ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΤΜΗΜΑ ΦΥΣΙΚΗΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΥΠΟΛΟΓΙΣΤΙΚΗΣ ΦΥΣΙΚΗΣ

ΤΜΗΜΑ ΦΥΣΙΚΗΣ Α.Π.Θ.
Πρόγραμμα Μεταπτυχιακών Σπουδών
Υπολογιστικής Φυσικής

Μηχανική Μάθηση σε Python
Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning)

Διακονίδης Θόδωρος
Ε.ΔΙ.Π
Εργαστήριο Θεωρ. Φυσικής
Τμήμα Φυσικής ΑΠΘ

Γιατί ονομάζεται έτσι; Σε τι διαφέρει από την επιβλεπόμενη; Πότε εφαρμόζεται;

- Γιατί με την εφαρμογή αυτών των μεθόδων, προσπαθούμε να ανακαλύψουμε τη δομή, συγκεκριμένες ομάδες στοιχείων (clustering) ή συσχετίσεις στοιχείων (association rules). Χωρίς να καθοδηγούμαστε εκ των προτέρων από τη γνώση αποτελεσμάτων (επιβλεπόμενη).

- Την εφαρμόζουμε όταν θέλουμε να κατηγοριοποιήσουμε τα δεδομένα μας χωρίς να γνωρίζουμε κάτι επιπλέον για αυτά. Να ανακαλύψουμε, μοτίβα, πληροφορία που είναι με μια πρώτη ματιά μη ανιχνεύσιμη.

ΤΜΗΜΑ ΦΥΣΙΚΗΣ Α.Π.Θ.
Πρόγραμμα Μεταπτυχιακών Σπουδών
Υπολογιστικής Φυσικής

Συσταδοποίηση (Clustering) κα αλγόριθμοι εφαρμογής

- **Ορισμός**

Είναι η διαδικασία ομαδοποίησης ενός συνόλου αντικειμένων σε μια ομάδα (συστάδα) με τέτοιο τρόπο, έτσι ώστε να είναι πιο συναφή μεταξύ τους σε σχέση με τα άλλα αντικείμενα των άλλων ομάδων.

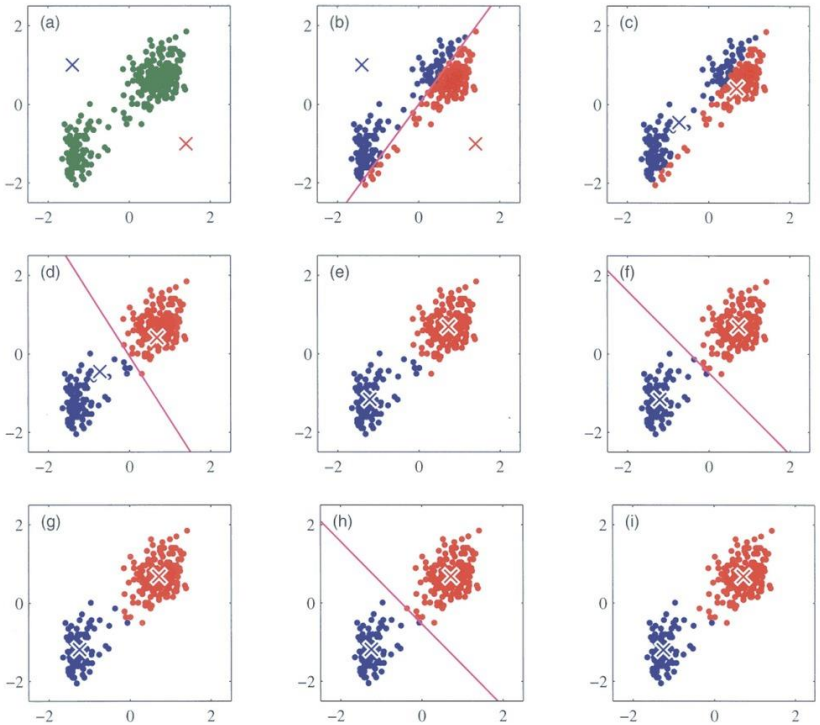
- **Με ποιους αλγόριθμους συσταδοποίησης θα ασχοληθούμε στην python:**

