מסחר אלקטרוני – 096211
תרגיל בית 2 – מערכות המלצה

מקרא שינויים

ירוק – תיקונים לאחר שעת קבלה 21.7 והערות של סטודנטים.

רקע

לאחר הצלחתכם המסחררת בתרגיל בית 1, פנתה אליכם הפעם חברת הענק "MePipe". לחברה יש אפליקציה בשם "MePipe Video" בעזרתה משתמשים צופים בסרטונים. אפליקציית MePipe Video גם ממליצה למשתמשים על סרטונים חדשים שהיא משערת שהם יאהבו. MePipe ביקשה שתסייעו לה ביצירת מערכת המלצת סרטונים חדשה.

לצורך ביצוע המשימה קיבלתם מסד נתונים של MePipe Video המציג נתונים אודות מספר השניות שמשתמשים ראו מסרטונים שונים.

משימה חלק 1

עליכם להשתמש בנתוני הצפיות שקיבלתם על מנת לחזות עבור כל זוג של משתמש וסרטון המופיעים בקובץ test.csv כמה שניות המשתמש יראה מהסרטון.

שימו לב! ככל שמספר השניות שמשתמש צפה בסרטון מסוים גדול יותר, סימן שהמשתמש אוהב את הסרטון יותר. לכן, ניתן להתייחס למספר השניות שהמשתמש צפה בסרטון כאל דירוג.

הסבר על הקבצים

כאמור, לצורך פתירת התרגיל קיבלתם שני מסדי נתונים:

- הקובץ user_clip.csv מתאר את הצפיות. הוא מכיל שלוש עמודות –
  - user_id - מזהה משתמש.
  - clip_id - מזהה הסרטון.
  - weight - מספר השניות שהמשתמש ראה את הסרטון.

בנוסף לקובץ הנתונים, קיבלתם את הקובץ test.csv המכיל את זוגות המשתמשים והסרטונים עבורם תצטרכו לחזות את מספר שניות הצפייה. גם קובץ זה מכיל שתי עמודות:

- user_id - מזהה משתמש.
- clip_id - מזהה הסרטון.

שימו לב כי עבור כל זוג של סרטון ומשתמש **שאינו** מופיע באף אחד מהקבצים user_clip ו-test מספר שניות הצפייה הוא 0.

הערכה וציון חלק 1

בחלק זה של תרגיל הבית תקבלו שתי משימות הכוללות ביצוע מדויק של שיטה שנראתה בביתה. עבור כל משימה תגישו קובץ פתרון (csv) נפרד כמתואר בחלק הבא. קובץ הפתרון יכיל את החיזויים שלכם עבור כל אחד מהזוגות של משתמש וסרטון מקובץ test.
הבהרה לשתי המשימות– הובהר בפורום לפני שעת הקבלה שיש להפוך תחזיות שליליות לאפסים כי אין משמעות לערכים שליליים במספר שניות ולכן נעקוב אחרי זה ולא אחרי מה שגור אמר בשעת הקבלה, שכן זה הובהר מוקדם יותר.

משימה 1:

נסמן ב-$\hat{r}_{u,i}$ את החיזוי שלכם עבור מספר השניות שמשתמש u ראה את סרטון i, וב-$r_{u,i}$ את הערך האמיתי.
החיזוי שלכם עבור הזוג (u,i) מקובץ ה-test הוא:

$$\hat{r}_{u,i} = r_{avg} + b_u + b_i$$

כאשר b_u, b_i הם משתנים המייצגים את ההטיות של המשתמשים השונים. נדגיש כי אלו לא בהכרח ההטיות שנראו בתרגולים וצריכות להלמד מהנתונים. בשביל לעשות את זה, החלטתם לבחור בפרמטרים הנ"ל שימזערו את הפונקציה f_1 המוגדרת למטה – שילוב ממושקל של MSE ורגולריזציה על גודל המשתנים.
עליכם למצוא את הפרמטרים שמביאים את פונקציית המטרה למינימום ולהשתמש בהם על מנת לבצע את החיזויים. דווחו את הערך של f_1 עבור הפרמטרים שבחרתם בקובץ ה pdf אותו אתם מגישים.
שימו לב – עבור סרטון או משתמש שנראים לראשונה בטסט בלי שנראו באימון, אתם יכולים להניח שהמשתנה המתאים להם הוא 0 ולעשות לפי זה את החיזוי.

$$f_1 = \sum_{(u,i)\in train} (r_{u,i} - r_{avg} - b_u - b_i)^2 + 0.1 \cdot \left(\sum_{u\in train} {b_u}^2 + \sum_{i\in train} {b_i}^2 \right)$$

בחיזוי ובפונקציית המטרה במשימה זו עליכם לא להתייחס לערכים חסרים, שכן זה שמשתמש לא ראה סרטון לא אומר שהוא לא אהב אותו.

