תרגיל בית 3 – מערכות לומדות

- תאריך הגשה: 22.1.2020
- המתרגל האחראי על התרגיל הוא תומר. שאלות יש להפנות אליו למייל – tlange@cs.technion.ac.il.
- קראו היטב את הוראות ההגשה שבסוף המסמך.
- כל עוד לא צויין אחרת, מותר לכם להשתמש בספריות חיצוניות כמו sklearn.
- התעדכנו ברשימת ה-FAQ שבאתר הקורס:
  - שאלה שכבר מופיעה ברשימה זו לא תיענה.
  - הנחיות שיופיעו ברשימה זו מחייבות את כל הסטודנטים.
- העבודה עלולה לקחת זמן רב ולכן מומלץ להימנע מדחייתה לרגע האחרון.

מבוא

תרגיל זה עוסק בבעיית סיווג בינארית. לאורך התרגיל נתנסה בבניית סוגים שונים של מסווגים בסיסיים (Decision Trees, KNNs) וניווכח בחשיבות התאמת המודל לבעיה שאותה רוצים לפתור. קשרים מן הסוגים שנראה קיימים בשפע בבעיות אמיתיות, ולכן אנו ממליצים לכם להשקיע מחשבה בפתרון התרגיל (וכמובן לשאול שאלות במידת הצורך).

הדאטה מכיל נתונים פיזיולוגיים של אנשים, כאשר התווית של כל דוגמה קובעת האם היא מייצגת אדם חולה (Outcome=1, נתייחס לדוגמה כזו כחיובית) או בריא (Outcome=0, נתייחס לדוגמה כזו כשלילית). הדאטה חולק עבורכם לשני קבצים: train.csv, test.csv, המכילים את קבוצות האימון והמבחן בהתאמה. ככלל, קבוצת האימון תשמש אותנו לבניית המסווגים, וקבוצת המבחן תשמש להערכת הביצועים שלהם.

חלק א' – עץ החלטה סטנדרטי (20 נקודות)

1. יהא דאטה המכיל דוגמאות אשר כל אחת מורכבת משתי תכונות מספריות (x,y) ותווית בינארית ('+' או '-'). נניח כי קיים ישר y=mx+n המפריד בין הדוגמאות החיוביות לבין הדוגמאות השליליות: לכל דוגמה עם ערכי תכונות (x=a, y=b), אם הדוגמה חיובית אז ma+n>b ואם הדוגמה שלילית אז ma+n<b.
דגש: שאלה זו עוסקת במבנה של עצי החלטה כפי שנלמד בכיתה.
   א. מהו התנאי על m,n המבטיח כי עץ החלטה יוכל להפריד בין הדוגמאות החיוביות והשליליות תוך פיצול יחיד?
   ב. איזו תכונה של מרחב ההיפותזות של עץ החלטה גורמת לכך שבמקרה הכללי הוא יידרש לפצל יותר מפעם אחת כדי להפריד בין הדוגמאות החיוביות והשליליות עבור דאטה מהסוג המתואר?
   ג. כיצד ניתן לשנות את המבנה של כלל הפיצול כך שבמקרה הכללי, עבור דאטה מהסוג המתואר, עץ החלטה יוכל להפריד בין הדוגמאות החיוביות והשליליות תוך פיצול יחיד?
   ד. יהא DT עץ החלטה לא גזום מסוג ID3, אשר נבנה תוך שימוש בדאטה מהסוג המתואר. האם ייתכן שקיים מסלול ב-DT מהשורש לאחד העלים, ובו שני צמתים המפצלים לפי אותה תכונה? אם לא, הוכיחו. אם כן, תנו דוגמה למקרה כזה.

נעבור עתה למימוש עץ החלטה מסוג ID3.

2. הגישו קובץ פייתון DT1.py אשר טוען את הדאטה, בונה עץ החלטה DT_1 לא גזום בעזרת קבוצת האימון ומדפיס את תוצאות הסיווג של קבוצת המבחן.

<u>פורמט הפלט</u> (מעתה – פורמט f1): יש להדפיס מטריצת בלבול באופן הבא –

```
[[TP FP]
[FN TN]]
```

טיפ: ניתן להשתמש ב-sklearn.metrics.confusion_matrix.

