



Πολυτεχνείο
Κρήτης

Σχολή Μηχανικών Παραγωγής και Διοίκησης

*Εισαγωγή στην Τεχνητή
Νοημοσύνη*

Μάθηση χωρίς επίβλεψη: Αλγόριθμοι
Ομαδοποίησης/Συσταδοποίησης (Clustering)

Διαχωριστικοί αλγόριθμοι (K-means)

Τσαφαράκης Στέλιος

Μηχανική μάθηση

- Ο άνθρωπος προσπαθεί να κατανοήσει το περιβάλλον του παρατηρώντας το και δημιουργώντας μια απλοποιημένη (αφαιρετική) εκδοχή του που ονομάζεται *μοντέλο*.
- Η δημιουργία ενός τέτοιου μοντέλου, ονομάζεται *επαγωγική μάθηση* (inductive learning) ενώ η διαδικασία γενικότερα ονομάζεται *επαγωγή* (induction).
- Επιπλέον ο άνθρωπος έχει τη δυνατότητα να οργανώνει & να συσχετίζει τις εμπειρίες & τις παραστάσεις του δημιουργώντας νέες δομές που ονομάζονται *πρότυπα* (patterns).
- Η δημιουργία μοντέλων ή προτύπων από ένα υπολογιστικό σύστημα με χρήση ενός συνόλου δεδομένων ονομάζεται *μηχανική μάθηση* (machine learning).

Είδη μηχανικής μάθησης

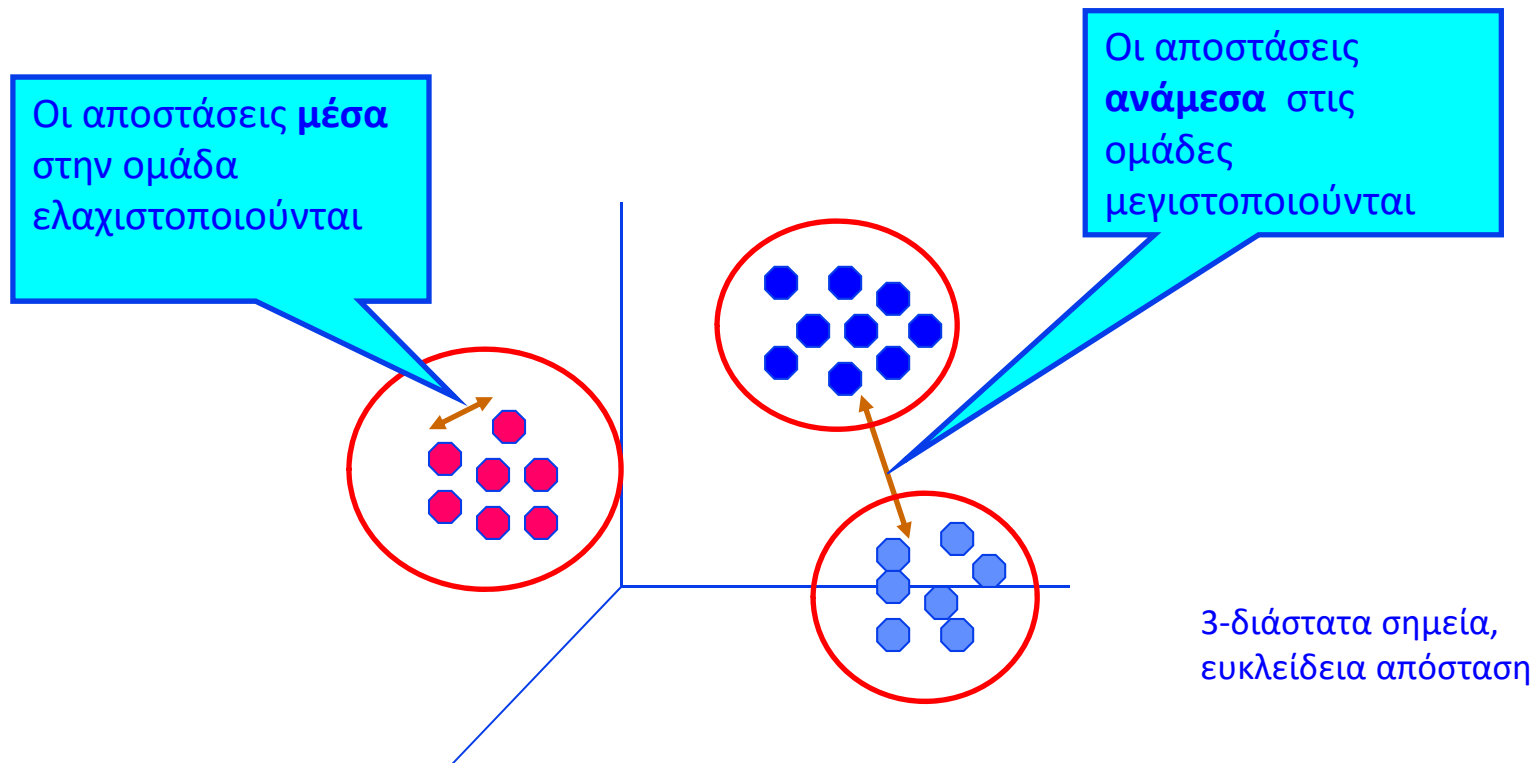
- Έχουν αναπτυχθεί πολλές τεχνικές μηχανικής μάθησης που χρησιμοποιούνται ανάλογα με τη φύση του προβλήματος & εμπίπτουν σε δυο είδη.
- Μάθηση με επίβλεψη (supervised learning) ή μάθηση με παραδείγματα (learning from examples):
 - το σύστημα καλείται να "μάθει" μια έννοια ή συνάρτηση η οποία αποτελεί περιγραφή ενός μοντέλου, μέσα από ένα σύνολο δεδομένων.
- Μάθηση χωρίς επίβλεψη (unsupervised learning) ή μάθηση από παρατήρηση (learning from observation):
 - το σύστημα πρέπει μόνο του να ανακαλύψει συσχετίσεις ή ομάδες σε ένα σύνολο δεδομένων, δημιουργώντας πρότυπα, χωρίς να είναι γνωστό αν υπάρχουν, πόσα και ποια είναι.

Μάθηση χωρίς επίβλεψη

- Στόχος η ανακάλυψη συσχετίσεων & ομάδων από τα δεδομένα με βάση μόνο τις ιδιότητές τους.
- Ως αποτέλεσμα προκύπτουν πρότυπα (περιγραφές).
- Κάθε πρότυπο περιγράφει ένα μέρος από τα δεδομένα.
- Παραδείγματα προτύπων πληροφόρησης είναι
 - οι κανόνες συσχέτισης (association rules) και
 - οι ομάδες (clusters), οι οποίες προκύπτουν από τη διαδικασία της ομαδοποίησης (clustering).
- Ομαδοποίηση: διαχωρισμός ενός συνόλου (συνήθως πολυδιάστατων) δεδομένων σε ομάδες, ώστε:
 - σημεία που ανήκουν στην ίδια ομάδα να μοιάζουν όσο το δυνατόν περισσότερο και
 - σημεία που ανήκουν σε διαφορετικές ομάδες να διαφέρουν όσο το δυνατόν περισσότερο.

