תרגיל בית 3:

Decision Trees Learning

מטרות התרגיל

- התנסות בהפעלת אלגוריתמי **למידה** על בעיות סיווג.
- היכרות עם השפעת **רעש** ($noise$) בדוגמאות על הלמידה.

הערות

- תאריך הגשה: 21.1.16.
- את המטלה יש להגיש **בזוגות בלבד!**
- שאלות בנוגע לתרגיל יש לשלוח לניצן: snussa@cs.technion.ac.il
- אנא עיינו ברשימת ה-FAQ המתעדכנת באתר לפני פנייה בשאלות דרך המייל.
- אנא עקבו בתשומת לב אחר הוראות ההגשה המצורפות בסוף התרגיל לפני הגשתו.
- כמו בתרגילים קודמים בקורס, גם בתרגיל זה הרצת הניסויים עשוייה לקחת זמן רב ולכן מומלץ מאוד להמנע מדחיית העבודה על התרגיל לרגע האחרון. **לא תינתנה דחיות על רקע זה.**

מבוא והנחיות

במטלה זו נעסוק בלמידה של **עצי החלטה** ובניית ועדות של עצי החלטה. מומלץ לחזור על שקפי ההרצאות הרלוונטיים לפני תחילת העבודה על התרגיל.

במהלך התרגיל תתבקשו להריץ מספר ניסויים ולנתח את תוצאותיהם. אנא בצעו **ניתוח מעמיק ומפורט** של התוצאות וצרפו אותו לדו"ח כפי שיוסבר בהמשך התרגיל.

חלק א' (78 נק')

מבוא:

בתרגיל זה נתמקד בועדות של עצי החלטה. עצי ההחלטה בְּוועדה יִבָּנו לפי משפחת האלגוריתמים TDIDT לבניית עצי החלטה שנלמדה בהרצאות. כזכור, אלגוריתמים אלה לומדים עצי סיווג מדוגמאות נתונות, ובכל שלב בתהליך הלמידה בוחרים תכונה ($feature$) לפיה הם יפצלו את הצומת הנוכחי.

בתרגיל זה, פיצול צומת מתקיים כל עוד יש בו יותר דוגמאות מחסם המינימום $m = 4$, כלומר בתהליך בניית העץ מבוצע "גיזום מוקדם" כפי שלמדתם בהרצאות. שימו לב כי פירוש הדבר הינו שהעצים הנלמדים אינם בהכרח עקביים עם הדוגמאות.

לאחר סיום הלמידה (של עץ יחיד), הסיווג של אובייקט חדש באמצעות העץ שנלמד מתבצע לפי **רוב הדוגמאות** בעלה המתאים. הסיווג של אובייקט חדש באמצעות <u>הועדה כולה</u> מתבצע לפי החלטת הרוב של העצים המרכיבים אותה.

יצירת ועדות:

כאשר יוצרים ועדה (יער) של עצי סיווג יש צורך שעצי הועדה יהיו שונים זה מזה על מנת שיוכלו להגיע ל<u>סיווג משותף</u> טוב יותר. במקרה של אלגוריתמי TDIDT, נרצה לגוון את העצים בועדה על סמך **המידע לפיו מתבצעת הלמידה** ועל-ידי **האופן בו נבחרת התכונה לפיצול** בכל צומת בעץ.

המידע לפיו מתבצעת הלמידה מורכב מקבוצת הדוגמאות (המתויגות) הניתנות לאלגוריתם וכן קבוצת התכונות ($features$) של האובייקטים אליהן האלגוריתם מתייחס. ניתן ליצור עצי החלטה שונים ע"י הגבלת המידע ממנו הם נבנים, כלומר על ידי בנייתם על סמך תת-קבוצה של הדוגמאות המתויגות או תת-קבוצה של התכונות של האובייקטים.

א. בבניית עץ על סמך תת-קבוצה של **הדוגמאות** המתויגות, נבחר (לפני בניית העץ) את תת הקבוצה של הדוגמאות (מתוך כל הדוגמאות הנתונות) שעליה תתבצע הלמידה, ואותה נכניס כקלט לאלגוריתם יצירת העץ. תתי קבוצות שונות זו מזו יביאו ללמידת עצים שונים זה מזה בועדת המסווגים.
נסמן ב-N את גודל קבוצת הדוגמאות המתויגות, וב-p את החלק היחסי של הדוגמאות ממנו נבנים העצים בשיטה זו ($0 < p \leq 1$), אזי כל עץ הנבנה בשיטה זו נבנה על סמך תת-קבוצה רנדומלית של דוגמאות מגודל $\lceil p \cdot N \rceil$. בתרגיל זה (כולו) ערכו של p יהיה קבוע: $p = 0.67$.

ב. לעומת זאת, בבניית עץ על סמך תת-קבוצה של **התכונות**, נחליט מראש (כלומר לפני בניית העץ) מהי קבוצת התכונות של האובייקטים אליה האלגוריתם יתייחס (ולפיה הוא יהיה רשאי לבצע פיצולים בצמתים). כאשר ניתן לאלגוריתם את קבוצת הדוגמאות לפיה הוא ילמד את העץ, נייצג אותן רק באמצעות קבוצת התכונות שבחרנו (ולא באמצעות כל התכונות הידועות לנו).

