תרגיל בית 3 – מבוא לבינה מלאכותית

סיון יוסבשוילי 318981586

חורף תשפ"א

1. תוצאת הדיוק :

```
out of 113 you are right about:
107
your accuracy is :
0.9469026548672567
```

2. **הטענה נכונה.** נסתכל על פונקציית הנרמול -

$$minmax(x) = \frac{x - x_{Min}}{x_{Max} - x_{Min}}$$

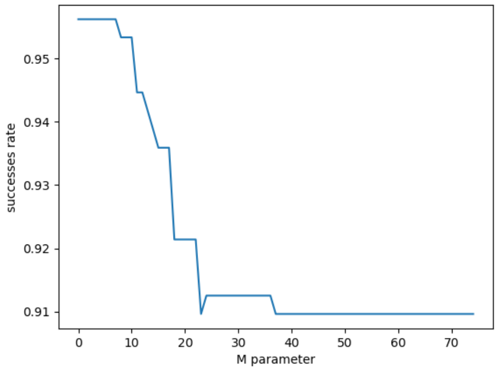
זוהי פונקציה לינארית רציפה, נסמנה $N(x)$. עבור ערכים $v_1 \le v_2 \le \cdots \le v_n$ מתקיים כי הפעלת הפונקציה משמרת סדר- $N(v_1) \le N(v_2) \le \cdots \le N(v_n)$ (המקדם המוביל הינו חיובי). נזכור כי הבחירה שנבצע בכל צומת לגבי ערך הפיצול עדיין תלוי בקבלת IG מקסימלי. נסתכל על צומת פיצול מקורי **לפני** שנרמלנו את הסט. פיצלנו לפי תכונה i וערך k וקיבלנו שתי קבוצות X, Y. מתקיים כי

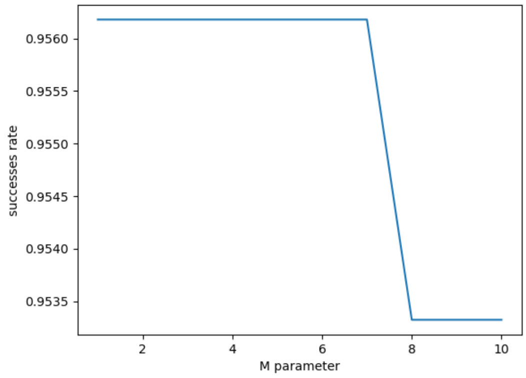
$$IG(f_i, X \cup Y) = entropy(X \cup Y) - \frac{|X|}{|X \cup Y|}Entropy(X) - \frac{|Y|}{|X \cup Y|}Entropy(Y)$$

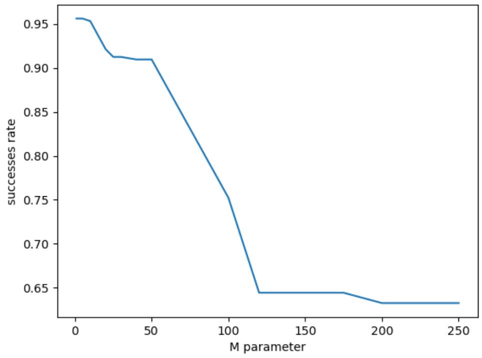
נשים לב כי בשל שימור הסדר, עדיין מתקיים עבור $N(k)$ (כאשר k היה ממוצע הערכים v_i, v_{i+1} כפי שמבצעים בחלוקה דינמית לתכונות רציפות, זאת בגלל לינאריות פונקציית הממוצע ולינאריות $N(x)$) כי נקבל את אותו הפיצול ל-X, Y . הפונקציה מעבירה את k ל-$N(k)$ ומתקיים כי $N(k)$ הינו ממוצע הערכים $N(v_i), N(v_{i+1})$, לכן למעשה גם לאחר הטרנספורמציה הלינארית, של נירמול המינמקס, עדיין הבחירות שמבצע אלגוריתם $ID3$ לא ישתנו, כי הערכים כולם שומרים על סדר ביניהם, (וכן ערכי הפיצול שנבחרו בהרצה המקורית), ומפה חישוב תוספת האינפורמציה אינו מושפע מהערכים של הדוגמאות אלא מגדלי הקבוצות לפי ערכי הפיצול, שכאמור אינם משתנים כי $N(x)$ משמרת סדר. נשים לב כי טיעון זה נכון לתיוג בינארי (האנטרופיה בכל צומת מחושבת לפי סכום של שני משתנים)! לכן, יתקבלו אותן ההחלטות שהתקבלו באלגוריתם המקורי ותוצאת הדיוק עבור אימון על קבוצת האימון ומבחן על קבוצת המבחן תישאר כמות שהיא.

