ΔΙΑΧΕΙΡΙΣΗ ΔΕΔΟΜΕΝΩΝ ΜΕΓΑΛΗΣ ΚΛΙΜΑΚΑΣ
Περικλής Ανδρίτσος
1/7/2024

ΟΔΗΓΙΕΣ

- Η εξέταση γίνεται με ανοιχτά βιβλία και σημειώσεις.
- Γράψτε το όνομα σας και τον Α.Μ. σας στην κόλλα που θα σας δοθεί.
- Αναφέρετε αν είσαστε φοιτητής ή φοιτήτρια ΤΕΙ ή ΠΑΔΑ
- Οι τελικές απαντήσεις σας δεν πρέπει να είναι γραμμένες με μολύβι.
- Στο τέλος της εξέτασης παραδώστε τόσο την εκφώνηση όσο και την κόλλα.
- **ΔΙΑΡΚΕΙΑ: 1 ώρα και 15 λεπτά.**

ΕΠΩΝΥΜΟ: _____ ΟΝΟΜΑ: _____ ΑΡ. ΜΗΤΡΩΟΥ: _____ ΕΧΩ ΠΑΡΑΔΟΣΕΙ ΕΡΓΑΣΙΑ ΣΕ ΠΑΛΑΙΟΤΕΡΟ ΕΞΑΜΗΝΟ ΧΡΟΝΙΑ : _____ ΒΑΘΜΟΣ :_____ ΤΕΙ: _____. ΠΑΔΑ: _____	Πρόβλημα 1 _____ (από 2) Πρόβλημα 2 _____ (από 2) Πρόβλημα 3 _____ (από 2) Πρόβλημα 4 _____ (από 2) Πρόβλημα 5 _____ (από 2) ΣΥΝΟΛΟ _____ (από 10)

ΠΡΟΒΛΗΜΑ 1 (2 μονάδες)

Σας δίνονται οι πραγματικές και οι προβλεπόμενες τιμές ενός προβλήματος παλινδρόμησης.

Πραγματικό: [12,15,18,21,24,27,30,33,36,39]
Προβλέφθηκε: [10,16,19,20,25,26,32,34,35,40]

Σημειώστε το σωστό Root Mean Squared Error (RMSE) για την αξιολόγηση του μοντέλου.

α) 0.095 β) 0.950 γ) 1.950 δ) 1.265 ε) Καμία από τις εναλλακτικές

ΠΡΟΒΛΗΜΑ 2 (2 μονάδες)

Θεωρήστε δύο διανύσματα, A = [2, 1, 3] και B = [1, 2, 2]. Υπολογίστε την ομοιότητα συνημίτονου (cosine similarity) μεταξύ των διανυσμάτων A και B.

α) 0,964 β) 0,816 γ) 0,891 δ) 0,707 ε) Καμία από τις εναλλακτικές

ΠΡΟΒΛΗΜΑ 3 (2 μονάδες)

Θεωρείστε τα παρακάτω τρία κείμενα εκφρασμένα ως σύνολο των shingles τους:
D1: {a, b, c} D2: {a, c, d} D3: {b, c, e}

Αφού κατασκευάσετε τον πίνακα shingles-by-documents θεωρώντας ότι:
- η γραμμή 1 του πίνακα αντιστοιχεί στο shingle a
- η γραμμή 2 του πίνακα αντιστοιχεί στο shingle b
- η γραμμή 3 του πίνακα αντιστοιχεί στο shingle c
- η γραμμή 4 του πίνακα αντιστοιχεί στο shingle d
- η γραμμή 5 του πίνακα αντιστοιχεί στο shingle e

επιλέξτε τον πίνακα υπογραφών (signature matrix) που αντιστοιχεί στην μετάθεση των γραμμών {4, 1, 2, 5, 3}

