מסחר אלקטרוני – 096211

תרגיל בית 2 – מערכת המלצה

מגישות:

סתיו ספקטור - 315818666

סתיו בר חיים – 316377449

אופן ניבוי ההשמעות

להלן השלבים שביצענו על מנת לחקור את נתוני השמעות העבר:

המרנו את קובץ ה "user_artist" לdataframe ולאחר מכן לdataset בשם "user_artist_data", כאשר הגדרנו מהו טווח הערכים האפשריים לניבוי הweight על ידי מציאת מספר השמעות מינימלי ומספר השמעות מקסימלי בקובץ ה "user_artist".

חילקנו את "user_artist_data" לtrain ולtest לצורך אימון על מנת למזער כמה שיותר את הloss שיתקבל. הגדרנו שגודל הtrain הוא 80% וגודל הtest הוא 20%.

ראשית, הפעלנו את האלגוריתם Knn אשר בונה את מטריצת הדימיון בשיטת cosine similarity על ידי מספר מקסימלי (k) של 40 אמנים דומים, זאת לאחר בדיקה שזהו הk המקסימלי האפשרי (של אמנים דומים).

אימנו את הtrain לפי אלגוריתם זה ובינינו את מטריצת הניבויים עבור הtest (של האימון) הפועלת לפי הנוסחה:

$$\hat{r}_{ui} = \frac{\sum_{j \in N_u^k(i)} \text{sim}(i,j) \cdot r_{uj}}{\sum_{j \in N_u^k(i)} \text{sim}(i,j)}$$

שנית, הפעלנו את האלגוריתם Baseline אשר מחשב את ההטיות b_u, b_i לפי שיטת least squares.

אימנו את הtrain לפי אלגוריתם זה ובינינו את מטריצת הניבויים עבור הtest (של האימון) הפועלת לפי הנוסחה:

$$\hat{r}_{ui} = b_{ui} = \mu + b_u + b_i$$

לאחר מכן, בנינו רשימה בגודל של test (טסט האימון) בשם "predictions" המאחסנת בתוכה את סכום הניבויים משתי השיטות הנ"ל, בהתאמה לזוג סדור- משתמש ואמן.

החלטנו לסכום את הניבויים משתי השיטות על מנת להשתמש בשיטה של Neighbourhood-based predictor שנלמדה בהרצאה ונוסחתה:

$$\hat{r}_{ui}^N = \left(r_{avg} + b_u + b_i\right) + \frac{\sum_{j \in \mathcal{L}_i} d_{ij} \tilde{r}_{uj}}{\sum_{j \in \mathcal{L}_i} |d_{ij}|}$$

בשלב זה, ביצענו חישוב של הloss (קנס) לפי הנוסחה שניתנה בתרגיל הבית.

לאחר מספר ריצות עם ערכים שונים, לדוגמה: שינוי מספר איטרציות של אלגוריתם הbaseline (n_epochs), גודל אימון ומבחן שונים וכדומה, הגענו למסקנה שהשיטה הנ"ל היא הטובה ביותר עבורנו.

אופן הניבוי לפי השיטה שנבחרה על קובץ ה test (שנתון בתרגיל הבית):

הגדרנו רשימה "rui_hat" שתאחסן את הניבויים הסופיים.

רצנו על זוגות סדורים של משתמשים ואמנים מקובץ test והפעלנו פונקציית predict לפי שני האלגוריתמים שהגדרנו וסכמנו ביניהם. את רשימה זו הכנסנו כעמודה חדשה בקובץ הtest.