משימה 2:

במשימה זו תעריכו את מספר שניות הצפייה של כל אחד מהזוגות באמצעות פירוק SVD וקירוב מדרגה נמוכה.
יש לבצע את הקירוב למימד בגודל $k = 20$. השתמשו בפירוק SVD כפי שנלמד בביתה כדי לייצר את האמדים $\hat{r}_{u,i}$.
העריכו את התחזיות שלכם בעזרת הפונקציה הבאה ודווחו את הערך שלה בקובץ ה-pdf.

$$f_2 = \sum_{(u,i)\in train} (\hat{r}_{u,i} - r_{u,i})^2$$

במשימה זו עליכם למלא חלקים חסרים במטריצת ה-(users, clips) באפסים.
שימו לב שהמטרה של שתי המשימות הוא מימוש מדויק של האלגוריתמים שהוצגו בהרצאה.
נזכיר כי המימוש שנראה בהרצאה כולל:

1. פירוק SVD של מטריצת הדירוגים החסרה כאשר שמים 0 בכניסות החסרות.
2. הורדה במימד המטריצה כפי שנראה בהרצאה.

3. התחזית לזוג של סרטון ומשתמש תהיה הכניסה במטריצה החדשה, כאשר המוטיבציה היא הורדת רעש מהדירוגים.

הציון של כל משימה הוא 25 נקודות, מימוש נכון יביא לניקוד מלא במשימה.

משימה חלק 2

חוקים חדשים של פרטיות המשתמש מגבילים את השימוש במודלים שעשיתם בשלב הראשון, לכן החברה פנתה לחברה חיצונית (לא אתם) שאחראית על איסוף נתונים "אנונימיים". חברה זו הצליחה ליצור אשכולות (קבוצות) של משתמשים לפי ההתנהגות שלהם ברשת, כאשר שני משתמשים יהיו באותו אשכול כאשר ההעדפות שלהם דומות. כעת, במקום התנהגות של משתמשים בודדים (המודל עליו דיברנו בהרצאה), ניתנת לכם התנהגות של אשכולות. בפרט, החברה החיצונית הצליחה לחזות נכונה את ההסתברות שמשתמש ששייך לאשכול מסויים יאהב קטגוריית סרטונים כלשהי. הסט של קטגוריות הסרטונים מסופק לכם גם הוא על ידי החברה החיצונית.

במשימה זו יש משתמשים מסוגים מסויימים, ומשימתכם היא להמליץ על ז'אנרים לכל משתמש חדש, שאת האשכול אליו הוא שייך אתם לא יודעים בוודאות. אתם כן יודעים את ההסתברות שהמשתמש שהגיע שייך לכל אשכול משתמשים.

מודל ההמלצה מוגדר כך: משתמש חדש (מסוג לא ידוע) פותח את MePipe Video ומחפש סרטונים לראות. מערכת ההמלצה (אותה תממשו) מממליצה למשתמש על ז'אנר אחד בכל שלב ומציגה לו סרטון מאותו ז'אנר. המשתמש נותן לייק לסרטון שקיבל בהסתברות מסוימת. אם המשתמש נתן לייק, הוא יישאר במערכת ותוכלו להמליץ על סרטון נוסף. אחרת, קיימת הסתברות מסויימת (התלויה בסוג המשתמש שהגיע והז'אנר שהומלץ) שהמשתמש יעזוב את המערכת. לאחר 15 המלצות המשתמש עוזב את המערכת בין אם הוא נתן לייק לסרטון האחרון ובין אם לא.

המטרה: למקסם את תוחלת הלייקים של משתמשים. תוחלת הלייקים של המשתמשים השונים מייצגת את הרווחה החברתית של המשתמשים במערכת, כלומר – את הסכום הממושקל של שביעות הרצון של המשתמשים, כאשר המשקול נקבע לפי גודל האשכולות השונים. התוחלת נלקחת על פני ההסתברות של סוגי משמשים שונים להגיע למערכת, ועל הסתברויות הלייקים ועזיבות של משתמשים לכל ז'אנר.