3. 1) פעולת הגיזום בעצי החלטה מגיעה מהצורך להתגבר על בעיית ה-overfitting. בשל היותו של אלגוריתם TDIDT עקבי עם קבוצת האימון, אנו אומנם מגיעים לשגיאת אימון אפס, אך עבור דוגמאות רועשות העץ יבצע התאמה עבורן. כדי להחליש את אפקט התאמת היתר נבצע גיזום של העץ (החלפת תת עץ בעלה). חשיבות פעולה זו גבוהה.

3) להלן שלושה גרפים המציגים את השפעת הפרמטר M על הדיוק עם סקאלות שונות של הפרמטר m :







עבור הערכים $m = 1{,}2{,}3{,}5$ (מבין אלה שנבדקו) התקבל הדיוק המקסימלי – 0.95618.

זה הגיוני, קבוצת האימון שלנו היא בסדר גודל של 340 דוגמאות. בשיטה בה עבדנו קבוצת האימון היא $\frac{4}{5}$ מגודל זה, בערך 272 דוגמאות לאימון. עבור ערכי m שקיבלנו, אנו יכולים למזער את אפקט התאמת היתר מבלי לפגוע מדי באמינות המסווג. עבור ערכי m גדולים מדי, אנו מוותרים על עקביות העץ ואילו עבור ערכי m שהצגתי, אנו מקבלים אפשרות סבירה לביטול דוגמאות רועשות (מתוך סט של 272 דוגמאות, סביר להניח שסדר גודל של עד 10 מהן יהיו רועשות, אחרת סט הדאטה כולו "מלוכלך" ולא מהימן).

עבור הרצת האלגוריתם עם m=1,2,3 על כל קבוצת האימון קיבלתי את אותם אחוזי דיוק כמו ההרצה ללא גיזום בסעיף 1 -

0.9469026548672567

לדוגמה, עבור הרצה של האלגוריתם עם פרמטר $m = 20$ (שאינו חזר כמיטבי מהניסוי) על כל קבוצת האימון, אני מקבלת שגיאה של -

0.9734513274336283, שהיא טובה אף מהדיוק הרגיל ללא גיזום וכן מהדיוק המתקבל מהניסוי.

הסיבה היא ששימוש בשיטת $k - cross$ מחפשת למעשה פרמטר שהכי "יציב". היא מייצרת לה כמה קבוצות טסטים במקום אחת וממצעת את הפרמטר שייתן את אחוז הדיוק המירבי עבור ה"מקרה הכללי". ספציפית עבור קבוצת המבחן שקיבלנו, יכולים להתקבל ערכי דיוק טובים יותר עבור m-ים אחרים, אך ניתן להניח כי במקרה הכללי והממוצע, ועם הרצת הניסוי על קבוצה רחבה ומייצגת של דוגמאות אימון השימוש ב-$k\ croos$ ייתן את ה-m ששואף לתוצאה המיטבית .

4. מזעור פונקציית ה- loss–

1)ערך ה-loss של המסווג הוא -

```
0.021238938053097345
```

.

הגיזום ($m = 1,2$) לא שיפר את אחוזי הדיוק.

