עבודת בית 3 – מבוא לבינה מלאכותית - 236501

אלעד אספיס

שאלה 1.2

תוצאות הדיוק שקיבלתי הן: 0.9469026548672567

שאלה 2 – הטענה נכונה

לפי הגדרה, פונקצית נירמול $MinMax$ מוגדרת באופן הבא: $x_{chang} = \frac{x-x_{min}}{x_{max}-x_{min}}$ ונסמנה ב-$\phi(x)$.

ראשית, נשים לב כי עבור $a < b$ מתקיים: $\phi(a) = \frac{a-x_{min}}{x_{max}-x_{min}} < \frac{b-x_{min}}{x_{max}-x_{min}} = \phi(b)$, דהיינו $\phi(x)$ הינה פונקציה אשר שומרת על הסדר.

כעת, ניזכר בכך שהדיוק של מסווג $ID3$ תלוי באופן ישיר בבחירת הסף $(threshold)$ בכך שאנו בוחרים תכונה עם תוספת האינפורמציה הגדולה ביותר. ניזכר כי כאשר אנו נמצאים בצומת E ומפצלים לפי תכונה f תוספת האינפורמציה מוגדרת ע"י $IG(f,E) = Entopy(E) - \frac{|E_1|}{|E|}Entropy(E_1) - \frac{|E_2|}{|E|}Entropy(E_2)$.

ובכן, בשלב קביעת הסף $(threshold)$ בהינתן $feature$ כלשהו ובהינתן סדרת ערכים $x_0 \le x_1 \le \cdots \le x_n$ המיוחסים לאותו ה-$feature$, לאחר הפעלת פונקציית הנרמול נקבל כי מתקיים $\phi(x_0) \le \phi(x_1) \le \cdots \le \phi(x_n)$ (ϕ שומרת סדר). הסף הנקבע בכל שלב הינו הממוצע בין $\phi(x_i)$ לבין $\phi(x_{i+1})$ ולכן לאחר הפעלת פונקציית הנרמול נקבל כי פיצול הקבוצות לפי הסף יהיה זהה. מכיוון שהטענה לעיל תקפה לכל צומת בעץ, נקבל כי תוספת האינפורציה לא תשתנה בכל צומת בעץ ההחלטות שאנו מפתחים.

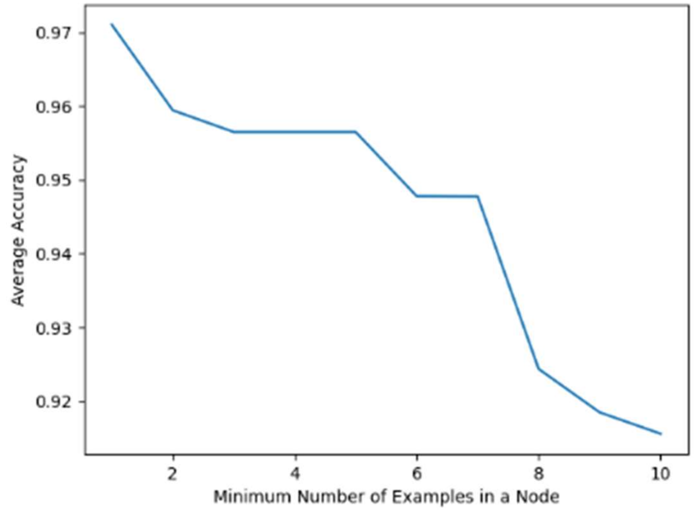
על-כן, ניתן לומר כי ההחלטות השונות שמסווג $ID3$ הנלמד על קבוצת האימון והנבחן על קבוצת המבחן אינן משתנות ולכן הדיוק אינו משתנה גם כן.

שאלה 3.1

חשיבות הגיזום היא גבוהה. הגיזום נעשה על מנת למנוע התאמת יתר (overfitting) בתוצאות המבחן. ניזכר כי עץ ההחלטות שאנו מפתחים הוא עקבי ועל-כן עבור דוגמאות רועשות נקבל עץ החלטות שגוי ולכן במקרה זה גיזום של העץ יפחית את תופעת ה-$overfitting$.

שאלה 3.3

iii.