1. Αλγόριθμος **Kmeans**
2. Ιεραρχικοί Αλγόριθμοι (**Hierarchical Clustering Algorithms**)
   a) Συσσωρευτικοί Αλγόριθμοι (Agglomerative Algorithms)
   b) Διαιρετικοί Αλγόριθμοι (Divisive Algorithms)

Αλγόριθμος kmeans

- Αποτελεί μια απλή και απέριττη προσέγγιση συσταδοποίησης της βάσης δεδομένων μας σε Ν **μη αλλεπικαλυπτόμενες ομάδες**.
- Απαραίτητη προϋπόθεση εφαρμογής η **προεπιλογή** του αριθμού των Ν αρχικών ομάδων.
- Βασίζεται στην ιδέα ότι, η ορθότερη ομαδοποίηση επιτυγχάνεται με την όσο το δυνατό μικρότερη **διακύμανση** εντός των ομάδων.
- Ουσιαστικά δηλαδή αναζητείται ομαδοποίηση με τέτοιο τρόπο έτσι ώστε το **άθροισμα όλων των εσωτερικών διακυμάνσεων των Ν ομάδων** να είναι το μικρότερο δυνατό.

Οπτικοποίηση των σταδίων του Kmeans



- Μεταθέτουμε τα κέντρα με βάση τα δεδομένα που επιλέχθηκαν για την κάθε ομάδα και επαναλαμβάνουμε τη διαδικασία.
- Η διαδικασία τερματίζεται όταν ο αλγόριθμος συγκλίνει. (Δεν υπάρχουν περαιτέρω αλλαγές στα μέλη της κάθε ομάδας).

Ιεραρχικοί Αλγόριθμοι

- Διαχωρίζονται σε 2 κατηγορίες:

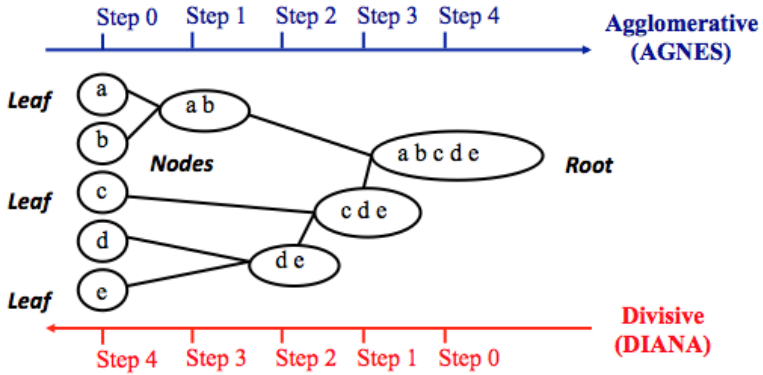
Συσσωρευτικοί Αλγόριθμοι (Agglomerative Algorithms) και AGNES

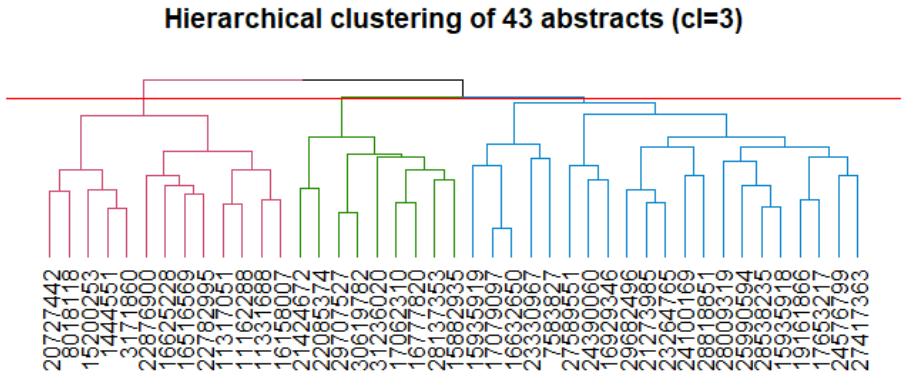
- **Λειτουργία**: Ξεκινάμε θεωρώντας κάθε δείγμα μας ως μια ομάδα. Φύλλο (leaf). Σε κάθε βήμα συγχωνεύονται δυο ομάδες και σχηματίζονται κλαδιά (branches). Αυτή η διαδικασία επαναλαμβάνεται, μέχρις ότου ο αλγόριθμος καταλήξει σε μια μοναδική συστάδα (tree).