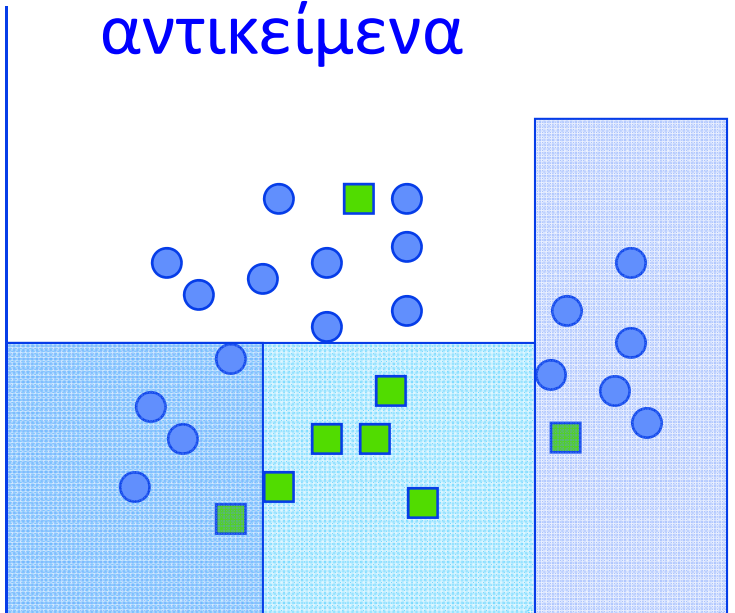
Ομαδοποίηση (Clustering)

Η εύρεση ομάδων αντικειμένων έτσι ώστε τα αντικείμενα σε κάθε ομάδα να είναι όμοια (ή να σχετίζονται) και διαφορετικά (ή μη σχετιζόμενα) από τα αντικείμενα των άλλων ομάδων



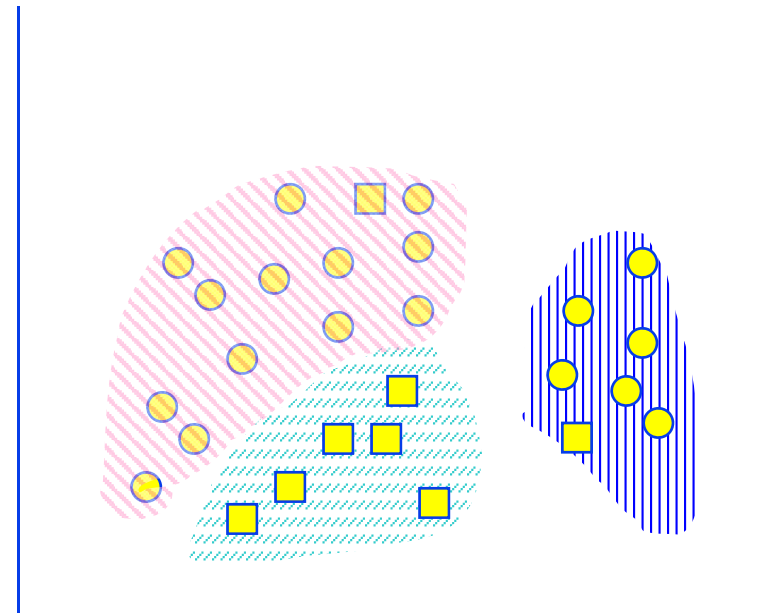
Ταξινόμηση (Classification)

- Μάθηση με επίβλεψη
- Μαθαίνει να προβλέπει την κλάση ενός αντικειμένου από ένα σύνολο από ταξινομημένα αντικείμενα



Ομαδοποίηση (Clustering)

- Μάθηση χωρίς επίβλεψη
- Ερευνά τη «φυσική» ομαδοποίηση των αντικειμένων με βάση μη ταξινομημένα δεδομένα



Παράδειγμα: Τμηματοποίηση αγοράς

- Δημιουργία ομάδων καταναλωτών με κοινά χαρακτηριστικά.
- Οι καταναλωτές που ανήκουν σε μια ομάδα (συστάδα) έχουν σε μεγάλο βαθμό παρόμοια χαρακτηριστικά.
- Τα χαρακτηριστικά μπορεί να είναι δημογραφικά ή παρόμοιες προτιμήσεις σχετικά με ένα προϊόν και τα χαρακτηριστικά του:
 - π.χ. στην Ελληνική αγορά υπάρχουν 2 ομάδες καταναλωτών σχετικά με τις προτιμήσεις τους για laptop.
 - Η 1^η ομάδα προτιμάει ελαφριά laptop με μικρή οθόνη, ισχυρό επεξεργαστή και μεγάλη μνήμη.
 - Η 2^η ομάδα προτιμάει laptop με μεγάλη οθόνη, ισχυρή κάρτα γραφικών και μεγάλο σκληρό δίσκο.

Δυσκολία ομαδοποίησης



Πόσες ομάδες χαρτιών υπάρχουν?

Δυσκολία καθορισμού αριθμού ομάδων



Πόσες Ομάδες?



2 ομάδες



4 ομάδες



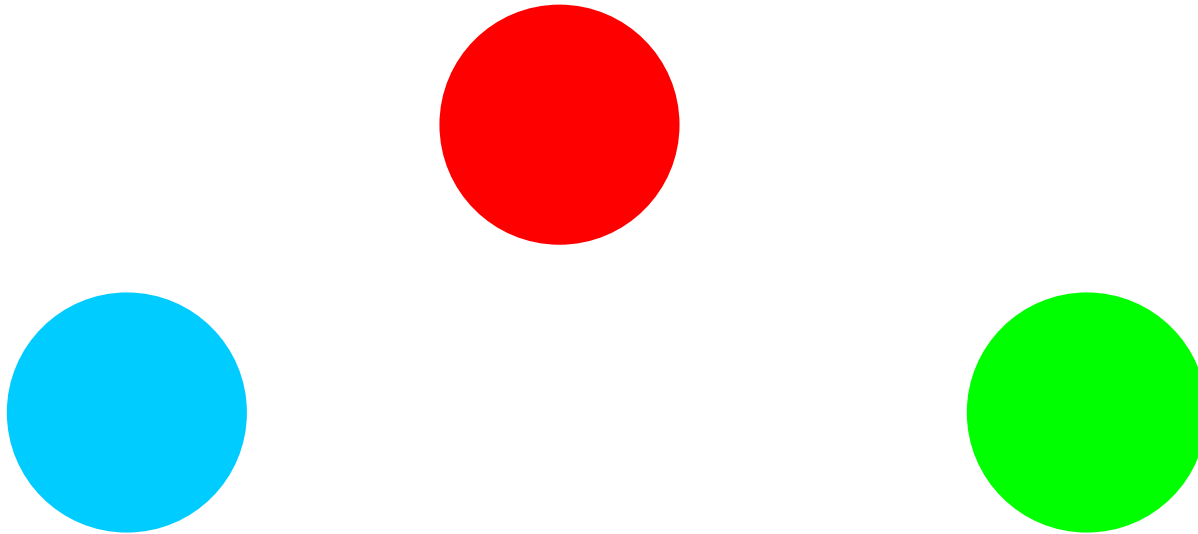
6 ομάδες

Είδη ομαδοποίησης

- Ομαδοποίηση Διαμέρισης - Διαχωριστική (Partitional Clustering)
- Ιεραρχική Ομαδοποίηση (Hierarchical Clustering)
- Ασαφής Ομαδοποίηση (Fuzzy Clustering)
- Ομαδοποίηση βασισμένη στα δίκτυα Kohonen (Kohonen Net Clustering)
- Ομαδοποίηση βασισμένη στην πυκνότητα (Density-based Clustering)
- Ομαδοποίηση βασισμένη σε πλέγμα (Grid-based Clustering)
- Ομαδοποίηση υποχώρων (Subspace Clustering).