<u>הבהרה</u>: בשיטה זו האלגוריתם לבניית העץ מקבל את **<u>כל</u>** הדוגמאות המתויגות.
נסמן ב-F את גודל קבוצת התכונות המקורית, וב-q את החלק היחסי של התכונות ממנו נבנים העצים בשיטה זו ($0 < q \leq 1$), אזי כל עץ הנבנה בשיטה זו נבנה על סמך תת-קבוצה רנדומלית של תכונות מגודל $\lceil q \cdot F \rceil$. בתרגיל זה (כולו) ערכו של q יהיה קבוע: $q = 0.67$.

האופן בו נבחרת התכונה לפיצול בכל צומת בעץ- תיתכנה שיטות רבות לבחירת התכונה לפיצול צומת. בתרגיל זה נתמקד בשלוש שיטות:

1. בחירת התכונה **הממקסמת את ערך ה-IG** ($Information\ Gain$) של הצומת. שיטה זו היא השיטה הנהוגה באלגוריתם ID3 השייך למשפחת האלגוריתמים TDIDT.
2. בחירה **רנדומלית** של תכונה לפיצול (מבין התכונות שלא נעשה לפיהן פיצול בַּמסלול מהשורש עד לצומת הנוכחי). בשיטה זו האלגוריתם בוחר באקראי תכונה לפיה יפצל את הצומת הנוכחי.
3. בחירה **סמי-רנדומלית** של תכונה לפיצול (מבין התכונות שלא נעשה לפיהן פיצול בַּמסלול מהשורש עד לצומת הנוכחי). בשיטה זו, ההסתברות לבחור תכונה מסוימת היא פרופרוציונלית לְתוספת האינפורמציה שלה, כלומר פרופרוציונלית לערך ה-IG ($Information\ Gain$) שיתקבל מהפיצול על פִּיהַ.

עבור צומת מסוים v בעץ, נסמן ב-E את קבוצת הדוגמאות שהוא מכיל ונסמן ב-$\{f_1, f_2, \dots, f_n\}$ את n התכונות שניתן לפצל אותו על פיהן (התכונות שלא נעשה לפיהן פיצול בַּמסלול מהשורש עד אליו). כמו כן נסמן ב-IG_i את ערך ה-$Information\ Gain$ המתקבל ע"י פיצול הצומת לפי התכונה ה-i, כלומר מתקיים (ביחס לסימונים שנלמדו בהרצאה):

$$InformationGain(f_i, E) = IG_i$$

בהינתן צומת v, נסמן ב-p_i את ההסתברות לפצל את v לפי התכונה ה-i בבחירה **סמי-רנדומלית** של תכונה לפיצול. נרצה שהסתברות זו תגדל ככל שתוספת האינפורמציה של התכונה ה-i (IG_i) תגדל. בנוסף, נרצה שהאלגוריתם יוכל לפצל צמתים לפי **כל תכונה** מבין $\{f_1, f_2, \dots, f_n\}$, על כן לכל תכונה f_i צריך להתקיים: $p_i > 0$. שימו לב כי האלגוריתם עשוי לפצל צמתים גם לפי תכונות שתוספת האינפורמציה שלהן **אפסית**, לכן גם עבור תכונות שערך ה-IG שלהן <u>שווה זהותית לאפס</u> צריך להתקיים $p_i > 0$.

שאלה 1 (הגדרת ההסתברות לפיצול צומת בבחירה סמי-רנדומלית):

<u>תזכורת</u>: הסימון $[n]$ שקול לסימון $\{1, 2, 3, \dots, n-1, n\}$.

<u>תזכורת נוספת</u>: $\{f_1, f_2, \dots, f_n\}$ הן n התכונות שניתן לפצל צומת נתון על פיהן (התכונות שלא נעשה לפיהן פיצול בַּמסלול מהשורש עד אליו). נבהיר כי קבוצה זו עשויה להשתנות בין הצמתים השונים בעץ.

נזכיר גם כי בהינתן צומת v, p_i היא ההסתברות לפצל **את** $\boldsymbol{v}$ לפי התכונה ה-i. עליכם להגדיר את הסתברות p_i זו. לשם כך עליכם:

א. לתת ביטוי (נוסחא) עבור ההסתברות p_i.

ב. להראות כי זוהי אכן פונקציית הסתברות, כלומר להראות כי מתקיים:

$$\forall i \in [n]: \quad 0 \le p_i \le 1$$

וכי מתקיים:

$$\sum_{i=1}^{n} p_i = 1$$

ג. להראות כי לכל צומת v, ההסתברות לפצל אותו לפי התכונה ה-i גְדֵלה ככל שתוספת האינפורמציה שלה (IG_i) גְדֵלה.

ד. להראות כי לכל צומת v ולכל תכונה f_i (גם עבור תכונה שעבורה מתקיים $IG_i = 0$), מובטח כי $p_i > 0$.

צרפו את התשובות לארבעת הסעיפים הנ"ל לדו"ח היבש של התרגיל.

סך הכל, נעסוק בִּשְׁתי דרכים ל<u>הגבלת המידע</u> לפיו מתבצעת הלמידה (הגבלת קבוצת הדוגמאות **<u>או</u>** הגבלת קבוצת התכונות) ובְּשלוש דרכים ל<u>בחירת התכונה</u> לפיצול צומת.