נשים לב כי בניסוי שביצעתי ככל שאנו מגדילים את m הדיוק קטן.

כמו כן, נבחין בכך שאכן תוצאות הגרף הגיוניות, שכן ככל שאנו מגדילים את m העץ נעשה פחות עקבי (המקרה האידיאלי הוא כאשר לא קיים גיזום כלל על פי הגרף).

iv. קיבלנו שהדיוק טוב עבור הערכים m=1,2, דהיינו שלא מתבצע גיזום כלל.

שאלה 3.4

הגיזום לא שיפר את הביצועים. הדיוק שקיבלתי הוא 0.9469026548672567

שאלה 4.1

הפעלתי את האלגוריתם ID3 עם גיזום מוקדם עבור m=1,2 וקיבלתי שערך ה-loss הוא 0.021238938053097345

הגיזום לא שיפר את הביצועים ביחס להרצה.

שאלה 4.2

ראשית, נשים לב כי לפי השאלה סיווג אדם חולה כבריא חמורה פי 10 כסיווג אדם בריא כחולה.

בשלב האימון נבצע $K - Fold\ cross\ validation$ על קבוצת האימון בלבד עם $k = 5$. מכיוון שברצוננו להקטין את ה-$loss$, נשמור את העץ בעל ה-$loss$ המינימלי, עליו נפעיל את פונקציית ה-$predict$ על קבוצת המבחן.

שיפור כללי:

בכל צומת הוספתי סף ($threshold$) נוסף באופן הבא: ראשית, חישבתי את הסף כפי שמחושב ב-$ID3$ (לצורך הנוחות נסמן $threhold_1 < threshold_2$ (יוסבר בהמשך)). כתוצאה מכך, יש לנו את הדוגמאות שערך התכונה (לפי מבצעים את הפיצול) גדול מהסף (נסמנם ב-$right - examples$) ויש את הדוגמאות שערך התכונה קטן מהסף (נסמנם ב-$left - examples$). כעת, אבחר את הקבוצה בעלת האנטרופיה הגדולה ביותר (דהיינו הקבוצה עם אי הוודאות הגדול ביותר) להיות הקבוצה התלויה בבחירת הסף השני. באופן זה, ע"י חלוקה של 3 קבוצות בכל צומת נוכל לסווג באופן יותר איכותי כל קבוצת מבחן אשר תהיה. בנוסף לכך, שלב הפרדיקציה מתבצע באופן הבא: נסמן את ערך הדוגמא לפי $Feature$ מסוים ב-x. כמו כן, נסמן $\varepsilon_i = 0.1 * threshold_i$ עבור $i = 1,2$.

נחלק למקרים:

1. במקרה בו יש שני בנים (בפרט קיים סף אחד) : אם x קרוב ל-$threshold$ עד כדי ε אז נכנס לשני המסלולים עד שנגיע לעלה וניקח את הסיווג הדומיננטי משני הערכים שהוחזרו. אחרת, במידה ו-x קטן מהסף נמשיך למסלול השמאלי, אחרת נכנס למסלול הימני.
2. במקרה בו יש שלושה בנים (במקרה זה קיימים שני ספים - $threshold_1 + threshold_2$) נפעל באופן דומה: אם x קרוב ל-$threshold_1$ עד כדי ε_1 נכנס למסלול האמצעי והשמאלי ובמידה וגם x קרוב ל-$threshold_2$ עד כדי ε_2 נכנס למסלול הימני ונחזיר את הסיווג הדומיננטי. אחרת:
    a. אם $x < threshold_1$ נחזיר את הסיווג של המסלול השמאלי.
    b. אם $threshold_1 \leq x < threshold_2$ נחזיר את הסיווג של המסלול האמצעי.
    c. אחרת (במקרה זה $threshold_2 \leq x$) נחזיר את הסיווג של המסלול הימני.

שאלה 4.3

ה-$loss$ שקיבלתי הוא `0.0008849557522123895`

הרצתי ניסויים על קביעת ה-ε (באמצעות $np.linspace(0.5,1.5, num = 100)$) על-מנת לקבל את ה-$loss$ המינימלי וקיבלתי עבור $\varepsilon = 0.07$ את ה-$loss$ המצוין לעיל.