ננסה לשפר את הביצועים באמצעות גיזום עץ ההחלטה DT_1. בשאלה זו נשתמש בגיזום מוקדם: בכל פעם שמתקבל צומת עם פחות מ-x דוגמאות, נעצור את פיתוח העץ ונהפוך את הצומת לעלה.

3. הציגו גרף המתאר את דיוק עצי ההחלטה הגזומים על קבוצת המבחן כתלות בגודל x עבור הערכים $\{3, 9, 27\}$. נתחו בקצרה את התוצאות שקיבלתם (מעתה נקרא לעצים אלו $DT_{x=3}, DT_{x=9}, DT_{x=27}$ בהתאמה).

אחת הסיבות לפופולריות של עצי החלטה טמונה ביכולת שלנו להבין את אופן פעולתם.

4. הציגו תרשים המתאר את מבנה עץ ההחלטה $DT_{x=27}$.

חלק ב' – דאטה לא מאוזן ושגיאה ממושקלת (40 נקודות)

במקרה הבינארי, דאטה לא מאוזן הוא דאטה שבו רוב הדוגמאות שייכות למחלקה אחת ומיעוט של הדוגמאות שייכות למחלקה השנייה. במקרה זה, יכולת ההכללה של עצי החלטה עלולה להיפגע.

5. נניח כי ברשותנו קבוצת אימון S לא מאוזנת, ובה p מהדוגמאות הן שליליות (כלומר עם תווית '-') ו-q=1-p מהדוגמאות הן חיוביות (כלומר עם תווית '+'), כאשר q<<1. בנוסף, ידוע כי ערך התווית של כל דוגמה בלתי תלוי בערכי התוויות של הדוגמאות האחרות. יהיו A עץ ID3 לא גזום ו-B עץ ID3 גזום, כאשר שניהם נבנו תוך שימוש ב-S, ו-x דוגמת מבחן כלשהי.
   א. כיצד תעריכו את ההסתברות (ביחס ל-p) ש-A יסווג את x כשלילית? הסבירו.
   ב. כיצד תעריכו את ההסתברות (ביחס ל-p) ש-B יסווג את x כשלילית? הסבירו.

אחת הדרכים להתמודד עם דאטה לא מאוזן היא להשוות בין גדלי המחלקות באמצעות "זריקה" של דוגמאות מהמחלקה הגדולה.

6. נניח שאיזנו את S באופן המתואר, ע"י זריקה <u>אקראית</u> של דוגמאות שליליות. נסמן ב-S' את קבוצת האימון החדשה שהתקבלה. יהא A' עץ ID3 לא גזום שנבנה תוך שימוש ב-S'.
   הוכיחו \ הפריכו: כל דוגמת מבחן x שסווגה ע"י A' כשלילית, תסווג גם ע"י A כשלילית.

בבעיות אמיתיות, לעתים קרובות ישנן שגיאות יותר "חמורות" משגיאות אחרות. בחלק זה נגדיר את המצב בו נבדק חולה תוייג כבריא (False Negative) כגרוע פי 4 מהמצב בו נבדק בריא תוייג כחולה (False Positive). לפיכך, נרצה לבנות עץ החלטה שיימזער את השגיאה הנתונה ע"י –

$$Error_w = 4 \times FN + FP$$

7. א. בהנחה שערכי התוויות בדאטה שלכם בלתי תלויים – מי צפוי להניב שגיאה $Error_w$ קטנה יותר על קבוצת המבחן, עץ גזום או עץ לא גזום? נמקו.
   ב. חשבו את השגיאה $Error_w$ של עץ ההחלטה הלא גזום DT_1 ושל עץ ההחלטה הגזום $DT_{x=27}$ על קבוצת המבחן. האם התוצאה מתיישבת עם תשובתכם לסעיף הקודם?

8. א. הסבירו אינטואיטיבית למה איזון הדאטה הנתון בתרגיל, לפני בניית עץ ההחלטה, עשוי להקטין את שגיאת המבחן $Error_w$.
   ב. נסמן ב-T,F את קבוצת הדוגמאות החיוביות ואת קבוצת הדוגמאות השליליות (בהתאמה) בדאטה הנתון לכם בתרגיל. הגישו קובץ פייתון BALANCED.py אשר טוען את הדאטה, מאזן את קבוצת האימון באמצעות בחירת |T| הדוגמאות השליליות <u>הראשונות</u> (וזריקת האחרות), בונה באמצעותה עץ לא גזום ומדפיס את תוצאות הסיווג של קבוצת המבחן בפורמט f1. האם התוצאה מתיישבת עם תשובתכם לסעיף הקודם?

