שם הקורס: מודלים למסחר אלקטרוני (096211)

מגישים: נימרוד סולומון (ת.ז 206574733) ומתן שילוני (ת.ז 208634469)

מספר תרגיל הבית: 2

תאריך הגשה: 09.06.2022

הקדמה

מסמך זה מפרט על אודות הגישה שלנו למציאת התחזיות לכמות ההשמעות של משתמש את האמן, עבור כל זוג של משתמש ואמן שקיבלנו בקובץ test.csv. במסמך נתאר כלים מרכזיים בהם השתמשנו, את שלבי העבודה ואת התובנות המרכזיות אליהן הגענו במהלכה עד לחיזוי הסופי (של משימה 1) המצורף בקובץ 206574733_208634469_task1.csv.

הספרייה Surprise

Surprise (Simple Python RecommendatIon System Engine) הינה ספריית python לבניית מערכות המלצה. הספרייה כוללת מספר רב של אלגוריתמי חיזוי שונים מוכנים לשימוש כגון baseline algorithms, neighborhood methods, matrix factorization-based (such as SVD) ונוספים. להלן קישור לאתר הספרייה surpriselib.com.

במהלך העבודה נעזרנו במימושים שקיימים בספרייה זו עבור שיטות ה-Neighborhood וה-SVD שנלמדו בהרצאה, וכן במתודות שקיימות בספרייה לבחירת היפר-פרמטרים עבור המודלים.

משימה 1: יישום החומר שנלמד בשיעורים לטובת פתרון תרגיל הבית

1. טרנספורמציה על הדאטה

- משימת העל במשימה הראשונה הייתה למזער את פונקציית ה-loss הבאה:

$$L = \sum_{(u,i)\in Testset} (\log_{10}(r_{u,i}) - \log_{10}(\hat{r}_{u,i}))^2$$

זוהי וריאציה של SSE: אם נסמן $\hat{y}_{u,i} = \log_{10}(\hat{r}_{u,i})$, $y_{u,i} = \log_{10}(r_{u,i})$ נקבל את פונקציית ה-SSE: $SSE = \sum_i (y_i - \hat{y}_i)^2$

- בספריית Surprise כל המודלים תומכים ב-RMSE (אשר מזעורו שקול למזעור SSE כפי שראינו בהרצאה) כמדד לטובת, למשל, בחירת המודל הטוב ביותר במסגרת תהליך של Cross Validation.
- לפיכך, נדרשנו לבצע טרנספורמציה על הדאטה, על מנת לקבל דאטה מעובד שמזעור RMSE עליו יהיה שקול למזעור של פונק' המטרה L.
- לפיכך, על כל ערך בטבלה user_artist בעמודה weight הפעלנו log בבסיס 10. מכאן שמזעור RMSE על ערכי ה-weight החדשים יביא למזעור של L על הדאטה המקורי.

2. בחירת המודל

- לאחר שביצענו את הטרנספורמציה הנ"ל על הדאטה, היה עלינו לבחור את המודל הטוב ביותר עבורו, במובן של המודל שמשיג שגיאת RMSE נמוכה ככל שאפשר על סט האימון.
- לשם כך בדקנו מספר מודלים, בהם מודלי ה-Neighborhood וה-SVD שראינו בהרצאה, וכן מודלים נוספים המהווים וריאציות שונות למודלים הנ"ל. לכל מודל חיפשנו פרמטרים שימזערו כאמור את ה-RMSE.
- לכל מודל (SVD, Neighborhood, etc.) מצאנו את ההיפר פרמטרים הטובים ביותר באמצעות שיטת cross-validation בשם gridsearchCV , אשר בודקת את כל השילובים האפשריים של היפר

פרמטרים למודל, מתוך קבוצת אפשרויות מוגדרת מראש של ערכים. עבור הפרמטרים הטובים ביותר שמרנו את ערך ה-RMSE הממוצע על סט ה-validation שהשיג כל מודל עם הפרמטרים הללו.

- השווינו בין ערכי ה-RMSE הממוצעים ששמרנו עבור כל מודל ובחרנו את המודל שהשיג את הערך המינימלי. להלן התוצאות שחזרו מהרצת הקוד:

```
           name  bestScore
4         SVDpp   0.359660
3           SVD   0.367830
1   KNNBaseline   0.377385
0      KNNBasic   0.489570
2  KNNWithMeans   0.523175

Best algorithm after GridSearchCV: SVDpp
Best RMSE: 0.35965999105520535
Best params for best algorithm: {'rmse': {'n_epochs': 30, 'lr_all': 0.007, 'reg_all': 0.02, 'verbose': False}}
```

נשים לב שהערכים לעיל ממוינים בסדר עולה, לכן **האלגוריתם שהחזיר את התוצאות הטובות ביותר עבור הדאטה הנתון הוא SVDpp** (וריאציה של SVD, להרחבות - surprise.prediction_algorithms.matrix_factorization.SVDpp).

3. <u>חיזוי באמצעות המודל הנבחר</u>

- לאחר בחירת המודל, אימנו אותו על כל סט האימון ואח"כ ביקשנו לחזות את הערכים הנדרשים בקובץ ה-test.
- מכיוון שהערכים עליהם אימנו את המודל היו כאמור "מנורמלים" באמצעות Log בבסיס 10, לתחזיות שהוציא המודל ביצענו את הפעולה ההפוכה: לכל תחזית $p_{i,u}$ שנתן המודל החזרנו כתחזית הסופית לקובץ שאותו הגשנו את $10^{p_{i,u}}$.
- נציין כי תהליך ה-CV שביצענו בעת בחירת המודלים מזער את הסיכוי ל-overfit, שכן המודל למד עם פרמטרים שונים עבור חלקים שונים של סט האימון שסופק לנו, ולבסוף המודל הנבחר למד על הדאטה כולו.

<u>הערות נוספות</u>

1. לצערנו לא הספקנו לבצע את משימה 2 האופציונאלית, לכן ביצענו את משימה 1 בלבד.
2. בקובץ 206574733_208634469.py ניתן למצוא שני חלקים: חלק של בחירת המודל כמפורט מעלה באמצעות cross-validation וחלק של שימוש במודל שנבחר לביצוע החיזוי. **לחלק של ה-cross validation לוקח זמן רב (כשעתיים במחשב שלנו) להתבצע, לכן ככל שאין צורך להריצו ספציפית, מומלץ להשאיר חלק זה בהערה** במהלך הרצת הקובץ.