Αν. Καθηγητής Π. Λουρίδας

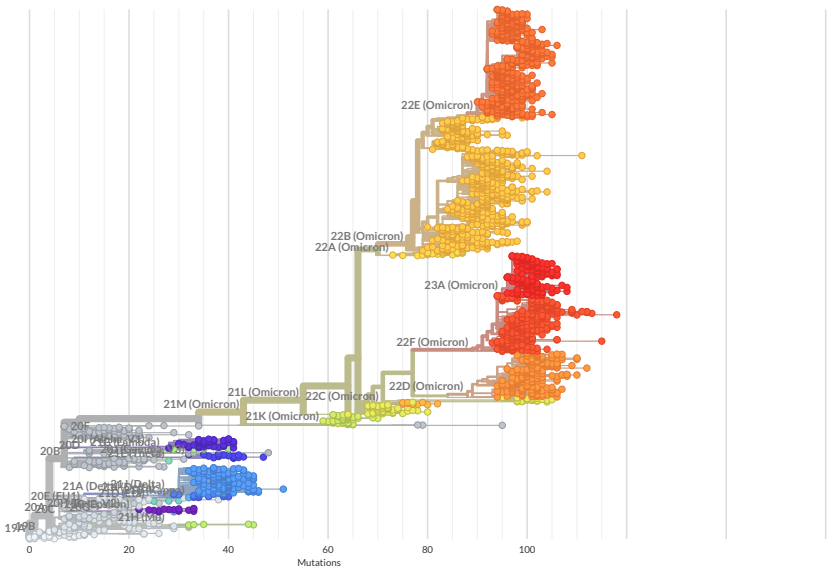
Τμήμα Διοικητικής Επιστήμης και Τεχνολογίας

Οικονομικό Πανεπιστήμιο Αθηνών

Δενδρογράμματα

Τα τελευταία πρόγραμμα κάτι κακό συνέβη στον κόσμο, το οποίο κόστισε στην ανθρωπότητα εκατομμύρια νεκρούς. Ταυτόχρονα, η ταχύτητα στην αντιμετώπιση του ιού Covid-19 ήταν μεγαλύτερη από οποιαδήποτε άλλη πανδημία στην ιστορία. Εμβόλια αναπτύχθηκαν γρηγορότερα από ποτέ, ενώ επιστήμονες σε όλον τον κόσμο συνεργάστηκαν με πυρετώδεις ρυθμούς για την παρακολούθηση της εξέλιξης της πορείας του ιού.

Ανάμεσα στα εργαλεία τα οποία χρησιμοποιήθηκαν ήταν διαγράμματα σαν αυτό που βλέπετε παρακάτω, και το οποίο προέρχεται από το εξαιρετικό nextstrain:



Αυτό είναι ένα *δενδρόγραμμα* (dendrogram), το οποίο δείχνει τις διάφορες μεταλλάξεις του ιού και τη σχέση της κάθε μετάλλαξης με άλλες. Το δενδρόγραμμα αυτό είναι ένα δένδρο που μεγαλώνει από τα αριστερά προς τα δεξιά. Τα φύλλα του δένδρου είναι διάφορες μεταλλάξεις του ιού. Τα φύλλα αυτά ενώνονται σε κλαδιά, ώστε ένα κλαδί να περιέχει τα φύλλα, δηλαδή τις μεταλλάξεις, που μοιάζουν περισσότερο μεταξύ τους. Τα κλαδιά επίσης ενώνονται μεταξύ τους με βάση την ομοιότητά τους, ώστε τελικά φτάνουμε στη ρίζα του δένδρου.

Τέτοιου είδους διαγράμματα παράγονται μέσω *συσταδοποίησης* (clustering). Σκοπός της συσταδοποίησης σε ένα σύνολο δεδομένων είναι η εύρεση ομάδων, ή συστάδων όπως αποκαλούνται, ώστε τα δεδομένα να συγκεντρώνονται σε ομάδες που αφενός

περιέχουν ομοειδή στοιχεία, αφετέρου η κάθε συστάδα να διαφέρει όσο το δυνατόν περισσότερο από τις άλλες συστάδες.