שימו לב כי המידע לפיו מתבצעת הלמידה והאופן בו נבחרת התכונה לפיצול בכל צומת בעץ הם **בלתי תלויים** האחד בשני. מכאן, שעל-ידי בחירה של <u>שיטה אחת</u> להגבלת המידע ו<u>שיטה אחת</u> לפיצול צמתים, ניתן להגיע לשישה אלגוריתמים שונים ליצירת עצים שונים בועדה. ועדת עצים שהעצים בה נוצרו ע"י בחירה כזו תיקרא **ועדה הומוגנית**, כלומר ועדה שכל העצים בה נוצרו על-ידי אלגוריתם TDIDT שהוגבל ע"י **בחירה מסוימת** של <u>שיטה אחת</u> להגבלת המידע ו<u>שיטה אחת</u> לפיצול צמתים. את שש הועדות ההומוגניות הנ"ל נסמן ע"י $Type = \{$א1, א2, א3, ב1, ב2, ב3$\}$, כאשר בכל סימון ה<u>אות</u> מציינת את אופן הגבלת המידע בלמידה, וה<u>מספר</u> מציין את האופן לפיו נבחרת התכונה לפיצול צומת. כך למשל שיטה א1 מתייחסת לבניית עצים הנבנים על סמך תת קבוצה של הדוגמאות המתויגות (א) ופיצול הצמתים בהם מתבצע על סמך בחירת תכונה הממקסמת את ה-$Information\ Gain$ (1).

:DATA SETS

בחלק זה של התרגיל (חלק א') נעסוק בִּשְׁנֵי $data-sets$.

1. ***Internet Advertisements*** העוסק בפרסומות ב-$web\ pages$. $Data-set$ זה מכיל 3279 דוגמאות, כאשר כל דוגמא מיוצגת ע"י 1558 תכונות (כמו למשל, גודל ה-$image$ או הופעתם של ביטויים מסוימים ב-URL). את ה-$data-set$ עצמו תוכלו להוריד מהקישור הבא: https://archive.ics.uci.edu/ml/datasets/Internet+Advertisements (שם גם תוכלו למצוא מידע נוסף על ה-$data-set$).

הערות: שלוש התכונות הראשונות של כל דוגמא עשויות להיות חסרות. על כן, אין להשתמש בהן בתהליך הלמידה (אין לפצל צמתים על פיהן). בנוסף, על מנת להקל את העבודה על $data-set$ זה (ולקצר את זמני הריצה של הניסויים) עליכם להשתמש רק ב-350 תכונות מתוך 1555 התכונות הנותרות. התכונות שתשתמשו בהן תיקבענה על סמך מספרי הזהות של שני המגישים. עליכם להשתמש בפונקציה $get\_ads\_features$ המסופקת לכם בקובץ $get\_features.py$ על מנת לקבל את האינדקסים שלהן. פונקציה זו תחזיר לכם קבוצת אינדקסים $Idxs$ שגודלה 350, כך שלכל $idx \in Idxs$ מתקיים: $3 \le idx \le 1557$. הקפידו להשתמש בקבוצת האינדקסים שניתנה לכם.

2. ***Har (Human Activity Recognition)*** העוסק בזיהוי הפעילות שאדם מבצע על סמך נתונים המתקבלים מחיישנים הנמצאים ב-$smartphone$ שלו. סה"כ קיימת התייחסות לשש פעילויות: הליכה, עלייה במדרגות, ירידה במדרגות, ישיבה, עמידה ושכיבה. $Data-set$ זה מכיל 10299 דוגמאות, כאשר כל דוגמא מיוצגת ע"י 561 תכונות (כמו למשל, מהירות ותאוצה). את ה-$data-set$ עצמו תוכלו להוריד מהקישור הבא:

https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones (שם גם תוכלו למצוא מידע נוסף על ה-$data-set$).

הערות: על מנת להקל את העבודה על $data-set$ זה, נרצה לסווג את הדוגמאות בו לשתי מחלקות בלבד: תנועה ומנוחה (במקום שש המחלקות המקוריות המוגדרות בו). לשם כך, כל דוגמא שהיתה מסווגת עד כה כ"הליכה", "עלייה במדרגות" או "ירידה במדרגות" תְּסֻוַּג כתנועה, וכל דוגמא שהיתה מסווגת עד כה כ"ישיבה", "עמידה" או "שכיבה" תְּסֻוַּג כמנוחה. המיפוי בין הסיווג עצמו לבין הספרה $digit \in \{1,2,\dots,6\}$ המציינת אותו נמצא בקובץ $activity\_labels.txt$. המיפוי החדש יצויין ע"י הספרות $\{0,1\}$, כאשר הספרה 0 תציין "מנוחה", והספרה 1 תציין "תנועה".

בנוסף, את הלמידה נבצע רק על הדוגמאות המופיעות בתיקייה $train$ המכילה 7352 דוגמאות. שימו לב כי הדוגמאות עצמן (ללא הסיווג שלהן) מופיעות בקובץ $X\_train.txt$ שבתיקייה $train$, בעוד הסיווגים המתאימים להן נמצאים בקובץ $y\_train.txt$ שבתיקייה זו. הבחנה זו רלוונטית (בין השאר) עבור השימוש בַּפונקציה get_noisy_folds המסופקת לכם.

Cross Validation:

בהינתן $data-set$ מסוים, נסמן את קבוצת כל הדוגמאות שבו ע"י $Examples$. כמו כן, עבור כל $data-set$ נתון נחלק את הדוגמאות שבו לעשרה $folds$ שיישארו **קבועים** לאורך כל הניסויים (עבור כל סוגי הועדות שנבחן) ונסמנם: $Folds = \{fold_1, fold_2, \dots, fold_{10}\}$. היותם של ה-$folds$ הנ"ל קבועים לאורך כל הניסויים היא **קריטית** לְשֵם יכולת ההשוואה בין הועדות השונות, ועליכם לוודא את קיומה.

