ΣΥΣΤΑΔΟΠΟΙΗΣΗ

Αναπλ. Καθηγ. Στελιος Ζήμερας
Τμημα Στατιστικης και Αναλογιστικων – Χρηματοοικονομικων Μαθηματικων
Πανεπιστημιο Αιγαιου
Σαμος

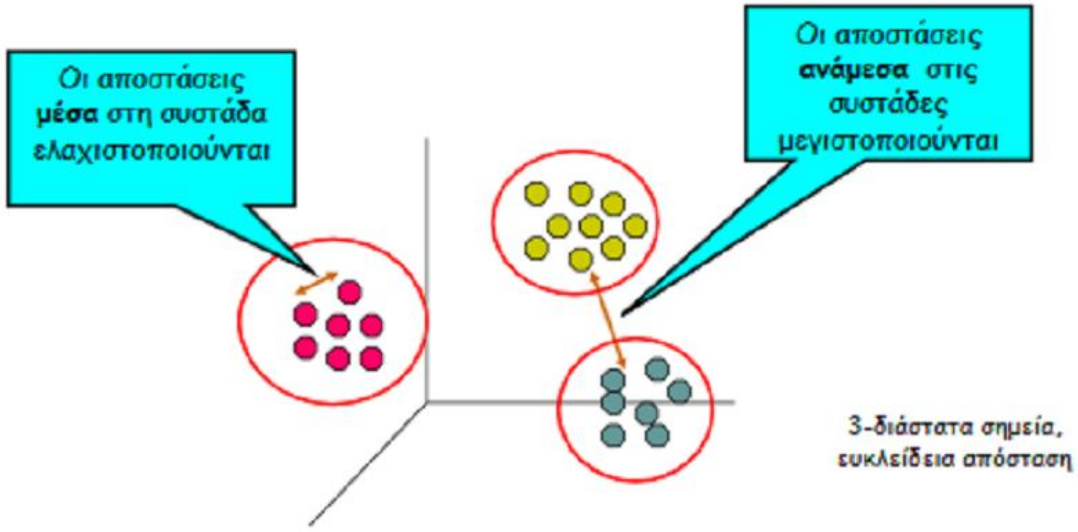
2021

Εισαγωγή

- ανάλυση των συστάδων
- Στόχος είναι η δημιουργία ενός μοντέλου, το οποίο να μπορεί να κατηγοριοποιήσει νέα δεδομένα σε κάποια από τις προϋπάρχουσες κλάσεις.
- Δοθέντων κάποιων δεδομένων χωρίς κλάσεις, οι αλγόριθμοι συσταδοποίησης ομαδοποιούν τα δεδομένα σε συστάδες, έτσι ώστε εγγραφές, οι οποίες ανήκουν στην ίδια συστάδα, να έχουν όμοια ή παραπλήσια χαρακτηριστικά.

Εισαγωγή

- Να παράγει ένα σύνολο από ομάδες με υψηλή εντός των ομάδων ομοιότητα ενώ παράλληλα να διατηρείται χαμηλή η ομοιότητα μεταξύ των διαφόρων ομάδων



Εισαγωγή

- Στο πρόβλημα της συσταδοποίησης μας δίνεται ένα σύνολο δεδομένων, χωρίς τις αντίστοιχες κλάσεις ή ετικέτες και χρειαζόμαστε κάποιον αλγόριθμο, ο οποίος θα ομαδοποιήσει αυτόματα τα δεδομένα σε συστάδες.
- Οι συστάδες που δημιουργούνται θέλουμε να διαχωρίζουν ορθά τα δεδομένα. Αυτό πρακτικά σημαίνει ότι μια συστάδα θέλουμε να απαρτίζεται από αντικείμενα, όπου κάθε αντικείμενο είναι πιο κοντά σε κάθε άλλο αντικείμενο της ίδιας συστάδας απ’ ότι σε κάποιο άλλο αντικείμενο διαφορετικής συστάδας.

Μέθοδοι

Αλγόριθμος k-means

Ο αλγόριθμος k-means ξεκινάει με k τυχαία σημεία, τα οποία ονομάζονται κεντροειδή της συστάδας και δηλώνουν το κέντρο βάρους της συστάδας. Το k υποδηλώνει σε πόσες συστάδες θέλουμε ο αλγόριθμος να δημιουργήσει. Ο αλγόριθμος εκτελεί επαναληπτικά δύο βήματα. Το πρώτο βήμα αφορά την ανάθεση σε κάποια συστάδα, ενώ το δεύτερο βήμα αφορά τον επαναπροσδιορισμό και τη μετατόπιση του κεντροειδούς κάθε συστάδας.

Αλγόριθμος k-means

- Στο πρώτο βήμα ο αλγόριθμος εξετάζει κάθε δείγμα σε σχέση με τα κεντροειδή των συστάδων. Με χρήση κάποιου μέτρου απόστασης, αναθέτει το εξεταζόμενο δείγμα στη συστάδα, της οποίας το κεντροειδές είναι το πλησιέστερο ως προς το συγκεκριμένο δείγμα.
- Στο δεύτερο βήμα, παίρνοντας τον μέσο όρο των δειγμάτων κάθε συστάδας, επανυπολογίζονται τα κεντροειδή της κάθε συστάδας, ώστε το κεντροειδές να είναι πιο αντιπροσωπευτικό στην πρόσφατα διαμοφωμένη συστάδα.

Αλγόριθμος k-means

- Ο αλγόριθμος εκτελεί επαναληπτικά αυτά τα δύο βήματα, μέχρις ότου τα κεντροειδή των συστάδων να με-τατοπίζονται ελάχιστα και σε απόσταση μικρότερη από κάποια δοθείσα τιμή κατωφλίου. Ως εναλλακτικό κρι-τήριο τερματισμού του αλγορίθμου μπορεί να χρησιμοποιηθεί και ο αριθμός επαναλήψεων του αλγορίθμου.

Αρχικοποίησε τυχαία τα k κεντροειδή των συστάδων $\mu_1, \mu_2, \ldots, \mu_k$.

Επανέλαβε{

Εξέτασε κάθε δείγμα και ανέθεσε το στη συστάδα με το πλησιέστερο κεντροειδές $\left(\mathbf{\min} \left| \mathbf{x}^{(i)} - \mu_k \right|^2 \right)$

Επανυπολόγισε τα κεντροειδή υπολογίζοντας το μέσο όρο των δειγμάτων της συστάδας

}

Αλγόριθμος k-means

Δοθέντος ενός συνόλου N παρατηρήσεων, ζητείται ο κωδικοποιητής C που αντιστοιχίζει αυτές τις παρατηρήσεις στα K clusters με τέτοιο τρόπο ώστε, μέσα σε κάθε συστάδα, ό μέσος όρος του μέτρου ανομοιότητας των αντιστοιχισμένων παρατηρήσεων ως προς το μέσο (mean) του cluster να ελαχιστοποιείται.

Έστω ότι το $\{x_i\}_{i=1}^{N}$ συμβολίζει ένα σύνολο πολυδιάστατων παρατηρήσεων το οποίο πρόκειται να διαμεριστεί σε ένα προτεινόμενο σύνολο K clusters, όπου το K είναι μικρότερο από τον αριθμό παρατηρήσεων N. έστω επίσης η σχέση:

$$j = C(i), \qquad i = 1,2, \dots, N \qquad (1)$$

που συμβολίζει ένα αντιστοιχιστή «πολλά-προς-ενα», που αποκαλείται κωδικοποιητής, ο οποίος αντιστοιχίζει την i-οστή παρατήρηση x_i στο j-οστό cluster σύμφωνα με έναν κανόνα που θα ορίσουμε. Για να υλοποιηθεί αυτή την κωδικοποίηση, χρειάζεται ένα μέτρο ομοιότητας μεταξύ κάθε ζεύγους διανυσμάτων x_i και $x_{i'}$, το οποίο συμβολίζεται ως $d(x_i,x_{i'})$. Όταν το μέτρο $d(x_i,x_{i'})$ είναι επαρκώς μικρό, αμφότερα τα x_i και $x_{i'}$ αντιστοιχίζονται στο ίδιο cluster. Διαφορετικά, αντιστοιχίζονται σε διαφορετικά clusters.