2) ראשית נאתחל את מסווגי $ID3$ המתוארים בתהליך זה עם **פרמטר $\mathbf{m = 25}$** יוסבר בהמשך).
ניקח את קבוצת האימון, ונתאמן עליה תוך שימוש בשיטת k cross- נבצע עבורה חלוקה ל-5 קבוצות. ניצור 5 מסווגים שיתאמנו כל פעם על $\frac{4}{5}$ מקבוצת האימון, ויבדקו על ידי $\frac{1}{5}$.

עבור כל מסווג נבדוק את ערך ה-$loss$ שלו ונחזיר את **המסווג עם ה-$\boldsymbol{loss}$ המינימלי**.

נבצע שני שינוים קלים באלגוריתם $ID3$ מתוך רצון לצמצם את ה-loss:

*כאשר נבחר ערך סף x לפיצול בצומת, אם הסיווג של מתחת לסף הוא אדם חולה, **נגדיר את סף הפיצול** להיות $(1+\delta)x$, ואם הסיווג מעל x הוא של חולה נבחר את סף הפיצול להיות $(1-\delta)x$. המחשבה העומדת מאחורי זה, היא שנעדיף לסווג דוגמאות מקו התפר בסיווג של חולה, בשל ההשפעה הקריטית של סיווג שגוי של חולה כבריא (אינטואיציה – אם סטודנט מצטיין הוא סטודנט עם ממוצע 90, מה נחשב סטודנט עם ממוצע 89?).

***בבחירת הערך הדיפולטיבי של צומת** (שצומת אב שולח לצמתי הילדים שלו) בפונקציית $majority\ class$ במקום להחזיר את התווית הדומיננטית, נחזיר את התווית 'חולה' עבור מצב בו $|healthy| \geq |sick| * (1+\alpha)$, כלומר נעדיף לתייג מצב שאינו חד משמעית מוטה כלפי אחת מהתגיות לכיוון תיוג 'חולה'.

ביצעתי ניסוי המשקלל את הפרמטרים m, α, δ וקיבלתי את ה-$loss$ המינימלי עבור $m = 25, \alpha = 0.1, \delta = 0.05$.
כמובן שפרמטרים אלו יהיו מדויקים יותר עבור סט דאטה גדול יותר, אך במסגרת הדאטה הניתן אלו הם הפרמטרים המיטביים.

טרם השיפור, המצב היה $|FP| = 4, |FN| = 2$,

ובאלגוריתם המשופר מתקיים כי $|FP| = 2, |FN| = 0$ וניתן להסיק כי המטרה העיקרית שלי (צמצום ערך FN כדי לצמצם את ה-$loss$) הושגה.

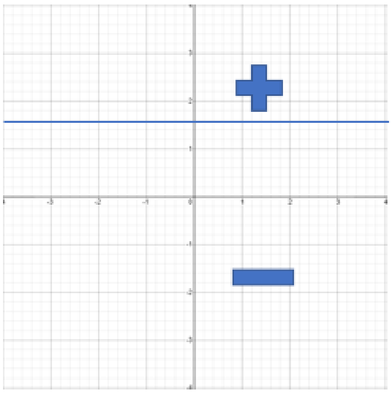
```
loss for ID3: 0.021238938053097345
now try to minimize loss, loss of costSensitiveID3 is:
0.0017699115044247779
```

.5

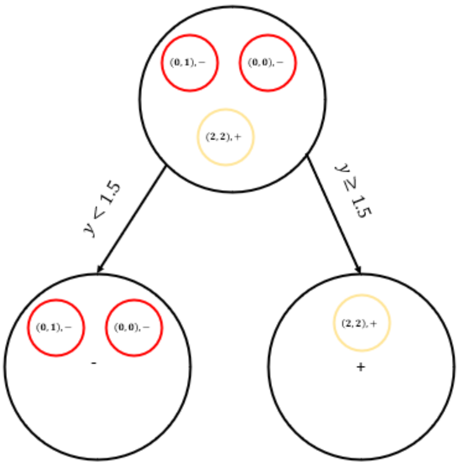
נציג דוגמאות למסווגים כנדרש:

