Εξόρυξη δεδομένων με χρήση τεχνικών μηχανικής μάθησης

Τμ. Μηχανικών Πληροφορικής και Υπολογιστών

Περιγραφή Αλγορίθμου Clustering

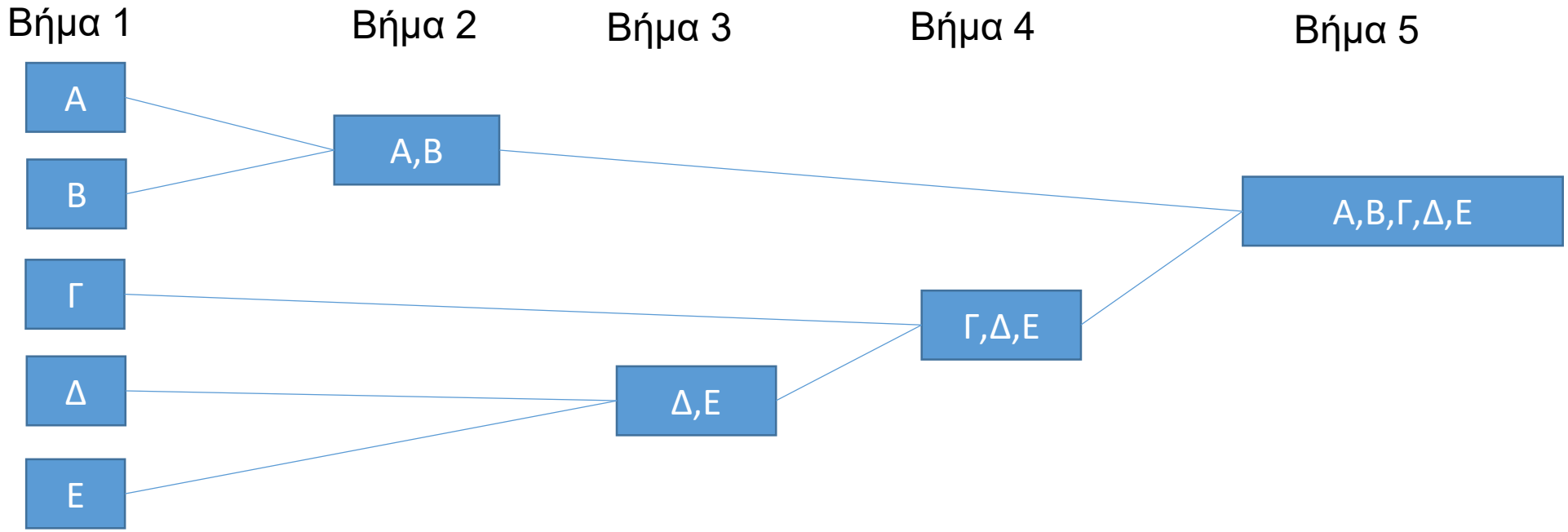
Αλγόριθμοι Ανάλυσης σε Συστάδες

Οι κατηγορίες αλγορίθμων που θα εξετάσουμε είναι

- Ιεραρχικής Ανάλυσης Συστάδων
- Διαχωριστικής Ανάλυσης Συστάδων

Ιεραρχικής Ανάλυσης Συστάδων

- Ας υποθέσουμε ότι έχουμε 5 αντικείμενα
- Στην αρχή έχουμε τόσα clusters όσα και αντικείμενα (5)
- Σε κάθε βήμα ομαδοποιήσαμε τα «όμοια» αντικείμενα με σκοπό τη δημιουργία νέων clusters .



Ιεραρχικής Ανάλυσης Συστάδων

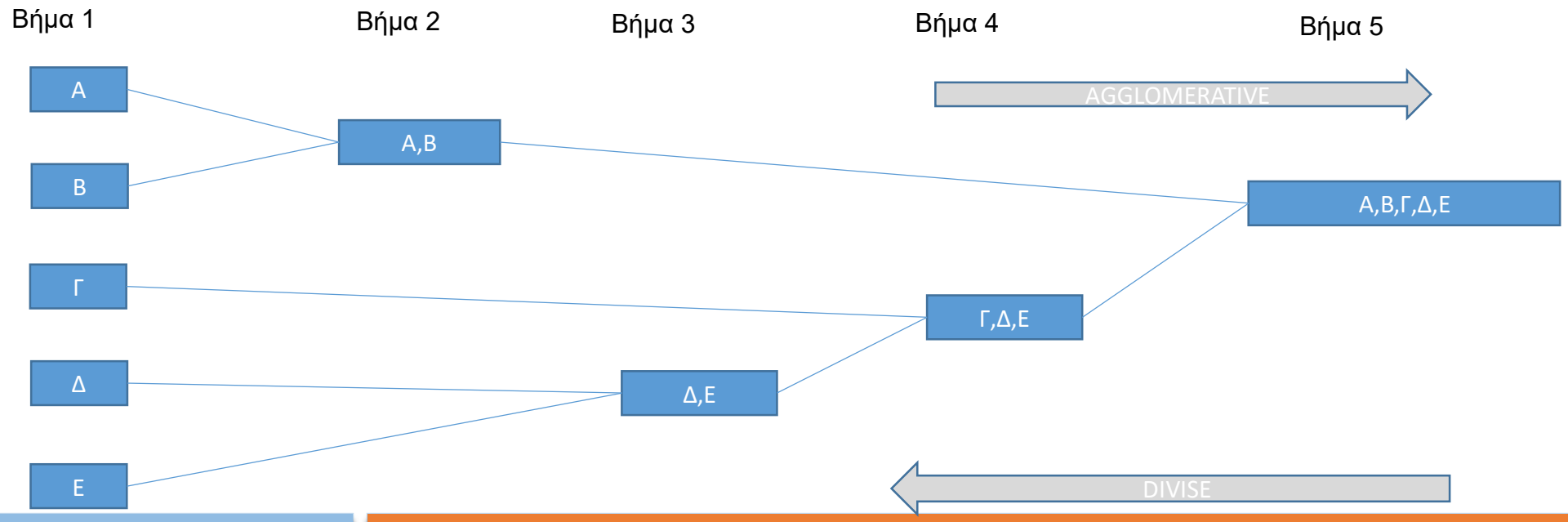
Βήματα Αλγορίθμου

1. Δημιουργώ τον πίνακα αποστάσεων για όλα τα αντικείμενα.
2. Βρίσκω ποια αντικείμενα έχουν τη μικρότερη απόσταση και τα ενώνω.
3. Αν δεν έχουν μπει όλα τα αντικείμενα σε μια ομάδα πηγαίνουμε ξανά στο βήμα 1.

Ιεραρχικής Ανάλυσης Συστάδων

Υπάρχουν 2 τύποι ιεραρχικής ανάλυσης

- Συσσωρευτικές (Agglomerative). Ξεκινάμε τη δημιουργία των clusters με κάθε cluster να αποτελείται από ένα αντικείμενο. Στην συνέχεια τα όμοια αντικείμενα τοποθετούνται μαζί δημιουργώντας νέα clusters μέχρι όλα τα αντικείμενα να καταλήξουν σε ένα cluster. Αυτή η μέθοδος είναι «από κάτω προς τα επάνω» (bottom up)
- Διαιρετικές (Divisive). Ξεκινά η διαδικασία με ένα cluster, και σε κάθε βήμα σπάει σε μικρότερα cluster μέχρι να φτάσουμε σε cluster όπου οι υποομάδες οι οποίες θα προκύψουν να έχουν τη μεγαλύτερη ανομοιότητα. Η μέθοδος είναι «από επάνω προς τα κάτω» (top down)



Ιεραρχική Ανάλυση

- Δεν υπάρχει συγκεκριμένος αριθμός clusters.
- Ο αριθμός των clusters προκύπτει από το σημείο στο οποίο θα σταματήσουμε το διαχωρισμό σε clusters.
- Η τοποθέτηση των αντικειμένων σε cluster πραγματοποιείται κάνοντας χρήση συναρτήσεων που μετράνε την ομοιότητα ή την απόσταση μεταξύ των αντικειμένων.