Αλγόριθμος k-means

Για τη βελτιστοποίηση της διαδικασίας συσταδοποίησης, εισάγουμε τη συνάρτηση κόστους :

$$J(C) = \frac{1}{2}\sum_{j=1}^{K}\sum_{C(i)=j}\sum_{C(i')=j} d(x_i, x_{i'}) \quad (2)$$

Για ένα προκαθορισμένο K, το ζητούμενο είναι να βρεθεί ο κωδικοποιητής C(i)=j για τον οποίο η συνάρτηση κόστους J(C) ελαχιστοποιείται. Επισημαίνεται ότι ο κωδικοποιητής C είναι άγνωστος, και σε αυτό οφείλεται η λειτουργική εξάρτηση ης συνάρτησης κόστους J από το C.

Στον αλγόριθμο K-means, χρησιμοποιεί το τετράγωνο της Ευκλείδειας νόρμας για τον ορισμό του μέτρου ομοιότητας μεταξύ των παρατηρήσεων x_i και $x_{i'}$, όπως αποδεικνύει η σχέση

$$d(x_i, x_{i'}) = \|x_i - x_{i'}\| \quad (3)$$

Άρα αντικαθιστώντας την Εξ.(3) στην Εξ.(2) προκύπτει :

$$J(C) = \frac{1}{2}\sum_{j=1}^{K}\sum_{C(i)=j}\sum_{C(i')=j} \|x_i - x_{i'}\| \quad (4)$$

Αλγόριθμος k-means

Δυο σημαντικά σημεία είναι :

1. Το τετράγωνο της Ευκλείδειας απόστασης μεταξύ των παρατηρήσεων x_i και $x_{i'}$ είναι συμμετρικό, δηλαδή:

$$\|x_i - x_{i'}\|^2 = \|x_{i'} - x_i\|^2$$

2. Το εσωτερικό άθροισμα στην Εξ.(4) ερμηνεύεται ως εξής: Για ένα δεδομένο x_i, ο κωδικοποιητής C αντιστοιχίζει στο cluster j όλες τις παρατηρήσεις $x_{i'}$ που είναι πλησιέστερα στην x_i. Εκτός από ένα συντελεστή κλιμάκωσης, το άθροισμα των παρατηρήσεων $x_{i'}$ που αντιστοιχίζεται είναι μια εκτίμηση του μέσου διανύσματος που αφορά το cluster j. Ο εν λόγω συντελεστής κλιμάκωσης είναι $1/N_j$, όπου N_j είναι ο αριθμός των σημείων δεδομένων μέσα στη συστάδα j.

Με βάση αυτά τα σημεία η Εξ.(4) μπορεί να απλοποιηθεί σε :

$$J(C) = \sum_{j=1}^{K} \sum_{C(i)=j} \|x_i - \hat{\mu}_j\|^2 \qquad (5)$$

όπου το $\hat{\mu}_j$ συμβολίζει το εκτιμώμενο μέσο διάνυσμα που σχετίζεται με τη συστάδα j. Ουσιαστικά, το μέσο $\hat{\mu}_j$ μπορεί να θεωρηθεί κέντρο της συστάδας j.

Αλγόριθμος k-means

Βήμα 1: Για ένα δεδομένο κωδικοποιητή , η συνολική διακύμανση συστάδας ελαχιστοποιείται ως προς το σύνολο μέσων συστάδας $\{\hat{\mu}_j\}_{j=1}^{K}$. Δηλαδή εκτελείται η ακόλουθη ελαχιστοποίηση:

$$min_{\{\hat{\mu}_j\}_{j=1}^{K}} \sum_{j=1}^{K} \sum_{C(i)=j} \|x_i - \hat{\mu}_j\|^2 \qquad για\ δεδομένο\ C$$

Βήμα 2: Αφού υπολογιστούν οι βελτιστοποιημένοι μέσοι των clusters $\{\hat{\mu}_j\}_{j=1}^{K}$ στο βήμα 1, στη συνέχεια βελτιστοποιούμε τον κωδικοποιητή ως εξής:

$$C(i) = \arg min_{1 \le j \le K} \|x_i - \hat{\mu}_j\|^2$$

Ξεκινώντας από κάποια αρχική επιλογή του κωδικοποιητή C, ο αλγόριθμος εναλλάσσεται μεταξύ αυτών των δυο βημάτων μέχρι να μην υπάρχει περαιτέρω αλλαγή στις αντιστοιχίσεις των συστάδων.

Μετρικές αποστάσεων

- στόχος της συσταδοποίησης είναι η ομαδοποίηση των αντικειμένων με βάση κάποιο μέτρο ομοιότητας.
- Ως ομοιότητα ορίζεται μια αριθμητική μέτρηση για το πόσο όμοια είναι δυο αντικείμενα ενώ ως μη ομοιότητα ορίζεται μια αριθμητική μέτρηση για το πόσο διαφορετικά είναι δυο αντικείμενα. Η ομοιότητα ή μη ομοιότητα μεταξύ δυο αντικειμένων υπολογίζεται συνήθως σύμφωνα με κάποια συνάρτηση απόστασης ανάμεσα στα δύο αντικείμενα.

Μετρικές αποστάσεων

Έστω δύο διανύσματα x, y του n-διάστατου χώρου R_n με $x = \{x_1,....,x_n\}$ και $y = \{y_1,....,y_n\}$.

$$dist(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$ **Ευκλείδεια Απόσταση**

$$dist(x,y) = \sum_{i=1}^{n}|x_i - y_i|$$ **Απόσταση Manhattan**

$$dist\,(x,y) = \sum_{i=1}^{n}[(|x_i - y_i|)^q]^{1/q}$$ **Απόσταση Minkowsky**

q=2 προκύπτει η ευκλείδεια και για q=1 η Manhattan.

$$dist(x,y) = \sqrt{(x-y)C^{-1}(x-y)^T}$$ **Απόσταση Mahalanobis**

C είναι ο αντίστοιχος covariance matrix

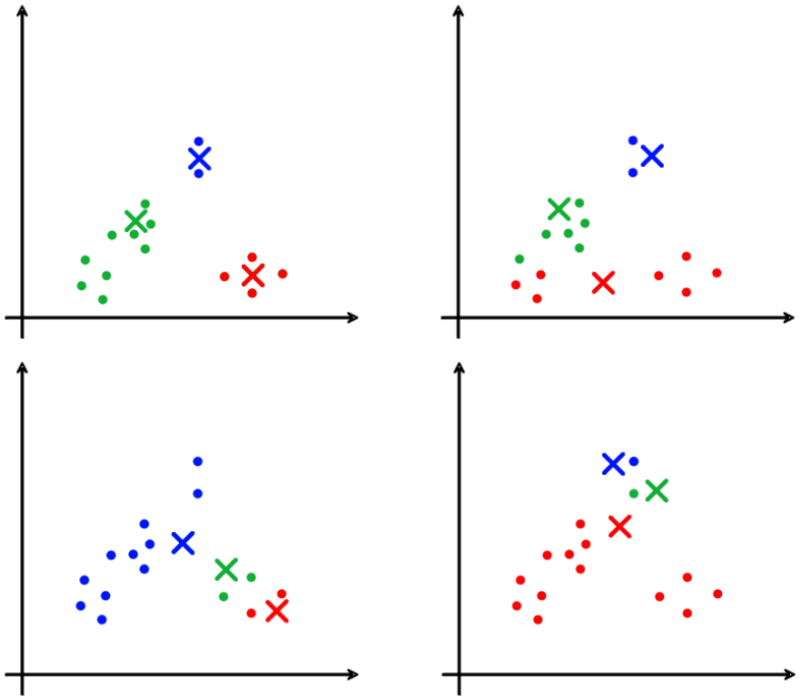
Αλγόριθμος k-means

- **Τυχαία Αρχικοποίηση Κεντροειδών**

Το πρώτο βήμα του αλγορίθμου k-means είναι η τυχαία αρχικοποίηση των k κεντροειδών των συστάδων.

