Τμήμα Μηχανικών Η/Υ & Πληροφορικής
Πανεπιστήμιο Ιωαννίνων

Λ.07 – Εξόρυξη Δεδομένων

2η Σειρά Ασκήσεων

Ensemble Learning και Ομαδοποίηση

Μπάτση Σοφία Α.Μ.: 372
Δημητριάδης Σωκράτης Α.Μ.: 359

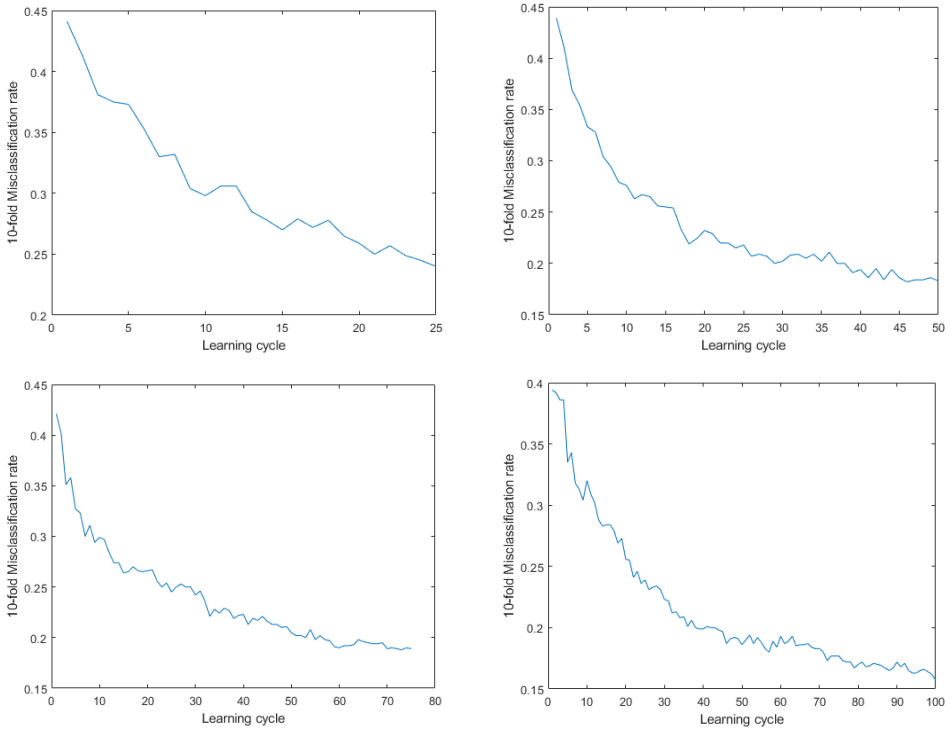
➢ Εισαγωγή

Στην παρούσα σειρά ασκήσεων, αντιμετωπίζεται το *πρόβλημα της ταξινόμησης* και το *πρόβλημα της ομαδοποίησης δεδομένων*. Η εύρεση λύσης στο πρώτο πρόβλημα, επιδιώκεται κάνοντας χρήση των μεθόδων ***Bagging*** και ***Random Forest***, του ***Ensemble Learning***, υπό ορισμένες παραμέτρους της λειτουργίας των. Για το δεύτερο πρόβλημα, χρησιμοποιήθηκαν οι μέθοδοι ***k-means, agglomerative*** και ***spectral clustering***, επίσης με τις ζητούμενες παραμέτρους λειτουργίας των. Οι μέθοδοι για το πρώτο πρόβλημα, εφαρμόστηκαν διαδοχικά στο δοθέν σύνολο δεδομένων της **1ης Σειράς Ασκήσεων**, ενώ για το δεύτερο εφαρμόστηκαν διαδοχικά στα σύνολα παραδειγμάτων που δόθηκαν για αυτό το σκοπό.

➢ Ταξινόμηση με τις μεθόδους *Bagging* και *Random Forest*:

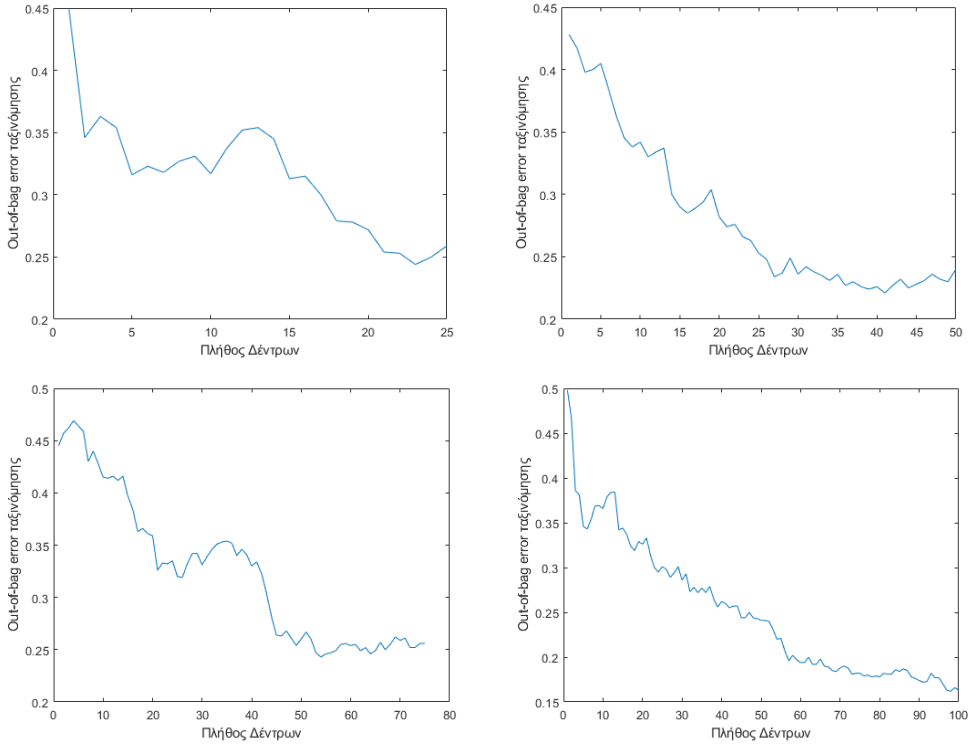
- ***Bagging*:**

  Για το δοθέν αρχείο δεδομένων, εκπαιδεύτηκαν τέσσερα μοντέλα ταξινόμησης με πλήθος ταξινομητών 25, 50, 75 και 100, αντίστοιχα. Η μέτρηση της απώλειας κάθε μοντέλου, έγινε με τη χρήση του cross-validation. Το ελάχιστο σφάλμα ταξινόμησης στις αντίστοιχες περιπτώσεις είναι 0.2150, 0.1670, 0.1610 και 0.1600. Οι αντίστοιχες γραφικές παραστάσεις του σφάλματος συναρτήσει του πλήθους ταξινομητών δίνονται παρακάτω:



- ***Random Forest***:

  Για το δοθέν αρχείο δεδομένων, εκπαιδεύτηκαν τέσσερα μοντέλα με πλήθος δέντρων απόφασης 25, 50, 75 και 100, αντίστοιχα. Η μέτρηση της ικανότητας γενίκευσης, έγινε με τη χρήση του out-of-bag error και ως ελάχιστο πλήθος φύλλων ορίστηκε το 'πέντε'(5). Το ελάχιστο oob σφάλμα ταξινόμησης των αντίστοιχων μοντέλων είναι 0.2440, 0.2210, 0.2430 και 0.1620. Οι αντίστοιχες γραφικές παραστάσεις του σφάλματος συναρτήσει του πλήθους ταξινομητών φαίνονται παρακάτω:



Παρατηρούμε σε σχέση με τα αποτελέσματα της **1ης Σειράς Ασκήσεων**, ότι τα τέσσερα μοντέλα ταξινόμησης, αποτελούμενα από 25, 50, 75 και 100 δέντρα απόφασης αντίστοιχα, έχουν καλύτερη απόδοση από τα μοντέλα που σχεδιάστηκαν με την ***k-NN*** μέθοδο (κατά επέκταση καλύτερη από ***Naïve-Bayes*** και ***SVM***). Ενδεικτικά, υπενθυμίζεται ότι οι ταξινομητές με τη μικρότερη απώλεια ήταν αυτοί με μετρική την Ευκλείδεια και πλήθος εννέα (9) γειτόνων, επιτυγχάνοντας μέγεθος απώλειας περίπου 24-26%. Τα μοντέλα της παρούσης αναφορικής παραγράφου, επιτυγχάνουν μικρότερες απώλειες στις περιπτώσεις των 50 και 100 ταξινομητών, ενώ περίπου ίσες απώλειες σε αυτές των 25 και 75.

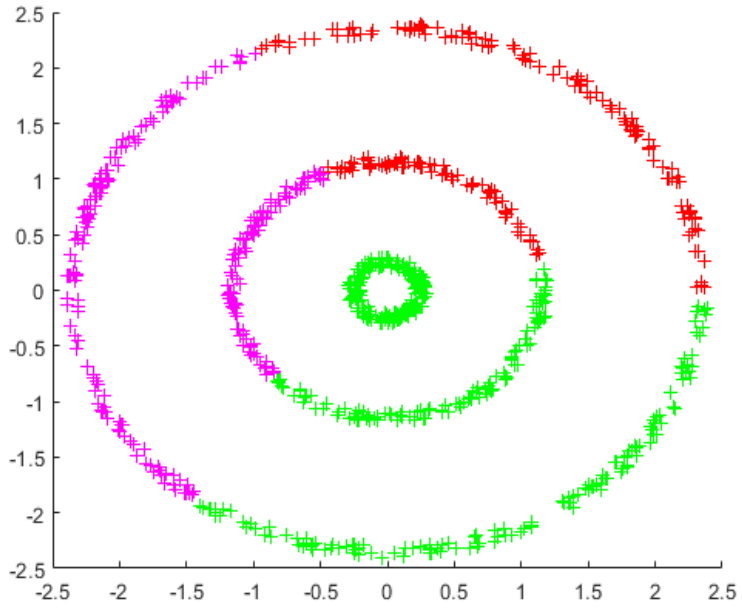
Συνεπώς, και συγκρίσει των ταξινομητών που προέκυψαν με τη μέθοδο ***Bagging***, σε κάθε περίπτωση πλήθους ταξινομητών, η απώλεια της μεθόδου είναι καλύτερη από αυτή της ***Random Forest***, με μοναδικούς ενδοιασμούς τις περιπτώσεις των 25 και 50 ταξινομητών. Κατά επέκταση, τα μοντέλα ταξινόμησης με τη μέθοδο ***Bagging***, είναι σαφώς αποδοτικότερα από αυτά που υλοποιήθηκαν στην **1η Σειρά Ασκήσεων** με εξαίρεση (ίσως) του μοντέλου των 25 ταξινομητών.

