαστάσεις.

Η Linear Discriminant Analysis (LDA) είναι στατιστική τεχνική που χρησιμοποιείται για την μείωση των διαστάσεων των δεδομένων εστιάζοντας στον διαχωρισμό των δειγμάτων ανάλογα με την κατηγορία που ανήκουν. Είναι ιδιαίτερα χρήσιμη σε περιπτώσεις ταξινόμησης και πρόβλεψης. Όπως και πριν, προκειμένου η μέθοδος να μην επηρεάζεται από τις διαφορετικές κλίμακες των μεταβλητών, τα δεδομένα κανονικοποιούνται ώστε να έχουν μέση τιμή μηδέν και τυπική απόκλιση ένα. Για να προσδιοριστεί ο βαθμός που διαφέρουν οι κλάσεις υπολογίζονται οι μέσες τιμές κάθε χαρακτηριστικού κάθε κλάσης. Οι ενδοκλασικές διακυμάνσεις δείχνουν πόσο διάσπαρτα είναι τα δείγματα εντός της κλάσης, ενώ οι ενδιάμεσες διακυμάνσεις δείχνουν πόσο διαφορετικές είναι οι κλάσεις μεταξύ τους. Έπειτα, προσδιορίζονται τα διανύσματα που μεγιστοποιούν την αναλογία της ενδιάμεσης προς την ενδοκλασική διακύμανση. Τα αρχικά δεδομένα προβάλλονται στο νέο χώρο που ορίζεται από τα διανύσματα αυτά, όπου οι κλάσεις είναι καλύτερα διαχωρισμένες.

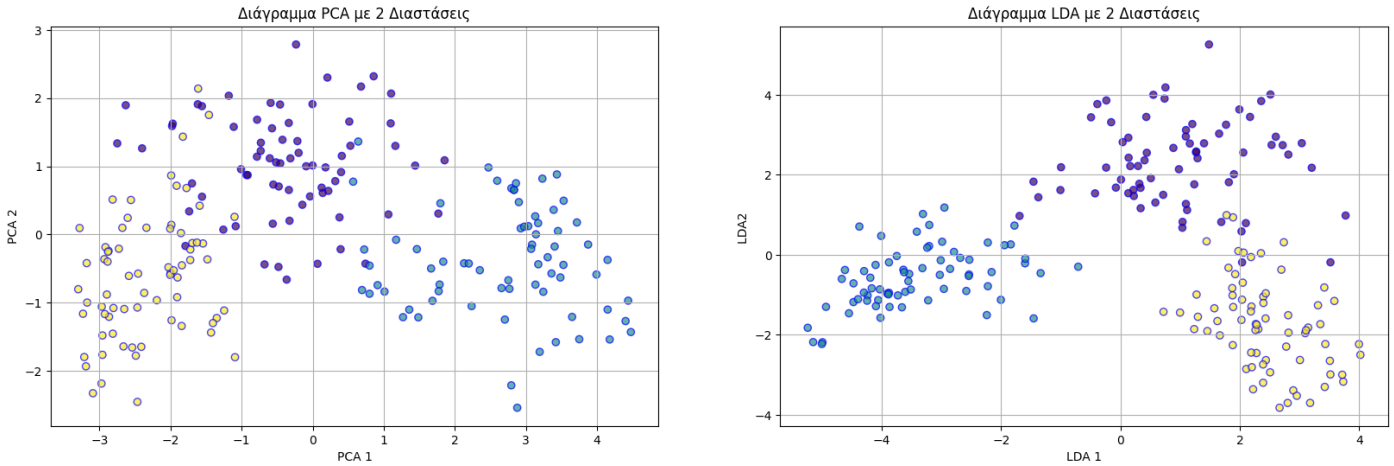
Σε αντίθεση με την PCA, που είναι μια μη επιβλεπόμενη τεχνική και εστιάζει στη μείωση της διαστατικότητας χωρίς να λαμβάνει υπόψη τις ετικέτες των δειγμάτων, η LDA είναι μια επιβλεπόμενη τεχνική και λαμβάνει υπόψη τις ετικέτες για να βρει τον καλύτερο διαχωρισμό μεταξύ των κλάσεων.



