הפקולטה לניהול
ע"ש קולר
אוניברסיטת תל אביב

Coller School
of Management
Tel Aviv University

בית הספר למוסמכים במינהל עסקים ע"ש ליאון רקנאטי

פרויקט סופי - סמסטר ב' תשפ"ה

מערכות המלצה

תאריך הגשה: 23.7.25

במידה ואתם מבצעים הנחות כלשהן, נא לציין זאת בסעיף הרלבנטי.

נתונים

יש להשתמש באחד (או יותר) ממאגרי הנתונים הבאים (להלן אופציה נוספת). שימו לב כי עליכם לבחור נתונים שמכילים לפחות את השדות הבאים timestamp, user id, item id, rating. במידה וברצונכם להשתמש במאגר נתונים אחר- אנא בקשו אישור מפורש ממני.

הגשה

יש להגיש את המטלה בזוגות או שלשות. יש לציין במפורש את ת"ז הסטודנטים ושמותיהם בראש המטלה.

על אחד מהסטודנטים להגיש את המטלה באתר הקורס.

בתרגיל מסכם זה תתבקשו לממש ולהעריך מספר אלגוריתמים של מערכות המלצה. אלגוריתם היברידי הינו אלגוריתם אשר משקלל מספר אלגוריתמים יחד, הן באמצעות Ensemble או באמצעות שילובם כאלגוריתם המלצה אחד אופטימלי. כמו כן תדרשו לממש מדדי הערכה מגוונים- ראו סעיף 4 ומאמר בנושא business metrics (יש להצדיק את הבחירה במדדים לאור הנתונים שבחרתם).

אפשר לממש את המערכת בכל שפה ובכל סביבה, כל עוד תספקו הוראות מדויקות כיצד להפעיל (בלי צורך בהתקנות מורכבות אצל הבודק).

1. ממשו את השיטה Load המקבלת את קובץ הנתונים ושנה מסוימת. השיטה תטען את כלל הנתונים, תבצע פעולת ניקוי בסיסית (למשל בחירת המשתמשים והפריטים בעלי לפחות k אינטרקאציות) ותחזיר את הנתונים של שנה זו בלבד. שיטה זו תסייע לכם לעבוד על מקטעי נתונים קטנים יותר.

2. ממשו פונקציה בשם Split_Train_Test אשר טוענת את הנתונים ומחלקת אותן ל-5 קבצי Train ו-Test לפי השנים לפי יחס 0.25-0.75 (25% מהנתונים האחרונים עבור Test ו-75% עבור ה-Train) ולפי חלון זמן שאתם תגדירו בעצמכם. הפונקציה תחזיר את עשרת הקבצים. כל הערכות הביצועים של האלגוריתמים יתבצעו על ה-training set ואילו בדיקות האיכות יתבצעו על גבי ה-test set. (מדובר במימוש של שיטה קיימת הנקראת time series cross-validation).

3. ממשו פונקציה בשם Split_Train_Test_Users אשר טוענת את הנתונים ומחלקת אותן לקבצי Train ו-Test לכל משתמש לפי לפי יחס זמנים 0.25-0.75 (25% נתונים אחרונים של כל user עבור Test ו-75% נתונים ראשוניים עבור ה-Train). הפונקציה תחזיר את שני הקבצים. כל הערכות הביצועים של האלגוריתמים יתבצעו על ה-training set ואילו בדיקות האיכות יתבצעו על גבי ה-test set.

4. ממשו את שיטת הבדיקה RMSE המחשבת את הטעות הממוצעת: $RMSE = \sqrt{\frac{\sum r_{u,i}(r_{u,i}-\hat{r}_{u,i})^2}{N}}$. ממשו 2 שיטות ranking (כגון MRR@k, NDCG@k וכו') נוספות וכן שיטת בדיקה נוספת שהינה beyond accuracy מתוך המאמר (על פי בחירתכם) והצדיקו את בחירתכם. סה"כ יש לממש בעצמכם 4 שיטות הערכה של איכות המודל.

5. עליכם לממש את השיטה TrainHybridModel המקבלת פרמטרים מותאמים למודלים השונים ומאמנת מודל

היברידי כלשהו המשלב לפחות 2 אלגוריתמי המלצה. להזכירכם, על המודל ההיברידי להתמודד עם אתגרים שונים הקשורים למטרות העסקיות של מערכות ההמלצה ולנתונים שלכם. הצדיקו את הבחירה באלגוריתמים לאור האתגרים בנתונים שלכם והשאלות העסקיות של הפרויקט שלכם. הקדישו מחשבה ל-Loss שתרצו שהאלגוריתם ההיברידי יספק.

6. צרו ויזואליזציות של לפחות 4 גרפים המתארים את איכות אימון המודל (training error) והכללת המודל (testing error) כתלות בפרמטר כלשהו, כגון (להלן כמה אפשרויות): גודל הנתונים (כמות משתמשים או כמות פריטים - Coverage), כמות latent factors או hyperparameters של המודל ההיברידי, גודל רשימת הפריטים (k), וכו'. הסבירו בקצרה את המסקנות מהגרפים שקיבלתם.

7. ממשו מודל המלצה נוסף (עפ"י בחירתכם) אשר מרחיב את המודל שיצרתם בסעיף 5 עם נתונים נוספים, כגון נתונים שהתקבלו ע"י הפעלת מודל מאומן (pre-trained model- ראו huggingface) או נתוני הקשר (context) או כל מאפיינים נוספים שלא השתמשתם בהם בסעיף 5. עליכם לממש את השיטה TrainExtendedHybridModel. הסבירו כיצד הרחבתם את המודל מסעיף 5 ומדוע בחרתם דווקא בנתונים הללו.

8. ממשו את השיטה PredictRating לכל אחד מהמודלים שבניתם בסעיפים 5,7. בהינתן user_id ומזהה של הפריט item_id הפונקציה תשתמש במודלים שבניתם בסעיפים הקודמים על מנת לחזות את הדירוג למשתמש. עבור המודל מסעיף 7 הפונקציה תקבל גם נתונים נוספים info שיהוו קלט למודל.

9. השוו את התוצאות של שני המודלים (על קבצי ה-Test שיצרתם בסעיפים 2-3) לפי שלושת המדדים שקבעתם בסעיף 4 והסבירו בפירוט את ההבדל בין התוצאות והמסקנות שהגעתם אליהם.

* ינתן **בונוס** של עד 5 נקודות לעבודה שתשפר באופן משמעותי את המדדים של המודל הקיים מסעיף 5 ע"י המודל המשופר שהצעתם בסעיף 7.

בהצלחה!