Διαιρετικοί Αλγόριθμοι (Divisive Algorithms) και DIANA

- **Λειτουργία:** Ξεκινάμε με όλα τα δείγματα να ανήκουν σε μια ενιαία συστάδα. Σε κάθε βήμα, μια ομάδα διασπάται σε δύο. Αυτό γίνεται επαναληπτικά, μέχρι να καταλήξουμε σε *n* ομάδες. (Ποιο πολύπλοκη διαδικασία).

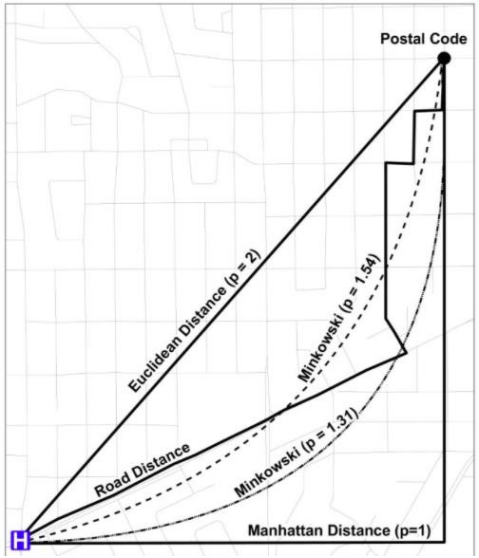
Οπτικοποίηση των σταδίων των HC Algorithms



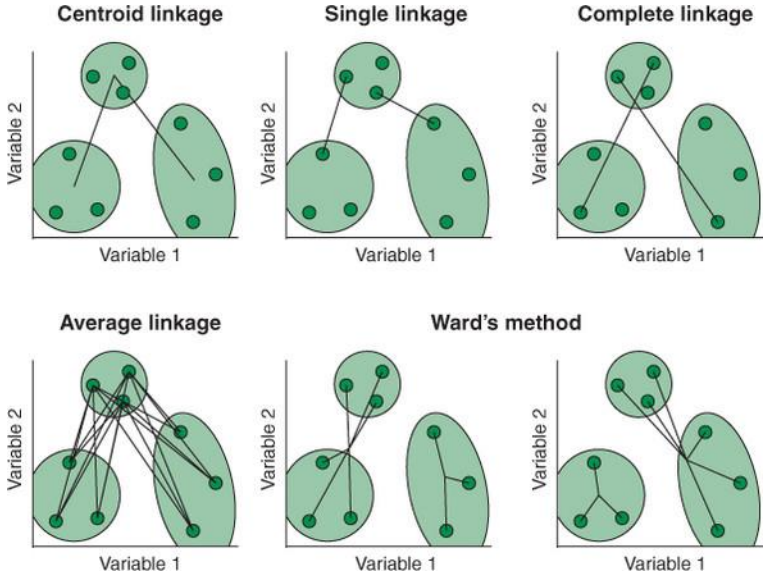


- Σε αντίθεση με τον kmeans δεν χρειάζεται να θέσουμε εκ των προτέρων τον αριθμό των clusters. Μπορούμε, ανάλογα με του που θα κόψουμε το δεντρο (tree) να προκύψει αυτόματα.