הערה: המשך השאלה עוסק בדאטה המקורי (ולא בדאטה המאוזן).

דרך נאיבית לנסות להקטין את $Error_w$ היא לשנות את הסיווג הנקבע על ידי העץ בדיעבד. לשם כך, נציע את הפרוטוקול הבא: בכל פעם שדוגמה מקבלת סיווג שלילי (בריא), נטיל מטבע בעל הסתברות p לתוצאת "עץ" והסתברות 1-p לתוצאת "פלי". אם יצא עץ, נשנה את הסיווג להיות חיובי; אחרת, נשאיר אותו שלילי.

9. א. איזה תנאי על הערכים TP, FP, FN, TN של DT_1 צריך להתקיים ע"מ שכדאי יהיה לקבוע p=1?
   ב. נסמן ב-T את קבוצת הדוגמאות החיוביות וב-F את קבוצת הדוגמאות השליליות מתוך קבוצת האימון. מהי תוחלת שגיאת <u>האימון</u> $Error_w$ של העץ DT_1 כאשר משתמשים בפרוטוקול המוצע (כתלות בערכי $p, |T|, |F|$)?
   ג. הציגו גרף המתאר את שגיאת <u>המבחן</u> $Error_w$ של העץ DT_1 כאשר משתמשים בפרוטוקול המוצע כתלות בגודל p עבור הערכים $\{0.05, 0.1, 0.2\}$. נתחו בקצרה את התוצאות שקיבלתם תוך השוואה עם ביצועי DT_1.

עתה ננסה להקטין את $Error_w$ באמצעות שינוי אופן בניית העץ. נגדיר את הכללים הבאים:

- כלל החלטה: בהינתן עלה v בעץ ההחלטה, נסמן את קבוצות דוגמאות האימון השייכות לעלה זה ב-T (חיוביות) ו-F (שליליות). עבור דוגמה x שהגיעה לעלה v בשלב המבחן, נציע כלל החלטה מהצורה הבאה:

$$C_1(x) = \begin{cases} P & \alpha|T| > |F| \\ N & else \end{cases}$$

- כלל פיצול: בכל פיצול נמקסם את תוספת האינפורמציה המחושבת ע"י **פונקציית האנטרופיה הממושקלת** (בינארית) הבאה:

$$E_w(X) = -[\delta p(P) \times \log\big(\delta p(P)\big) + (1-\delta)p(N) \times \log\big((1-\delta)p(N)\big)]$$

10. מהם ערכי α, δ ששימוש בהם עשוי למזער את $Error_w$?

11. הגישו קובץ פייתון DT2.py אשר טוען את הדאטה, בונה עץ החלטה DT_2 גזום עם x=9 במטרה למזער את $Error_w$ על קבוצת המבחן ומדפיס את התוצאות בפורמט f1.
    - הערה: אינכם צריכים לממש את כללי ההחלטה והפיצול הממושקלים באופן מפורש. השתמשו בפיצ'רים המובנים של sklearn.

12. איזו מניפולציה ניתן לבצע על הדאטה, כך שפעולת עץ ID3 סטנדרטי גזום עם x=9 על הדאטה החדש תהיה זהה לפעולת העץ DT_2 (שנבנה תוך שימוש בדאטה הישן ובכללים שהוצעו)? הסבירו.

חלק ג' – KNN ובחירת תכונות (40 נקודות)

13. יהיו S קבוצת האימון שהוגדרה בשאלה 5, C מסווג KNN שנבנה תוך שימוש ב-S, ו-x דוגמת מבחן כלשהי. כיצד תעריכו את ההסתברות (ביחס ל-p) ש-C יסווג את x כשלילית כתלות בגודל k? הסבירו.

כפי שנלמד בשיעור, KNN רגיש לתחומי הערכים של התכונות. בצעו נרמול באמצעות ההפרש בין המקסימום למינימום באופן שהוצג בשיעור. מעתה נשתמש בדאטה המנורמל להמשך התרגיל.