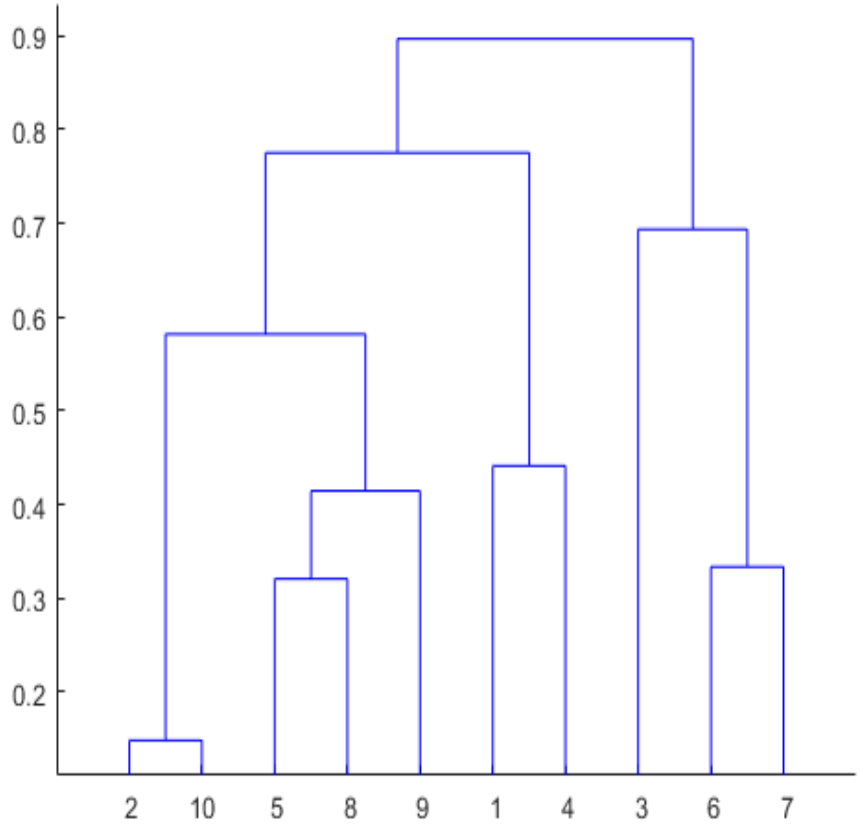
Ιεραρχική Ανάλυση

Η απόσταση μεταξύ των clusters μετριέται με τις παρακάτω μεθόδους

- Single link (MIN). Ψάχνουμε να βρούμε τη μικρότερη απόσταση μεταξύ ενός αντικειμένου σε ένα cluster και ενός αντικειμένου σε ένα άλλο cluster.
- Complete link (MAX). Ψάχνουμε να βρούμε τη μεγαλύτερη απόσταση μεταξύ ενός αντικειμένου σε ένα cluster και ενός αντικειμένου σε ένα άλλο cluster.
- Average. Εξετάζουμε τον Μ.Ο. της απόσταση μεταξύ ενός αντικειμένου σε ένα cluster και ενός αντικειμένου σε ένα άλλο cluster.

Δενδρόγραμμα Ιεραρχικής Ανάλυσης

- Τα Δενδρογράμματα αποτελούν τον πιο διαδεδομένο τρόπο αναπαράστασης της διαδικασίας των διαδοχικών συγχωνεύσεων ή διασπάσεων.
- Στο κάτω μέρος του δενδρογράμματος βρίσκονται τα αντικείμενα και κάθε κόμβος του δένδρου αντιπροσωπεύει μια συστάδα.
- Μέσω του Δενδρογράμματος ο χρήστης μπορεί να έχει εικόνα των συστάδων που δημιουργούνται και να αποφασίσει το σημείο στο οποίο θα τερματίσει τη διαδικασία συγχώνευσης ή δημιουργίας νέων clusters.



k-Means

- Στόχος της μεθόδου K-means αποτελεί ο διαμοιρασμός ενός συνόλου αντικειμένων σε ένα προκαθορισμένο αριθμό clusters κατά τέτοιο τρόπο έτσι ώστε να δημιουργούνται ομάδες ομοειδών αντικειμένων.
- Για τις ανάγκες διαμοιρασμού των αντικειμένων στα clusters, υπολογίζουμε κάθε φορά την απόσταση του αντικειμένου από το κέντρο του cluster και το αντικείμενο εντάσσεται στη συστάδα με το πλησιέστερο κέντρο.

k-Means

Ο αλγόριθμος K-means αποτελείται από τα παρακάτω βήματα

1. Αρχικά ο χρήστης καθορίζει τον αριθμό των clusters (*k).*
2. *Επιλέγονται τυχαία K αντικείμενα τα οποία αποτελούν και τα πρώτα κέντρα των clusters.*
3. Στην συνέχεια κάθε αντικείμενο εντάσσεται στο *cluster* του οποίου το κέντρο είναι πλησιέστερα του.
4. Για τον υπολογισμό της απόστασης συνήθως χρησιμοποιείται η Ευκλείδεια απόσταση.
5. Κάθε φορά που εντάσσεται ένα νέο αντικείμενο τα κέντρα επαναυπολογίζονται.

k-Means

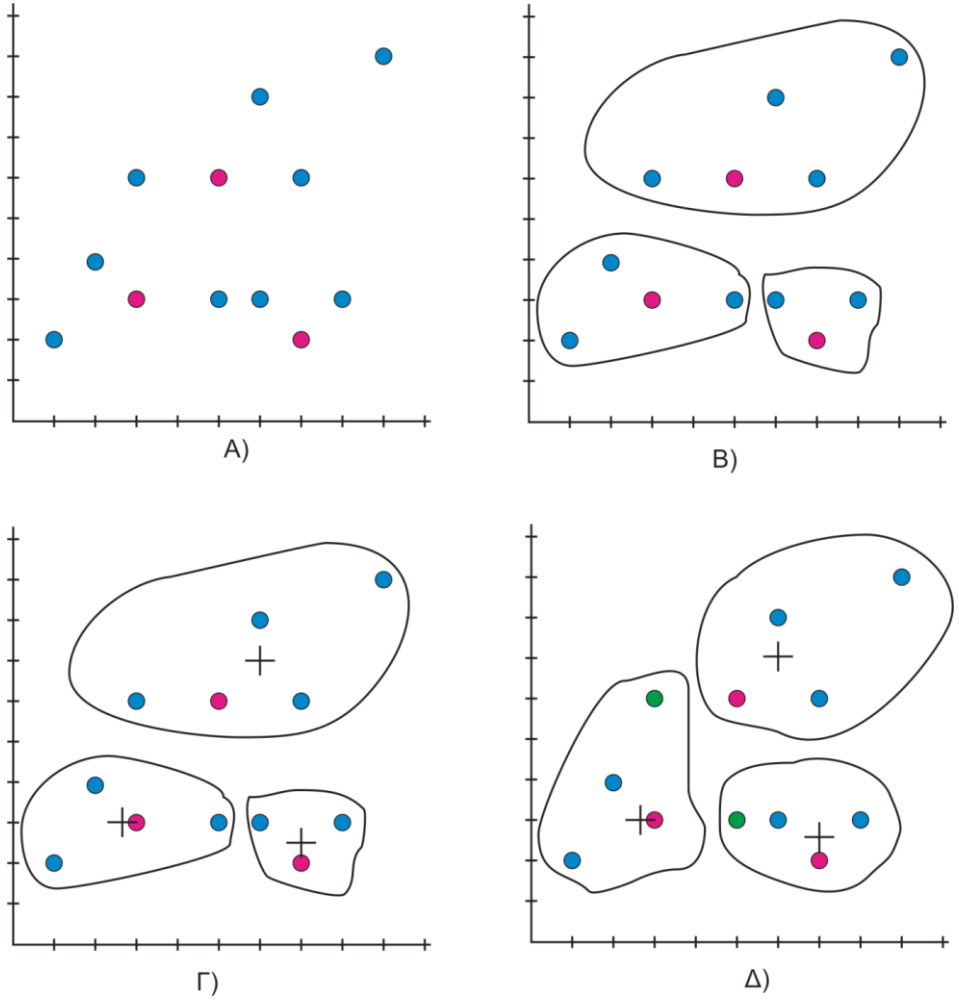
Πλεονεκτήματα

- Απαιτείται λιγότερος χρόνος σε σχέση με τις ιεραρχικές μεθόδους.