Αρκετές φορές μια «κακή» αρχικοποίηση μπορεί να οδηγήσει σε κακής ποιότητας συστάδες στην πορεία

καλή περίπτωση



λιγότερο καλή

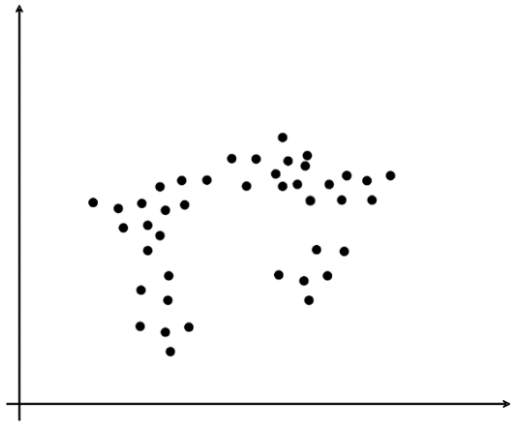
Αλγόριθμος k-means

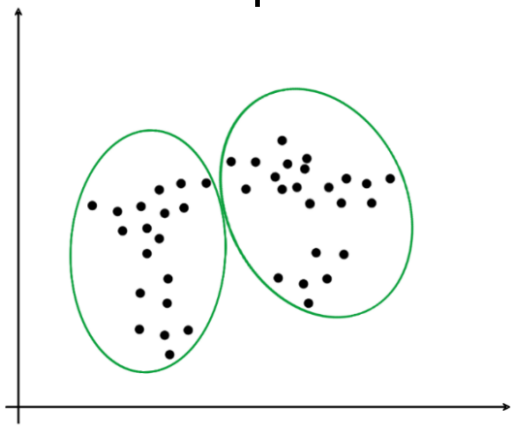
Επιλογή του Αριθμού Συστάδων

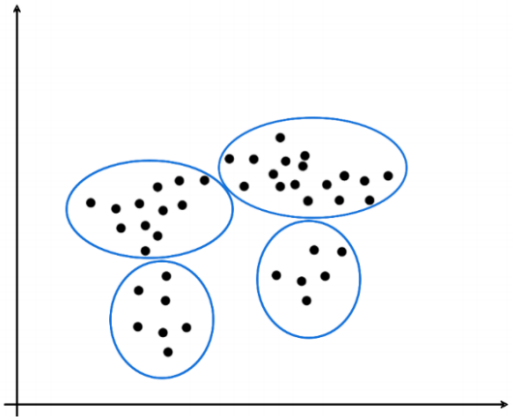
Ένα από τα μειονεκτήματα του αλγορίθμου k-means είναι το γεγονός ότι δεν υπάρχει κάποιος αυτοματοποιημένος τρόπος επιλογής του k, δηλαδή του αριθμού των συστάδων. Ο αριθμός των συστάδων δίνεται ως είσοδος από τον χρήστη και η επιλογή του σωστού αριθμού επαφίεται στη δική του γνώση και εμπειρία. Συνεπώς, η διαδικασία επιλογής του αριθμού συστάδων, ενδεχομένως, να απαιτήσει την εξερεύνηση και μελέτη των δεδομένων, για παράδειγμα, μέσα από οπτικοποιήσεις, προκειμένου να καταλήξουμε στον σωστό αριθμό συστάδων.

Αλγόριθμος k-means

Η εκτίμηση του αριθμού των clusters σε ένα σύνολο δεδομένων, η μεταβλητή που συμβολίζεται συνήθως με K







$$K \approx \sqrt{\frac{N}{2}}\ , \qquad N \text{ ο αριθμός των δεδομένων}$$

Αλγόριθμος k-means

Σε αυτή την εργασία χρησιμοποιήθηκε η «μέθοδος του αγκώνα» (Elbow Method), η οποία λειτουργεί ως εξής :

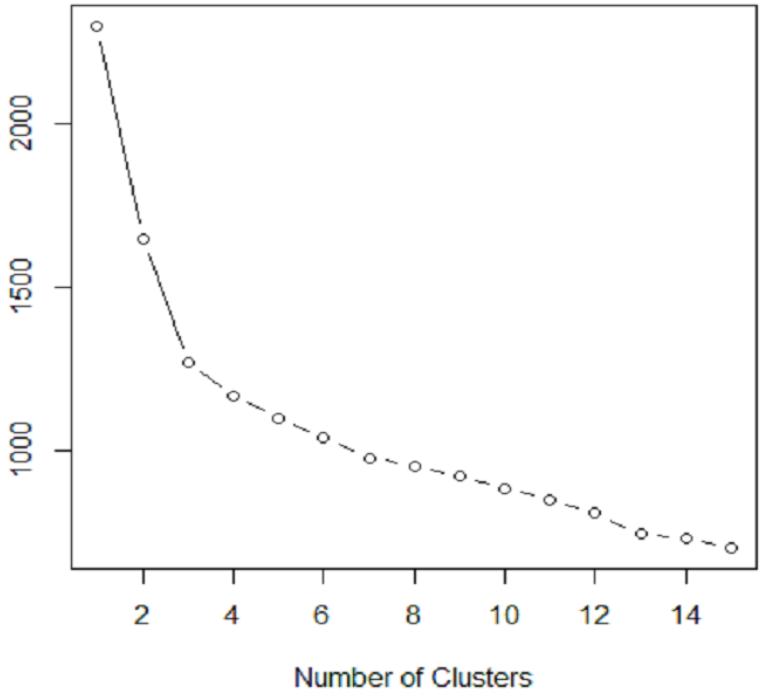
Εκτελούμε τον αλγόριθμο συσταδοποίησης (συγκεκριμένα εδώ τον αλγόριθμο K-means) για πολλές τιμές του K, από πολύ μικρές έως πολύ μεγάλες. Για κάθε εκτέλεση, υπολογίζεται ένας δείκτης εκτίμησης των clusters. Εν προκειμένω, αυτός ο δείκτης είναι το άθροισμα του τετραγώνου των σφαλμάτων (SSE), που ορίζεται ως το άθροισμα των τετραγώνων των αποστάσεων ανάμεσα σε κάθε δεδομένο και τον κεντροειδή του cluster στον οποίο ανήκει. Δηλαδή :

$$SSE = \sum_{i=1}^{K} \sum_{x \in c_i} dist(x, c_i)^2$$

Στη συνέχεια αν κανείς σχεδιάσει τη γραφική παράσταση του SSE συναρτήσει του αριθμού των clusters K, θα παρατηρήσει ότι ο δείκτης SSE, δηλαδή το σφάλμα μειώνεται όσο το K αυξάνεται, το οποίο είναι αναμενόμενο με βάση τα όσα αναφέραμε παραπάνω. Θα παρατηρήσει επίσης ότι για κάποιο K, η γραφική παράσταση θα παρουσιάζει μια έντονη γωνία, εξ' ου και το όνομα της μεθόδου. Αυτή η τιμή του K είναι που επιλέγεται ως βέλτιστη για την εκτέλεση του αλγορίθμου. Ακολουθεί ένα παράδειγμα μιας τέτοια γραφικής παράστασης που δείχνει αυτή την απότομη μεταβολή:

Αλγόριθμος k-means

- Δυστυχώς, για την επιλογή του αριθμού των συστάδων δεν υπάρχει κάποιος γενικός κανόνας, ο οποίος να λειτουργεί εγγυημένα και για όλες τις περιπτώσεις. Ένα απλό και πρακτικό τέχνασμα, το οποίο μπορεί να βοηθήσει σε ορισμένες περιπτώσεις, είναι «ο κανόνας του αγκώνα»



Ιεραρχικοί Αλγόριθμοι Συσταδοποίησης

- Οι ιεραρχικοί αλγόριθμοι διακρίνονται σε δύο υποκατηγορίες: τους συσσωρευτικούς και τους διαιρετικούς.
- Οι αλγόριθμοι μπορούν να αναπαρασταθούν πλήρως με δενδρογράμματα, δηλαδή με δενδρικά διαγράμματα, τα οποία παρουσιάζουν τη διάταξη των συστάδων που δημιουργήθηκαν από την ιεραρχική συσταδοποίηση.
- Ουσιαστικά, κάθε επίπεδο ενός δενδρογράμματος ορίζει ένα βήμα του αλγορίθμου. Το βασικό πλεονέκτημα των ιεραρχικών αλγορίθμων είναι ότι δεν χρειάζεται να υποθέσουμε ένα συγκεκριμένο αριθμό συστάδων, αφού οποιοσδήποτε αριθμός μπορεί να επιτευχθεί, απλά κόβοντας το δενδρόγραμμα στο κατάλληλο επίπεδο.