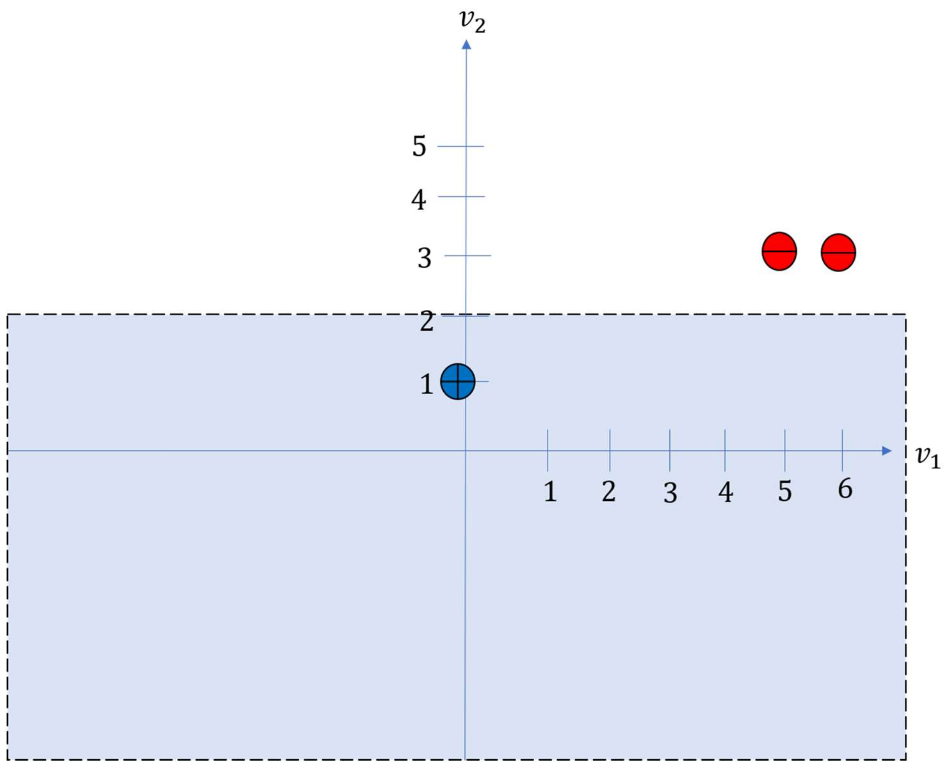
שאלה 5

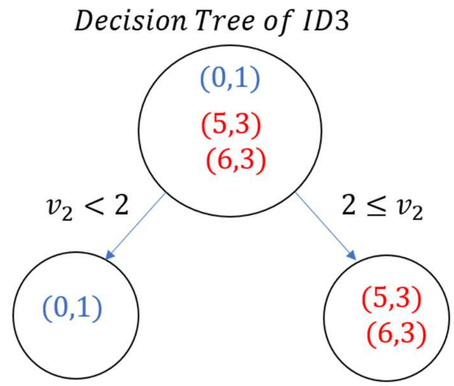
<u>סעיף א'</u>

ניקח $D=\{((0,1),+),((5,3),-),((6,3),-)\}$

נשים לב כי למידת עץ $ID3$ תניב מסווג אשר עונה נכון עבור כל דוגמת מבחן אפשרית, בעוד אם ניקח את הדוגמא החיובית $(4,3)$ למידת KNN תניב כי דוגמא זו היא שלילית כי עבור כל K מספר השכנים השליליים גדול ממספר השכנים החיוביים עבור הנקודה $(4,5)$.

