# תרגיל בית 3 – מבוא ללמידה

## הנחיות כלליות:

- תאריך ההגשה: 26/01/2021 ב23:59
- התרגיל הוא להגשה **ביחידים.**
- המתרגל האחראי על התרגיל: רפאל גד, נא לפנות בשאלות אך ורק למייל ai.technion@gmail.com .
- הקוד שלכם ייבדק אוטומטית וגם ידנית, רמאות כלשהי תוביל לועדת משמעת. התרגיל מהווה חלק משמעותי מהציון הסופי שלכם בקורס, כל תקשורת בין סטודנטים בכל פלטפורמה שהיא בנוגע לתרגיל אסורה ומהווה רמאות. קבוצת הקורס תפעיל אמצעי בינה מלאכותית מתקדמים לעלות על הרמאים.
- אנא הקפידו על ההנחיות, כל הפרה תגרום להורדת ניקוד לרבות אי עמידה בציפיות פורמט הפלט של כל קובץ הנדרש מכם.
- בכל יום חמישי בשעה 14:30 רפאל יקיים כיתה הפוכה בה תוכלו לשאול שאלות לגבי התרגיל.
- בחלק מהסעיפים קיימת הגבלת שורות לפתרון שלכם. אם אתם חושבים שבסעיף מסוים הגבלה זו לא ריאלית אנא שילחו לרפאל נימוק על כך במייל ואם הנימוק יתקבל נשנה את ההגבלה לכלל הסטודנטים.
- הקוד שלכם צריך להתייחס לקבציי הדאטה כנמצאים בתיקייה הנוכחית ולא בתת תיקייה. אין לשנות את שמות קבצי הדאטה.
- בסעיפים הרטובים בתרגיל מסופקות לכן מספר מועט יחסית של הנחיות ביחס לתרגילים הקודמים. כל מימוש העונה על הדרישות יתקבל. יש לכם יד חופשית בתכנון הקוד שלכם. אנו ממליצים לממש כל מסווג כ- class שמממש פונקציות fit(X, Y) שמבצעת אימון, ו- predict(X) שמבצעת את הפרדיקציה.
- בחלק מהשאלות תתבקשו להמציא שיפור לאלגוריתם למידה, כתוצאה מכך יש בהן דרגות חופש, הציון בשאלות אלו יינתן בהתאם לטיב הפתרון, הן מבחינת יצירתיות והן מבחינת תוצאות אמפיריות. יש לתאר בצורה ברורה, פורמלית ומדויקת את הפתרון.
- מותר להשתמש בספריות ,sklearn, pandas ,numpy, random, matplotlib, argparse,abc,typing, all the built in packages in python אך כמובן שאין להשתמש באלגוריתמי הלמידה, או בכל אלגוריתם או מבנה נתונים אחר המהווה חלק מאלגוריתם למידה אותו תתבקשו לממש.

## רקע:

*לתרגיל מצורף קובץ נתונים על מחלה מסוימת, כאשר כל שורה מתארת אדם. העמודה הראשונה מציינת האם האדם חולה (M) או בריא (B). שאר העמודות מציינות כל מיני תכונות רפואיות של אותו אדם (התכונות קצת מורכבות ואינכם צריכים להתייחס למשמעות שלהן כלל). בשאלות הבאות תתבקשו לממש אלגוריתמי למידה שונים מנת לאפשר חיזוי מדויק ככל האפשר של היות אדם חולה במחלה.*

## ***בהצלחה!***

1. (15 נק') **אלגוריתם ID3:**

1.1. ממשו את אלגוריתם ID3 כפי שנלמד בהרצאה.
שימו לב שכל התכונות רציפות. אתם מתבקשים להשתמש בשיטה של חלוקה דינמית המתוארת בהרצאה. כאשר בוחנים ערך סף לפיצול של תכונה רציפה, דוגמאות עם ערך השווה לערך הסף משתייכות לקבוצה עם הערכים הגדולים מערך הסף. במקרה שיש כמה תכונות אופטימליות בצומת מסוים בחרו את התכונה בעלת האינדקס המקסימלי.
המימוש צריך להופיע בקובץ בשם ID3.py אשר בהרצתו מאמן עץ על קבוצת האימון בעזרת המימוש שלכם ומדפיס את הדיוק שלו על קבוצת המבחן <u>ללא כל מלל או הדפסות אחרות (ככה יהיו גם כל שאר ההדפסות שלכם בתרגיל)</u>.