כאשר נרצה לחשב את ערך הדיוק של אלגוריתם למידה מסוים או ועדה הומוגנית מסוג כלשהו (ועדה שמאופיינת ע"י **מספר העצים** בה וכן **בחירה מסוימת** של <u>שיטה אחת</u> להגבלת המידע ו<u>שיטה אחת</u> לפיצול צמתים) נפעל באופן הבא $(*)$:

לכל $i \in \{1,2,\dots,10\}$:

א. הפעל את אלגוריתם הלמידה (או לְמַד ועדה הומוגנית מהסוג הרצוי) כאשר קבוצת הדוגמאות המשמשת ללמידה (קבוצת האימון) היא $train_i = Examples/fold_i$, כלומר קבוצת כל הדוגמאות שאינן שייכות ל-$fold_i$. $(**)$

ב. חַשֵב את אחוזי הדיוק $(accuracy)$ של הועדה שנלמדה על קבוצת הבדיקה שתהא ה-$fold$ ה-i, כלומר: $test_i = fold_i$. נסמן אחוזי דיוק אלה ב-acc_i.

כעת, הגדר את ערך הדיוק של אלגוריתם הלמידה (או הועדה ההומוגנית מהסוג הנ"ל) להיות הממוצע בין ערכי הדיוק שנמצאו, כלומר:

$$accuracy = \frac{1}{10} \cdot \sum_{i=1}^{10} acc_i$$

<u>הערות</u>:

$(*)$ שימו לב כי הלמידה וכן החישוב של אחוזי הדיוק מתבצעים לאחר שהחלוקה ל-$folds$ כבר בוצעה.

$(**)$ שימו לב כי בפועל, נלמד את עצי ההחלטה על קבוצת אימון שונה במעט (יפורט בהמשך) ולא על הקבוצה $train_i$ עצמה.

Noise:

היתרון הגדול של ועדות מסווגים על פני אלגוריתמי למידה אחרים הוא יכולת ההתמודדות שלהן עם קבוצת אימון מורעשת. נזכיר כי קבוצת דוגמאות **מורעשת** היא קבוצת דוגמאות שמכילה דוגמאות רועשות, ודוגמא (מתויגת) רועשת היא דוגמא שהסיווג/התיוג הנתון שלה <u>שגוי</u> (שונה מהסיווג <u>האמיתי</u> שלה).

בחלק א' נבצע את הלמידה של עצי ההחלטה על קבוצת אימון מורעשת, שנייצר אותה באמצעות הרעשה מלאכותית של הדוגמאות.

בהינתן קבוצת דוגמאות $Examples$ מ-$data-set$ מסוים, נרצה ליצור עבורה קבוצה $Examples'$ בה **חלק מהדוגמאות** הן רועשות. לשם כך, לכל $fold_i \in Folds$ ניצור את הגרסה המורעשת שלו, $fold'_i$, ע"י הרעשת חלק מהדוגמאות שב-$fold_i$. ההרעשה תתבצע באמצעות הפונקציה get_noisy_folds המסופקת לכם, עם פרמטר רעש קבוע שערכו: $noise = 0.3$.

באופן זה נקבל עשרה $folds$ מורעשים: $Folds' = \{fold'_1, fold'_2, \dots, fold'_{10}\}$ שגם הם ישארו קבועים לאורך כל הניסוי, והקבוצה $Examples'$ תוגדר להיות:

$$Examples' = \bigcup_{i=1}^{10} fold'_i$$

כעת, כאשר נרצה ללמוד ועדת מסווגים, נבצע את הלמידה של הועדה ה-i על קבוצת האימון:

$$train'_i = Examples' / fold'_i$$

(ו**<u>לא</u>** על קבוצת האימון $train_i$ המקורית!).

שימו לב, כי על מנת לחשב את אחוזי הדיוק יש עדיין להשתמש בקבוצות הבדיקה ש**<u>אינן מורעשות</u>**, זאת על מנת לקבל מדד אמין על אחוזי הדיוק שהושגו בפועל. כלומר, אחוזי הדיוק של הועדה ה-i יחושבו על קבוצת הבדיקה $test_i = fold_i$. כמו כן, ערך הדיוק של הועדה ההומוגנית מהסוג הנ"ל עדיין מוגדר כַּממוצע בין ערכי הדיוק שנמצאו, כלומר:

$$accuracy = \frac{1}{10} \cdot \sum_{i=1}^{10} acc_i$$

קבצים המסופקים לכם:

1. $get\_features.py$ – המכיל את הפונקציה get_ads_features שתשמש אתכם בעיבוד המקדים של $Internet\ Advertisements$.
2. $noise.py$ – המכיל את הפונקציה get_noisy_folds שתשמש אתכם במהלך הניסויים לחלוקת הדוגמאות לעשרה $folds$ "רגילים" ועשרה $folds$ מורעשים.

הרצת הניסוי:

נזכיר כי גודל ועדה של עצי סיווג הוא מספר העצים הנמצאים בה.