Ιεραρχικοί Αλγόριθμοι Συσταδοποίησης

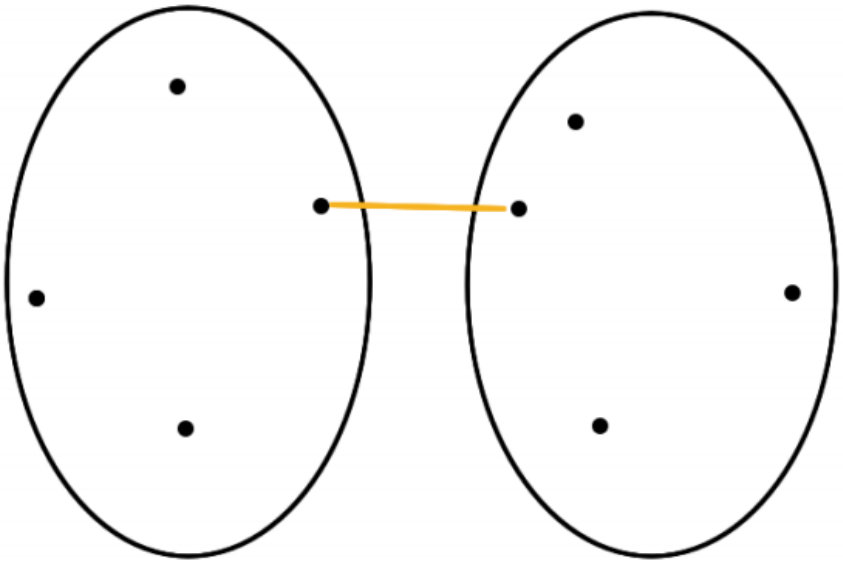
Ορισμός Απόστασης Συστάδων

Οι κυριότεροι είναι οι εξής:

- Ελάχιστης απόστασης ή απλού συνδέσμου (single link).
- Μέγιστης απόστασης ή πλήρους συνδέσμου (complete link).
- Μέσου όρου της συστάδας (group average).
- Απόσταση κεντρικών σημείων.
- Μέθοδος του Ward.

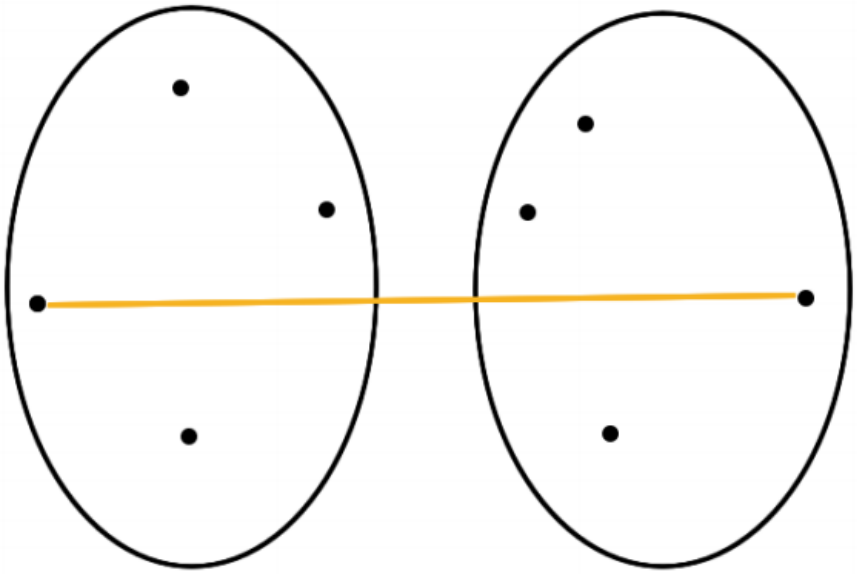
Ορισμός Απόστασης Συστάδων

- Με βάση το κριτήριο απλού συνδέσμου, η ομοιότητα μεταξύ δύο συστάδων βασίζεται στα δύο πιο όμοια (πιο γειτονικά) σημεία στις διαφορετικές συστάδες δηλαδή στα σημεία με την ελάχιστη απόσταση μεταξύ τους. Είναι γνωστή και ως **μέθοδος συσταδοποίησης κοντινότερου γείτονα.** Τα προτερήματα αυτής της μεθόδου είναι ότι δημιουργούνται συνεχόμενες συστάδες, ενώ μπορεί να χειριστεί μη ελλειπτικά σχήματα. Το βασικό μειονέκτημα είναι η ευαισθησία στον θόρυβο και στις ακραίες τιμές (outliers).



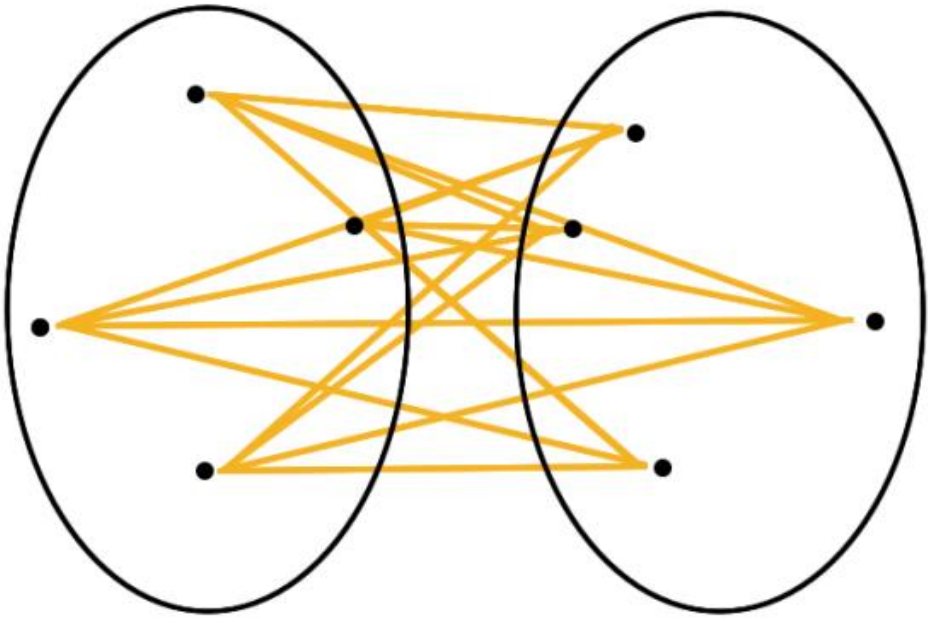
Ορισμός Απόστασης Συστάδων

- Με βάση το κριτήριο πλήρους συνδέσμου, η ομοιότητα μεταξύ δύο συστάδων βασίζεται στα δύο πιο ανόμοια (πιο απόμακρα) σημεία στις διαφορετικές συστάδες δηλαδή στα σημεία με τη μέγιστη απόσταση μεταξύ τους. Το βασικό πλεονέκτημα αυτού του τρόπου σύνδεσης είναι η μικρή ευαισθησία στον θόρυβο και στις ακραίες τιμές (outliers). Τα μειονεκτήματα που έχει είναι ότι τείνει να διασπά μεγάλες συστάδες και ότι οδηγεί, συνήθως, σε κυκλικά σχήματα.



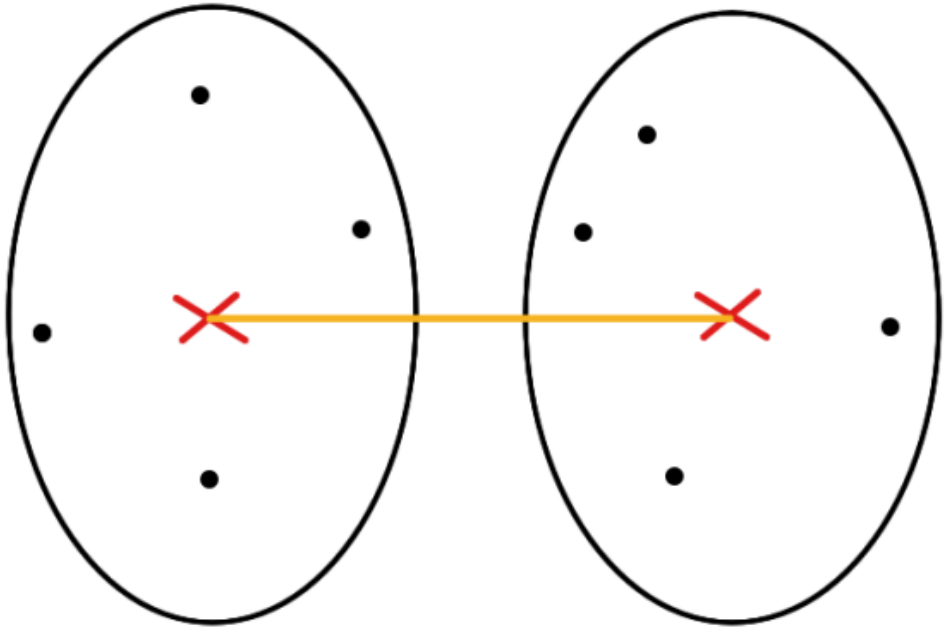
Ορισμός Απόστασης Συστάδων

- Ο μέσος όρος συστάδων είναι ουσιαστικά η μέση τιμή των αποστάσεων μεταξύ κάθε πιθανού ζεύγους μεταξύ των σημείων των δύο συστάδων. Βρίσκεται κάπου ανάμεσα στην ελάχιστη και τη μέγιστη απόσταση. Έχει μικρότερη ευαισθησία σε θόρυβο και σε ακραίες τιμές (outliers), αλλά ευνοεί τις συστάδες με κυκλικό σχήμα.