Καλώς διαχωρισμένες ομάδες

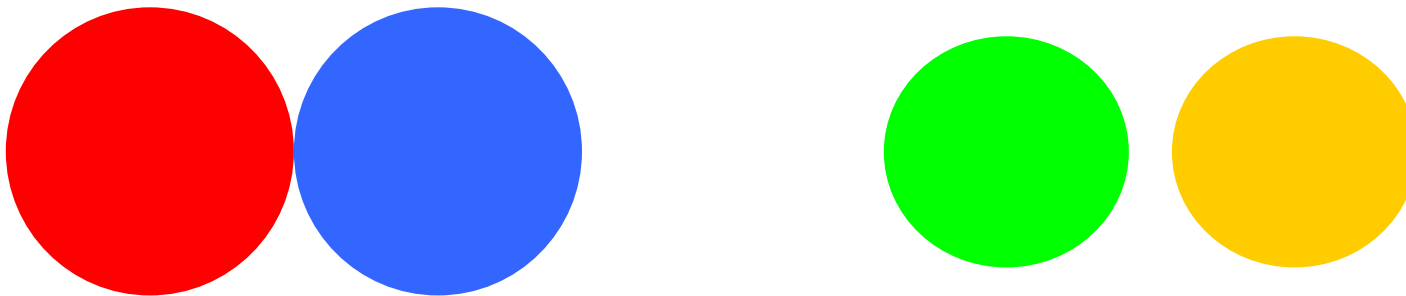
- Μια ομάδα είναι ένα σύνολο σημείων ώστε κάθε σημείο να είναι κοντινότερο σε (ή πιο όμοιο με) όλα τα άλλα σημεία της ομάδας από ότι σε οποιοδήποτε άλλο σημείο που δεν ανήκει στην ομάδα.



3 καλώς διαχωρισμένες ομάδες

Ομάδες βασισμένες σε κέντρο

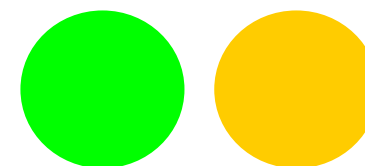
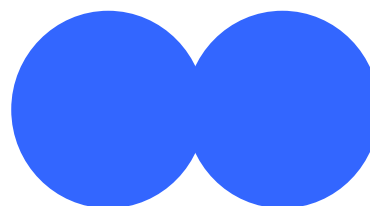
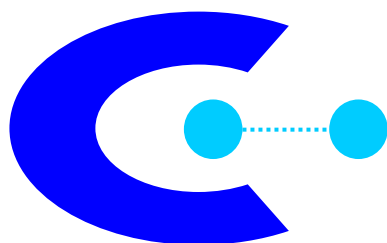
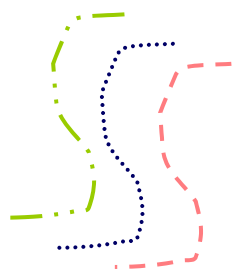
- Μια ομάδα είναι ένα σύνολο σημείων ώστε κάθε σημείο να είναι κοντινότερο στο (ή πιο όμοιο με το) κέντρο της ομάδας σε σχέση με το κέντρο οποιασδήποτε άλλης ομάδας.
- Το κέντρο της ομάδας είναι:
- Ο μέσος όρος των σημείων της ομάδας (centroid)
- Το πιο αντιπροσωπευτικό σημείο της ομάδας (medoid)



4 ομάδες βασισμένες σε κέντρο

Συνεχείς Ομάδες

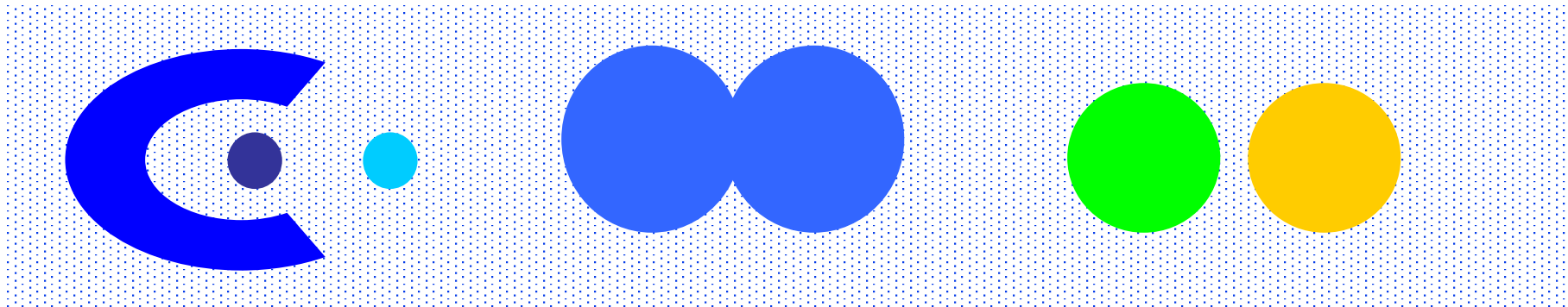
- Μια ομάδα είναι ένα σύνολο σημείων ώστε κάθε σημείο να είναι πιο κοντά σε ένα ή περισσότερα σημεία της ομάδας από ότι σε οποιοδήποτε άλλο σημείο εκτός ομάδας.
- Χρησιμοποιούνται σε περιπτώσεις ομάδων με μη κανονικά σχήματα.
- Προβλήματα με θόρυβο.



8 συνεχείς ομάδες

Ομάδες βασισμένες στην πυκνότητα

- Μια ομάδα είναι μια πυκνή περιοχή σημείων η οποία διαχωρίζεται από άλλες πυκνές περιοχές σημείων.
- Χρησιμοποιούνται σε περιπτώσεις ομάδων με μη κανονικά σχήματα ή αλληλοπλεκόμενες.
- Καλύτερη αντιμετώπιση θορύβου.



6 ομάδες βασισμένες στην πυκνότητα

- **Διαχωριστική Ομαδοποίηση (Partitional)**

- Ένας διαμερισμός των αντικειμένων σε μη επικαλυπτόμενα υποσύνολα (ομάδες) τέτοιος ώστε κάθε αντικείμενο ανήκει σε ακριβώς ένα υποσύνολο.
- Προσπαθεί να βρει τον καλύτερο διαχωρισμό ενός συνόλου δεδομένων σε ένα συγκεκριμένο αριθμό ομάδων.
- Ξεκινάει από ένα διαχωρισμό και τον βελτιώνει επαναληπτικά

- **Ιεραρχική Ομαδοποίηση (Hierarchical)**

- Προσπαθεί με ιεραρχικό τρόπο να ανακαλύψει τον αριθμό και τη δομή των ομάδων.
- Κάθε ομάδα έχει υποομάδες οργανωμένες σε ένα ιεραρχικό δέντρο (εμφωλευμένες)

Απόσταση μεταξύ σημείων

- Η ομαδοποίηση απαιτεί κάποιο μέτρο της ομοιότητας ή διαφοράς μεταξύ των δεδομένων (απόσταση).
- Έστω ένα σύνολο δεδομένων D και 2 δεδομένα του x, y που περιγράφονται από m χαρακτηριστικά $(x_1, x_2, \dots, x_m), (y_1, y_2, \dots, y_m)$.
- Τυπικά μέτρα απόστασης μεταξύ των δύο δεδομένων είναι η απόσταση Μανχάταν και η Ευκλείδεια απόσταση:

$$d(x, y) = \sum_i |x_i - y_i|$$

Απόσταση Μανχάταν

$$d(x, y) = \sum_i \sqrt{(x_i - y_i)^2}$$

Ευκλείδεια απόσταση

- Αν κάποια χαρακτηριστικά είναι διακριτά, τότε η απόσταση τους θεωρείται 0 αν έχουν την ίδια τιμή και 1 αν έχουν διαφορετικές τιμές (απόσταση Hamming).