נסמן ב- $Size = \{11, 21, 31, 41, 51, 61, 71, 81, 91, 101\}$ את קבוצת הגדלים של הועדות. כמו כן נסמן ב- $Type$ את ששת הסוגים של הועדות ההומוגניות האפשריות, $Type = \{$א1, א2, א3, ב1, ב2, ב3$\}$ (כפי שתואר בהסבר על יצירת הועדות).

בחלק א', לכל $data - set$:

א. בצעו עיבוד מקדים על ה-$data - set$ כפי שפורט לעיל בחלק העוסק ב-$Data\ Sets$.

ב. חלקו את הדוגמאות בַּ-$data - set$ לעשרה $folds$ וכן צרו מהם עשרה $folds$ רועשים. לשם כך עליכם להסתייע בפונקציה get_noisy_folds המסופקת לכם.

ג. חשבו את מידת הדיוק $(accuracy)$ של עץ בודד שנוצר באמצעות הרצה של אלגוריתם ID3 שנלמד בכתה ע"י שימוש ב- $cross - validation$. שימו לב שלְשֵׁם כך עליכם בעצם **ללמוד** עשרה עצים שונים ולְמַצֵּעַ את מידת הדיוק שלהם. נסמן את מידת הדיוק שהושגה כאן ע"י $acc(ID3)$.

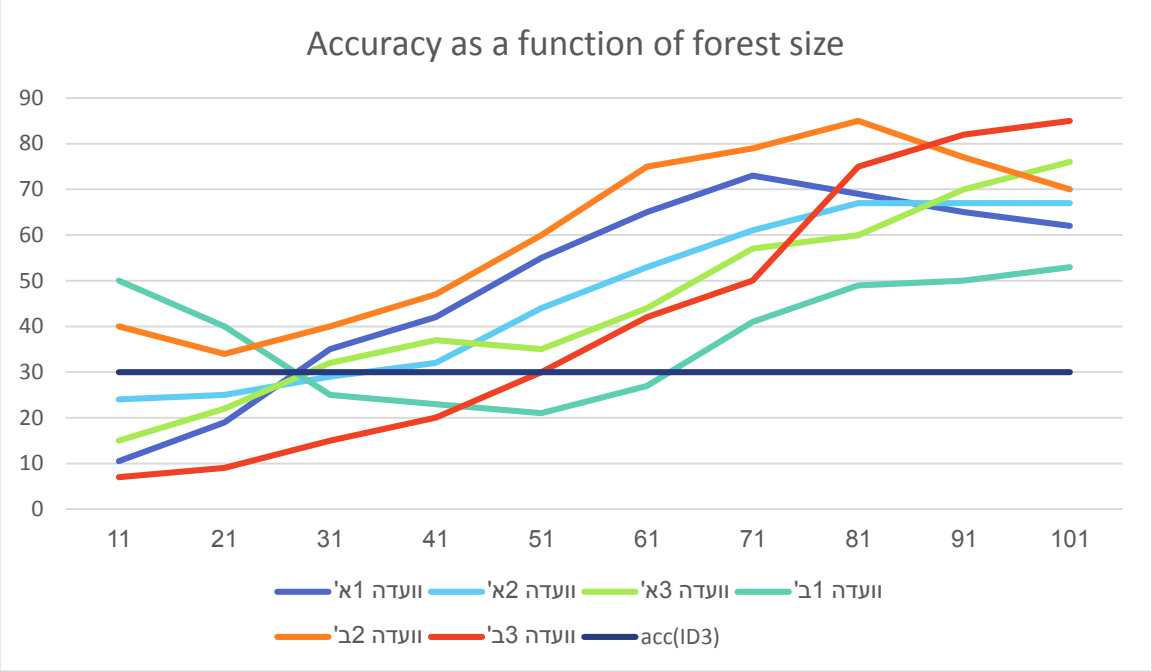
ד. לכל סוג ועדה $t \in Type$:

לכל גודל ועדה $s \in Size$:

חשבו את מידת הדיוק $(accuracy)$ של הועדה מסוג t מגודל s ע"י שימוש ב- $cross - validation$. שימו לב שלְשֵׁם כך עליכם בעצם **ללמוד** 10 ועדות שונות מסוג t מגודל s ולְמַצֵּעַ את מידת הדיוק שלהן.

לבסוף, לכל $data - set$, עליכם להציג גרף המתאר את **הדיוק** (באחוזים) של ששת האלגוריתמים כפונקציה של **גודל** הועדות ההומוגניות שנלמדו. הוסיפו לגרף קו אופקי המתאר את מידת הדיוק של עץ סיווג בודד שנוצר באמצעות הרצה של אלגוריתם ID3. קו זה מקביל לציר ה-x וגובהו $acc(ID3)$.

כלומר סך הכל, בחלק א' עליכם להציג שני גרפים מהצורה $(*)$:



כאשר לכל גרף כזה עליכם לענות בִּמְפורט על הסעיפים הבאים:

א. נתחו את ההשפעה של גודל הועדה על הדיוק המתקבל. יש להסביר את מגמות העלייה (או ירידה) בַּדיוק של הגדלים של הועדות השונות, כפי שהן מוצגות בגרף.

ב. השוו בין השיטות השונות ל**הגבלת המידע** עפ"י הגרף. עבור אילו גדלים של ועדה יש העדפה להגבלת קבוצת הדוגמאות ועבור אילו גדלים של ועדה יש העדפה להגבלת קבוצת התכונות? מדוע לדעתכם זה קורה?