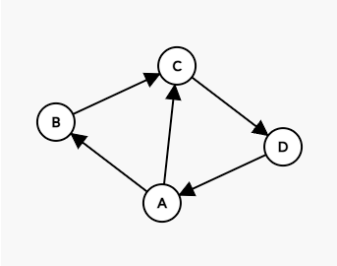
a)
D1	D2	D3
2	2	1

b)
D1	D2	D3
1	1	1

c)
D1	D2	D3
1	2	1

d)
D1	D2	D3
2	1	1

ΠΡΟΒΛΗΜΑ 4 (2 μονάδες)
Θεωρήστε έναν κατευθυνόμενο γράφο με τέσσερις κόμβους (A, B, C, D) και την ακόλουθη δομή:



Θέλουμε να υπολογίσουμε τις τιμές του PageRank για αυτούς τους κόμβους χρησιμοποιώντας τη μέθοδο του Power Iteration. Σας δίνονται οι αρχικές τιμές PageRank: [0,4, 0,2, 0,3, 0,1]

Για να υπολογίσουμε τις ενημερωμένες τιμές του PageRank μετά από μία επανάληψη, χρησιμοποιούμε τον τύπο:

$$PR(t+1) = (1 - d) + d * (PR(t) / \text{Outgoing Links})$$

όπου PR(t+1) είναι η ενημερωμένη τιμή PageRank, PR(t) είναι η τρέχουσα τιμή PageRank, OutgoingLinks είναι ο αριθμός των εξερχόμενων ακμών από τον κόμβο και d είναι ο παράγοντας απόσβεσης (dumping factor).
Υποθέστε έναν παράγοντα απόσβεσης 0,85 και υπολογίστε τις ενημερωμένες τιμές PageRank μετά από μία επανάληψη.

α) [0,32, 0,32, 0,405,0,235] β) [0,32, 0,32, 0,235, 0,405]
γ) [0,29, 0,29, 0,311, 0,145] δ) [0,252, 0,165, 0,275, 0,174] ε) Καμία από τις εναλλακτικές

ΠΡΟΒΛΗΜΑ 5 (2 μονάδες)

Σας δίνονται τα παρακάτω διανύσματα A και B:

```
A=[1,0,1,1,0,1]
B=[0,1,1,0,1,1]
```

Ποια είναι η τιμή της απόστασης Jaccard (Jaccard distance) των A και B ?
ΠΡΟΣΟΧΗ: Για να πάρετε όλες τις μονάδες γράψτε αναλυτικά τους υπολογισμούς σας στην κόλλα σας.

ΛΥΣΕΙΣ ΘΕΜΑΤΩΝ

ΠΡΟΒΛΗΜΑ 1 (2 μονάδες)

Για να υπολογίσουμε το Root Mean Squared Error (RMSE) μεταξύ των πραγματικών και των προβλεπόμενων τιμών, χρησιμοποιούμε τον ακόλουθο τύπο:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

όπου:

- y_i είναι οι πραγματικές τιμές,
- $\hat{y}_i$ είναι οι προβλεπόμενες τιμές,
- n είναι ο αριθμός των δεδομένων.

Πρώτα, υπολογίζουμε τις τετραγωνικές διαφορές για κάθε ζεύγος πραγματικών και προβλεπόμενων τιμών:

$$(12 - 10)^2 = 4$$
$$(15 - 16)^2 = 1$$
$$(18 - 19)^2 = 1$$
$$(21 - 20)^2 = 1$$
$$(24 - 25)^2 = 1$$
$$(27 - 26)^2 = 1$$
$$(30 - 32)^2 = 4$$
$$(33 - 34)^2 = 1$$
$$(36 - 35)^2 = 1$$
$$(39 - 40)^2 = 1$$

Συνοψίζουμε αυτές τις τετραγωνικές διαφορές:

$$4 + 1 + 1 + 1 + 1 + 1 + 4 + 1 + 1 + 1 = 16$$

Στη συνέχεια, βρίσκουμε το μέσο όρο αυτών των τετραγωνικών διαφορών διαιρώντας με τον αριθμό των δεδομένων (10):

$$\frac{16}{10} = 1.6$$

Τέλος, παίρνουμε την τετραγωνική ρίζα αυτού του μέσου όρου για να βρούμε το RMSE:

$$\text{RMSE} = \sqrt{1.6} \approx 1.265$$

Επομένως, το σωστό RMSE είναι περίπου 1.265 και η σωστή απάντηση είναι: δ) 1.265

ΠΡΟΒΛΗΜΑ 2 (2 μονάδες)