Ορισμός Απόστασης Συστάδων

- Ή απόσταση κεντρικών σημείων είναι η απόσταση μεταξύ των κέντρων των συστάδων. Το πρόβλημα με αυτή την απόσταση είναι ότι δεν έχει μονότονη αύξηση. Έτσι, δύο συστάδες που συγχωνεύονται μπορεί να έχουν μικρότερη απόσταση από συστάδες, οι οποίες έχουν συγχωνευτεί σε προηγούμενα βήματα.



Ορισμός Απόστασης Συστάδων

- Η βασική ιδέα πίσω από τη μέθοδο του Ward είναι ότι η απόσταση μεταξύ δύο συστάδων, C_i και C_j, είναι ίση με το πόσο θα αυξηθεί το άθροισμα των τετραγώνων της απόστασης των στοιχείων της κάθε συστάδας από το αντίστοιχο κεντροειδές (της κάθε συστάδας) μετά τη συγχώνευση τους, C_{ij}, δηλαδή:

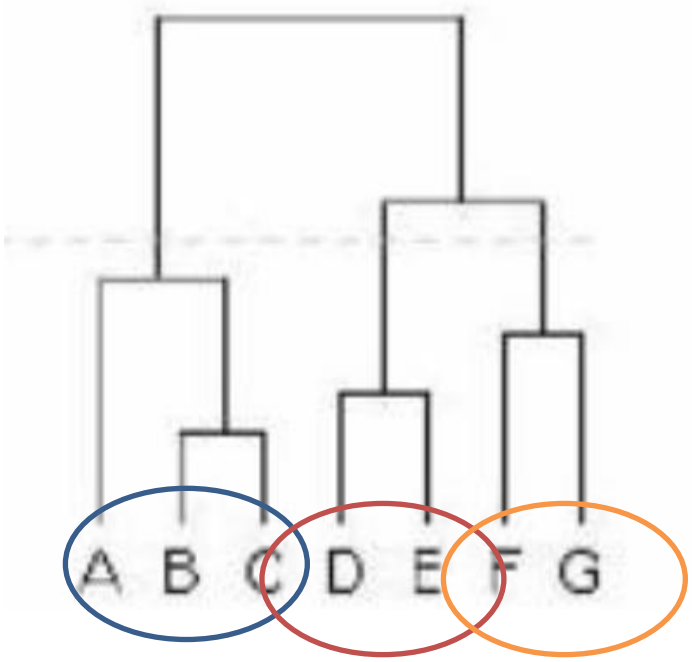
$$D_w(C_i, C_j) = \sum_{x \in C_i} (x - r_i)^2 + \sum_{x \in C_j} (x - r_j)^2 - \sum_{x \in C_{ij}} (x - r_{ij})^2$$

όπου r_i είναι το κεντροειδές της συστάδας C_i, r_j είναι το κεντροειδές της συστάδας C_j, και r_{ij} είναι το κεντροειδές της συστάδας C_{ij}, που προκύπτει από τη συγχώνευσή τους.

Ιεραρχική Συσταδοποίηση (Hierarchical Clustering)

- Στην ιεραρχική συσταδοποίηση τα στιγμιότυπα δίνονται με μορφή δενδρογράμματος. Στα δενδρογράμματα αυτά, επιλέγεται ένα επίπεδο που θα «κλαδευτούν». Το σημείο που θα κλαδευτεί κάποιο δενδρόγραμμα δείχνει τον αριθμό των clusters που θα προκύψουν καθώς και τα σημεία που περιέχει το κάθε cluster. Οι ιεραρχικές τεχνικές είναι είτε συσσωρευτικές (bottom-up), είτε διαιρετικές (top-down).

Ιεραρχική Συσταδοποίηση (Hierarchical Clustering)



Διαφαίνονται τρία clusters. Το πρώτο περιέχει τα σημεία A, B και C, το δεύτερο τα σημεία D και E, και το τρίτο τα σημεία F και G.

Ιεραρχική Συσταδοποίηση (Hierarchical Clustering)

Στη γενική της μορφή, η ιεραρχική συσταδοποίηση λειτουργεί ως εξής:

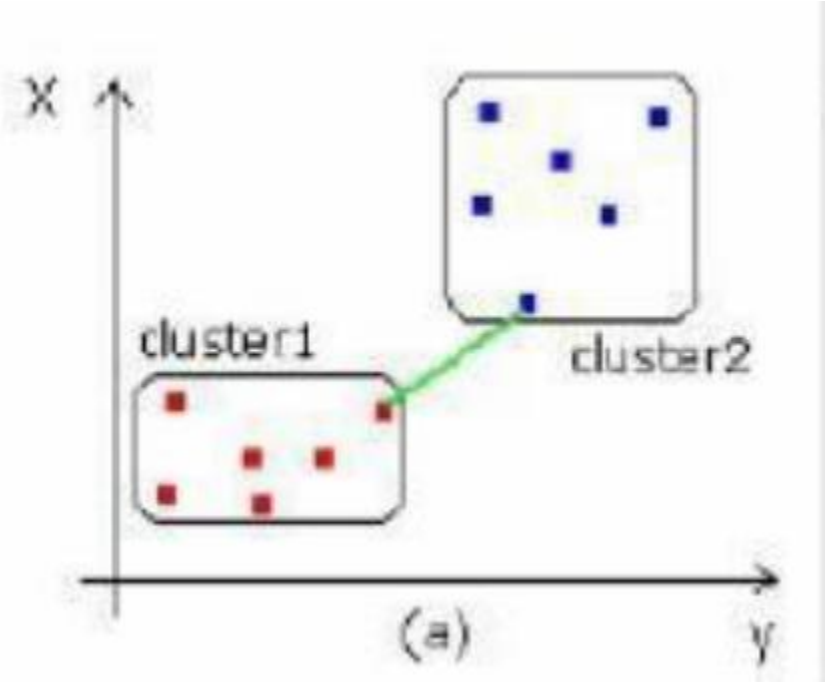
1. Αρχικά κάθε σημείο θεωρείται σαν μια ομάδα. Ν δηλαδή υπάρχει ένα σύνολο από Ν σημεία τα οποία πρέπει να ομαδοποιηθούν, τότε αρχικά υπάρχουν Ν ομάδες, που η καθεμία περιέχει ένα μόνο σημείο. Μετρώνται οι μεταξύ τους αποστάσεις.
2. Βρίσκεται το πιο κοντινό ζευγάρι ομάδων. Το ζευγάρι αυτό συγχωνεύεται σε ένα. Πλέον υπάρχει μια ομάδα λιγότερη.
3. Υπολογίζονται εκ νέου οι αποστάσεις των ομάδων μεταξύ τους.
4. Επαναλαμβάνονται τα βήματα 2 και 3 έως ότου και τα Ν σημεία να τοποθετηθούν σε μια και μοναδική ομάδα.
5. Τέλος, σχεδιάζεται το αντίστοιχο δενδρόγραμμα και επιλέγεται σε ποιο σημείο θα κλαδευτεί.

