Εξόρυξη δεδομένων με χρήση τεχνικών μηχανικής μάθησης

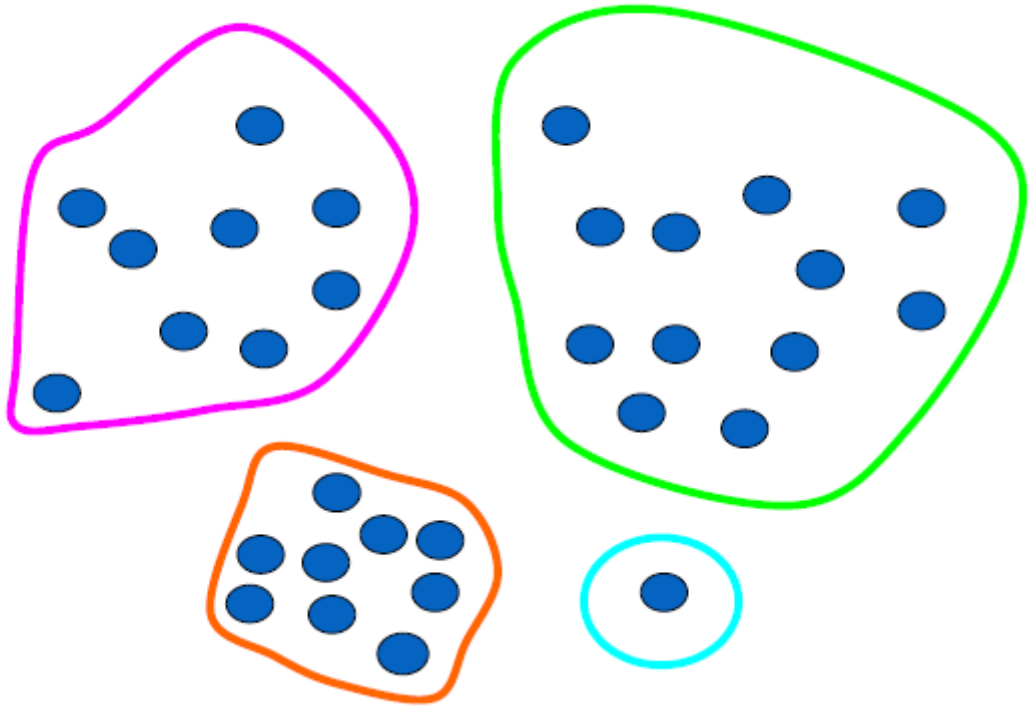
Τμ. Μηχανικών Πληροφορικής και Υπολογιστών

Ομαδοποίηση (clustering)

Τι είναι η Ομαδοποίηση (clustering)

- Σκοπός όλων των clustering αλγορίθμων αποτελεί ο διαχωρισμός ενός πλήθους σημείων ή αντικειμένων σε ομοειδείς ομάδες (clusters).
- Προς την κατεύθυνση αυτή χρησιμοποιούνται συναρτήσεις οι οποίες υπολογίζουν την απόσταση μεταξύ των σημείων.
- Με βάση την απόσταση μεταξύ των σημείων δημιουργούνται ομάδες από αντικείμενα τα οποία εμφανίζουν τη μικρότερη δυνατή απόσταση. Δηλαδή 2 αντικείμενα που ανήκουν σε διαφορετικά clusters θα πρέπει να εμφανίζουν μεγαλύτερη απόσταση σε σχέση με αυτά που βρίσκονται στο ίδιο το δικό τους cluster.

Τι είναι η Ομαδοποίηση (clustering)



Τι είναι η Ομαδοποίηση (clustering)

Προκειμένου να επιτύχουμε το διαχωρισμό του συνόλου μας σε αντιπροσωπευτικά clusters (ομάδες) θα πρέπει τα αντικείμενα εντός των clusters να έχουν

- Υψηλές τιμές ομοιότητας εντός του cluster.
- Χαμηλές τιμές ομοιότητας με τα άλλα αντικείμενα εκτός του cluster.