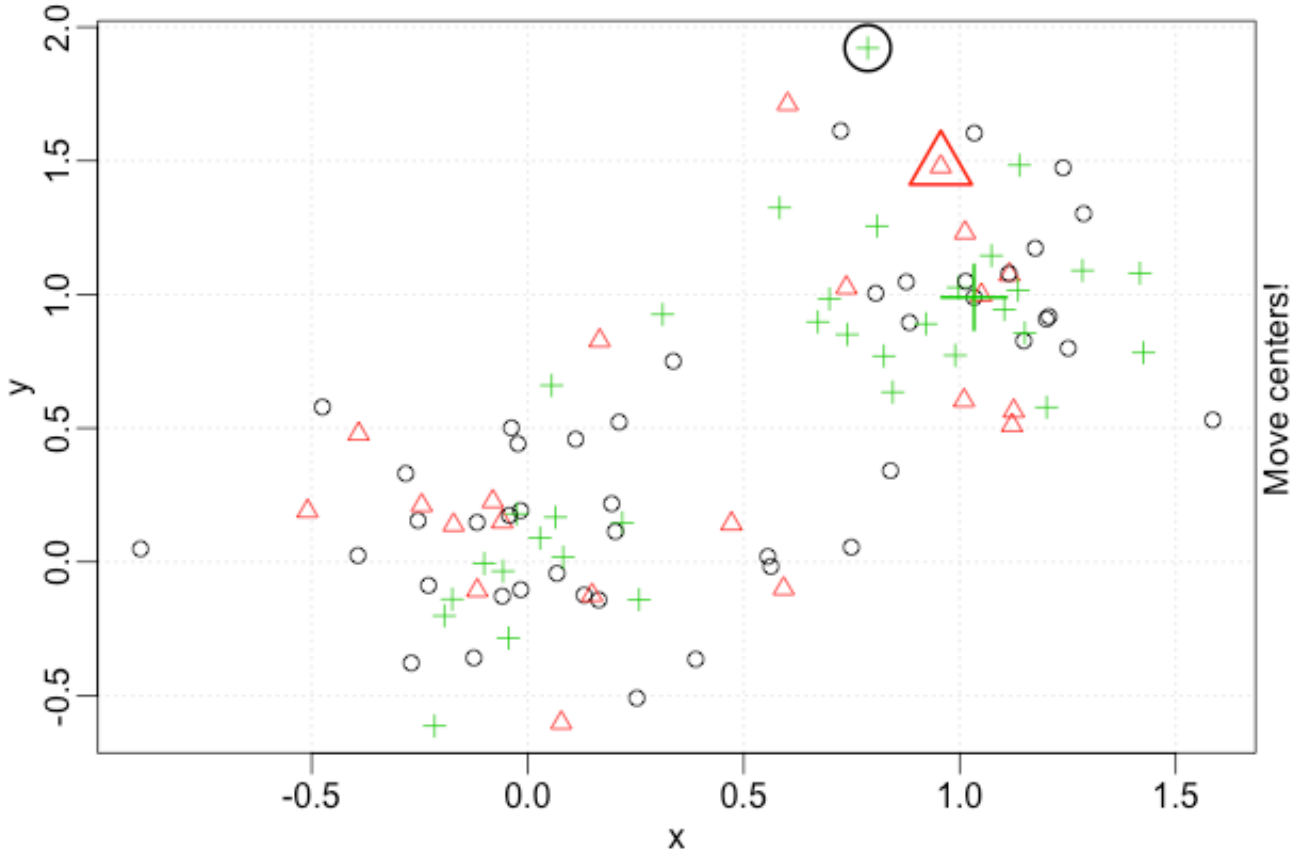
Μειονεκτήματα

- Η επιλογή των αρχικών κέντρων καθορίζει το τελικό αποτέλεσμα.
- Τα αποτελέσματα επειρεάζονται από την ύπαρξη ακραίων τιμών (outliers).
- Ο αριθμός των clusters καθορίζεται στην αρχή

