

## פרויקט

אופן ההגשה:	ההגשה בזוגות או בשלשות. הגשת jupyter notebook וקובץ דו"ח (word או PDF) דרך ה moodle, שמות הקבצים יהיו מספר תעודות הזרות של המגישים.
בודק תרגילים:	מר אלון זולפי zolfi@post.bgu.ac.il
תאריך הגשה:	9.5.2021

**מטרות:** תרגול מעשי של תהליך פיתוח אלגוריתמים שונים למערכות המלצה. הבנת היתרונות והחסרונות של השיטות השונות. הכרות עם ספריות Python נפוצות למערכות המלצה.

**נתונים:** הורידו את ה [movie lens](#) dataset 100000. השתמשו בקובץ u1 ל train (u1.base) ול test (u1.test). בצעו את האנליזה במחברת jupyter, יש להגיש דו"ח מפורט המכיל את הניתוחים והתשובות בקובץ word או PDF. השתמשו באותה חלוקה ל train ול test בה השתמשתם בתרגיל מספר 2 בשאר התרגילים לצורך השוואה. לצורך השוואה. האימון מתבצע על סמך קובץ ה train והבדיקה על קובץ ה test.

**הערה:** ישנו נספח בסוף המסמך המפרט את הדרישות הטכניות ומתאר את הסטנדרט הנדרש בהגשת הפרויקט.

## חלק א – ניתוח מידע

## תרגיל 1

א. עבור כל סרט חשבו את ה rating הממוצע. הציגו התפלגות (היסטוגרמה) של ה rating הממוצע בציר x, ומספר הסרטים בעלי rating זה בציר y. מיינו את הסרטים על פי ה rating והציגו את שלושת הסרטים בעלי הדירוג הממוצע הגבוה ביותר.

ב. חיזרו על סעיפים א עבור אוכלוסיית הנשים בנפרד ואוכלוסיית הגברים בנפרד. האם קיימים הבדלים בערכים ממוצעים בין שתי האוכלוסיות? מי הם שלושת הסרטים בעלי הדירוג הגבוה ביותר בקרב נשים? בקרב גברים? בעלי הפער הגבוה ביותר בין גברים לנשים? האם זיהיתם מכנה משותף בין הסרטים?

ג. בדקו את התפלגות הקטגוריות (genre) של הסרטים בעלי הדירוג הגבוה ביותר והנמוך ביותר עבור האוכלוסייה כולה, ועבור אוכלוסיות שונות (על פי נתונים דמוגרפים). ציינו כיצד בחרתם את הסרטים בעלי דירוג גבוה ודירוג נמוך. האם יש הבדל בין האוכלוסיות? מהם ההבדלים העיקריים?

ד. מי הם הסרטים הפופולאריים ביותר? כיצד תמדדו פופולאריות של סרט?

ה. חשבו את ה sparsity של ה dataset, ואת מספר ה ratings הממוצע למשתמש

1.

**חלק ב – המלצות לא אישיות****תרגיל 2**

- א. בנו מודל חיזוי דירוג rating לכל סרט על סמך ממוצע הדירוג שחישבתם בסעיף א' תרגיל 1. חשבו את ה MAE הממוצע של הדירוג החזוי לעומת הדירוג בפועל עבור קבוצת הבדיקה. שימו לב שבמודל זה, הדירוג החזוי לכל סרט יהיה זהה לכל המשתמשים.
- ב. חיזרו על סעיפים א + ב עבור חישובים שונים לגברים ולנשים. באיזה שיטה קבלתם תוצאות טובות יותר? מדוע?

**חלק ג – המלצות אישיות****תרגיל 3**

- לביצוע תרגיל זה השתמשו בספריית [Turi Create](#) שהצגנו בכיתה
- א. ממשו מודל לחיזוי rating לסרט עבור user על פי מודל ה item , matrix factorization , item content 1 similarity.
- ב. חשבו MAE ממוצע לקבוצת הבדיקה עבור כל אחת מהשיטות.
- ג. השוו את התוצאות שקבלתם על פי המודלים השונים לבין התוצאות שקבלתם בתרגיל מספר 2 (המלצות לא אישיות). איזה מהשיטות עבדה בצורה טובה יותר? מדוע? בהשוואה, התייחסו לתוצאות ה MAE ומשך האימון.

**תרגיל 4**

- לביצוע תרגיל זה השתמשו בספריית [keras](#) שהצגנו בכיתה
- ג. ממשו מודל לחיזוי rating לסרט עבור user על פי מודל neural collaborative filtering שהוצג בכיתה. התחילו ממודל עם שכבת hidden אחת.
- ד. חשבו MAE ממוצע לקבוצת הבדיקה. בידקו את התוצאות עבור אפשרויות שונות של פרמטרים של המודל: מספר השכבות, גודל כל שכבה, ה optimizer, ה loss function, ה activation function של שכבות הביניים. סה"כ בידקו 3 אפשרויות שונות לבחירתכם.
- ד. השוו את התוצאות שקבלתם על פי המודלים השונים. איזה מהשיטות עבדה בצורה טובה יותר? מדוע? בהשוואה, התייחסו לתוצאות ה MAE ומשך האימון.