ג. השוו בין השיטות השונות ל**פיצול צמתים** עפ"י הגרף. עבור אילו גדלים של ועדה יש העדפה לפיצול צמתים על-סמך $maximal - InformationGain$? עבור אילו גדלים יש העדפה לפיצול על-סמך בחירה אקראית של תכונות? עבור אילו גדלים יש העדפה לפיצול על-סמך בחירה סמי-רנדומלית של תכונות? מדוע לדעתכם זה קורה?

ד. השוו בין הועדות השונות (ללא ההתייחסות לעץ הסיווג היחיד). עבור סעיף זה, אלגוריתם למידה של ועדה מתקבל ע"י בחירה מסוימת של <u>שיטה אחת</u> להגבלת המידע ו<u>שיטה אחת</u> לפיצול צמתים. האם קיים אלגוריתם למידה של ועדה (מבין ששת האלגוריתמים שחקרתם) העדיף על פני האלגוריתמים האחרים? האם קיימת תלות בין גודל הועדה לבין האלגוריתם העדיף <u>עבור גודל זה</u>? מדוע לדעתכם זה המצב?

ה. השוו בין למידה של עץ סיווג יחיד לפי אלגוריתם $ID3$ לבין למידה של ועדות מסווגים. האם קיים אלגוריתם (מבין השִבְעה) העדיף תמיד על פני האלגוריתמים האחרים? האם קיימים סוגי ועדות שהינם עדיפים על-פני העץ הבודד עבור כל הגדלים שנבדקו בניסוי? האם קיים גודל ועדה מסויים אשר החל ממנו השימוש בועדות עדיף תמיד על-פני השימוש בעץ בודד? האם קיימת תלות בין סוג הועדה לבין הגודל הנדרש לשם הפגנת עדיפות על העץ הבודד? מדוע לדעתכם זהו המצב?

לכל גרף כנ"ל צרפו גם את **טבלת הנתונים** שיצרו אותו, למשל $(*)$:

גודל הועדה	ועדה1 א'	ועדה2 א'	ועדה3 א'	ועדה1 ב'	ועדה2 ב'	ועדה3 ב'	Acc(ID3)
11	10.5	24	15	50	40	7	30
21	19	25	22	40	34	9	30
31	35	29	32	25	40	15	30
41	42	32	37	23	47	20	30
51	55	44	35	21	60	30	30
61	65	53	44	27	75	42	30
71	73	61	57	41	79	50	30
81	69	67	60	49	85	75	30
91	65	67	70	50	77	82	30
101	62	67	76	53	70	85	30

$(*)$ שימו לב כי הגרף והטבלה שהצגנו כאן נוצרו על סמך ערכים שרירותיים שקבענו לצורך **המחשה בלבד**. לא מובטח שהתוצאות שתקבלו בהרצת הניסויים שלכם תהיינה דומות לערכים המוצגים בגרף ובטבלה שלעיל.

חלק ב' (22 נק')

בחלק ב' נבחן את ההשפעה של מידת **הרעש** בנתונים על מידת **הדיוק** של ששת האלגוריתמים מחלק א', וכן על מידת הדיוק של סיווג באמצעות עץ סיווג **יחיד**.

לשם כך בכל חלק זה גודל הועדות יהיה **קבוע** וערכו יהיה $s = 101$ עצי-החלטה. כמו כן, גם בחלק ב' פיצול צומת מתקיים כל עוד יש בו יותר דוגמאות מחסם המינימום $m = 4$.

יצירת ועדות:

בחלק ב' נשתמש בְּועדות **הומוגניות** מששת הסוגים שהוצגו בחלק א' (שימו לב כי לא ניתן להשתמש באותן הועדות ממש מחלק א', אלא יש לבצע למידה מחדש לפי הפרמטרים של חלק זה).

בחלק ב' נבחן גם את מידת הדיוק של סיווג לפי **עץ סיווג יחיד** שנבנה על-סמך אלגוריתם הלמידה ID3.

Data Sets:

בחלק זה של התרגיל (חלק ב') תחקרו $data - set$ יחיד אותו תבחרו כרצונכם מבין שני ה-$data - sets$ בהם השתמשתם בחלק א' של התרגיל.

Noise & Cross Validation:

בחלק ב' נבצע את הלמידה של עצי ההחלטה על קבוצת דוגמאות **מורעשת**, כאשר בהרצות שונות נשנה את **מידת הרעש** המוכנס לנתונים על פיהם תתבצע הלמידה.

מבחינה טכנית, כמו בחלק א', בהינתן קבוצת דוגמאות $Examples$ מ-$data-set$ מסוים, ניצור עבורה קבוצה $Examples'$ בה **חלק מהדוגמאות** הן רועשות. מידת הרעש תיקבע ע"י שינוי ערכו של פרמטר הקלט $noise$ של הפונקציה get_noisy_folds (מידת הרעש עולָה ככל שערכו עולֶה). הערכים איתם נבצע את הלמידה יהיו כל הערכים בין אפס ל-0.3 בקפיצות של 0.05 (כולל ערכי הקצה) כלומר:

$$noise \in \{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$$

כמו בחלק א', נחלק את שתי קבוצות הדוגמאות ($Examples$ ו-$Examples'$) לעשרה $folds$ שישארו **<u>קבועים</u>** **<u>לאורך כל הניסוי</u>** ונסמנם: $Folds = \{fold_1, fold_2, \dots, fold_{10}\}$ ו- $Folds' = \{fold'_1, fold'_2, \dots, fold'_{10}\}$ בהתאמה (כאשר כמו בחלק א', $fold'_i$ נוצר מ-$fold_i$ ע"י הרעשה של חלק מהדוגמאות שבו).