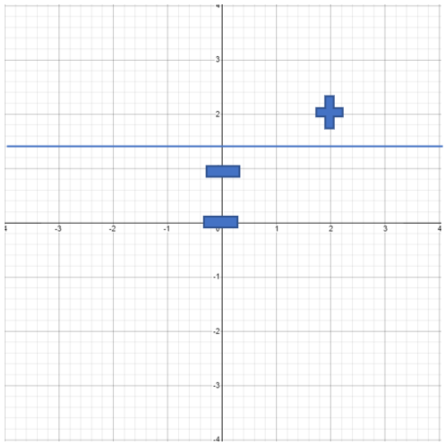
א. נסתכל על מסווג מהצורה - $(x,y) \to \{1,0\}$. המחזירה 1 אמ"מ $y \geq 1.5$. עבור קבוצת האימון $D = \{\big((0,0),-\big),\big((0,1),-\big),((2,2),+)\}$, נקבל -

מסווג מטרה -



ומסווגים -

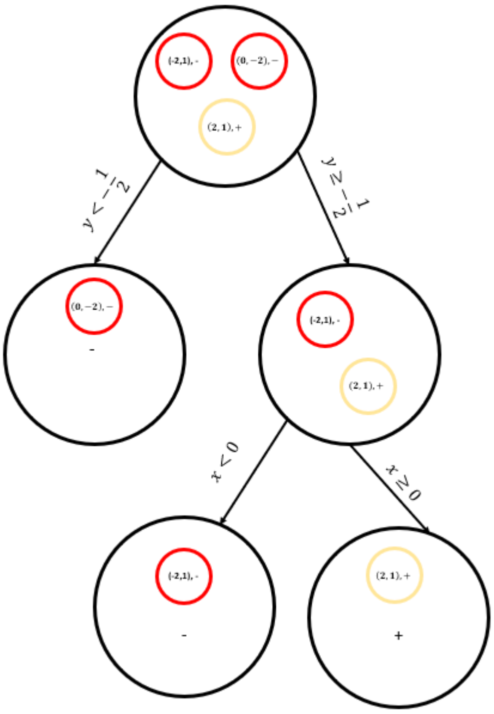


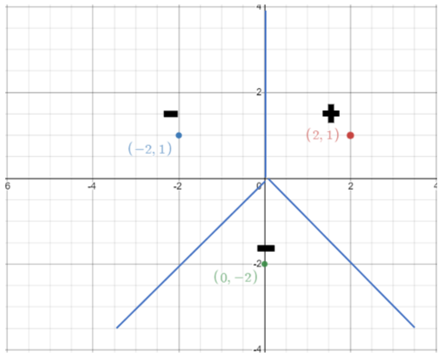


עבור דוגמת המבחן $(-1000,2),+$ לכל $k \geq 1$ התווית הדומיננטית של דוגמאות אימון בקבוצת k השכנים הקרובים ביותר תהיה של דוגמאות המורות על תוית '-' – סיווג שגוי.

נשים לב כי $ID3$ הניב את מסווג המטרה.

ב. נבחר $D=\{\left((-2,1),-\right),\left((0,-2),-\right),\left((2,1),+\right)\}$. נגדיר $A=\{(x,y)|x>0,y>-x\}$ ואנו מסתכלים על מסווג שקובע אילו נקודות שייכות ל-A. מתקבלים המסווגים הבאים -

