Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Αλγοριθμική Επιστήμη Δεδομένων 2025-26

Διδάσκοντες: Θ. Σούλιου

1η Σειρά Ασκήσεων

Άσκηση 1.

Μελετήστε τον αλγόριθμο Spectral Clustering από το άρθρο: https://arxiv.org/pdf/0711.0189

Έστω ότι έχουμε 4 σημεία χωρισμένα σε 2 "φυσικά" clusters:

- Cluster 1: x_1, x_2
- Cluster 2: x_3, x_4

Η **similarity matrix** W (ή weighted adjacency matrix) δίνεται από:

$$W = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0.05 & 0 \\ 0 & 0.05 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

*Σημείωση: Οι τιμές δείχνουν την **ομοιότητα (similarity)** μεταξύ των σημείων. Μέσα στο ίδιο cluster η ομοιότητα είναι μεγάλη, ενώ μεταξύ clusters είναι μικρή. Μπορούμε να σκεφτούμε την W και ως adjacency matrix ενός weighted graph, όπου οι τιμές των edges αντιπροσωπεύουν την ομοιότητα.*

α)Υπολογίστε τη διαγώνια μήτρα βαθμών D και τον unnormalized Laplacian: $L = D - W$.

β)Βρείτε τα $k = 2$ ιδιοδιανύσματα που αντιστοιχούν στις μικρότερες ιδιοτιμές του L.

γ)Αναπαραστήστε τα δεδομένα σε νέο χώρο χρησιμοποιώντας τα ιδιοδιανύσματα και εφαρμόστε k-means clustering και

δ) Εξηγήστε γιατί το Spectral Clustering μπορεί να διαχωρίσει δεδομένα που δεν είναι γραμμικά διαχωρίσιμα (non-linearly separable data). Σημείωση: τα non-linearly separable data δεν μπορούν να διαχωριστούν με ένα γραμμικό όριο, όπως μια ευθεία γραμμή. Το Spectral Clustering μετασχηματίζει τα δεδομένα σε έναν νέο χώρο χρησιμοποιώντας τα ιδιοδιανύσματα του Graph Laplacian, όπου τα clusters γίνονται γραμμικά διαχωρίσιμα.

Άσκηση 2.

Να λύσετε τις ασκήσεις 7.2.2, 7.2.3, 7.2.5, 7.3.2 και 7.4.1 από το βιβλίο MMDS.

Άσκηση 3. Δίνεται γράφημα με κορυφές: 1, 2, 3, 4, 5, 6, 7, 8, 9 και ακμές:

$$1 \to 2, 3$$
$$2 \to 3, 4$$
$$3 \to 1, 5$$
$$4 \to 5, 6$$
$$5 \to 6, 7$$
$$6 \to 3, 7, 8$$
$$7 \to 4, 8$$
$$8 \to 9$$
$$9 \to 5$$

καθώς και οι τιμές:

SimRank decay factor: $C = 0.8$, RWR restart probability: $\alpha = 0.85$, Αρχικά similarity (SimRank): $s(a, a) = 1$, όλοι οι άλλοι: 0, RWR ξεκινά από ένα ή περισσότερους seed nodes (επιλογή δική σας)

Επιλέξτε 2–3 seed nodes για το RWR και αιτιολογήστε την επιλογή σας (πώς επηρεάζει η θέση τους την κατανομή πιθανοτήτων;). Πως θα αλλάξει η επιλογή των seeds αν θέλουμε να εντοπίσουμε κόμβους "κεντρικούς" ή "απομονωμένους";

SimRank

Υπολογίστε την αρχική similarity για τα ζεύγη: $(2, 5), (3, 6), (4, 7), (6, 8), (7, 9)$ για 2 iterations. Δείξτε όλα τα βήματα χρησιμοποιώντας τον τύπο:

$$s(a, b) = \frac{C}{|I(a)||I(b)|} \sum_{i \in I(a)} \sum_{j \in I(b)} s(i, j)$$

όπου $I(a)$ οι εισερχόμενοι κόμβοι του a.

Εξηγήστε ποιοι κόμβοι φαίνονται πιο όμοιοι και γιατί. Σκεφτείτε ζεύγη που δεν συνδέονται άμεσα αλλά έχουν υψηλό SimRank. Τι σημαίνει για το γράφημα;

Random Walk with Restart (RWR)

Χρησιμοποιώντας τα seed nodes που επιλέξατε, υπολογίστε 1 iteration των πιθανοτήτων για όλους τους κόμβους.

Ποιοι κόμβοι είναι πιο "κοντά" στα seeds;

Ποιοι κόμβοι μπορεί να έχουν υψηλό PageRank αλλά χαμηλή πιθανότητα RWR; Τι συμπέρασμα βγάζουμε;

Σύγκριση

Συγκρίνετε τα αποτελέσματα SimRank και RWR: Πότε οι δύο μετρικές συμφωνούν για ζεύγη κόμβων; Πότε διαφέρουν; Ποιες δομές του γράφου προκαλούν διαφορές; Προσπαθήστε να βρείτε μια προσέγγιση για να επιλέγετε seeds έτσι ώστε η RWR να εντοπίζει πιο αξιόπιστους ή σημαντικούς κόμβους.

Άσκηση 4.

Να λύσετε τις ασκήσεις 5.2.2, 5.2.3, 5.2.4, 5.3.1, 5.4.1 και 5.4.2 του βιβλίου MMDS.

Προθεσμία υποβολής και οδηγίες. Οι απαντήσεις θα πρέπει να υποβληθούν έως τις 10/5/2026, σε ηλεκτρονική μορφή. Συνιστάται *θερμά* να αφιερώσετε ικανό χρόνο για να λύσετε τις ασκήσεις μόνοι σας προτού καταφύγετε σε οποιαδήποτε *θεμιτή* βοήθεια (διαδίκτυο, βιβλιογραφία, συζήτηση με συμφοιτητές). Σε κάθε περίπτωση, οι απαντήσεις θα πρέπει να είναι *αυστηρά* ατομικές (δηλαδή όχι 'copy-paste'). Για να βαθμολογηθείτε θα πρέπει να παρουσιάσετε σύντομα τις λύσεις σας σε ημέρα και ώρα που θα ανακοινωθεί αργότερα.

Καλή επιτυχία!