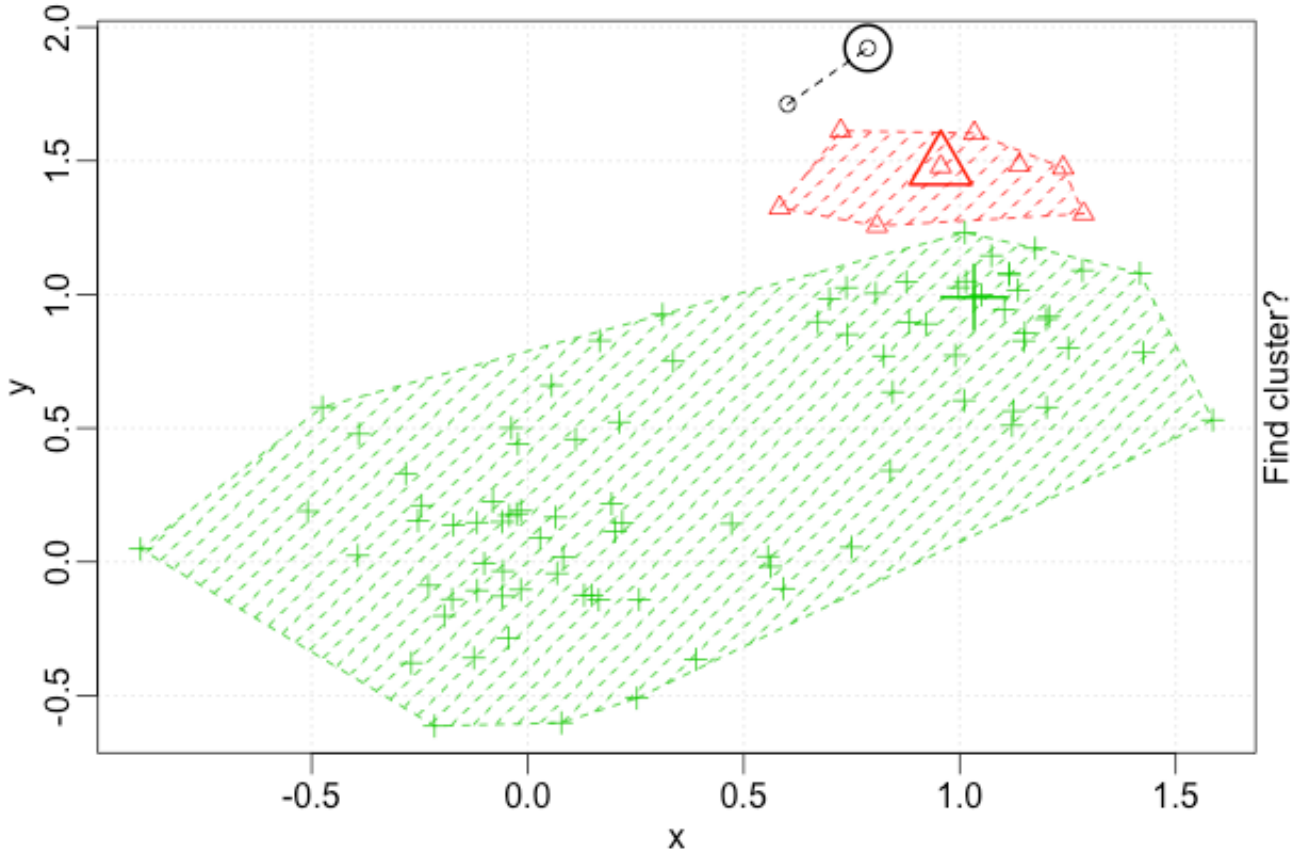
k-Means



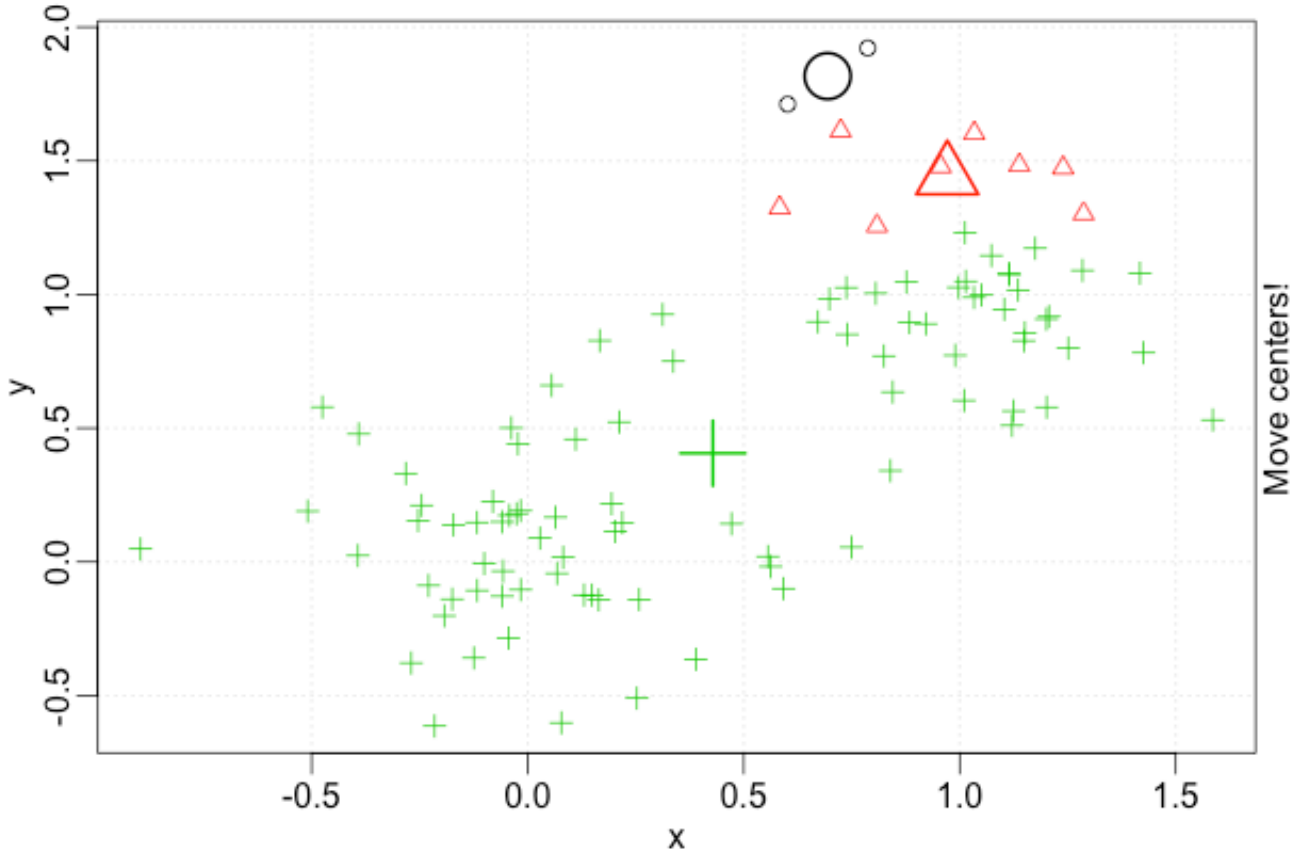
K-means



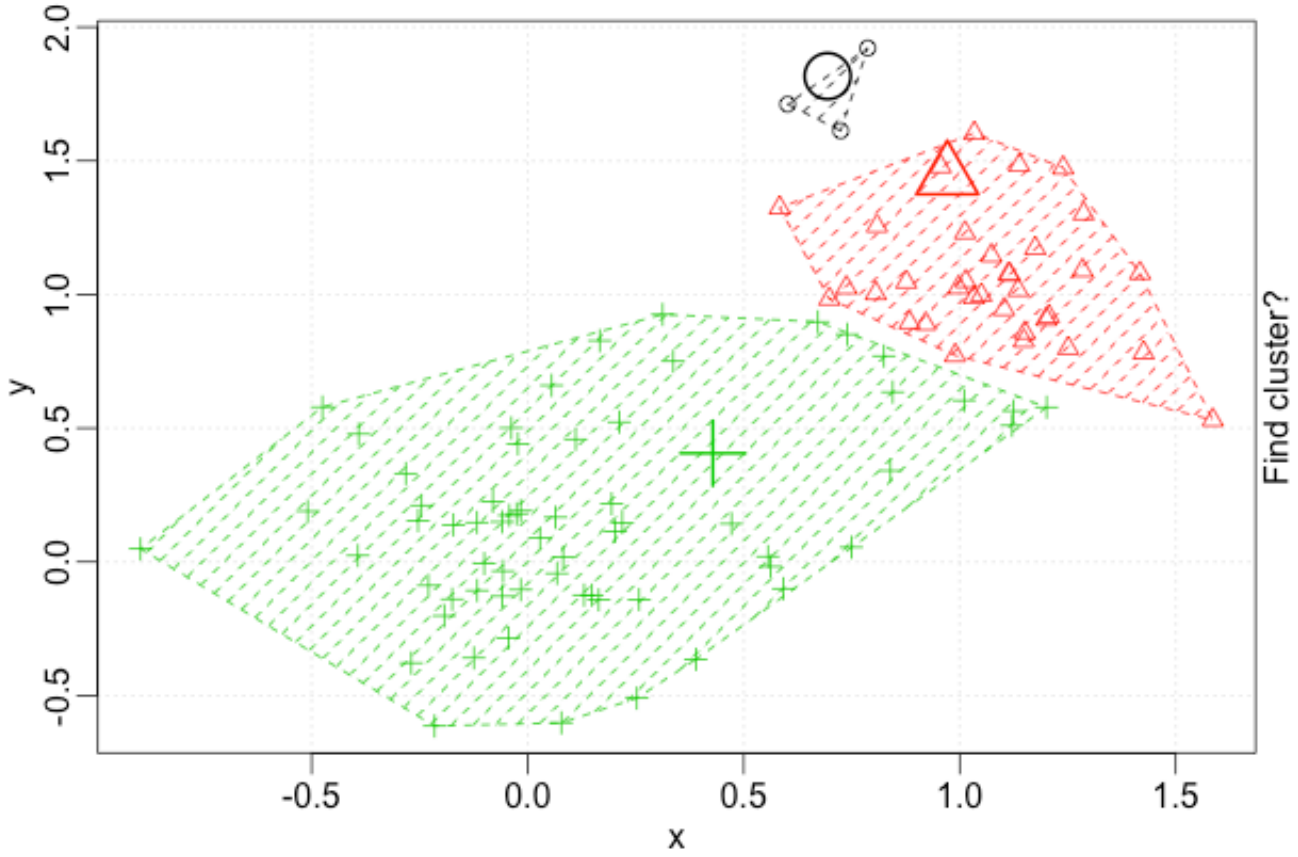