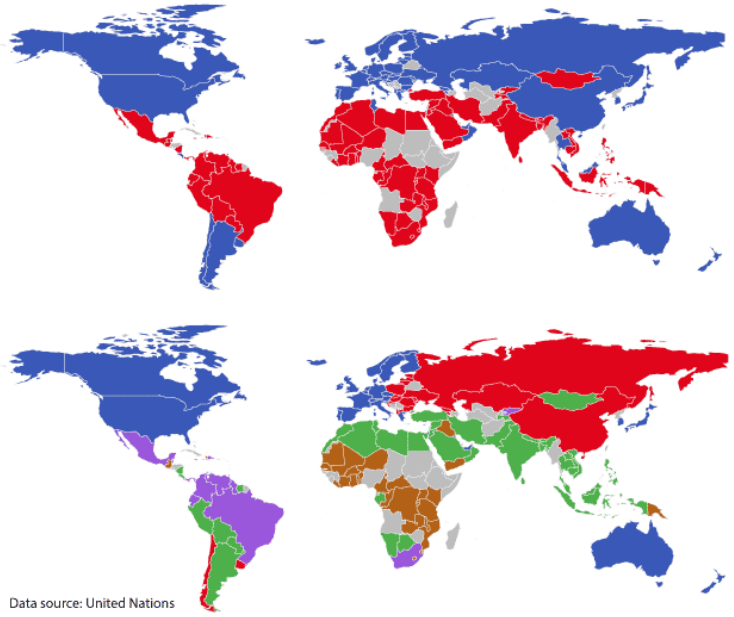
Ο διαχωρισμός σε clusters εξαρτάται από

- Το μέτρο ομοιότητας
- Τη μέθοδο που θα χρησιμοποιήσουμε και
- Τα κριτήρια που θέτει ο χρήστης με βάση την εμπειρία του

Παραδείγματα Χρήσης της Ομαδοποίησης

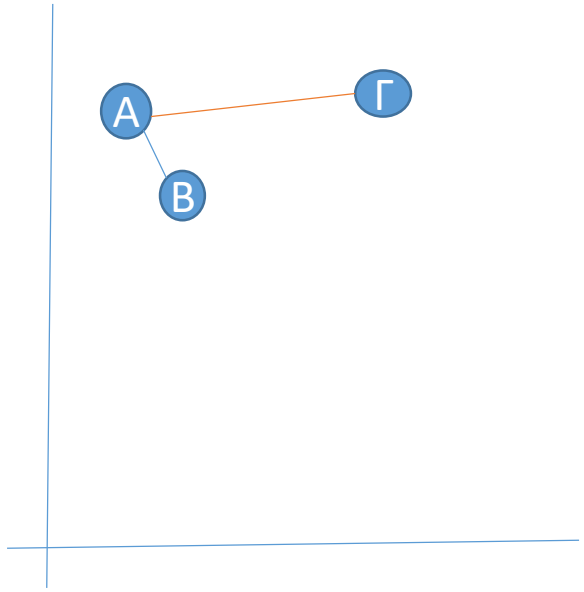
Στο τομέα του Marketing

Στην ομαδοποίηση πληθυσμών (*GDP growth, gender equality, number of years of school and median age***)**



Μέτρα Ομοιότητας

- Ένας τρόπος καθορισμού του βαθμού ομοιότητας δύο παρατηρήσεων είναι με τη χρήση της απόστασης τους