$$f(v_1,v_2)=\begin{cases}0 & v_2<2\\ 1 & 2\le v_2\end{cases}$$





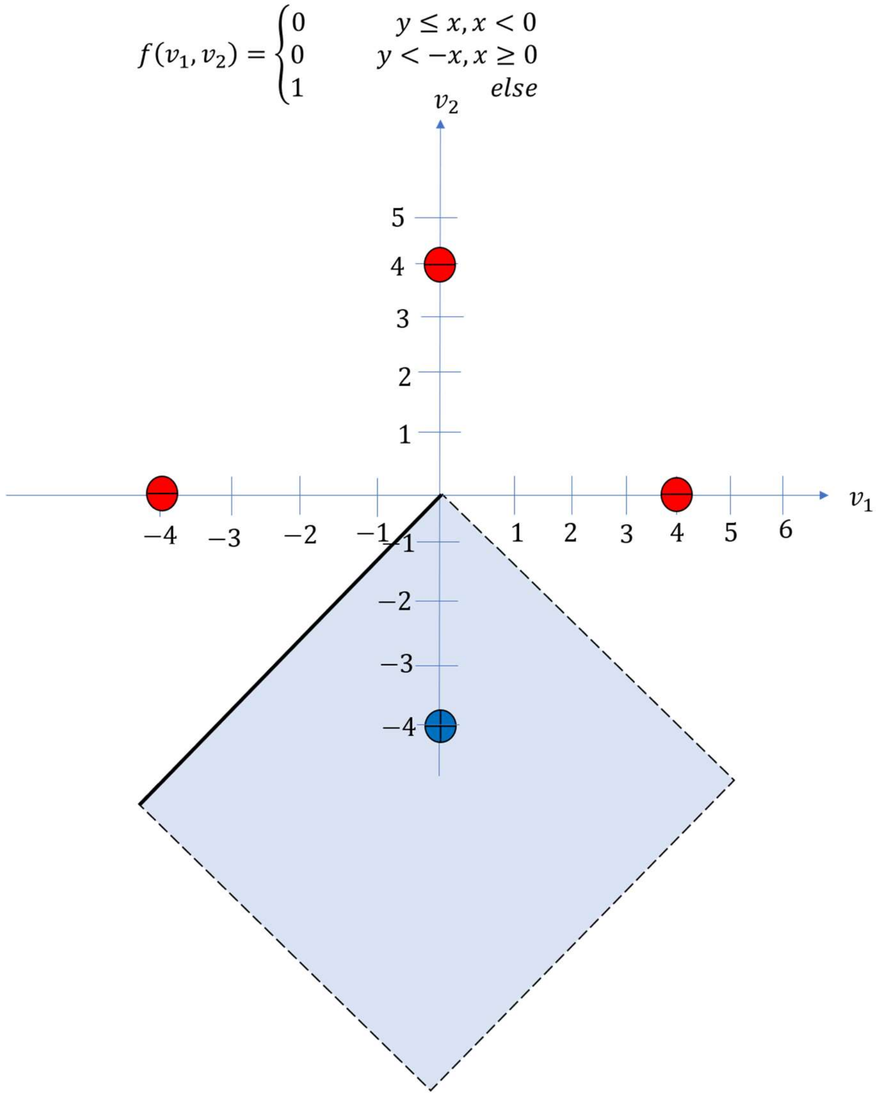
<u>סעיף ב'</u>

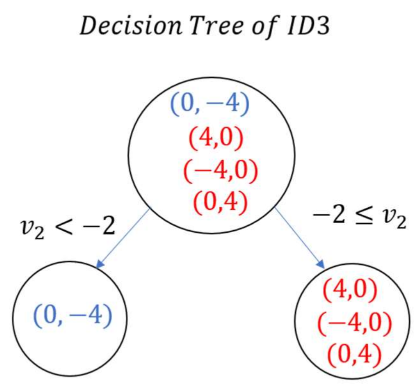
ניקח $D = \{((0,4),+),((4,0),-),((-4,0),-),((0,4)-)\}$

נשים לב כי למידת עץ $ID3$ תסווג את הדוגמא $(-3,-1)$ כדוגמא חיובית, בניגוד למסווג המטרה, ולכן יטעה.

