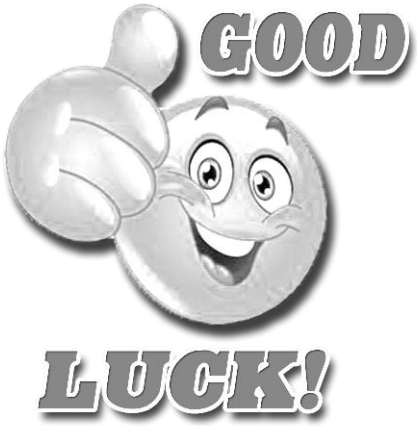
תרגיל בית 3 – מערכות לומדות

- תאריך הגשה: 1.7.2020
- המתרגל האחראי על התרגיל הוא תומר. שאלות יש להפנות אליו למייל – tlange@cs.technion.ac.il.
- קראו היטב את הוראות ההגשה שבסוף המסמך.
- לאורך כל התרגיל, עליכם לממש את המסווגים בעצמכם. מותר להשתמש בספריות החיצוניות הבאות לכל צורך שהוא מלבד מימוש המסווגים עצמם: numpy, pandas, sklearn, matplotlib.
- 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 tree, KNN) והרחבות שלהם. השאלות התיאורטיות שבתרגיל דומות לאלו שתראו במבחן, ולכן אנו ממליצים לכם להשקיע בהן מחשבה רבה.

הדאטה מכיל מדדים שנאספו מצילומים שנועדו להבחין בין גידול שפיר לגידול ממאיר. כל דוגמה מכילה 30 מדדים כאלה, ותווית בינארית diagnosis הקובעת את סוג הגידול (0=שפיר, 1=ממאיר). כל התכונות (מדדים) רציפות.

הדאטה חולק עבורכם לשתי קבוצות: קבוצת אימון (train.csv) וקבוצת מבחן (test.csv). בכלל, קבוצת האימון תשמש אותנו לבניית המסווגים, וקבוצת המבחן תשמש להערכת ביצועיהם.

בהצלחה!

נתחיל בשאלה תיאורטית בנושא עצי החלטה.

1. יהא D מאגר אימון עבורו קיימת פונקציה f כך שקיימות שתי תכונות **נומינאליות** a,b המקיימות –

$$\forall x \in D:\ x.a = \mathrm{f}(x.b)$$

נסמן ב-D_a את המאגר המתקבל מ-D ע"י הסרת התכונה a, וב-D_b את המאגר המתקבל מ-D ע"י הסרת התכונה b. יהיו T עץ ID3 לא גזום הנבנה תוך שימוש ב-D, T_a עץ ID3 לא גזום הנבנה תוך שימוש ב-D_a ו-T_b עץ ID3 לא גזום הנבנה תוך שימוש ב-D_b. הוכיחו \ הפריכו:

א. T ו-T_a בהכרח יסווגו באופן זהה כל דוגמת מבחן.
ב. T ו-T_b בהכרח יסווגו באופן זהה כל דוגמת מבחן.

עתה נעבור למימוש עץ החלטה מסוג ID3. תזכורת: יש לממש את המסווג בעצמכם.

2. הגישו קובץ פייתון DT.py אשר טוען את הדאטה, בונה עץ החלטה ID3 לא גזום בעזרת קבוצת האימון ומדפיס את דיוק (accuracy) הסיווג של קבוצת המבחן.

הערות:

א. במקרה של שוויון בתוספת האינפורמציה, פצלו לפי התכונה עם האינדקס הנמוך יותר.
ב. כאשר מפצלים לפי תכונה i כלשהי לקבוצה S1 שבה לכל דוגמה ערך i גבוה מערך הסף וקבוצה S2 שבה לכל דוגמה ערך i נמוך מערך הסף, ערך הסף ייקבע להיות **הממוצע** של ערך התכונה i המינימלי בקבוצה S1 וערך התכונה i המקסימלי בקבוצה S2.

לסיכום חלק זה, נבחן את ההשפעה של גיזום מוקדם על ביצועי העץ. בכל פעם שמתקבל צומת עם x דוגמאות או פחות, נעצור את פיתוחו ונהפוך אותו לעלה. בשלב המבחן, הסיווג ייקבע לפי החלטת הרוב בעלה. במקרה של שוויון, הסיווג ייקבע להיות True.

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.