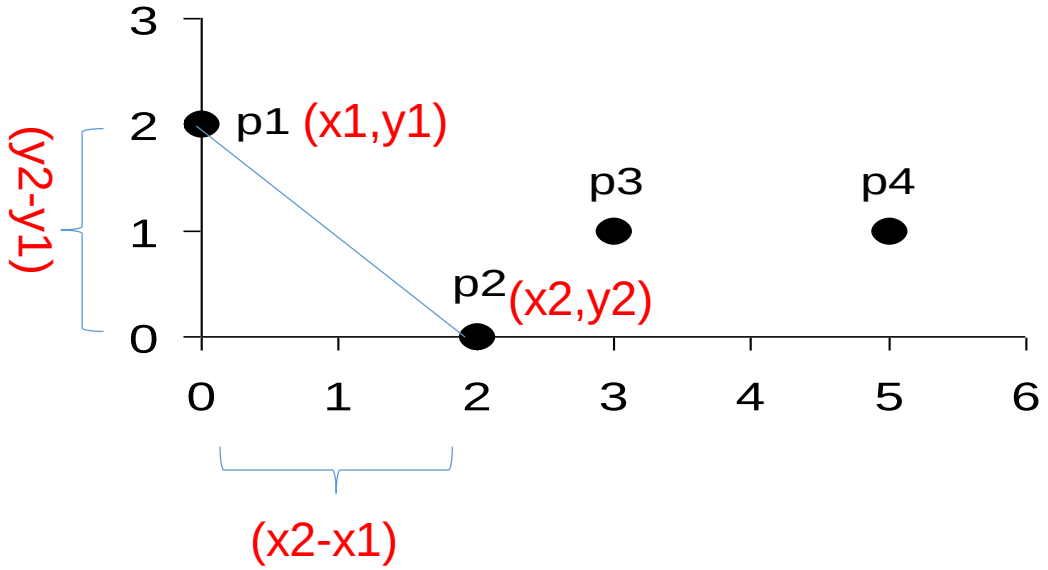
Euclidean Distance

- Ευκλείδεια Απόσταση

$$\boldsymbol{dist} = \sqrt{\sum_{\boldsymbol{k}=1}^{\boldsymbol{n}} (\boldsymbol{p_k} - \boldsymbol{q_k})^2}$$

Όπου n είναι ο αριθμός των διαστάσεων και p_k και q_k είναι οι τιμές αντίστοιχα των πεδίων p και q.

Euclidean Distance



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

Distance	p1	p2	p3	p4
p1	0	2,828	3,162	5,099
p2	2,828	0	1,414	3,162
p3	3,162	1,414	0	2
p4	5,099	3,162	2	0

- Η απόσταση μεταξύ των σημείων p1,p2 ορίζεται ως

$$D(p1, p2) = \sqrt{(x2-x1)^2 + (y2-y1)^2} = \sqrt{(2-0)^2 + (0-2)^2} = \sqrt{8} = 2.828$$

Euclidean Distance

Σε περίπτωση που οι τιμές για τις οποίες καλούμαστε να υπολογίσουμε την απόσταση έχουν μεγάλη διαφορά κλίμακας τότε θα πρέπει στις τιμές να γίνει κανονικοποίηση.

- Έστω τα παρακάτω δεδομένα

	Βάρος	Ύψος
p1	100	1.90
p2	54	1.60
p3	51	1.70
p4	88	1.80

- Αν θέλουμε να μετρήσουμε την απόσταση μεταξύ p1,p3 τότε αυτή είναι

$$D(p1, p3) = \sqrt{(x3 - x1)^2 + (y3 - y1)^2} = \sqrt{(100 - 51)^2 + (1.90 - 1.70)^2} = \sqrt{49,20}$$

- Το ύψος παίρνει τιμές στο διάστημα 1,70-1.90 ενώ το βάρος στο διάστημα 51-100 οπότε λόγω της διαφοράς κλίμακας η απόσταση εξαρτάται σε μεγάλο βαθμό από το βάρος. Οπότε σε αυτές τις περιπτώσεις θα πρέπει να κάνουμε κανονονικοποιήση των δεδομένων.