Εικόνα 7: Διάγραμμα LDA για δύο διαστάσεις

Να συγκρίνετε την απεικόνιση αυτή με την αντίστοιχη που παράγεται από τη μέθοδο PCA. Ποια τα κυριότερα ποιοτικά χαρακτηριστικά των απεικονίσεων που παράγουν οι δύο μέθοδοι και σε τι οφείλονται? Εξηγήστε.

Τα δύο διαγράμματα για την PCA και την LDA φαίνονται στην εικόνα 8. Η απεικόνιση με PCA, όπου οι κύριε συνιστώσες επιλέγονται ώστε να μεγιστοποιείται η διακύμανση, επικεντρώνεται στην ανάδειξη των δομικών πληροφοριών των δεδομένων. Αντίθετα, η απεικόνιση με LDA, όπου οι κύριες συνιστώσες επιλέγονται ώστε να διαχωρίζονται τα δεδομένα λαμβάνοντας υπόψιν τις ετικέτες, επικεντρώνεται στον όσο το δυνατό πιο καθαρό διαχωρισμό των δεδομένων. Συνεπώς, η PCA ενδείκνυται σε εφαρμογές για μείωση των διαστάσεων προς οπτικοποίηση των δομικών πληροφοριών δεδομένων, ενώ η LDA σε εφαρμογές ξεκάθαρου διαχωρισμού των δεδομένων δεδομένου των ετικετών τους.



Εικόνα 8: Διάγραμμα PCA (αριστερά) και LDA (δεξιά) για τις δύο κύριες συνιστώσες.

ΕΡΩΤΗΜΑ Δ

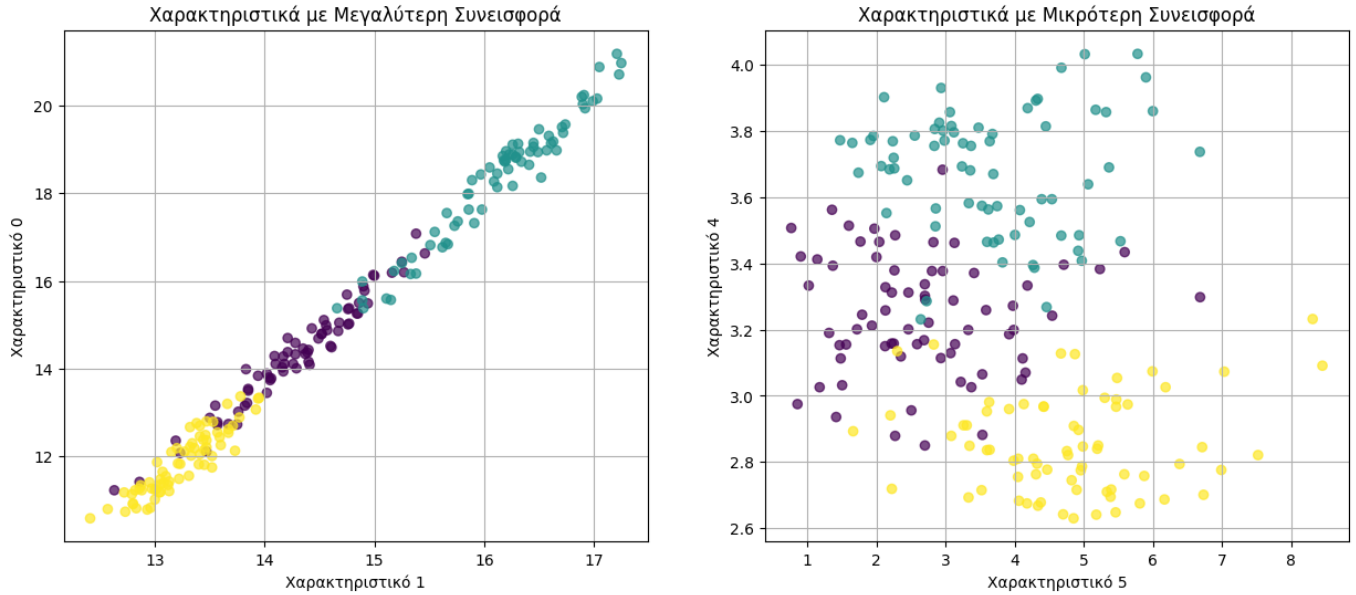
Με βάση τον πίνακα προβολής που παράγεται από την LDA στο προηγούμενο ερώτημα, ποια είναι τα δύο χαρακτηριστικά (features) που συνεισφέρουν περισσότερο στη διάκριση μεταξύ των κλάσεων και ποια τα δύο που συνεισφέρουν λιγότερο απ' όλα? Δημιουργείστε δυο δυσδιάστατες απεικονίσεις των δεομένων χρησιμοποιώντας το καθένα από τα δύο ζεύγη χαρακτηριστικών που καταδείξατε. Σχολιάστε.