K-means



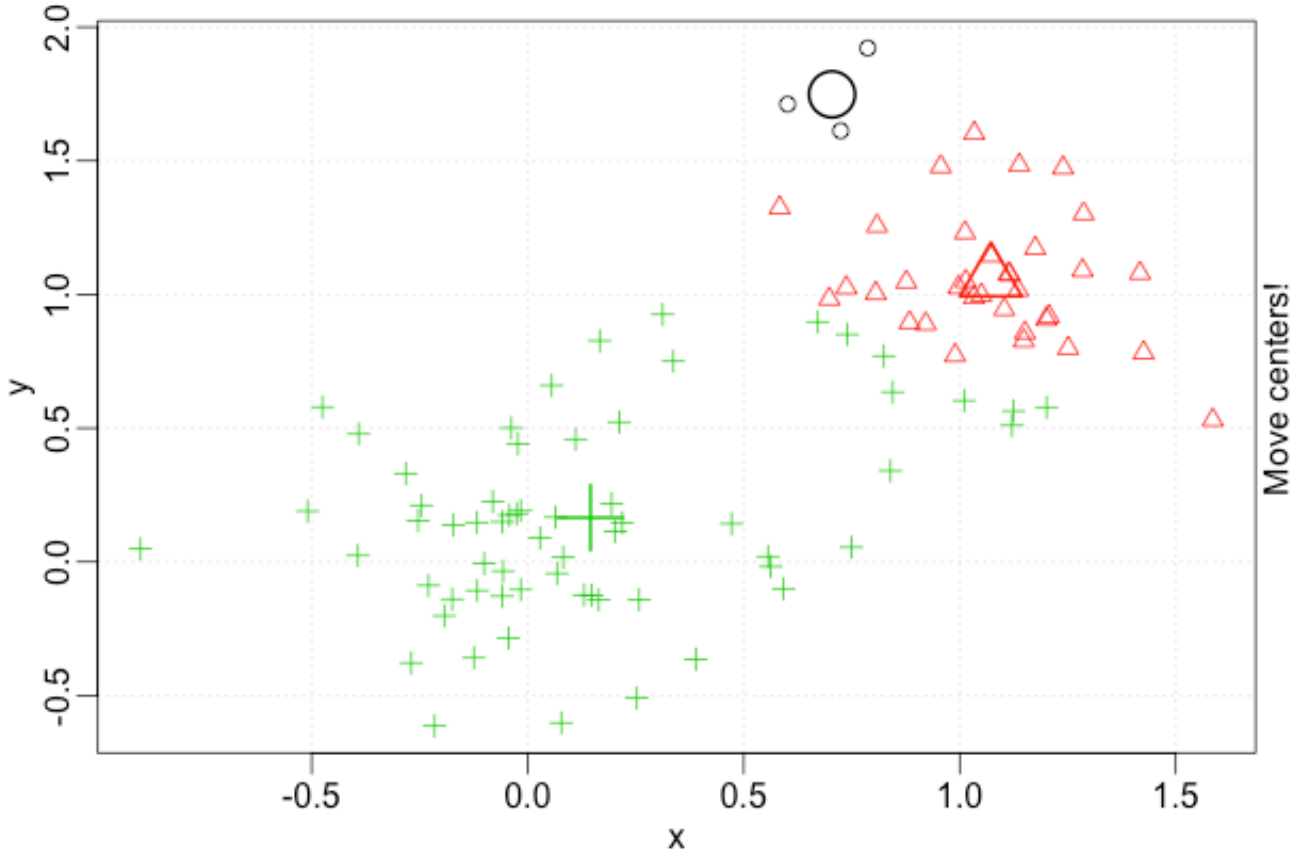
K-means



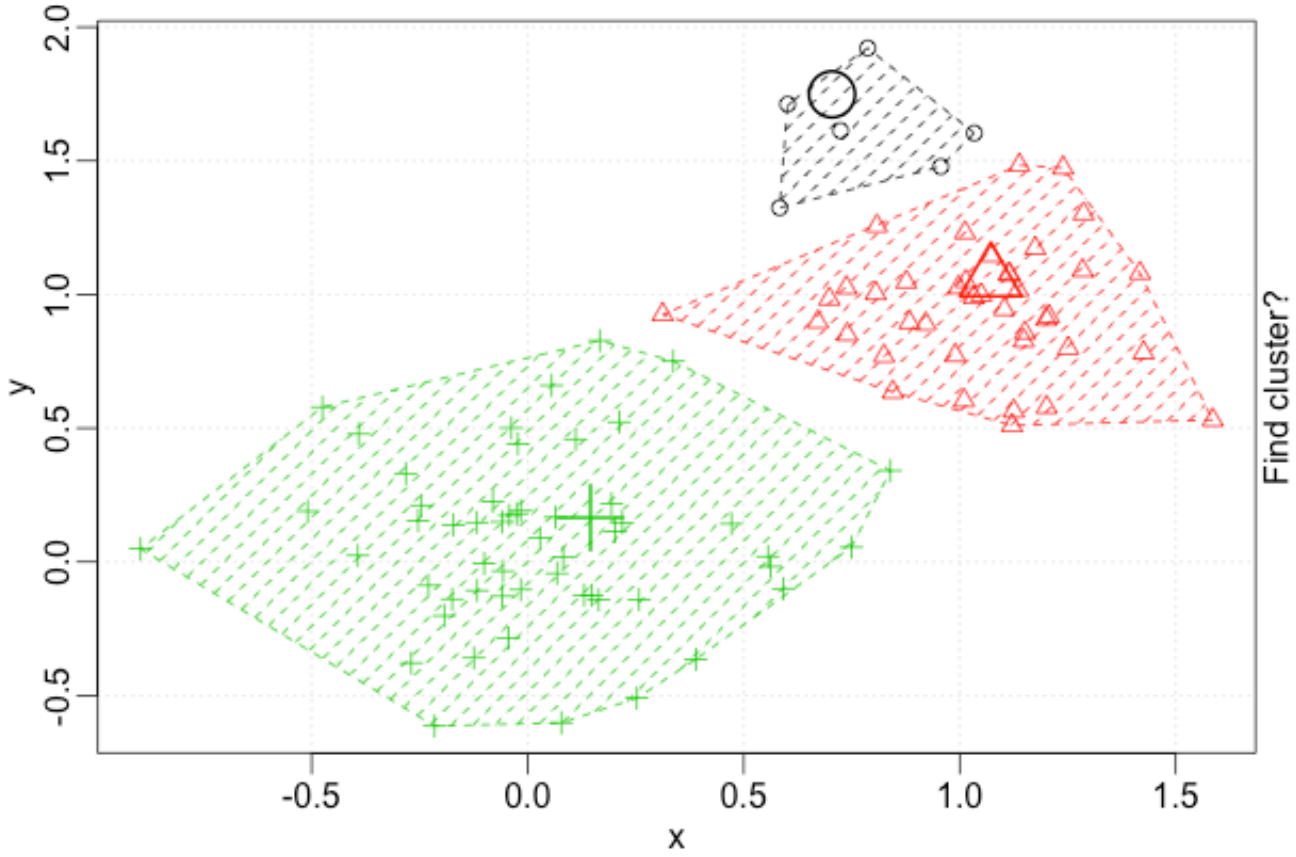
K-means



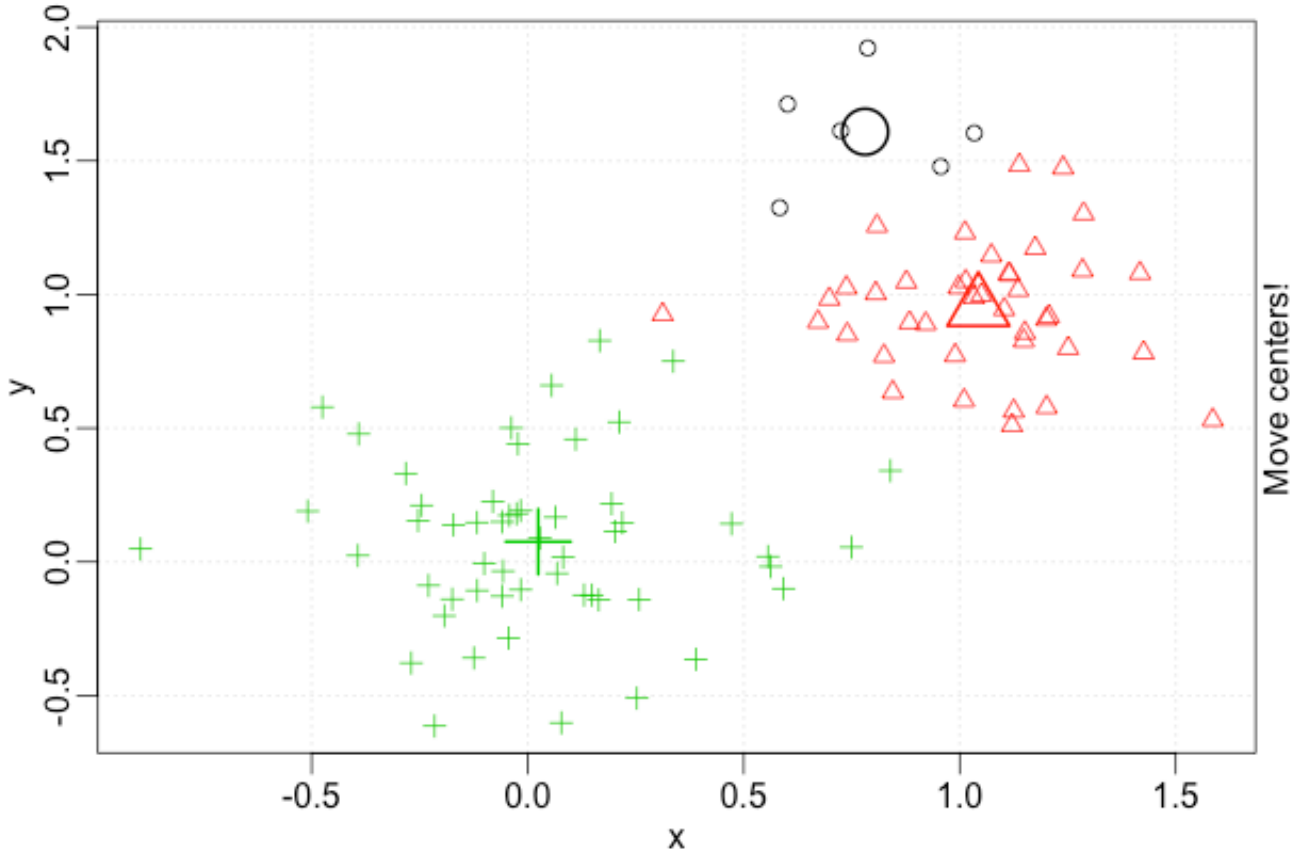