כמו כן, עבור $K = 1$ למידת מסווג KNN תניב מסווג אשר עונה נכון בכל דוגמאת מבחן אפשרית. נשים לב כי עבור כאשר $x \geq 0$ עבור הנקודות $y = -x$ (אשר מרחקן זהה לנקודות $(4,0)$ ו-$(0,-4)$) נקבל כי הסיווג שייבחר הוא סיווג שלילי (בדומה לסיווג המטרה), מכיוון שיש העדפה לבחור ב-v_1 גדול יותר. כמו כן, כאשר $x < 0$ עבור הנקודות $y = x$ (אשר מרחקן זהה לנקודות $(-4,0)$ ו-$(0,-4)$), נקבל כי הסיווג הוא חיובי (בדומה לסיווג המטרה), מכיוון שיש העדפה לבחור בערך v_1 גדול יותר. נשים לב כי בראשית (בנקודה $(0,0)$ הסיווג שייבחר הוא שלילי (בדומה לסיווג המטרה), מכיוון שיש העדפה לבחור ב-v_1 גדול יותר.

עבור שאר הנקודות, ניתן לראות כי הסיווג של KNN אכן יהיה דומה לסיווג המטרה, כנדרש.





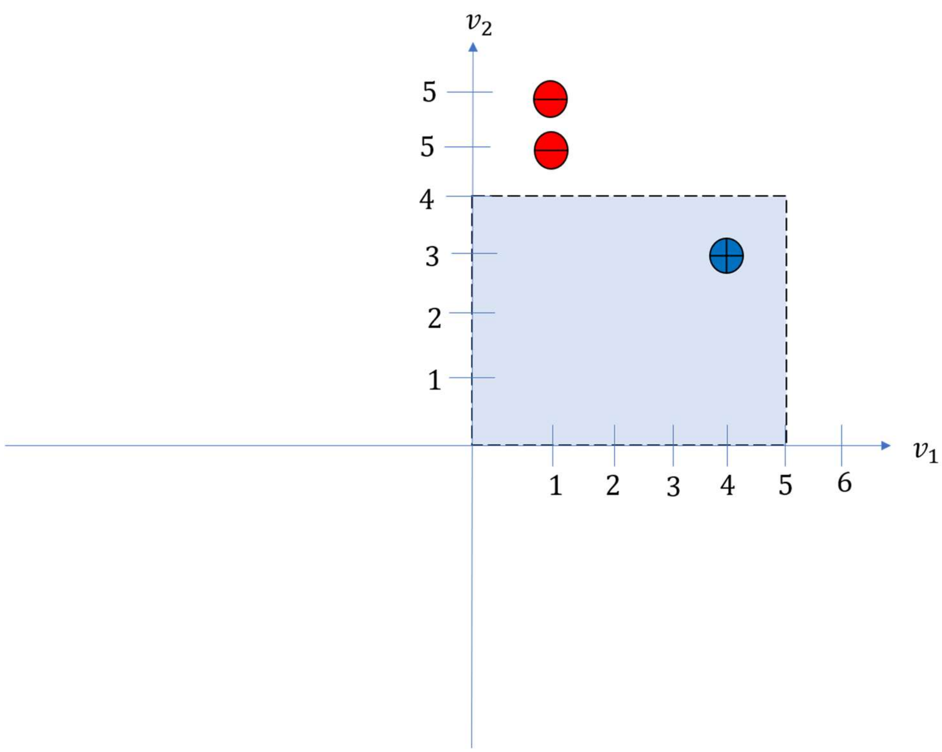
<u>סעיף ג'</u>

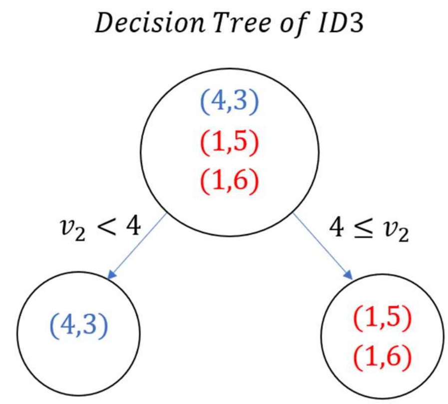
ניקח $D = \{((3,4),+),((5,1),-),((6,1)-)\}$

למידת עץ $ID3$ תניב מסווג אשר לדוגמא עבור הדוגמא השלילית $(-1,-1)$ נקבל כי הסיווג הינו חיובי מכיוון ש- $-1 < 4$.