כאשר נרצה ללמוד ועדת מסווגים, נבצע את הלמידה של הועדה ה-i על קבוצת האימון **המורעשת** $train'_i = Examples'/fold'_i$, אחוזי הדיוק של הועדה ה-i יחושבו על קבוצת הבדיקה $test_i = fold_i$ וערך הדיוק של הועדה ההומוגנית מהסוג הנ"ל עדיין מוגדר כַּממוצע בין ערכי הדיוק שנמצאו, כלומר:

$$accuracy = \frac{1}{10} \cdot \sum_{i=1}^{10} acc_i$$

הרצת הניסוי:

נסמן ב-$Noise = \{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$ את קבוצת הערכים המשמשת לקביעת היקף הרעש שיוכנס לנתונים. כמו כן, בדומה לחלק א', נסמן ב-$Type$ את **שבעת הסוגים** של המסווגים האפשריים (שש ועדות הומוגניות אפשריות וכן עץ סיווג יחיד), $Type = \{1\text{א}, 2\text{א}, 3\text{א}, 1\text{ב}, 2\text{ב}, 3\text{ב}, \boldsymbol{ID3}\}$.

בחלק ב', עבור ה-$data-set$ שבחרתם:

לכל ערך רעש $noise \in Noise$:

א. חלקו את הדוגמאות בַּ-$data-set$ לעשרה $folds$ וכן צרו מהם עשרה $folds$ רועשים (לפי הפרמטר $noise$). לשם כך עליכם להסתייע בפונקציה get_noisy_folds המסופקת לכם.

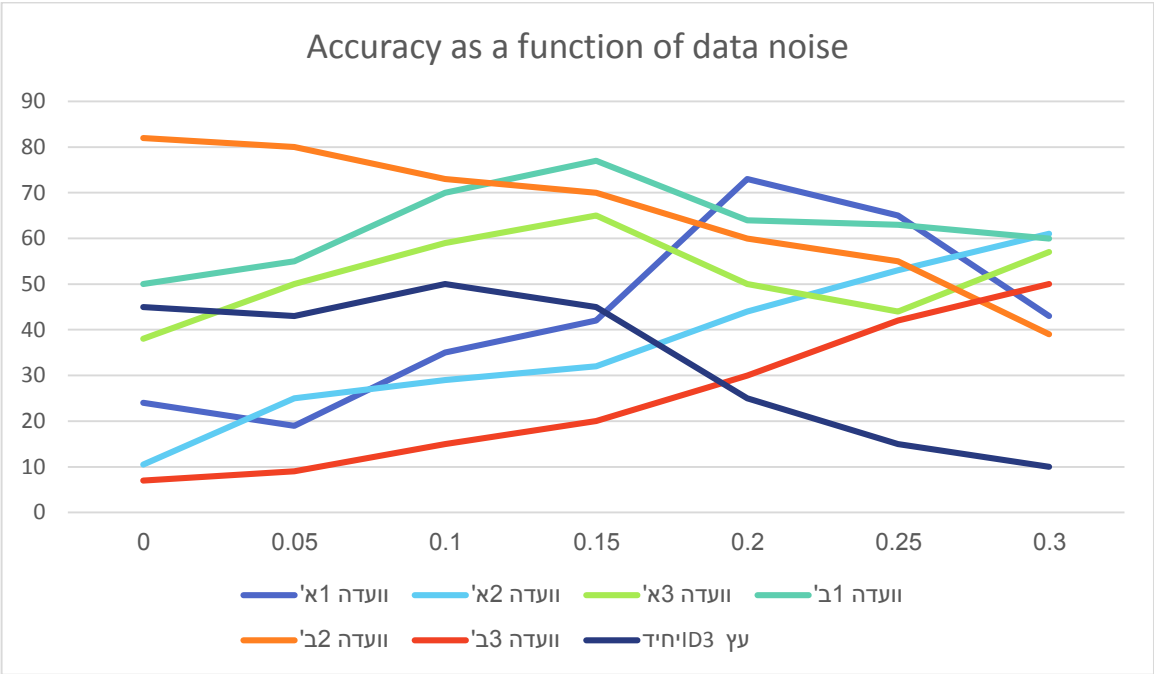
ב. לכל סוג מסווג (ועדה או עץ יחיד) $t \in Type$:

חשבו את מידת הדיוק ($accuracy$) של המסווג (ועדה/עץ יחיד) מסוג t ע"י שימוש ב- $cross-validation$. שימו לב שלשם כך עליכם בעצם **ללמוד** 10 מסווגים

שונים מסוג t וּלְמַצֵּעַ את מידת הדיוק שלהם. כמו כן עליכם לשים לב כי כל המסווגים נלמדים על סמך אותם ה-$folds$ שיצרתם בשלב א' לעיל.

לבסוף, עליכם להציג גרף המתאר את **הדיוק** (באחוזים) של שבעת האלגוריתמים כפונקציה של מידת **הרעש** (המבוטאת ע"י ערכו של הפרמטר $noise$) עבור ה-$data-set$ שבחרתם.

