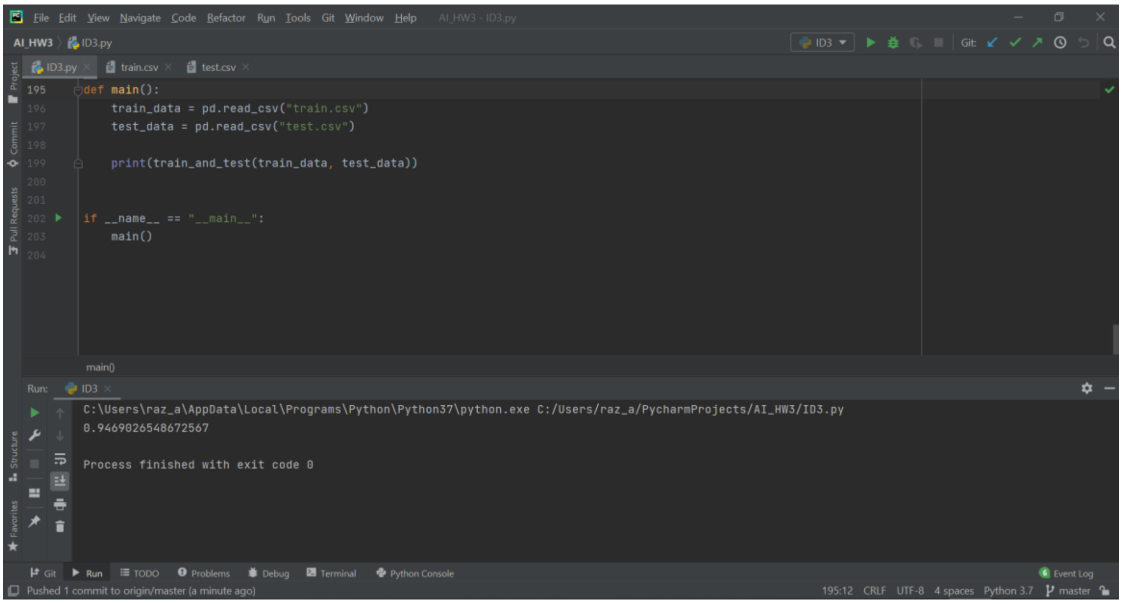
תרגיל בית 3 – חלק יבש

שאלה 1

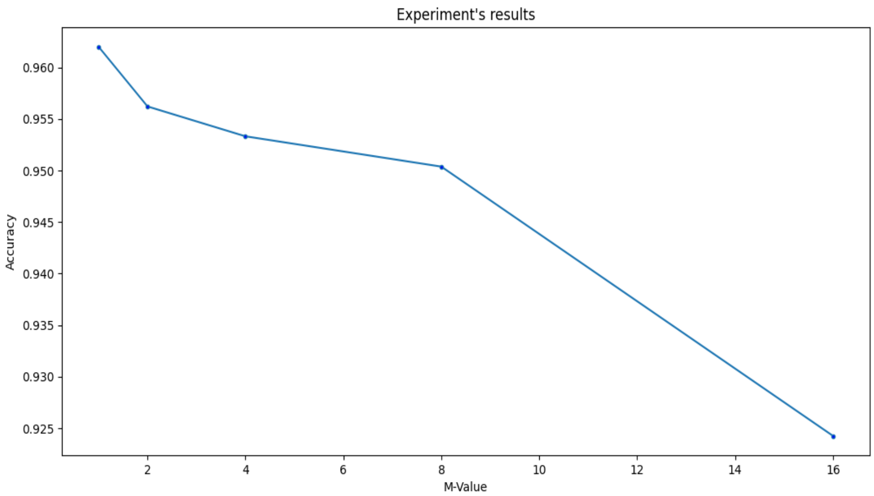
להלן צילום מסך של תוצאת הדיוק (התוצאה היא 0.9469~):



שאלה 3

1. הגיזום באופן כללי נועד כדי להקטין את העצים, בין אם זה לפני או אחרי בנייתם, במטרה להחליש את אפקט התאמת היתר (overfitting), שזו תופעה שאנו מנסים למנוע באופן זה שנשאף להגדיל את שגיאת האימון בשביל להקטין את שגיאת המבחן.

