Ανάλυση σε Ομάδες (Cluster Analysis)

Περιεχόμενα:

1. Εισαγωγή και Παράδειγμα

2. Ιεραρχικές μέθοδοι
   (Hierarchical methods).

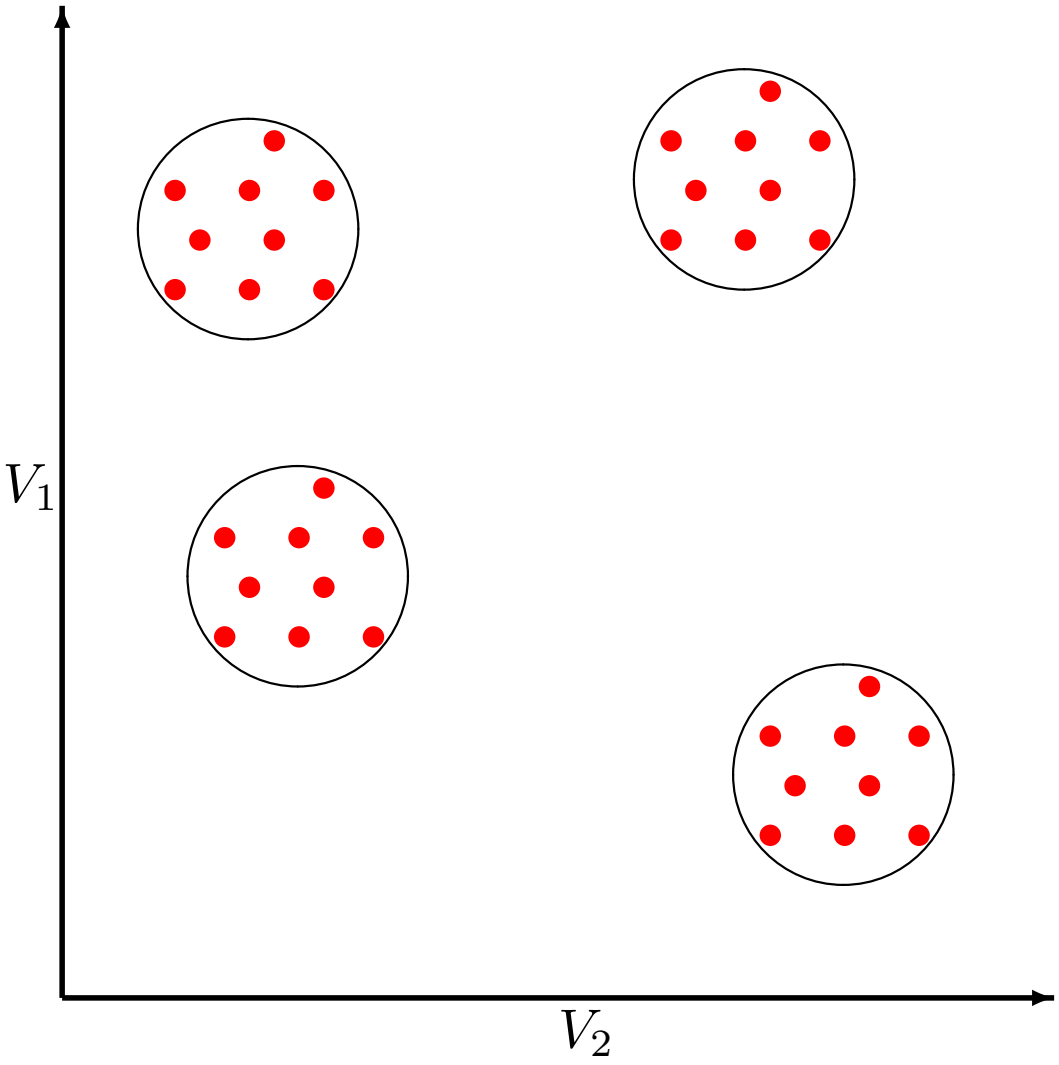
3. Μη-ιεραρχικές μέθοδοι
   (Non-hierarchical methods).

4. Παραδείγματα.

Ανάλυση Ομάδων

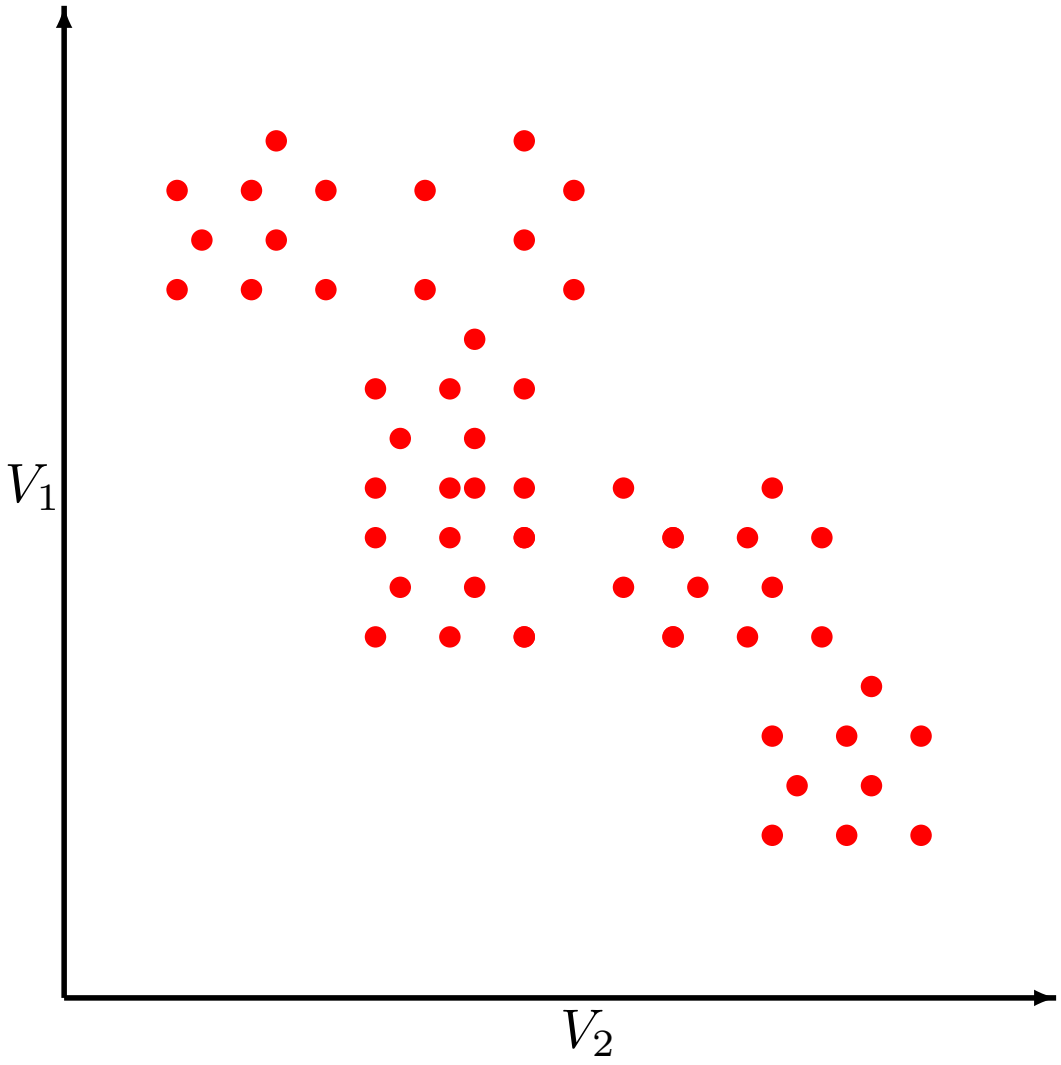
- Σκοπός της ανάλυσης σε ομάδες είναι η ομαδοποίηση δεδομένων σε ομοιογενείς ομάδες βάση διαφόρων μεταβλητών. Τα στοιχεία κάθε ομάδας είναι όμοια μεταξύ τους βάση αυτών των μεταβλητών και διαφέρουν από τα δεδομένα σε άλλες ομάδες.
- Οι ομάδες εσωτερικά παρουσιάζουν μεγάλη ομοιογένεια ενώ μεταξύ των ομάδων υπάρχει μεγάλη ανομοιογένεια. Έτσι, σε μια επιτυχή κατηγοριοποίηση, οι παρατηρήσεις μέσα στην ομάδα θα βρίσκονται η μια κοντά στην άλλη, όταν αυτές αναπαρασταθούν γεωμετρικά. Δύο διαφορετικές ομάδες θα βρίσκονται μακριά η μία από την άλλη.
- Η ανάλυση σε ομάδες ονομάζεται επίσης *ομαδοποίηση* ή *αριθμητική ταξονομία (numerical taxonomy)*.
- Εμείς θα ασχοληθούμε με μεθόδους σχηματισμού ομάδων όπου κάθε παρατήρηση ανήκει σε μία μόνο ομάδα.
- Για την ανάλυση ομάδων θεωρούμε ότι δεν υπάρχουν εκ των προτέρων πληροφορίες σχετικά με τις ομάδες ή τις παρατηρήσεις. Τα δεδομένα προτείνουν τις ομάδες που θα πρέπει να δημιουργηθούν και δεν καθορίζονται από πριν.

Μια ιδανική ομοδοποίηση



- Πιο πάνω φαίνεται μια ιδανική ομαδοποίηση όπου οι ομάδες διαχωρίζονται ευδιάκριτα βάση δύο μεταβλητών (V_1: ποιότητα του προϊόντος, και V_2: τιμή αγαθού). Σημειώστε ότι κάθε καταναλωτής εμπίπτει μόνο σε μια ομάδα και δεν υπάρχει επικάλυψη.

Μια πρακτική ομαδοποίηση



- Η πιο πάνω ομαδοποίηση ειναι πιο πιθανόν να παρατηρηθεί στην πράξη. Τα όρια των ομάδων δεν είναι ευδιάκριτα και η κατηγοριοποίηση κάποιων παρατηρήσεων δεν είναι εύκολη. Αυτό συμβαίνει γιατί κάποιες παρατηρήσεις θα μπορούσαν να ανήκουν σε διάφορες ομάδες.

Υποθέσεις στην ανάλυση ομάδων

- Η αναλυση ομάδων δεν θεωρείται στατιστική μέθοδος (statistical inference technique) όπου οι δειγματικοί παράμετροι μπορούν να αντιπροσωπεύσουν του πληθυσμούς. Αντίθετα η ανάλυση σε ομάδες θεωρείται μια ουσιαστική μεθοδολογία μέτρησης των χαρακτηριστικών των παρατηρήσεων που αφορούν στη δομή τους.

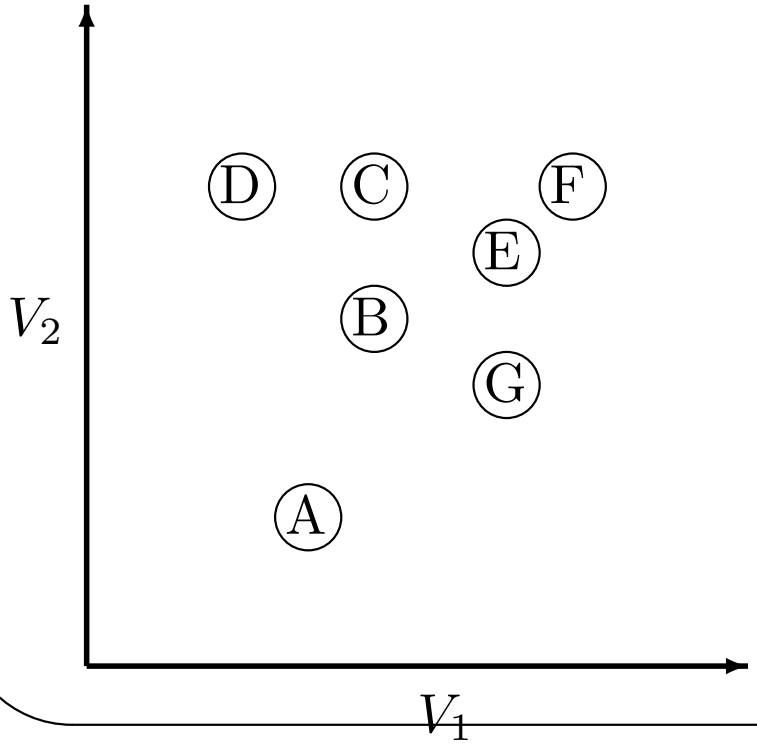
- Η μεθοδολογία αν και έχει ισχυρές μαθηματικές ιδιότητες, δεν στηρίζεται σε στατιστικές βάσεις. Οι υποθέσεις κανονικότητας, γραμμικότητας και ομοσκεδαστικότητας (normality, linearity and homoscedasticity) που είναι σαφώς σημαντικές σε άλλες μεθοδολογίες, δεν ισχύουν στην ανάλυση ομάδεων. Ο ερευνητής πρέπει να δώσει σημασία στο πόσο αντιπροσωπευτικό είναι το δείγμα.

- Η ανάλυση ομάδων χρησιμοποιείται για την απεικόνηση του δείγματος. Έτσι μεγάλη έμφαση πρέπει να αποδείδεται στο πόσο αντιπροσωπευτικό είναι το δείγμα και αν μπορεί να χρησιμοποιηθεί στη γενικοποίηση των αποτελεσμάτων του πληθυσμού.

Ανάλυση ομάδων: Παράδ. δύο μεταβλητών

- Υποθέστε ότι ένας λειτουργός μάρκετιγκ θέλει να κατηγοριοποιήσει μια μικρή αγορά βάση της αφοσίωσης που δείχνουν σε συγκεκριμένες μάρκες (λογότυπα) και σε καταστήματα. Για το σκοπό αυτό έχει επιλεγεί ένα μικρό δείγμα 7 καταναλωτών. Δύο παράμετροι για μέτρηση της αφοσίωσης έχου επιλεγεί σε κλίμακα από 0-ως-10: V_1 (αφοσίωση στο κατάστημα) και V_2 (αφοσίωση στη μάρκα).

Παράμετρος ομαδοποίησης	Καταναλωτές						
	A	B	C	D	E	F	G
V_1	3	4	4	2	6	7	6
V_2	2	5	7	7	6	7	4



- Ο βασικός σκοπός της ανάλυσης σε ομάδες είναι να ορίσει τη δομή των δεδομένων με τη κατηγοριοποίση όμοιων παρατηρήσεων σε ομάδες.

- Πώς μετρούμε την ομοιογέια;

  Χρειαζόμαστε μια μέθοδο που να συγκρίνει ταυτόχρονα παρατηρήσεις βάση των μεταβλητών ομαδοποίησης (V_1 και V_2). Π.χ. Απόσταση.

- Πώς δημιουργούνται οι ομάδες;

  Ανεξάρτητα από τη μέθοδο ομοιογένειας που λαμβάνεται υπόψη, σκοπός είναι η κατηγοριοποίηση των πιο όμοιων παρατηρήσεων σε ομάδες. Έτσι, η διαδικασία καθορίζει τις ομάδες των παρατηρήσεων.

- Πόσες ομάδες πρέπει να συντάξουμε;

  Υπάρχουν αρκετοί κανόνες που μπορούν να χρησιμοποιηθούν. Όμως, ο βασικός είναι η ανάθεση της *μέσης* ομοιογένειας στις ομάδες έτσι ώστε: καθώς ο μέσος αυξάνεται, η ομάδα να γίνεται λιγότερο ομοιογενής.

Δίλημμα: λιγότερες ομάδες έναντι λιγότερης ομοιογένειας.

Καθώς ο αριθμός των ομάδων μειώνεται, η ομοιογένεια μέσα στις ομάδες μειώνεται. Θα πρέπει να υπάρξει μια ισορροπία μεταξύ του αριθμού των ομάδων και του βαθμού ομοιογένειας μέσα στις ομάδες.

Proximity Matrix (Μήτρα Εγγύτητας)

- Η ομοιογένεια μετριέται βάση της Ευκλείδιας απόστασης (ευθεία) ανα δύο παρατηρήσεις. Η απόσταση μεταξύ δύο σημείων (x_1, y_1) και (x_2, y_2) δίνεται από:

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

Όσο μικρότερη είναι η απόσταση, τόσο μεγαλύτερη είναι η ομοιγένεια, ενώ όσο μεγαλύτερη είναι ενδυκνύει ανομοιογένεια.

- Ως μέτρο ομοιογένειας χρησιμοποιούμε το άθροισμα του τετραγώνου των σφαλμάτων, Error Sum of Squares (ESS).

Άθροισμα τετραγώνου σφαλμάτων (ESS)

Το ESS είναι το άθροισμα του *τετραγώνου της απόστασης* των παρατηρήσεων από το μέσο της ομάδας.

Π.χ. $E = (6, 6)$, $F = (7, 7)$ και $G = (6, 4)$. Ο μέσος του E και F δίνεται από:

$$M = (\frac{6+7}{2}, \frac{6+7}{2}) = (6.5, 6.5).$$

Το ESS του $\{E, F\}$ δίνεται από:

$$\begin{aligned} \text{ESS}_{\{E,F\}} &= \Big((6-6.5)^2 + (6-6.5)^2\Big) \\ &\quad + \Big((7-6.5)^2 + (7-6.5)^2\Big)\Big) \\ &= 0.5^2 + 0.5^2 + 0.5^2 + 0.5^2 \\ &= 1. \end{aligned}$$

Ο μέσος των E, F και G δίνεται από:

$$M = (\frac{6+7+6}{3}, \frac{6+7+4}{3}) = (\frac{19}{3}, \frac{17}{3}).$$

Επίσης:

$$\text{ESS}_{\{E,F,G\}} = 5\frac{1}{3}.$$

Παρατηρ.	Παρατηρήσεις A	B	C	D	E	F	G
A	0						
B	3.16	0					
C	5.10	2.00	0				
D	5.10	2.83	2.00	0			
E	5.00	2.27	2.24	4.12	0		
F	6.40	3.61	3.00	5.00	1.41	0	
G	3.61	2.24	3.61	5.00	2.00	3.16	0

- Υπάρχου 21 αποστάσεις (εάν εξαιρέσουμε τις μηδενικές αποστάσεις). Γενικά για n παρατηρήσεις υπάρχουν $n(n-1)/2$ αποστάσεις.

- Υπάρχουν διαφορετικοί μέθοδοι δημιουργίας ομάδων. Εδώ θα θεωρήσουμε την ιεραρχική μέθοδο.

  Στο πιο κάτω παράδειγμα η ακόλουθη διαδικασία χρησιμοποιείται:

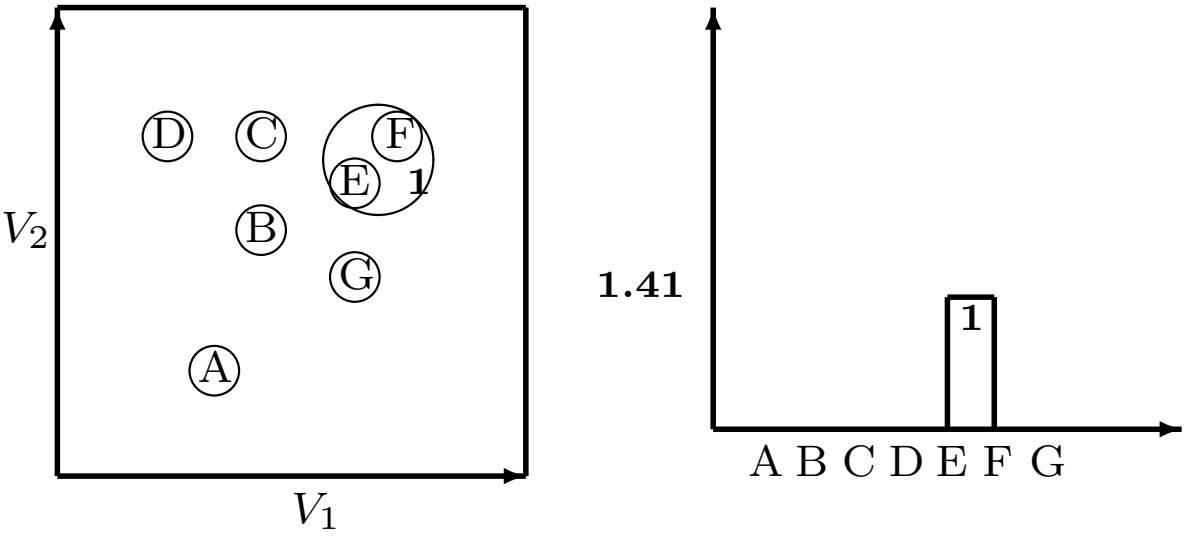
  1. Θεωρούμε ότι κάθε παρατήρηση ορίζει μία ομάδα.
  2. Εντοπίζουμε τις δύο πιο όμοιες (κοντινές) παρατηρήσεις, που δεν ανήκουν ήδη στην ίδια ομάδα, και τις εντάσσουμε σε μια ομάδα.
  3. Επαναλαμβάνουμε το Βήμα 2 μέχρις ότου όλες οι παρατηρήσεις να είναι στην ίδια ομάδα.