Υπολογίζω την συνεισφορά κάθε χαρακτηριστικού με την συνάρτηση `lda.coef_` η οποία επιστρέφει τους συντελεστές για κάθε χαρακτηριστικό κάθε κλάσης. Σημειώνω ότι η συνεισφορά καθορίζεται από την απόλυτη τιμή, οπότε σε περίπτωση που περιέχει και αρνητικές τιμές εφαρμόζω την `np.abs`. Οι συνεισφορές αυτές ταξινομούνται σε φθίνουσα και προσδιορίζονται τα χαρακτηριστικά με την μέγιστη και την ελάχιστη συνεισφορά.

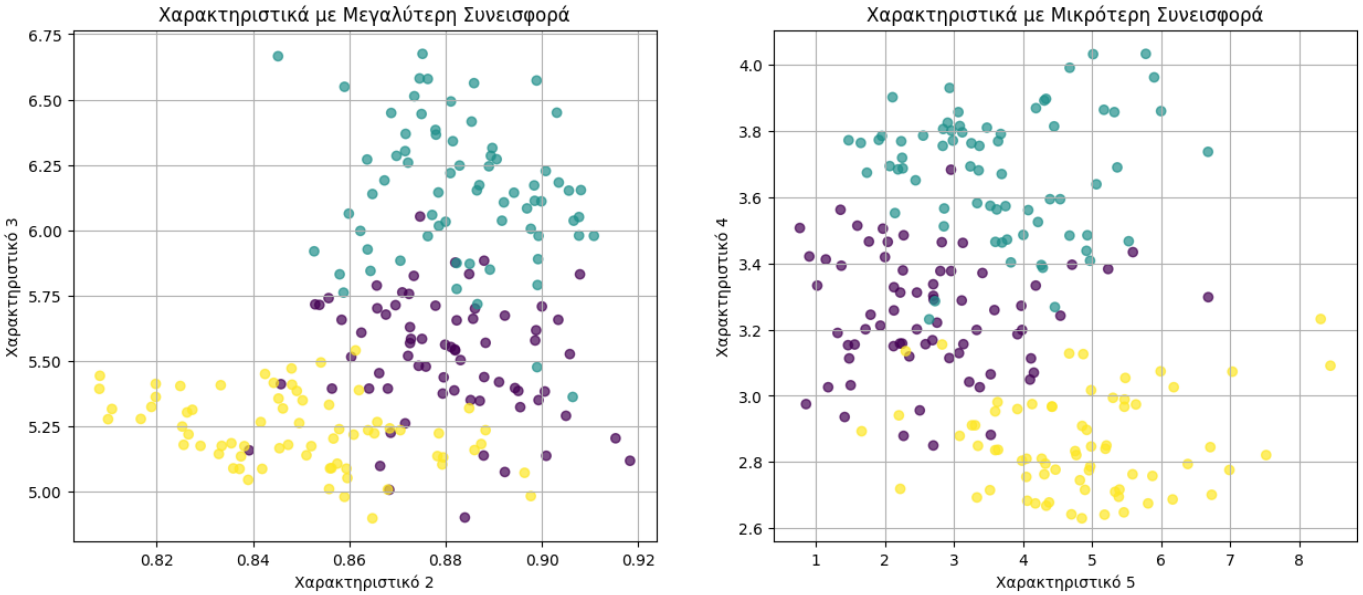
Για κανονικοποιημένα δεδομένα, τα χαρακτηριστικά με την μέγιστη συνεισφορά είναι το 1 και το 0 ενώ εκείνα με την μικρότερη συνεισφορά είναι το 4 και το 5. Αξιοσημείωτο είναι ότι εάν δεν κανονικοποιήσω τα δεδομένα διαφοροποιούνται μονάχα τα χαρακτηριστικά με την μέγιστη συνεισφορά, πλέον είναι τα 2 και 3.

