תרגיל בית 4

חלק א' – רטוב (60%) :

סט הנתונים שנתון לנו הוא "movie_metadata" והוא זמין להורדה בקבצים של התרגיל. סט הנתונים מתאר מדגם בסדר גודל של 5,043 סרטים עם 26 תכוניות ולכל סרט ציון לפי ה-IMDB.

למידע נוסף על הנתונים, ראו את הקישור :

https://www.kaggle.com/bobirino/movie-metadata

תיאור המשימות :

בתרגיל זה נעסוק בבעיית סיווג בינארי. נגדיר שסרט מקבל את התיוג 1 אם הוא סרט מצליח, אחרת התיוג שלו הוא 0. נגדיר שסרט הוא מצליח אם ציון ה-IMDB שלו שווה ל-7 ומעלה.

התכוניות בסט הנתונים מתחלקים לשני סוגים: נומריים וקטגוריאליים. כאשר :

Categorical Features		Numerical Features	
1.	actor_1_name	1.	actor_1_facebook_likes
2.	actor_2_name	2.	actor_2_facebook_likes
3.	actor_3_name	3.	actor_3_facebook_likes
4.	color	4.	aspect_ratio
5.	content_rating	5.	budget
6.	country	6.	cast_total_facebook_likes
7.	director_name	7.	director_facebook_likes
8.	genres	8.	duration
9.	language	9.	facenumber_in_poster
10.	movie_imdb_link	10.	gross
11.	plot_keywords	11.	movie_facebook_likes
		12.	num_critic_for_reviews
		13.	num_user_for_reviews
		14.	num_voted_users
		15.	title_year

דרך אחת להתמודד עם תכוניות קטגוריאליות היא על ידי המרה של כל תכונית קטגוריאלית לקבוצת אינדיקטורים (נקראים גם dummy variables). המשמעות היא שבסופו של דבר מספר התכוניות שעליהם תפעילו את האלגוריתם שלכם, יהיה מספר גדול מאוד <u>וכבר לא</u> 26 כמו בסט הנתונים המקורי.

נסתכל למשל על התכונית language :

$$\{Aboriginal, Arabic, Aramic, \dots, , Zulu\} \in language$$

התכונית הקטגוריאלית language מכילה x משתנים קטגוריאליים, ולכן <u>נחליף</u> עמודה זו ב-x עמודות כשכל אחת מייצגת אינדיקטור. נגדיר שבעמודה Col_{l_i} יופיע הערך 1 ברשומות של סרטים ששפתם היא l_i, אחרת יופיע הערך 0.

תשימו לב שישנן תכוניות קטגוריאליות שבהן סרט יקבל את הערך 1 **בדיוק** בעמודה אחת מבין קבוצת העמודות שהחליפו את התכונית (למשל התכונית language), וישנן תכוניות שבהן סרט יכול לקבל את הערך 1 **יותר** מעמודה אחת מבין קבוצת העמודות שהחליפו את התכונית (למשל התכונית genres).

1) עבור סט הנתונים יש לבצע 5-fold cross validation. תשימו לב כי עליכם לבצע חלוקה ל-train ו-test בלבד, בלי validation.

עליכם לבנות מעטפת גנרית ללמידה והסקה של שני אלגוריתמי סיווג שלמדתם בהרצאה :

- K-Nearest Neighbors (KNN) classifier – נבחר $K = 10$
- Rocchio classifier

לבסוף עליכם להציג עבור כל מסווג :

- Precision
- Recall
- Accuracy

כאשר נתייחס ל-positive class עבור סרטים עם תיוג 1 (סרטים מוצלחים).

(50%)

2) משימה תחרותית :

עליכם למקסם את מדד ה-accuracy עבור שני האלגוריתמים שבסעיף (1). לצורך חישוב המדד, יש לבצע **גם בסעיף זה** 5-fold cross validation על סט הנתונים.

בסעיף זה יש לכם חופש פעולה מלא, לרבות בשלב ה-data processing, בחירת התכוניות, בחירת מספר השכנים באלגוריתם KNN וכיו"ב.
עליכם לגלות יצירתיות ולהסביר במפורט את תהליך ביצוע המשימה בקובץ word או pdf.

לצורך המשימה באפשרותכם להשתמש בשלב הלמידה וההסקה בחבילה *sklearn* בלבד. בכל שלב אחר בתוכנית שלכם, כל פונקציה כשירה לשימוש לרבות פונקציות שלא הועברו בתרגולים ובמעבדות ובתנאי שסיפקתם עבורן תיעוד מתאים או התייחסתם אליהם בקובץ ה-word או ה-pdf. משימה שלא תספק הסבר מתאים תקבל ניקוד מופחת.