  Σημειώστε ότι υπάρχουν εναλλακτικές μέθοδοι που μπορούν να χρησιμοποιηθούν στην ένωση δύο ομάδων.

Στάδιο 1

Παρατηρ.	Παρατηρήσεις A	B	C	D	E	F	G
A	0						
B	3.16	0					
C	5.10	2.00	0				
D	5.10	2.83	2.00	0			
E	5.00	2.27	2.24	4.12	0		
F	6.40	3.61	3.00	5.00	1.41	0	
G	3.61	2.24	3.61	5.00	2.00	3.16	0

Προσδιορίζουμε τις δύο κοντινότερες παρατηρήσεις (E και F) τις εντάσσουμε σε μια ομάδα.



$\mathrm{ESS}_{\{E,F\}} = 1.$

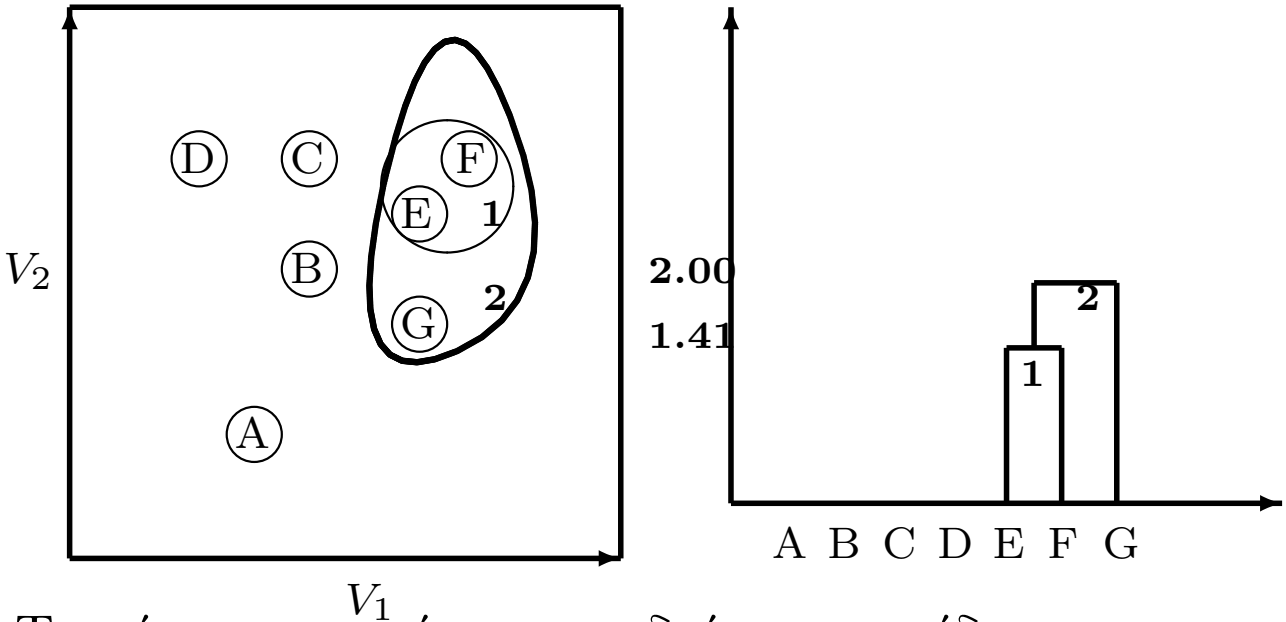
Έτσι, το συνολικό ESS της διαδικασίας ομαδοποίησης:

$$\begin{aligned}\mathrm{ESS} =& \mathrm{ESS}_{\{A\}} + \mathrm{ESS}_{\{B\}} + \mathrm{ESS}_{\{c\}} + \mathrm{ESS}_{\{D\}} \\ &+ \mathrm{ESS}_{\{E,F\}} + \mathrm{ESS}_{\{G\}} = 0+0+0+0+1+0 = 1.\end{aligned}$$

Στάδιο 2

Παρατηρ.	Παρατηρήσεις A	B	C	D	$\{E,F\}$	G
A	0					
B	3.16	0				
C	5.10	2.00	0			
D	5.10	2.27	2.00	0		
$\{E,F\}$	5.00	2.34	2.24	4.12	0	
G	3.61	2.24	3.61	5.00	2.00	0

Βρίσκουμε το επόμενο ζευγάρι παρατηρήσεων (ομάδων) που βρίσκονται πιο κοντά: $G-\{E,F\}$, $C-D$ και $B-C$ και τα τρία με απόσταση 2.00. Επιλέγουμε να ενώσουμε τις ομάδες $\{G\}$ και $\{E,F\}$ για τη δημιουργία της ομάδας $\{E,F,G\}$.

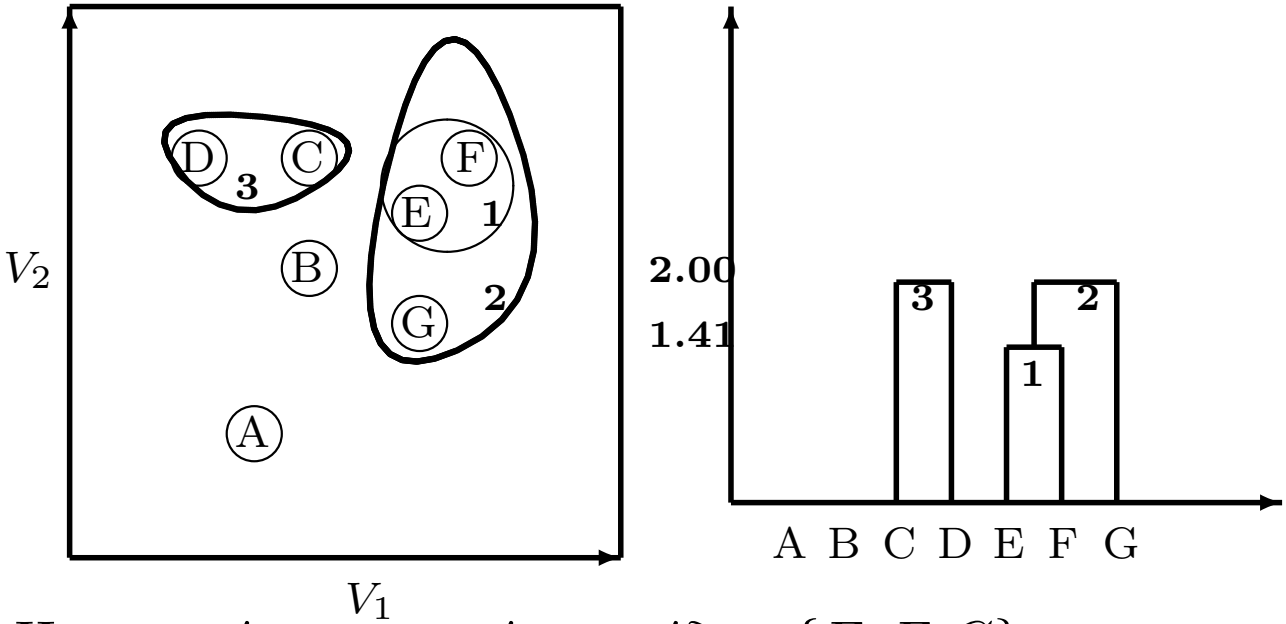


Το μέτρο ομοιογένειας τις δεύτερης ομάδας $\{E,F,G\}$ είναι $\mathrm{ESS}_{\{E,F,G\}}=5.33$.

Step 3

Παρατηρ.	Παρατηρήσεις A	B	C	D	$\{E,F,G\}$
A	0				
B	3.16	0			
C	5.10	2.00	0		
D	5.00	2.83	2.00	0	
$\{E,F,G\}$	3.61	2.24	2.24	4.12	0

Βρίσκουμε το επόμενο ζευγάρι παρατηρήσεων (ομάδων) που βρίσκονται πιο κοντά: $C-D$ και $B-C$ και τα τρία με απόσταση 2.00. Επιλέγουμε να ενώσουμε τις παρατηρήσεις C και D στην ομάδα $\{C,D\}$.



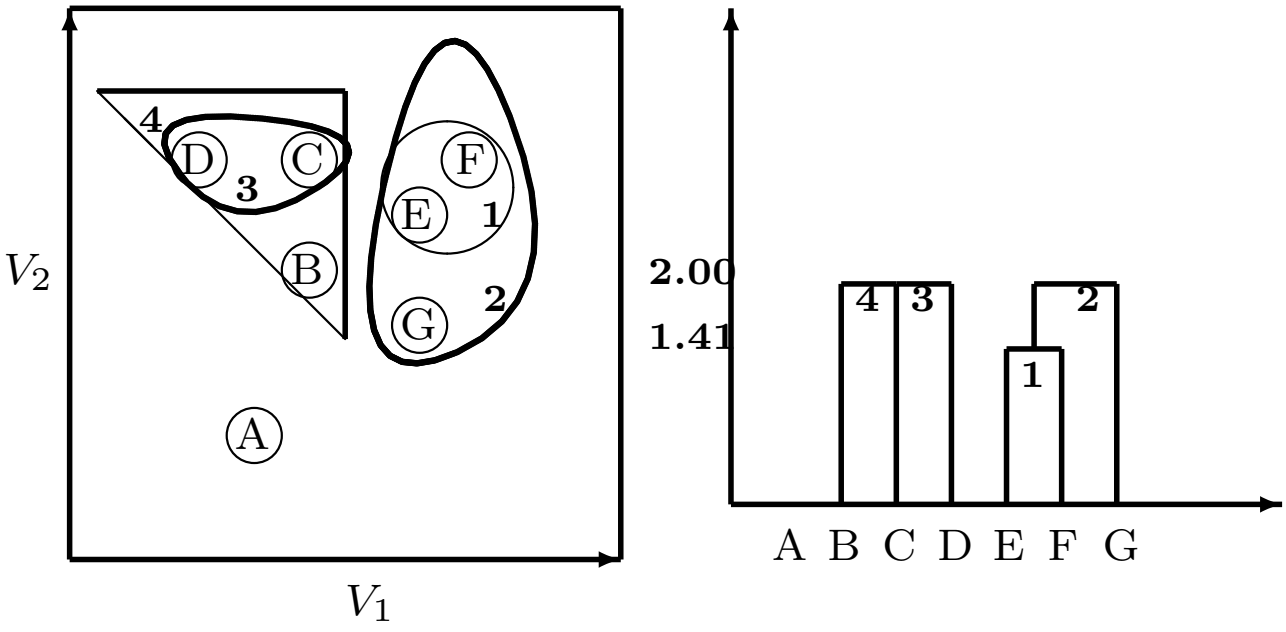
Η ομοιογένεια των νέων ομάδων $\{E,F,G\}$ και $\{C,D\}$ είναι το άθροισμα των ΕΕΣ. Έτσι,

$$\text{ΕΕΣ} = \text{ΕΕΣ}_{\{E,F,G\}} + \text{ΕΕΣ}_{\{C,D\}} = 5.33 + 2 = 7.33.$$

Στάδιο 4

Παρατηρ.	Παρατηρήσεις A	B	$\{C,D\}$	$\{E,F,G\}$
A	0			
B	3.16	0		
$\{C,D\}$	5.00	2.00	0	
$\{E,F,G\}$	3.61	2.24	2.24	0

Βρίσκουμε το επόμενο κοντίτερο ζευγάρι παρατηρήσεων (ομάδων): Ενώνουμε B και $\{C,D\}$ σε μια νέα ομάδα $\{B,C,D\}$.

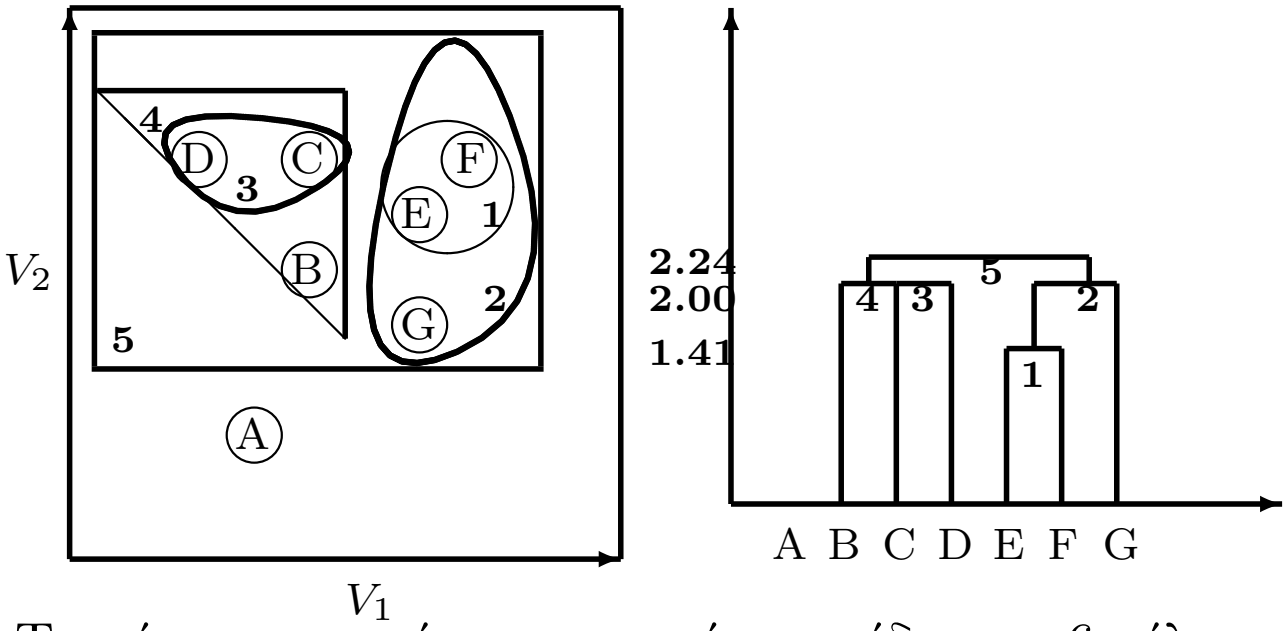


Η ομοιογένεια των νεών ομάδων $\{E,F,G\}$ και $\{B,C,D\}$ δίνεται από EES = $\text{EES}_{\{E,F,G\}} + \text{EES}_{\{B,C,D\}} = 5.33 + 5.33 = 10.67.$

Στάδιο 5

Παρατηρ.	Παρατηρήσεις		
	A	$\{B,C,D\}$	$\{E,F,G\}$
A	0		
$\{B,C,D\}$	3.16	0	
$\{E,F,G\}$	3.61	2.24	0

Η μικρότερη απόσταση είναι 2.24. Έτσι, ενώνουμε τις δύο ομάδες με 3 μέλη $\{B,C,D\}$ και $\{E,F,G\}$ και έχουμε $\{B,C,D,E,F,G\}$.



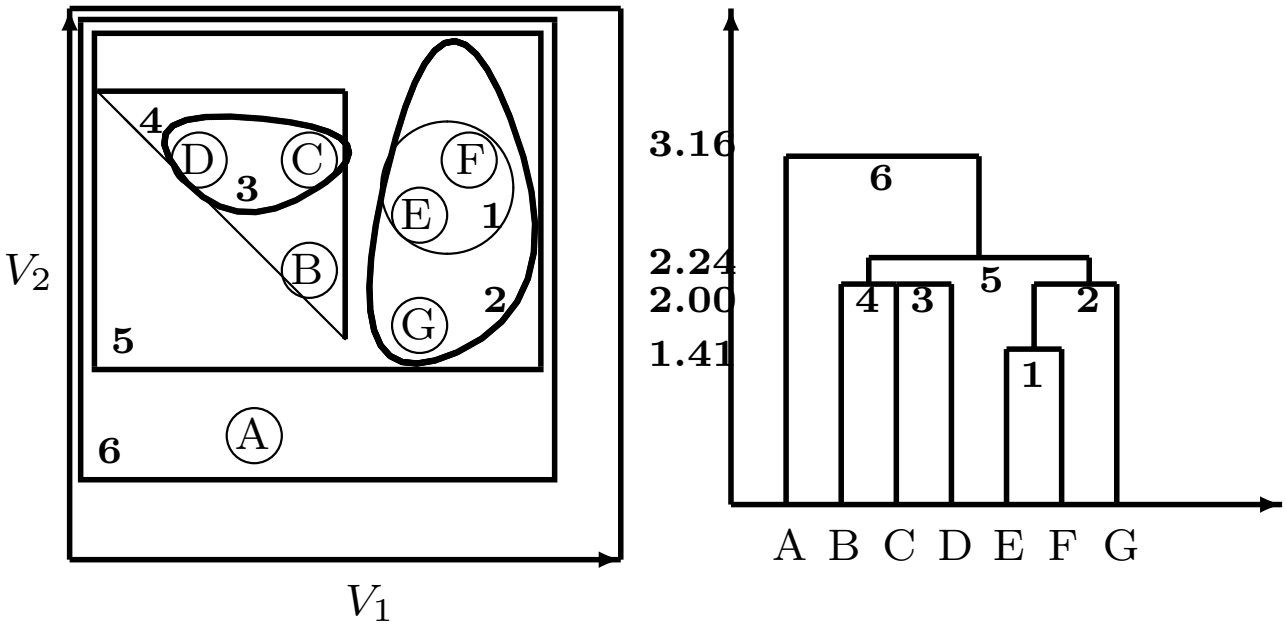
Το μέτρο ομοιογένειας των νέων ομάδων με 6-μέλη $\{B,C,D,E,F,G\}$ δίνεται από:

$$\text{EES} = \text{EES}_{\{B,C,D,E,F,G\}} = 24.83.$$

Στάδιο 6

Παρατηρ.	Παρατηρήσεις A	Παρατηρήσεις $\{B,C,D,E,F,G\}$
A	0	
$\{B,C,D,E,F,G\}$	3.16	0

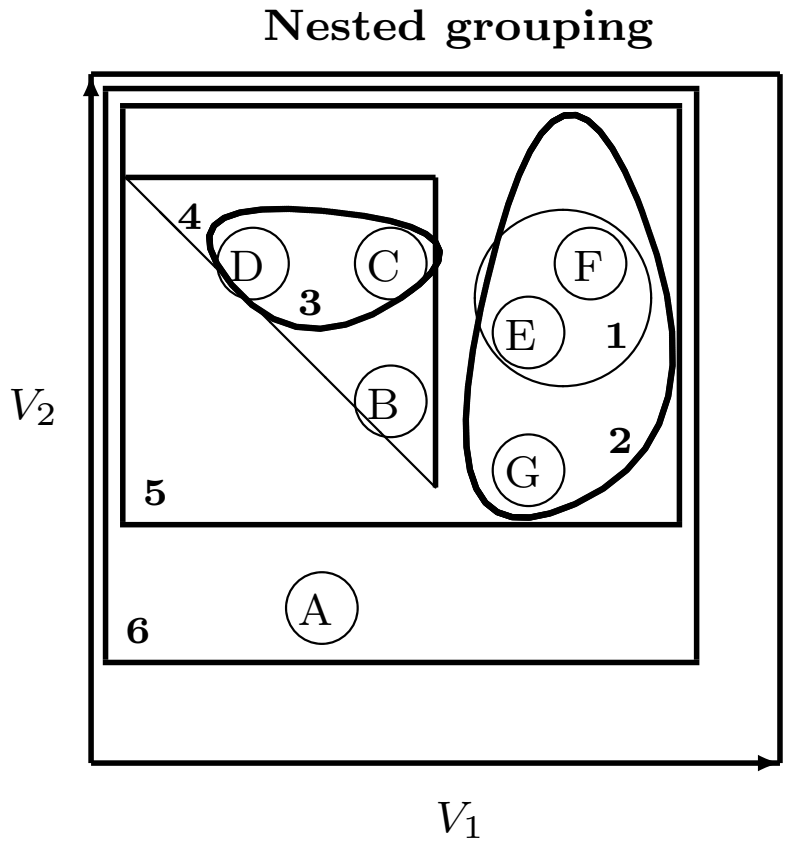
Το τελικό στάδιο ενώνει όλες τις παρατηρήσεις σε μια μόνο ομάδα, δηλαδή: $\{A,B,C,D,E,F,G\}$.