Ιεραρχική Συσταδοποίηση (Hierarchical Clustering)

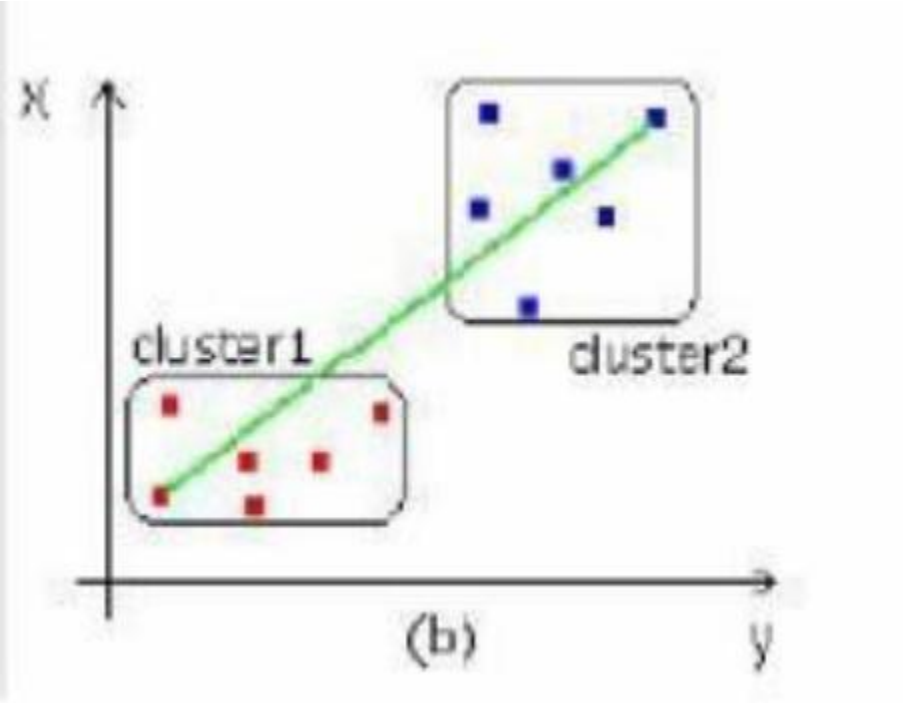
- Το βήμα 3 μπορεί να πραγματοποιηθεί με διάφορους τρόπους. Στην ιεραρχική συσταδοποίηση απλού συνδέσμου (simple-linkage), θεωρείται ως απόσταση μεταξύ δύο ομάδων η μικρότερη απόσταση μεταξύ όλων των ζευγών των προτύπων με στοιχεία κι από τις δύο ομάδες. Στην συσταδοποίηση ολοκληρωμένου συνδέσμου (complete-linkage), θεωρείται ως απόσταση μεταξύ δύο ομάδων η μεγαλύτερη απόσταση μεταξύ όλων των ζευγών των προτύπων με στοιχεία κι από τις δύο ομάδες. Στην συσταδοποίηση μέσου συνδέσμου (average-linkage) τέλος, θεωρείται ως απόσταση μεταξύ δύο ομάδων η μέση απόσταση μεταξύ όλων των ζευγών των προτύπων με στοιχεία κι από τις δύο ομάδες.

Ιεραρχική Συσταδοποίηση (Hierarchical Clustering)

απλού συνδέσμου



ολοκληρωμένου συνδέσμου



Παράδειγμα

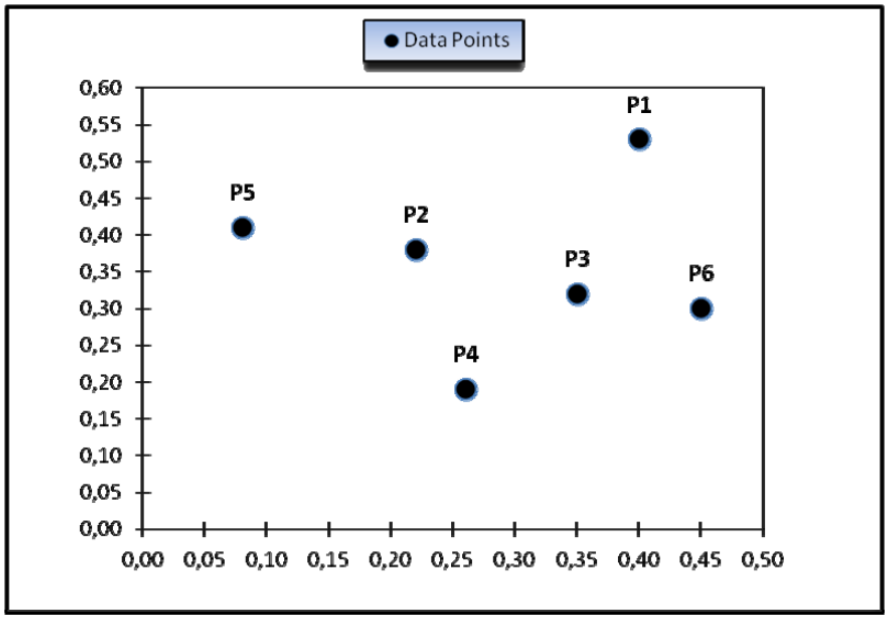
Πρόβλημα: Ας υποθέσουμε ότι η βάση δεδομένων D προς επεξεργασία δίνεται από τον πίνακα παρακάτω. Ακολουθώντας την μέθοδο απλού δεσμού να βρούμε τις συστάδες στην βάση δεδομένων D χρησιμοποιώντας ως μέτρο την Ευκλείδεια απόσταση.

	X	**Y**
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30

Παράδειγμα

Λύση:

Βήμα 1. Κάνουμε την γραφική παράσταση των σημείων στον ν-διάστατο χώρο (όπου ν είναι ο αριθμός των χαρακτηριστικών), στην περίπτωσή μας είναι 2, το x και y. Επομένως κάνουμε την γραφική παράσταση των σημείων p1, p2, … p6 στον δυσδιάστατο χώρο:



Παράδειγμα

Βήμα 2. Υπολογίζουμε την απόσταση κάθε σημείου από όλα τα άλλα χρησιμοποιώντας τον τύπο της ευκλείδειας απόστασης και τοποθετούμε τα νούμερα στον πίνακα αποστάσεως.

Ανακαλούμε ότι ο τύπος της ευκλείδειας απόστασης μεταξύ δύο σημείων i και j είναι:

$$d(i,j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \ldots + |x_{ip} - x_{jp}|^2}$$

όπου x_{i1} είναι η τιμή του χαρακτηριστικού 1 για i και x_{j1} είναι η τιμή του χαρακτηριστικού 1 για j, και τα λοιπά, για όσα χαρακτηριστικά έχουμε.

Παράδειγμα

$$d(p1,p2) = \sqrt{|x_{p1} - x_{p2}|^2 + |y_{p1} - y_{p2}|^2}$$

$$= \sqrt{|0.40 - 0.22|^2 + |0.53 - 0.38|^2}$$

$$= \sqrt{|0.18|^2 + |0.15|^2}$$

$$= \sqrt{0.0324 + 0.0225}$$

$$= \sqrt{0.0549}$$

$$= 0.2343$$

Παράδειγμα

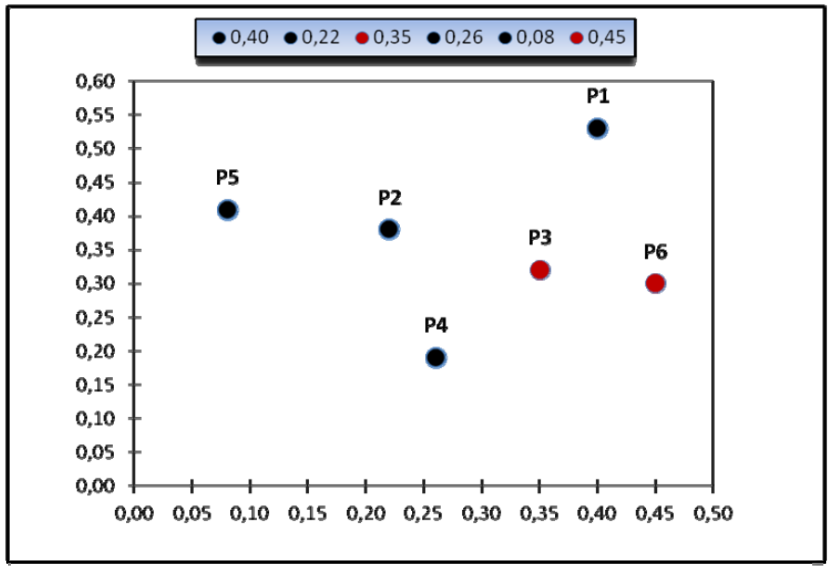
Πίνακας Αποστάσεως.

	p1	p2	p3	p4	p5	p6
p1	0	0.2343	0.2159	0.3677	0.3418	0.2354
p2	0	0	0.1432	0.1942	0.1432	0.2435
p3	0	0	0	0.1581	0.2846	0.1020
p4	0	0	0	0	0.2843	0.2195
p5	0	0	0	0	0	0.3860
p6	0	0	0	0	0	0

Παράδειγμα

Βήμα 3 Βρίσκουμε στον πίνακα τις δύο συστάδες με την μικρότερη απόσταση, και τις συγχωνεύουμε σε μία. Υπολογίζουμε ξανά τις τιμές για τον πίνακα αποστάσεως αφού αυτές οι δύο συστάδες είναι πλέον μία (δεν υφίστανται πλέον σαν μονάδες).