1.2. אמנו את האלגוריתם על קבוצת האימון ובדקו אותו על קבוצת המבחן. צרפו צילום מסך של תוצאת הדיוק בדו"ח.

2. (4 נק') הוכח/הפרך: בהינתן דאטה כלשהו עם תכונות רציפות ותיוגים בינאריים המחולק לקבוצת אימון ומבחן, הפעלה של פונקציית נירמול MinMax הנלמד בתרגול על הדאטה אינה משפיעה על דיוק של מסווג ID3 הנלמד על קבוצת האימון והנבחן על קבוצת המבחן. [אורך התשובה מוגבל ל20שורות]

3. (12 נק') **גיזום מוקדם.**

3.1. הסבירו מה החשיבות של הגיזום באופן כללי ואיזה תופעה הוא מנסה למנוע? [אורך התשובה מוגבל ל3 שורות]

3.2. ממשו את הגיזום המוקדם כפי שהוגדר בהרצאה. הפרמטר M מציין את מספר המינימלי בעלה לקבלת החלטה. על המימוש של הגיזום המוקדם להימצא גם כן בקובץ ID3.py.

3.3. בצעו כיוונון לפרמטר M על קבוצת האימון:

i. בחרו לפחות חמישה ערכים שונים לפרמטר M.

ii. עבור כל ערך, חשבו את הדיוק של האלגוריתם על ידי K-fold cross validation על קבוצת האימון בלבד. כדי לבצע את חלוקת קבוצת האימון ל- K קבוצות יש להשתמש בפונקציה sklearn.model_selection.KFold עם הפרמטרים n_split=5, shuffle=True ו- random_state שווה למספר תעודת הזהות שלכם. (כל כיוונון פרמטרים בתרגיל יעשה בצורה דומה).

iii. השתמשו בתוצאות שקיבלתם כדי ליצור גרף המציג את השפעת הפרמטר M על הדיוק. צרפו את הגרף בדו"ח.

iv. הסבירו את הגרף שקיבלתם. לאיזה גיזום קיבלתם התוצאה הטובה ביותר ומהי תוצאה זו?

על מימוש כיוונון הפרמטר להימצא בפונקציה בשם experiment בקובץ ID3.py. הוסיפו הערה בתחילת פונקציה זו לגבי איך להריץ את הפונקציה שלכם לקבלת הגרף. שימו לב שבריצה של הקובץ ID3.py עדיין אך ורק הדיוק של ID3 ללא גיזום צריך להיות מודפס.

3.4. השתמשו באלגוריתם ID3 עם הגיזום המוקדם כדי ללמוד מסווג מתוך **כל** קבוצת האימון ולבצע חיזוי על קבוצת המבחן. השתמשו בערך ה- M האופטימלי שמצאתם בסעיף 3.3. ציינו בדו"ח את הדיוק שקיבלתם? האם הגיזום שיפר את הביצועים ביחס להרצה ללא גיזום בשאלה 1?

4. (20 נק') **למידה מוכוונת מחיר.**

במקרה שלנו, בשל אופי הבעיה, סיווג אדם חולה כבריא חמורה פי 10 מסיווג של אדם בריא כחולה. לפיכך, הדיוק שהצגתם בשאלות 1 ו- 3 אינו משקף היטב את טיב המסווג שלכם. בעזרת משקול סוגי השגיאות בהתאם למה שלמדתם בתרגול, נגדיר פונקציה חדשה שתייצג את טיב ביצועי המסווגים:

$$loss(C) := \frac{0.1FP + FN}{n}$$

כשאר FP הינו מספר הסיווגים השגויים של בריאים כחולים על ידי המסווג C, FN הינו מספר הסיווגים השגויים של חולים כבריאים על ידי המסווג C ו- $n$ הוא מספר הדוגמאות בקבוצת המבחן.

בשאלה זו תתאימו את אלגוריתם ID3 ללמידה מכוונת מחיר.(אם כבר פתרתם עם KNN התשובות גם כן מתקבלות)

4.1. מדדו את ערך ה- loss של ID3: חזרו על סעיף 3.4 ובמקום הדיוק תמדדו את ערך ה- loss.

4.2. תארו דרך לגרום לID3 ללמוד מסווג אשר ממזער את פונקציית ה- loss שהוצגה כאן בצורה טובה יותר מאשר האלגוריתם הרגיל.