## תרגיל 5

א. בנוסף לנתוני צפיה, ה dataset מכיל נתונים נוספים על סרטים וצופים כגון קטגורית הסרט, מין הצופה, גיל הצופה וכו'. הציעו שני מודלים המשלבים מאפייני סרט ו/או צופה עם נתוני צפיה. בהצעתכם, התייחסו לבחירת המאפיינים הנוספים, ארכיטקטורת המודל המוצע, ועיבוד הנתונים הנדרש על מנת להזין את הנתונים למודל. אחד המודלים צריך להיות מבוסס על ספריית [DeepCTR](#) שהוצגה בכיתה.


- ב. בחרו שני מאפיינים של סרט ו/או צופה. ממשו את המודלים שהצעתם.
- ג. חשבו ממוצע MAE לקבוצת הבדיקה עבור המודל שהצעתם. בידקו מספר אפשרויות של פרמטרי המודל, ושילוב המאפיינים.
- ד. השוו את התוצאות שקבלתם על ידי שימוש במודלים השונים.
- ה. לסיכום, דונו בתוצאות, מהו המודל המומלץ על ידכם לחיזוי rating? מדוע?

**בהצלחה!**

## נספח

1. יש להגיש דו"ח בנוסף למחברת ה-jupyter notebook שאתם מגישים, הדו"ח יכלול:
  - מבוא – סקירה קצרה של כל מה שעשיתם בפרויקט.
  - תיאור הדאטה והצגת התוצאות והתובנות משלב ניתוח המידע.
  - תיאור האלגוריתמים שבחרתם ונימוק הבחירה שלכם בכל אלגוריתם ואילו מאפיינים השתמשתם (כאשר אתם נדרשים לבחור).
  - דיון בתוצאות/מסקנות עבור כל שלב בפרויקט.
  - יש לכתוב כל הנחה לא טריוויאלית שהנחתם בפרויקט ולנמק אותה.
  - נא לא להעתיק לדו"ח קטעי קוד.
2. יש לתעד את הקוד במחברת – אתם נדרשים לתעד את הפונקציות: קלט, פלט, מטרת הפונקציה, הסבר לאלגוריתמים. לפי סטנדרט [8PEP](#) ),
3. יש להציג תוצאות של שלב ה-exploration, והחיזוי בצורה ברורה ולהשתמש בכלים ויזואליים כגון גרפים וטבלאות. גרפים וויזואליזציה – כאשר אתם נדרשים להציג תוצאות, לבחור מודלים או לתאר התפלגות של המידע, יש להציג זאת בצורה גרפית במידת האפשר. יש להציג בצורה גרפית את המידע הבא:
  - התפלגות המידע – לדוגמא באמצעות היסטוגרמה.
  - איור המתאר מודלים שבחרתם. דוגמא – איור המתאר ארכיטקטורת מודל deep learning שאתם בחרתם וגרפים המתארים את אימון מודלים אלו על מנת לנתח את ההתכנסות.
  - תוצאות חיזוי –טבלה המציגה recall precision, confusion matrix וכו'.
  - יש לשים לב שהגרפים מפורטים. שמות משמעותיים לצירים, כותרות, קנה המידה הגיוני, הערכים בצירים הגיוניים וניתנים לקריאה ולהקפיד על נראות כללית ברורה של הגרפים.
  - כאשר משווים בין תוצאות מודלים **חובה** להציג טבלהגרף השוואה שמראה באופן ברור את ההבדלים.
4. הערות כלליות:
  - יש להגיש מחברת jupyter notebook לאחר הרצה (שניתן לראות את התוצאות) – מחברת שלא הורצה לא תיבדק!
  - אנא וודאו שהקוד רץ בסביבת Google Colab (<https://colab.research.google.com/>).
  - סיפריית Turi Create עובדת רק בסביבת linux, על מנת להקל עליכם – יש להשתמש ב-google colab – מאחר והוא פועל בlinux ויאפשר לכם להשתמש בספרייה.

- יש לדאוג לכך שהמחברת תהיה מוכנה להרצה ואין שגיאות בעת הרצת המחברת, אנא וודאו זאת לפני שאתם מגישים את הפרויקט – שימו לב שנתביים לא לוקאליס למחשב שלכם וכו'. במידה ונריץ את הפרויקט שלכם אנו נריץ אותו בgoogle colab.
  - כדי להימנע מהגשת הdataset וכדי שנוכל להריץ את הקוד, אנא הוסיפו בתחילת הקוד משתנה אשר ייצג את מיקום הdataset, בתחילת המחברת לדוגמא:
- `data_path = "example/example/data_set.csv"`
- במידה והשתמשתם בספריות חיצוניות יש להוסיף שורות התקנה שלהם בcell הראשון במחברת, לדוגמא:

```
 !pip install LIBRARY_NAME==X.X.X
```

כאשר LIBRARY\_NAME מייצג את שם הספרייה וX.X.X מייצג את הגרסה שבה השתמשתם.