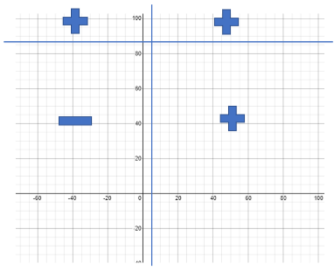




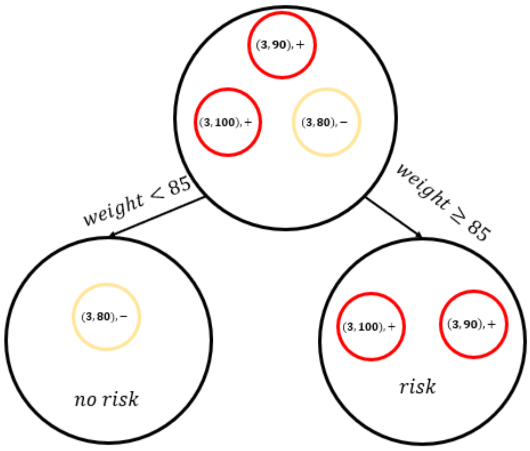
נשים לב כי על פי הגדרת הקבוצה A והגדרת KNN, עבור $k=1$, לכל דוגמת מבחן יתקיים כי מסווג KNN הינו מסווג המטרה.

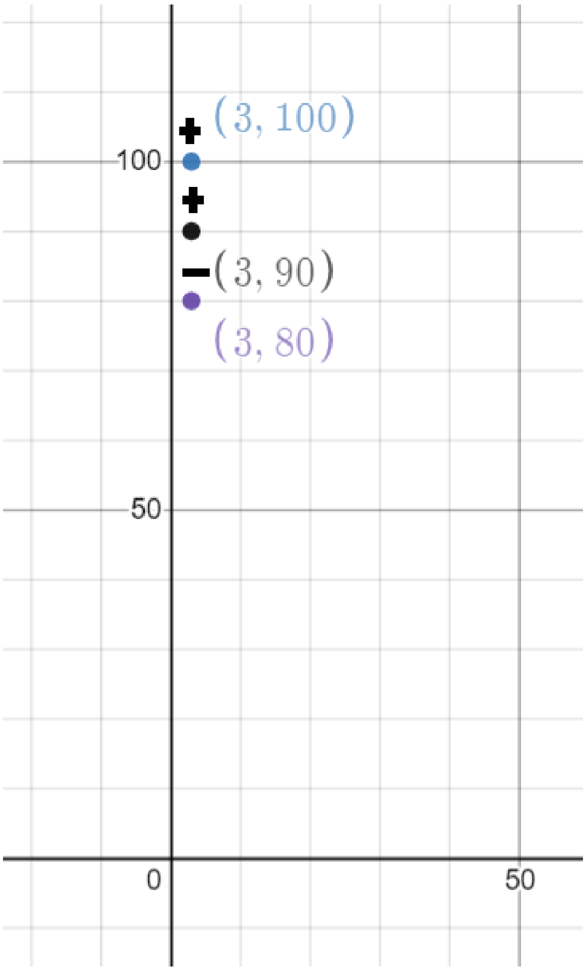
מנגד, עבור דוגמת המבחן – $\left(0.1,-\frac{1}{2}\right)$ שמין הסתם אינה שייכת ל-A מסווג $ID3$ יחזיר כי הנקודה שייכת לקבוצה. לכן קיימת לפחות דוגמת מבחן אחת עבורה $ID3$ טועה.

ג. לצורך הדוגמה הבאה נניח שאדם בסיכון ללקות בהתקף לב אם אחוז השומן שלו בדם הוא מעל 5 **או** אם הוא שוקל מעל 85. נסתכל על מסווגים מהצורה $(fat_{percent}, weight) \to \{+,-\}$, כאשר + מציין שיש סיכוי ללקות בהתקף לב ו- - אין. עבור קבוצת האימון -$D = \{((3,80),-),((3,90),+),((3,100),+)\}$ מסווג המטרה הוא :