Η ομοιότητα συνημίτονου (cosine similarity) μεταξύ δύο διανυσμάτων Α και Β υπολογίζεται ως:

$$\text{cosine similarity} = \frac{A \cdot B}{\|A\|\|B\|}$$

1. Υπολογίζουμε το εσωτερικό γινόμενο των Α και Β:

$$A \cdot B = 2 \cdot 1 + 1 \cdot 2 + 3 \cdot 2 = 2 + 2 + 6 = 10$$

2. Υπολογίζουμε το μέγεθος του διανύσματος Α :

$$\|A\| = \sqrt{2^2 + 1^2 + 3^2} = \sqrt{4 + 1 + 9} = \sqrt{14}$$

3. Υπολογίζουμε το μέγεθος του διανύσματος Β ό:

$$\|B\| = \sqrt{1^2 + 2^2 + 2^2} = \sqrt{1 + 4 + 4} = \sqrt{9} = 3$$

4. Υπολογίζουμε την ομοιότητα συνημίτονου:

$$\text{cosine similarity} = \frac{A \cdot B}{\|A\|\|B\|} = \frac{10}{\sqrt{14} \cdot 3} = \frac{10}{3\sqrt{14}}$$

Κάνοντας πράξεις η ομοιότητα ισούται με 0.891 και η σωστή απάντηση είναι η γ)

ΠΡΟΒΛΗΜΑ 3 (2 μονάδες)

Για να βρούμε τον πίνακα υπογραφών (signature matrix) που αντιστοιχεί στην μετάθεση των γραμμών {4, 1, 2, 5, 3}, ακολουθούμε τα παρακάτω βήματα:

1. **Κατασκευή πίνακα shingles-by-documents**:

   - D1: {a, b, c}
   - D2: {a, c, d}
   - D3: {b, c, e}

   Ο πίνακας shingles-by-documents (sparse matrix) είναι:

	$D1$	$D2$	$D3$
a	1	1	0
b	1	0	1
c	1	1	1
d	0	1	0
e	0	0	1

2. **Κατασκευή πίνακα shingles-by-documents**:

   Για κάθε κείμενο, διατρέχουμε την αντίστοιχη στήλη του παραπάνω πίνακα ακολουθώντας τη σειρά γραμμών που δίνεται από τη μετάθεση {4, 1, 2, 5, 3}

   1. Κείμενο D1: Πρώτη γραμμή που έχουμε μη μηδενική τιμή είναι στη γραμμή 1 (της μετάθεσης)
   2. Κείμενο D2: Πρώτη γραμμή που έχουμε μη μηδενική τιμή είναι στη γραμμή 2 (της μετάθεσης)
   3. Κείμενο D3: Πρώτη γραμμή που έχουμε μη μηδενική τιμή είναι στη γραμμή 1 (της μετάθεσης)

   Σωστή απάντηση είναι η c)

ΠΡΟΒΛΗΜΑ 4 (2 μονάδες)

PR(A) = (1-0.85)+0.85*(0.4/2) = 0.32
PR(B) = (1-0.85)+0.85*(0.2/1) = 0.32
PR(C) = (1-0.85)+0.85*(0.3/1) = 0.405
PR(D) = (1-0.85)+0.85*(0.1/1) = 0.234

Σωστή απάντηση είναι η α)

ΠΡΟΒΛΗΜΑ 5 (2 μονάδες)

Ο τύπος είναι

$$d_J(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

Η τομή ισούται με τον αριθμό των θέσεων στα διανύσματα όπου και τα δύο είναι 1
Η ένωση ισούται με τον αριθμό των θέσεων στα διανύσματα όπου τουλάχιστον ένα από τα δύο είναι 1

Ένωση = 2, Τομή = 6, άρα: 1-(2/6) = 4/6 = 2/3 = 0.67