הסבר לגבי מתן הניקוד: תוצאות המשימה ידורגו בסדר יורד של מדד ה-accuracy שתקבלו.
הרשימה הממוינת תחולק לקבוצות כך שבכל קבוצה יש 20 הגשות, כאשר הגשה בהגדרה יכולה להיות של זוג סטודנטים או הגשה של יחיד (במידה וניתן לכך אישור). הקבוצה הראשונה תקבל את כל 10 הנקודות, הקבוצה השנייה תקבל 9 נקודות, וכך הלאה כך שההפרש בין הקבוצות הוא נקודה אחת.
לידיעתכם, פתרון למשימה 1 תקף גם כפתרון למשימה 2 (למשל במקרה שבו לא הצלחתם לקבל ביצועים טובים יותר).

(10%)

פלט התוכנית צריך להיות מהצורה הבאה :

Question 1:
KNN classifier: value of precision, value of recall, value of accuracy
Rocchio classifier: value of precision, value of recall, value of accuracy

Question 2:
KNN classifier: value of accuracy
Rocchio classifier: value of accuracy

בין התוצאות של Question 1 ו- Question 2 יש רווח של שורה אחת.

דרישות למימוש :

*** לצורך המימוש עליכם להשתמש בחבילה של *sklearn*.

המימוש חייב להכיל לפחות את המודולים, המחלקות והמתודות המתאימות שבפרק זה.

1) main.py – ממשק ראשי לריצת התוכנית כפי שנדרש בפרק "תיאור המשימות".

שורת הפקודה (שורה אחת) לצורך ריצת התוכנית במעבדת ההוראה צריכה להיות:

python3 /home/student/**your_path**/main.py
/home/student/**your_path**/movie_metadata.csv

כאשר **your_path** זהו הנתיב שבו בחרתם לשמור את קטעי הקוד של התוכנית שלכם.

2) data.py – ממשק לאיסוף הנתונים.

מחלקה Data:

function members:

2.1) **preprocess** – מתודה לביצוע הנקודות הבאות :

I. עליכם לקרוא את התכוניות בסט הנתונים לרבות התיוגים, **למעט** התכוניות הבאות:
   - content_rating
   - movie_imdb_link
   - plot_keywords

II. אין לקלוט רשומות שמכילות לפחות תכונית אחת עם ערך חסר. תשימו לב כי לא מופיעים בקובץ ערכי NaN אלא שהתאים הם ריקים ממש.
הערה: יש לבצע סעיף זה רק לאחר שבצעתם את סעיף I. הסיבה היא שלא נרצה להתעלם מרשומות שהערכים החסרים שלהן הם רק בעמודות הלא רלוונטיות.

III. במקרה שבו שם של סרט מופיע יותר מפעם אחת בעמודה movie_title, נקלוט אך ורק את הרשומה העליונה ביותר על פי סדר הרשומות שבסט הנתונים המקורי. אין לקלוט את כל שאר הרשומות שמופיעות מתחתיה עם השם של אותו הסרט.

IV. נטפל בתכוניות קטגוריאליות כפי שתואר בפרק "תיאור המשימות".

V. נבצע נרמול של סט הנתונים אך ורק לתכוניות **ללא** dummy variables.
נסתכל למשל על התכונית ה-i. יהיה μ ממוצע התצפיות ו-σ סטיית התקן. תהי תצפית שערך התכונית ה-i שלה הוא x. לכן הערך המנורמל של x הוא $\frac{x-\mu}{\sigma}$ כאשר σ היא סטיית התקן המדגמית (במכנה מופיע n-1).

2.2) **split_to_k_Folds** – תפקידה לבצע חלוקה של סט הנתונים ל-k-folds.
עליכם להשתמש בקריאה הבאה:

sklearn.model_selection.KFold(n_splits=5, shuffle=False, random_state=None)

3) algorithm_runner.py – ממשק להפעלת אלגוריתמי סיווג.

מחלקה AlgorithmRunner:

<u>data members</u>:

algorithm – אובייקט של המודול sklearn דרכו נבצע את פעולות האלגוריתם (כזכור ב-Python כל פונקציה היא אובייקט). תשימו לב כי בכל מופע של אובייקט מטיפוס המחלקה AlgorithmRunner פעיל מסווג אחד בלבד ולא שניהם.

<u>function members</u>:

fit – מעטפת לצורך קריאה לפונקציה fit של האובייקט algorithm.

predict – מעטפת לצורך קריאה לפונקציה predict של האובייקט algorithm.

מתודות המעטפת fit ו-predict צריכות להפעיל מתוכן קריאה למתודות fit ו-predict באמצעות algorithm. שתי המתודות הן תקפות לשני המסווגים ב-sklearn.