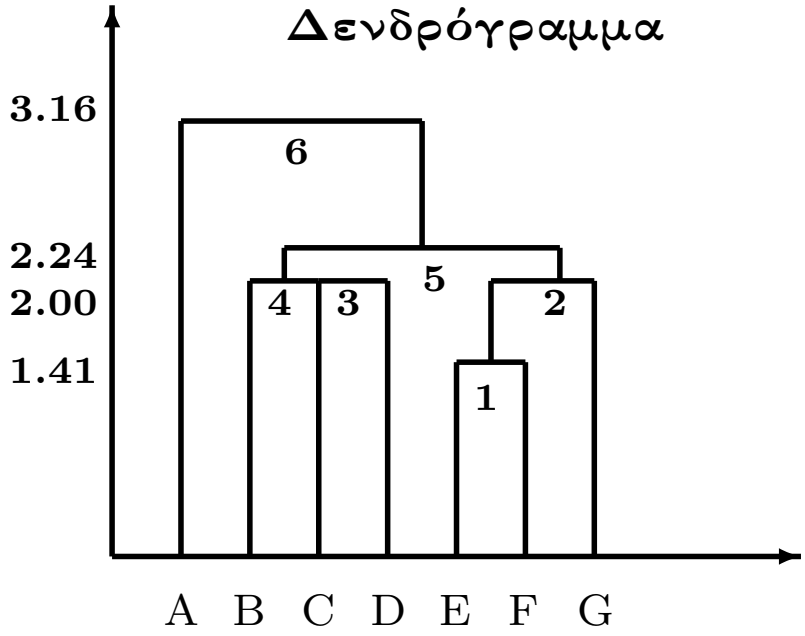


Η ομοιογένεια της τελικής ομάδας 7-μελών $\{A,B,C,D,E,F,G\}$ είναι

$$\text{EES} = \text{EES}_{\{A,B,C,D,E,F,G\}} = 41.43.$$

Nested grouping και Δενδρογράμματα





Συσσωρευτικό Σχέδιο			Λύση Ομαδοποίησης		
Στάδιο	Ελάχ. Απόστ. Μη-Ομαδο. Παρατηρ. *	Ζεύγη Παρατ.	Ομάδα	# Ομάδων	Συνολ. Ομοιογ. (ΕΕΣ)
0	Αρχ. Λύση		$\{A\}\{B\}\{C\}\{D\}$ $\{E\}\{F\}\{G\}$	7	0
1	1.41	$E-F$	$\{A\}\{B\}\{C\}\{D\}$ $\{E,F\}\{G\}$	6	1
2	2.00	$E-G$	$\{A\}\{B\}\{C\}\{D\}$ $\{E,F,G\}$	5	5.33
3	2.00	$C-D$	$\{A\}\{B,C\}\{D\}$ $\{E,F,G\}$	4	7.33
4	2.00	$B-C$	$\{A\}\{B,C,D\}$ $\{E,F,G\}$	3	10.67
5	2.24	$B-E$	$\{A\}\{B,C,D,E,F,G\}$	2	24.83
6	3.16	$A-B$	$\{A,B,C,D,E,F,G\}$	1	41.43

* Ευκλείδεια απόσταση μεταξύ παρατηρήσεων

Σκοπός της ομαδοποίησης είναι η δημιουργία απλής δομής που όμως να δημιουργεί ομοιογενείς ομάδες. Καθώς μειώντεται ο αριθμός των ομάδων, μια αύξηση στη γενική ομοιογένεια υποδειλώνει ότι δύο ομάδες δεν είναι και τόσο όμοιες.

Στο πιο πάνω παράδειγμα το μέτρο συνολικής ομοιογένειας αυξάνεται όταν αρχικά ενώνουμε δύο παρατηρήσεις (Στάδιο 1). Επίσης αυξάνεται όταν δημιουργείται η πρώτη ομάδα 3-μελών (Στάδιο 2). Στα επόμενα δύο στάδια (3 και 4) η συνολική ομοιογένεια δεν μεταβάλλεται σημαντικά. Αυτό υποδειλώνει ότι δημιουργούμε ομάδες με την ίδια ομοιογένεια όπως αυτών που έχει ήδη δημιουργηθεί. Στο Στάδιο 5 όπου ενώνονται οι δύο ομάδες 3-μελών παρατηρούνται μεγάλες αυξήσεις στο ΕΣΣ. Αυτό υποδηλώνει ότι η ένωση των δύο αυτών ομάδων δημιουργεί μια νέα ομάδα λιγότερο ομοιογενή. Συμπερασματικά, η ομαδοποίηση στο Στάδιο 4 είναι καλύτερη από αύτη του Σταδίου 5.

Επιπλέον παρατηρούμε ότι στο Στάδιο 6 το ESS αυξάνεται ελαφρός. Αυτό δείχνει ότι ακόμα και αν η δεδομένη παρατήρηση παραμένει διαχωρισμένη μέχρι το τελευταίο στάδιο, η ενσωμάτωση της επηρεάζει την ομοιογένεια.

Έτσι, η γενική εικόνα υποδηλώνει ότι η προτεινόμενη λύση 3-ομάδων στο στάδιο 4 είναι η καταλληλότερη: Δημιουργούνται δύο ομάδες ίσου μεγέθους και μία ακόμη παρατήρηση.

- Η εξαγωγή συμπερασμάτων εξαρτάται σε μεγάλο βαθμό από την κρήση του κάθε ερευνητή.
- Η τελική απόφαση και την επιλογή του αριθμού των ομάδων εξαρτάται από τον εκάστοτε ερευνητή.
- Στις περισσότερες ερευνητικές μελέτες μάρκετιγκ χρησιμοποιούνται περισσότερες από μία μεταβλητές και πολύ περισσότερες παρατηρήσεις. Αυτό αναγάγει την πολυπλοκότητα του προβλήματος ομαδοποίησης.
- Δίνονται τα διανύσματα $X = (X_1, X_2, \ldots, X_n)$ και $Y = (Y_1, Y_2, \ldots, Y_n)$, τότε η (Ευκλείδια) απόσταση μεταξύ των X και Y δίνεται από:

$$\sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2 + \cdots + (X_n - Y_n)^2}.$$

Π.χ. θεωρούμε $X = \begin{pmatrix} 9 \\ 3 \\ 1 \end{pmatrix}$ και $Y = \begin{pmatrix} 10 \\ 2 \\ 9 \end{pmatrix}$.

Η Ευκλείδεια απόσταση δίνεται από:

$$\sqrt{(9-10)^2 + (3-2)^2 + (1-9)^2} = \sqrt{66} = 8.12.$$

Παράδ.: μισθός και αριθμός αυτοκινήτων

Η συσχέτηση του μισθού και του αριθμού των αυτοκινήτων που κατέχουν τρεις ιδιοκτήτες A, B και C δίνεται από:

	Αρχ. Παρατ.		Προς. μέσου Παρατ.		Τυποπ. Παρατ.	
Ιδιο.	Μισθ. $	αυτο.	$	αυτο.	$	αυτο.
A	20000	1	0	-1/3	0	$-1/\sqrt{3}$
B	30000	2	10000	2/3	1	$2/\sqrt{3}$
C	10000	1	-10000	-1/3	-1	$-1/\sqrt{3}$
Μέσος	10000	4/3	0	0	0	0
(SD)	10000	$1/\sqrt{3}$	10000	$1/\sqrt{3}$	1	1

	Αρχ.		Προς. μέσου		Τυποπ.	
Ιδιο.	Από.	Κατ.	Από.	Κατ.	Από.	Κατ.
A-B	10000	1	10000	1	$\sqrt{4}=2$	2
A-C	10000	1	10000	1	1	1
B-C	20000	2	20000	2	$\sqrt{7}=2.65$	3

Σημειώστε ότι για τα αρχικά και τα τυποποιημένα δεδομένα, οι απόστασεις μεταξύ A και B δίνονται, αντίστοιχα, από:

$$d(A,B)=\sqrt{(20000-30000)^2+(1-2)^2}=10000.00005$$
$$=10000$$

και $$d(A,B)=\sqrt{(0+1)^2+(\tfrac{-1-2}{3})^2}=\sqrt{4}=2.$$

Τυποποίηση δεδομένων

- Εάν οι μεταβλητές μετριούνται με πολύ διαφορετικές μονάδες, τότε η ομαδοποίηση θα επιρεαστεί από τις μονάδες μέτρησεις. Σε αυτή την περίπτωση, πριν την ομαδοποίησηση, είναι απαραίτητη η τυποποίηση των παρατηρήσεων κάθε μεταβλητής. Στην τυποποίηση κάθε μεταβλητή θα έχει μέσο μηδέν και τυπική απόκλιση ένα.

Παράδειγμα

Υποθέστε ότι τρείς καταναλωτές A, B και C συγκρίνονται βάση δύο μεταβλητών: Πιθανότητα αγοράς μάρκας X (σε ποσοστό) και Χρόνος που ξοδεύεται στην παρακολούθηση διαφημίσεων αυτού του λογότυπου X (σε λεπτά ή δευτερόλεπτα). Δίνονται τα ακόλουθα δεδομένα:

	Χρόνος διαφημίσεων		
	Αγορά		
Ιδιοκτ.	Πιθανότητα (%)	(λεπτά)	(δευτερόλ.)
A	60	3.0	180
B	65	3.5	210
C	63	4.0	240

Χρησιμοποιώντας τις πιο πάνω πληροφορίες υπολογίζουμε τις αποστάσεις. Επίσης υπολογίζονται το τετράγωνο (απόλυτη) της Ευκλείδειας απόστασης και η Ευκλείδεια απόσταση. Οι μικρότερες αποστάσεις υποδειλώνουν μεγαλύτερη ομοιογένεια. Αυτές δίνονται πιο κάτω:

	Απόσταση σε **λεπτά** παρακολούθησεις			
	Απλη Ευκλ. Από,		Τετρ. και Απόλυτη Ευκλ, Από.	
Ιδιοκτ.	Τιμή	Κατ.	Τιμή	Κατ.
A-B	5.025	3	25.25	3
A-C	3.162	2	10.00	2
B-C	2.062	1	4.25	1

Οι πιο όμοιες παρατηρήσεις (αυτές με τη μικρότερη απόσταση) είναι B και C, και ακολούθως είναι A και C. Το A και B είναι τα λιγότερο όμοια στοιχεία. Βάση των δύο μεθόδων μέτρησης της απόστασης η κατάταξη είναι η ίδια. Όμως η σχετική ομοιογένεια, ή αλλιώς απόκλιση, μεταξύ των στοιχείων είναι πιο εμφανείς βάση του τετραγώνου της Ευκλείδειας απόστασης.

Η κατάταξη ομοιογένειας μπορεί να αλλάξη άρδην όταν αλλάξουν οι μονάδες μέτρησεις μιας μεταβλητής. Για παράδειγμα, αν μετρήσουμε το χρόνο παρακολούθησης διαφημίσεων σε δευτερόλεπτα αντί για λεπτά, τότε παρατηρούνται διαφορετικές κατατάξεις:

	Απόσταση σε **λεπτά** παρακολούθησεις			
	Απλή Ευκλ. Απόστ.		Τετρ. και Απόλυτη Ευκλ. Απόστ.	
Ιδιοκτ.	Τιμή	Κατ.	Τιμή	Κατ.
A-B	30.41	2	925	2
A-C	60.07	3	3609	3
B-C	30.06	1	904	1

Η κατάταξη ομοιογένειας έχει αλλάξει δραματικά. Αν και το B και το C είναι τα πιο ομοιογενή στοιχεία, το ζευγάρι A-B είναι το ακολούθως πιο ομοιο και έχει σχεδόν την ίδια ομοιογένεια με το ζευγάρι B-C.

Οι πιο πάνω διαφορές στην κατάταξη οφείλονται στο ότι η μονάδα μέτρησεις του χρόνου παρακολούθησης επηρέασε σε μεγάλο βαθμό τους υπολογισμούς. Αυτό κατέστησε την άλλη μεταβλητή (πιθανότητα αγοράς) λιγότερο σημαντική παράμετρο για την κατάταξη.

Τώρα θεωρούμε ότι τα δεδομένα τυποποιούνται. Για το σκοπό αυτό αφαιρούμε το μέσο και διαιρούμε με την τυπική απόκλιση[a]. Σημειώστε ότι ο μέσος και η τυπική απόκλιση της πιθανότητα αγοράς δίνονται αντίστοιχα από $(0.60 + 0.65 + 0.63)/3 = 0.627$ και 0.025. Παρομοίως, ο μέσος και η τυπική απόκλιση του χρόνου παρακολούθησης σε λεπτά (δευτερόλεπτα) δίνονται από 3.5 (210) και 0.5 (30), αντίστοιχα.

	Τυποπ. Τιμές		Απλή Ευκλ.		Τετρ. Ευκελ.	
Στοιχ.	Αγορά	Min/Sec	Από.		Από.	
Ζεύγ.	Πιθ.	Χρόνου Παρακολ.	Τιμές	Κατ.	Τιμή	Κατ.
A-B	-1.06	-1.0	2.22	2	4.95	2
A-C	0.93	0.0	2.33	3	5.42	3
B-C	0.13	1.0	1.28	1	1.63	1

Η κατάταξη για τις δύο μετρήσεις είναι η ίδια, όμως η σχετική ομοιογένεια ή διασπορά, μεταξύ των στοιχείων είναι πιο έκδηλη για το τετράγωνο της Ευκλείδειας απόστασης.

[a] $\mathrm{SD} = \sqrt{(x_i - \bar{x})^2/(n-1)}$

Ιεραρχική διαδικασία: συσσωρευτική μέθοδος

- Οι Ιεραρχικές διαδικασίες περιλαμβάνουν τη δημιουργία μιας δενδροειδούς διάταξης. Υπάρχουν δύο διαδικασίες ιεραρχικής ομαδοποίησης: οι συσσωρευτικοί (the agglomerative)και οι διαχωριστικοί μέθοδοι (divisive methods).

- Η συσσωρευτική μέθοδος ξεκινά ορίζοντας κάθε μία παρατήρηση ως μια ομάδα. Στα επόμενα στάδια, οι δύο κοντινότερες ομάδες (ή παρατηρήσεις) δημιουργούν μια νέα ομάδα. Σε κάποιες περιπτώσεις μία παρατήρηση μπορεί να ομαδοποιηθεί με την αρχική ομάδα. Σε κάποιες άλλες περιπτώσεις δύο ομάδες που δημιουργηθήκαν σε προηγούμενα στάδια μπορούν να είναι ομάδα. Τελικά όλες οι παρατηρήσεις ομαδοποιούνται σε μια μεγάλη ομάδα. Λόγω αυτής της διαδικασίας οι μέθοδοι ονομάζονται συσσωρευτικές.

- Ένα βασικό χαρακτηριστικό των ιεραρχικών διαδικασιών είναι ότι η ομαδοποίηση σε ένα στάδιο σχετίζεται με την ομαδοποίηση του προηγούμενου σταδίου. Είναι το λεγόμενο δένδρο ομοιογένειας. Π.χ. μια λύση έξι ομάδων προέρχεται από την ενωποίηση δύο ομάδων του

σταδίου των επτά ομάδων. Έτσι, μπορούμε εύκολα να ακολουθήσουμε την πορεία ομαδοποίησης που ακολουθεί μια παρατήρηση από την αρχή της διαδικασίας μέχρι το τέλος.

Αυτή η διαδικασία ονομάζεται **Δενδρόγραμμα**.

- Η διαδικασία που ακολουθείται στη **διαχωριστική μέθοδο** είναι εκ διαμέτρου αντίθετη από αυτή που ακολουθείται στη συσσωρευτική μέθοδο.

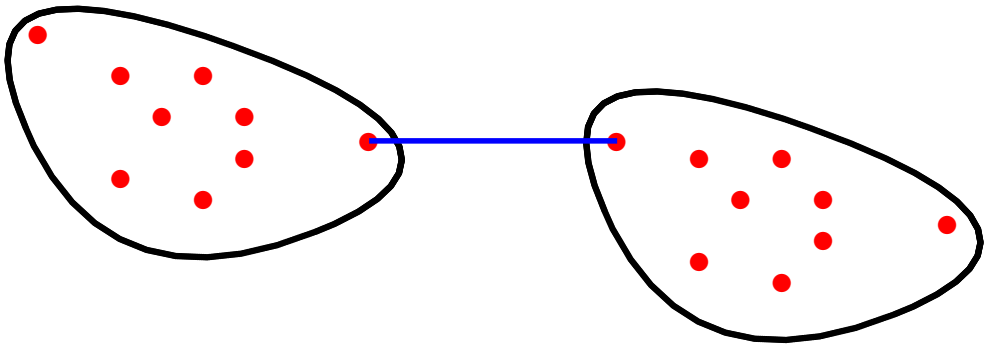
Μέθοδοι δημιουργίας ομάδων

Υπάρχουν πέντε μέθοδοι για τη δημιουργία ομάδων στη συσσωρευτική μέθοδο. Οι μέθοδοι διαφέρουν στον τρόπο υπολογισμού της απόστασης μεταξύ των ομάδων:

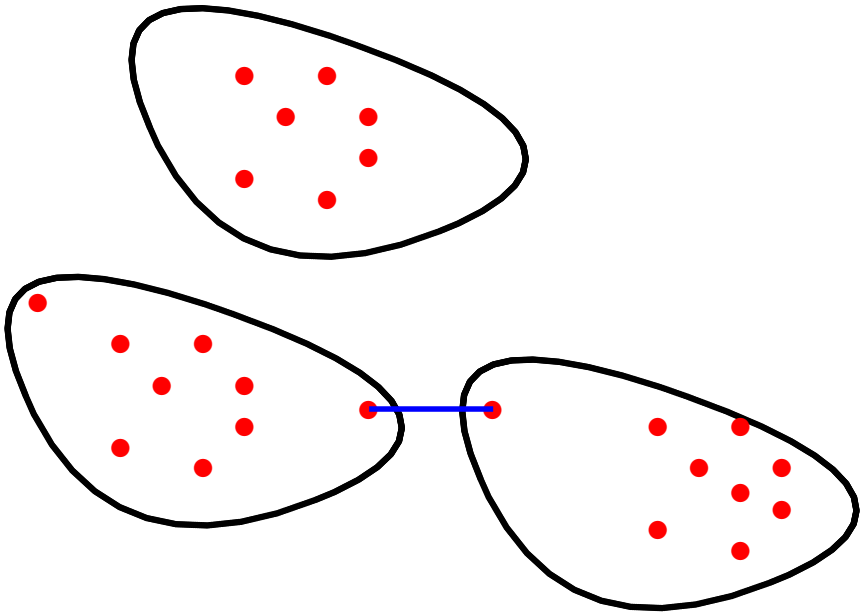
- Απλή Σύνδεση (Single Linkage)

Η απλή σύνδεση στηρίζεται στην ελάχιστη απόσταση. Εντοπίζει το ζεύγος με τη μικρότερη απόσταση και τα κατατάσσει στην πρώτη ομάδα. Στη συνέχεια εντοπίζεται το επόμενο κοντίτερο ζεύγος και είτε ενσωματώνεται στην αρχική ομάδα, είτε δημιουργείται μια δεύτερη ομάδα. Η διαδικασία συνεχίζεται μέχρις ότου όλα τα στοιχεία είναι σε μια ομάδα.

Η απόσταση μεταξύ δύο οποιονδήποτε ομάδων είναι η μικρότερη απόσταση από ένα σημείο μιας ομάδας σε ένα οποιοδήποτε σημείο μιας άλλης ομάδας. Εάν δύο ομάδες ικανοποιούν την πιο πάνω αρχή μπορούν να ενωθούν σε οποιοδήποτε στάδιο της διαδικασίας.



Στην περίπτωση που δύο ομάδες δεν εχουν οριοθετηθεί σωστά, πιθανό να δημιουργηθούν προβλήματα με τη μέθοδο *Απλής Σύνδεσης*. Σε τέτοιες περιπτώσεις δημιουργούνται ατέλευτες συνδέσεις μεταξύ των ομάδων.