למידת מסווג מסוג KNN תניב מסווג אשר לדוגמא עבור הדוגמא חיובית $(1,3.5)$ הסיווג שנקבל יהיה שלילי עבור כל ערך K.

$$f(v_1, v_2) = \begin{cases} 0 & 0 < v_1 < 4, 0 < v_2 \leq 5 \\ 1 & else \end{cases}$$





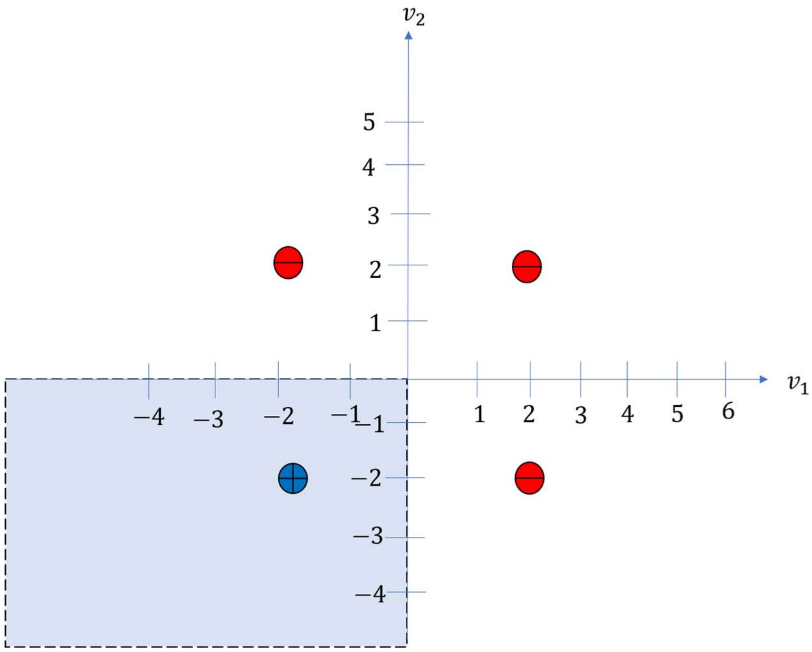
סעיף ד'

ניקח $D = \{((-2,-2),+),((-2,2),-),((2,-2)-),((2,2)-)\}$

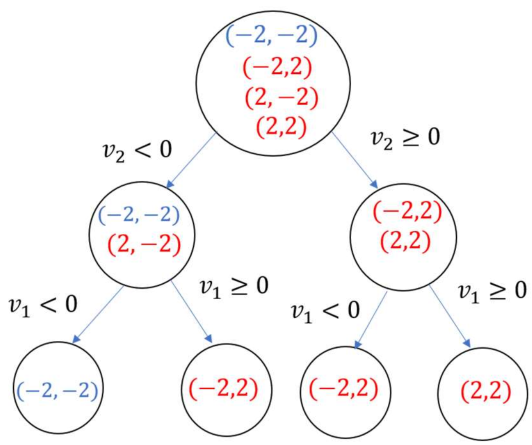
נשים לב כי למידת עץ $ID3$ תניב מסווג אשר עונה נכון עבור כל דוגמת מבחן אפשרית.

בלמידת מסווג KNN עבור $K = 1$ מכיוון שהתחום החיובי הוא פתוח ואיננו כולל את השפה נקבל כי הוא עונה נכון, כנדרש.

$$f(v_1, v_2) = \begin{cases} 0 & v_1 < 0, v_2 < 0 \\ 1 & else \end{cases}$$



Decision Tree of ID3



שאלה 6.1

ביצעתי ניסויים עבור פרמטרים טובים: לקחתי 3 רשימות $N = [50,60,70]$ ובנוסף $K = [31,41,51]$, $P = [0.3,0.4,0.5,0.6,0.7]$ וקיבלתי כי עבור $N = 70, K = 41$, $p = 0.7$ תוצאות הדיוק הן הכי גבוהות. על-כן, לקחתי את ערכים אלו עבור הרצת התוכנית.