4.3. ממשו הצעתכם בקובץ CostSensitiveID3.py. על קובץ זה להדפיס בהרצתו את ה- loss שקיבלתם מהרצת האלגוריתם המשופר כאשר הוא לומד על קבוצת האימון ונבחן על קבוצת המבחן. אם ביצעתם ניסויים לקביעת פרמטרים לאלגוריתם שלכם, פרטו זאת. כמו כן צרפו בדו"ח את ה-loss שקיבלתם. שימו לב, אינכם צריכים לדאוג מכך שהשיפורים שלכם יפגעו בדיוק ובלבד שהם ישפרו את הloss.

5. (9 נק') נגדיר דטה סט $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ שבו n דוגמאות מתויגות עם סיווג בינארי $y_i \in \{0,1\}$. כל דוגמה היא וקטור תכונות המורכב משתי תכונות רציפות $x_i = (v_{1,i}, v_{2,i})$.
הניחו כי קיים מסווג מטרה $f(x): R^2 \to \{0,1\}$ שאותו אנו מעוניינים ללמוד (הוא אינו ידוע לנו) וכן שהדוגמאות ב-$D$ עקביות עם מסווג המטרה (כלומר שאין דוגמאות רועשות ב- $D$).
בסעיפים הבאים, עבור KNN, הניחו פונק' מרחק אוקלידי.
כמו כן, הניחו שאם קיימות נקודות במרחב כך שעבורן יש מספר דוגמאות במרחק זהה, קודם מתחשבים בדוגמאות עם ערך $v_1$ מקסימלי ובמקרה של שוויון בערך של $v_1$, מתחשבים קודם בדוגמאות עם ערך $v_2$ מקסימלי. הניחו כי אין דוגמאות זהות לחלוטין (כלומר גם עם ערך $v_1$ זהה וגם עם ערך $v_2$ זהה).
מסווג ID3 כאן צריך לתאום לכל תנאי קצה שהצגנו למימוש שלכם.
בכל סעיף, **הציגו מקרה המקיים את התנאים המוצגים בסעיף, הסבירו במילים, וצרפו תיאור גרפי (ציור) המתאר את המקרה (הכולל לפחות תיאור מסווג המטרה והדוגמאות שבחרתם)** . סמנו דוגמאות חיוביות בסימן '+' (פלוס) ודוגמאות שליליות בסימן '-' (מינוס). בכל אחת מתתי הסעיפים הבאים אסור להציג מסווג מטרה טריוויילי, דהיינו שמסווג כל הדוגמאות כחיוביים או כל הדוגמאות כשליליים. [2 שורות לכל סעיף, אין הגבלה על הגרפים, יש להימנע ממלל ופתרון שאינו מוגדר היטב כמתבקש לא מקבל ניקוד]

סעיף (א)
הציגו מסווג מטרה $f(x): R^2 \to \{0,1\}$ וקבוצת אימון בעלת לכל היותר 10 דוגמאות כך שלמידת עץ ID3 תניב מסווג אשר עונה נכון עבור כל דוגמת מבחן אפשרית (כלומר יתקבל מסווג המטרה), אך למידת KNN תניב מסווג שעבורו קיימת לפחות דוגמת מבחן אחת עליה הוא יטעה, לכל ערך K שייבחר. (2 נק')

סעיף (ב)
הציגו מסווג מטרה $f(x): R^2 \to \{0,1\}$ וקבוצת אימון בעלת לכל היותר 10 דוגמאות כך שלמידת מסווג KNN עבור ערך K מסוים תניב מסווג אשר עונה נכון עבור כל דוגמת מבחן אפשרית (כלומר יתקבל מסווג המטרה), אך למידת עץ ID3 תניב מסווג אשר עבורו קיימת לפחות דוגמת מבחן אפשרית אחת עליה הוא יטעה. (2 נק')

סעיף (ג)
הציגו מסווג מטרה $f(x): R^2 \to \{0,1\}$ וקבוצת אימון בעלת לכל היותר 10 דוגמאות כך שלמידת מסווג KNN עבור ערך K מסוים תניב מסווג אשר עבורו קיימת לפחות דוגמת מבחן אפשרית אחת עליה הוא יטעה, וגם למידת עץ ID3 תניב מסווג אשר עבורו קיימת לפחות דוגמת מבחן אחת אפשרית עליה הוא יטעה. (2 נק')