Για τα διαγράμματα με την μεγαλύτερη συνεισφορά παρατηρώ ότι οι κατηγορίες διαχωρίζονται με μεγάλη σαφήνεια, σε αντίθεση με το διάγραμμα με την μικρότερη συνεισφορά. Συνοψίζοντας, τα χαρακτηριστικά με την μεγαλύτερη συνεισφορά παρέχουν πιο σαφή διαχωρισμό συγκριτικά με εκείνα με την μικρότερη συνεισφορά, όπως ακριβώς αναμενόταν. Γίνεται επομένως κατανοητό ότι η επιλογή των κατάλληλων διαστάσεων για ανάλυση και οπτικοποίηση των δεδομένων είναι πολύ σημαντική.

Στο σημείο αυτό σημειώνω ότι ο αλγόριθμος της LDA επιδιώκοντας να μειώσει τις διαστάσεις προκειμένου να μεγιστοποιήσει την διακριτική ικανότητα μεταξύ των κλάσεων, εάν εφαρμοστεί σε κανονικοποιημένα δεδομένα για διδιάστατη απεικόνιση, ο προσανατολισμός των δεδομένων θα είναι η ευθεία. Το αποτέλεσμα αυτό οφείλεται στην κανονικοποίηση των δεδομένων ώστε κάθε χαρακτηριστικό να συμβάλλει ισομερώς στον διαχωρισμό των κλάσεων.



Εικόνα 9: Απεικόνιση δεδομένων για χώρο που ορίζεται από τα χαρακτηριστικά με μέγιστη (αριστερά) και ελάχιστη (δεξιά) συνεισφορά για εφαρμογή LDA σε κανονικοποιημένα δεδομένα



Εικόνα 10: : Απεικόνιση δεδομένων για χώρο που ορίζεται από τα χαρακτηριστικά με μέγιστη (αριστερά) και ελάχιστη (δεξιά) συνεισφορά για εφαρμογή LDA σε μη κανονικοποιημένα δεδομένα