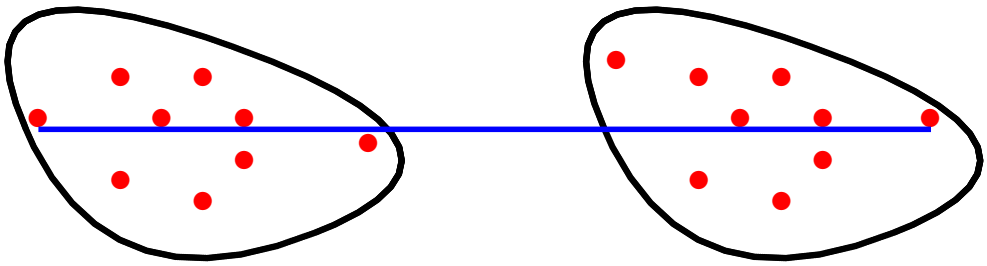
- Πλήρης Σύνδεση (Complete Linkage)

  Η μέθοδος πλήρους σύνδεσης είναι παρόμοια με τη μέθοδο απλού συνδέσμου με τη διαφορά ότι το κριτήριο ομαδοποίησης βασίζεται στη μέγιστη απόσταση. Και στις δύο περιπτώσεις δύο ομάδες συγχωνεύονται για να δημιουργήσουν μια ομάδα όταν

η απόσταση αυτή, όπως και να ορίζεται είναι ελάχιστη. Η μέθοδος ονομάζεται έτσι γιατί η απόσταση δύο ομάδων είναι η μέγιστη (ελάχιστη ομοιογένεια) από τις αποστάσεις μεταξύ όλων των ζευγών στοιχείων από τις δύο ομάδες. Η μέθοδος αυτή δημιουργεί καλύτερες, πιο συμπαγείς ομάδες. Αντίθετα η μέθοδος απλού συνδέσμου έχει την τάση να δημιουργεί σκόρπιες, επιμήκης ομάδες.

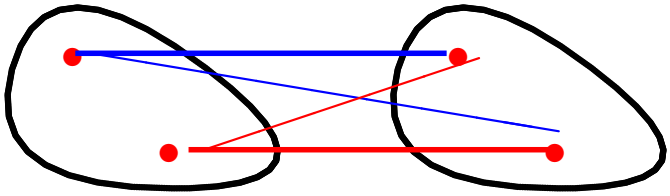
Στην περίπτωση της ελάχιστης απόστασης μόνο τα *κοντίτερα* (πιο ομοιογενή) ζευγάρια ενώνονται. Αλλά στην πλήρη σύνδεση τα πιο ακραία (λιγότερο όμοια) ζευγάρια ενώνονται.

Έτσι δίνεται πιο πολλή βαρύτητα στα πιο απόμακρα σημεία.



- Μέσος Σύνδεσμος (Average Linkage)

  Στη μέθοδο μέσου συνδέσμου το κριτήριο ομαδοποίησης είναι η μέση απόσταση μεταξύ όλων των ζευγών παρατηρήσεων, όπου ένα μέλος ζεύγους προέρχεται από καθεμιά από τις ομάδες. Έτσι, το κριτήριο ομαδοποίησης στηρίζεται σε όλα τα μέλη των ομάδων και όχι σε ένα ζεύγος ακραίων τιμών. Η μέθοδος αυτή θεωρείται πολύ δημοφιλής αν και υπολογιστικά είναι πιο απαιτητική.



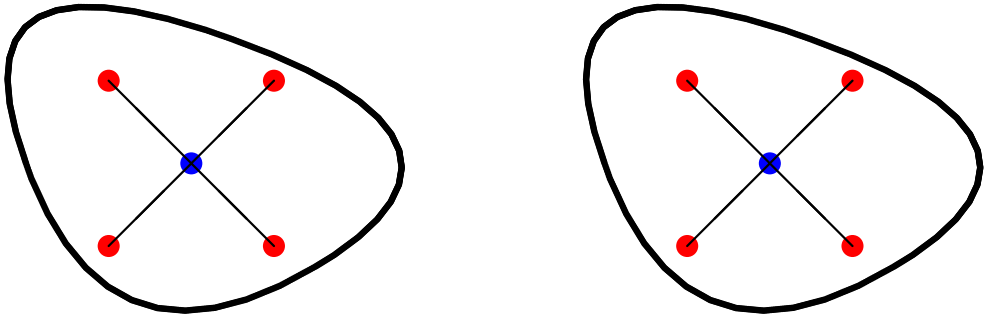
- Μέθοδος Ward's

  Η μέθοδος αυτή κατατάσσεται στις μεθόδους διασποράς (variance methods) όπου σκοπός είναι η δημιουργία ομάδων που να ελαχιστοποιούν τη διακύμανση μέσα σε μια ομάδα. Η μέθοδος Ward's είναι μια ευρέως διαδεδομένη μέθοδος διασποράς.

Αρχικά, για κάθε ομάδα, υπολογίζεται ο μέσος όλων των μεταβλητών.

Στη συνέχεια, για κάθε στοιχείο, υπολογίζεται η Ευκλείδεια απόσταση από το κέντρο (μέσο) της ομάδας. Οι αποστάσεις αθροίζονται για όλα τα στοιχεία.

Σε κάθε στάδιο ομαδοποιούνται οι δύο ομάδες με τη μικρότερη αύξηση στο άθροισμα τετραγωνικού λάθους εσωτερικά της ομάδας.



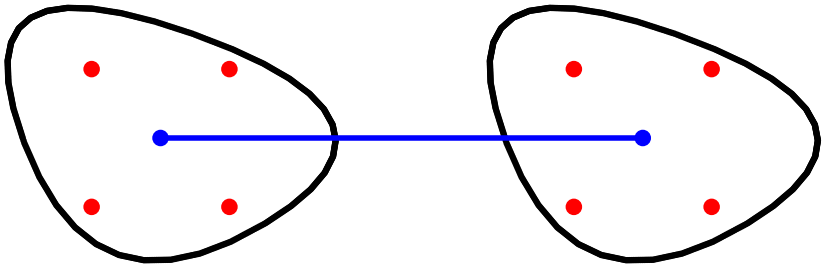
Η μέθοδος αυτή τίνει να ενώνει ομάδες με μικρό αριθμό παρατηρήσεων. Επίσης μεροληπτεί προς τη δημιουργία ομάδων με τον ίδιο αριθμό παρατηρήσεων.

- Κεντροειδής μέθοδος (Centroid method)

  Στην κεντροειδή μέθοδο απόσταση μεταξύ δύο ομάδων θεωρείται η απόσταση μεταξύ των κέντρων τους. Το κέντρο μιας ομάδας ορίζεται ως ο μέσος όρος των παρατηρήσεων μια μεταβλητής σε μια ομάδα.

  Κάθε φορά που συντάσσεται μια νέα ομάδα το κέντρο επαναυπολογίζεται.

  Πλεονέκτημα αυτής της μεθόδου θεωρείται το γεγονός ότι επηρεάζεται λιγότερο από ακραίες παρατηρήσεις.



- Από τις ιεραρχικές μεθόδους, η μέθοδοι μέσου συνδέσμου και η Ward's αποδίδουν καλύτερα από τις άλλες μεθόδους.

Μέθοδος Ward's

Η μέθοδος Ward's είναι μια ιεραρχική διαδικασία ομαδοποίησης που βασίζεται στην εσωτερική διακύμανση μια ομάδας, παρά στο σύνδεσμο της.

Υποθέστε ένα δείγμα n παρατηρήσεων που καταμερίζεται σε g ομάδες, η ith ομάδα περιέχει n_i παρατηρήσεις με μέσο $\bar{x}_i$.

Το άθροισμα του τετραγώνου των αποκλίσεων μέσα σε μια ομάδα (within group sum of squared deviations) για αυτό το διαχωρισμό σε $g-$ομάδες είναι:

$$W = \sum_{j=1}^{g} \sum_{i=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \ ,$$

όπου x_{ij} είναι η jth παρατήρηση της ομάδας i.

Η τιμή του W μπορεί να υπολογιστεί για οποιαδήποτε κατηγοριοποίηση.

Ο τύπος W/n δίνει τη συγκεντρωτική διακύμανση της κατάταξης.

Τα στάδια που ακολουθούνται για την ομαδοποίηση Ward είναι τα ακόλουθα:

1. Ξεκινούμε με n ομάδες, όπου κάθε παρατήρηση αποτελεί και μία ομάδα. Στο στάδιο αυτό $W = 0$.

2. Σε κάθε στάδιο μειώνουμε τον αριθμό των ομάδων κατά ένα μέχρις ότου να καταλήξουμε στις δύο αυτές ομάδες που η ένωση τους θα μας δώσει την ελάξιστη αύξηση στο W.

3. Συνεχίζουμε την διαδικασία για $n - 1$ στάδια (συγχώνευσεις) μέχρι να παραμείνει μόνο μια ομάδα.

Παράδειγμα

Θεωρείστε την περίπτωση με μία μεταβλητή. Δίνεται ένα δείγμα 6 παρατηρήσεων με τιμές $1, 2, 5, 7, 9, 10$. Η μέθοδος Ward ομαδοποιεί τις παρατηρήσεις με την εξής σειρά:

Στάδιο	# Ομάδων	Ομάδες	W
0	6	$(1),(2),(5),(7),(9),(10)$	0.00
1	5	$(1,2),(5),(7),(9),(10)$	0.50
2	4	$(1,2),(5),(7),(9,10)$	1.00
3	3	$(1,2),(5,7),(9,10)$	3.00
4	2	$(1,2),(5,7,9,10)$	15.25
5	1	$(1,2,5,7,9,10)$	67.33

Οι συνδιασμοί στο σταδιο 1:

ζεύγος	κέντρο	W	ζεύγος	κέντρο	W
$(1,2)$	1.5	0.5	$(5,7)$	6.0	2.0
$(1,5)$	3.0	8.0	$(5,9)$	7.0	8.0
$(1,7)$	4.0	18.0	$(5,10)$	7.5	12.5
$(1,9)$	5.0	32.0	$(7,9)$	8.0	2.0
$(1,10)$	5.5	40.5	$(7,10)$	8.5	4.5
$(2,5)$	3.5	4.5	$(9,10)$	9.5	0.5
$(2,7)$	4.5	12.5			
$(2,9)$	5.5	24.5			
$(2,10)$	6.0	32.0			

Πιθανοί συνδ. 4-ομάδων	W
$(1,2,5),(7),(9),(10)$	8.67
$(1,2,7),(5),(9),(10)$	
$(1,2,9),(5),(7),(10)$	
$(1,2,10),(5),(7),(9)$	
$(1,2),(5,7),(9),(10)$	
$(1,2),(5,9),(7),(10)$	
$(1,2),(5,10),(7),(9)$	
$(1,2),(5)(7,9),(10)$	
$(1,2),(5)(7,10),(9)$	$0.5+0+(7-8.5)^2+(10-8.5)^2=5$
$(1,2),(5)(7),(9,10)$	$0.5+0+0+0.5=1$

Πιθανοί συνδ. 3-ομάδων	W
$(1,2,5),(7),(9,10)$	$8.67+0+0.5=9.17$
$(1,2,7),(5),(9,10)$	
$(1,2,9,10),(5),(7)$	
$(1,2),(5,7),(9,10)$	$0.5+2.0+0.5=3.0$
$(1,2),(5,9,10),(7)$	
$(1,2),(5),(7,9,10)$	5.17

Πιθανοί συνδ. 2-ομάδων	W
$(1,2,5,7),(9,10)$	15.25
$(1,2,9,10),(5,7)$	
$(1,2),(5,7,9,10)$	

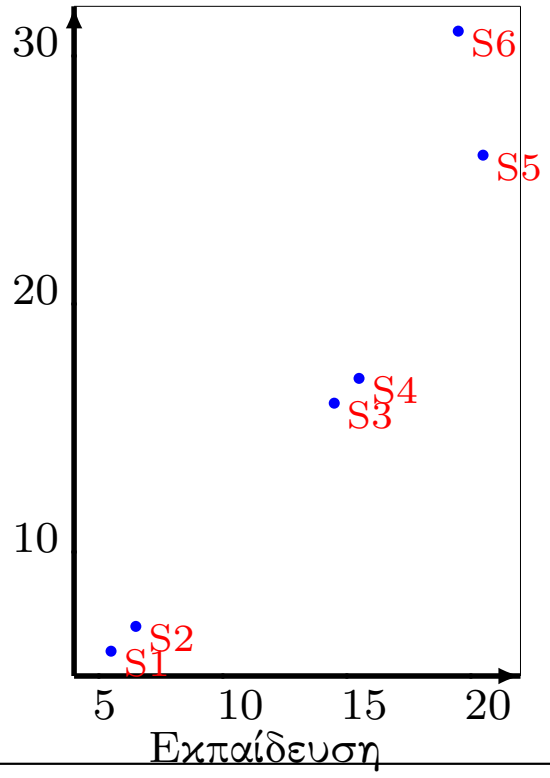
Πιθανοί συνδ. 1-ομάδας	W
$(1,2,5,7,9,10)$	67.33

Ποιο είναι το ESS για την περίπτωση των δύο ομάδων $\{(1,2,5),(7,9,10)\}$?

Παράδ.: Κεντροειδής και Ward's μέθοδοι

Θεωρείστε τα ακόλουθα υποθετικά δεδομένα. Περιλαμβάνουν τον αριθμό αναγνώρισης ταυτότητας συστήματος (sid), το εισόδημα ($χιλιάδες) και η εκπαίδευση (σε χρόνια).

Sid	Εισόδημα	Εκπαίδευση
S1	5	5
S2	6	6
S3	15	14
S4	16	15
S5	25	20
S6	30	19



Κεντροειδής

Ο πίνακας ομοιογένειας δίνεται ως ακολούθως:

	S1	S2	S3	S4	S5	S6
S1	0					
S2	2	0				
S3	181	145	0			
S4	221	181	2	0		
S5	625	557	136	106	0	
S6	821	745	250	212	26	0

Για τον υπολογισμό την Ευκλείδειας απόστασης:
Π.χ.

$d^2(S1, S2) = (5-6)^2 + (5-6)^2 = 2$
$d^2(S2, S4) = (6-16)^2 + (6-15)^2 = 181$

a. Πέντε Ομάδες

Η πρώτη ομάδα δημιουργείται με τη σύμπτυξη των στοιχείων S1 και S2. Το κέντρο της πρώτης ομάδας είναι το κέντρο των στοιχείων S1 και S2. Η νέα ομάδα έχει μέση εκπαίδευση 5.5 χρόνων (=(5+6)/2) και μέσο εισόδημα 5.5 χιλιάδες δολάρια.

Ο ακόλουθος πίνακας παρουσιάζει τις πέντε ομάδες που έχουν δημιουργηθεί και τον αντίστοιχο πίνακα ομοιότητας: (χρησιμοποιώντας το τετράγωνο της Ευκλείδεια απόστασης).

	{S1,S2}	S3	S4	S5	S6
{S1,S2}	0				
S3	162.50	0			
S4	200.50	2	0		
S5	590.50	135.96	106	0	
S6	782.50	250	212	26	0

E.g.
$d^2(\{S1, S2\}, S5) = (5.5-25)^2+(5.5-20)^2 = 590.50.$

b. Τέσσερις ομάδες

Τα στοιχεία S3 και S4 βρίσκονται κοντίτερα και είναι τα πιο όμοια. Έτσι τα ομαδοποιούμε. Η ομάδα αντιπροσωπεύεται από το κέντρο των στοιχείων. δηλαδή΄ $(15.5, 14.5)$.

	{S1,S2}	{S3,S4}	S5	S6
{S1,S2}	0			
{S3,S4}	181	0		
S5	590.50	120.50	0	
S6	782.50	230.50	26	0

E.g. $d^2(\{S3, S4\}, S5) =$
$(15.5 - 25)^2 + (14.5 - 20.0)^2 = 120.5.$

c. Τρεις ομάδες

Τα στοιχεία S5 και S6 έχουν την μικρότερη απόσταση και έτσι ενώνονται για τη δημιουργία της 3ης ομάδας. Το κέντρο του νέου συνόλου είναι $(27.5, 19.5)$.

	{S1,S2}	{S3,S4}	{S5,S6}
{S1,S2}	0		
{S3,S4}	181	0	
{S5,S6}	680	169	0

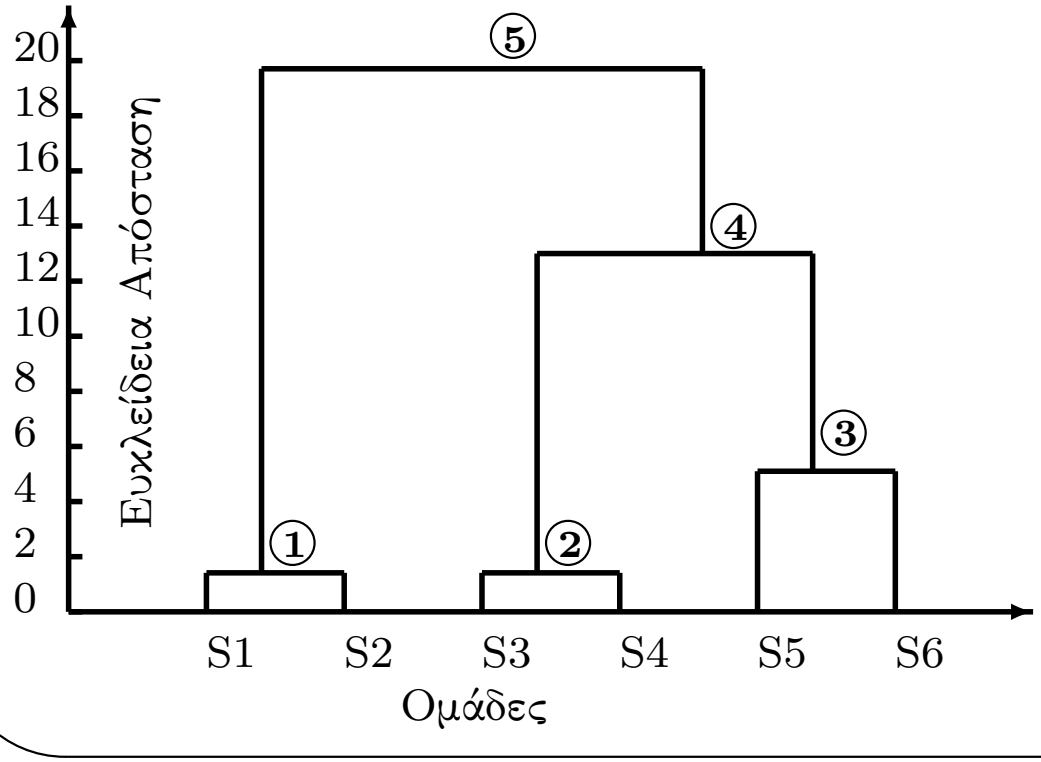
d. Δύο ομάδες

{S3, S4} και {S5,S6} έχουν τη μικρότερη απόσταση και ενώνονται για να δημιουργήσουν μια νέα ομάδα με κέντρο $(21.5, 17)$.

Το Τετράγωνο της Ευκλείδειας απόστασης μεταξύ των ομάδων {S3, S4, S5,S6} και {S1,S2} είναι 388.25.

e. Μία ομάδα

Δημιουργείται με την ένωση όλων των στοιχείων σε μια ομάδα.



Μέθοδος Ward's

Για τη μέθοδο Ward's στην ουσία δεν υπολογίζουμε αποστάσεις μεταξύ των ομάδων. Ανταυτού δημιουργούμε ομάδες βελτιστοποιώντας της ομοιογένειας εσωτερικά της ομάδας. Το άθροισμα τετραγώνων μέσα στην ομάδα χρησιμοιείται σαν μέτρο ομοιογένειας (αυτό είναι γνωστό ως το άθροισμα τετραγωνικού λάθους ESS).

a. Πέντε ομάδες

Αρχικά κάθε παρατήρηση είναι και μία ομάδα. Έτσι το ESS είναι 0. Για τη δημιουργία των 5 ομάδων χρειαζόμαστε 1 ομάδα με δύο στοιχεία και άλλες 4 ομάδες με 1 στοιχείο.

Π.χ. μπορούμε να έχουμε μία ομάδα που να αποτελείται από τα στοιχεία S1 και S2.
$\text{ESS} = (5-5.5)^2+(5-5.5)^2+(6-5.5)^2+(6-5.5)^2 = 1$

Π.χ. μπορούμε να έχουμε μία ομάδα που να αποτελείται από τα στοιχεία S2 και S5.
$\text{ESS} = (6-15.5)^2 + (6-13)^2 + (25-15.5)^2 + (20-13)^2 = 278.5.$

Ο πίνακας δίνει όλες τις πιθανές λύσεις 5 ομάδων και τα αντίστοιχα ESS. Βασισμένοι στο κριτήριο της ελαχιστοποίηση του ESS επιλέγουμε μια από τις λύσεις 1 ή 10. Η επιλογή αυτή γίνεται τυχαία. Ας υποθέσουμε ότι επιλέγουμε τη λύση 1.

b. Τέσσερις ομάδες

Π.χ. η ομάδα αποτελείται από τα στοιχεία S1, S2 και S6:

$$\begin{aligned}\text{ESS} &= (5-13.66)^2 + (5-10)^2 + (6-13.66)^2 \\ &\quad + (6-10)^2 + (30-13.66)^2 + (19-10)^2 \\ &= 522.66\end{aligned}$$

Η λύση ομαδοποίησης 5 είναι αυτή που ελαχιστοποιεί το ESS.