הסבר על הקבצים

במשימה זו תקבלו שני קבצים:

1. ID1_ID2_part2.py – בקובץ זה נמצאת המחלקה Recommender שבה תממשו את מערכת ההמלצה שלכם. במחלקה תצטרכו לממש שלוש פונקציות:
   - __init__ - בפונקציה זו תבצעו את מירב החישובים שלכם לגבי המלצת הז'אנרים במהלך סימולציה. הפונקציה מוגבלת בשתי דקות ריצה. מתוך הבנה שמבנה המטריצה עשוי להיות גורם חשוב עבור מבנה ההמלצות, תוכלו לבצע חישוב לא מקוון כדי להחליט על סדרת ההמלצות
   - recommend – בפונקציה זו תמליצו למשתמש על ז'אנר. זמן הריצה של הפונקציה מוגבל בעשירית שנייה.
   - update – פונקציה זו מקבלת את תוצאת הסיבוב הקודם, ומשמשת לעדכון המערכת שלכם בתוצאה שלו. זמן הריצה של הפונקציה מוגבל בעשירית שנייה.

שתי הפונקציות האחרונות פועלות "בזמן אמת" ולכן צריכות להיות בעלות זמני ריצה מהירים. אם תחרגו מזמן זה, המשתמש יצא מהמערכת לפני שתספיקו להמליץ על פריט או לנתח את תגובתו.

2. simulate.py – קובץ המכיל סימולציה של מערכת ההמלצה, הקובץ מייבא את המחלקה Recommender מ-submit.py ומריץ סימולציה יחידה של מערכת ההמלצה שלכם.

הפונקציה __init__ מקבלת שתי מטריצות L, S ווקטור p, המגדירים מקרה פרטי של הבעיה.

- L – כניסה (i,j) במטריצה זו מייצגת את ההסתברות שמשתמש מסוג j ייתן לייק לז'אנר i.
- S - כניסה (i,j) במטריצה זו מייצגת את ההסתברות שמשתמש מסוג j יישאר במערכת אם הוא לא נתן לייק לז'אנר i.
- p – פריור על סוגי המשתמשים, הושג על ידי מידע דמוגרפי בסיסי של המשתמש. כניסה i בוקטור זה מייצגת את ההסתברות שהמשתמש שהגיע למערכת הוא מסוג i.

פרמול פונקציית המטרה: נגדיר את המשתנים המקריים הבאים:

$$M - distribution\ over\ the\ users\ clusters$$
$$L^{m,r} - distibution\ of\ likes\ from\ user\ of\ cluster\ m\ and\ recommendation\ strategy\ s$$

המטרה שלכם – למצוא את אסטרגיית ההמלצות s שתמקסם את:
$$E_{m\sim M}E[L^{m,r}]$$

לנוחותכם, ננתח כעת דוגמאות בסיסיות הלקוחות מתוך הקובץ *simulate.py*, שאותן תצטרכו לפתור.

<u>דוגמה 1:</u>

```
# Instance 1
L1 = np.array([[0.8, 0.7, 0.6], [0.79, 0.69, 0.59], [0.78, 0.68, 0.58]])
S1 = np.array([[0.56, 0.46, 0.36], [0.55, 0.45, 0.35], [0.54, 0.44, 0.34]])
p1 = np.array([0.35, 0.45, 0.2])
```

בדוגמה זו, ההסתברות שמשתמש מסוג (1) יאהב את קטגוריית הפריטים מסוג (2) היא הכניסה ה *(2,1)* במטריצה *L* – 0.79. באופן דומה, ההסתברות שמשתמש מסוג (3) ישאר במערכת לאחר שלא אהב פריט מקטגורייה (1) היא 0.36. ההסתברות שהמשתמש שהגיע למערכת הוא מאשכול (2) היא הכניסה השנייה בפריור – 0.45. נתקדם צעד אחד קדימה ונגיד כי ההסתברות שהמשתמש עם הפריור הנוכחי יאהב פריט מקטגוריה (1) ניתנת לחישוב לפי התנייה על סוג המשתמש, כלומר המכפלה הפנימית של השורה הראשונה של *L* עם הפריור.
בדוגמה זו, ניתן לראות כי הפעולה הראשונה יותר טובה משאר הפעולות– היא גם מבטיחה לייק בהסתברות יותר גבוהה משאר הפעולות לכל אשכול משתמשים, וגם ההסתברות שהמשתמש ישאר במערכת הינה יותר גדולה משאר הפעולות לכל אשכול משתמשים.