תוצאות הדיוק שקיבלתי הן `0.9823008849557522` `0.9734513274336283`

שאלה 7.1

השיפורים שביצעתי הם:

- בכל דוגמת מבחן עבורה קיבלתי k עצים שה-$centroid$ הכי קרובים לדוגמת המבחן אני מבצע משקול ערכים: דהיינו, העץ עבורו ה-$centroid$ שלו הכי קרוב לדוגמת המבחן מקבל את המשקל הגבוה ביותר. העץ השני עבורו ה-$centroid$ שלו הוא השני הכי קרוב לדוגמת המבחן מקבל את המשקל השני בערכו.
  באופן כללי, העץ ה-i עבורו ה-$centroid$ שלו הוא ה-i הכי קרוב לדוגמת המבחן מקבל את המשקל ה-i בערכו (עבור $1 \leq i \leq k$). האינטואיציה פה היא שככל שה-$centroid$ קרוב יותר לדוגמת המבחן רצוי לקבל משקל גבוה יותר כי תוצאות קרובות לדוגמת המבחן ככל הנראה דומה לסיווג דוגמת המבחן, ויש מן ההגיון שעצים שה-$centroid$ שלהם רחוק מן דוגמת המבחן יקבלו משקל נמוך כי אינם בהכרח דומים לדוגמאת המבחן (כי תוצאות אשר רחוקות לדוגמת המבחן ככל הנראה אינן קשורות לסיווג דוגמת המבחן ולכן נמשקל אותן בהתאם – במשקל נמוך יותר).
- הגדלת מספר הספים ל-2 ספים (במקום סף אחד) כפי שביצעתי בסעיף 4.2. האינטואיציה כאן היא למעשה להקטין את חוסר הוודאות בכל צומת ע"י הרחבת הטווחים. כמו כן, הגדרתי שדוגמאות אשר קרובות "מאוד" (ע"י פרמטר שקבעתי כפי שתואר בסעיף 4.2). האינטואיציה כאן היא להתחשב בדוגמאות אשר קרובות מאוד – הדבר דומה לבן אדם בן 89. הוא איננו בן 90 אבל למעשה הוא קרוב מאוד לגיל 90 ולכן נרצה להתחשב בתוצאות מסוג זה.
- העדפה לתכונות רלוונטיות – בז מן חישוב ה-$centroid$ קבעתי כי יהיה משקל גבוה לתכונות שבהם עץ ההחלטות פיצל לפיו מאשר תכונות שלא השתמש בו בכלל (אשר יקבל משקל נמוך יותר). במימוש הגדרתי שתכונה אשר עץ ההחלטות השתמש תקבל משקל של 2, בעוד משקלן של תכונות שעץ ההחלטות לא השתמש יהיה 1.
- כמו כן, ביצעתי ניסוי על הפרמטר m (פרמטר הגיזום) וקיבלתי כי עבור $m = 3$ הדיוק הוא הגבוה ביותר.

שאלה 7.2

הניסוי שביצעתי הוא שינוי הערכים N, p, K. לקחתי 2 רשימות $N = [50,60,70]$ ובנוסף $K = [31,41,51]$. בנוסף, כל פעם שיניתי את הערך p.קיבלתי כי עבור $N = 60, K = 31$ הדיוק הוא הכי גבוה. כמו כן, ביצעתי ניסויים עבור ערך האפסילון. ביצעתי שינוי ערכים ועבור $\varepsilon = 0.1$ קיבלתי את הדיוק המיטבי. בנוסף, ביצעתי ניסויים על פרמטר הגיזום ועבור $m = 1$ הדיוק גבוה.

הדיוקים שקיבלתי הם `0.9911504424778761` `0.9823008849557522`

לדוגמא, עבור $p = 0.7$ קיבלתי שהדיוקים שקיבלתי הם:

```
k:  21 , N:  50
0.9911504424778761
k:  21 , N:  60
0.9823008849557522
k:  21 , N:  70
0.9911504424778761
k:  31 , N:  50
0.9911504424778761
k:  31 , N:  60
0.9911504424778761
k:  31 , N:  70
0.9823008849557522
k:  41 , N:  50
0.9823008849557522
k:  41 , N:  60
0.9823008849557522
k:  41 , N:  70
0.9823008849557522
```

כך, למעשה כל פעם שיניתי את הערך p על-מנת להגיע לדיוק מקסימלי.