Table 7.6 **Ward's Method**

Cluster Solution	Members in Cluster 1	2	3	4	5	ESS
(a) All Possible Five-Cluster Solutions						
1	S1,S2	S3	S4	S5	S6	1.0
2	S1,S3	S2	S4	S5	S6	90.5
3	S1,S4	S2	S3	S5	S6	110.5
4	S1,S5	S2	S3	S4	S6	312.5
5	S1,S6	S2	S3	S4	S5	410.5
6	S2,S3	S1	S4	S5	S6	72.5
7	S2,S4	S1	S3	S5	S6	90.5
8	S2,S5	S1	S3	S4	S6	278.5
9	S2,S6	S1	S3	S4	S5	372.5
10	S3,S4	S1	S2	S5	S6	1.0
11	S3,S5	S1	S2	S4	S6	68.0
12	S3,S6	S1	S2	S4	S5	125.0
13	S4,S5	S1	S2	S3	S6	53.0
14	S4,S6	S1	S2	S3	S5	106.0
15	S5,S6	S1	S2	S3	S4	13.0
(b) All Possible Four-Cluster Solutions						
1	S1,S2,S3	S4	S5	S6		109.333
2	S1,S2,S4	S3	S5	S6		134.667
3	S1,S2,S5	S3	S4	S6		394.667
4	S1,S2,S6	S3	S4	S5		522.667
5	S1,S2	S3,S4	S5	S6		2.000
6	S1,S2	S3,S5	S4	S6		69.000
7	S1,S2	S3,S6	S4	S5		126.000
8	S1,S2	S4,S5	S3	S6		54.000
9	S1,S2	S4,S6	S3	S5		107.000
10	S1,S2	S5,S6	S3	S4		14.000

Συσσωρευτικό σχέδιο

Το αποτελέσματα των ιεραρχικών μεθόδων συνοψίζονται στο συσσορευτικό σχέδιο. Θεωρείστε τα ακόλουθα 5 στοιχεία που ομαδοποιούνται με τη μέθοδο της πλήρους συσχέτισης. Ο πίνακας δίνει το βαθμό ανομοιογένειας:

	Στ. 1	Στ. 2	Στ. 3	Στ. 4	Στ. 5
Στοιχείο 1	0				
Στοιχείο 2	1.0	0			
Στοιχείο 3	2.0	3.0	0		
Στοιχείο 4	8.0	9.0	10.0	0	
Στοιχείο 5	11.0	12.0	13.0	5.0	0.0

Το συσσωρευτικό σχέδιο δίνεται πιο κάτω:

Στάδιο	Σύνδες. Ομάδων		Επίπεδο ή
	Ομάδα 1	Ομάδα 2	Συντελεστής
1	1	2	1.0
2	1	3	3.0
3	4	5	5.0
4	1	4	13.0

Ο συσσωρευτικός συντελεστής είναι η απόσταση μεταξύ ενός ζεύγους στοιχείων που ενώνονται.

Το συσσωρευτικό σχέδιο περιέχει τις ακόλουθες πληροφορίες:

- Στο αρχικό στάδιο τα στοιχεία 1 και 2 και κατά συνέπεια οι ομάδες τους ενώνονται (στάδιο v1).

- Στο δεύτερο στάδιο τα στοιχεία 1 και 3 και κατά συνέπεια οι ομάδες τους συγξωνεύονται σε βαθμό 3.0. Το σχέδιο παρουσιάζει μόνο το πρώτο στοιχείο κάθε ομάδας. Η ομάδα 1ου και 2ου σταδιου αποτελείται από δύο στοιχεία, το 1 και το 2 που είχαν ενωθεί στο πρώτο στάδιο.

- Στο στάδιο 3 οι ομάδες των στοιχείων 4 και 5 ενώνονται σε επίπεδο 5.0.

- Στο 4ο στάδιο οι ομάδες των στοιχείων 1 και 4 συγχωνεύονται. Όλα τα στοιχεια ανήκουν σε μια μόνο ομάδα γιατί η ομάδα 1 αποτελείται από τα στοιχεία 1, 2 και 3 και η ομάδα 2 αποτελείται από τα στοιχεία 4 και 5. Οι ομάδες ενώνονται σε επίπεδο 13.0.

Το σχέδιο δεν μας δίνει πληροφορίες σχετικά με το ποια στοιχεία ανήκουν σε μια ομάδα. Όμως οι πληροφορίες μπορούν να εξαχθούν από το σχέδιο.

Ο πρωταρχικός στόχος είναι να πάρουμε πληροφορίες σχετικά με τη διαδικασία και τον αριθμό των ομάδων. Τα συσσωρευτικά επίπεδα συνέχεια αυξάνονται (αν χρησιμοποιούμε την ανομοιογένεια) ή μειώνονται (αν αναλύουμε την ομοιογένεια).

Ορισμένα προγράμματα υπολογιστών παρέχουν ακόμη περισσότερες πληροφορίες. Η SPSS παρουσιάζει το στάδιο κατά το οποίο μια ομάδα εμφανίζεται και το στάδιο κατά το οποίο ενώνεται με μια άλλη ομάδα:

	Clusters Combined		Agglom.	Stage Cluster First Appears		Next
Stage	Clust 1	Clust 2	Coeffic.	Clust 1	Clust 2	Stage
1	1	2	1.0	0	0	2
2	1	3	3.0	1	0	4
3	4	5	5.0	0	0	4
4	1	4	13.0	2	3	0

Δυστυχώς η SPSS δεν παρουσιάζει τον αριθμό των ομάδων, την αύξηση ανομοιογένειας ή την μείωση και τους δεσμούς. Οι δεσμοί δεν εμφανίζονται όταν περισσότερα από ένα ζευγάρια των *πιο όμοιων* ομάδων υπάρχουν σε ένα συγκεκριμένο στάδιο.

Ο πίνακας μπορεί να χρησιμοποιηθεί για τον εντοπισμό παρατηρήσεων που εισέρχονται στην διαδικασία ομαδοποίησης πολύ αργά - πιθανές ακραίες παρατηρήσεις.

Πόσες ομάδες πρέπει να δημιουργήσουμε

- Μια από τις σημαντικότερες αποφάσεις στην ανάλυση ομαδοποίηση είναι ο προσδιορισμός του τελικού αριθμού των ομάδων που πρέπει να δημιουργηθούν (stopping rule). Δυστυχώς δεν υπάρχει μια αντικειμενική διαδικασία ή κανόνας για αυτή την επιλογή. Ούτε μπορεί να χρησιμοποιηθεί κάποιο στατιστικό κριτήριο για διεξαγωγή συμπερασμάτων. Βέβαια υπάρχουν κάποιες εξειδικευμένες διαδικασίες, που όμως είναι πολύπλοκες.

- Μια ιδέα είναι η εξέταση σε κάθε στάδιο μέτρων ομοιογένειας ή απόστασης μεταξύ των ομάδων. Ο αριθμός των ομάδων καθορίζεται όταν το μέτρο ομοιογένειας υπερβεί μια συγκεκριμένη τιμή ή όταν οι τιμές σε συνεχόμενα στάδια αυξηθούν απότομα.

- Όταν παρουσιαστεί μια μεγάλη αύξηση επιλέγουμε τον αμέσως προηγούμενο αριθμό ομάδων με την λογική ότι ο νέος συνδιασμός έχει προκαλέσει ουσιαστική μείωση στην ομοιογένεια. Αυτή η μέθοδος δίνει μια σχετικά ακριβή επιλογή ομάδων σε εμπειρικές μελέτες.

- Συχνά ο αριθμός των ομάδων καθορίζεται βάση του δενδρογράμματος. Ο χρήστης σημειώνει τον αριθμό των μικρών λόφων (ομάδων) που ενώνουν στοιχεία σε μικρή απόσταση.

  Κάποιος μπορεί επίσης να παρατηρήσει τα συσσωρευτικά επίπεδα. Διαβάζουμε τα επίπεδα από πάνω προς τα κάτω, ξεκινώντας από το στάδιο 1.

  Ο χρήστης βρίσκει της απότομες αυξήσεις (αν χρησιμοποιεί την ανομοιογένεια) ή μειώσεις (αν χρησιμοποιεί την ομοιογένεια) στα συσσωρευτικά επίπεδα.

  Στο τελευταίο παράδειγμα που εξηγήσαμε παρουσιάζεται μια απότομη αύξηση στο στάδιο 3 και 4. Έτσι, το στάδιο 3 θεωρείται το καλύτερα στάδιο για να σταματήσουμε. Ο λόγος είναι ότι στο στάδιο 4 το συσσωρευτικό επίπεδο αυξάνεται ραγδαία.

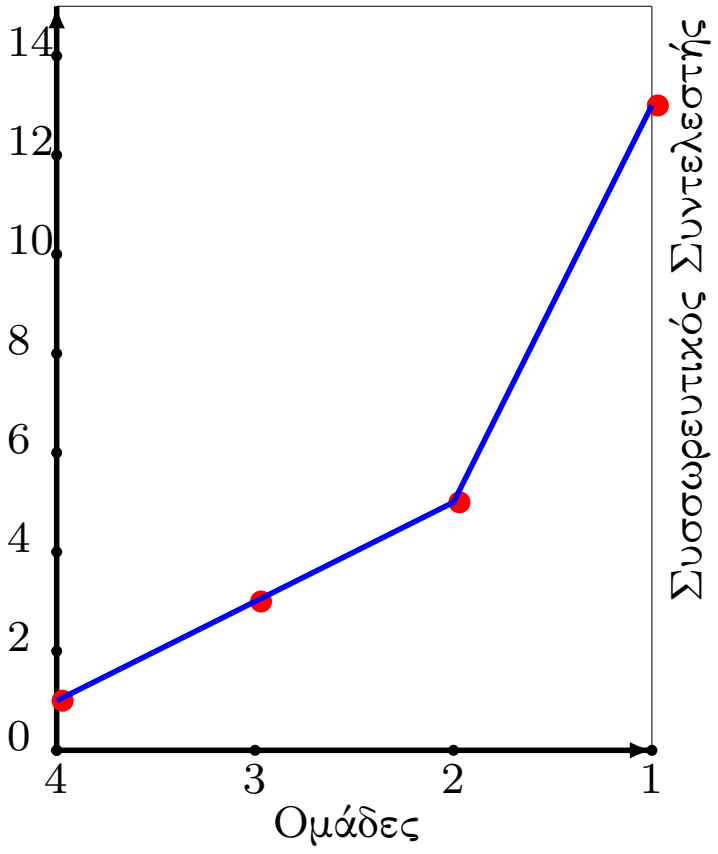
  Το στάδιο 3 προτείνει λύση δύο ομάδων. Η ομάδα 1 περιέχει τα στοιχεία 1, 2 και 3 , ενώ η ομάδα 2 τα στοιχεία 4 και 5.

  Η μέθοδος που επεξηγήθηκε πιο πάνω να αναπαρασταθεί στο λεγόμενο inverse scree test

(αντίστροφη διαγνωστική εξέταση). Για το σκοπό αυτό δημιουργείται ένα γράφημα.

Η εξέταση αυτή δεν είναι στατιστικού περιεχομένου. Ο άξονας x υποδεικνύει τον αριθμό των ομάδων. Ο άξονας y δείχνει το συσσωρευτικό επίπεδο. Μια απότομη αύξηση στο σχέδιο συσσέρευσης, οδηγεί σε μια απότομη καμπή (elbow knick).

Στο παράδειγμα που χρησιμοποιήσαμε, η καμπή εμφανίζεται στη 2η ομάδα (όπως δείχνει το γράφημα). Έτσι επιλέγουμε να δημιουργήσουμε 2 ομάδες.



- Πολλές φορές ειναι καλύτερα να υπολογίζουμε ένα αριθμό διαφορετικών λύσεων (π.χ. 2, 3, 4 ομάδες) και έπειτα να επιλέγουμε τον κατάλληλο αριθμό είτε βάση απλών κριτηρών, κρίσης, κοινής λογικής ή θεωρητικών κριτηρίων.

  Η λύση ομαδοποίησης μπορεί να βελτιωθεί αισθητά εάν έχουμε σαν βάση τις εννοιολογικές πτυχές του προβλήματος.

Παράδειγμα

Αρ. Ομάδων	Συσσερευτ. Συντελεστής	Ποσοστιαία Αλλαγή στο Συντελ. στο επόμενο στάδιο
10	258.7	8.8
9	281.4	8.4
8	305.0	9.2
7	331.1	9.6
6	364.9	9.1
5	398.1	12.1
4	446.3	17.2
3	523.0	17.6
2	615.0	61.8
1	994.8	–

Ο συντελεστής ομαδοποίησης (συσσώρευσης) δείχνει μια κάπως μεγάλη αύξηση από τις 4 στις 3 ομάδες (523.0-446.3=76.6), από τις 3 στις 2 ομάδες (615.0-523.0 = 92.0), και από 2 στη 1 ομάδα (994.8-615.0 = 379.8). Για τον προσδιορισμό της αύξησης ομοιογένειας σε σχετικούς όρους, υπολογίζουμε την ποσοστιαία αλλαγή στο συντελεστή ομαδοποίησης. Η μεγαλύτερη ποσοστιαία αύξηση παρουσιάζεται από τις 2 στη 1 ομάδα. Η αμέσως μεγαλύτερη αλλαγή είναι η ποσοστιαία αύξηση από τις 4 στις 3 ομάδες.

Μειονεκτήματα ιεραρχικών διαδικασιών

- Χρονοβόρες (για ένα σύνολο δεδομένων με n στοιχεία, η διαδικασία προτείνει n διαφορετικές λύσεις).

- Δεν είναι δυνατή η τροποποίηση προηγούμενων κατατάξεων. Ένα στοιχείο που αρχικά κατατάχθηκε σε μία ομάδα, δεν πρόκειται να ανατεθεί σε καμία άλλη ομάδα. Δηλαδή δεν μπορεί να διορθωθεί μια λανθασμένη προηγούμενη απόφαση.

Παράδειγμα

Κάποιοι καταναλωτές έχουν ζητηθεί να δηλώσουν τον βαθμό συμφωνίας τους με τις ακόλουθες δηλώσεις. (1=διαφωνώ, 7=συμφωνώ):

- V1: Τα ψώνια είναι διασκεδαστικά.
- V2: Τα ψώνια βλάπτουν το πορτοφόλι.
- V3: Συνδυάζω ψωνια και φαγητό.
- V4: Προσπαθώ να κάνω τις καλύτερες επιλογές αγαθών.
- V5: Δεν ενδιαφέρομαι για ψώνια.
- V6: Εξοικονομω πολλά με την σύγκριση τιμών.

V1	V2	V3	V4	V5	V6
6.00	4.00	7.00	3.00	2.00	3.00
2.00	3.00	1.00	4.00	5.00	4.00
7.00	2.00	6.00	4.00	1.00	3.00
4.00	6.00	4.00	5.00	3.00	6.00
1.00	3.00	2.00	2.00	6.00	4.00
6.00	4.00	6.00	3.00	3.00	4.00
5.00	3.00	6.00	3.00	3.00	4.00
7.00	3.00	7.00	4.00	1.00	4.00
2.00	4.00	3.00	3.00	6.00	3.00
3.00	5.00	3.00	6.00	4.00	6.00
1.00	3.00	2.00	3.00	5.00	3.00
5.00	4.00	5.00	4.00	2.00	4.00
2.00	2.00	1.00	5.00	4.00	4.00
4.00	6.00	4.00	6.00	4.00	7.00
6.00	5.00	4.00	2.00	1.00	4.00
3.00	5.00	4.00	6.00	4.00	7.00
4.00	4.00	7.00	2.00	2.00	5.00
3.00	7.00	2.00	6.00	4.00	3.00
4.00	6.00	3.00	7.00	2.00	7.00
2.00	3.00	2.00	4.00	7.00	2.00

SPSS:

```
Analyze/Classify/Hierarchical Cluster
/Statistics  (tick on proximity matrix)
 / Plots      (tick on dendrogram)
 / Method     (select cluster method e.g
   within-groups linkage and interval
   e.g Euclidean distance)
```

Proximity Matrix

Case	City Block Distance 1	2	3	4	5	6	7	8	9	10	11	12	13
1	.000	16.000	6.000	13.000	17.000	3.000	5.000	5.000	12.000	16.000	14.000	5.000	17.000
2	16.000	.000	16.000	13.000	5.000	13.000	11.000	15.000	6.000	10.000	4.000	11.000	3.000
3	6.000	16.000	.000	15.000	19.000	7.000	7.000	3.000	16.000	18.000	16.000	7.000	15.000
4	13.000	13.000	15.000	.000	16.000	10.000	10.000	14.000	13.000	5.000	15.000	8.000	12.000
5	17.000	5.000	19.000	16.000	.000	14.000	12.000	18.000	5.000	13.000	3.000	14.000	8.000
6	3.000	13.000	7.000	10.000	14.000	.000	2.000	6.000	11.000	13.000	13.000	4.000	14.000
7	5.000	11.000	7.000	10.000	12.000	2.000	.000	6.000	11.000	13.000	11.000	4.000	12.000
8	5.000	15.000	3.000	14.000	18.000	6.000	6.000	.000	17.000	17.000	17.000	6.000	16.000
9	12.000	6.000	16.000	13.000	5.000	11.000	11.000	17.000	.000	10.000	4.000	11.000	9.000
10	16.000	10.000	18.000	5.000	13.000	13.000	13.000	17.000	10.000	.000	12.000	11.000	9.000
11	14.000	4.000	16.000	15.000	3.000	13.000	11.000	17.000	4.000	12.000	.000	13.000	7.000
12	5.000	11.000	7.000	8.000	14.000	4.000	4.000	6.000	11.000	11.000	13.000	.000	12.000
13	17.000	3.000	15.000	12.000	8.000	14.000	12.000	16.000	9.000	9.000	7.000	12.000	.000
14	16.000	14.000	18.000	3.000	17.000	13.000	13.000	17.000	14.000	4.000	16.000	11.000	13.000
15	7.000	15.000	9.000	10.000	14.000	6.000	8.000	8.000	13.000	13.000	15.000	6.000	16.000
16	16.000	12.000	18.000	5.000	15.000	13.000	13.000	17.000	12.000	2.000	14.000	11.000	11.000
17	5.000	15.000	11.000	10.000	14.000	6.000	6.000	8.000	13.000	13.000	15.000	6.000	16.000
18	16.000	10.000	18.000	9.000	13.000	15.000	15.000	19.000	10.000	6.000	10.000	13.000	9.000
19	16.000	16.000	18.000	5.000	19.000	15.000	15.000	17.000	16.000	6.000	18.000	11.000	15.000
20	17.000	5.000	17.000	16.000	6.000	16.000	14.000	18.000	5.000	13.000	5.000	14.000	8.000

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	14	16	2.000	0	0	3
2	6	7	2.000	0	0	7
3	10	14	3.000	0	1	8
4	2	13	3.000	0	0	14
5	5	11	3.000	0	0	10
6	3	8	3.000	0	0	15
7	6	12	4.000	2	0	9
8	4	10	4.333	0	3	11
9	1	6	4.333	0	7	13
10	5	9	4.500	5	0	12
11	4	19	5.250	8	0	17
12	5	20	5.333	10	0	14
13	1	17	5.750	9	0	15
14	2	5	6.500	4	12	18
15	1	3	6.900	13	6	16
16	1	15	7.429	15	0	19
17	4	18	8.200	11	0	18
18	2	4	13.556	14	17	19
19	1	2	14.375	16	18	0

Μη-ιεραρχικές μέθοδοι

- Μειονεκτήματα: Είναι απαραίτητο να γνωρίζουμε τον αριθμό των ομάδων εκ των προτέρων.

  1. Επιλέγουμε ένα αρχικό αριθμό k ομάδων.
  2. Κατατάσσουμε κάθε παρατήρηση στην κοντινότερη ομάδα.
  3. Επανατοποθετούμε κάθε παρατήρηση σε μια από τις ομάδες k βάση κανόνα.
  4. Η διαδικασία σταματά όταν δεν υπάρχουν δυνατότητες επανατοποθέτησης. Διαφορετικά επιστρέφουμε στο στάδιο 2.

- Οι περισσότεροι μη-ιεραρχικοί αλγόριθμοι διαφέρουν:

  1. Στη μέθοδο επιλογής αρχικού αριθμού κεντροειδών.
  2. Στον κανόνα επανατοποθέτησης παρατηρήσεων.