דוגמה 2:

```
# Instance 2
L2 = np.array([[0.9, 0.75], [0.64, 0.5]])
S2 = np.array([[0.2, 0.4], [0.7, 0.8]])
p2 = np.array([0.3, 0.7])
```

בדוגמה זו ההסתברויות שמשתמש מאשכול מסויים יאהב פריט מקטגוריה מסויימת מחושבת בצורה דומה לדוגמה הקודמת, אך פה אפשר לראות מקרה מעניין – הפעולה הראשונה מבטיחה לייק מהמשתמש בהסתברויות יותר גבוהות מהפעולה השנייה, אך במידה והמשתמש לא יאהב את הפריט שהומלץ לו, ההסתברויות שהמשתמש ישאר במערכת יותר נמוכות מאשר בפעולה השנייה. זה חשוב מכיוון שאם המשתמש נשאר במערכת הוא עדיין יכול לאהוב פריטים בתורים הבאים, ואם הוא עוזב את המערכת, כבר לא נקבל ממנו יותר לייקים. לכן בעת המלצה צריך לחשוב על הטריידאוף הזה ולנתח אותו בצורה שתביא את פונקציית המטרה למקסימום.

הערכה וציון חלק 2

עליכם לממש מחלקת Recommender שתעבוד לכל מקרה פרטי של הבעיה (כלומר תעבוד עבור L, S, p לא ידועים). ההערכה של המחלקה שלכם תתבצע באופן הבא:
בקובץ simulate.py יש חמישה מקרים פרטיים של הבעיה. עליכם לעבור ערך סף עבור כל אחד מהמקרים:

1. 4.65
2. 5.6
3. א. 12.4, ב. 6.1, ג. 6.77
4. 5.43
5. 6.4

סך הניקוד עבור תוחלת גדולה מערך הסף בכל המקרים הפרטיים הוא 25 נקודות.

בנוסף, עבור מקרה מספר 5 יהיה חלק תחרותי של 10 נקודות, שבו נשווה את הביצועים שלכם אחד מול השני. חוץ מזה, נעשה תחרות של 15 נקודות (+ בונוסים) בין מערכות ההמלצה שלכם על מקרים שלא נחשוף לכם. אתם יכולים להניח שלא נריץ את מערכת ההמלצה שלכם על מטריצות גדולות מ-10x10.

הגשה

ההגשה מפוצלת לשני עמודי הגשה במודל, אחד לחלק הראשון של התרגיל (משימת הפרדיקציה) ואחד לחלק השני (משימת ההמלצה הסדרתית).
בהגשה של החלק הראשון עליכם להגיש:

1. שני קבצי csv עבור המשימות בחלק הראשון של התרגיל. על הקבצים להיות באותו פורמט בדיוק כמו זה של קובץ הדוגמא המצורף ושמם ID1_ID2_taskX.csv כאשר ID1 ו-ID2 הם מספרי תעודת הזהות

של המגישים (עבור סטודנטים המגישים לבד שם הקובץ צריך להיות ID1.csv), ובמקום הספרה X מצוין מספר המשימה בהתאם למספור המופיע בחלק הקודם.

2. קובץ py שכותרתו ID1_ID2_part1.py המכיל את הפתרון שלכם לחלק הראשון של התרגיל. שימו לב כי הקוד שאתם מגישים צריך לסיים את ריצתו תוך שעתיים (בסה"כ), ולייצר בדיוק את קבצי ה-csv של שתי המשימות.
3. קובץ pdf שבו כתובים הערכים של פונקציות המטרה f_1, f_2 (לא אמור להיות יותר מ2 שורות).

בהגשה של החלק השני עליכם להגיש:

1. קובץ הסבר בפורמט pdf, ששמו ID1_ID2.pdf, באורך של **לכל היותר עמוד**, הכולל את ההסבר שלכם לפתרון החלק השני של תרגיל הבית. הקובץ צריך לפרט את הגישה שלכם לפתרון הבעיה.
2. הקובץ שצורף למשימה ID1_ID2_part2.py לאחר שהשלמתם את המחלקה Recommender בתוכו.

כל הקבצים של כל אחד מהחלקים אמור להיות בתיקיית זיפ עם השם ID1_ID2_X, כאשר X זהו מספר החלק.