כלומר סך הכל, בחלק ב' עליכם להציג גרף יחיד מהצורה $(*)$:



כאשר עבורו עליכם לענות בִּמְפורט על הסעיפים הבאים:

א. נתחו את ההשפעה של מידת הרעש על הדיוק המתקבל. יש להסביר את מגמות העלייה (או ירידה) בַּדיוק של מידות הרעש השונות, כפי שהן מוצגות בגרף.

ב. השוו בין הועדות השונות (ללא ההתייחסות לעץ הסיווג היחיד). עבור סעיף זה, אלגוריתם למידה של ועדה מתקבל ע"י בחירה מסוימת של <u>שיטה אחת</u> להגבלת המידע ו<u>שיטה אחת</u> לפיצול צמתים. האם קיים אלגוריתם למידה של ועדה (מבין ששת האלגוריתמים שחקרתם) העדיף על פני האלגוריתמים האחרים? האם קיימת תלות בין מידת הרעש לבין האלגוריתם העדיף <u>עבור מידת רעש זו</u>? מדוע לדעתכם זה המצב?

ג. השוו בין למידה של עץ סיווג יחיד לפי אלגוריתם $ID3$ לבין למידה של ועדות מסווגים. האם קיים אלגוריתם למידה של ועדה/עץ (מבין שבעת האלגוריתמים שחקרתם) העדיף (תמיד) על פני האלגוריתמים האחרים? האם קיימים סוגי ועדות שהינם עדיפים על-פני העץ הבודד עבור כל מידות

הרעש שנבדקו בניסוי? האם קיימת מידת רעש מסויימת אשר החל ממנה השימוש בועדות עדיף תמיד על-פני השימוש בעץ בודד? האם קיימת תלות בין סוג הועדה לבין מידת הרעש הנדרשת לשם הפגנת עדיפות על העץ הבודד? מדוע לדעתכם זהו המצב?

לכל גרף כנ"ל צרפו גם את **טבלת הנתונים** שיצרו אותו, למשל (*):

מידת הרעש	ועדה 1א'	ועדה 2א'	ועדה 3א'	ועדה 1ב'	ועדה 2ב'	ועדה 3ב'	עץ ID3 יחיד
0	24	10.5	38	50	82	7	45
0.05	19	25	50	55	80	9	43
0.1	35	29	59	70	73	15	50
0.15	42	32	65	77	70	20	45
0.2	73	44	50	64	60	30	25
0.25	65	53	44	63	55	42	15
0.3	43	61	57	60	39	50	10

(*) שימו לב כי הגרף והטבלה שהצגנו כאן נוצרו על סמך ערכים שרירותיים שקבענו לצורך **המחשה בלבד**. לא מובטח שהתוצאות שתקבלו בהרצת הניסויים שלכם תהיינה דומות לערכים המוצגים בגרף ובטבלה שלעיל.

חלק ג'- שאלת בונוס (עד 7 נק')

לאור הניסויים שערכנו בְּשני חלקיו של התרגיל, הַציעו אלגוריתם ליצירת ועדה שאיננה בהכרח הומוגנית. נמקו את בחירתכם (אולי בעזרת ניסוי ☺).

בונוס יינתן עבור תשובות מנומקות ויצירתיות במיוחד.

הוראות הגשה

- הגשת התרגיל תתבצע **אלקטרונית בלבד**.
- עליכם להגיש קובץ ארכיון יחיד בשם: AI3_<id1>_<id2>.zip (ללא הסוגריים המשולשים). קובץ זה יכיל:
  - קובץ בשם readme.txt בפורמט הבא:

```
name1 id1 email1
```

name2 id2 email2

- קובץ בשם AI_HW3.PDF המכיל את דו"ח הניסויים שערכתם, תשובות לחלק היבש והערות לקוד שהגשתם (כולל תפקיד כל קובץ הנמצא בתיקייה שהגשתם).
- כל **חבילה חיצונית** בה השתמשתם, זאת על מנת שיהיה אפשר להריץ את הקוד שלכם על כל מחשב. עליכם לציין בקובץ AI_HW3.PDF באילו חבילות כנ"ל השתמשתם בהרצת הניסויים.
- כל **קוד עזר** שכתבתם/השתמשתם בו לשם הרצת הניסויים או יצירת הגרפים.

- אין להעתיק את הקבצים המסופקים לכם אל תוך תיקיית ההגשה. הניחו כי קבצים אלו יהיו זמינים בעת בדיקת התרגיל.
- שימו לב שכל הפנייה למיקום קובץ/תיקייה כלשהם בקוד תהיה רלטיבית ($relative\ path$) ולא אבסולוטית, כך שהקוד יעבוד כפי שהוא על כל מחשב בכל מיקום שנבחר לתיקיית הפרוייקט. הקפידו לבדוק זאת לפני ההגשה!
- "המצאת" נתונים לצורך בניית הגרפים **אסורה** ותוביל לדיון בבית הדין המשמעתי של הטכניון.
- אתם רשאים לעשות שימוש בכל קוד שתמצאו ברשת, אך כל קוד חיצוני **מחייב הצהרה מפורשת** על המקור שלו בקובץ AI_HW3.PDF. אי-קיום דרישה זו מהווה עבירה משמעתית!
- הקפידו על קוד **<u>ברור, קריא ומתועד!</u>** עליכם לתעד כל חלק שאינו טריוויאלי בקוד שלכם. בפרט, אם התשמשתם בקוד שנמצא ברשת וביצעתם בו שינויים, עליכם לתעד זאת.

בהצלחה!!