Μέθοδοι επιλογής αριθμού κεντροειδών

- Επιλέξτε ένα αρχικό αριθμό k κεντροειδών τα οποία αποτελούν και τα μόνα στοιχεία των ομάδων.
- Χρησιμοποιήστε την πρώτη παρατήρηση ως το πρώτο κεντροειδές. Η δεύτερη παρατήρηση που απέχει μια καθορισμένη απόσταση από το πρώτο κεντροειδές είναι το δεύτερο κεντροειδές, το τρίτο πρέπει να απέχει μια καθορισμένη απόσταση από το προηγούμενο κεντροειδές, ...
- Επιλέγουμε τυχαία k παρατηρήσεις.
- Εξηγήστε την επιλογή κεντροειδών με ένα κανόνα.
- Χρησιμοποιήστε μια ευρετική μέθοδο που προσδιορίζει κέντρα ομάδων που βρίσκονται όσο το δυνατό μακρύτερα το ένα από το άλλο.
- Δίνεται από τον ερευνητή.

Κανόνες επανατοποθέτησης παρατηρήσεων

1. Υπολογίστε τα κεντροειδή
   - Ανακατανέμετε τα στοιχεία στην ομάδα με το κοντινότερο κεντροειδές.
   - Υπολογίστε εκ νέου τα κεντροειδή αφού επανατοποθετήσετε τις παρατηρήσεις.
   - Αν η αλλαγή στα κεντροειδή είναι μεγαλύτερη από μια συγκεκριμένη τιμή, αναδιανεμετε τα στοιχεία κτλ.
2. Η μέθοδος είναι η ίδια όπως την προηγούμενη μόνο που τώρα υπολογίζουμε νέα κεντρειδοί μετά από κάθε επανατοποθέτηση.

Η πρώτη μέθοδος ονομάζεται k-means.

Μέθοδος K-means

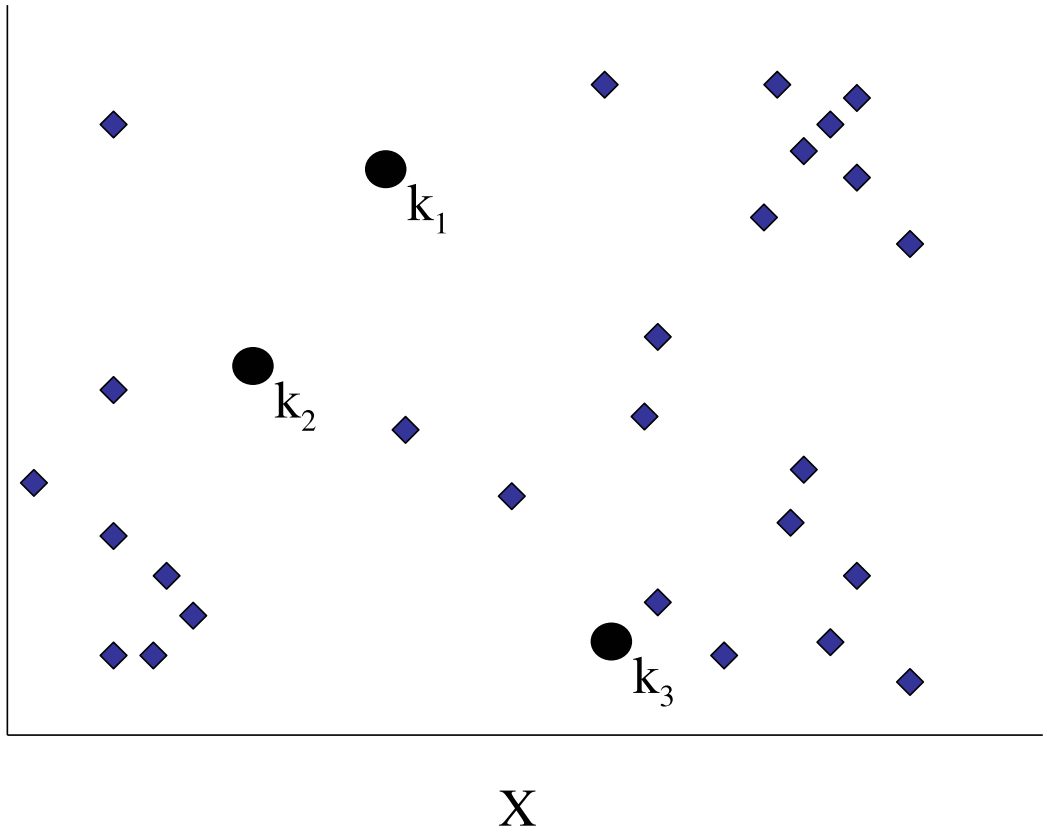
1. Αρχικά επιλέγουμε τυχαία τα κέντρα των ομάδων.
2. Τοποθετούμε τα στοιχεία στην κοντινότερη ομάδα (π.χ. κοντίτερα στο κεντροειδές της ομάδας).
3. Επαναπροσδιορίζουμε το νέο κεντοειδές της ομάδας.
4. Επανατοποθετούμε τα στοιχεία στις ομάδες.
5. Επαναλαμβάνουμε το στάδιο 2 μέχρις ότου δεν υπάρχει καμία αλλαγή.

- Υπάρχουν n^k πιθανές κατατάξεις n στοιχείων σε k ομάδες.
- Στην πράξη είναι δύσκολο να επιτευχθεί βέλτιστη κατάταξη.
- Η κατάταξη εξαρτάται από τον αρχικό αριθμό ομάδων.
- Δοκιμάστε διαφορετικούς αριθμούς ομάδων.

k-Means Example (I)

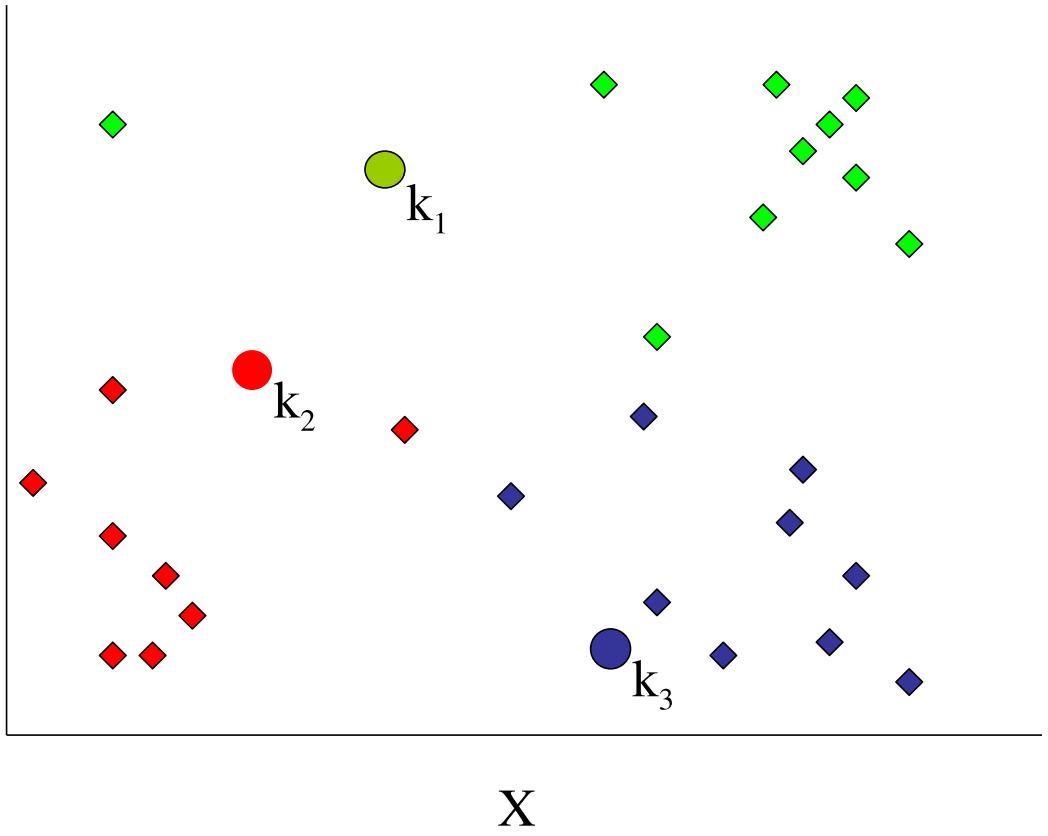
K-means: Παράδειγμα 1

Pick 3 initial cluster centers (randomly)



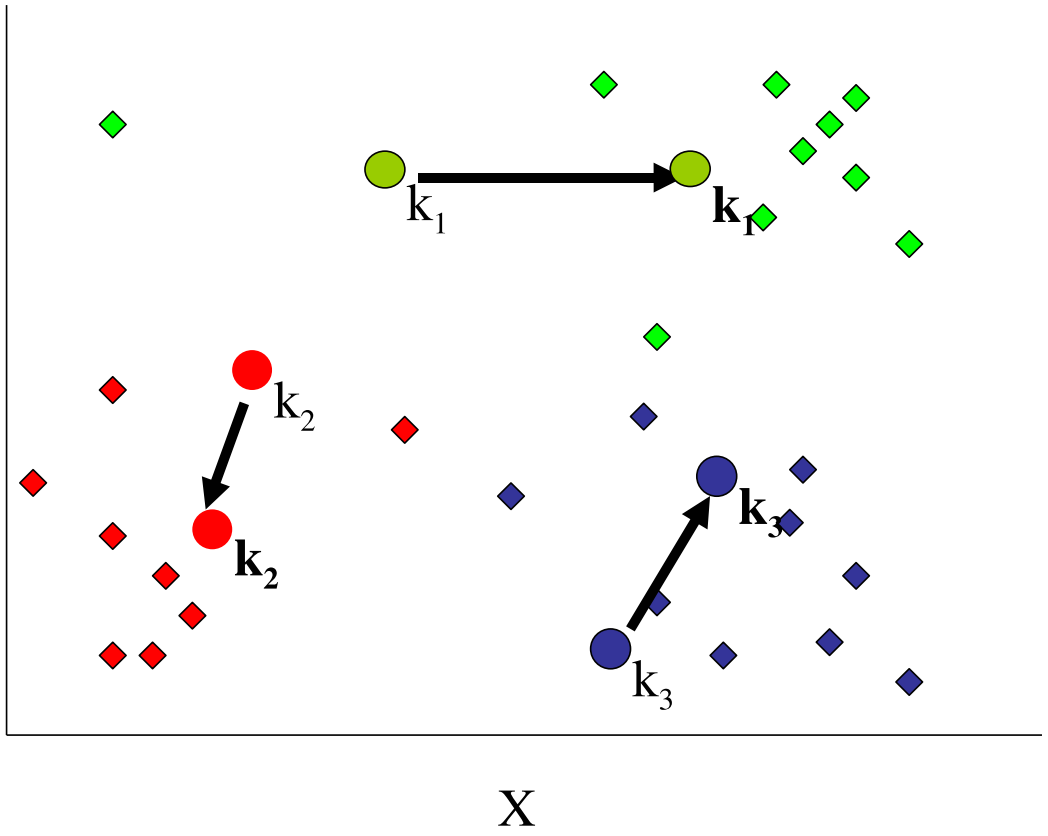
k-Means Example (II)

Assign each point to the closest cluster center



k-Means Example (III)

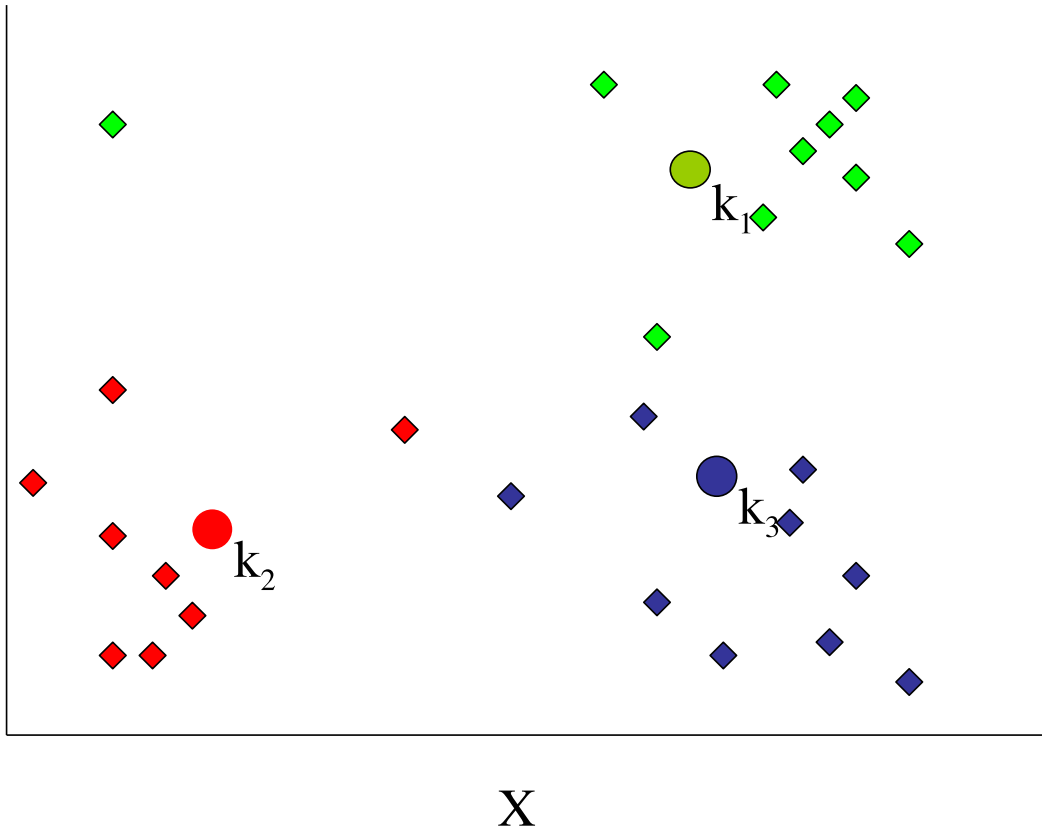
Move each cluster center to the mean of each cluster



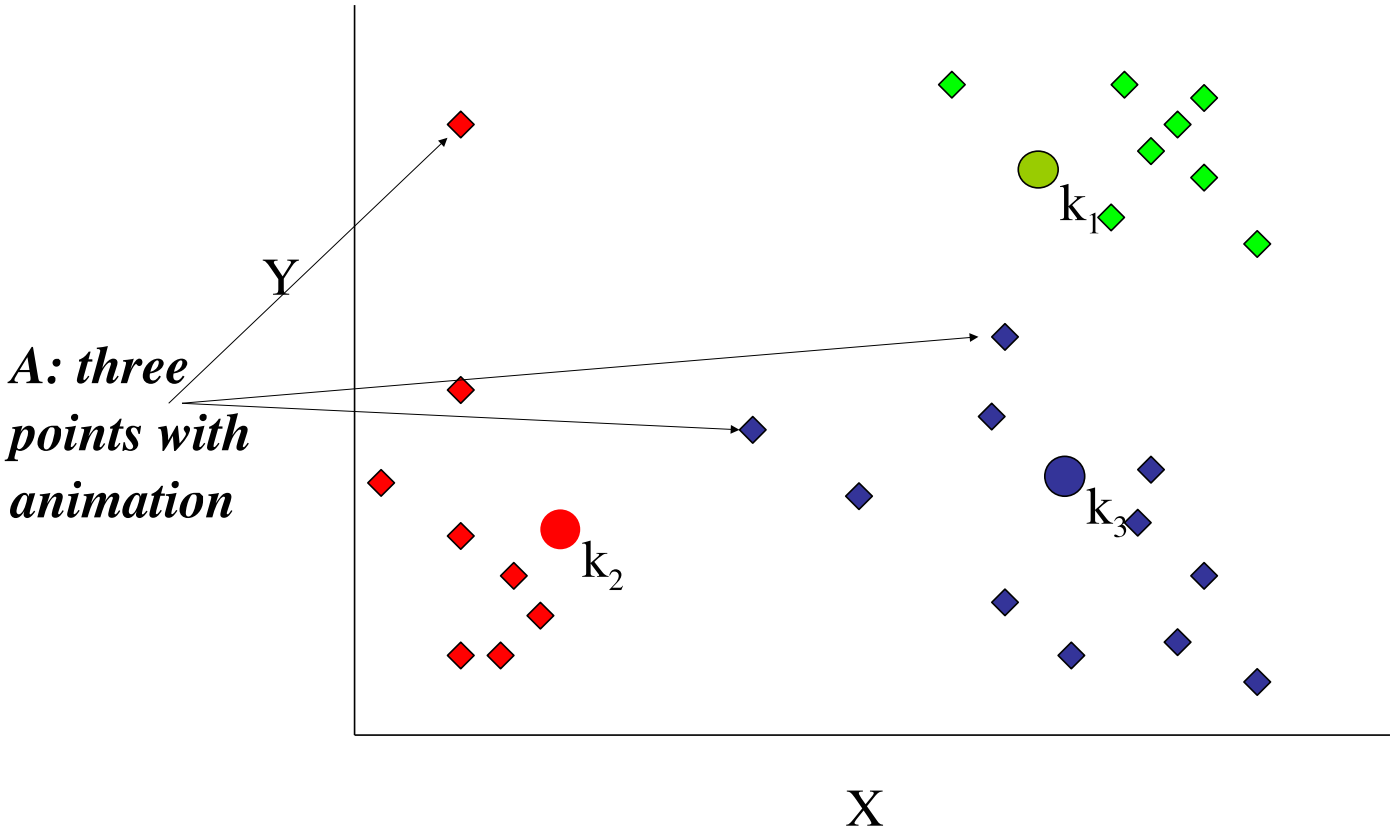
k-Means Example (IV)

Reassign points closest to a different new cluster center

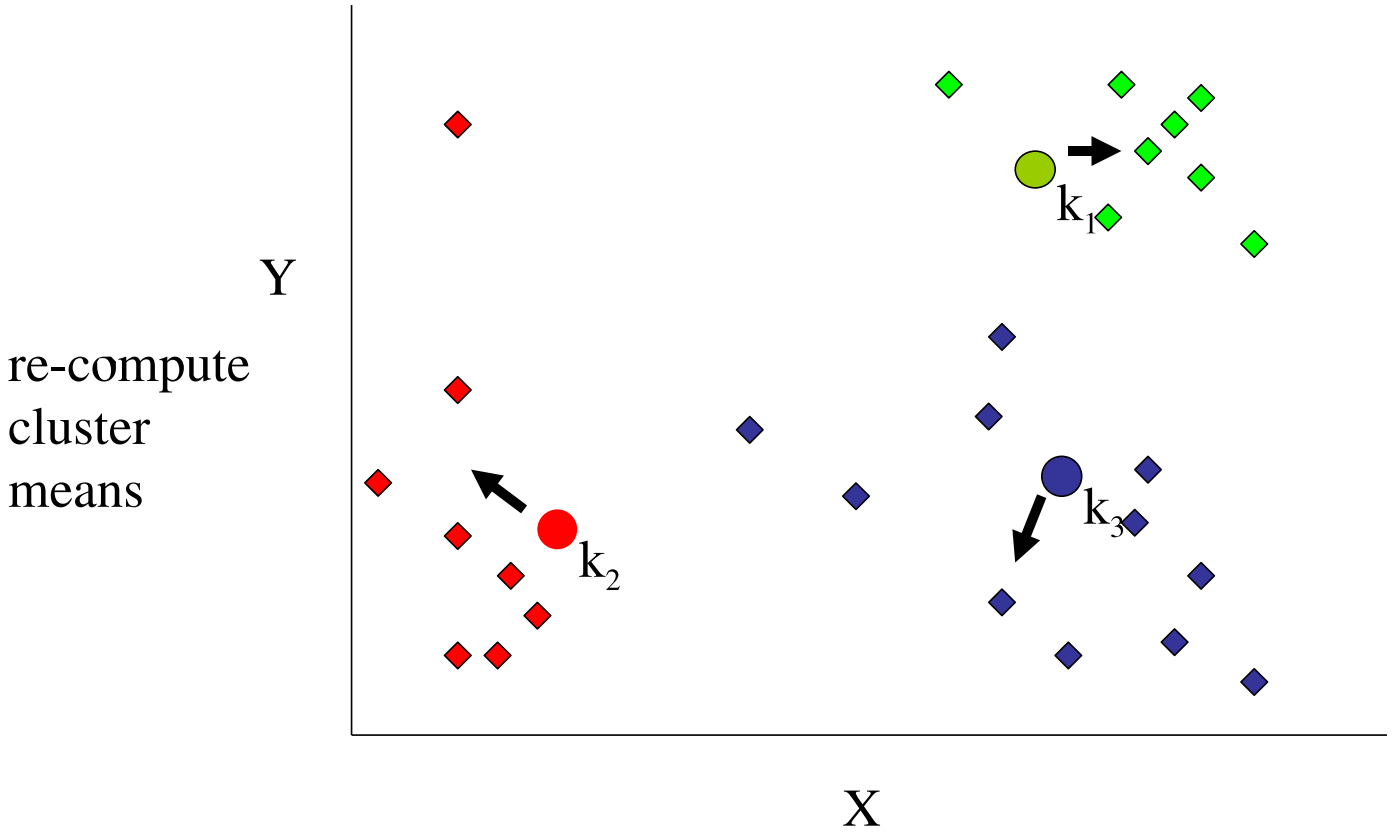
Q: Which points are reassigned?