Υπάρχουν διάφοροι μέθοδοι συσταδοποίησης. Μια κατηγορία μεθόδων συσταδοποίησης είναι η *ιεραρχική συσταδοποίηση* (hierarchical clustering). Στην ιεραρχική συσταδοποίηση, οι συστάδες ανήκουν σε μία ιεραρχία, όπου μεγαλύτερες συστάδες περιέχουν μικρότερες συστάδες. Αυτό απεικονίζεται σε ένα δενδρόγραμμα, όπως αυτό που είδαμε. Η ρίζα του δένδρου αντιστοιχεί σε μία συστάδα που περιέχει το σύνολο των δεδομένων. Κάθε συστάδα περιέχει τις συστάδες που προκύπτουν από τα κλαδιά του, κ.ο.κ. μέχρι να φτάσουμε στα φύλλα.

Η ιεραρχική συσταδοποίηση πραγματοποιείται ως εξής:

1. Αρχικά τοποθετούμε κάθε ένα στοιχείο σε μία συστάδα μόνο του. Άρα έχουμε τόσες συστάδες όσο είναι το πλήθος των δεδομένων.
2. Όσο έχουμε παραπάνω από δύο συστάδες:
    1. Βρίσκουμε τις δύο συστάδες που είναι πιο όμοιες μεταξύ τους.
    2. Ενώνουμε αυτές τις δύο συστάδες και δημιουργούμε μια νέα συστάδα.
    3. Αφαιρούμε τις δύο συστάδες που χρησιμοποιήσαμε και προσθέτουμε τη νέα συστάδα στις συστάδες που έχουμε στη διάθεσή μας.

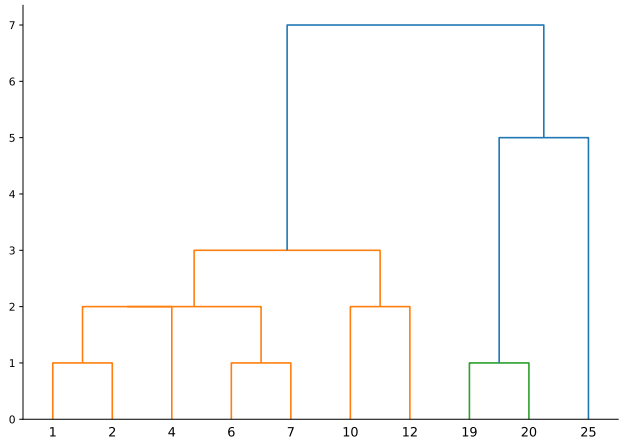
Για να δουλέψει η παραπάνω προσέγγιση, θα πρέπει να ορίσουμε τον τρόπο με τον οποίο μετράμε την ομοιότητα των συστάδων, ώστε να μπορούμε να βρούμε σε κάθε επανάληψη τις δύο συστάδες που είναι πιο όμοιες μεταξύ τους. Η μέτρηση της ομοιότητας ανάγεται στη μέτρηση της απόστασης που έχουν δύο συστάδες: όσο πιο μακριά είναι η μία από την άλλη, τόσο πιο ανόμοιες είναι. Υπάρχουν λοιπόν διάφορες μετρικές με τις οποίες μετράμε την απόσταση που έχουν δύο συστάδες. Εμείς θα δούμε κάποιες από αυτές. Στη συνέχεια, θα συμβολίζουμε με $d(\cdot, \cdot)$ την απόσταση δύο συστάδων και με $\text{dist}(\cdot, \cdot)$ την απόσταση δύο σημείων (παρατηρήσεων, δεδομένων) που ανήκουν σε συστάδες.

Η *απλή* (single) μέθοδος, ορίζει την απόσταση μεταξύ δύο συστάδων u, v ως:

$$d(u, v) = \min_{i \in u, j \in v} (\text{dist}(u[i], v[j]))$$

δηλαδή η απόσταση των δύο συστάδων είναι ίση με τη μικρότερη απόσταση μεταξύ δύο οποιονδήποτε στοιχείων των δύο συστάδων.