ΑΣΚΗΣΗ 4

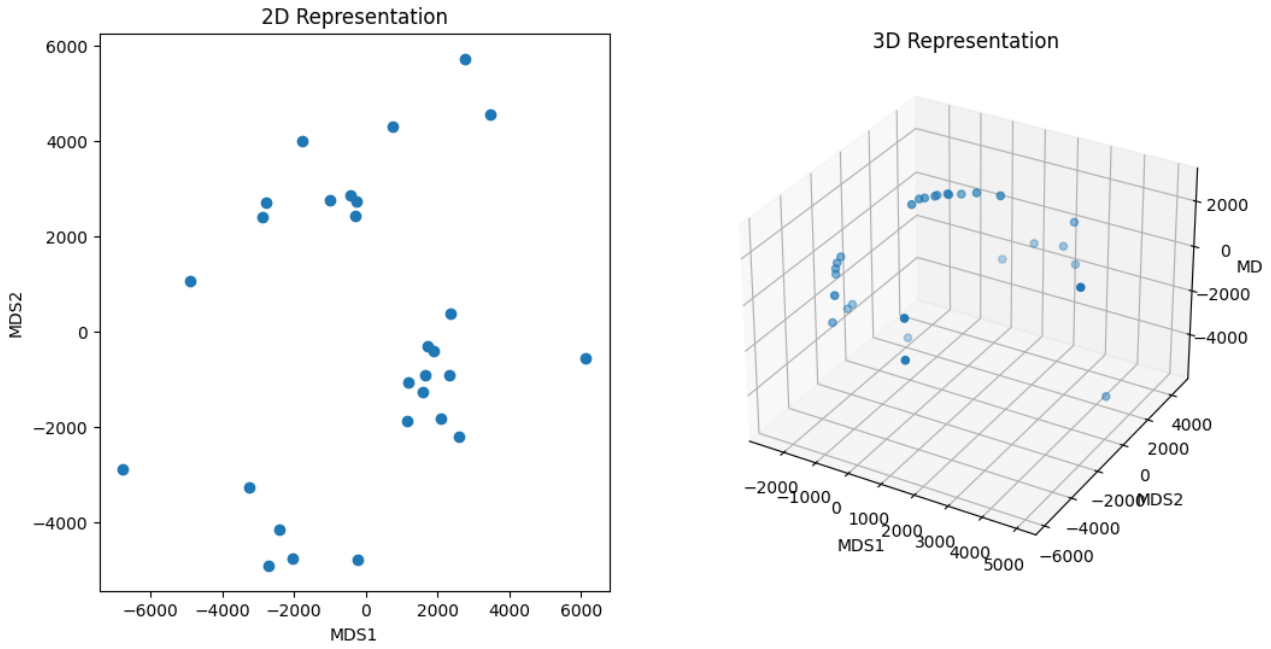
Κατεβάστε από https://drive.google.com/file/d/1SoOZUfqH3ek9_sNjEQE3bi8fJ1_EnAUL/view το dataset «Air Distances Between Cities in Statute Miles» που περιέχει την απόσταση μεταξύ πόλεων είτε του κόσμου είτε των Ηνωμένων Πολιτειών Αμερικής μετρούμενων σε μίλια κατά μήκος μέγιστων κύκλων.

Για την ανάγνωση των δεδομένων ακολουθώ την ίδια διαδικασία με τις προηγούμενες ασκήσεις.

ΕΡΩΤΗΜΑ Α

Χρησιμοποιείστε τη μέθοδο classical MDS για να δημιουργήσετε μία διανυσματική αναπαράσταση των πόλεων του κόσμου (Distance_Matrix_world) στις δύο και στις τρεις διαστάσεις. Απεικονίστε τις αναπαραστάσεις αυτές σε κατάλληλο διάγραμμα και σχολιάστε το αποτέλεσμα.

Η Classical Multidimentional Scaling (MDS) είναι στατιστική τεχνική που χρησιμοποιείται για την μείωση των διαστάσεων των δεδομένων εστιάζοντας στον διαχωρισμό των δεδομένων διατηρώντας όσο το δυνατόν καλύτερα τις αποστάσεις μεταξύ των σημείων. Εφόσον διατίθεται ο πίνακας αποστάσεων κάθε ζεύγους, η μέθοδος υπολογίζει την ευκλείδεια απόσταση μεταξύ των διαφόρων σημείων και αποθηκεύει τα αποτελέσματα σε έναν πίνακα. Από τον πίνακα αυτόν υπολογίζονται οι ιδιοτιμές και τα αντίστοιχα ιδιοδιανύσματα που αντιστοιχούν στον αριθμό των διαστάσεων που επιθυμούμε προς αναπαράσταση. Η διαδικασία αυτή υλοποιείται με την συνάρτηση `MDS` της `sklearn`.



Εικόνα 11: 2D (αριστερά) και 3D (δεξιά) αναπαράσταση των αποστάσεων των πόλεων του κόσμου με την μέθοδο MDS.