Ο αλγόριθμος των K-μέσων (K-means)

- Διαχωριστικός αλγόριθμος
 - Κάθε ομάδα συσχετίζεται με ένα κεντρικό σημείο (κέντρο - centroid)
 - Κάθε σημείο ανατίθεται στην ομάδα με το κοντινότερο κέντρο
 - Ο αριθμός των ομάδων, K , είναι είσοδος στον αλγόριθμο
-

Επέλεξε K σημεία ως τα αρχικά κέντρα

Επανάλαβε

1. Ανάθεσε όλα τα αρχικά σημεία στο κοντινότερο τους από τα K κέντρα
2. Επανα-υπολόγισε το κέντρο κάθε ομάδας

Έως να μην υπάρχει αλλαγή στο διαχωρισμό των σημείων στις ομάδες

Ψευδοκώδικας K-means

Είσοδος: Σύνολο σημείων $D=\{x_1, x_2, \dots, x_n\}$, Αριθμός ομάδων K

Έξοδος: ομάδες C_i ,

1. Δημιούργησε (συνήθως τυχαία) τα αρχικά κέντρα m_i ,
 $i=1,2,\dots,K$
2. Επανάλαβε
 3. Για κάθε σημείο $x_j \in D$ ($j=1,2,\dots,n$)
 4. Υπολόγισε την απόσταση του x_j από όλα τα κέντρα m_i
($i=1,2,\dots,K$)
 5. Ανέθεσε το x_j στην ομάδα C_i με το κοντινότερο κέντρο m_i
 6. Υπολόγισε το νέο κέντρο της κάθε ομάδας C_i ως το μέσο διάνυσμα των σημείων που ανήκουν στην ομάδα
7. Πήγαινε στο βήμα 3
8. Μέχρι κανένα σημείο να μην αλλάζει ομάδα

Παράδειγμα

- ✧ Να εφαρμοστεί ο K-means για 4 ομάδες στα παρακάτω μονοδιάστατα δεδομένα: 1, 2, 3, 6, 8, 12, 27, 31, 35, 38, 39, 45, 49, 85, 87, 91, 99, 102, 112, 158, 163, 171, 188, 195, 197, 201
- ✧ Δίνονται αρχικά κέντρα: 20, 49, 90, 130

1^η επανάληψη

Ομάδες								
1 ^η (20)	1	2	3	6	8	12	27	31
2 ^η (49)	35	38	39	45	49			
3 ^η (90)	85	87	91	99	102			
4 ^η (130)	112	158	163	171	188	195	197	201

Νέα Κέντρα
11,25
41,2
92,8
173,12

2^η επανάληψη

1 ^η (11,25)	1	2	3	6	8	12	
2 ^η (41,2)	27	31	35	38	39	45	49
3 ^η (92,8)	85	87	91	99	102	112	
4 ^η (173,12)	158	163	171	188	195	197	201

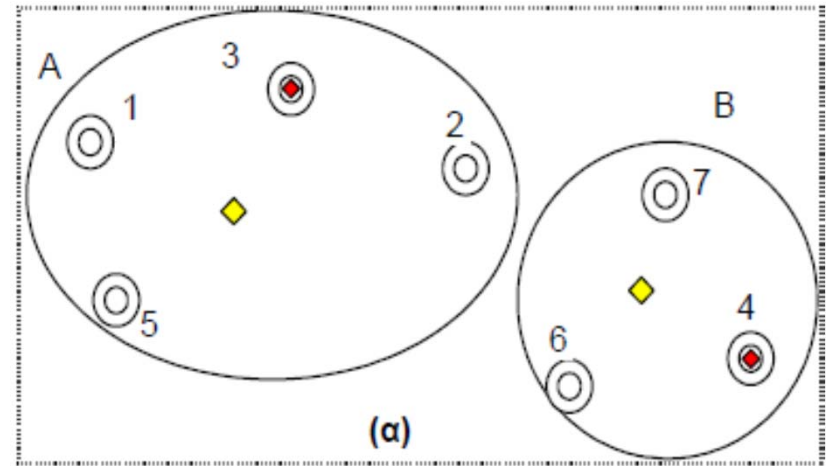
Νέα Κέντρα (τελικά)
5,33
37,71
96
181,85

K-means: Παρατηρήσεις

- Τα αρχικά κεντρικά σημεία συνήθως επιλέγονται τυχαία
- Οι ομάδες που παράγονται μπορεί να διαφέρουν από το ένα τρέξιμο του αλγορίθμου στο άλλο
- Η ομοιότητα των σημείων υπολογίζεται με βάση κάποια απόσταση που εξαρτάται από το είδος των σημείων (π.χ. Ευκλείδεια απόσταση)
- Επειδή η απόσταση υπολογίζεται συχνά πρέπει να είναι σχετικά απλή
- Το κέντρο μιας ομάδας είναι (συνήθως) το μέσο (mean) των σημείων της (το οποίο μπορεί να μην είναι ένα από τα δεδομένα εισόδου)

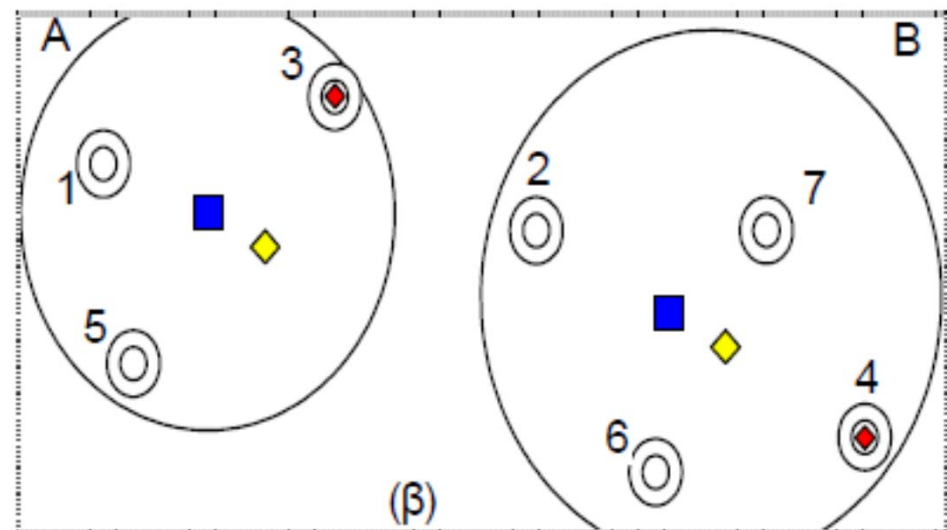
Παράδειγμα με 2-διάστατα σημεία & $K=2$

- Έστω ότι ο αλγόριθμος εκτελείται με $k=2$ για τα 7 σημεία του Σχήματος (α):
- Αρχικά επιλέγονται τυχαία, έστω τα σημεία 3 & 4, ως κέντρα των δύο ομάδων A & B
- Κάθε σημείο από τα υπόλοιπα ανατίθεται στην ομάδα της οποίας το κέντρο είναι πιο κοντά
- Άρα τα σημεία 1, 2, 3 και 5 θα ανήκουν στην ομάδα A ενώ τα 4, 6 και 7 στην B.
- Ξαναυπολογίζονται τα κέντρα κάθε ομάδας & με αυτόν τον τρόπο ολοκληρώνεται ένας κύκλος υπολογισμών.
- Τα νέα κέντρα απεικονίζονται με κίτρινο ρόμβο.