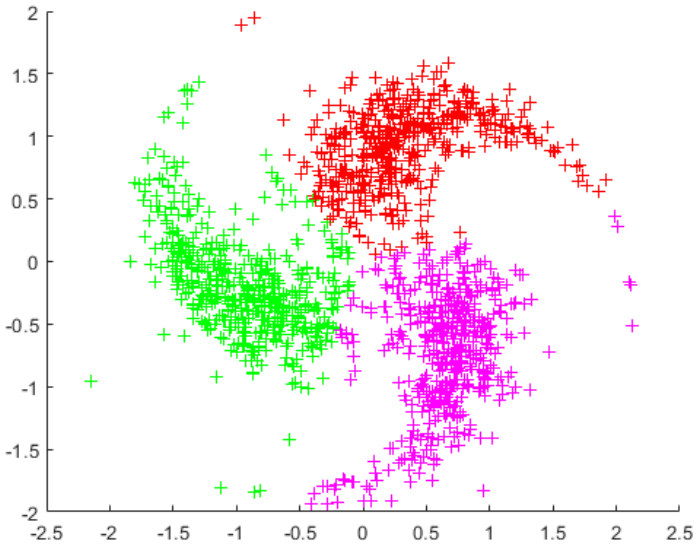
➢ **Ομαδοποίηση:**

Η εφαρμογή των μεθόδων ομαδοποίησης, έγινε στα σύνολα δεδομένων που προαναφέρθηκαν, με πλήθος ομάδων το πραγματικό (αναγραφόμενος ακέραιος σε κάθε σύνολο). Έτσι για κάθε μέθοδο, προέκυψαν τα εξής αποτελέσματα:

- ***k-means*:**

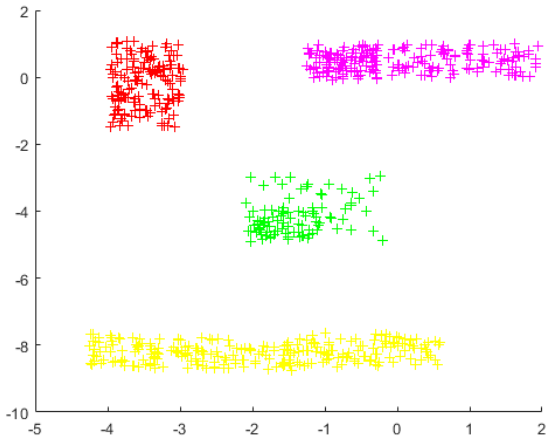
Με k = 3, για τα σύνολα '3rings' και '3wings', έχουμε αντίστοιχα τις εξής απεικονίσεις:





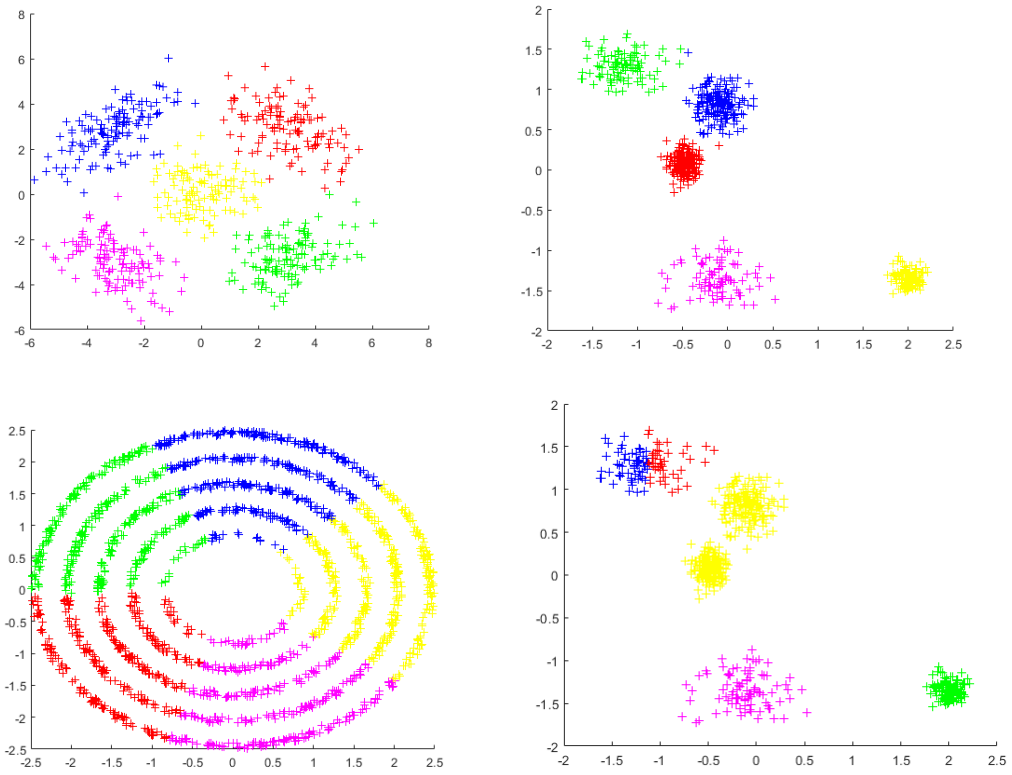
Στην πρώτη απεικόνιση, φαίνεται η λανθασμένη ομαδοποίηση που έχει γίνει, καθώς οι ομάδες δεν καταφέρνουν να διαχωριστούν ορθώς. Στην δεύτερη, αν και διακρίνονται καλύτερα οι ομάδες, παρατηρούνται ατέλειες στις άκρες και στο εσωτερικό των ‘’φτερών’’.

Με k = 4 και για το σύνολο '4rectangles' έχουμε:



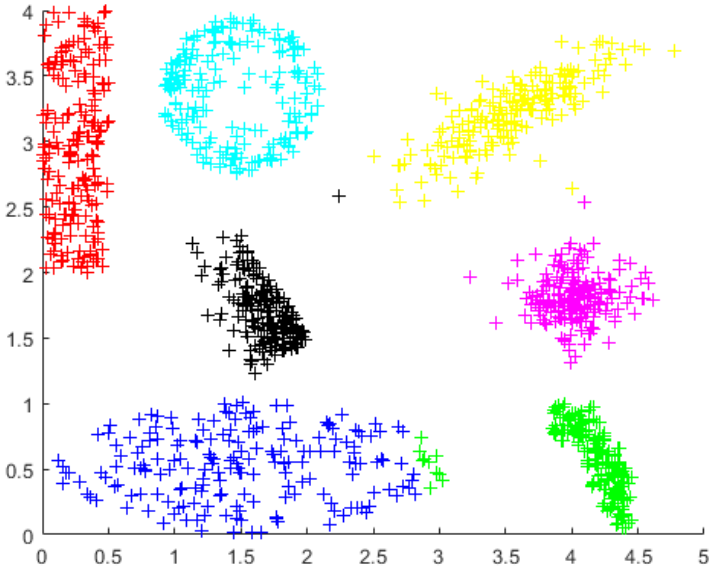
Σε αυτή την περίπτωση λοιπόν, φαίνεται η σωστή διάκριση των ομάδων.

Με k = 5 για τα σύνολα '5clusters', '5Gaussians' και '5rings', έχουμε τις αντίστοιχες γραφικές απεικονίσεις:

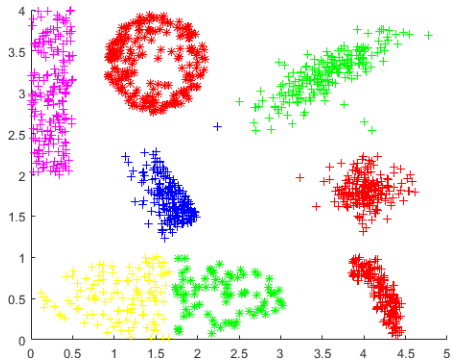


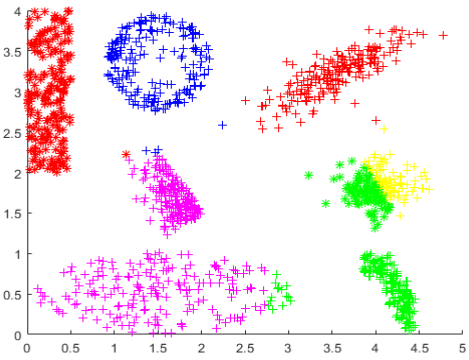
Στην πρώτη απεικόνιση, οι ομάδες διακρίνονται σχεδόν επιτυχώς. Στην τρίτη, όπως ήταν αναμενόμενο, ο ***k-means*** δεν καταφέρνει να διαχωρίσει τους δακτυλίους. Στην δεύτερη απεικόνιση, φαίνεται ότι οι ομάδες διαχωρίζονται σχεδόν επιτυχώς. Αυτό όμως δε συμβαίνει σε κάθε εκτέλεση του αλγορίθμου, όπως άλλωστε μπορεί να παρατηρηθεί στην τελευταία απεικόνιση.

Με k = 7 για το σύνολο '7clusters', έχουμε:



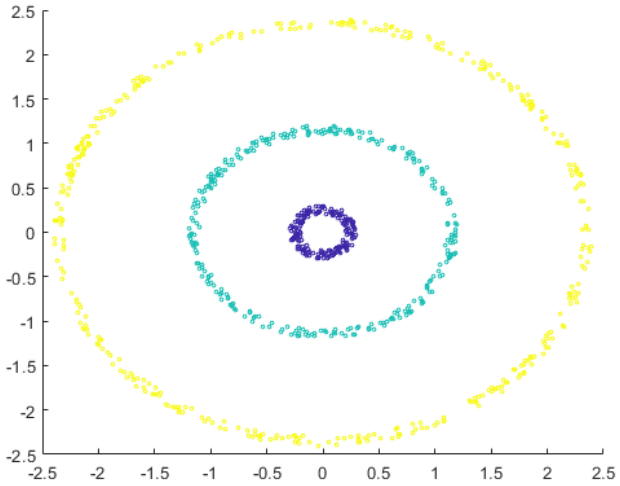
Φαίνεται σε αυτή την εκτέλεση ότι οι ομάδες διαχωρίζονται σχετικά πλήρως. Όμως αυτός δεν είναι ο κανόνας, όπως παρατηρήθηκε σε προηγούμενες περιπτώσεις. Για παράδειγμα στις παρακάτω γραφικές απεικονίσεις, διακρίνεται η αδυναμία σωστής ομαδοποίησης:

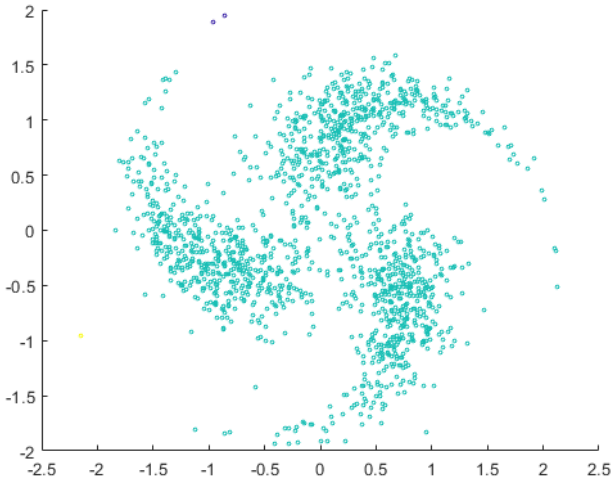


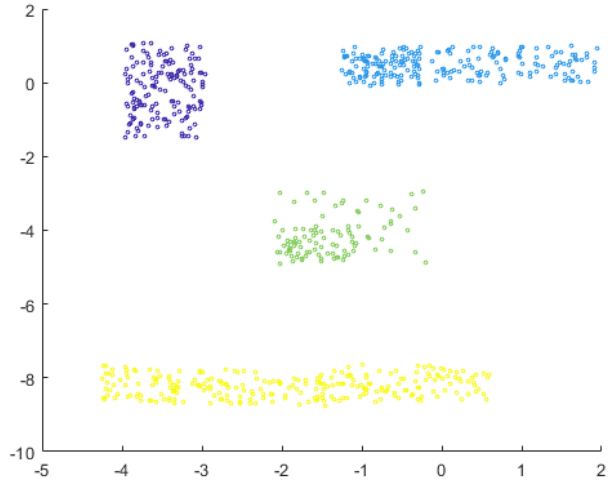


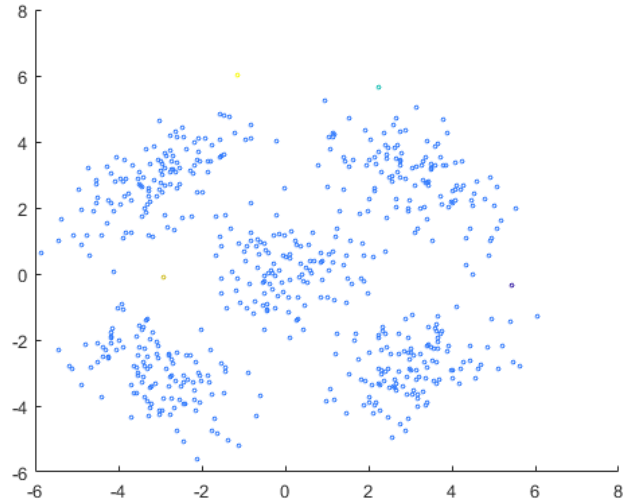
Σημειώνεται ότι με πράσινο και κόκκινο χρώμα, αναγράφονται δύο ειδών ομάδες για κάθε χρώμα. Η διαφοροποίηση είναι στο σύμβολο (+ ή *).

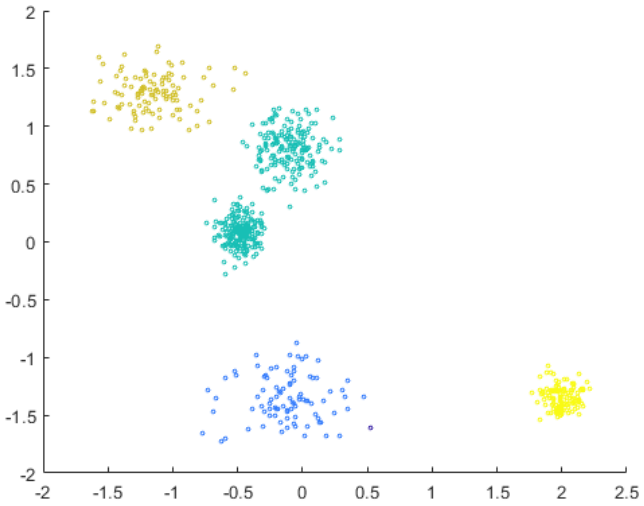
- ***Agglomerative Clustering***:
  Χρησιμοποιώντας ως μετρική απόστασης την Ευκλείδεια και πλήθος ομάδων το πραγματικό, παρατηρήθηκαν τα εξής αποτελέσματα σε κάθε μέθοδο για το αντίστοιχο σύνολο δεδομένων:
  - ***Single Link***:

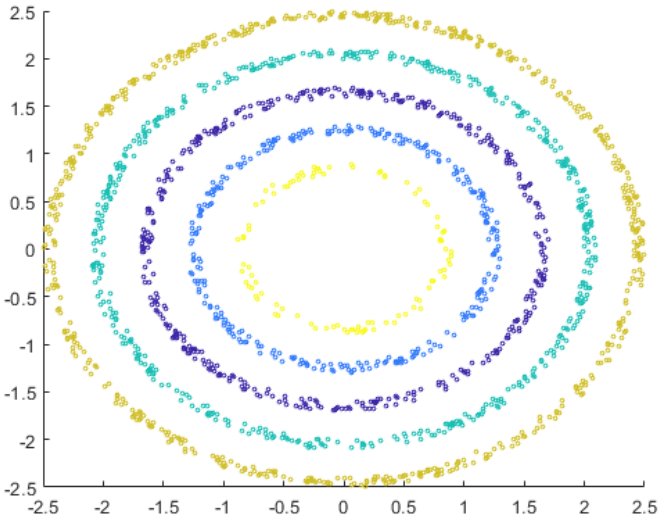


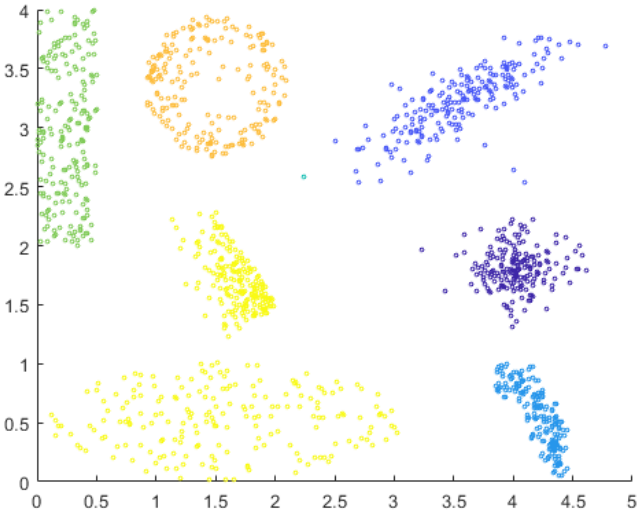




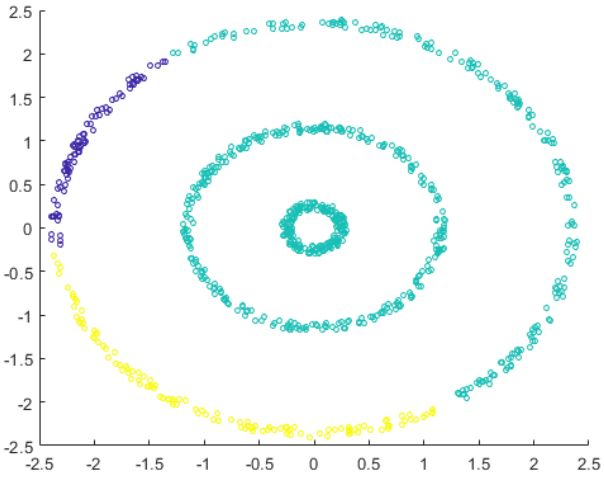

8
6
4
2
0
-2
-4
-6
-6
-4
-2
0
2
4
6
8

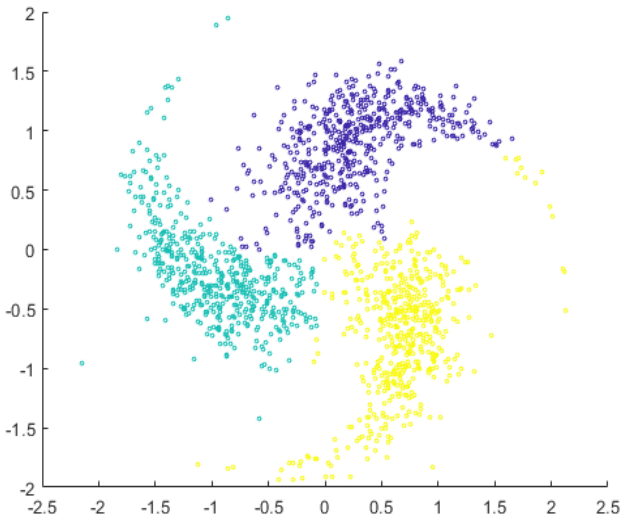

2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5

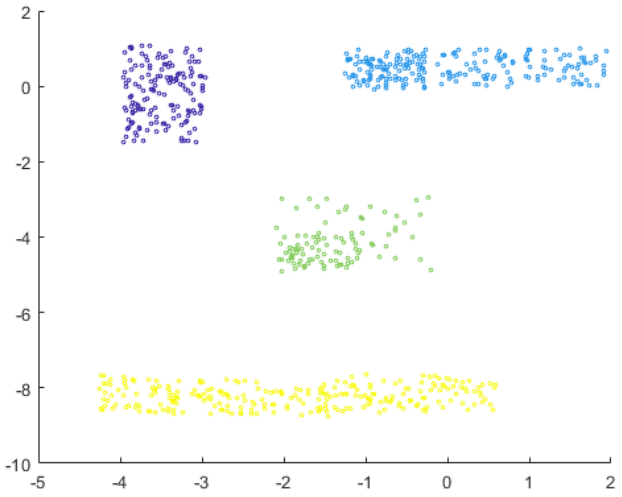

2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5

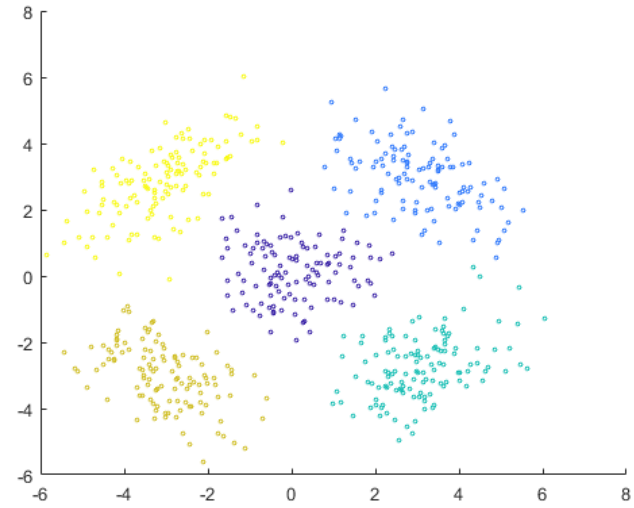


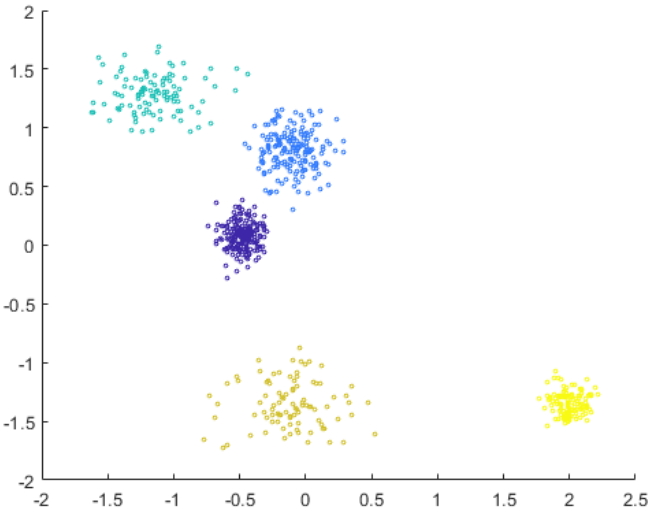
- ***Average Link***:

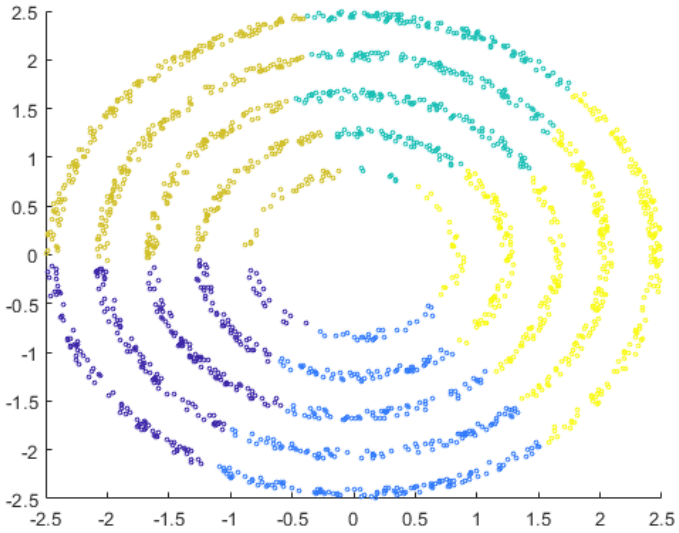


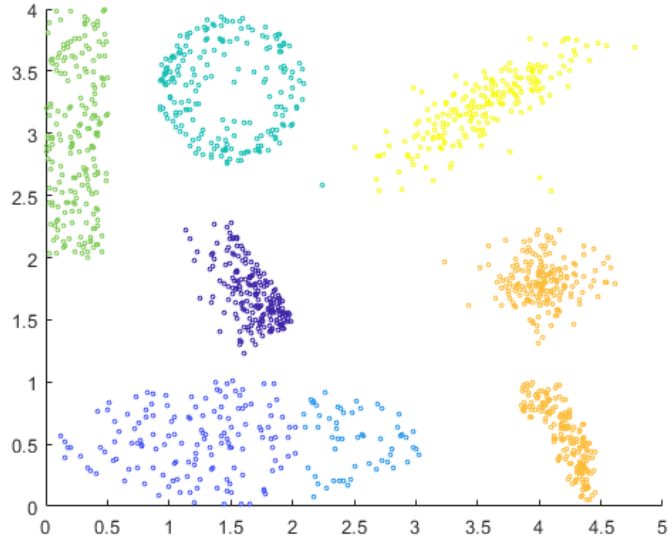



2
0
-2
-4
-6
-8
-10
-5
-4
-3
-2
-1
0
1
2


8
6
4
2
0
-2
-4
-6
-6
-4
-2
0
2
4
6
8


2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5





Παρατηρούνται λοιπόν, ορισμένες διαφορές σε σχέση με την ομαδοποίηση που πραγματοποίησε ο ***k-means***, όσον αφορά τον διαχωρισμό των ομάδων.

Πιο συγκεκριμένα, ο ***Agglomerative Single Link*** ομαδοποιεί σωστά τα σύνολα '3rings', '4rectangles', '5Gaussians', '5rings' και '7clusters', ενώ αποτυγχάνει πλήρως στα '3wings' και '5clusters'. Συνεπώς, "υπερτερεί" του ***k-means*** στα σύνολα με δακτύλιους, ενώ σε σύνολα με κοντινές παρατηρήσεις ή με ομάδες που δεν είναι σαφές ότι ξεχωρίζουν, τείνει να τις συμπεριλαμβάνει στην ίδια ομάδα.

Ο ***Agglomerative Average Link*** ομαδοποιεί σωστά τα σύνολα, '3wings' '4rectangles', '5clusters', '5Gaussians' και '7clusters', ενώ αποτυγχάνει στα '3rings' και '5rings'. Αυτό, ήταν και το αναμενόμενο, λαμβάνοντας υπόψην τον τρόπο υπολογισμού της απόστασης μεταξύ των ομάδων.

- ***Spectral Clustering με RBF πυρήνα***:

Για τις διάφορες τιμές του sigma, 0.1, 0.5 και 1, έχουμε για τα δοθέντα σύνολα τα αντίστοιχα αποτελέσματα:

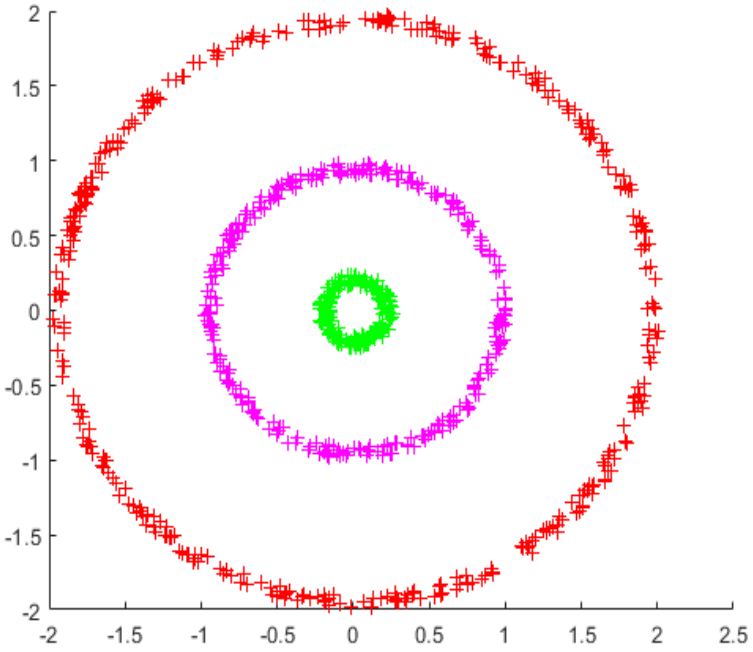
Τα '3rings', '4rectangles', '5rings', ομαδοποιούνται με sigma 0.1, ενώ στις άλλες περιπτώσεις αποτυγχάνει η μέθοδος.

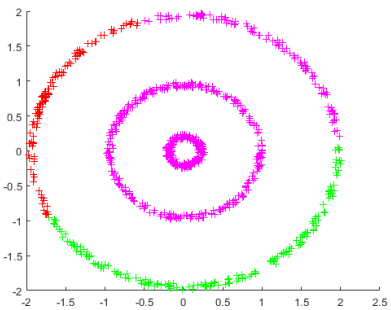
Τα '3wings', '5clusters', '5Gaussians' ομαδοποιούνται σωστά (ή σχεδόν σωστά) με sigma 0.5 και 1.

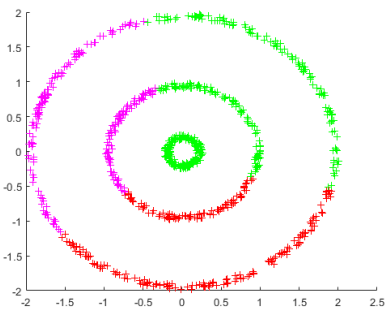
Το '7clusters' δεν ομαδοποιείται επιτυχώς για καμία από αυτές τις τιμές του sigma.

Οι γραφικές απεικονίσεις των μεθόδων για τα αντίστοιχα σύνολα δεδομένων και για το αντίστοιχο sigma από όπου προέκυψαν, καθώς επίσης και τα συμπεράσματα, δίνονται παρακάτω:

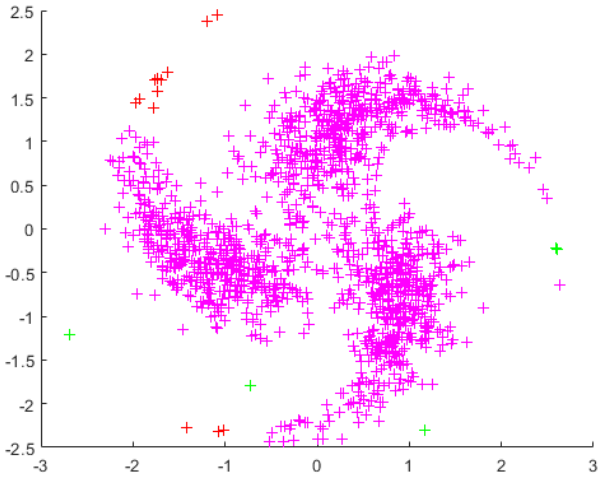
- '3rings':

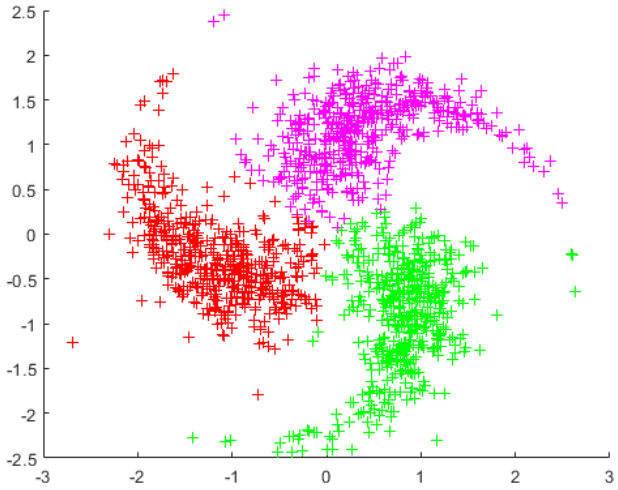


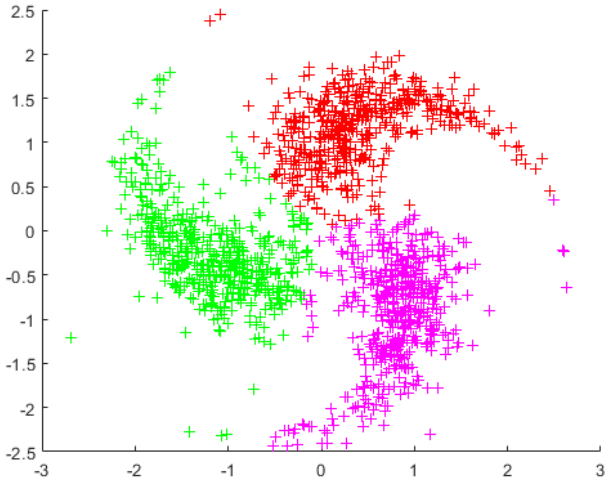




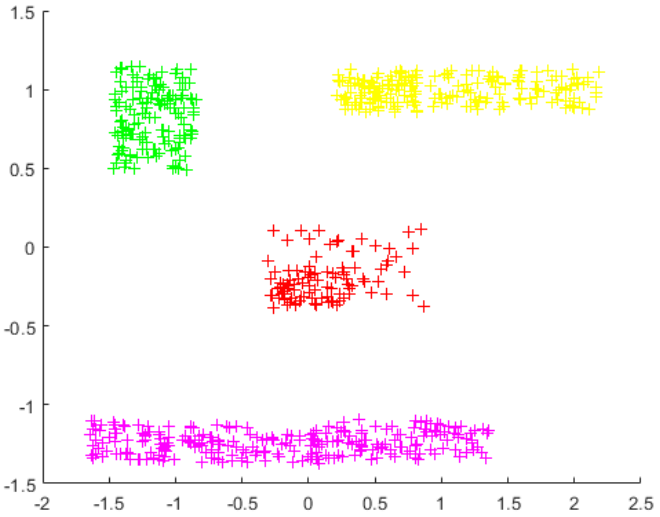
- ‘3wings’:

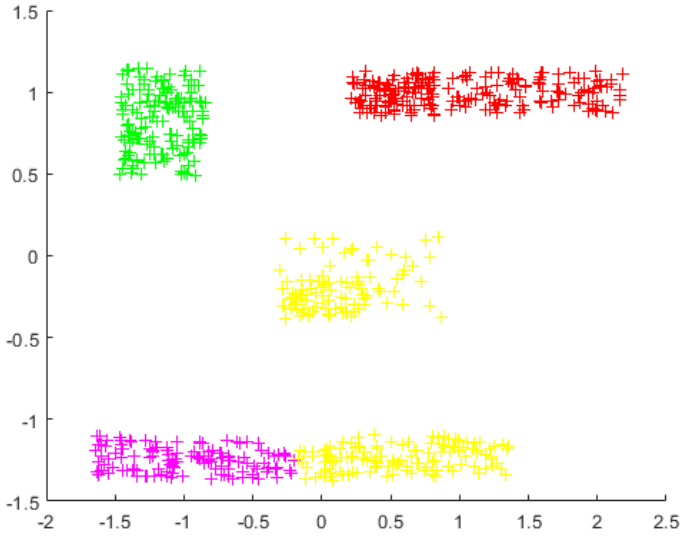

2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-2
-1
0
1
2
3

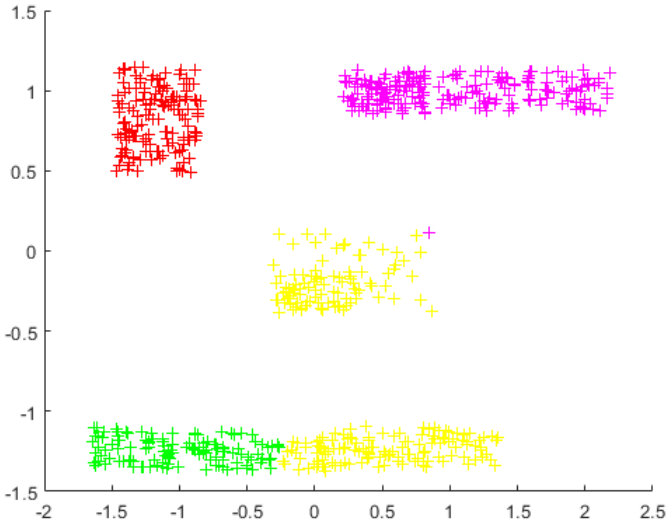

2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-2
-1
0
1
2
3


2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-2
-1
0
1
2
3

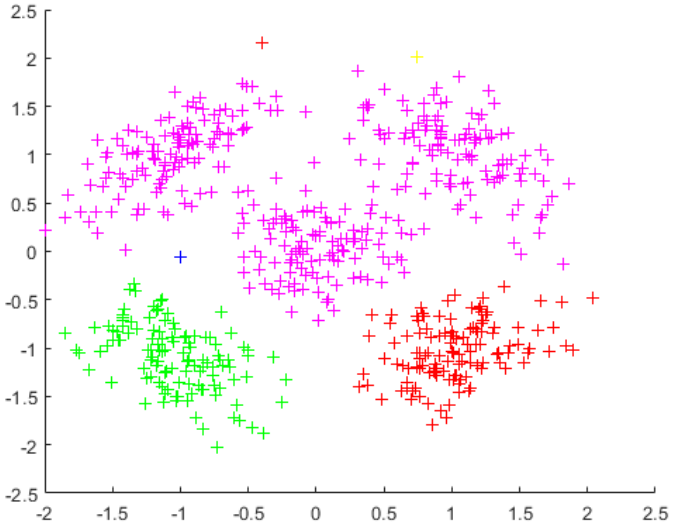
- ‘4rectangles’:

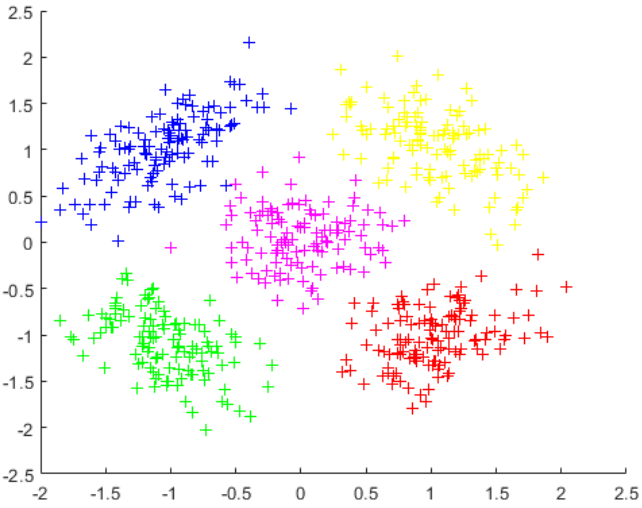


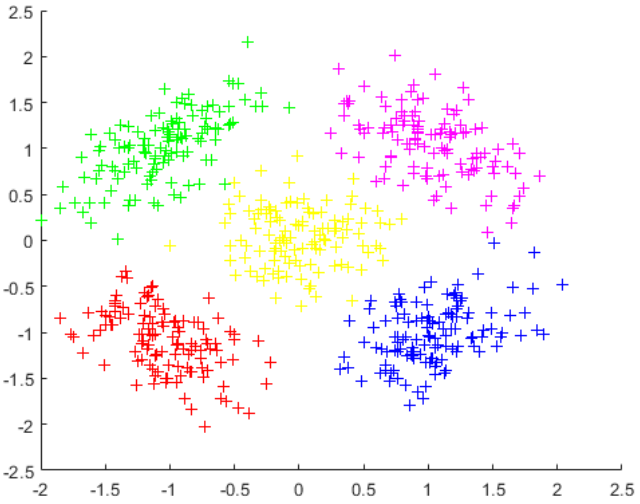




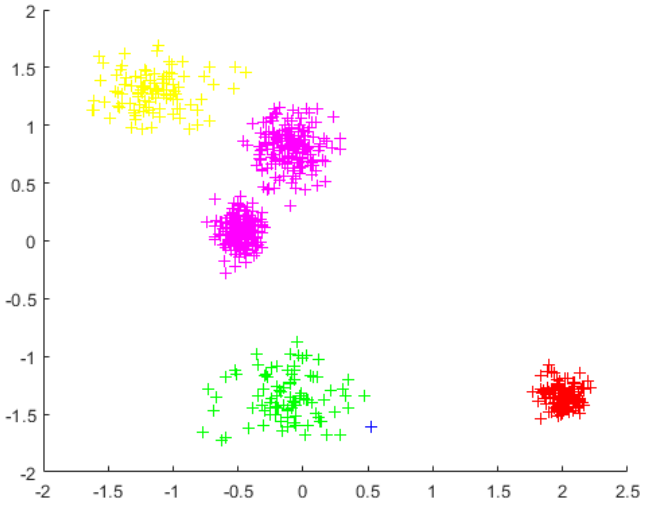
- ‘5clusters’:

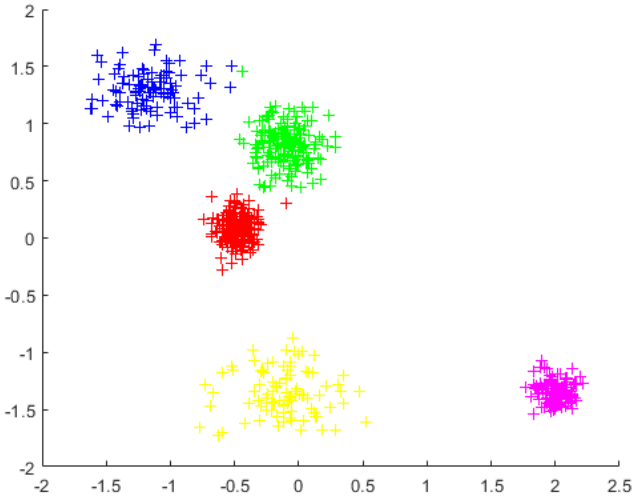

2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5

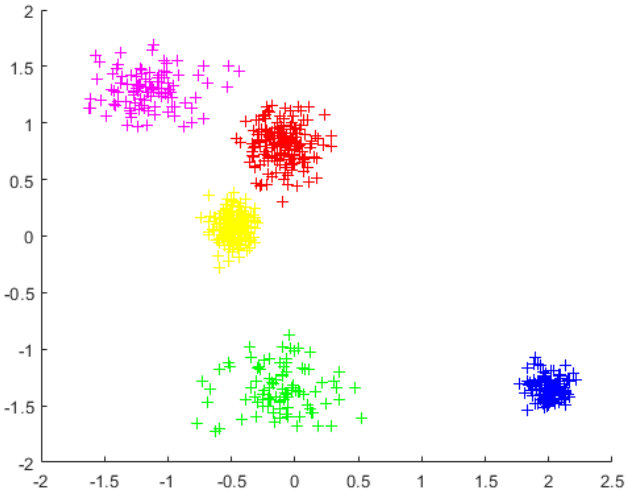

2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5


2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5

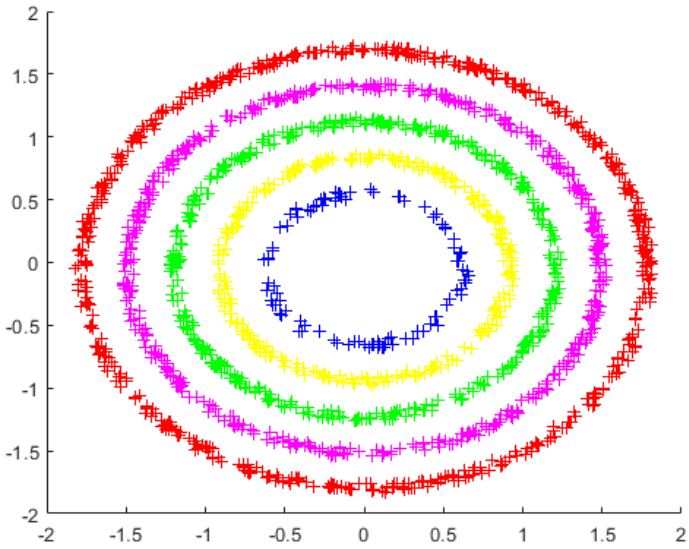
- ‘5Gaussians’:

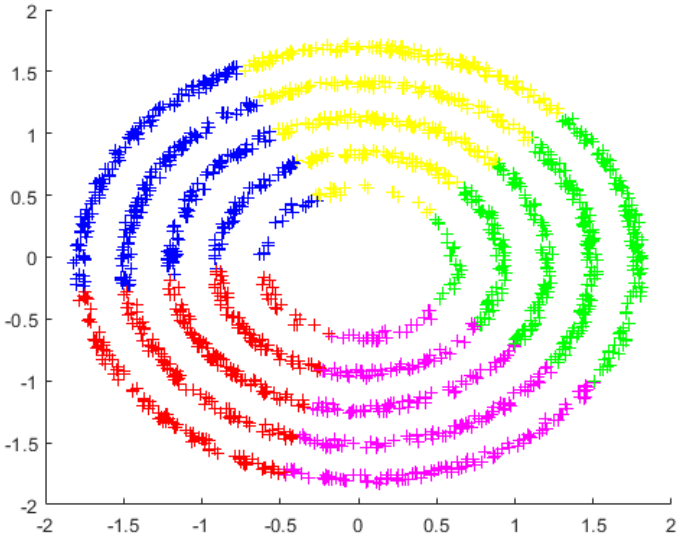

2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5

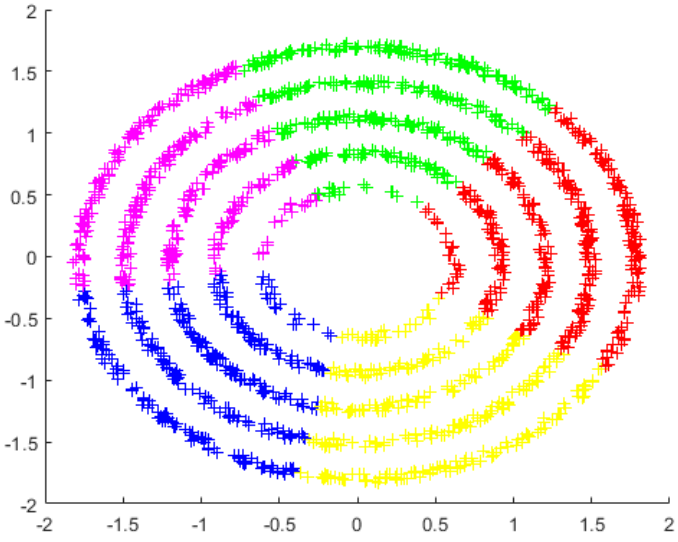

2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5


2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5

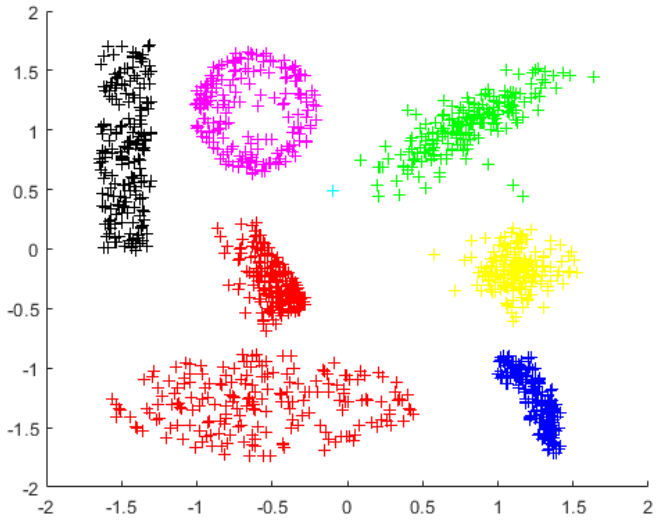
- ‘5rings’

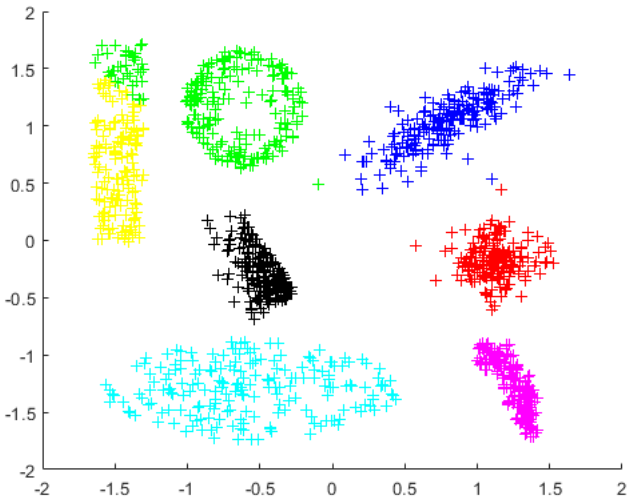

2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

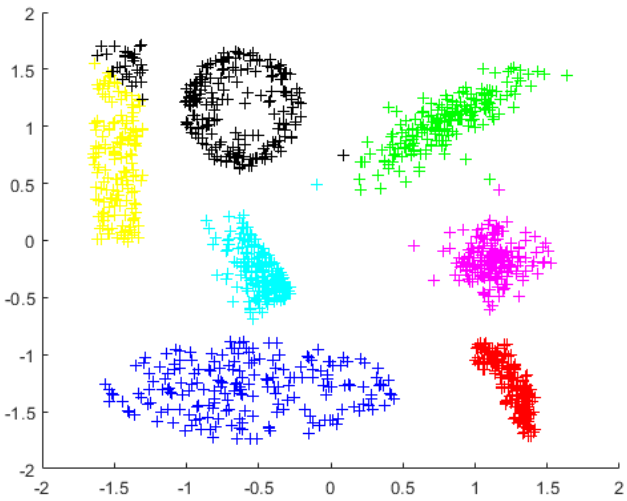

2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2


2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

- ‘7clusters’:


2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

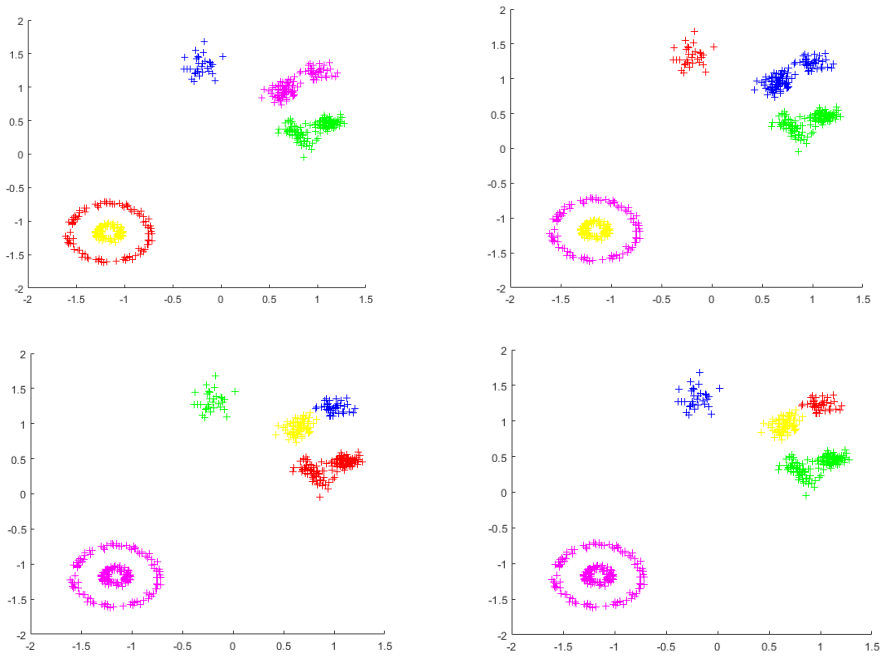

2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2


2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

Για την εύρεση του κατάλληλου sigma εντός του διαστήματος [0.1, 0.4], που ομαδοποιεί σωστά το σύνολο δεδομένων 'gaussian_rings', έγινε η εξής προσέγγιση:

- Αρχικά, εκτελέστηκε ο αλγόριθμος '***spectral***' με είσοδο το αρχείο δεδομένων 'gaussian_rings', όπου ως παράμετρος πλήθους ομάδων τέθηκε υποθετικά ένας ακέραιος (συγκεκριμένα το '3') και ως παράμετρος sigma το 0.1.
- Έχοντας λάβει τα αποτελέσματα της εκτέλεσης του προαναφερθέντος αλγορίθμου και την γραφική απεικόνιση, έγινε η εκτίμηση του πραγματικού πλήθους των ομάδων.
- Υποθέτοντας ότι το φαινομενικό πλήθος των ομάδων (πέντε (5)) είναι και το πραγματικό, υλοποιήθηκε μια επαναληπτική διαδικασία δοκιμής παραμέτρων sigma στο δεδομένο διάστημα με βήμα 0.01.
- Διαπιστώθηκε ότι από το 0.1, δηλαδή την αρχή του διαστήματος, έως και το 0.21, ο αλγόριθμος 'spectral', κατάφερνει να βρει τη λύση.

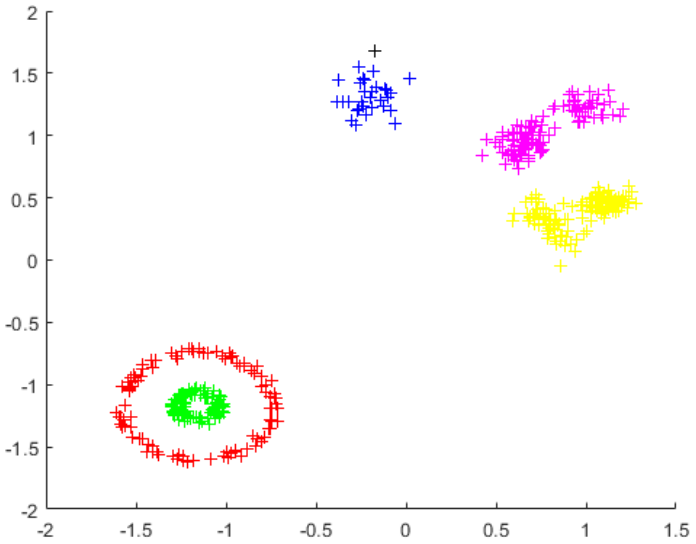
Ώστε για οποιαδήποτε τιμή του sigma εντός του [0.1, 0.21], ο αλγόριθμος 'spectral' βρίσκει τη ζητούμενη λύση. Ενδεικτικά παρατίθενται οι παρακάτω γραφικές απεικονίσεις της λειτουργίας της μεθόδου για τις τιμές του sigma 0.1, 0.2, 0.3 και 0.4 αντίστοιχα:

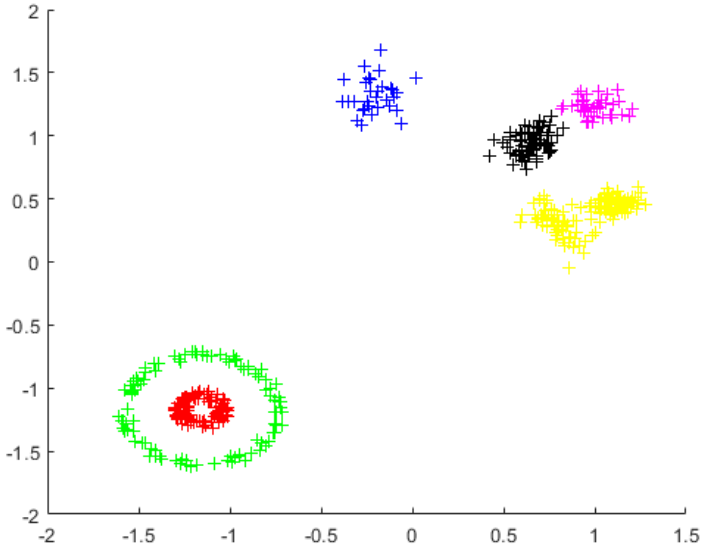


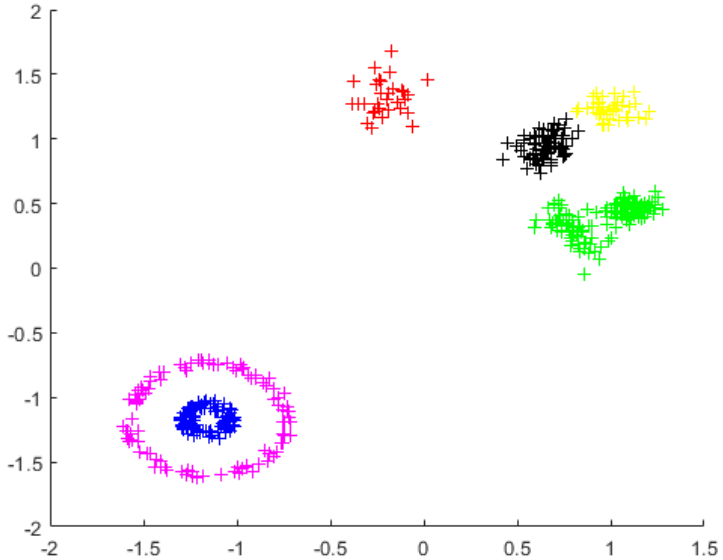
Παρατηρείται, ότι για sigma 0.3 και 0.4, η μέθοδος αποτυγχάνει να ομαδοποιήσει σωστά τα δεδομένα.