Στην εικόνα 11 κάθε σημείο απεικονίζει μία πόλη. Εφόσον η μείωση των διαστάσεων προς απεικόνιση έγινε με την μέθοδο MDS, η σχετική θέση και η απόσταση των σημείων αντανακλούν τις πραγματικές αποστάσεις. Η διαφορά των δύο απεικονίσεων έγκειται στο γεγονός ότι η τριδιάστατη απεικόνιση προσφέρει πιο πλήρη εικόνα των αποστάσεων καθώς η απεικόνιση γίνεται σε περισσότερες διαστάσεις.

ΕΡΩΤΗΜΑ Β

Δημιουργείστε για τις πόλεις μία διανυσματική αναπαράσταση με τις μέγιστες διαστάσεις d που μπορεί να σας επιστρέψει ο αλγόριθμος MDS.

Ο μέγιστος αριθμός διαστάσεων d είναι n-1, όπου n το πλήθος των πόλεων. Στην παρούσα εργασία, εφόσον έχω 28 πόλεις, ο μέγιστος αριθμός διαστάσεων είναι 27.

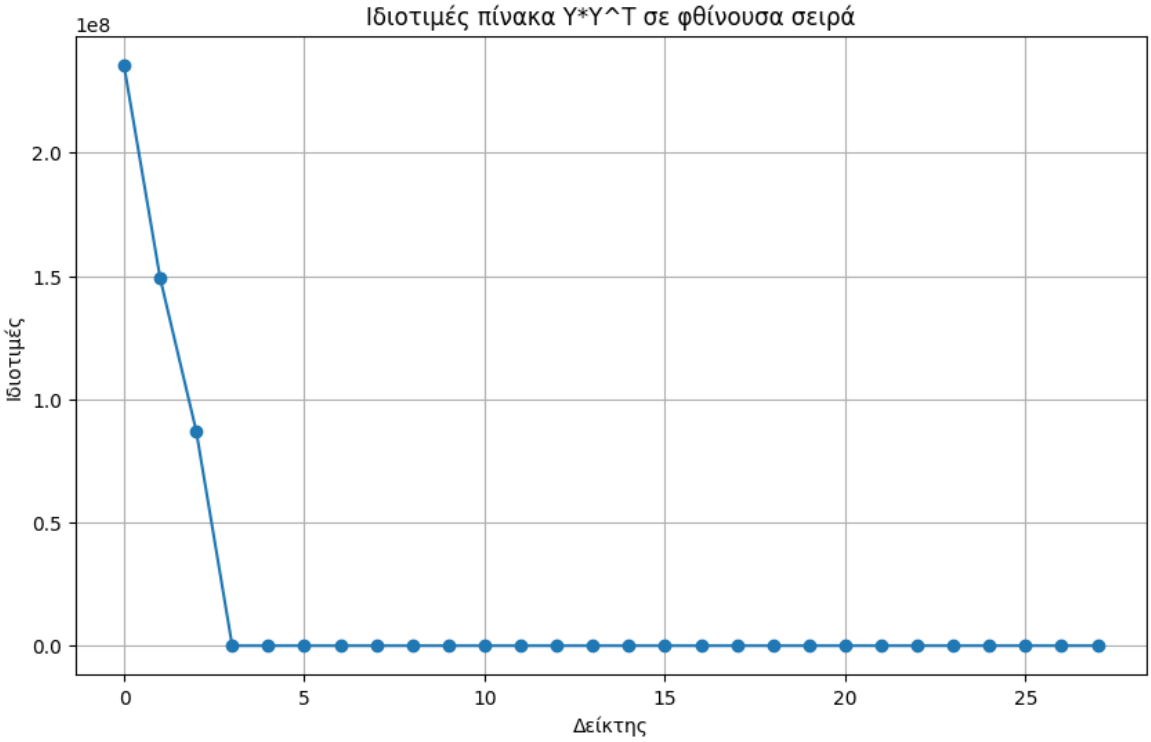
Εάν $Y \in R^{N \times d}$ ο πίνακας με τις αναπαραστάσεις για τις N πόλεις, να δημιουργήστε το διάγραμμα των ιδιοτιμών του πίνακα $Y \times Y^T$ σε φθίνουσα σειρά.