סעיף (ד)
הציגו מסווג מטרה $f(x): R^2 \to \{0,1\}$ וקבוצת אימון בעלת לכל היותר 10 דוגמאות כך שלמידת מסווג KNN עבור ערך K מסוים תניב מסווג אשר עונה נכון עבור כל דוגמת מבחן אפשרית (כלומר יתקבל מסווג המטרה), וגם למידת עץ ID3 תניב מסווג עונה נכון עבור כל דוגמת מבחן אפשרית (כלומר יתקבל מסווג המטרה). (2 נק')

6. (20 נק') **יער.**
ננסה עכשיו לשפר את הביצועים שלנו על ידי שימוש בועדות כפי שלמדתם בהרצאה. נציע כאן סוג חדש של יער שתממשו בעצמכם בשם KNN-decision-tree.
האלגוריתם תחילה ילמד N (פרמטר) עצי החלטה, כל אחד עם תת קבוצה שונה של דוגמאות מקבוצת

האימון. בעת סיווג דוגמא חדשה האלגוריתם יבחר K (פרמטר) עצים מתוך הN שנלמדו, יסווג את הדוגמא לפי כל אחד מK העצים ויחזיר את הסיווג הנפוץ ביותר.
<u>בחירת דוגמאות לכל עץ מתוך N העצים שנלמד:</u> גודל קבוצת האימון יסומן ב-n. עבור כל עץ מ-N העצים הגרילו (באופן רנדומלי) p*n דוגמאות מקבוצת האימון, כאשר p הוא פרמטר לאלגוריתם וערכו בין 0.3 ל0.7.
<u>בחירת K עצים בהם נשתמש לסיווג דוגמא:</u> עבור כל עץ, נחשב centroid: וקטור ממוצע של כל הדוגמאות שבהם השתמשתם לבניית העץ (למשל, אם p=0.5,n=200 השתמשנו ב100 דוגמאות לבניית עץ מסוים. כל "פיצ'ר" בcentroid של עץ זה יהיה ממוצע 100 הפיצ'רים המתאימים ב100 הדוגמאות בהם השתמשנו).
לאחר שחישבנו centroid לכל עץ (תוכלו לעשות זאת כבר בזמן האימון) כאשר נרצה לסווג דוגמא חדשה נבחר את K העצים שהcentroid שלהם קרוב ביותר לדוגמא שאנו רוצים לסווג. למדידת מרחק נשתמש במרחק אוקלידי.

6.1. עליכם לממש בקובץ KNNForest.py את אלגוריתם הלמידה KNN-decision-tree. בהרצת הקובץ האלגוריתם צריך להתאמן על כל קבוצת האימון, ולהדפיס את הדיוק שלו על קבוצת המבחן ללא כל מלל או הדפסות. עליכם לבחור בעצמכם את כל הפרמטרים השונים לאלגוריתם. מומלץ לבצע ניסויים כדי למצוא ערכי פרמטרים טובים. הסבירו אילו ניסויים ביצעתם, וציינו מהו הדיוק המקסימלי שקיבלתם על קבוצת המבחן. שימו לב שמטרת האלגוריתם הינה להשיג דיוק (רגיל) מקסימלי ולא למזער את פונקציית ה- loss.

7. (20 נק') **שיפור ליער.**

7.1. תארו שיפור לאלגוריתם KNN-decision-tree. השיפור צריך להתייחס למקרה הכללי, ולא ללמידה מוכוונת מחיר בלבד.

7.2. ממשו הצעתכם בקובץ ImprovedKNNForest.py. על קובץ זה להדפיס בהרצתו את הדיוק שקיבלתם מהרצת האלגוריתם המשופר כאשר הוא לומד על קבוצת האימון ונבחן על קבוצת המבחן. אם ביצעתם ניסויים לקביעת פרמטרים לאלגוריתם שלכם, פרטו זאת. כמו כן צרפו בדו"ח את הדיוק שקיבלתם.

## הוראות הגשה:

- הגשה ביחידים בלבד.
- הגישו קובץ zip בודד המכיל את כל הקבצים שהתבקשתם לממש יחד עם קבצים נוספים כרצונכם בלי קבצי הדאטה שצורפו לכם.
- ודאו שכל הקבצים שלכם מחזירים מה שביקשנו מכם, שאינם פולטים גרפים, ושלא שכחתם להתייחס לנקודות מסויימות.
- נוהל האיחורים מופיע בסילבוס הקורס.

הרבה בריאות והצלחה.