Κοιτώντας τον πίνακα αποστάσεως παραπάνω, βλέπουμε ότι τα σημεία p3 και p6 είναι αυτά τα οποία έχουν τη μικρότερη απόσταση μεταξύ όλων των άλλων - 0.1020 ,Επομένως συγχωνεύουμε αυτά τα δύο σε μία συστάδα και υπολογίζουμε ξανά τον πίνακα αποστάσεως.



Παράδειγμα

Πίνακας Αποστάσεως

	p1	p2	{p3,p6}	p4	p5
p1	0	0.2343	0.2159	0.3677	0.3418
p2	0	0	0.1432	0.1942	0.1432
{p3,p6}	0	0	0	0.1581	0.2846
p4	0	0	0	0	0.2843
p5	0	0	0	0	0

Αφού συγχωνεύσαμε τα σημεία (p3, p6) μαζί σε μία συστάδα θα έχουμε μία εγγραφή στον πίνακα αποστάσεως για αυτά.

Επομένως δε θα έχουμε πλέον τα σημεία p3 ή p6 ξεχωριστά. Θέλουμε έτσι λοιπόν να υπολογίσουμε την απόσταση της νέας μας συστάδας - (p3, p6) από όλα τα υπόλοιπα σημεία. Ανακαλούμε το γεγονός ότι στην μέθοδο απλού δεσμού η ομοιότητα μεταξύ δύο συστάδων καθορίζεται από την κοντινότερη απόσταση μεταξύ δύο οποιοδήποτε σημείων που ανήκουν στις δύο αυτές συστάδες. Επομένως, η απόσταση για παράδειγμα της (p3, p6) από το p1 υπολογίζεται ως εξής:

Παράδειγμα

dist((p3, p6), p1) = MIN (*dist*(p3, p1) , *dist*(p6, p1))

= MIN (0.2159 , 0.2354)

= 0.2159

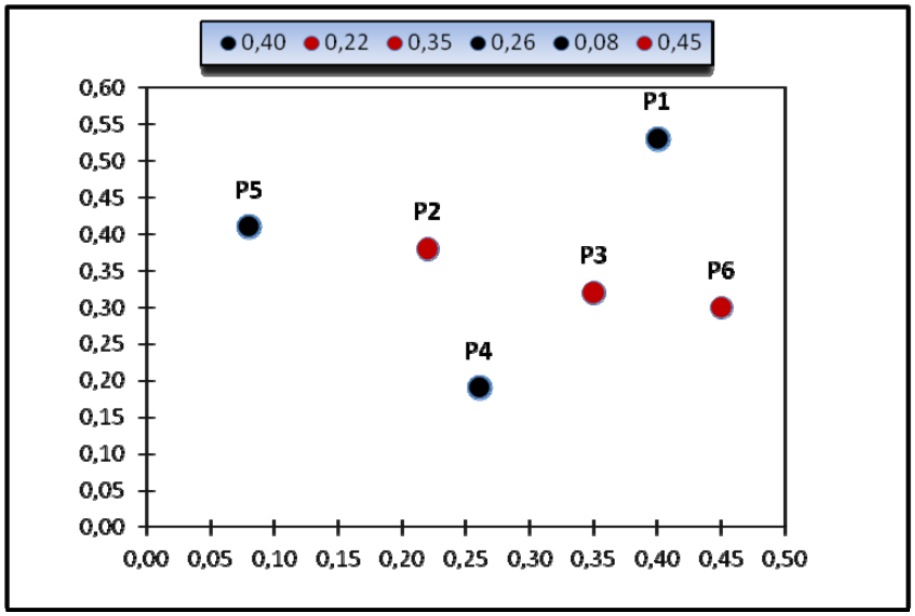
Πίνακας Αποστάσεως.

	p1	p2	p3	p4	p5	p6
p1	0	0.2343	0.2159	0.3677	0.3418	0.2354
p2	0	0	0.1432	0.1942	0.1432	0.2435
p3	0	0	0	0.1581	0.2846	0.1020
p4	0	0	0	0	0.2843	0.2195
p5	0	0	0	0	0	0.3860
p6	0	0	0	0	0	0

Παράδειγμα

Βήμα 4 Επανάληψη του βήματος 3 μέχρι όλες οι συστάδες (ή γενικά όλα τα σημεία) να συγχωνευθούν σε μία συστάδα.

α. Έτσι, κοιτώντας τον πιο πρόσφατο πίνακα αποστάσεως παραπάνω, βλέπουμε πως τα σημεία p2 και p5 είναι εκείνα με την μικρότερη απόσταση - 0.1432, επίσης όμως βλέπουμε ότι και τα σημεία p2 και (p3,p6) έχουν την ίδια απόσταση - 0.1432. Στην περίπτωση αυτή, μπορούμε να επιλέξουμε οποιοδήποτε από τα δύο. Ας επιλέξουμε τα p2 και (p3,p6).Τα συγχωνεύουμε σε μία συστάδα και υπολογίζουμε ξανά τον πίνακα αποστάσεως.



Παράδειγμα

Πίνακας Αποστάσεως

	p1	{p3,p6,p2}	p4	p5
p1	0	0.2159	0.3677	0.3418
{p3,p6,p2}	0	0	0.1581	0.1432
p4	0	0	0	0.2843
p5	0	0	0	0

Παράδειγμα

Αφού συγχωνεύσαμε τα p2 και (p3, p6) μαζί σε μία συστάδα έχουμε πλέον μία εγγραφή για αυτά στον πίνακα αποστάσεως.

Τα σημεία p2 και (p3,p6) δεν υπάρχουν πλέον σαν μονάδες. Επομένως πρέπει να υπολογίσουμε την απόσταση της νέας συστάδας από όλα τα υπόλοιπα σημεία/συστάδες. Η απόσταση μεταξύ των (p3, p6,p2) και p5 υπολογίζεται ως εξής:

dist((p3, p6,p2), p5) = MIN (*dist*(p3, p5) , *dist*(p6, p5), *dist*(p2, p5))

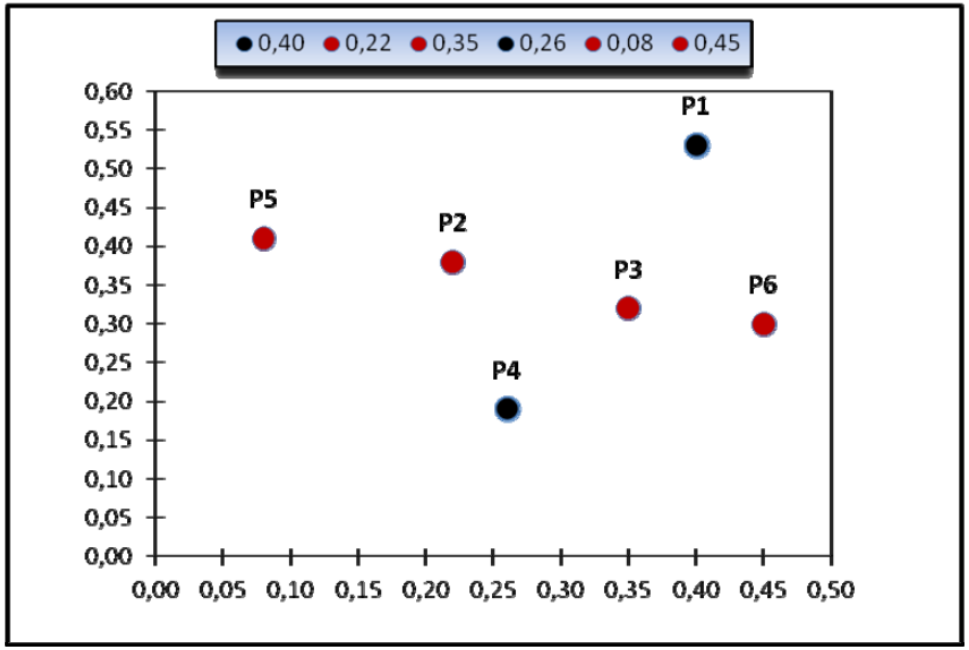
= MIN (0.2846 , 0.3860, 0.1432)

= 0.1432

Παράδειγμα

β. Αφού έχουμε και άλλες διαθέσιμες συστάδες για συγχώνευση, συνεχίζουμε να επαναλαμβάνουμε το βήμα 3.

Έτσι, βλέποντας τον πιο πρόσφατο πίνακα αποστάσεως παραπάνω διαπιστώνουμε ότι οι συστάδες (p3,p6,p2) και p5 έχουν τη μικρότερη απόσταση όλων - 0.1432. Επομένως συγχωνεύουμε αυτά τα δύο σε μία συστάδα και υπολογίζουμε ξανά τον πίνακα αποστάσεως.