Αν τα αρχικά μας δεδομένα είναι τα $[7, 10, 4, 20, 2, 25, 19, 6, 12, 1]$, τότε το δενδρόγραμμα που προκύπτει με αυτήν τη μέθοδο είναι:

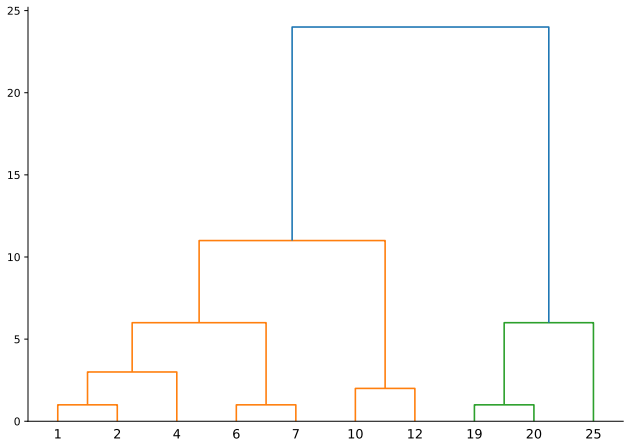


Στο δενδρόγραμμα αυτό, στον οριζόντιο άξονα έχουμε τα δεδομένα και στον κάθετο άξονα έχουμε την απόσταση μεταξύ των συστάδων, η οποία αντιστοιχεί στο μήκος κάθε κλαδιού του δένδρου.

Η *πλήρης* (complete) μέθοδος, ορίζει την απόσταση μεταξύ δύο συστάδων u, v ως:

$$d(u, v) = \max_{i \in u, j \in v} (\text{dist}(u[i], v[j]))$$

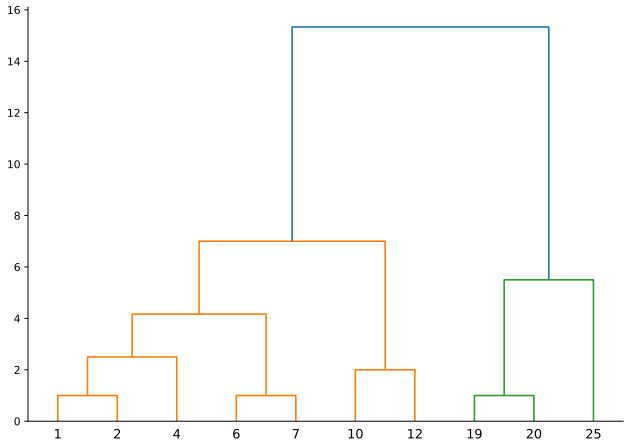
δηλαδή η απόσταση των δύο συστάδων είναι ίση με τη μεγαλύτερη απόσταση μεταξύ δύο οποιονδήποτε στοιχείων των δύο συστάδων. Το δενδρόγραμμα που προκύπτει για το παράδειγμά μας είναι:



Η *μέση* (average) μέθοδος, ορίζει την απόσταση μεταξύ δύο συστάδων u, v ως:

$$d(u,v) = \sum_{i \in u, j \in v} \frac{\text{dist}(u[i], v[j])}{|u||v|}$$

δηλαδή η απόσταση των δύο συστάδων είναι ίση με τη μέση απόσταση των στοιχείων των δύο συστάδων. Το δενδρόγραμμα που προκύπτει για το παράδειγμά μας είναι:



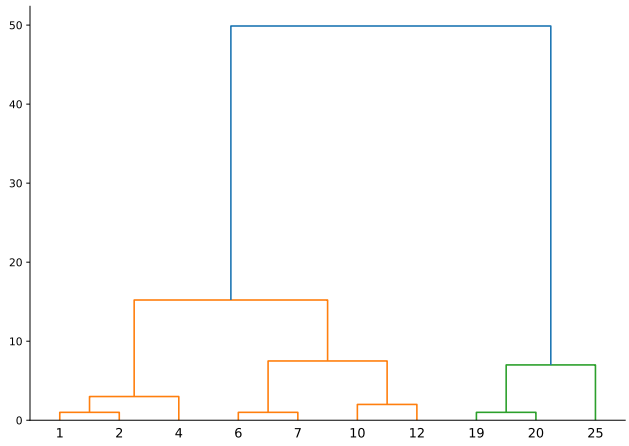
Μπορείτε να παρατηρήσετε ότι στην περίπτωση αυτή οι συστάδες είναι οι ίδιες

με αυτές που προκύπτουν από την πλήρη μέθοδο, αν και οι αποστάσεις είναι διαφορετικές.