בהינתן עץ החלטה T, וקטור $\varepsilon > 0$ ודוגמת מבחן x, כלל אפסילון-החלטה שונה מכלל ההחלטה הרגיל שנלמד בשיעור באופן הבא: נניח שמגיעים לצומת בעץ המפצל לפי ערכי התכונה i, עם ערך הסף v_i. אם מתקיים $|x_i - v_i| \leq \varepsilon_i$ אזי ממשיכים **בשני** המסלולים היוצאים מצומת זה, ואחרת ממשיכים לבן המתאים בדומה לכלל ההחלטה הרגיל. לבסוף, מסווגים את הדוגמה x בהתאם לסיווג הנפוץ ביותר של הדוגמאות הנמצאות **בכל העלים** אליהם הגענו במהלך הסיור על העץ (במקרה של שוויון – הסיווג ייקבע להיות True). נתחיל בשתי שאלות תיאורטיות.

4. יהא T עץ החלטה לא גזום, ויהא T' העץ המתקבל מ-T באמצעות גיזום מאוחר שבו הוסרה הרמה התחתונה של T (כלומר כל הדוגמאות השייכות לזוג עלים אחים הועברו לצומת האב שלהם).
הוכיחו \ הפריכו: **בהכרח** קיים ε כך שהעץ T עם כלל אפסילון-החלטה והעץ T' עם כלל ההחלטה הרגיל יסווגו כל דוגמת מבחן בצורה זהה.

5. יהא מאגר נתונים D, ויהא T_l עץ ID3 שנבנה תוך שימוש ב-D ונגזם ע"י הפסקת פיצול צמתים החל מעומק l. סטודנט בקורס שם לב שכאשר משתמשים בכלל אפסילון-החלטה, ריבוי המסלולים האפשריים עלול להאריך מאוד את שלב הסיווג (inference). הוא הציע לבנות באמצעות D עץ ID3 אחר, נסמנו T'_l, אשר נגזם באותו אופן, ונבנה כך: בכל פעם שמגיעים לצומת בעץ המפצל לפי ערכי התכונה i עם ערך הסף v_i, כל דוגמת אימון בצומת המקיימת $|x_i - v_i| \leq \varepsilon_i$ תועבר **לשני** הבנים שלו. בשלב המבחן, הסיווג ייקבע על פי כלל ההחלטה הרגיל (ולא כלל אפסילון-החלטה).
הוכיחו \ הפריכו: העץ T_l עם כלל אפסילון-החלטה והעץ T'_l עם כלל ההחלטה הרגיל בהכרח יסווגו כל דוגמת מבחן באופן זהה.
הבהרה: שני העצים משתמשים באותו וקטור ε.

עתה נממש עץ המשתמש בכלל אפסילון-החלטה. נסמן ב-v את הוקטור המכיל את סטיות התקן של ערכי התכונות בקבוצת האימון, כלומר האלמנט v_i הוא סטיית התקן של ערכי התכונה ה-i. נגדיר את אפסילון באופן הבא: $\varepsilon_i = 0.1 \times v_i$. נסמן ב-T_9 את עץ ההחלטה הגזום עם x=9.

6. הגישו קובץ פייתון DT_epsilon.py אשר טוען את הדאטה, בונה את T_9 בעזרת קבוצת האימון ומדפיס את דיוק הסיווג של קבוצת המבחן תוך שימוש בכלל האפסילון-החלטה שתואר לעיל.
טיפ: הקריאה numpy.std(a) מחזירה את סטיית התקן של האלמנטים במערך a.

עתה נעבור לדון במסווג KNN. נתחיל בשאלה תיאורטית.

7. בהינתן מאגר אימון DS, נסמן ב-$KNN_{DS}(x)$ את תוצאת הסיווג של דוגמת מבחן x ע"י מסווג KNN המשתמש במאגר האימון DS.

יהיו שני מאגרים DS_1, DS_2 **ללא חזרות** (בכל מאגר בנפרד אין דוגמה שמופיעה יותר מפעם אחת). נגדיר:

$$DS_1 \cup DS_2 = \{x | x \in DS_1 \ or \ x \in DS_s\}$$
$$DS_1 \cap DS_2 = \{x | x \in DS_1 \ and \ x \in DS_s\}$$

ונגדיר את המאגר $DS_1 + DS_2$ על ידי –

$$count_x(DS_1 + DS_2) = count_x(DS_1) + count_x(DS_2)$$

כאשר $count_x(DS)$ שווה למספר המופעים של אלמנט x במאגר DS.