3. להלן הגרף המציג את השפעת הפרמטר M על הדיוק עבור M=1,2,4,8,16:



כפי שניתן לראות, הגרף נמצא במגמת ירידה, מה שמצביע על כך שככל שנגדיל את הפרמטר M כך נקבל דיוק נמוך יותר. זה הגיוני מכיוון שעבור M גדול יותר האלגוריתם אמור לגזום יותר תתי עצים ובעקבות זה יגדיל את שגיאת האימון. כפי שניתן לראות, הM האופטימלי הוא 1 עם דיוק של 0.962~, שזה הדיוק הגבוה ביותר עם שגיאת האימון הנמוכה ביותר. עבור M=1 אין גיזום כי עבור דוגמא אחת תמיד נקבל צומת שלפי האלגוריתם יחזיר מיד את הסיווג שלו.

4. עבור הערך האופטימלי M=1 קיבלנו בדיוק את אותה התוצאה משאלה 1.

<u>שאלה 4</u>

1. מימשתי את הפונקציה ()get_loss אשר מחזירה את ערך ה-loss של ID3, וערך ה-loss שיצא מהרצת הפונקציה עם M=1 לפי סעיף 3.4 הינו 0.021238938~.

2. מכיוון שזה מאוד חמור שאדם חולה יסווג כבריא מבחינת ה-loss, אז חשבתי על לנסות למזער את ההפרש בין מספר הדוגמאות של אנשים בריאים לעומת אנשים חולים בכל צומת כאשר יש יותר דוגמאות של בריאים מחולים, וכך נוריד את ההסתברות לסיווג אדם חולה כבריא. ניתן לממש זאת ע"י גיזום מוקדם עם היפר-פרמטר N שמוגדר להיות מספר מינימלי של הפרש דוגמאות בין בריאים לחולים, והאלגוריתם המשופר יגזום רק כאשר יש יותר דוגמאות של בריאים מדוגמאות של חולים.

3. לאחר מימוש דרך זו, ביצעתי ניסויים לקביעת פרמטר N האופטימלי מN=1 ועד 100, ואלו התוצאות שקיבלתי:

loss בקירוב	N
0.021238938	1-2
0.022123893	3-7
0.004424779	8-9
0.002654867	>=10

לכן עבור N=10 נקבל את האלגוריתם המשופר עם ערך אופטימלי loss=0.002654867, שהוא נמוך משמעותית מזה שקיבלנו בסעיף 1.