K-means



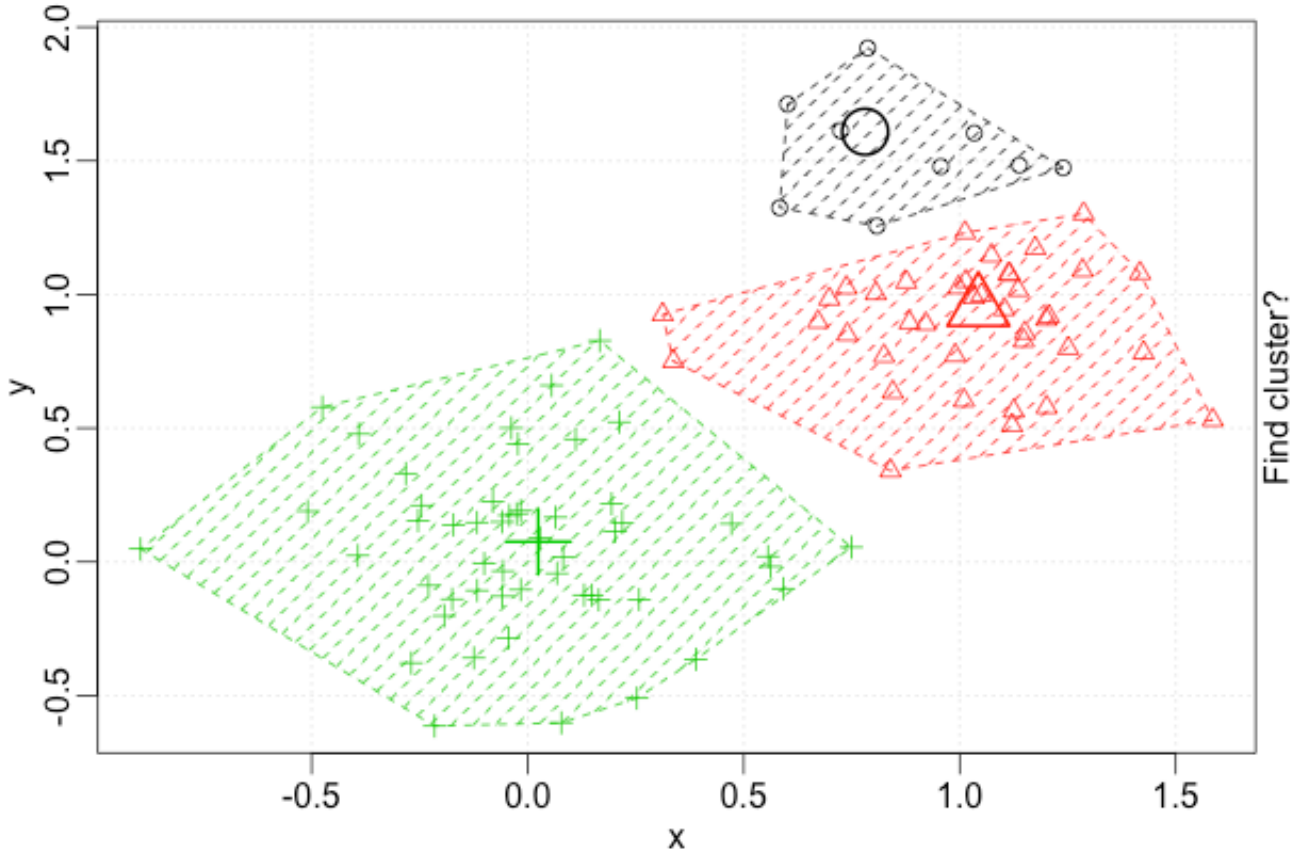
K-means



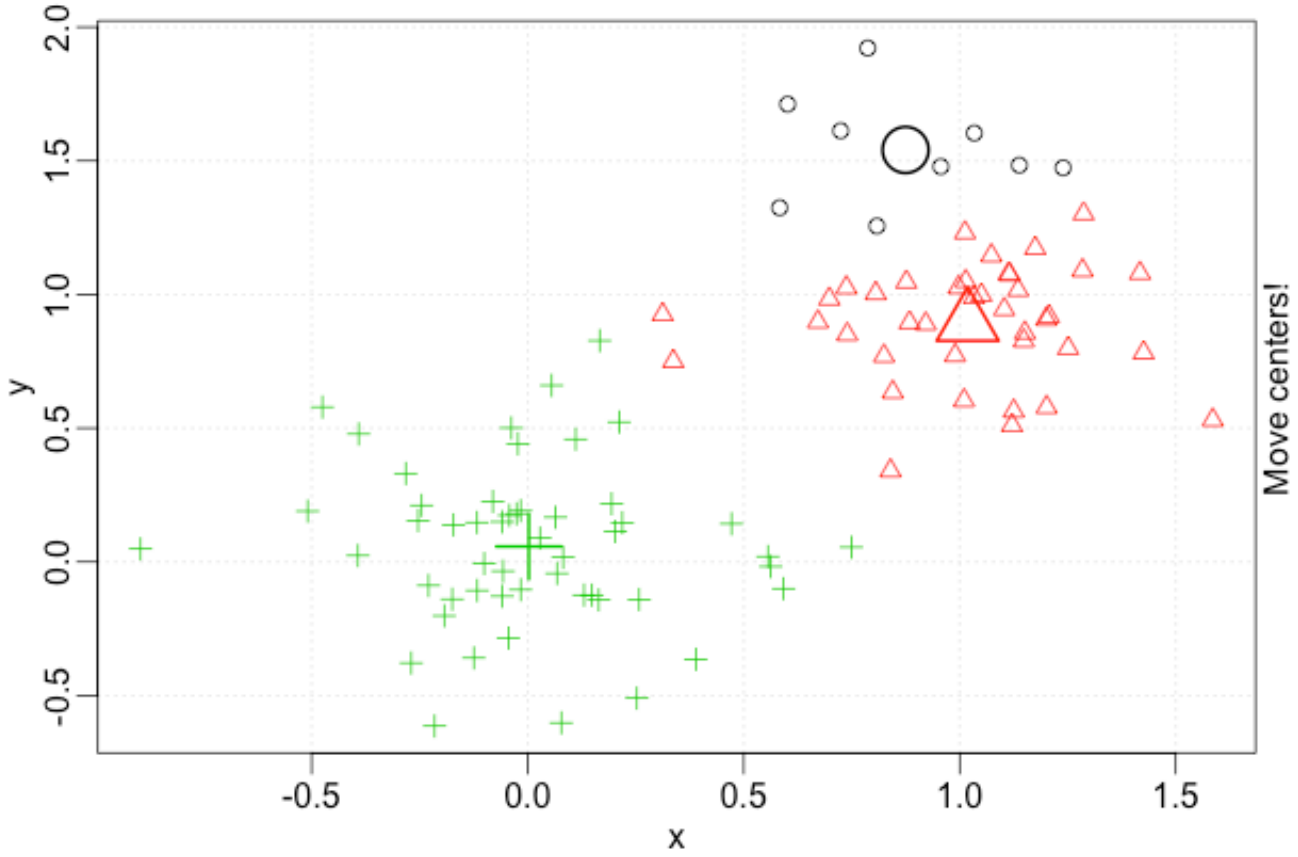
K-means