Η μέθοδος *Ward* ορίζει την απόστατη μεταξύ της νέας συστάδας u που προκύπτει από τη συγχώνευση δύο συστάδων s, t και μιας άλλης συστάδας v ως εξής:

$$d(u,v) = \frac{|v|+|s|}{|v|+|s|+|t|}|d(v,s)| + \frac{|v|+|t|}{|v|+|s|+|t|}|d(v,t)| - \frac{|v|}{|v|+|s|+|t|}|d(s,t)|$$

Η μέθοδος πήρε το όνομά της από τον Joe H. Ward, Jr. που την επινόησε. Η μέθοδος Ward επιλέγει να συγχωνεύσει δύο συστάδες ώστε συνολικά οι συστάδες που θα έχουμε να εμφανίζουν τη μικρότερη δυνατή εσωτερική διακύμανση. Το δενδρόγραμμα που προκύπτει για το παράδειγμά μας είναι:



Παρά το ότι αυτές οι μέθοδοι είναι διαφορετικές μεταξύ τους, μπορούν όλες να υπολογιστούν με μία ενιαία προσέγγιση, η οποία ονομάζεται Lance-Williams, από το όνομα των ερευνητών που την επινόησαν.

Σύμφωνα με τους Lance-Williams, ενεργούμε ως εξής:

- Η απόσταση μεταξύ δύο αρχικών συστάδων u, v (αυτές που περιέχουν ένα από τα αρχικά στοιχεία) είναι ίση με $|i-j|$ όπου $i \in u$ είναι το μοναδικό στοιχείο της u και $j \in v$ είναι το μοναδικό στοιχείο της v.
- Η απόσταση μιας νέας συστάδας u που δημιουργούμε από τη συγχώνευση δύο άλλων συστάδων s, t από κάθε άλλη συστάδα v προκύπτει από τον τύπο:

$$d(u,v) = \alpha_i d(s,v) + \alpha_j d(t,v) + \beta d(s,t) + \gamma |d(s,v) - d(t,v)|$$

όπου τα $\alpha_i, \alpha_j, \beta, \gamma$ εξαρτώνται από την εκάστοτε μέθοδο.

Επομένως έχοντας τις αποστάσεις μεταξύ των αρχικών συστάδων, απλώς υπολογίζουμε τις αποστάσεις των συστάδων από κάθε νέα συστάδα όταν τη δημιουργούμε.

Για τις μεθόδους που αναφέραμε, οι τιμές των συντελεστών είναι όπως στον πίνακα που ακολουθεί:

	α_i	α_j	β	γ
single	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
complete	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
average	$\frac{\lvert s\rvert}{\lvert s\rvert+\lvert t\rvert}$	$\frac{\lvert t\rvert}{\lvert s\rvert+\lvert t\rvert}$	0	0
Ward	$\frac{\lvert s\rvert+\lvert v\rvert}{\lvert s\rvert+\lvert v\rvert+\lvert t\rvert}$	$\frac{\lvert t\rvert+\lvert v\rvert}{\lvert s\rvert+\lvert v\rvert+\lvert t\rvert}$	$-\frac{\lvert v\rvert}{\lvert s\rvert+\lvert v\rvert+\lvert t\rvert}$	

Σκοπός της εργασίας είναι η υλοποίηση προγράμματος σε Python για την εφαρμογή συσταδοποίησης σε δεδομένα *σύμφωνα με την προσέγγιση Lance-Williams που περιγράφηκε παραπάνω.* Τα δεδομένα μας θα αποτελούνται από ακέραιους αριθμούς. Γενικότερα, τα δεδομένα μπορεί να είναι διανύσματα (το κάθε διάνυσμα αντιστοιχεί σε ιδιότητες των οντοτήτων που μας ενδιαφέρουν). Η λογική που περιγράψαμε δεν διαφέρει, το μόνο που χρειάζεται να προσέξουμε είναι ο ορισμός της απόστασης δύο στοιχείων· συνήθως είναι η Ευκλείδια απόσταση. Αλλά όπως είπαμε εδώ θα περιοριστούμε σε δεδομένα που είναι απλοί ακέραιοι.