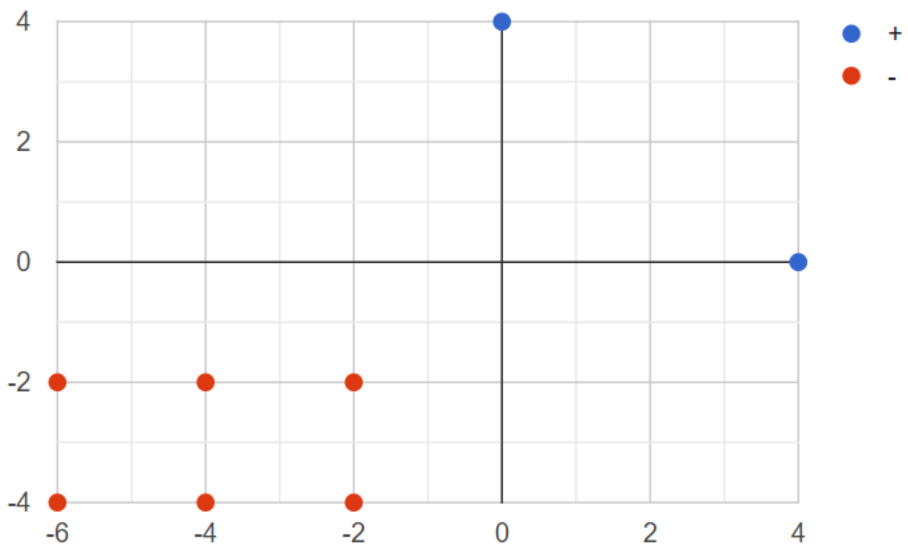
<u>שאלה 5</u>

<u>סעיף א</u>

מסווג המטרה שבחרתי: $f\left(v_{1,i},v_{2,i}\right)=\begin{cases}1 & v_{2,i}\geq -1\\ 0 & v_{2,i}<-1\end{cases}$

קבוצת האימון שבחרתי: $D=\left\{\begin{matrix}((-6,-4),0),((-4,-4),0),((-2,-4),0),((-6,-2),0),\\ ((-4,-2),0),((-2,-2),0),((4,0),1),((0,4),1)\end{matrix}\right\}$

תיאור גרפי:



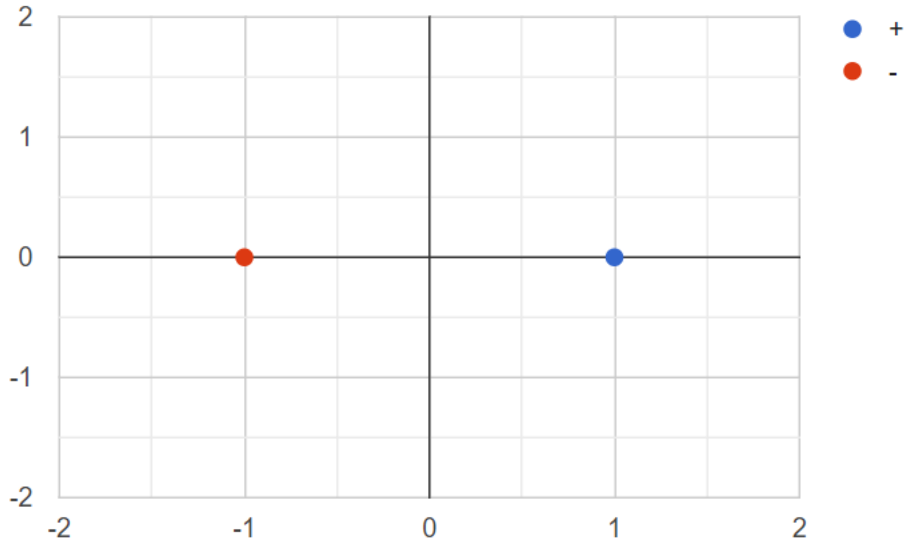
עבור למידת עץ ID3 נקבל שערך הסף שממקסם את הIG הוא $v_2=-1$ ולכן העץ יכיל רק שורש ושני בנים שמפריד באופן טהור לחלוטין את קבוצת הדוגמאות ולכן נקבל את מסווג המטרה.
עבור למידת KNN, דוגמה אשר תבחן בנקודה $(-6,0)$ תחזיר "-" לכל K שנבחר (ניתן לראות זאת בברור לפי הגרף המצורף כי הדוגמאות "+" הרבה יותר רחוקות מהדוגמה הנבחנת, יחסית לנקודות ה"-") וזאת לא בהתאם למסווג מטרה שהצגתי כי היה אמור להחזיר "+".

<u>סעיף ג</u>

מסווג המטרה שבחרתי: $f(v_{1,i},v_{2,i})=\begin{cases}1 & v_{1,i}=1\wedge v_{2,i}=0\\ 0 & \text{else}\end{cases}$

קבוצת האימון שבחרתי: $D=\{((1,0),1),((-1,0),0)\}$

תיאור גרפי:



עבור למידת מסווג KNN עבור K=1 ולמידת עץ ID3, דוגמת מבחן עם $x=(2,0)$ נקבל עבורה $y=0$ כלומר האלגוריתמים יסווגו זאת "-" בניגוד למסווג מטרה שהצגתי שיסווג זאת בתור $y=1$.