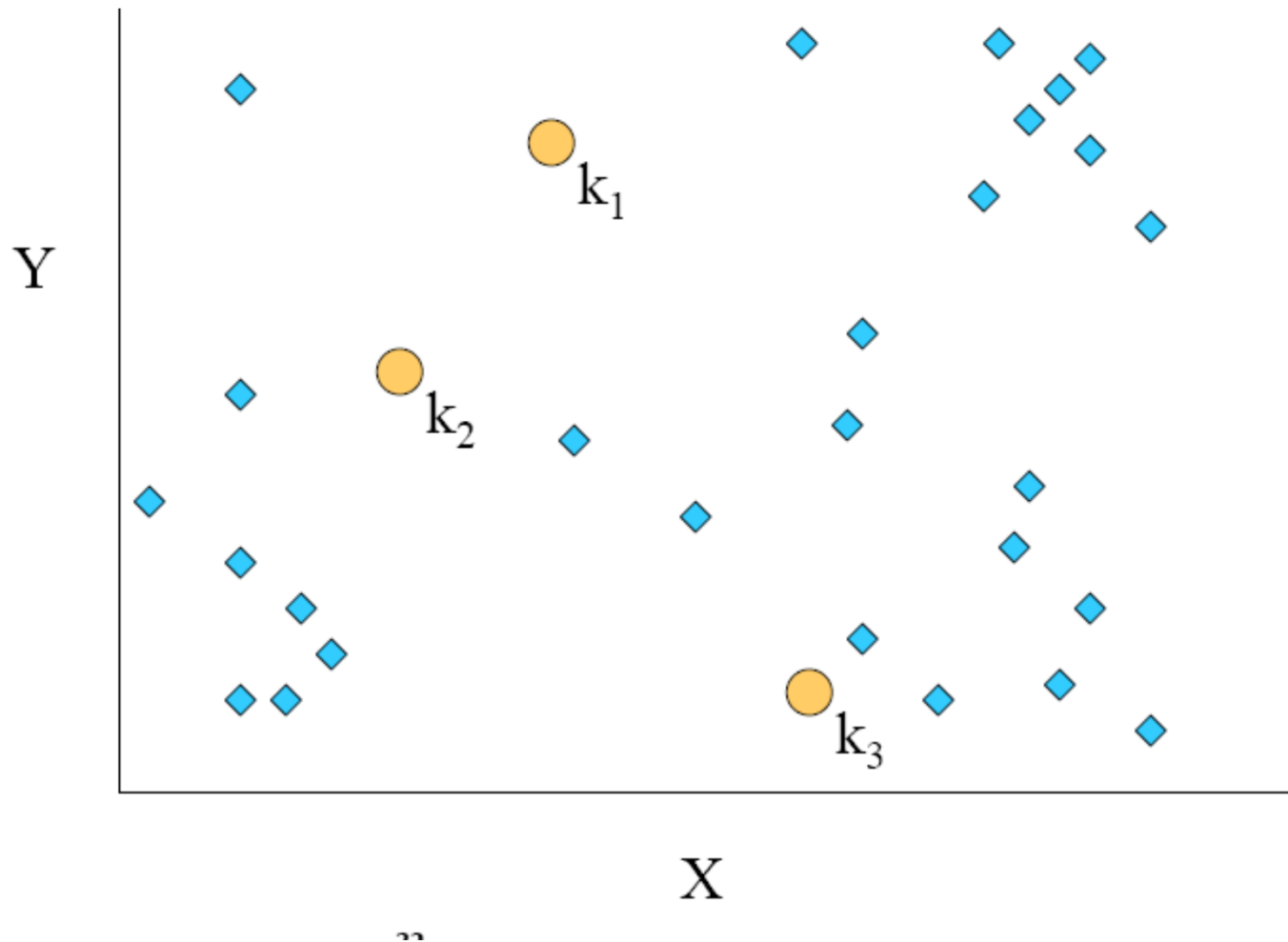
Παράδειγμα με 2-διάστατα σημεία & $K=2$

- Ο παραπάνω κύκλος υπολογισμών επαναλαμβάνεται χρησιμοποιώντας αυτήν τη φορά τις αποστάσεις των σημείων από τα νέα κέντρα.
- Μια μεταβολή που συντελείται είναι ότι το σημείο 2 αλλάζει ομάδα και ανήκει πλέον στη Β.
- Επιπλέον, τα κέντρα των ομάδων μετατοπίζονται σε νέες θέσεις που σημειώνονται με μπλε τετράγωνο.
- Κατά την τρίτη επανάληψη των παραπάνω υπολογισμών, δε συντελείται καμία μεταβολή