k-Means Example (V)

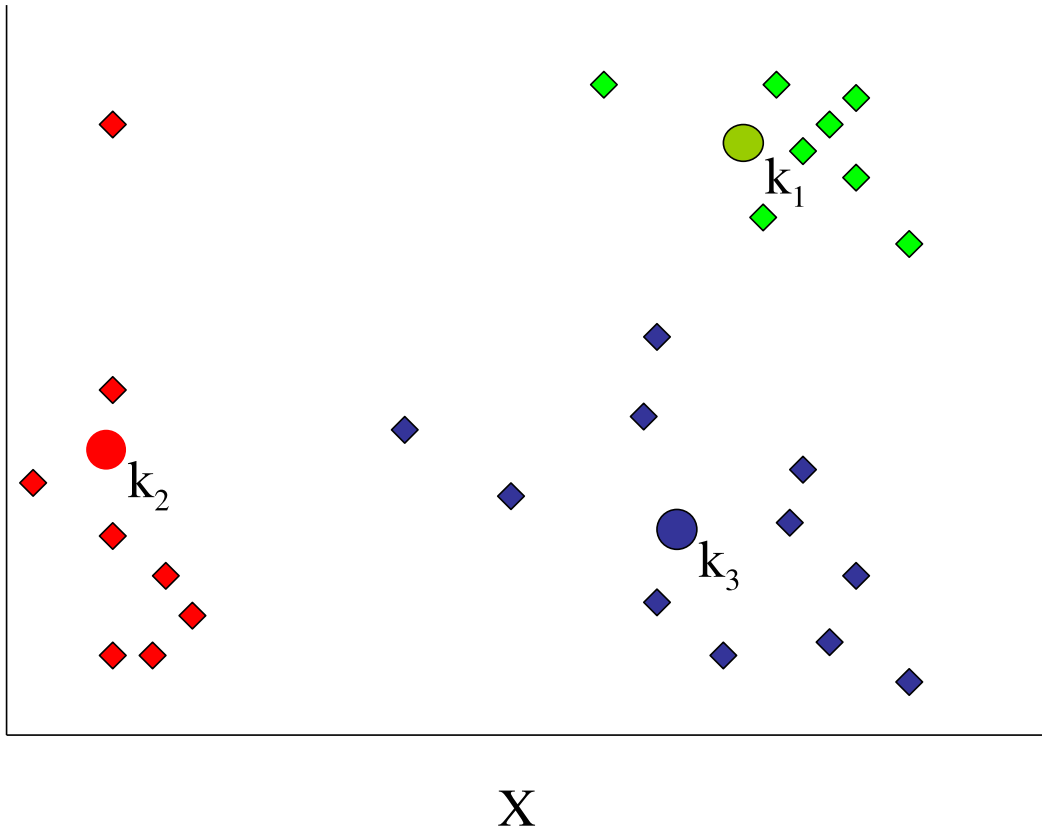


k-Means Example (VI)



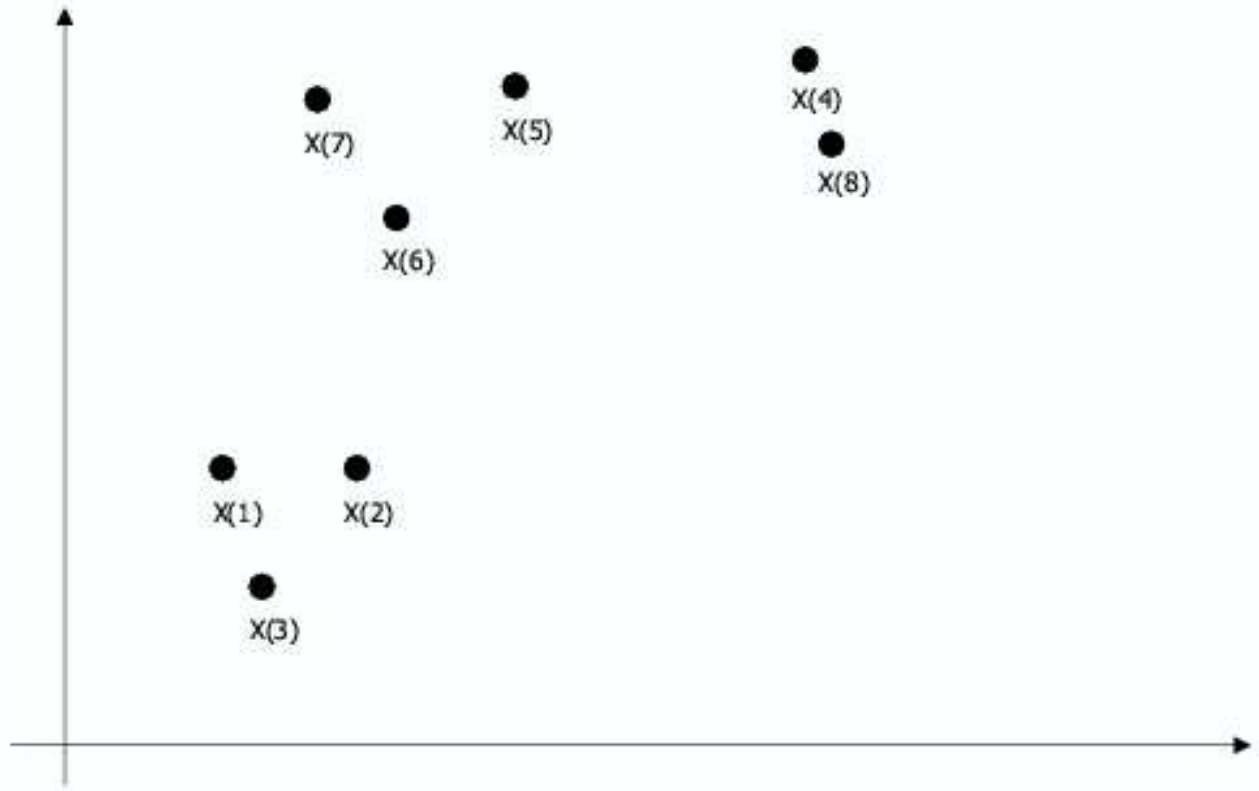
k-Means Example (VII)