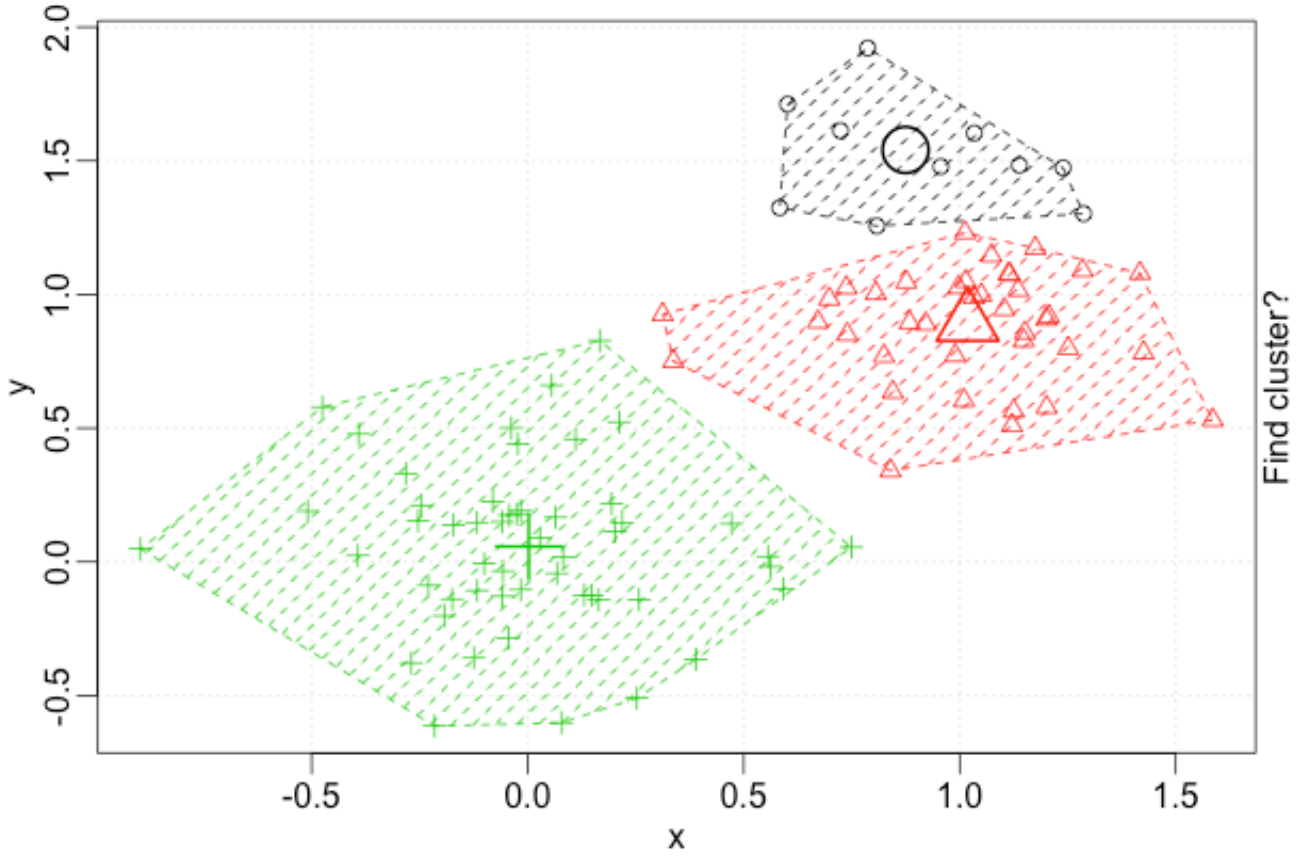
K-means



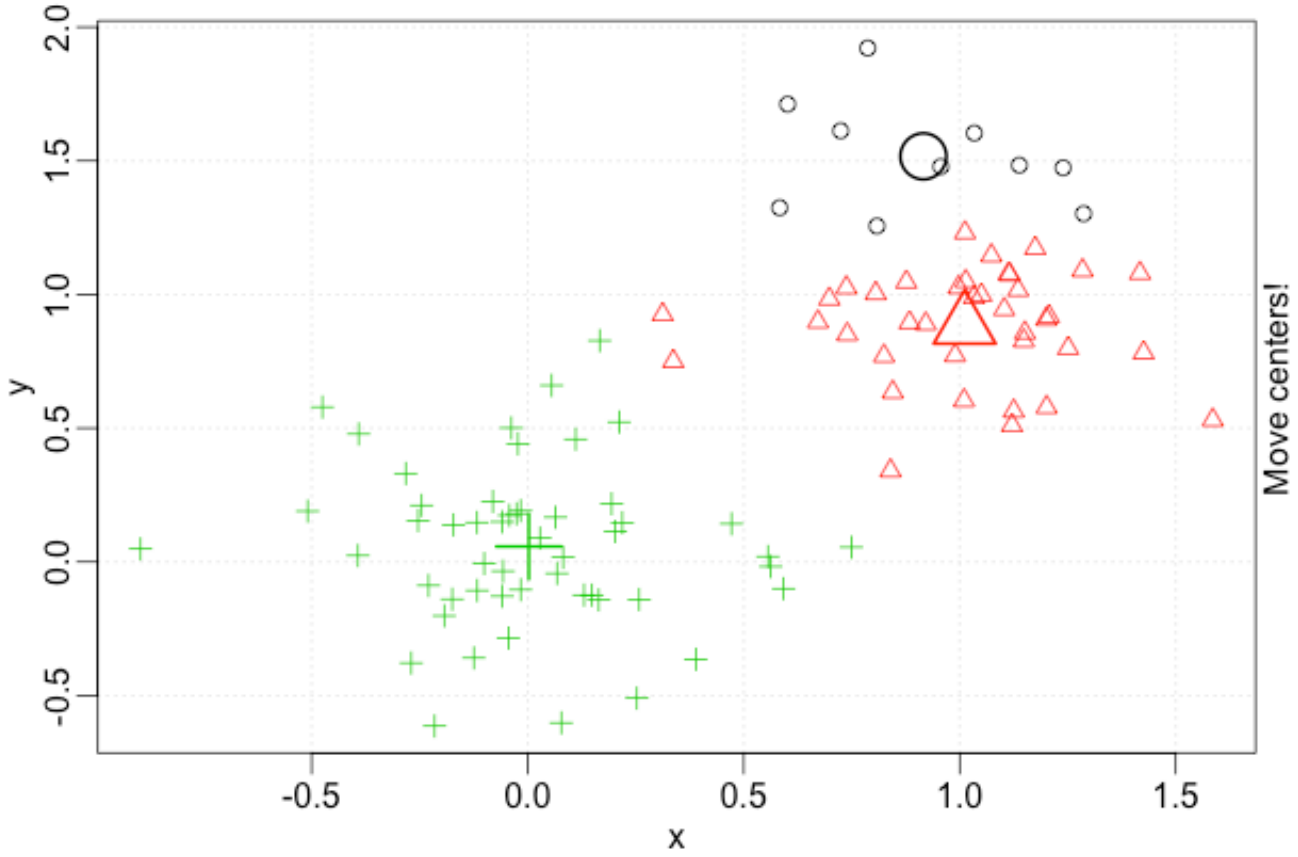
K-means



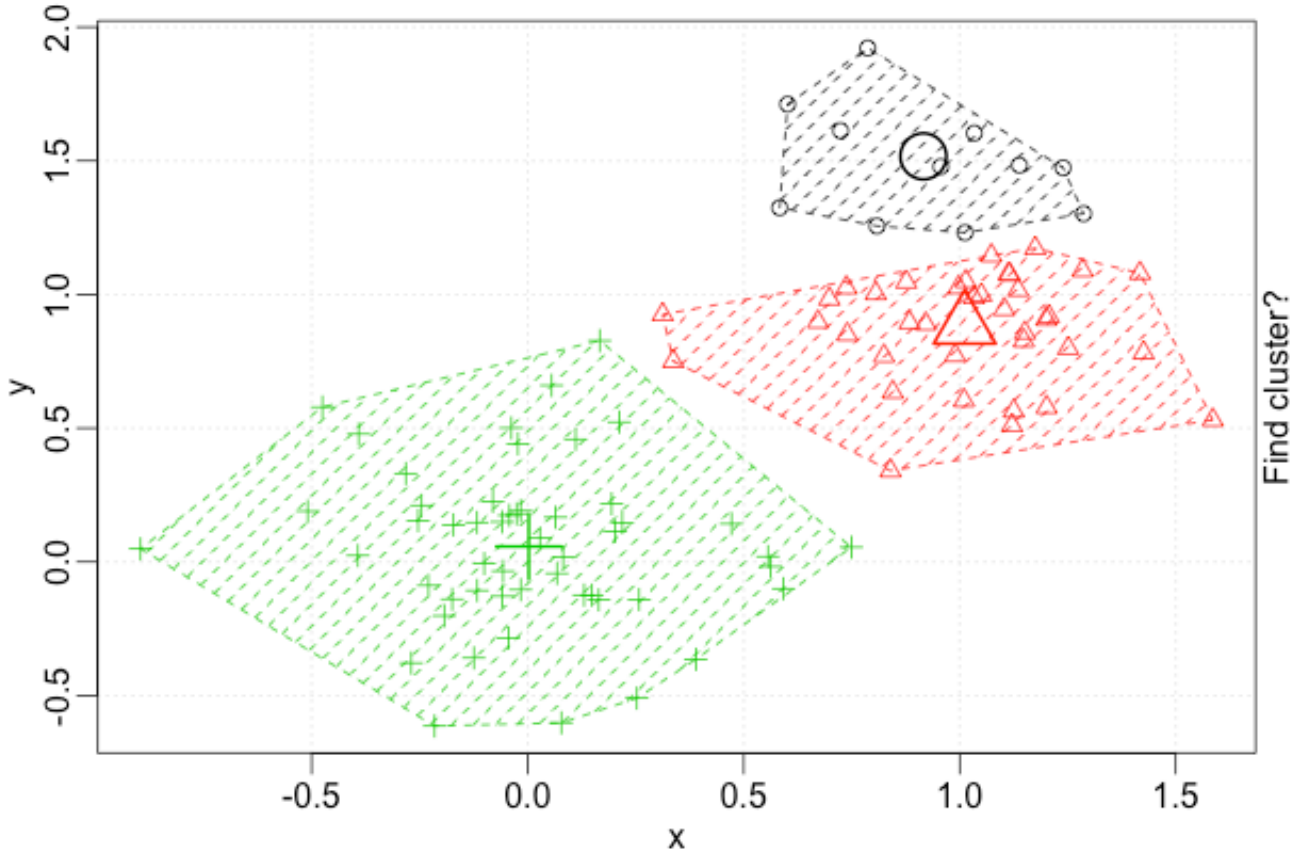
K-means