Υποθέτοντας εναλλακτικά, ότι οι ομάδες μπορούν να διαχωριστούν περεταίρω σε έξι (6), η παράμετρος sigma κυμαίνεται από 0.15 έως και 0.25. Αυτό προκύπτει θέτοντας ως όρισμα του αλγορίθμου '***spectral***' το 6, στο πλήθος των ομάδων και ύστερα με την επαναληπτική μέθοδο, αναζητώντας το κατάλληλο sigma βάσει των αποτελεσμάτων.

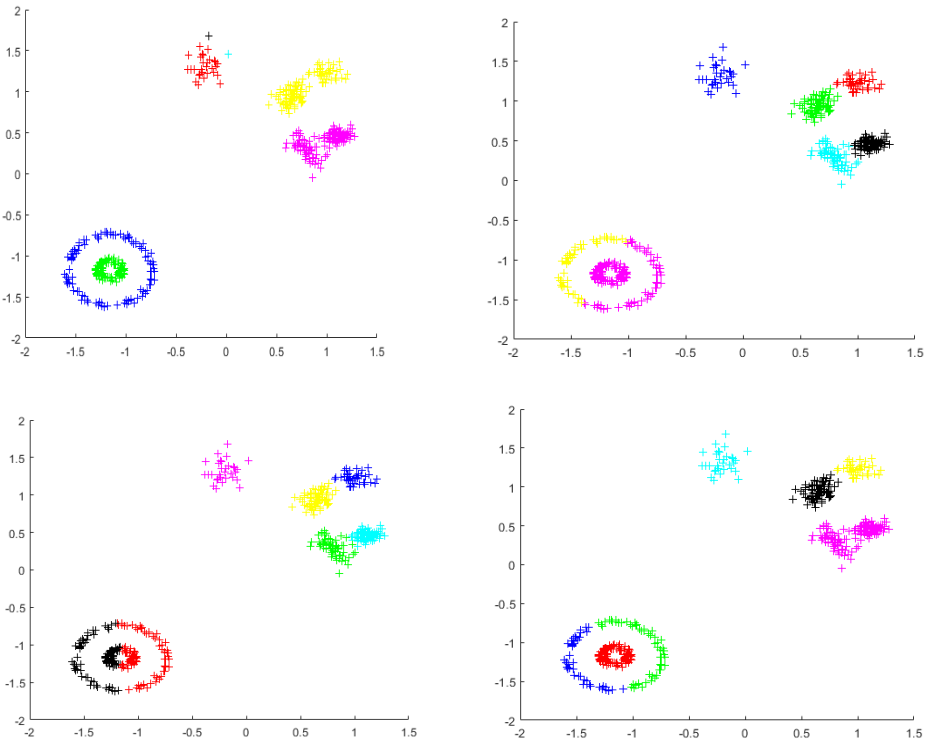
Ενδεικτικά, παρατίθενται ορισμένες από τις γραφικές απεικονίσεις που προέκυψαν κατά τις δοκιμές, με πρώτη την αποτυχία της μεθόδου και τελευταίες τις επιτυχίες:

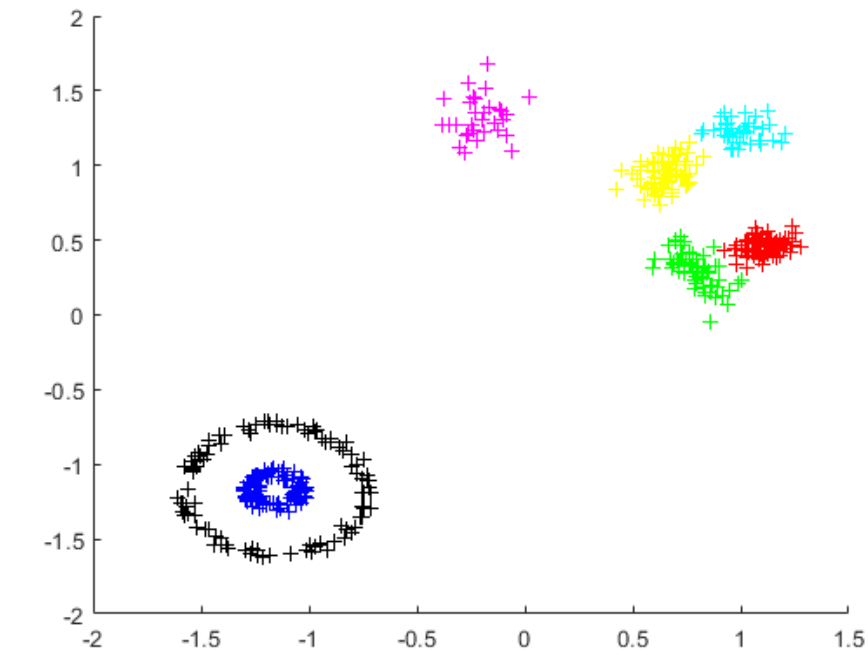

2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1.5
-1
-0.5
0
0.5
1
1.5


2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1.5
-1
-0.5
0
0.5
1
1.5


2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1.5
-1
-0.5
0
0.5
1
1.5

Τέλος, αν υποτεθεί ότι το πλήθος των ομάδων είναι επτά (7), τότε το κατάλληλο sigma βρίσκεται σχετικά πιο δύσκολα και έχει πολύ μικρότερο εύρος (της τάξης των δεκάτων). Αρκετές επαναλήψεις και δοκιμές, κατέληξαν στην τιμή 0.25 για την εύρεση λύσης σε αυτή την περίπτωση. Παρακάτω φαίνονται κάποιες από τις δοκιμές και τα αντίστοιχα αποτελέσματα αυτών, με πρώτες τις αποτυχημένες προσπάθειες και τελευταία την επιτυχημένη:





Σημειώνεται, ότι η ***evalclusters*** με κριτήριο αξιολόγησης το silhouette, εκτιμάει ότι το πραγματικό πλήθος των ομάδων, σε αυτό το σύνολο, είναι το 7.

Στη συνέχεια, για τα αρχεία δεδομένων ‘3wings’, ‘4rectangles’, ‘5clusters’, ‘5Gaussians’ και ‘7clusters’, δηλαδή των συνόλων που δεν περιέχουν δακτυλίους, εκτιμούμε μέσω της μεθόδου ‘***evalclusters***’ του ***Matlab***, το πλήθος των ομάδων κάθε συνόλου. Θεωρείται ότι αλγόριθμος ομαδοποίησης είναι ο ***k-means*** και κριτήριο αξιολόγησης το silhouette.

Με βάση τα παραπάνω έχουμε ότι οι ομάδες στα αντίστοιχα σύνολα δεδομένων είναι 3, 5, 5, 5 και 8. Σημειώνεται ότι το εύρος αναζήτησης πλήθους ομάδων ορίστηκε τυχαία, στις αντίστοιχες περιπτώσεις, το [1,7], [2,9], [2,8], [3,10] και [4,12], χωρίς βέβαια αυτό να παίζει κάποιο ιδιαίτερο ρόλο.

Παρατηρείται λοιπόν, ότι στα σύνολα ‘3wings’, ‘5clusters’ και ‘5Gaussians’, η εκτίμηση των ομάδων είναι σωστή. Αντιθέτως, στα υπόλοιπα σύνολα, η μέθοδος αποτυγχάνει στην ορθή εκτίμηση των ομάδων. Αυτό οφείλεται στο κριτήριο αξιολόγησης των ταξινομητών ***k-means*** για τις διάφορες τιμές του k στα αντίστοιχα διαστήματα.

- **Παράρτημα**:

```
%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%
%&%                                                                 %&%
%&%              Λ.07 - ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ                          %&%
%&%                 2Η Σειρά Ασκήσεων                               %&%
%&%                                                                 %&%
%&%                    Υλοποίηση:                                   %&%
%&%                                                                 %&%
%&%              Μπάτση Σοφία     Α.Μ.:372                         %&%
%&%          Δημητριάδης Σωκράτης Α.Μ.:359                         %&%
%&%                                                                 %&%
%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%&%
```

```
clear all;
clc;

% Φόρτωση του αρχείου δεδομένων/καταχωρήσεων
load ('BatsiDimitriadisData.mat');
```

