אלגו' בביולוגיה חישובית (76558)
תרגיל 1: עימוד רצפים ותכנון דינמי
תאריך הגשה: 17/11/2024

1. מספר העימודים האפשריים

כמוטיבציה לחיפוש אלגוריתמים יעילים לעימוד רצפים, נרצה להעריך את גודל מרחב החיפוש, כלומר להעריך את מספר העימודים האפשריים בין שני רצפים s, t, באורך n בסיסים כל אחד.
הוכיחו כי מספר העימודים האפשריים הוא לפחות אקספוננציאלי ב-n.

2. סגמנטציה של סדרת נתונים

בשאלה זו, נרצה לפתח אלגוריתם תכנון דינמי לסגמנטציה של סדרת נתונים למקטעים (סגמנטים).
תהי X סדרה באורך n:

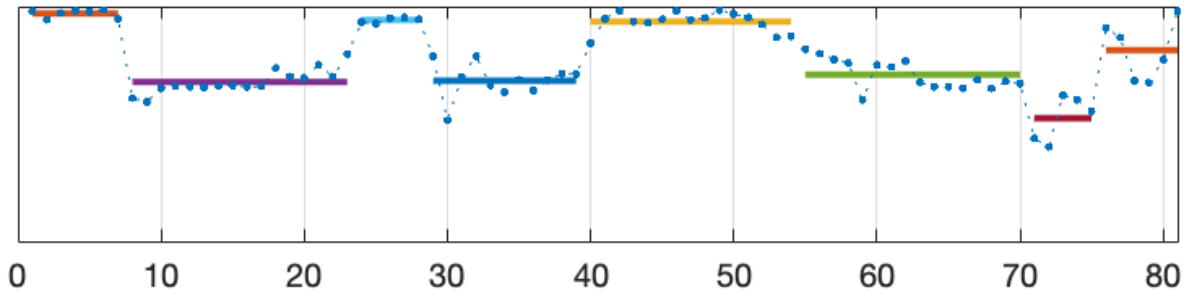
$$\{x_i\}_1^n \in \mathbb{R}$$

ונרצה לחלק את הסדרה ל-k סגמנטים רצופים שיכסו את הסדרה במלואה:

$$[x_1, x_a]_1, [x_{a+1}, x_b]_2, [x_{b+1}, x_c]_3, [x_{c+1}, x_d]_4, \ldots, [x_{f+1}, x_n]_k$$

באופן שימזער את סכום ריבועי המרחקים של כל נקודה x_i מהערך הממוצע μ_j בסגמנט אליה היא שייכת, ובתוספת קנס רגולריזציה p לכל בלוק.

$$\arg\min \sum_1^n (x_i - \mu_j)^2 + k \cdot p$$



לשם כך נשים לב, שהפתרון האופטימלי לסדרה באורך n, מושג על ידי שרשור הפתרון האופטימלי לסדרה באורך m כלשהו (m<n) עם הסגמנט האחרון בסדרה $[x_{m+1},x_n]$. באופן דומה, הסקור יהיה סכום הסקור האופטימלי עד m, בתוספת ריבועי המרחקים עבור הסגמנט האחרות, ועוד הקנס p. לשם היעילות, ניתן להניח ש**אורך כל סגמנט חסום בגודל q כלשהו** (משתנה קלט).

עליכם לכתוב תכנית:

```
segs, cost = segment(x, p, q)
```

שתקבל כקלט את הסדרה x, את הקנס p, ואת גודל הסגמנט המקסימלי q, ותחזיר טבלה k x 3, עם התחלה, הסוף והערך הממוצע של כל סגמנט, ועם הסקור הכללי של הסדרה.

מומלץ לשמור מערך c בגודל n המכיל את הציון האופטימלי לסגמנטציה של הרישא באורך i של הסדרה x, וכן מערך t בגודל n המכיל את ה-traceback. כלומר הערך במקום ה-i, יהיה המיקום של התחלת הסגמנט האחרון (זה שמסתיים בעמדה i) בפתרון האופטימלי של הרישא.

לאחר שתמלאו את שני המערכים מההתחלה ועד לסופם, תוכלו להשתמש במערך t כדי לשחזר את הסגמנטציה האופטימלית, מהסוף אחורה.

ניתוח זמן ריצה תיאורטי ומעשי, כולל ניתוח מעמיק של תזמוני ריצות שונות, כפונקציה של n, של אורך הסגמנט המקסימלי q, ושל מספר הסגמנטים k. אנא דיגמו להנאתכם סדרות סינתטיות עם פרמטרים שונים, והוסיפו גרפים מפורטים.

בונוס (5 נק') - סגמנטציה למידע רב-ערוצית. שנו את התכנית כך שתוכל לקבל מטריצה x, ובה d סדרות מתואמות באורך n כל אחת, אשר מייצגות מידע סדרתי רב-ערוצי. הסתכלות שקולה, כל איבר בסדרה הוא d-מימדי:

$$\{x_i\}_1^n \in \mathbb{R}^d$$

במקרה זה, הסגמנטציה תהיה משותפת לכל d הסדרות, אולם יהיו d ערכי ממוצע $\{\mu_j\}$ לסגמנט ה-j (אחד לכל אחד מ-d הערוצים/המימדים), והסקור הכללי יהיה שווה לסכום הסקורים של d הסדרות.

ניתן להגיש את התרגיל **בזוגות** (ולא יותר).

יש לעבוד על התרגיל באופן עצמאי (ולא להעתיק).

את התשובות יש להגיש בקובץ ex1.tar.gz הכולל:

- קובץ בשם ex1.pdf ובו תשובות לשאלות התיאורטיות
- קובץ בשם segment.py עם התכנית שכתבתם (ניתן להוסיף קבצי python נוספים).

אנא הקפידו על **תשובות מלאות ופורמליות** ועל **תיעוד והסברים ברורים** בגוף הקוד. יש לשים דגש על נכונות הקוד, על אלגנטיות, ועל יעילות האלגוריתם. זמן הריצה הצפוי, לקלט באורך אלפי בסיסים, הוא שניות בודדות (או פחות).

השימוש בכלי עזר לתיכנות מבוססי בינה מלאכותית (דוגמת copilot) מותר (בחלק התכנותי בלבד!), אולם עליכם **להצהיר על כך**, לפרט באיזה כלים השתמשתם, מה הפרומפטים שהכנסתם, ולתאר במילים באופן מפורט את תהליך העבודה על התרגיל. אין להשתמש בכלי AI כדי לענות על השאלות התיאורטיות.

דרישות טכניות

את הקובץ segment.py יהיה ניתן להפעיל מהטרמינל באמצעות הפקודה

```
python3 segment.py --filepath in.txt --penalty 0.5 --maxlen 10
```

על התוכנה להדפיס את הסגמנטציה כך שכל סגמנט מופיע בשורה נפרדת (מיקום התחלה, מיקום הסוף) לצד הערך הממוצע של הסגמנט (מופרדים ברווחים):

1 9 0.945
10 22 0.631
23 29 0.953
...
40.238

וכן הלאה, ובעקבותם הסקור הכללי של הסדרה (מודפס בשורה האחרונה של הפלט).

הערות

- יש להדפיס את הערכים המספריים בדיוק של 3 ספרות אחרי הנקודה.
- מומלץ להיעזר בשלד התרגיל, אשר מכיל פונקציות לקריאת הקבצים ולהדפסת התוצאות.
- מצורף קובץ דאטה לדוגמה, וקובץ נוסף שמתאים לסעיף הבונוס.