לגביי KNN, לפי פונק' מרחב אוקלידי ברור כי $(2,0)$ יותר קרובה ל$(1,0)$ מאשר $(-1,0)$ ולכן יסווג זאת כמו $(1,0)$, כלומר בתור "+" במקום "-".

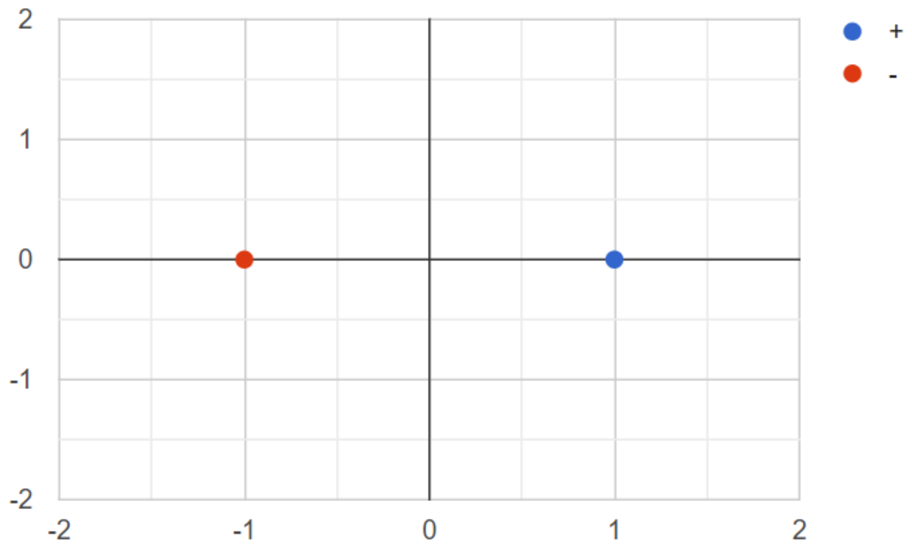
לגביי ID3, התכונה בעלת הIG המקסימלי שיבחר היא v_1 כי רק היא מפרידה את שתי הדוגמאות בקבוצת אימון עם ערך סף 0 (v_2 זהה בשתי הדוגמאות ולכן לא יופרדו). אז כל דוגמה שנבחן תסווג כ"+" כאשר היא גדולה-שווה מערך הסף, ובפרט תסווג את דוגמת המבחן כ"+", בניגוד למסווג שהצגתי שעבור דוגמת המבחן אמור להחזיר "-".

<u>סעיף ד</u>

מסווג המטרה שבחרתי: $f(v_{1,i},v_{2,i})=\begin{cases}1 & v_{1,i}\geq 0\\ 0 & v_{1,i}<0\end{cases}$

קבוצת האימון שבחרתי: $D=\{((1,0),1),((-1,0),0)\}$

תיאור גרפי:



למידת מסווג KNN עבור K=1 ולמידת עץ ID3 מניבות את המסווג הנ"ל אשר עונה נכון על כל דוגמת מבחן אפשרית.

לגביי KNN, ברור כי דוגמאות שעבורן $v_1>0$ יותר קרובות לדוגמה בקבוצת אימון שהיא "+" ולכן יסווגו גם כ"+", ודוגמאות שעוברן $v_1<0$ יותר קרובות לדוגמה בקבוצת אימון שהיא "-" ולכן יסווגו גם כ"-", וזאת בהתאם למסווג שבחרתי. לכל דוגמה שעבורה $x_i=(0,v_{2,i})$ נקבל שהמרחק יהיה זהה בינה לבין שתי הדוגמאות שבקבוצת האימון ולכן מתחשבים קודם בערך v_1 המקסימלי שהוא 1, כלומר ה"+". לכן נקבל $y_i=1$ בהתאם למסווג שהצגתי.