הוכיחו \ הפריכו: אם $3NN_{DS_1 \cup DS_2}(x) = 3NN_{DS_1 \cap DS_2}(x) = TRUE$, אזי $3NN_{DS_1 + DS_2}(x) = TRUE$.

כזכור, מסווג KNN רגיש לתחומי ערכי התכונות. בתרגיל זה ננרמל את הדאטה לפי ההפרש בין המקסימום למינימום, כפי שנלמד בהרצאה. זכרו שאת ערכי הקיצון יש לחשב באמצעות קבוצת האימון בלבד.

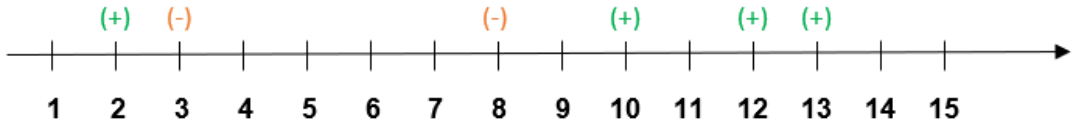
נעבור למימוש מסווג KNN ובחינת ההשפעה של הפרמטר K. תזכורת: יש לממש את המסווג בעצמכם.

8. הגישו קובץ פייתון KNN.py אשר טוען את הדאטה, מנרמל אותו, בונה מסווג KNN סטנדרטי עם k=9, ומדפיס את דיוק הסיווג על קבוצת המבחן.

9. הציגו גרף המתאר את דיוק מסווג ה-KNN על קבוצת המבחן כתלות בגודל k עבור הערכים {1,3,9,27}. נתחו בקצרה את התוצאות שקיבלתם.

בחלק האחרון נשלב בין עץ עם כלל אפסילון-החלטה לבין מסווג KNN. כרגיל, נתחיל בשאלה תיאורטית.

10. נתונה קבוצת האימון הבאה, המורכבת מ-6 דוגמאות בעלות תכונה רציפה אחת –



נניח כי אנו בונים שני מסווגים: עץ ID3 לא גזום עם כלל אפסילון-החלטה, ומסווג KNN עם k=3. בשאלה זו נניח כי כאשר עץ ההחלטה קובע ערך לפיצול, הוא בוחר את ממוצע הערכים של שתי הדוגמאות הרלוונטיות. כמו כן, בשלב ההחלטה – אם יש שוויון, עץ ההחלטה בוחר סיווג בצורה רנדומלית. האם קיים ערך של ε **המבטיח** ששני המסווגים יסווגו את כל דוגמאות האימון בצורה זהה? אם כן, מצאו ערך כזה והראו זאת. אם לא, הסבירו למה לא.

לסיום, נעבור למימוש המסווג המשולב.

11. הגישו קובץ פייתון KNN_epsilon.py אשר מבצע את הפעולות הבאות:

א. טוען את הדאטה, מנרמל אותו, ובונה את T_9 בעזרת קבוצת האימון.
ב. לכל דוגמת מבחן, מוצא את העלים הרלוונטיים לכלל האפסילון החלטה שתואר לעיל.
ג. קובע את הסיווג של כל דוגמה תוך שימוש ב-KNN עם k=9 על כל הדוגמאות הנמצאות בעלים אלו.
ד. מדפיס את דיוק הסיווג של קבוצת המבחן כולה.

הערות:

- במקרה שבקבוצת העלים הרלוונטיים יש פחות מ-9 דוגמאות, הסיווג ייקבע לפי החלטת הרוב.
- במקרה של שוויון בהחלטת הרוב, הסיווג ייקבע להיות True.

הוראות הגשה

- הגשת התרגיל תתבצע אלקטרונית בזוגות בלבד.

  מותר לממש פונקציות עזר, להוסיף קבצי קוד משלכם, ולהשתמש בספריות numpy, pandas, sklearn, matplotlib לכל צורך שהוא, מלבד מימוש המסווגים עצמם.
- 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.
- ✓ כל קבצי הקוד שנדרשתם לממש בתרגיל:
  DT.py, DT_epsilon.py, KNN.py, KNN_epsilon.py
- ✓ כל קוד עזר שמימשתם בתרגיל.