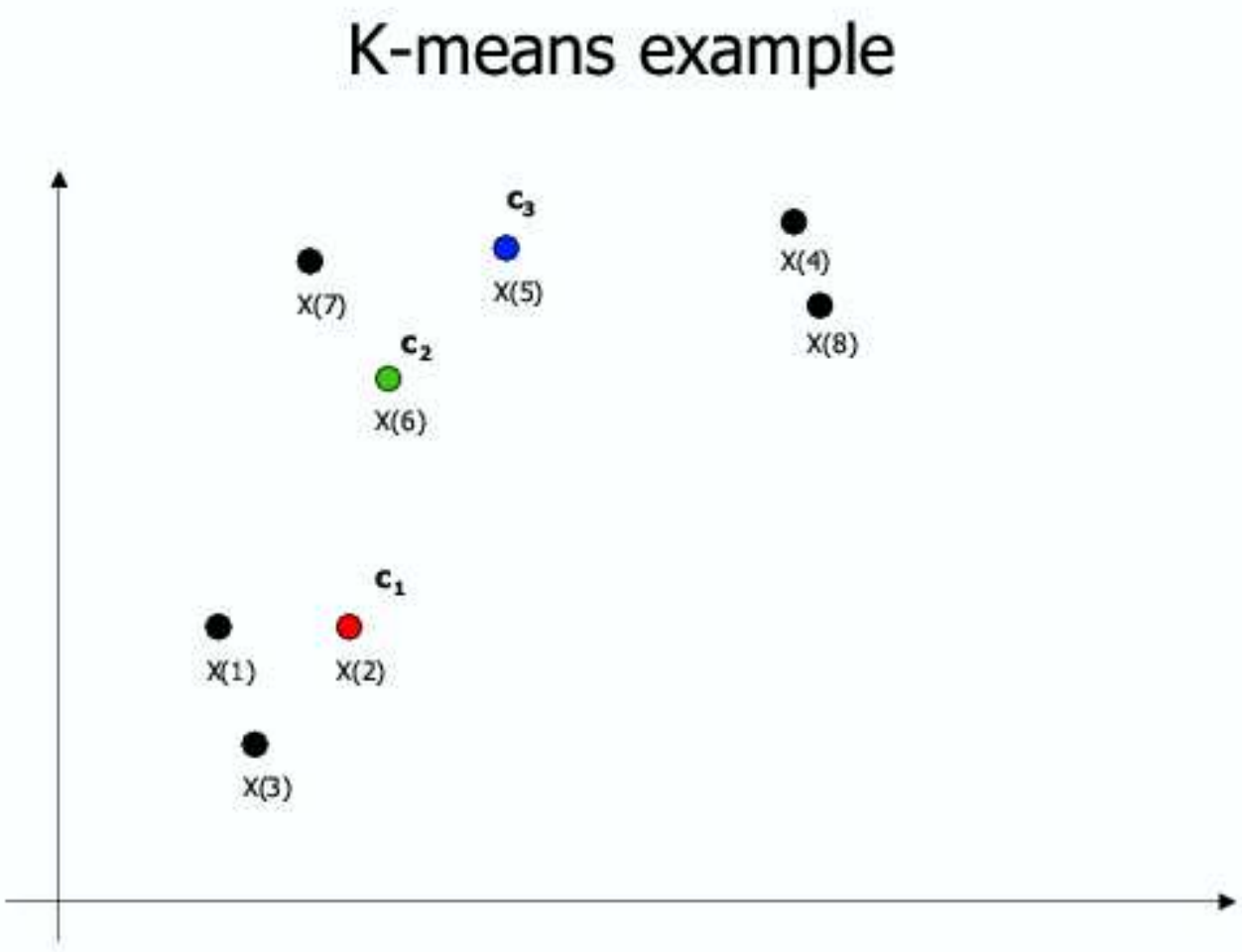
move cluster centers to cluster means

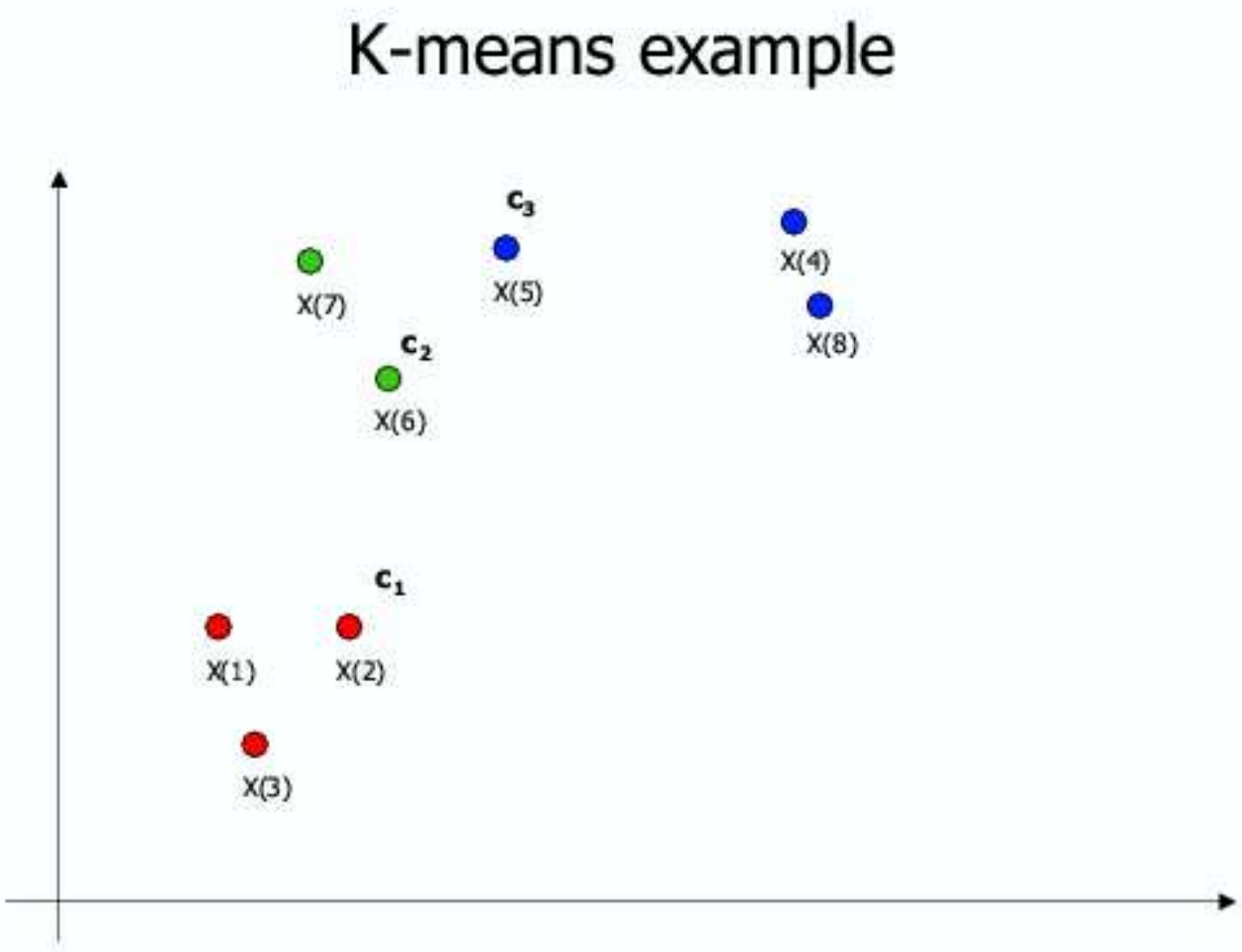


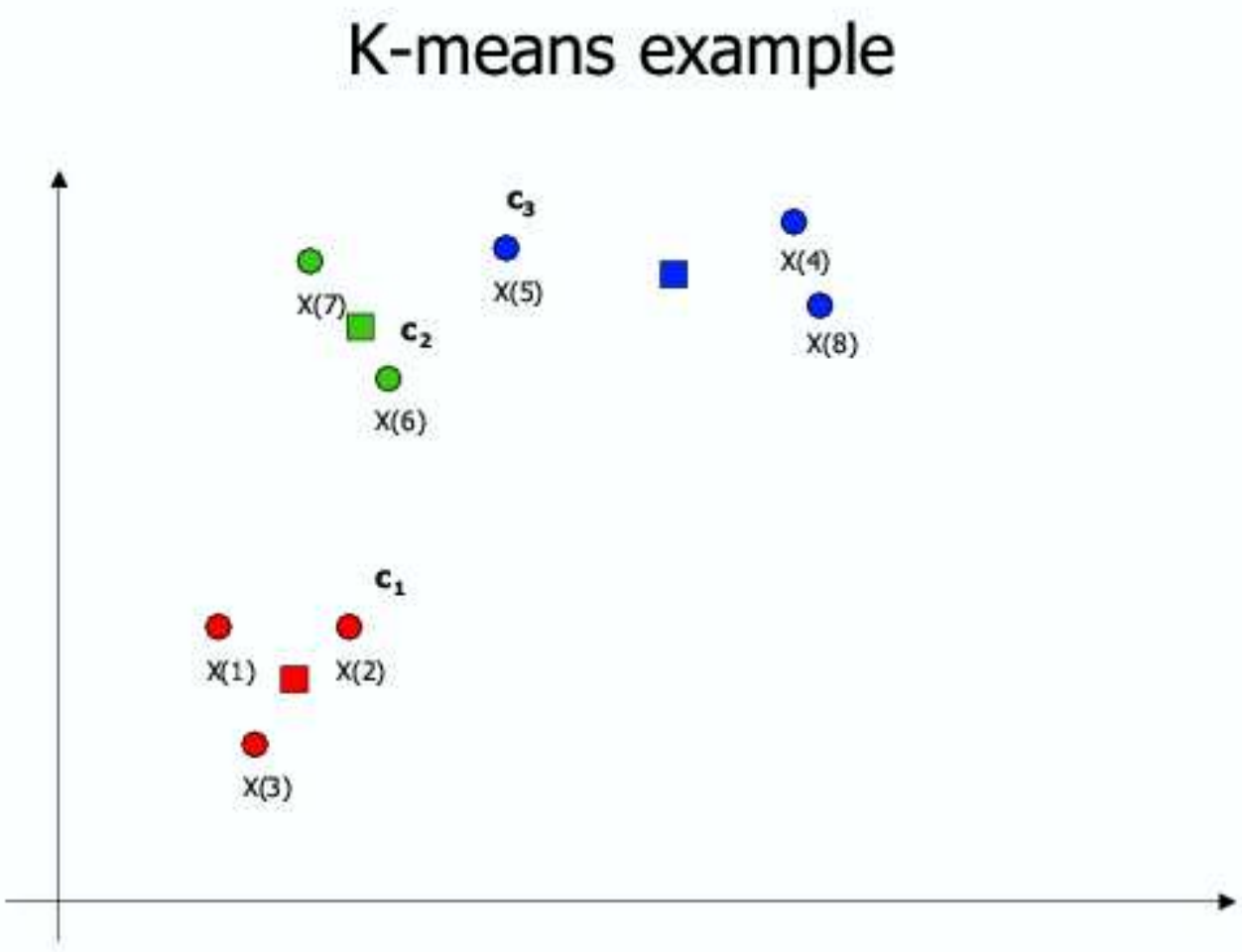
K-means: Παράδειγμα 2

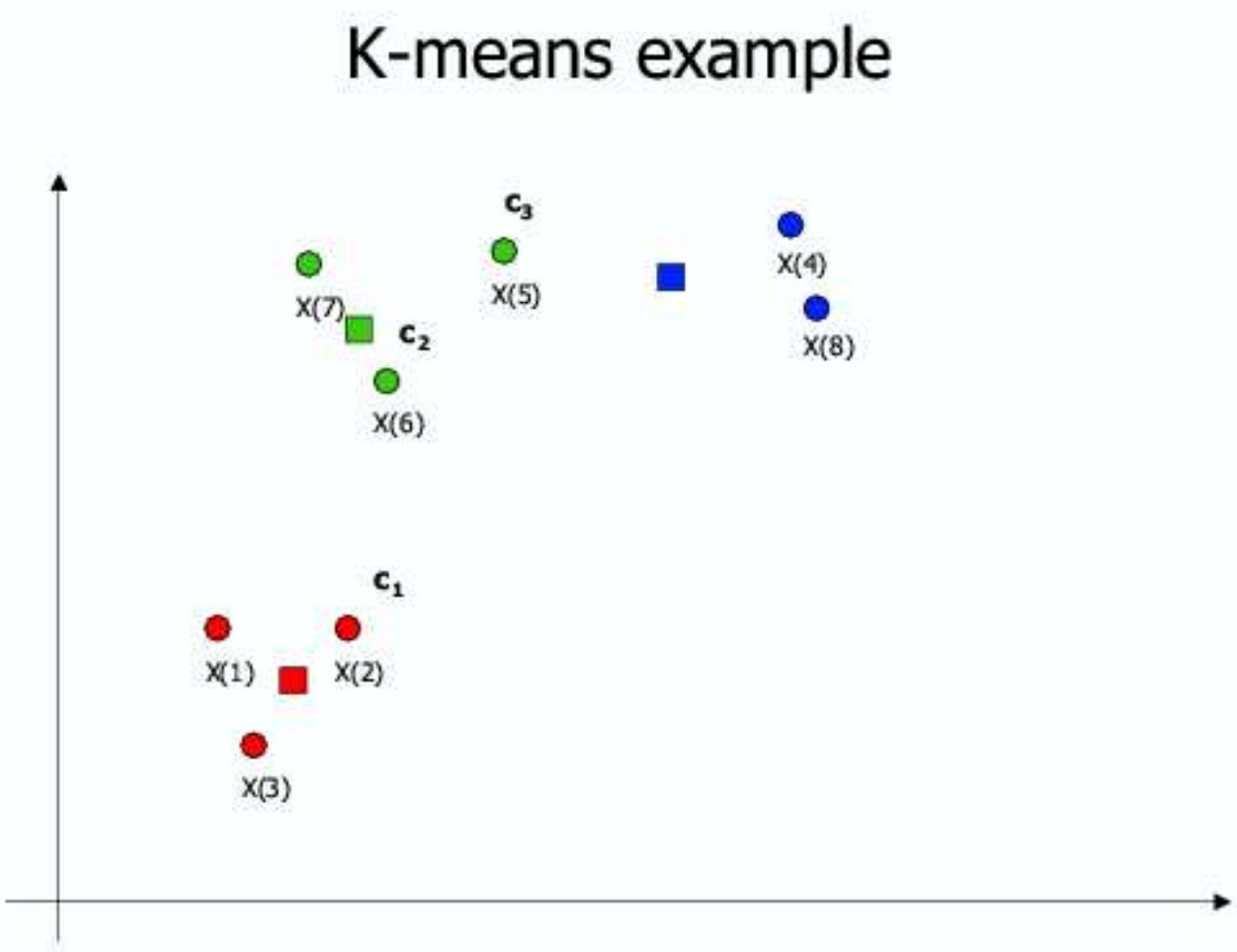
K-means example



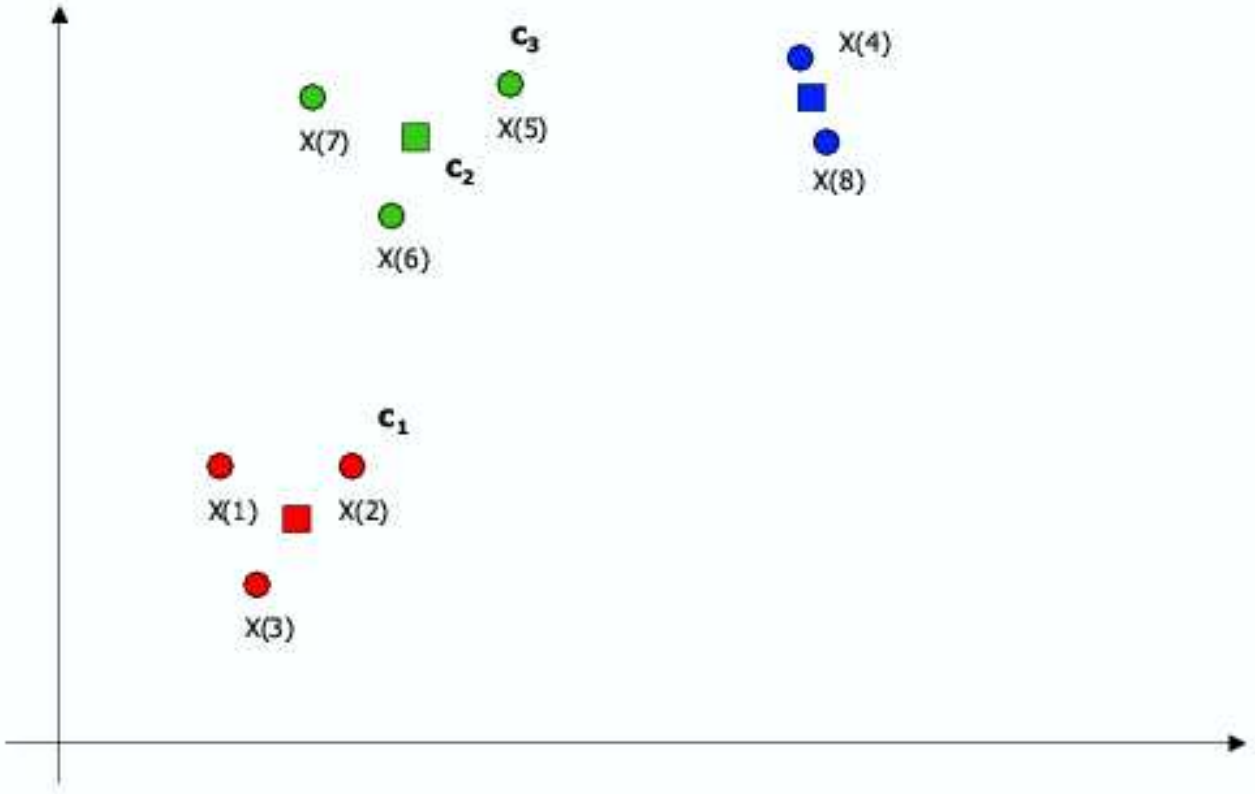

K-means example
c3
X(7)
X(5)
X(4)
X(8)
c2
X(6)
c1
X(1)
X(2)
X(3)


K-means example
c3
X(7)
X(5)
X(4)
X(8)
c2
X(6)
c1
X(1)
X(2)
X(3)


K-means example
c3
X(5)
X(4)
X(8)
X(7)
c2
X(6)
c1
X(1)
X(2)
X(3)


K-means example
c3
X(5)
X(4)
X(7)
c2
X(8)
X(6)
c1
X(1)
X(2)
X(3)

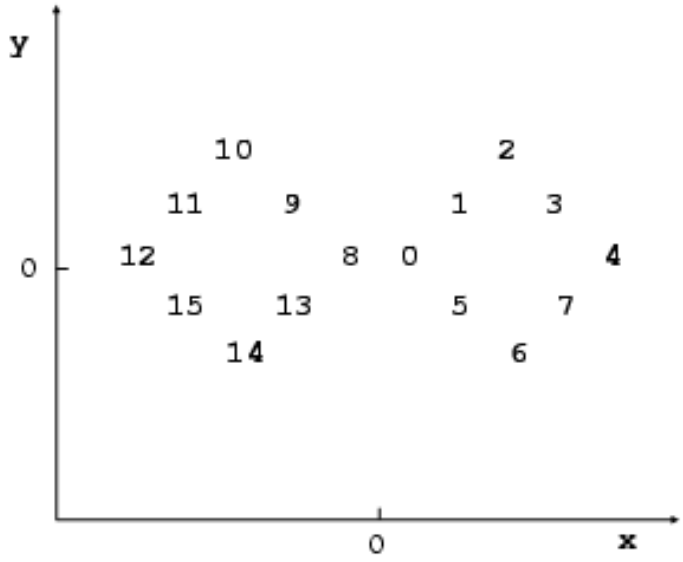
K-means example



K-means: Παράδειγμα 3

k-means: Example

Id	x	y
0:	1.0	0.0
1:	3.0	2.0
2:	5.0	4.0
3:	7.0	2.0
4:	9.0	0.0
5:	3.0	-2.0
6:	5.0	-4.0
7:	7.0	-2.0
8:	-1.0	0.0
9:	-3.0	2.0
10:	-5.0	4.0
11:	-7.0	2.0
12:	-9.0	0.0
13:	-3.0	-2.0
14:	-5.0	-4.0
15:	-7.0	-2.0



- find the best 2 clusters

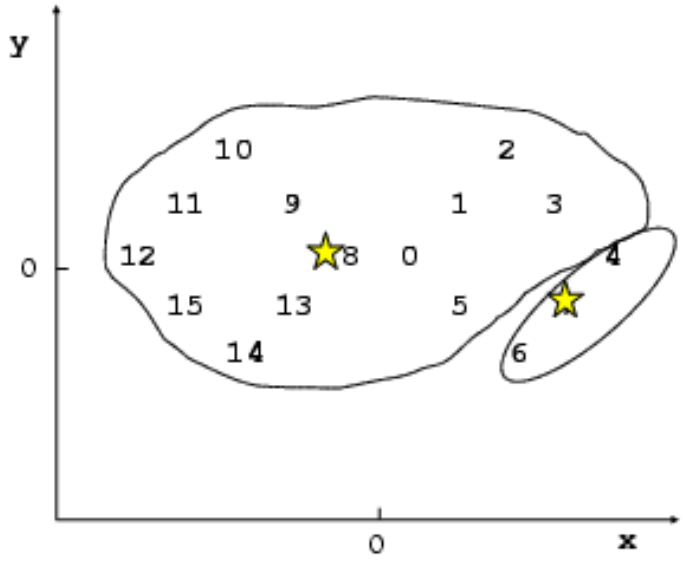
Seed: (9 0) (8 1)

Clustering: (4 6 7) (0 1 2 3 5 8 9 10 11 12 13 14 15)
Cluster Centers: (7.0 -2.0) (-1.61538 0.46153)
Average Distance: 4.35887

Seed: (9 0) (8 1)

Clustering: (4 6 7) (0 1 2 3 5 8 9 10 11 12 13 14 15)
Cluster Centers: (7.0 -2.0) (-1.61538 0.46153)
Average Distance: 4.35887

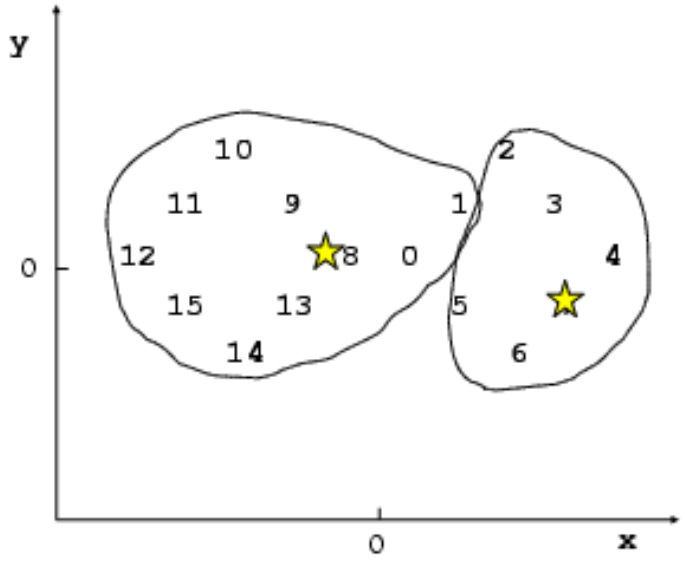
Clustering: (2 3 4 5 6 7) (0 1 8 9 10 11 12 13 14 15)



Seed: (9 0) (8 1)

Clustering: (4 6 7) (0 1 2 3 5 8 9 10 11 12 13 14 15)
Cluster Centers: (7.0 -2.0) (-1.61538 0.46153)
Average Distance: 4.35887

Clustering: (2 3 4 5 6 7) (0 1 8 9 10 11 12 13 14 15)
Cluster Centers: (6.0 -0.33334) (-3.6 0.2)
Average Distance: 3.6928



Seed: (9 0) (8 1)
Clustering: (4 6 7) (0 1 2 3 5 8 9 10 11 12 13 14 15)
Cluster Centers: (7.0 -2.0) (-1.61538 0.46153)
Average Distance: 4.35887
Clustering: (2 3 4 5 6 7) (0 1 8 9 10 11 12 13 14 15)
Cluster Centers: (6.0 -0.33334) (-3.6 0.2)
Average Distance: 3.6928
Clustering: (1 2 3 4 5 6 7) (0 8 9 10 11 12 13 14 15)
y
x
0
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Seed: (9 0) (8 1)
Clustering: (4 6 7) (0 1 2 3 5 8 9 10 11 12 13 14 15)
Cluster Centers: (7.0 -2.0) (-1.61538 0.46153)
Average Distance: 4.35887
Clustering: (2 3 4 5 6 7) (0 1 8 9 10 11 12 13 14 15)
Cluster Centers: (6.0 -0.33334) (-3.6 0.2)
Average Distance: 3.6928
Clustering: (1 2 3 4 5 6 7) (0 8 9 10 11 12 13 14 15)
Cluster Centers: (5.57143 0.0) (-4.33334 0.0)
Average Distance: 3.49115
y
x
0
0
1
2
3
4
5
6
7
0
8
9
10
11
12
13
14
15

Seed: (9 0) (8 1)
Clustering: (4 6 7) (0 1 2 3 5 8 9 10 11 12 13 14 15)
Cluster Centers: (7.0 -2.0) (-1.61538 0.46153)
Average Distance: 4.35887
Clustering: (2 3 4 5 6 7) (0 1 8 9 10 11 12 13 14 15)
Cluster Centers: (6.0 -0.33334) (-3.6 0.2)
Average Distance: 3.6928
Clustering: (1 2 3 4 5 6 7) (0 8 9 10 11 12 13 14 15)
Cluster Centers: (5.57143 0.0) (-4.33334 0.0)
Average Distance: 3.49115
Clustering: (0 1 2 3 4 5 6 7) (8 9 10 11 12 13 14 15)
y
x
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Seed: (9 0) (8 1)
Clustering: (4 6 7) (0 1 2 3 5 8 9 10 11 12 13 14 15)
Cluster Centers: (7.0 -2.0) (-1.61538 0.46153)
Average Distance: 4.35887
Clustering: (2 3 4 5 6 7) (0 1 8 9 10 11 12 13 14 15)
Cluster Centers: (6.0 -0.33334) (-3.6 0.2)
Average Distance: 3.6928
Clustering: (1 2 3 4 5 6 7) (0 8 9 10 11 12 13 14 15)
Cluster Centers: (5.57143 0.0) (-4.33334 0.0)
Average Distance: 3.49115
Clustering: (0 1 2 3 4 5 6 7) (8 9 10 11 12 13 14 15)
Cluster Centers: (5.0 0.0) (-5.0 0.0)
Average Distance: 3.41421
y
x
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Seed: (9 0) (8 1)

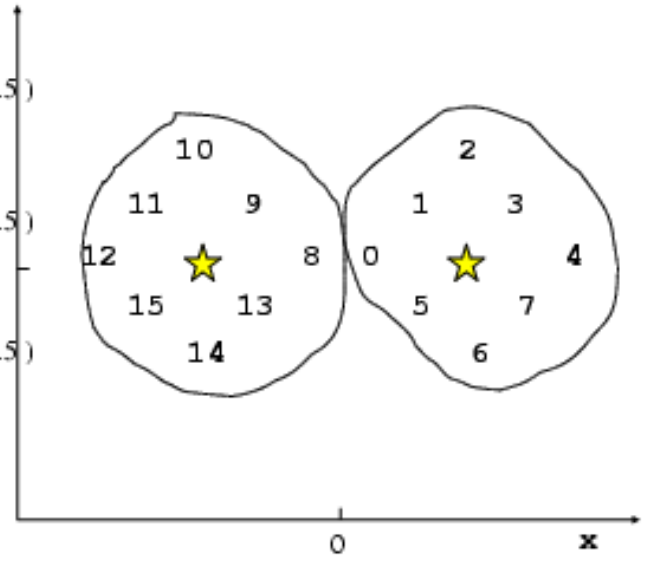
Clustering: (4 6 7) (0 1 2 3 5 8 9 10 11 12 13 14 15)
Cluster Centers: (7.0 -2.0) (-1.61538 0.46153)
Average Distance: 4.35887

Clustering: (2 3 4 5 6 7) (0 1 8 9 10 11 12 13 14 15)
Cluster Centers: (6.0 -0.33334) (-3.6 0.2)
Average Distance: 3.6928

Clustering: (1 2 3 4 5 6 7) (0 8 9 10 11 12 13 14 15)
Cluster Centers: (5.57143 0.0) (-4.33334 0.0)
Average Distance: 3.49115

Clustering: (0 1 2 3 4 5 6 7) (8 9 10 11 12 13 14 15)
Cluster Centers: (5.0 0.0) (-5.0 0.0)
Average Distance: 3.41421

Clustering: (0 1 2 3 4 5 6 7) (8 9 10 11 12 13 14 15)
No improvement.



Σύνοψη

- Ιεραρχικές μέθοδοι.

Στον παρελθόν είχα καθιερωθεί οι Ward's και οι μέθοδος μέσης σύνδεση (average linkage) ως οι πιο δημοφιλείς ιεραρχικές μέθοδοι.

Οι ιεραρχικές μέθοδοι έχουν το πλεονέκτημα ότι υλοποιούνται γρήγορα και δίνουν τα αποτελέσματα σε λιγότερο υπολογιστικό χρόνο. Βέβαια καθώς βελτιώνεται η δύναμη των υπολογιστών, οι προσωπικοί υπολογιστές μπορούν να επεξεργαστούν μεγάλες βάσεις δεδομένων σχετικά εύκολα.

Οι ιεραρχικές μέθοδοι δεν προορίζονται για την ανάλυση μεγάλων δειγμάτων. Καθώς το μέγεθος του δείγματος αυξάνεται, αυξάνεται δραματικά και η ποσότητα ηλεκτρονικής μνήμης που χρειάζονται.

Π.χ. ένα δείγμα 400 παρατηρήσεων χρειάζεται μέγεθος μνήμης για 80000 ομοιογένειες. Αυτές αυξάνονται στις 125000 όταν το δείγμα αυξηθεί στις 500 παρατηρήσεις.

Αυτές οι ανάγκες υπερβαίνουν τις δυνατότητες των περισσοτέρων προσωπικών υπολογιστών και έτσι σε κάποιοες περιπτώσεις οι ιεραρχικές μέθοδοι δεν μπορούν να εφαρμοστούν.

Σε τέτοιες περιπτώσεις μπορούμε να μείωσουμε το μέγεθος του δείγματος χρησιμοποιώντας ένα μόνο μέρος του που έχει επιλεγεί με τυχαία δειγματοληψία. Όμως το ερώτημα που εγείρεται είναι κατά πόσο το μικρό αυτό δείγμα μπορεί να αντιπροσωπεύσει το αρχικό.

Η ιεραρχικές μέθοδοι μπορεί να οδηγήσουν σε παραπλαπλανητικά αποτελέσματα καθώς ανεπιθύμητοι συνδιασμοί που δημιουργήθηκαν στα αρχικά στάδια, παρουσιάζονται μέχρι το τέλος της ανάλυσης και οδηγούν σε μη ρεαλιστικά αποτελέσματα. Πέραν αυτού, ανησυχία προκαλεί η μεγάλη επίδραση των ακραίων παρατηρήσεων ιδίως στη μέθοδο πλήρους συνδέσης. Για αποφυγή του φενομένου συνίσταται η επανάληψη της ανάλυσης πολλές φορές, όπου κάθε φορά θα διαγράφονται προβληματικές παρατηρήσεις ή ακραίες τιμές.

Βέβαια πρέπει να λαμβάνεται υπόψη ότι η διαγραφή παρατηρήσεων, ακόμη και αυτές που δεν είναι ακραίες, μπορεί να διαστρεβλώσει τα αποτελέσματα. Έτσι η διαγραφή παρατηρήσεων πρέπει να γίνεται με σύνεση.

- Μη-ιεραρχικές μέθοδοι.

  Οι μη-ιεραρχικές μέθοδοι γίνονται όλο και πιο αποδεκτές και η εφαρμογή τους εξαπλώνεται ραγδαία. Όμως η εφαρμογή τους εξαρτάται από την επιλογή των κεντροειδών. Αυτό γίνεται είτε βάση πρακτικών κανόνων είτε βάση θεωρητικού υπόβαθρου.

  Οι μη-ιεραρχικές μέθοδοι πλεονεκτούν έναντι των ιεραρχικών. Τα αποτελέσματα είναι λιγότερο επιρρεπή σε ακραίες τιμές, στις μεθόδους μέτρησης της απόστασης και στα λανθασμένα ή άχρηστα δεδομένα. Τα πλεονεκτήματα αυτά είναι εμφανή μόνο όταν χρησιμοποιούνται μη-τυχαία κεντροειδή σημεία. Στην αντίθετη περίπτωση, δηλαδή χρήση τυχαίων κεντροειδών, οι μη-ιεραρχικές μέθοδοι είναι κατώτερες των ιεραρχικών. Διαφορετικά κεντροειδή μπορούν να οδηγήσουν σε διαφορετικές ομάδες.

Ποιες ομάδες όμως πρέπει να επιλεγούν; Η σωστότερη δομή μπορεί να επιλεγεί μόνο με ανάλυση και επαλήθευση.

- Συνδυασμός μεθόδων.

  Εναλλακτικά μπορούν να χρησιμοποιηθούν και οι δύο μέθοδοι (ιεραρχική και μη-ιεραρχική). Έτσι μπορούμε να επωφεληθούμε από τα πλεονεκτήματα και των δύο μεθόδων. Πρώτον, οι ιεραρχικές μέθοδοι μπορούν να προσδιορίσουν τον αριθμό των ομάδων, τα κέντρα των ομάδων και να προσδιορίσουν τις εμφανείς ακραίες παρατηρήσεις. Αφού διαγραφούν οι ακραίες παρατηρήσεις, οι υπόλοιπες παρατηρήσεις μπορούν να ομαδοποιηθούν βάση μιας μη-ιεραρχικής μεθόδου. Σε τέτοια περίπτωση τα κέντρα των ομάδων, που απορρέουν από την ιεραρχική μέθοδο, μπορούν να χρησιμοποιηθούν ως τα αρχικά κεντροειδή σημεία. Με αυτό τον τρόπο, τα πλεονεκτήματα των ιεραρχικών μεθόδων ενισχύονται από την ικανότητα των μη-ιεραρχικών μεθόδων *καθορίζουν με ακρίβεια* τα αποτελέσματα. Αυτό πραγματοποιείται με την ευχέρεια αλλαγής των μελών μιας ομάδας.