בס"ד

בינה מלאכותית - תרגיל 2

<u>הבעיה</u> – נתונים הקבצים הבאים:

- קובץ train.txt בו נתונות דוגמאות הכוללות ערכי מאפיינים שונים ואת הסיווג של כל דוגמא.
- קובץ test.txt בו נתונות דוגמאות אשר את הסיווג שלהן צריך לחזות.

חיזוי זה ייעשה באמצעות האלגוריתמים Decision Tree, KNN ו- naïve base.
יש להריץ כל אחד מאלגוריתמים אלו על קובץ train.txt.

כתוב תוכנית הקוראת מקובץ train.txt את סט הדוגמאות כאשר, **השורה** הראשונה של קובץ זה כוללת את שמות השדות (כלומר, את המאפיינים של הנתונים)
הערכים האפשריים של כל מאפיין הם הערכים שמופיעים בעמודת המאפיין בקובץ ה-train (לא יהיו ערכי מאפיינים שלא יפיעו בקובץ).
העמודה האחרונה בכל שורה הינה הסיווג (ה-class).
כל הערכים מופרדים ב <tab>.

דוגמא לקטע מקובץ ה-train.txt אותו תקבלו:

pclass age sex survived

crew adult male no

1st adult male no

2nd adult female yes

כפי שניתן לראות בקובץ זה, ישנם 3 מאפיינים (pclass, age, sex) והסיווג הוא (survived).
הערכים האפשריים היחידים של מאפיין sex בדוגמא זו הם: male, female

<u>Decision Tree</u>

כתבו פונקציה שמיישמת את אלגוריתם ID3.
את העץ שנבנה מהאלגוריתם יש להדפיס לקובץ בשם output_tree.txt בפורמט הבא:

<attribute_name>=<attribute_value>
<tab>|<attribute_name>=<attribute_value>:class

לדוגמה:

age = child

|pclass = crew: yes

|pclass = 1st: yes

|pclass = 2nd: yes

|pclass = 3rd: no

age = adult

|pclass = crew: no

|pclass = 1st: yes

|pclass = 2nd: no

|pclass = 3rd: no

שים לב שדוגמה זו אינה מכילה את כל התכונות ורק מדגימה את הרעיון הכללי של פורמט הפלט.

<u>KNN</u>

כתבו פונקציה הממשת את אלגוריתם KNN כאשר K=5. חישוב המרחק יעשה באמצעות מרחק hamming.

<u>Naïve Base</u>

כתבו פונקציה הממשת את חיזוי naïve base.

לאחר בנית המודל לכל אחד מהאלגוריתמים, יש לבצע חיזוי לדוגמאות שבקובץ test.txt.

הדפס לקובץ output.txt את החיזוי בפורמט הבא:

כותרת:

Num**<tab>**DT **<tab>**KNN>**<tab>**naiveBase

ומתחתיה החיזוי של כל אחת מהדוגמאות:

<example_number>**tab** <DT_prediction>**tab**<KNN_prediction>**tab**<naiveBase_prediction>

בשורה האחרונה בקובץ יכתב הaccuracy של כל אחד מהחיזויים (באותו פורמט).
כלומר, רמת הדיוק של כל אחד מאלגוריתמים. כמה כל אחד מהאלגוריתמים חזה נכון ביחס לסיווגים בקובץ ה TEST

יש להגיש:

- קובץ details.txt בו יש לכתוב את שם המגיש באנגלית באותיות קטנות בשורה הראשונה ובשורה השניה את מספר ת.ז.
- קובץ py_ex2 או java_ex2 אשר יכיל את הקוד. (יש לתעד את הקוד)

לצורך הדוגמא נתונים הקבצים הבאים:

train.txt – המכיל הדוגמאות המתארות מאפיינים שונים של נוסעים שהיו על הטיטאניק, והאם שרדו או לא.

test.txt - מכיל דוגמאות מתוייגות נוספות

output_tree – כמתואר לעייל. הקובץ להדגמת הפורמט בלבד.

Output.txt – הקובץ להדגמת הפורמט בלבד.

(**<u>הערה!!!</u>:** אין להשתמש HARD CODED בנתונים ספציפיים.)

בהצלחה!