עתה נממש מסווג KNN (ללא שימוש בספריות כגון sklearn).

14. הגישו קובץ פייתון KNN1.py אשר טוען את הדאטה, בונה מסווג KNN סטנדרטי עם k=9, ומדפיס את תוצאות הסיווג על קבוצת המבחן בפורמט f1. מהי השגיאה $Error_w$ של מסווג זה?

השאלה הבאה (בלבד) עוסקת בבחירת תכונות. נסמן את אוסף התכונות בדאטה ב-A.

15. א. הגישו קובץ פייתון OPT.py אשר טוען את הדאטה ומדפיס תת קבוצה אופטימלית של התכונות $T \subseteq A$ אשר שימוש בה ממקסם את הדיוק של מסווג KNN עם k=9 על קבוצת המבחן.

<u>פורמט הפלט</u> (מעתה – פורמט f2): יש להדפיס רשימה עם האינדקסים של התכונות הנבחרות, (האינדקס של התכונה הראשונה הוא אפס) –

[ind1, ind2, …, indk]

ב. הגישו קובץ פייתון SFS.py אשר טוען את הדאטה ומדפיס בפורמט f2 תת קבוצה "טובה" של תכונות $T' \subseteq A$, אשר שימוש בה מגדיל את הדיוק (ביחס לשימוש בכל התכונות) של מסווג KNN עם k=9 על קבוצת המבחן, תוך שימוש ב*בחירה מקומית לפנים* (כפי שנלמד בשיעור).

לסיום, ננסה לבנות מסווג KNN שיימזער את $Error_w$: בהינתן דוגמת מבחן x, נמצא את k השכנים הקרובים אליה ביותר, ונשתמש בכלל החלטה המעניק לשכן חיובי משקל גבוה פי ארבעה מהמשקל המוענק לשכן שלילי (גם בחלק זה – אין להשתמש בספריות כגון sklearn).

16. הגישו קובץ פייתון KNN2.py אשר טוען את הדאטה, בונה מסווג KNN עם k=9 המבוסס על כלל ההחלטה שתואר לעיל, ומדפיס את תוצאות הסיווג על קבוצת המבחן בפורמט f1.

17. הציגו גרף המתאר את $Error_w$ של מסווג המשתמש בכלל ההחלטה שתואר לעיל כתלות בגודל k, עבור הערכים $\{1, 3, 9, 27\}$. נתחו את התוצאות תוך התייחסות לחוסר האיזון של הדאטה.

הוראות הגשה

- הגשת התרגיל תתבצע אלקטרונית בזוגות בלבד.
- מותר לממש פונקציות עזר, להוסיף קבצי קוד משלכם, ולהשתמש בספריות חיצוניות כמו sklearn, אלא אם צויין אחרת.
- אין להגיש את קבצי הנתונים – הניחו כי הם זמינים בתיקייה הנוכחית (current folder).
- הקפידו על הפניות רלטיביות לקבצים\תיקיות (relative path).
- הקוד שלכם ייבדק (גם) באופן אוטומטי ולכן יש להקפיד על הפורמט המבוקש.
- המצאת נתונים לצורך בניית הגרפים אסורה ומהווה עבירת משמעת.
- הקפידו על קוד קריא ומתועד.

יש להגיש קובץ זיפ יחיד בשם AI3_<id1>_<id2>.zip (ללא סוגריים משולשים), שמכיל:

✓ קובץ בשם readme.txt שמכיל את פרטי המגישים בפורמט הבא:

```
Name1 ID1 Email1
Name2 ID2 Email2
```

✓ קובץ בשם AI_HW3.PDF המכיל את תשובותיכם לשאלות היבשות.

✓ קובץ requirements.txt שמכיל כל חבילה חיצונית בה השתמשתם ואינה מותקנת ב-Anaconda Python. יש לוודא שניתן להתקין את החבילות באמצעות הפקודה – `$ pip install -r requirements.txt`.

✓ כל קבצי הקוד שנדרשתם לממש בתרגיל:
בחלק של עצי החלטה – DT1.py, BALANCED.py, DT2.py
בחלק של שכנים קרובים ביותר – OPT.py, SFS.py, KNN1.py, KNN2.py

✓ כל קוד עזר שמימשתם בתרגיל.