לגביי ID3, התכונה בעלת הIG המקסימלי שיבחר היא v_1 כי רק היא מפרידה את שתי הדוגמאות בקבוצת אימון עם ערך סף 0 (v_2 זהה בשתי הדוגמאות ולכן לא יופרדו). אז כל דוגמה שנבחן תסווג כ"+" כאשר היא גדולה-שווה מערך הסף, אחרת תסווג כ"-", וזאת בהתאם למסווג שהצגתי.

שאלה 6

ביצעתי ניסויים למציאת ערכי פרמטרים עבור דיוק מקסימלי באופן הבא:

לכל $p \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$, לכל $1 \le N \le 5$ ולכל $1 \le K \le N$ הרצתי את אלגוריתם הלמידה KNN-decision-tree. ערכי הפרמטרים שקיבלתי עבורם דיוק מעל 0.98 היו:

p	N	K	accuracy
0.3	4	3	0.9823008849557522
0.4	1	1	0.9823008849557522
0.5	1	1	0.9911504424778761
0.5	5	5	0.9911504424778761
0.7	4	4	0.9823008849557522

עבור ערכי הפרמטרים לעיל הרצתי שוב 10 פעמים עבור כל שורה וחישבתי את ממוצע הדיוק מתוך ה10 הרצות בכל שורה. הערכים עבורם קיבלתי את הממוצע הגבוה ביותר הם p=0.3 N=4 K=3 עם דיוק מקסימלי של 1 בהרצות הנוספות. כמו כן, בכל עץ השתמשתי באלגוריתם ID3 ללא גיזום כדי להגיע למצב של דיוק אופטימלי כפי שהוסבר בשאלה 3.

שאלה 7

1. ניתן לשפר את האלגוריתם בכך שבמקום לדגום באופן אקראי $p*n$ דוגמאות מקבוצת האימון, אנו נדגום $p*n$ דוגמאות בכל איטרציה כך שדוגמאות שנבחרו באיטרציה ה-$i \in [N]$ הן הדוגמאות שנבחרו הכי מעט פעמים מכל האיטרציות הקודמות ל-i. באופן זה נמנע מעצים שעלולים להיות דומים מידיי בהרצה מסוימת בגלל האקראיות שבבחירת הדוגמאות, וכך בעצם אנו משפרים את הדיוק של המקרה הכללי עבור בחירה כלשהי של פרמטרים. במימוש אנו נגדיל את קבוצת האימון שבגודל n פי $N \cdot p$ בצורה ציקלית (אם השורה האחרונה בקבוצה היא 343 אז השורה ה344 תהיה כמו שורה 1, והשורה ה345 תהיה כמו שורה 2 וכן הלאה) ואז העץ הראשון יכיל את $n \cdot p$ הדוגמאות הראשונות, והעץ השני יכיל את $n \cdot p$ הדוגמאות השניות וכך הלאה עד העץ ה-N. באופן זה יש מספיק דוגמאות לכל עץ וגם קיימים מספיק עצים שיחסית שונים אחד מהשני, ובחישוב ה-centroid עבור העצים השונים נקבל וקטורים שאינם דומים בסבירות גבוהה.
2. בדקתי את כל הפרמטרים מהשאלה הקודמת וקיבלתי עבור האפשרות p=0.3 N=5 K=5 ועבור מספר נוסף של אפשרויות את הדיוק הכי גבוה שהוא 0.97345~. אמנם היו מספר ערכי פרמטרים עבורם קיבלתי דיוק יותר גבוה בשאלה 6 אך אלו מקרים נדירים מתוך כמות הרצות גבוהה, כאשר במימוש הנוכחי קיבלתי דיוק שלא נופל מ0.92 עבור אף ערכי פרמטרים וללא אקראיות כלל, לעומת המימוש בשאלה 6 שבו הדיוק היה לעיתים פחות מ0.9, כלומר הייתה שונות גדולה שנבעה ככל הנראה מהאקראיות במימוש הרגיל ותוקן במימוש המשופר.