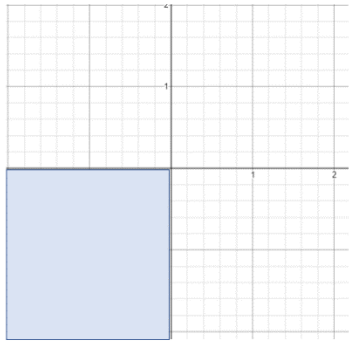
ונקבל את המסווגים הבאים -



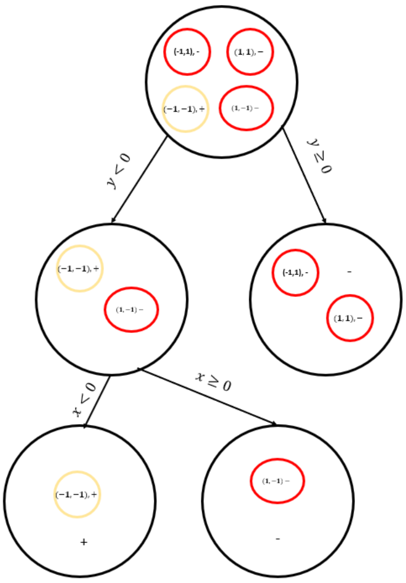


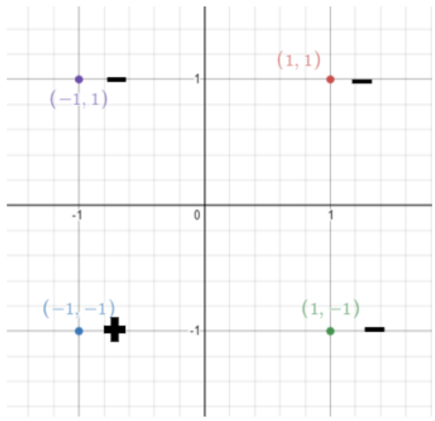
ששניהם מסווגים לא טוב את דוגמת המבחן $(10,80), +$.

ד. נסתכל על מסווג מטרה מהצורה-



המסווג כחיוביות את כל הדוגמאות ברביע הרביעי (עם $x < 0 \ \&\& \ y < 0$).
עבור קבוצת הדוגמאות $D = \{((-1,-1),+),((1,1),-),((1,-1),-),((-1,1),-)\}$, נקבל את המסווגים -





ולכן **קיים** $\boldsymbol{k = 1}$ כך שלכל דוגמת מבחן
המסווג יחזיר את התשובה הנכונה (ישירות מהגדרת מרחק אוקלידי).
המסווג ID3 שנקבל גם הוא מסווג את כל הנקודות נכונה (ניתן לראות לפי החלוקה לתחומים).

*הערה – בכל השאלות בניית המסווגים נעשתה לפי הגדרות החלק הרטוב וההבהרות בחלק היבש.

6. בבחירת הפרמטרים לאלגוריתם ביצעתי ניסוי דו שלבי -
* בשלב הראשון, היו לי 3 לולאות מקוננות עבור ערכי n, p, k עם טווח ערכים גדול אך קפיצות די גדולות.
עבור $n = range(10,100,10), p = [0.3, 0.4, 0.5, 0.6, 0.7], k = range(3, 99, 15)$. למעשה איתחלתי את העץ עם הפרמטרים הללו כל פעם ובדקתי את אחוז הדיוק שלו על קבוצת המבחן לאחר שהתאמן על קבוצת המבחן. ניסוי זה נתן לי תמונה ראשונית של **טווח** הערכים בהם צריכים להימצא הפרמטרים.
*בשלב השני, עבור טווחי הערכים שקיבלתי, בחנתי ערכים בדידים עבור n, k בתחום זה, וביצעתי בהתאם כיוונון לפרמטר p.
בסופו של דבר, בחרתי את הפרמטרים $n = 60, p = 0.69, k = 51$ ואני מקבלת אחוז דיוק על קבוצת המבחן את הדיוקים הבאים (לא דטרמיניסטי בגלל הבחירה האקראית של הדוגמאות)-

```
0.9911504424778761    0.9734513274336283    0.9823008849557522
```

עבור $k{\sim}50$ קיבלתי את אחוזי הדיוק המקסימליים בקומבינציות שבחרתי, והשיקול שלי בבחירתו להיות 51 היא שיהיה אי זוגי כדי שתהיה קביעה משמעית לגבי $majority\ class$.
הדיוק המקסימלי הינו 0.9915.
חשוב לציין שפרמטרים אלו כוונו ידנית בהתאם לסט הדאטה שקיבלנו. כדי למצוא פרמטרים מיטביים למקרה הכללי יש להשתמש בניסויים על סט דאטה גדול יותר.