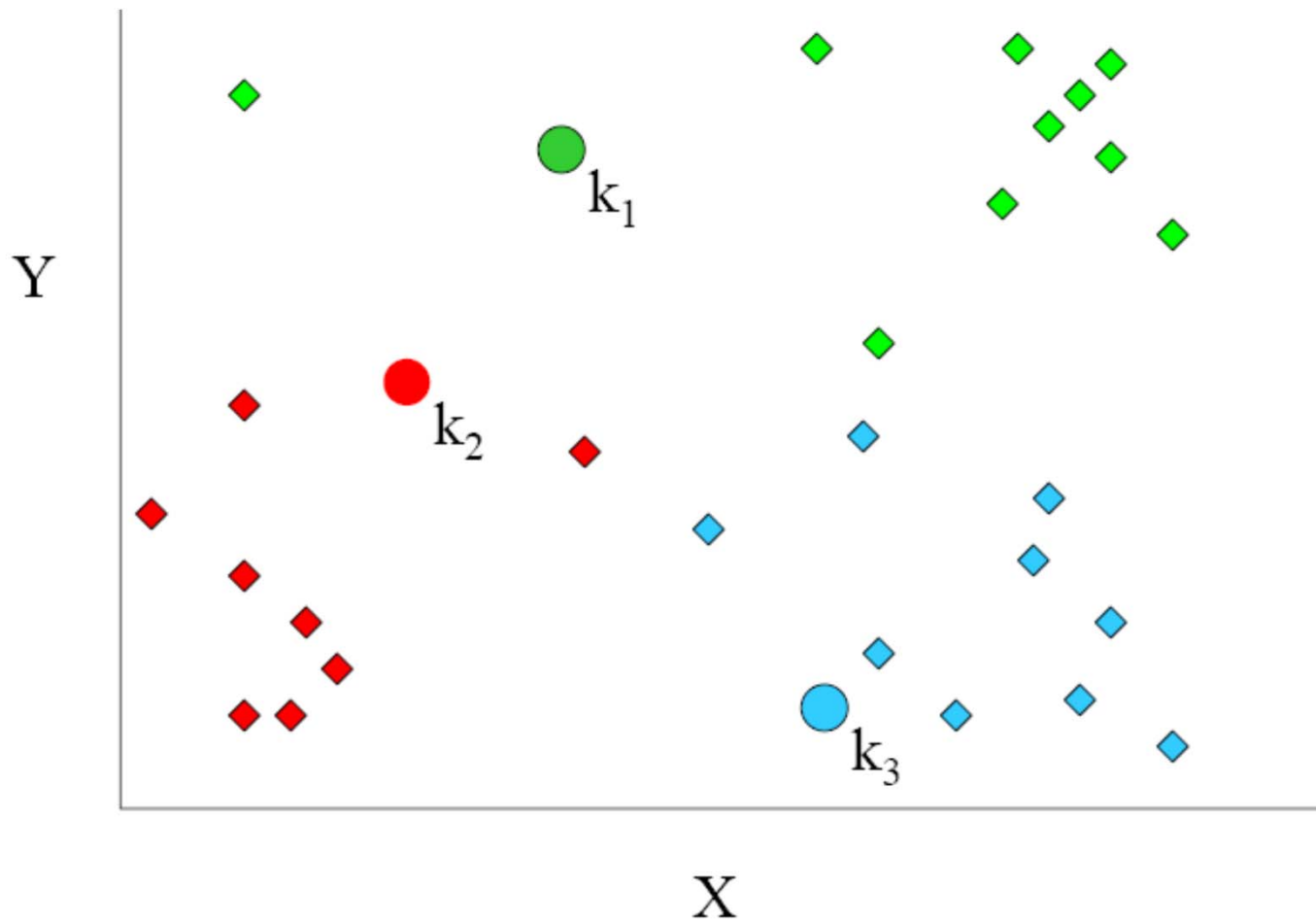


Παράδειγμα με $K=3$

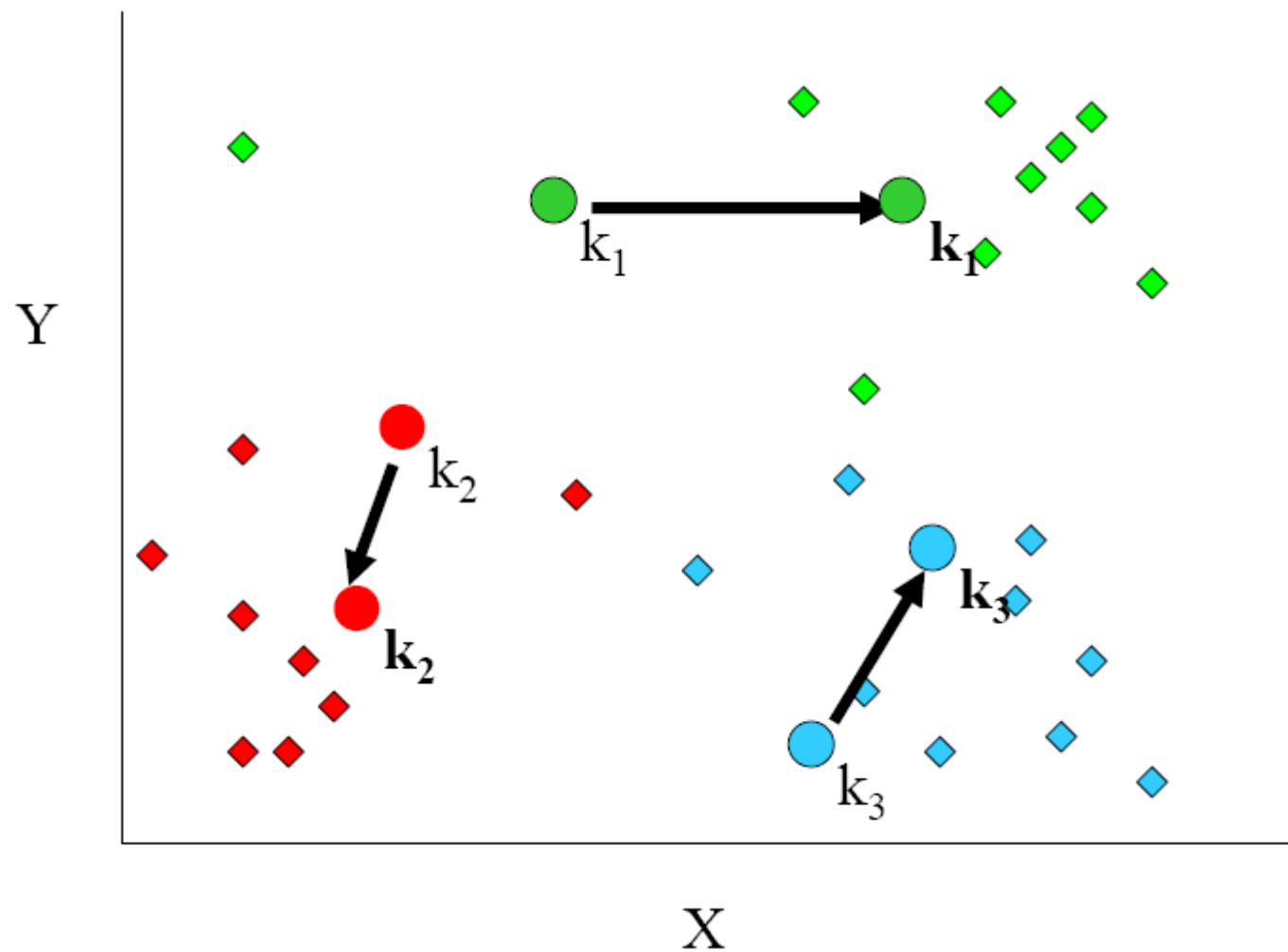
- Επιλέγουμε τυχαία αρχικά κέντρα K_1 K_2 K_3 σημεία τα οποία δεν ανήκουν στα δεδομένα.