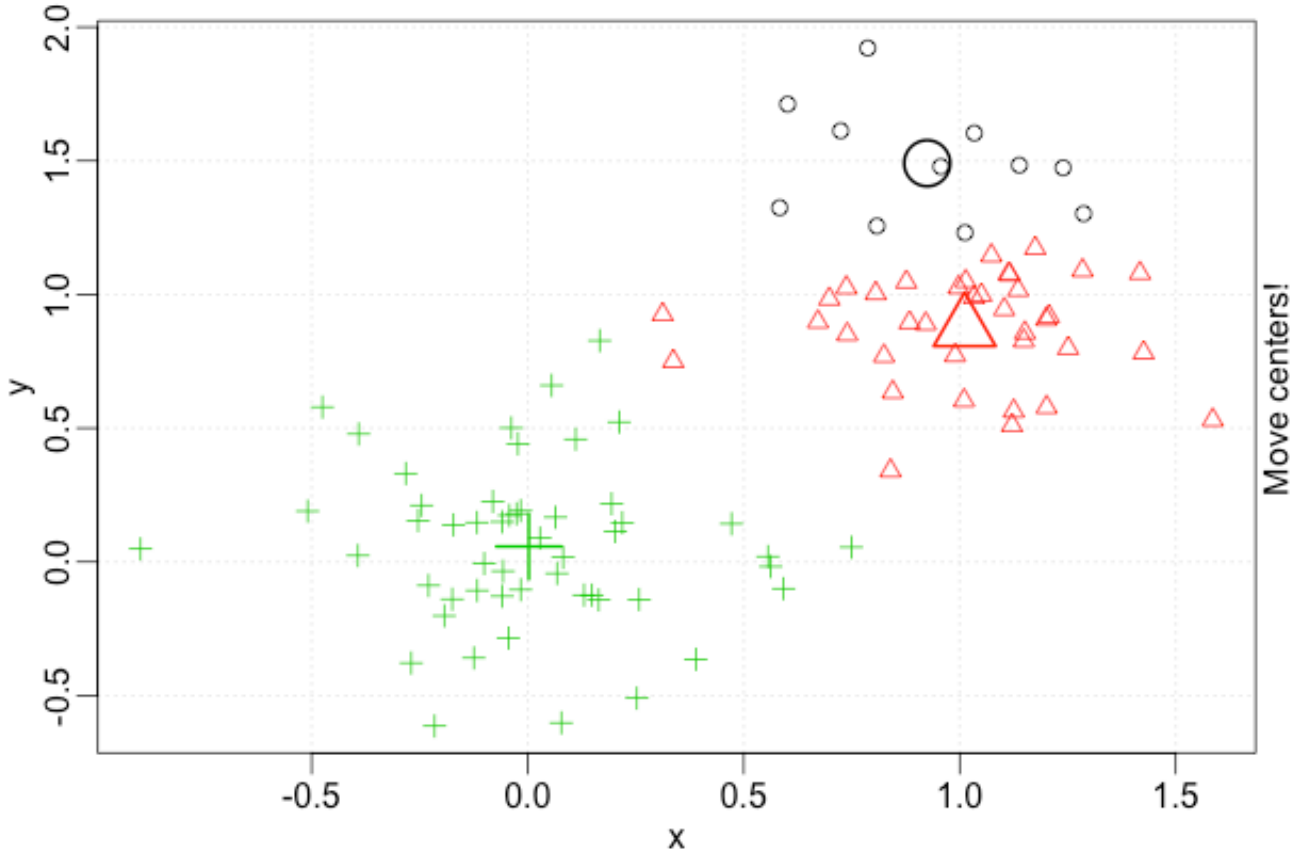
K-means



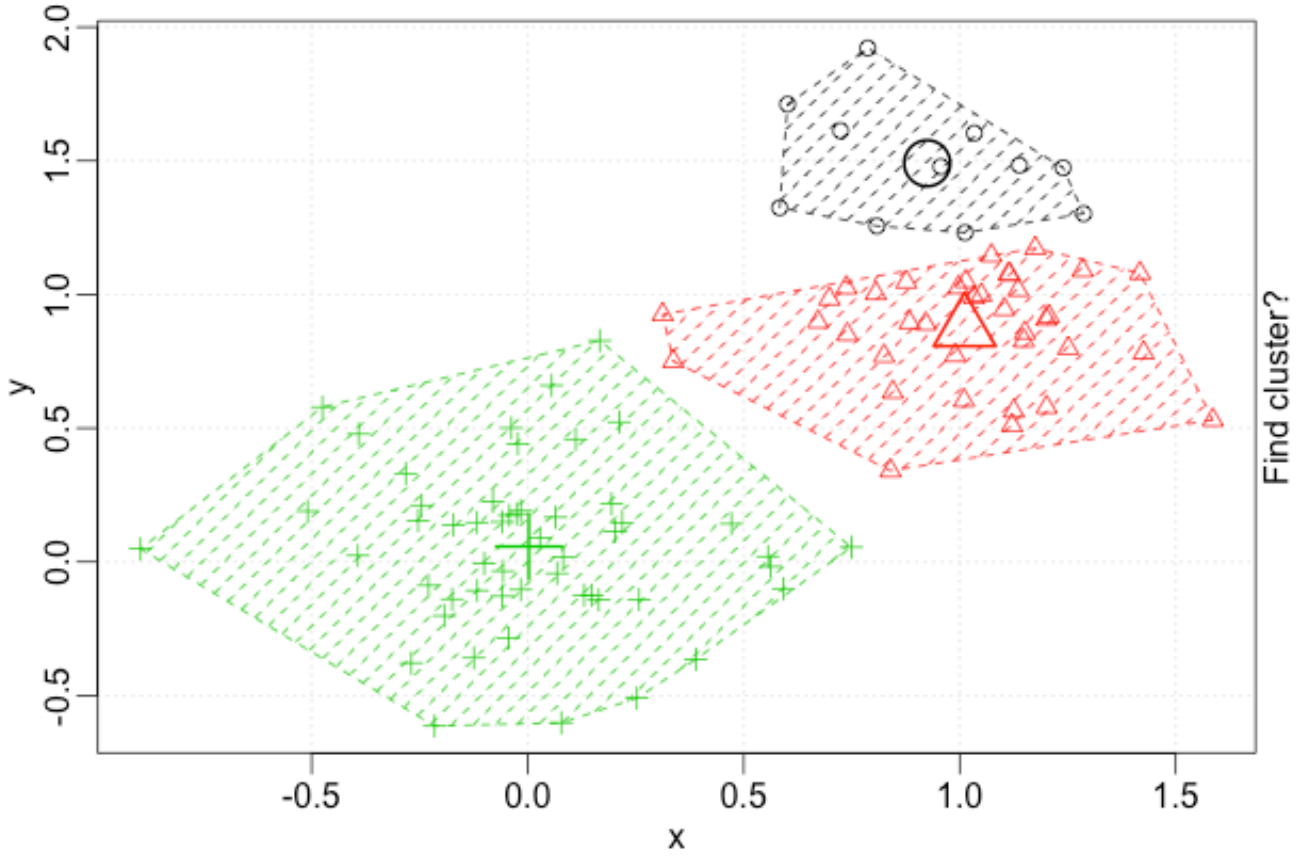
K-means



K-means



K-means



ΤΕΛΟΣ ΕΝΟΤΗΤΑΣ