7. המימוש המשופר שלי מתייחס למספר גורמים :

a. <u>דיוק העץ –</u> כאשר אני בוחנת את $p \cdot n$ הדוגמאות, אני אקח את $(1-p) \cdot n$ הדוגמאות הנותרות ואבחן את המסווג שבניתי עליהן. כעת עבור כל מסווג יש לי דיוק עבור קבוצת המבחן שיצרתי. בבחירת k העצים, אבחר את העצים על פי מיון שלוקח בחשבון הן את הקרבה אליי במרחב ה-n ממדי והן את אחוז הדיוק. כעת הבחירה שלי לוקחת בחשבון גם פרמטר של דיוק של המסווג.

b. <u>תכונות רלוונטיות –</u> בחישוב המרחק האוקלידי בין וקטור ה- $centroid$ לבין הדוגמא, עבור תכונה שהינה רלוונטית בבניית העץ, אני אמשקל את המרחק בין הסנטרואיד לדוגמה במימד זה כך שיהיה דומיננטי יותר מאשר במימד בו התכונה לא רלוונטית (בגלל שמחפשים מרחק מינימלי בין דוגמאות, חילקתי את המרחק במימד זה במספר חיובי גדול מ-1). באופן זה, אני בוחרת עצים שהם קרובים ביותר לדוגמה וכן בעלי קרבה רלוונטית (אינטואיציה – אם זהו מסווג לקבלה לטכניון, ואחת התכונות היא מספר עוקבים באינסטגרם, כנראה שהעץ שלי לא יתחשב בה במהלך האימון ולכן לא ארצה להשתמש בה באומדן של קרבה בין דוגמאות).

c. <u>נרמול ערכי התכונות -</u>האלגוריתם שלי משתמש בנירמול $minmax$ על קבוצת האימון וכן על קבוצת המבחן כדי לתת משקל שווה במרחק האוקלידי לכל התכונות.

d. <u>משקול knn -</u>בחישוב התווית הדומיננטי בקרב ה-k הקרובים ביותר, ניתן משקל גדול יותר להחלטה של עץ קרוב יותר, מתוך הבנה שה"דעה" שלו חשובה יותר מה"דעה" של מישהו רחוק יותר. בלולאה שלי על k העצים הקרובים ביותר, כאשר הם ממויינים לפי קרבה, נתתי משקל של $k_{param} - i + 1$ לתוית של העץ ה-i.

e. <u>בחירת פרמטר m –</u> בניסויים שביצעתי, פרמטר m האידיאלי ליער המשופר הינו $m = 10$.

כעת קיבלתי את אחוזי הדיוק הבאים, כאשר ניתן לראות כי מדי פעם המסווג מגיע לאחוזי דיוק של 1.0 :

```
0.9734513274336283    0.9823008849557522    0.9911504424778761    1.0
```

* כדי לבחון שיפור אמיתי ולנטרל את פרמטר האי דטרמינסטיות של בחירת הדוגמאות האקראיות הרצתי ניסוי עם יצירת יער רגיל ויער משופר על אותן דוגמאות, ובב- 75% מהמקרים היער המשופר היה עדיף.

* בנוסף , בהרצת 20 ניסויים של היער הרגיל אחוז הדיוק הממוצע היה

```
avg accuracy is: 0.9792035398230088
```

ואילו בהרצת 20 ניסויים של היער המשופר ממוצע הדיוק היה

```
avg accuracy after 20 runs is: 0.9893805309734514
```

ובממוצע בהרצת 20 ניסויים על המשופר, על שלושה מקבלים אחוז דיוק של 1.0 .