Ανάθεση σημείων στα κοντινότερα κέντρα

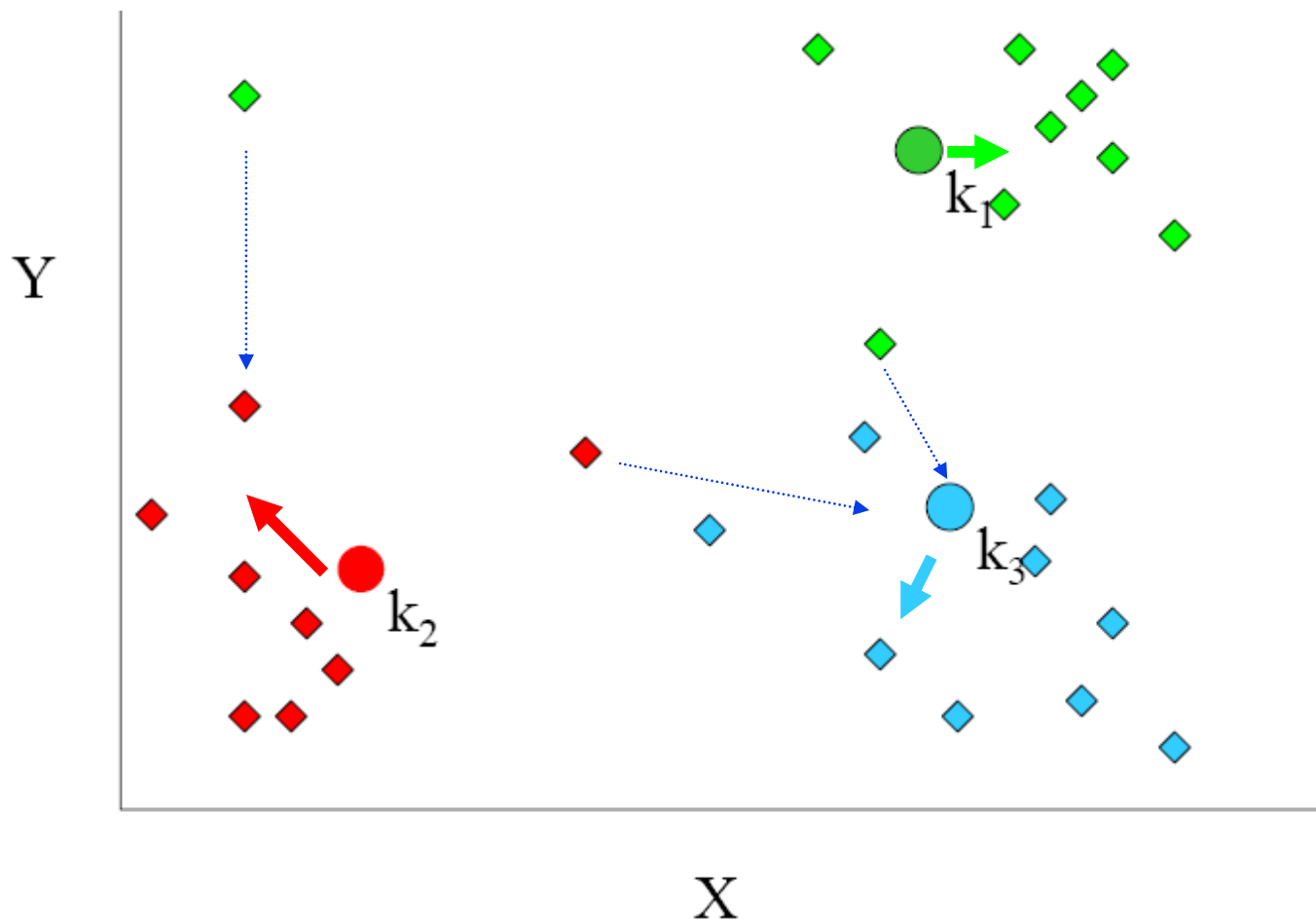


Επαν-υπολογισμός κέντρων

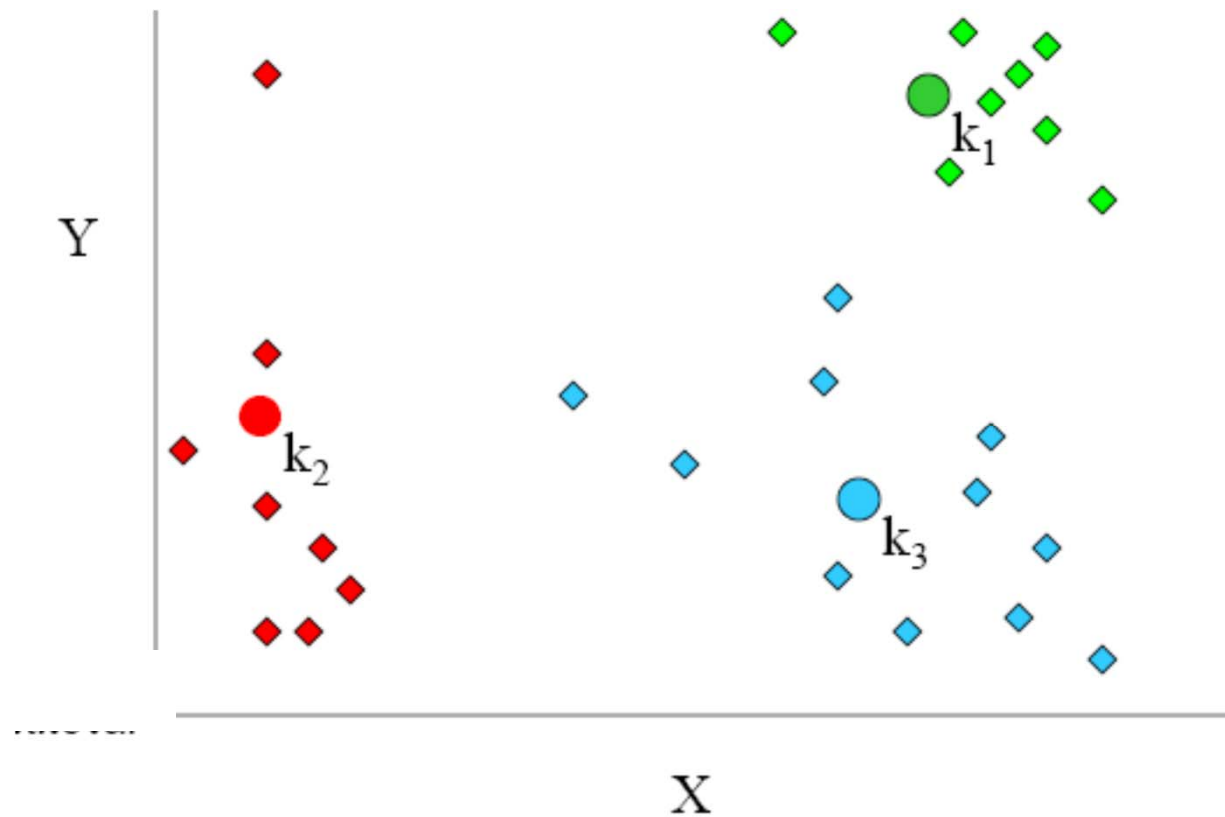


Νέα ανάθεση των σημείων

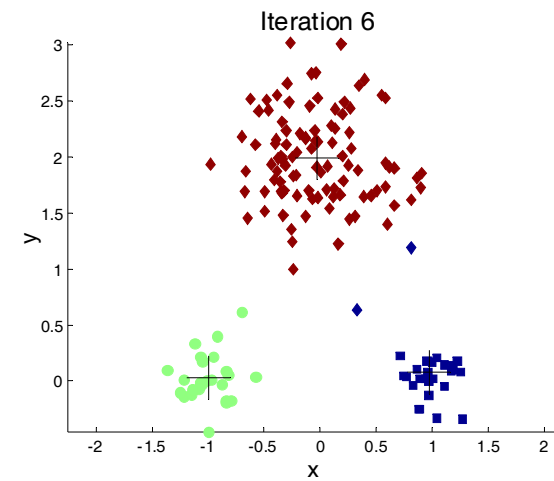
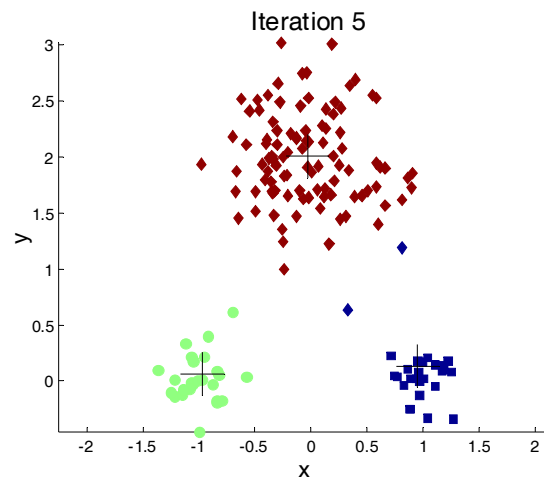
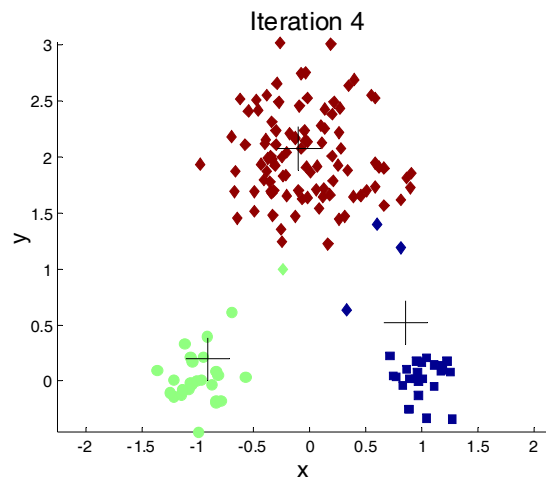
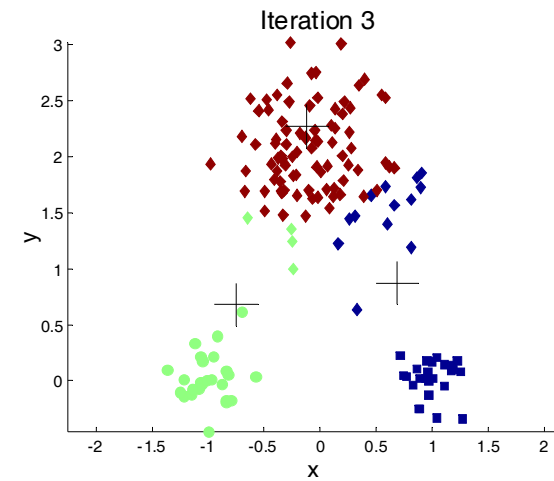
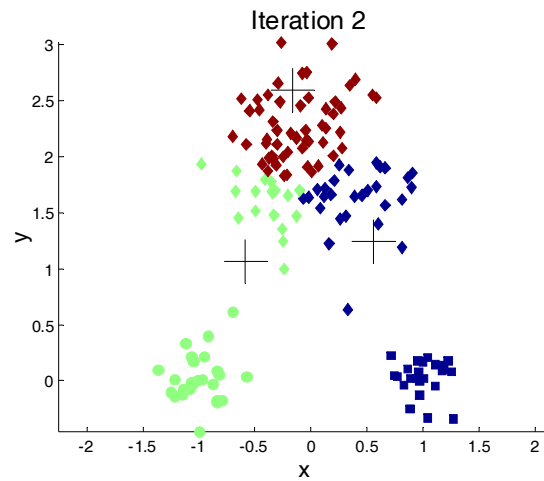
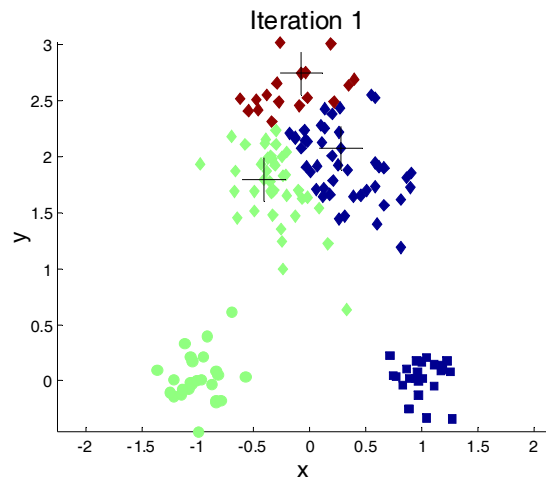
Νέα κέντρα