Απόσταση Manhattan

- Η απόσταση Manhattan D_{AB} μεταξύ 2 σημείων ορίζεται ως:

$$D_{AB} = \sum_{i=1}^{n} | a_i - b_i |$$

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

- Παράδειγμα
  - **p1** = (0,2)
  - **p2** = (2,0)

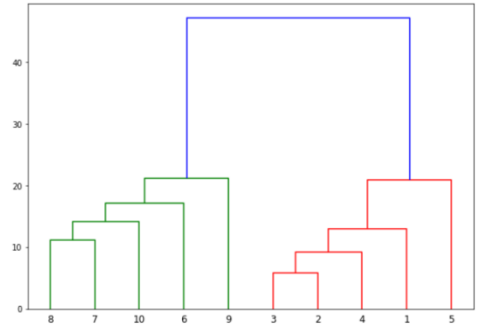
$$D(p1, p2) = | 2 - 0 | + | 0 - 2 | = 4$$

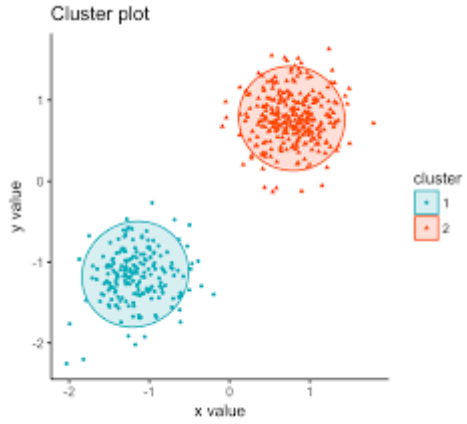
Μέθοδοι εκτέλεσης διαδικάσίας Ομαδοποίησης (clustering)

Οι 4 από τους πιο διαδεδομένους τρόπους ανάλυσης ενός συνόλου σε συστάδες είναι:

Ιεραρχικές μέθοδοι. Οι ιεραρχικές μέθοδοι (hierarchical methods) δημιουργούν μια ιεραρχία από συστάδες. Στο κατώτατο επίπεδο υπάρχουν τα αντικείμενα τα οποία ενώνονται σε ομάδες στο επόμενο επίπεδο. Στο τελευταίο επίπεδο βρίσκεται μια υπερσυστάδα, η οποία περιέχει όλα τα αντικείμενα.

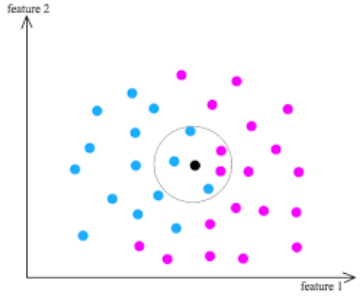
Διαχωριστικές μέθοδοι. Σε αυτή την μέθοδο τα αντικείμενα επιμερίζονται σε k συστάδες. Το πλήθος των συστάδων προκαθορίζεται από τον χρήστη. Ο πιο γνωστός αλγόριθμος διαχωριστικής ΑΣ είναι ο k-Means.



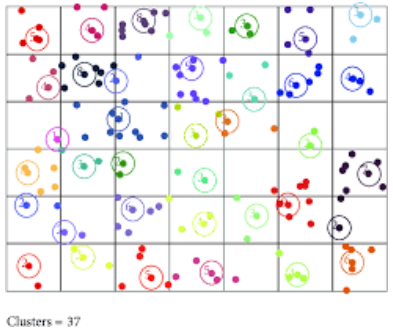


Μέθοδοι εκτέλεσης διαδικάσίας Ομαδοποίησης (clustering)

Μέθοδοι βασισμένες στην πυκνότητα. Στις βασισμένες στην πυκνότητα μεθόδους (density based methods) ελέγχεται η πυκνότητα των αντικειμένων στον χώρο και δημιουργούνται συστάδες, οι οποίες καλύπτουν τις πυκνές περιοχές.



Μέθοδοι πλέγματος. Σε αυτή τη μέθοδο τα δεδομένα επιμερίζονται σε κελιά τα οποία με τη σειρά τους συγκροτούν ένα πλέγμα. Σε αυτή την περίπτωση η δημιουργία των συστάδων καθορίζεται από τα κελιά στα οποία ανήκουν τα αντικείμενα.



ΤΕΛΟΣ ΕΝΟΤΗΤΑΣ