הפקולטה לניהול
ע"ש קולר
אוניברסיטת תל אביב

Coller School
of Management
Tel Aviv University

בית הספר למוסמכים במינהל עסקים ע"ש ליאון רקנאטי

מטלה 2 - סמסטר ב' תשפ"ה

מערכות המלצה

תאריך הגשה: 23.6.25

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

<u>נתונים</u>

יש להשתמש במאגר הנתונים של Yelp המופיע באתר הקורס.

<u>הגשה</u>

יש להגיש את המטלה <u>בזוגות</u>. יש לציין במפורש את ת"ז הסטודנטים ושמותיהם בראש המטלה.
על אחד מהסטודנטים להגיש את המטלה באתר הקורס.

בתרגיל זה תתבקשו לממש מערכת המלצה מבוססת מודל. כמו כן תדרשו לממש את חישוב הטעות הממוצעת (RMSE) ומדד הערכה נוסף שתבחרו (יש להצדיק את הבחירה במדד).
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 ה-csv ושנה מסוימת. השיטה תטען את כלל הנתונים ותחזיר את הנתונים של שנה זו בלבד.

2. ממשו פונקציה בשם Split_Train_Test אשר טוענת את הנתונים ומחלקת אותן לקבצי Train ו-Test לפי השנים לפי יחס 1:3 (שנה עבור Test ושנתיים עבור ה-Train). הפונקציה תחזיר את שני הקבצים. כל הערכות הביצועים של האלגוריתמים יתבצעו על ה-training set ואילו דיווח איכות המודל יתבצע על גבי ה-test set.

3. 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 גם שיטת בדיקה נוספת על פי בחירתכם והצדיקו את בחירתכם. אין להשתמש בספריות חיצוניות במימוש זה.

4. ממשו את מודל MF שמתואר ב- http://sifter.org/~simon/journal/20061211.html. הסברים נוספים על המודל ניתן למצוא בספר הקורס (בסיסי) עמודים 151-152. עליכם לממש את השיטה TrainBaseModel המקבלת את מספר ה-latent features (גודל הווקטורים p,q) ומאמנת מודל בסיסי. להזכירכם, בהינתן מספר הפיצ'רים:
   a. חלקו את ה-training set ל-train ול-validation (ישנו חלק נוסף – test – בו נשתמש מאוחר יותר לבדיקת איכות בסעיף 8). את כל האימון נבצע בהינתן ה-train בלבד.
   b. חשבו את μ – ממוצע הדירוגים.
   c. אתחלו את b_u, b_i, p_u, q_i – באופן רנדומלי (לערכים קטנים מאוד חיוביים ושליליים).
   d. עברו על הדירוגים $r_{u,i}$ שב-training set ובצעו לכל דירוג
      i. $e_{u,i}=r_{u,i}-\mu-b_i-b_u-p_u \cdot q_i$
      ii. עדכנו את הפרמטרים:
         1) $b_u=b_u+\gamma \cdot (e_{ui}-\lambda \cdot b_u)$
         2) $b_i=b_i+\gamma \cdot (e_{ui}-\lambda \cdot b_i)$
         3) $q_i=q_i+\gamma \cdot (e_{ui} \cdot p_u - \lambda \cdot q_i)$

4) $p_u=p_u+\gamma \cdot (e_{ui} \cdot q_i - \lambda \cdot p_u)$

iii. רצוי להשתמש בפרמטרים קטנים עבור γ ו-λ (באיזור ה-0.05).

e. חשבו את הטעות (RMSE) על גבי ה-validation set. עצרו כאשר הטעות גדלה מהאיטרציה הקודמת. אם הטעות קטנה – חזרו על שלב d.

5. צרו ויזואליזציה של הוקטורים p,q בקובץ ה-Train בדו-מימד (ניתן להשתמש ב-TSNE להקטנה לדו מימד) ונסו לתאר במילים לפחות 3 תובנות מהגרף שקיבלתם. ניתן לבחור לבצע את הויזואלזציה עבור מחצית מהנתונים (אין צורך לבצע על סט הנתונים המלא).

6. ממשו מודל נוסף (עפ"י בחירתכם) מבוסס תוכן (content). עליכם לממש את השיטה TrainContentModel. הסבירו איזה מודל מבוסס תוכן מימשתם ומדוע בחרתם להשתמש בו.

7. ממשו את השיטה PredictRating לכל אחד מהמודלים שבניתם בסעיפים 4,6. בהינתן user_id ומזהה של בית העסק business_id הפונקציה תשתמש במודלים שבניתם בסעיפים הקודמים על מנת לחזות את הדירוג למשתמש. עבור המודל שמימשתם בשאלה 6 תוסיפו מידע הקשור למשתמש או ל-Business.

8. השוו את התוצאות של שני המודלים (על קובץ ה-Test) לפי המדדים שקבעתם בסעיף 3 והסבירו במילים את ההבדל בין התוצאות והמסקנות שהגעתם אליהם.

* ינתן **בונוס** של עד 5 נקודות לעבודה שתרוץ באופן המהיר ביותר ושתקבל את ה-RMSE הטוב ביותר עבור המודלים שפיתחתם בסעיפים 4,6. הצעות לשיפורים שכדאי לנסות עבור המודל בסעיף 4:

a. נסו ערכים שונים עבור λ ,γ. נסו להתחיל עם γ גדול ולהפחית אותו לאט.
b. נסו לאתחל את הערכים בדרכים שונות.
c. נסו ערכים שונים עבור גדלי הוקטורים p,q.

בהצלחה!