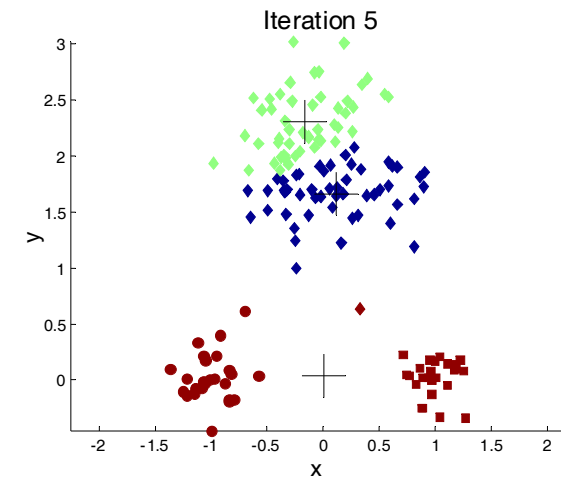
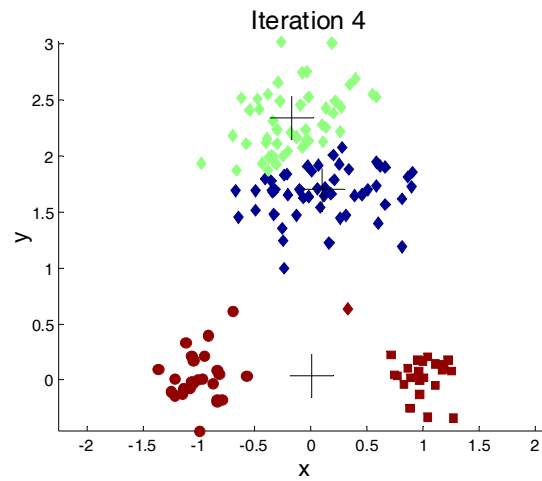
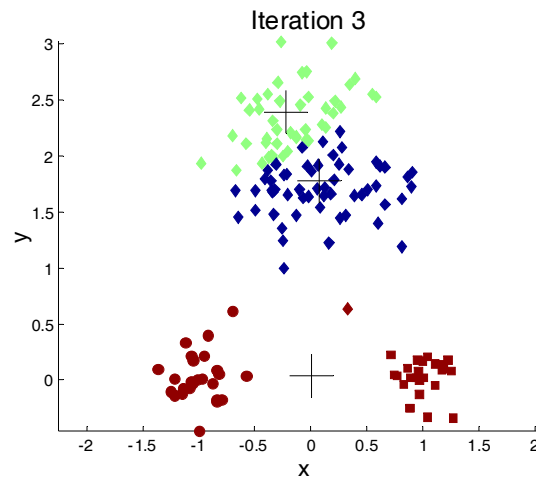
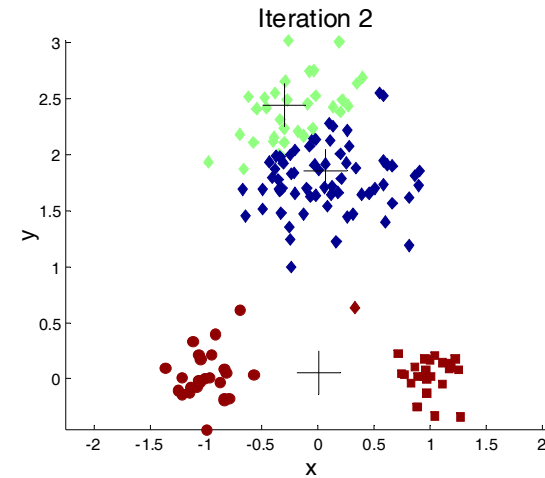
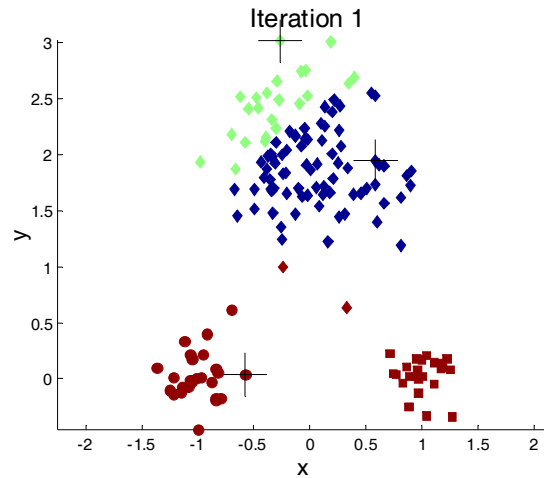
Σύγκλιση αλγορίθμου - τερματισμός



K-means: Επιλογή αρχικών σημείων



K-means: Επιλογή αρχικών σημείων



K-means: Λύσεις για την επιλογή αρχικών σημείων

- Πολλαπλά τρεξίματα
 - Βοηθά, αλλά πολλές περιπτώσεις
- Χρήση αρχικά κάποιας ιεραρχικής τεχνικής
- Επιλογή παραπάνω από k αρχικών σημείων και μετά επιλογή k από αυτά τα αρχικά κεντρικά σημεία (π.χ. τα πιο απομακρυσμένα μεταξύ τους)
- Σταδιακή επιλογή
 - Επιλογή του πρώτου σημείου τυχαία ή ως το μέσο όλων των σημείων
 - Για καθένα από τα υπόλοιπα αρχικά σημεία επέλεξε αυτό που είναι πιο μακριά από τα μέχρι τώρα επιλεγμένα αρχικά σημεία