Απαιτήσεις Προγράμματος

Κάθε φοιτητής θα εργαστεί σε αποθετήριο στο GitHub. Για να αξιολογηθεί μια εργασία θα πρέπει να πληροί τις παρακάτω προϋποθέσεις:

- Για την υποβολή της εργασίας θα χρησιμοποιηθεί το ιδιωτικό αποθετήριο του φοιτητή που δημιουργήθηκε για τις ανάγκες του μαθήματος και του έχει αποδοθεί. Το αποθετήριο αυτό έχει όνομα του τύπου `username-algo-assignments`, όπου `username` είναι το όνομα του φοιτητή στο GitHub. Για παράδειγμα, το σχετικό αποθετήριο του διδάσκοντα θα ονομαζόταν `louridas-algo-assignments` και θα ήταν προσβάσιμο στο https://github.com/dmst-algorithms-course/louridas-algo-assignments. *Τυχόν άλλα αποθετήρια απλώς θα αγνοηθούν.*
- Μέσα στο αποθετήριο αυτό θα πρέπει να δημιουργηθεί ένας κατάλογος `assignment-2023-1`.
- Μέσα στον παραπάνω κατάλογο το πρόγραμμα θα πρέπει να αποθηκευτεί με το όνομα `lance_williams.py`.
- Δεν επιτρέπεται η χρήση έτοιμων βιβλιοθηκών γράφων ή τυχόν έτοιμων υλοποιήσεων των αλγορίθμων, ή τμημάτων αυτών, εκτός αν αναφέρεται ρητά ότι επιτρέπεται.

- Επιτρέπεται η χρήση δομών δεδομένων της Python όπως στοίβες, λεξικά, σύνολα, κ.λπ.
- Επιτρέπεται η χρήση των παρακάτω βιβλιοθηκών ή τμημάτων τους όπως ορίζεται:
  - `sys.argv`
  - `argparse`
- Το πρόγραμμα θα πρέπει να είναι γραμμένο σε Python 3.
- Υπάρχουν διάφοροι τρόποι να υλοποιηθεί ιεραρχική συσταδοποίηση· εσείς θα πρέπει να υλοποιήσετε τον τρόπο των Lance-Williams που περιγράφεται εδώ. Υλοποιήσεις άλλων μεθόδων δεν θα γίνουν δεκτές.
- Στην υλοποίηση του προγράμματος θα χρειαστεί να διατρέχετε τις συστάδες προκειμένου να βρείτε το ζευγάρι που θα συγχωνεύσετε. Για να είναι τα αποτελέσματά σας ίδια με τα αποτελέσματα αναφοράς που θα δείτε στα παραδείγματα, θα πρέπει οι συστάδες να είναι ταξινομημένες πριν τις διατρέξετε.
- Η έξοδος του προγράμματος θα πρέπει να περιλαμβάνει μόνο ό,τι φαίνεται στα παραδείγματα. *Η φλυαρία δεν επιβραβεύεται.*

Το πρόγραμμα θα καλείται ως εξής (όπου `python` η κατάλληλη εντολή στο εκάστοτε σύστημα):

```
python lance_williams.py  method input_filename
```

Η σημασία των παραμέτρων είναι η εξής:

- Η παράμετρος `method` δίνει τη μέθοδο συσταδοποίησης και οι επιτρεπτές τιμές είναι `single`, `complete`, `average`, `ward`.
- Η παράμετρος `input_filename` δίνει το όνομα του αρχείου από όπου θα διαβαστούν τα δεδομένα προς συσταδοποίηση. Τα δεδομένα θα χωρίζονται μεταξύ τους με έναν κενό χαρακτήρα.

Η έξοδος του προγράμματος θα δείχνει την πορεία της συσταδοποίσης, όπως στα παραδείγματα που ακολουθούν.