Στο διάγραμμα της εικόνας 12 φαίνονται οι ιδιοτιμές του πίνακα $Y \times Y^T$ ταξινομημένες σε φθίνουσα σειρά. Ο κάθετος άξονας δηλώνει την τιμή ενώ ο οριζόντιος την θέση της στην ταξινομημένη σειρά. Οι πρώτες

ιδιοτιμές (αριστερό μέρος του διαγράμματος) έχουν τις μεγαλύτερες τιμές, άρα οι συγκεκριμένες διαστάσεις περιέχουν σημαντική πληροφορία για τα δεδομένα. Προχωρώντας προς τα δεξιά, οι ιδιοτιμές μικραίνουν, δηλώνοντας ότι οι αντίστοιχες διαστάσεις συμβάλουν λιγότερο στην συνολική δομή έναντι των πρώτων. Το διάγραμμα αυτό βοηθάει στην κατανόηση του πόσες διαστάσεις είναι σημαντικές για να περιγράψουν τα δεδομένα διατηρώντας μονάχα την σημαντική πληροφορία.

Το διάγραμμα αυτό χρησιμοποιείται ως ένδειξη της βέλτιστης διάστασης αναπαράστασης, διατηρώντας τόσες διαστάσεις όσες οι σημαντικές ιδιοτιμές. Με βάση αυτό, πόσες διαστάσεις εκτιμάτε ότι είναι οι βέλτιστες για τα δεδομένα του αρχείου Distance_Matrix_world;

Ο βέλτιστος αριθμός των διαστάσεων προς αναπαράσταση καθορίζεται από το σημείο όπου οι ιδιοτιμές δεν παρουσιάζουν δραστική μείωση. Στο παρόν πρόβλημα, τα δεδομένα μπορούν να περιγράφουν χωρίς περιττή πληροφορία και θόρυβο με μόλις 3 διαστάσεις, καθώς οι επόμενες ιδιοτιμές έχουν μηδενική τιμή.



Εικόνα 12: Διάγραμμα ιδιοτιμών πίνακα $Y \times Y^T$ σε φθίνουσα σειρά για τα δεδομένα των πόλεων του κόσμου

BONUS

Για ποιο λόγο υπάρχουν μη μηδενικές ιδιοτιμές για περισσότερες των 3 διαστάσεων στο παραπάνω πρόβλημα;

Το μήκος μέγιστων κύκλων αναφέρεται στην απόσταση μεταξύ δύο σημείων πάνω στην επιφάνεια της σφαίρας υπολογισμένη κατά μήκος του συντομότερου δρόμου. Σε αντίθεση με τις ευθείες αποστάσεις που μετράμε σε έναν επίπεδο χάρτη, οι αποστάσεις μέγιστων κύκλων λαμβάνουν υπόψη την καμπυλότητα της Γης και παρέχουν μια πιο ακριβή μέτρηση για μεγάλες αποστάσεις.

Οι αποστάσεις του αρχείου, λοιπόν, είναι μη ευκλείδειες, άρα απαιτούν περισσότερες διαστάσεις για να αποτυπωθούν στον ευκλείδειο χώρο. Οι τιμές των ιδιοτιμών δεν είναι ακριβώς μηδέν, αλλά κοντά στο μηδέν, δηλώνοντας ότι οι διαστάσεις αυτές περιέχουν σημαντική πληροφορία για την δομή των δεδομένων.