דגשים נוספים:

1) עליכם לכתוב את הקוד בהתאם לדגשים והסטנדרטים לפי PEP8. לשימושכם המסמך "Code Quality Requirements" תחת הכותרת "מעבדת הוראה והנחיות טכניות" באתר ה-moodle של הקורס.
קוד אשר לא יעמוד בסטנדרטים הנדרשים, יקבל ניקוד מופחת.
2) ניתן להוסיף מתודות נוספות, במידה ותמצאו לנכון. יש להימנע מכפילויות קוד.
3) ניתן להשתמש במתודות שהן built-in בשפה. קרי, מתודות אשר **<u>לא</u>** דורשות ייבוא של ספריות.
4) יש לתת שמות בעלי משמעות לכל משתנה.
5) חובה לתעד את הקוד באנגלית. בפרט עליכם לכתוב עבור כל מתודה docstring.

חלק ב' – יבש (40%) :

1) טענה: בהינתן 3 תצפיות בלבד, האלגוריתמים Single Link, Complete Link, Average Link יניבו תוצאה זהה.
הוכיחו או הפריכו את הטענה. (8%)

2) הוכיחו את הטענה הבאה :
נתון מסמך d ומאגר מסמכים C. נניח שכל המסמכים במאגר C מיוצגים על ידי וקטורי משקולות $tf.idf$ מנורמלים. דירוג המסמכים בסדר יורד לפי cosine similarity שקול לדירוגם בסדר עולה לפי Euclidian distance. כלומר, יחס הסדר בשתי השיטות נשמר :

$$|d_j - d_k| > |d_j - d_m| \iff cos(d_j, d_k) < cos(d_j, d_m)$$

כאשר d_j, d_k, d_m הם המסמכים ה-j, k, m בהתאמה במאגר C. (12%)

3) יהיו מסמך d ושאילתה q. נגדיר $score$ של מסמך d ביחס לשאילתה q, באופן הבא :

$$score(d; q) = \sum_{t \in q} wf_{t,d} \;\; ; \; wf_{t,d} = log(tf_{t,d}) + 1$$

כאשר t הוא מילה בשאילתה q, ו-$wf_{t,d}$ הוא וריאציה של מדד ה-tf כפי שהגדרתם בהרצאה.
נציע לדרג את המסמכים בסדר יורד לפי פונקציית score זו. ציינו 2 יתרונות ו-2 חסרונות לשימוש בפונקציה. (8%)

4) להלן שאלות נכון/לא נכון. נמקו את תשובתכם.

א. ככל שה-recall גדל אז בהכרח ה-precision גדל, ולהיפך. (4%)

ב. במידה וקיבלנו accuracy של 100% במדגם האימון, בסבירות גבוהה שה-accuracy במדגם המבחן יהיה גם כן גבוה. (4%)

ג. ערך ה-$tf.idf$ (או $tf * idf$) של מילה יהיה גבוה יותר ככל שמספר הפעמים שהמילה הופיעה במסמכים שמכילים אותה יורד, וככל שהשכיחות שלה במאגר עולה. (4%)

הוראות הגשה

- התרגיל להגשה בזוגות או ביחידים (נדרש אישור להגשה ביחידים).
- לפני ההגשה, חובה לוודא שהתוכנית עובדת במעבדת ההוראה ולא בסביבה אחרת.
- ההגשה חייבת להכיל <u>קובץ אחד</u> (קובץ zip) :

1. שם הקובץ חייב להיות hw4_xxxxxxxx_yyyyyyyy.zip כאשר xxxxxxxx ו-yyyyyyyy הם מספרי תעודות הזהות של המגישים, כולל ספרת ביקורת.
הקובץ מכיל את כל קבצי הקוד. אין להכיל בקובץ ה-zip תיקייה ובתוכה קבצי הקוד, אלא את קבצי הקוד עצמם.
<u>הערה</u> : עליכם לוודא שהתוכנית מתחילה לפעול מקובץ "main.py" בלבד.
בפרט, עליכם להגיש את התוכנית כשהיא ניתנת להרצה <u>אך ורק</u> דרך שורת הפקודה כפי שמתואר בסעיף 1 על מתודת ה-main בפרק "דרישות למימוש".
2. תשובות לחלק היבש (אם קיים) בקובץ word או pdf. יש לציין בבירור את מספר השאלה עליה ניתנת התשובה.

- ההגשה היא אלקטרונית בלבד, דרך אתר ה-moodle של הקורס. תרגילים שיוגשו בכל דרך אחרת לא ייבדקו.
- אין להגיש את אותו הקובץ פעמיים. התרגיל יוגש ע"י אחד מבני הזוג.
- שימו לב שההגשה תיחסם בדיוק בשעה 23:55 ביום ההגשה. מומלץ להגיש לפחות שעה לפני המועד האחרון.
- ניתן להגיש כמה פעמים. רק ההגשה האחרונה תישמר.
- תרגיל בית שלא יוגש לפי הוראות ההגשה – לא ייבדק (כלומר יקבל ציון 0).
- לצורך תרגיל הבית ייפתח פורום. ניהול שאלות ומתן תשובות בנושא התרגיל יתבצע דרך הפורום בלבד.

בהצלחה !