Παραδείγματα

Παράδειγμα 1

Αν ο χρήστης του προγράμματος δώσει:

```
python lance_williams.py single example.txt
```

τότε το πρόγραμμα θα διαβάσει τα δεδομένα από το αρχείο `example.txt` και στην έξοδο θα εμφανίσει:

```
(1) (2) 1.00 2
(6) (7) 1.00 2
```

```
(19) (20) 1.00 2
(1 2) (4) 2.00 3
(1 2 4) (6 7) 2.00 5
(10) (12) 2.00 2
(1 2 4 6 7) (10 12) 3.00 7
(19 20) (25) 5.00 3
(1 2 4 6 7 10 12) (19 20 25) 7.00 10
```

Κάθε γραμμή περιγράφει τη δημιουργία μίας συστάδας. Σε παρενθέσεις εμφανίζονται οι δύο συστάδες που συγχωνεύονται. Μετά ακολουθεί η απόσταση των δύο συγχωνευόμενων συστάδων με ακρίβεια δύο δεκαδικών ψηφίων και το πλήθος των στοιχείων της νέας συστάδας που προκύπτει.

Παράδειγμα 2

Αν ο χρήστης του προγράμματος δώσει:

```
python lance_williams.py complete example.txt
```

τότε το πρόγραμμα θα διαβάσει τα δεδομένα του προηγούμενου παραδείγματος και στην έξοδο θα εμφανίσει:

```
(1) (2) 1.00 2
(6) (7) 1.00 2
(19) (20) 1.00 2
(10) (12) 2.00 2
(1 2) (4) 3.00 3
(1 2 4) (6 7) 6.00 5
(19 20) (25) 6.00 3
(1 2 4 6 7) (10 12) 11.00 7
(1 2 4 6 7 10 12) (19 20 25) 24.00 10
```

Παράδειγμα 3

Αν ο χρήστης του προγράμματος δώσει:

```
python lance_williams.py average example.txt
```

τότε το πρόγραμμα θα διαβάσει τα δεδομένα του προηγούμενου παραδείγματος και στην έξοδο θα εμφανίσει:

```
(1) (2) 1.00 2
(6) (7) 1.00 2
(19) (20) 1.00 2
(10) (12) 2.00 2
(1 2) (4) 2.50 3
(1 2 4) (6 7) 4.17 5
(19 20) (25) 5.50 3
(1 2 4 6 7) (10 12) 7.00 7
(1 2 4 6 7 10 12) (19 20 25) 15.33 10
```

Παράδειγμα 4

Αν ο χρήστης του προγράμματος δώσει:

```
python lance_williams.py ward example.txt
```

τότε το πρόγραμμα θα διαβάσει τα δεδομένα του προηγούμενου παραδείγματος και στην έξοδο θα εμφανίσει:

```
(1) (2) 1.00 2
(6) (7) 1.00 2
(19) (20) 1.00 2
(10) (12) 2.00 2
(1 2) (4) 3.00 3
(19 20) (25) 7.00 3
(6 7) (10 12) 7.50 4
(1 2 4) (6 7 10 12) 15.21 7
(1 2 4 6 7 10 12) (19 20 25) 49.89 10
```

Περισσότερες Πληροφορίες

Οι Lance και Williams δημοσίευσαν την εργασία τους στο Lance and Williams (1966). Για περισσότερα όσον αφορά τις τιμές των παραμέτρων $\alpha_i, \alpha_j, \beta, \gamma$, δείτε το άρθρο του Cormack (1971). Για τη μέθοδο Ward, δείτε το άρθρο Ward (1963).

Cormack, R. M. 1971. "A Review of Classification." *Journal of the Royal Statistical Society* 134 (3): 321–67.

Lance, G. N., and W. T. Williams. 1966. "A Generalized Sorting Strategy for Computer Classifications." *Nature* 212 (5058): 218. https://doi.org/10.1038/212218a0.

Ward, Joe H. Jr. 1963. "Hierarchical Grouping to Optimize an Objective Function." *Journal of the American Statistical Association* 58 (301): 236–44.

Καλή Επιτυχία!