Άσκηση 1: Ταξινόμηση με τη μέθοδο Bagging και με random forest:

```
% Α) Μέθοδος Bagging για Ensemble Learning:
%
% ->Μέτρηση της ικανότητας γενίκευσης:
% Κάνοντας χρήση του cross-validation
% ->Πλήθος ταξινομητών ανά μέθοδο:
% 25,50,75 και 100 αντίστοιχα
Bagging25Model =
fitcensemble(X,classes,'Method','Bag','NumLearningCycles',25,'CrossVal','on');

Bagging50Model =
fitcensemble(X,classes,'Method','Bag','NumLearningCycles',50,'CrossVal','on');

Bagging75Model =
fitcensemble(X,classes,'Method','Bag','NumLearningCycles',75,'CrossVal','on');

Bagging100Model =
fitcensemble(X,classes,'Method','Bag','NumLearningCycles',100,'CrossVal','on');

kfL25 = kfoldLoss(Bagging25Model,'Mode','cumulative');
figure;
plot(kfL25);
ylabel('10-fold Misclassification rate');
xlabel('Learning cycle');
disp(min(kfL25));

kfL50 = kfoldLoss(Bagging50Model,'Mode','cumulative');
figure;
plot(kfL50);
ylabel('10-fold Misclassification rate');
xlabel('Learning cycle');
disp(min(kfL50));
```

```
kfL75 = kfoldLoss(Bagging75Model,'Mode','cumulative');
figure;
plot(kfL75);
ylabel('10-fold Misclassification rate');
xlabel('Learning cycle');
disp(min(kfL75));

kfL100 = kfoldLoss(Bagging100Model,'Mode','cumulative');
figure;
plot(kfL100);
ylabel('10-fold Misclassification rate');
xlabel('Learning cycle');
disp(min(kfL100));

% Για οποιοδήποτε k, το αντίστοιχο k-fold
% cross-validation μπορεί να πραγματοποιηθεί
% επιλέγοντας τις παραμέτρους για κάθε μοντέλο:
% cvp = cvpartition(PartitionSize,'KFold',k)
% fitcensemble(...,'CVPartition',cvp,...)
% ---------------------------------------


%  Β) Μέθοδος Random Forest για Ensemble Learning:
%
% ->Μέτρηση της ικανότητας γενίκευσης:
% Κάνοντας χρήση του out-of-bag error
% ->Πλήθος δέντρων ταξινόμησης ανά μέθοδο:
% 25,50,75 και 100 αντίστοιχα
RF25Model =
TreeBagger(25,X,classes,'OOBPrediction','On','MinLeafSize',5,'Method','classification'
);

RF50Model =
TreeBagger(50,X,classes,'OOBPrediction','On','MinLeafSize',5,'Method','classification'
);

RF75Model =
TreeBagger(75,X,classes,'OOBPrediction','On','MinLeafSize',5,'Method','classification'
);

RF100Model =
TreeBagger(100,X,classes,'OOBPrediction','On','MinLeafSize',5,'Method','classification
');

% Η επιλογή 'OOBPrediction' σε 'on' , διατηρεί την
% πληροφορία για το ποιες παρατηρήσεις είναι
% out-of-bag κάθε δέντρο απόφασης.
% Αυτή η πληροφορία μπορεί να χρησιμοποιηθεί από
% την συνάρτηση oobPrediction για να υπολογισθούν
% οι πιθανότητες των προβλεπόμενων κλάσσεων κάθε
% δέντρου του Ensemble μοντέλου.

% Προβολή ενδεικτικών δεντρικών αναπαραστάσεων
% των αντίστοιχων μοντέλων
view(RF25Model.Trees{14},'Mode','graph')
view(RF50Model.Trees{30},'Mode','graph')
view(RF75Model.Trees{55},'Mode','graph')
view(RF100Model.Trees{80},'Mode','graph')
```

```
figure;
oobError25 = oobError(RF25Model);
plot(oobError25)
xlabel 'Πλήθος Δέντρων';
ylabel 'Out-of-bag error ταξινόμησης';
min(oobError25)

figure;
oobError50 = oobError(RF50Model);
plot(oobError50)
xlabel 'Πλήθος Δέντρων';
ylabel 'Out-of-bag error ταξινόμησης';
min(oobError50)

figure;
oobError75 = oobError(RF75Model);
plot(oobError75)
xlabel 'Πλήθος Δέντρων';
ylabel 'Out-of-bag error ταξινόμησης';
min(oobError75)

figure;
oobError100 = oobError(RF100Model);
plot(oobError100)
xlabel 'Πλήθος Δέντρων';
ylabel 'Out-of-bag error ταξινόμησης';
min(oobError100)

% Η σύγκριση με τους ταξινομητές που
% υλοποιήθηκαν στην 1η σειρά ασκήσεων
% έχει γίνει στην γραπτή αναφορά
```

Άσκηση 2: Ομαδοποίηση:

```
% Φόρτωση των αρχείων δεδομένων/καταχωρήσεων
% όλων των παραδειγμάτων
X1 = load('3rings.mat');
X1 = X1.X;

X2 = load('3wings.mat');
X2 = X2.X;

X3 = load('4rectangles.mat');
X3 = X3.X;

X4 = load('5clusters.mat');
X4 = X4.X;

X5 = load('5Gaussians.mat');
X5 = X5.X;

X6 = load('5rings.mat');
X6 = X6.X;

X7 = load('7clusters.mat');
X7 = X7.X;

GR = load('gaussian_rings.mat');
GR = GR.X;

clear X;

% Εφαρμογή των μεθόδων ομαδοποίησης στα παραπάνω
% σύνολα δεδομένων, με πλήθος ομάδων το πραγματικό:
% 1) k-means

% Με k = 3 για τα σύνολα '3rings'
% και '3wings' διαδοχικά

[K_3M_X1,C1] = kmeans(X1,3);

plot_max10_clusters(X1,K_3M_X1);

[K_3M_X2,C2] = kmeans(X2,3);

plot_max10_clusters(X2,K_3M_X2);

% Με k = 4 για το σύνολο '4rectangles'

[K_4M_X3,C3] = kmeans(X3,4);

plot_max10_clusters(X3,K_4M_X3);
```

```
% Με k = 5 για τα σύνολα '5clusters',
% '5Gaussians' και '5rings' αντίστοιχα

[K_5M_X4,C4] = kmeans(X4,5);

plot_max10_clusters(X4,K_5M_X4);

[K_5M_X5,C5] = kmeans(X5,5);

plot_max10_clusters(X5,K_5M_X5);

[K_5M_X6,C6] = kmeans(X6,5);

plot_max10_clusters(X6,K_5M_X6);

% Με k = 7 για το σύνολο '7clusters'

[K_7M_X7,C7] = kmeans(X7,7);

plot_max10_clusters(X7,K_7M_X7);
% ----------------------------

% 2) Agglomerative Clustering (single/average link)

% Επιλέγεται ως μετρική η Ευκλείδεια
% (Συνήθεις είναι οι 'chebychev','mahalanobis',
% 'hamming' κ.α.), για τα σύνολα όπως προηγουμένως:

% Σε 3 clusters τα σύνολα '3rings' και '3wings' αντίστοιχα
Z1S = linkage(X1,'single','euclidean');
Z1A = linkage(X1,'average','euclidean');

c1S = cluster(Z1S,'maxclust',3);
c1A = cluster(Z1A,'maxclust',3);

figure;
scatter(X1(:,1),X1(:,2),5,c1S)
figure;
scatter(X1(:,1),X1(:,2),5,c1A)

Z2S = linkage(X2,'single','euclidean');
Z2A = linkage(X2,'average','euclidean');
c2S = cluster(Z2S,'maxclust',3);
c2A = cluster(Z2A,'maxclust',3);

figure;
scatter(X2(:,1),X2(:,2),5,c2S)
figure;
scatter(X2(:,1),X2(:,2),5,c2A)
```

```
% Σε 4 clusters το σύνολο '4rectangles'
Z3S = linkage(X3,'single','euclidean');
Z3A = linkage(X3,'average','euclidean');
c3S = cluster(Z3S,'maxclust',4);
c3A = cluster(Z3A,'maxclust',4);

figure;
scatter(X3(:,1),X3(:,2),5,c3S)
figure;
scatter(X3(:,1),X3(:,2),5,c3A)

% Σε 5 clusters τα σύνολα '5clusters',
% '5Gaussians' και '5rings' αντίστοιχα
Z4S = linkage(X4,'single','euclidean');
Z4A = linkage(X4,'average','euclidean');
c4S = cluster(Z4S,'maxclust',5);
c4A = cluster(Z4A,'maxclust',5);

figure;
scatter(X4(:,1),X4(:,2),5,c4S)
figure;
scatter(X4(:,1),X4(:,2),5,c4A)

Z5S = linkage(X5,'single','euclidean');
Z5A = linkage(X5,'average','euclidean');
c5S = cluster(Z5S,'maxclust',5);
c5A = cluster(Z5A,'maxclust',5);

figure;
scatter(X5(:,1),X5(:,2),5,c5S)
figure;
scatter(X5(:,1),X5(:,2),5,c5A)

Z6S = linkage(X6,'single','euclidean');
Z6A = linkage(X6,'average','euclidean');
c6S = cluster(Z6S,'Maxclust',5);
c6A = cluster(Z6A,'Maxclust',5);

figure;
scatter(X6(:,1),X6(:,2),5,c6S)
figure;
scatter(X6(:,1),X6(:,2),5,c6A)

% Σε 7 clusters το σύνολο '7clusters'
Z7S = linkage(X7,'single','euclidean');
Z7A = linkage(X7,'average','euclidean');
c7S = cluster(Z7S,'Maxclust',7);
c7A = cluster(Z7A,'Maxclust',7);

figure;
scatter(X7(:,1),X7(:,2),5,c7S)
figure;
scatter(X7(:,1),X7(:,2),5,c7A)
% ---------------------------
```

```
% 3) Spectral Clustering (sigma=0.1, 0.5 και 1)

% Σε 3 clusters τα σύνολα '3rings' και '3wings' αντίστοιχα
SpecX11 = spectral(X1,3,0.1);
SpecX12 = spectral(X1,3,0.5);
SpecX13 = spectral(X1,3,1);

SpecX21 = spectral(X2,3,0.1);
SpecX22 = spectral(X2,3,0.5);
SpecX23 = spectral(X2,3,1);

% Σε 4 clusters το σύνολο '4rectangles'
SpecX31 = spectral(X3,4,0.1);
SpecX32 = spectral(X3,4,0.5);
SpecX33 = spectral(X3,4,1);

% Σε 5 clusters τα σύνολα '5clusters',
% '5Gaussians' και '5rings' αντίστοιχα
SpecX41 = spectral(X4,5,0.1);
SpecX42 = spectral(X4,5,0.5);
SpecX43 = spectral(X4,5,1);

SpecX51 = spectral(X5,5,0.1);
SpecX52 = spectral(X5,5,0.5);
SpecX53 = spectral(X5,5,1);

SpecX61 = spectral(X6,5,0.1);
SpecX62 = spectral(X6,5,0.5);
SpecX63 = spectral(X6,5,1);

% Σε 7 clusters το σύνολο '7clusters'
SpecX71 = spectral(X7,7,0.1);
SpecX72 = spectral(X7,7,0.5);
SpecX73 = spectral(X7,7,1);
% -------------------------

% Για το σύνολο 'gaussian_rings', αναζητούμε
% το κατάλληλο sigma εντός του διαστήματος
% [0.1, 0.4] για το οποίο ο αλγόριθμος spectral
% δίνει τη σωστή λύση ομαδοποίησης. Αυτό, το
% επιτυγχάνουμε ως εξής:
for i=0.1:0.05:0.4

    SpecGR = spectral(GR,5,i);
    % Αλλάζοντας το 5, με 6 ή 7, μπορείτε να δείτε
    % το πως αλλάζει το κατάλληλο sigma της μεθόδου
    % Επιπλέον, μπορείτε να μικρύνετε το βήμα
    % της επανάληψης στο 0.01 και να λάβετε
    % με μια μικρή καθυστέρηση, μεγαλύτερη ακρίβεια
end
% -------------------------------------------------
```

```
% Για τα αρχεία δεδομένων που δεν περιέχουν δακτυλίους,
% εκτιμούμε μέσω της evalclusters το πλήθος των ομάδων
% κάθε συνόλου. Αλγόριθμος ομαδοποίησης είναι ο k-means
% και κριτήριο αξιολόγησης το silhouette.
evaX2 = evalclusters(X2,'kmeans','silhouette','KList',[1:7]);
evaX3 = evalclusters(X3,'kmeans','silhouette','KList',[2:9]);
evaX4 = evalclusters(X4,'kmeans','silhouette','KList',[2:8]);
evaX5 = evalclusters(X5,'kmeans','silhouette','KList',[3:10]);
evaX7 = evalclusters(X7,'kmeans','silhouette','KList',[4:12]);

disp(evaX2.OptimalK);
disp(evaX3.OptimalK);
disp(evaX4.OptimalK);
disp(evaX5.OptimalK);
disp(evaX7.OptimalK);
```

Published with MATLAB® R2018a