מסחר אלקטרוני – תרגיל בית 2 חלק 1

רקע והוראות כלליות

כזכור, מערכת המלצה עם קבוצת משתמשים U וקבוצות פריטים I, מקבלת סט של דירוגים $r_{u,i}$, ומטרתה לתת תחזית $\hat{r}_{u,i}$ לדירוגים חדשים. כמקובל, נקרא לדירוגים שמערכת ההמלצה מקבלת סט האימון (*train*), ולדירוגים שהמערכת אמורה לחזות סט המבחן (*test*).

באופן פורמלי, סט האימון של מערכת ההמלצה הוא אוסף דירוגים $r_{u,i}$ עבור כל זוג של משתמש וסרט $(u,i) \in train \subseteq U \times I$. סט המבחן הוא אוסף זוגות $(u,i) \in test \subseteq U \times I$ עבורם המערכת צריכה לתת תחזית $\hat{r}_{u,i}$.

בחלק זה תממשו שני אלגוריתמים למערכות המלצה שנלמדו בכיתה, ובעזרתם תצטרכו לחזות את הדירוגים של סט המבחן ולדווח את מדד ה MSE על סט האימון. הנתונים איתם תעבדו הם דירוגים של סרטים, שהם מספרים שלמים בין 1 ל-5.

שימו לב: סט האימון וסט המבחן זרים לחלוטין, כלומר אין חפיפה בין זוגות המשתמשים והפריטים שמופיעים בהם ($train \cap test = \emptyset$). יחד עם זאת, כל משתמש וכל פריט שמופיעים בסט המבחן מופיעים גם בסט האימון. כלומר, אין מקרים של משתמשים או פריטים חדשים (במציאות זה לא תמיד נכון, למשל במקרה של סרט חדש שיצא, או משתמש חדש שמצטרף).

כזכור, מדד ה *MSE* על סט האימון מוגדר על ידי:

$$\frac{1}{|train|} \sum_{(u.i) \in train} \left(r_{u,i} - \hat{r}_{u,i}\right)^2$$

משימה 1

בחלק זה עליכם לממש מודל המלצה פשוט, המבוסס על הטיות קבועות לכל משתמש ולכל פריט (ראו שקופיות 20-26 בהרצאה 5). את ההטיה של משתמש u נסמן ב- b_u, ואת ההטיה של פריט i נסמן ב- b_i. תחזית הדירוג של המשתמש u לסרט i מחושבת לפי הנוסחה:

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

כאשר r_{avg} הוא ממוצע הדירוגים בסט האימון. מטרת המודל היא לאמוד את הטיות המשתמשים והפריטים בכדי למזער את פונקציית השגיאה הבאה:

$$L_1 = \sum_{(u,i) \in train} \left(r_{u,i} - r_{avg} - b_u - b_i\right)^2 + \lambda\left(\sum_{u \in U} b_u^2 + \sum_{i \in I} b_i^2\right)$$

עליכם לאמוד את ההטיות הממזערות את L_1 עבור $\lambda = 1$, ובעזרתן לחזות את הדירוגים של סט המבחן בקובץ ולחשב את ה-MSE על סט האימון (בהמשך יתואר כיצד להגיש תחזיות אלה).

משימה 2

בחלק זה עליכם לבצע קירוב של מטריצת הדירוגים באמצעות פירוק SVD מדרגה נמוכה, עם דרגת קירוב $k = 10$ (ראו שקופיות 46-56 בהרצאה 5). עליכם להשלים את השלבים הבאים:

1) בניית מטריצת הדירוגים מתוך סט האימון, כך שכל שורה מייצגת משתמש וכל עמודה מייצגת פריט. השלימו דירוגים שלא נמצאים בסט האימון לאפס.
2) מציאת קירוב של המטריצה באמצעות SVD, תוך שימוש בדרגה נמוכה k = 10. השתמשו באלגוריתמים עבור מטריצות ספרסיות (sparse), דוגמת scipy.sparse.linalg.svd.
3) חישוב תחזיות הדירוגים לפי הכניסות הרלוונטיות במטריצה המקורבת.

כמו במשימה 1, עליכם לדווח את התחזיות שלכם ואת ה-MSE.

קבצים מצורפים

סט האימון נמצא בקובץ ה-train.csv csv בפורמט הבא:

```
user id,item id,rating
97,1436,5
97,1113,3
97,664,4
97,1608,3
97,69,3
97,815,4
97,295,1
97,262,5
97,143,2
```

סט הבדיקה נמצא בקובץ ה-test.csv csv בפורמט הבא:

```
user id,item id
97,25
97,149
97,431
97,1263
97,1177
```

בשני הקבצים, user id הוא המזהה הייחודי של משתמש, ו-item id הוא המזהה הייחודי של סרט. העמודה "rating" בקובץ train.csv (שאינה נמצא בקובץ test.csv) מכילה את הדירוג שהמשתמש הנתון נתן לפריט הנתון.

ניקוד

מימוש נכון של כל אחת מהמשימות מזכה ב-25 נקודות (סך הכל 50).

ספריות מותרות לשימוש

ניתן להשתמש בספריות הסטנדרטיות, ב-numpy, scipy ו-pandas בלבד.

דגשים חשובים

על התחזיות להיות בתחום של בין 1 ל-5. עבור סט המבחן, אם תחזיות חורגות מעבר ל-5 או מתחת ל-1 עליכם לתת תחזיות של 5 או 1, בהתאמה. לצורך חישוב ה-MSE על סט האימון, אין לעשות זאת. לתחזיות מותר להיות ערכים לא שלמים, ואין לעגל אותן.

הוראות הגשה

שימו לב - הגשה לא תקינה תכלול הורדת ציון!

עליכם להגיש תיקיית ZIP בשם HW2_PART1_ID1_ID2.zip המכילה את הקבצים הבאים בלבד.

- task1.py – קובץ המכיל את הקוד עם הפתרון של משימה 1.
- task2.py – קובץ המכיל את הקוד עם הפתרון של משימה 2.

בכל אחד מהקבצים האלה אמורה להיות פונקציה בשם solve(), שמכילה את עיקר הפתרון של אותה משימה.

- pred1.csv – התחזיות על סט המבחן עבור משימה 1
- pred2.csv – התחזיות על סט המבחן עבור משימה 2

על התחזיות להיות באותו פורמט כמו train.csv ולהכיל את הדירוגים של כל הזוגות מסט האימון (נמצאים ב- test.csv).

- mse.txt – קובץ המכיל את ה-MSE עבור סט האימון עבור שתי המשימות.

בקובץ זו צריכות להיות אך ורק שתי שורות, כאשר בשורה הראשונה רשום ציון ה-MSE של משימה 1 ובשורה השנייה רשום את ציון ה-MSE של משימה 2. מבנה קובץ לדוגמה:

```
1    1.0
2    2.0
```