Συγκρίνετε το αντίστοιχο διάγραμμα για τα δεδομένα του αρχείου Distance_Matrix_US. Που οφείλεται αυτή η διαφορά και πως σχετίζεται με τη φύση του προβλήματος και των δεδομένων;

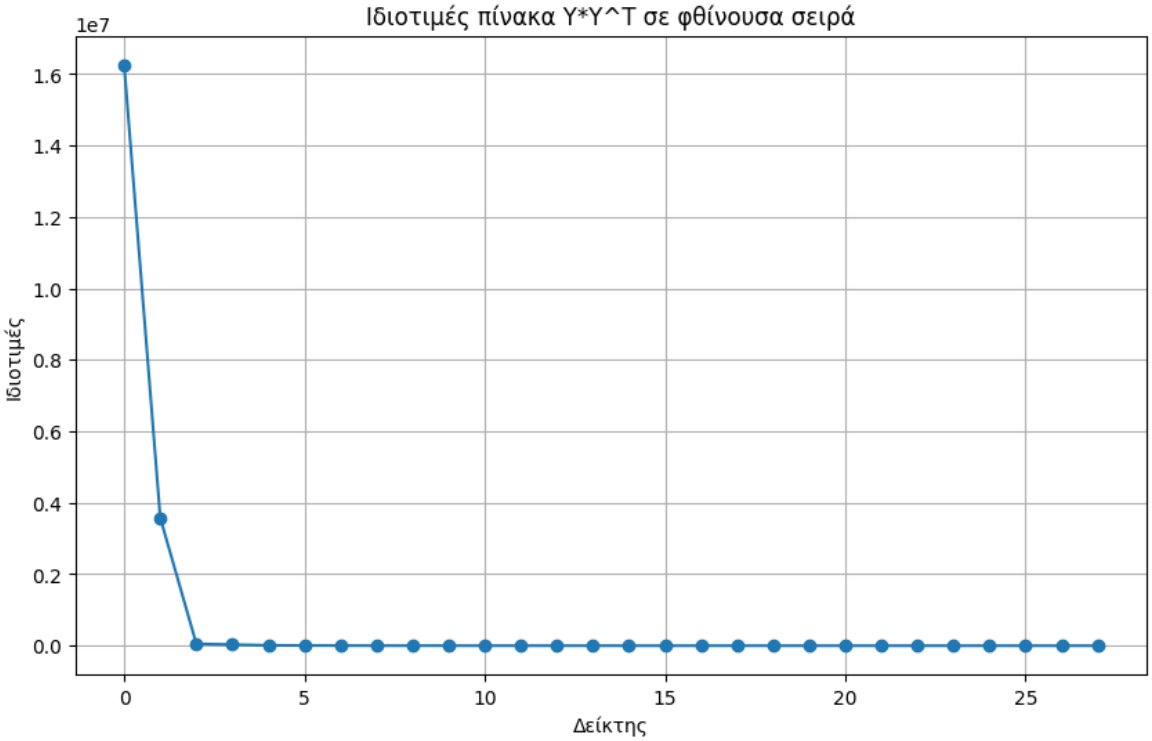
Σύμφωνα με την εικόνα 13, τα δεδομένα για τις πόλεις των ΗΠΑ μπορούν να περιγράφουν χωρίς περιττή πληροφορία και θόρυβο με μόλις 2 διαστάσεις, καθώς οι επόμενες ιδιοτιμές έχουν μηδενική τιμή. Η διαφορά των βέλτιστων διαστάσεων για τα δύο σύνολα δεδομένων, πόλεις του κόσμου και πόλεις των ΗΠΑ, οφείλεται στους ακόλουθους λόγους:

- **Προβολή μήκους μέγιστου κύκλου**

Οι αποστάσεις μεταξύ πόλεων σε παγκόσμιο επίπεδο μπορεί να είναι πιο περίπλοκες και να μην αντιπροσωπεύονται εύκολα σε έναν ευκλείδειο χώρο, λόγω της καμπυλότητας της Γης.

- **Διαφορετική Διακύμανση**

Το πρώτο σύνολο δεδομένων καλύπτει πόλεις σε παγκόσμιο επίπεδο περιλαμβάνοντας μεγαλύτερες αποστάσεις και περισσότερη διακύμανση σε αυτές έναντι του δεύτερου συνόλου που περιορίζεται σε γεωγραφικό πλάτος και μήκος.



Εικόνα 13: Διάγραμμα ιδιοτιμών πίνακα $Y \times Y^T$ σε φθίνουσα σειρά για τα δεδομένα των πόλεων των ΗΠΑ