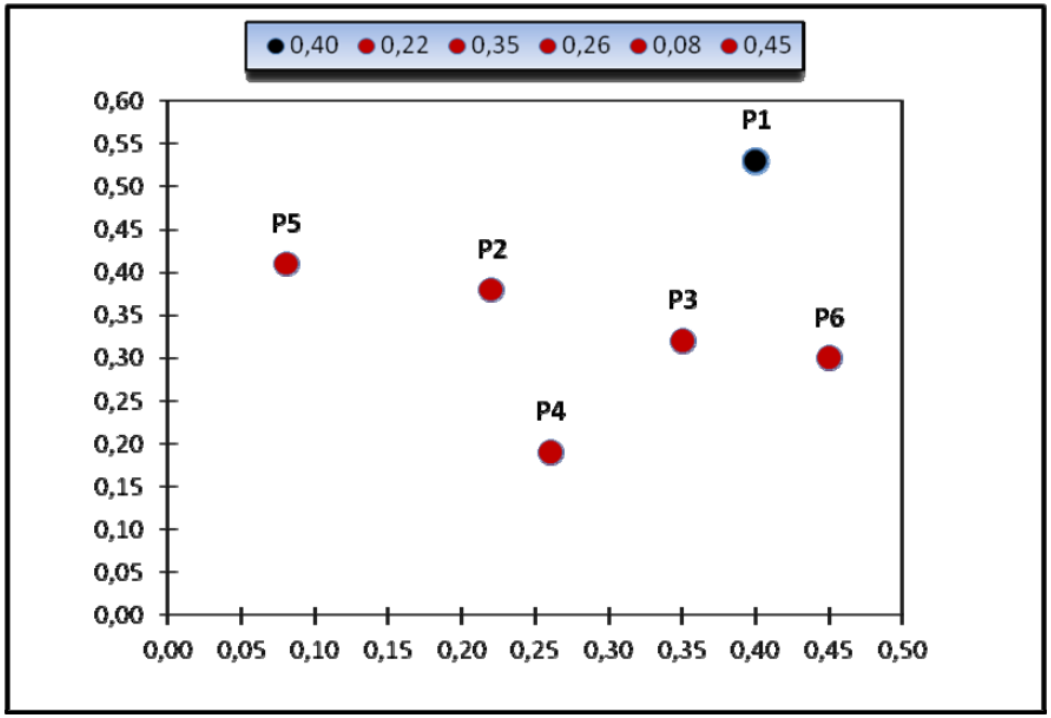
Παράδειγμα

	p1	{p3,p6,p2,p5}	p4
p1	0	0.2159	0.3677
{p3,p6,p2,p5}	0	0	0.1581
p4	0	0	0

ν. Αφού έχουμε και άλλες διαθέσιμες συστάδες για συγχώνευση, συνεχίζουμε να επαναλαμβάνουμε το βήμα 3.

Επομένως, βλέποντας τον πιο πρόσφατο πίνακα αποστάσεως παραπάνω, παρατηρούμε πως τα (p3, p6, p2, p5) και p4 έχουν τη μικρότερη απόσταση όλων - 0.1581 . Έτσι συγχωνεύουμε αυτά τα δύο σε μία συστάδα και υπολογίζουμε ξανά τον πίνακα αποστάσεως.

Παράδειγμα


0,40
0,22
0,35
0,26
0,08
0,45
0,60
0,55
0,50
0,45
0,40
0,35
0,30
0,25
0,20
0,15
0,10
0,05
0,00
0,00 0,05 0,10 0,15 0,20 0,25 0,30 0,35 0,40 0,45 0,50
P1
P2
P3
P4
P5
P6

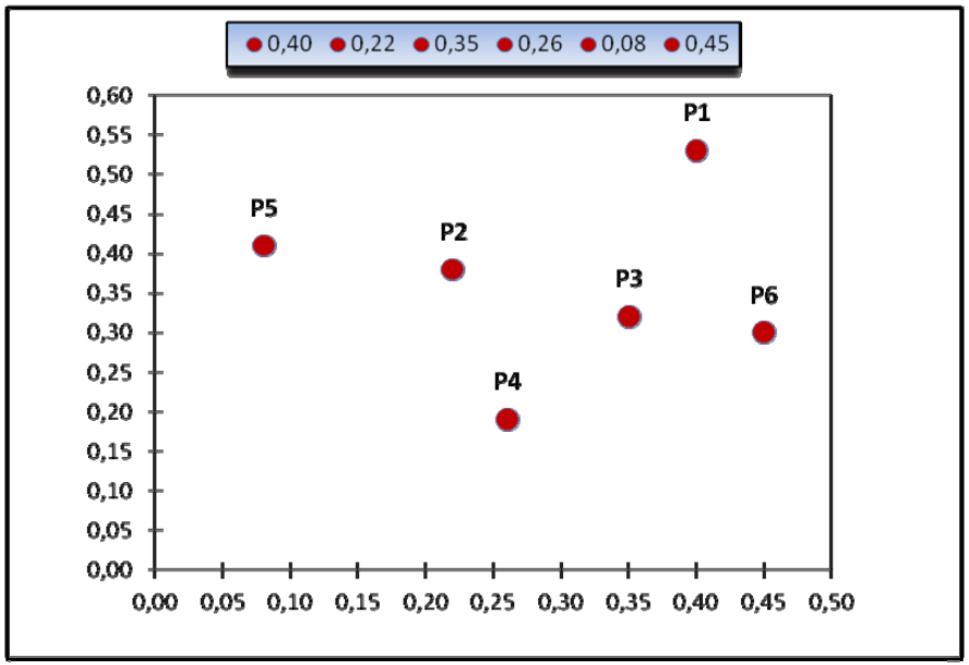
Παράδειγμα

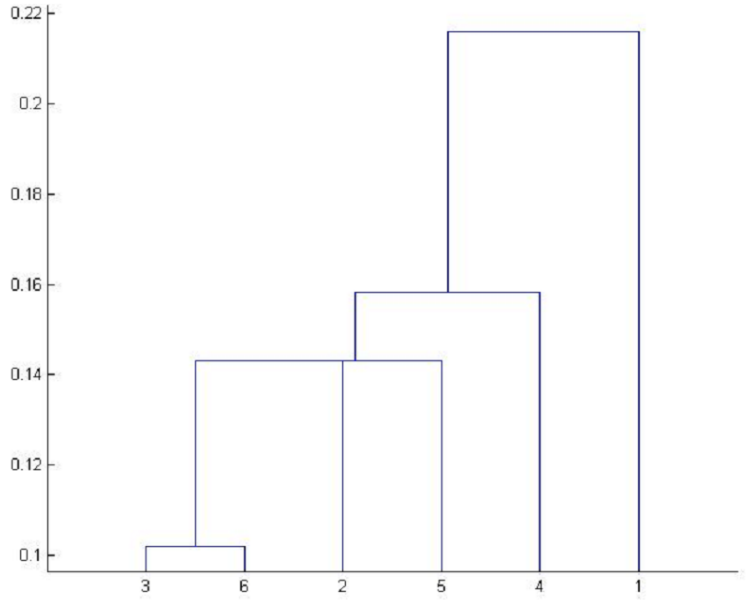
	p1	{p3,p6,p2,p5,p4}
p1	0	0.2159
{p3,p6,p2,p5,p4}	0	0

δ. Αφού έχουμε και άλλες διαθέσιμες συστάδες για συγχώνευση, συνεχίζουμε να επαναλαμβάνουμε το βήμα 3.

Επομένως, βλέποντας τον πιο πρόσφατο πίνακα αποστάσεως παραπάνω, παρατηρούμε ότι τα (p3, p6, p2, p5, p4) και p1 έχουν τη μικρότερη απόσταση όλων - 0.2159 (η τελευταία που απομένει στον πίνακα). Έτσι συγχωνεύουμε αυτά τα δύο σε μία συστάδα. Τώρα πλέον δεν υπάρχει λόγος να υπολογίσουμε τον πίνακα αποστάσεως διότι δεν υπάρχουν άλλες συστάδες για συγχώνευση.

Παράδειγμα


0,40
0,22
0,35
0,26
0,08
0,45
0,60
0,55
0,50
0,45
0,40
0,35
0,30
0,25
0,20
0,15
0,10
0,05
0,00
0,00 0,05 0,10 0,15 0,20 0,25 0,30 0,35 0,40 0,45 0,50
P1
P2
P3
P4
P5
P6


0.22
0.2
0.18
0.16
0.14
0.12
0.1
3
6
2
5
4
1