Αποστάσεις & Ομαδοποιήσεις δειγμάτων



Συγκριτικό διαφόρων αποστάσεων



Συγκριτικό διαφόρων ομαδοποιήσεων

Βίντεο επεξήγησης των ομαδοποιήσεων

Επιλέγοντας τον καταλληλότερο αριθμό συστάδων

Ένα από τα δυσκολότερα σημεία στην ομαδοποίηση των δεδομένων είναι η επιλογή του αριθμού των ομάδων. Υπάρχουν διάφοροι αυτοματοποιημένοι τρόποι. Το πιο σημαντικό είναι να γνωρίζουμε, να έχουμε οπτική επαφή με τα δεδομένα μας.

ΤΜΗΜΑ ΦΥΣΙΚΗΣ Α.Π.Θ.
Πρόγραμμα Μεταπτυχιακών Σπουδών
Υπολογιστικής Φυσικής

Μέθοδοι επιλογής συστάδων

- Οι πιο σημαντικοί είναι οι ακόλουθοι:

1. Elbow method
2. Silhouette method
3. Gap statistic method

Ο κανόνας του αγκώνα (The elbow method)

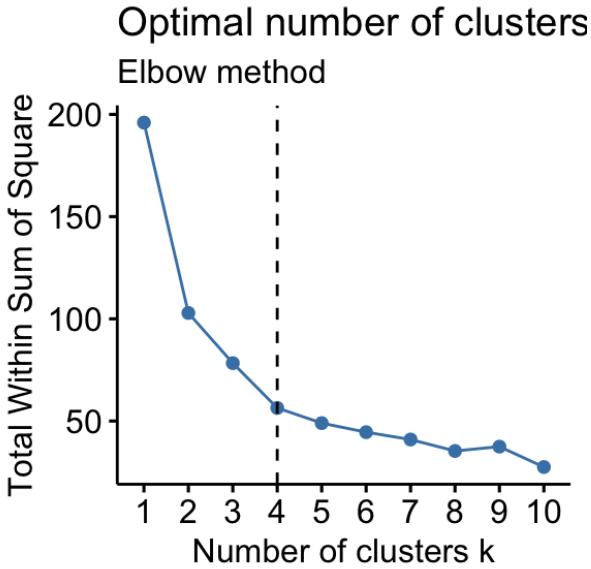
- Η μέθοδος βασίζεται στον υπολογισμό του WSS (Within-cluster Sum of Square) διακύμανση, που υπολογίζει το πόσο συνεπτυγμένη, συμπαγής είναι η ομάδα.

(**kmeans:** η ορθότερη ομαδοποίηση επιτυγχάνεται με την όσο το δυνατό μικρότερη διακύμανση εντός των ομάδων).

Ο κανόνας του αγκώνα (The elbow method)

- Ο καταλληλότερος αριθμός ομάδων ορίζεται ως εξής:

1. Υπολογίζουμε τον αλγόριθμο ομαδοποίησης για διαφορετικό αριθμό ομάδων k.
2. Για κάθε k υπολογίζουμε το WSS.
3. Κατασκευάζουμε την καμπύλη WSS-k.
4. Το σημείο που εμφανίζει καμπή στο διάγραμμα θεωρείται ο καταλληλότερος αριθμός ομάδων.



Average Silhouette Method

- Υπολογίζει την ποιότητα της συσταδοποίησης. Το πόσο καλά κάθε δείγμα ανήκει σε μια ομάδα.

- Το κάθε δείγμα παίρνει τιμή ανάλογη, από -1 έως 1. Ανάλογα με την μέση απόστασή του από τα σημεία της ίδιας ομάδας a(i), και αυτά άλλων, b(i).

$$s(\mathrm{i}) = \frac{b(\mathrm{i}) - \mathrm{a}(\mathrm{i})}{\max\{\mathrm{a}(\mathrm{i}), \mathrm{b}(\mathrm{i})\}}$$

- Δείγμα κοντά στην τιμή:

  1: Ομαδοποιήθηκε σωστά 0: είναι μεταξύ 2 ομάδων

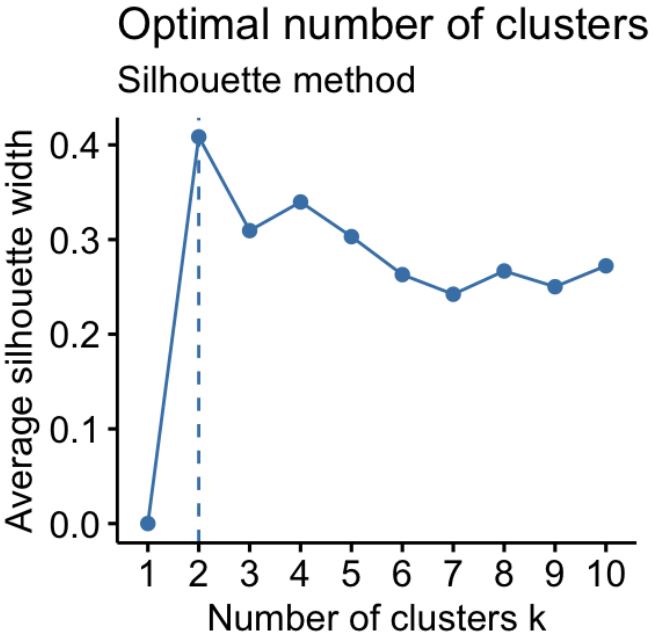
  -1: Ανήκει πιθανόν σε άλλη ομάδα.

- Υψηλό Average Silhouette Width υποδεικνύει καλή ομαδοποίηση.

Average Silhouette Method (Η Διαδικασία)

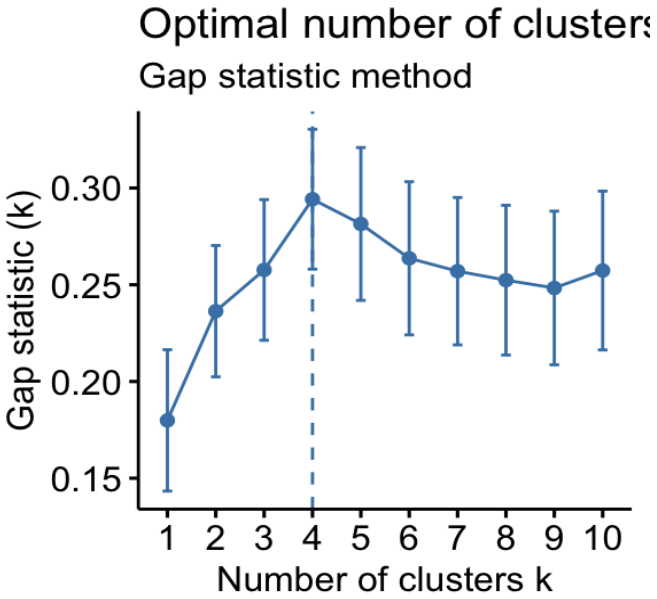
Η διαδικασία είναι η ακόλουθη:

1. Υπολογίζουμε τον αλγόριθμο ομαδοποίησης για διαφορετικό αριθμό ομάδων k.
2. Για κάθε k υπολογίζουμε το ASW.
3. Κατασκευάζουμε την καμπύλη ASW-k.
4. Το σημείο που εμφανίζει μέγιστο θεωρείται ο καταλληλότερος αριθμός ομάδων.



Μέθοδος Gap Statistic

- Υπολογίζει ουσιαστικά την διαφορά στην αναμενόμενη διακύμανση καθεμιάς ομάδας (από το σύνολο των ομάδων που διαχωρίσαμε τα δεδομένα) σε σχέση με τη διακύμανση μιας ομάδας που δημιουργήθηκε με τυχαίο τρόπο.
- Ο σκοπός μας είναι να διαλέξουμε τις k ομάδες που μεγιστοποιούν το $Gap_n(k)$.



Σημείωση: Για όσους ενδιαφέρονται περαιτέρω, η μέθοδος αναπτύσσεται αναλυτικά στην παρακάτω δημοσίευση: https://web.stanford.edu/~hastie/Papers/gap.pdf