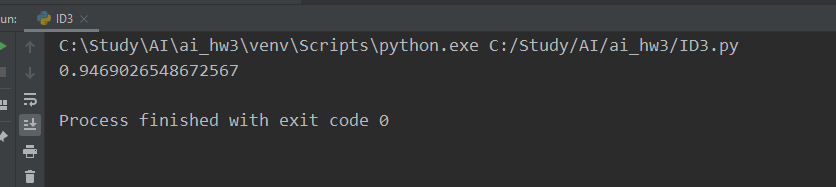
מבוא לבינה מלאכותית תרגיל בית 3

יובל גושן 205810179

# שאלה 1

1.2 התקבלה תוצאת דיוק של על קבוצת המבחן.  


# שאלה 2

הטענה נכונה. נראה זאת בשני שלבים:  
1. העץ שנבנה בשלב האימון מתפצל לפי תכונות זהות אבל עם ערכי threshold המתאימים לנרמול.

נראה שנרמול MinMax מונוטוני, כלומר משמר יחס סדר, נניח שיש 2 דוגמאות ופיצ'ר f עבור דוגמא 1: ועבור דוגמא 2: וערכי מקסימום עבור אותו פיצ'ר בכל סט הדוגמאות כך שמתקיים . נפעיל סדרה של פעולות מונוטוניות השומרות על אי השוויון :

לכן בכל פעם שנרצה לפצל עץ, ונבחר פיצ'ר וThreshold שהוא ממוצע בין כל 2 ערכים כפי שנלמד בהרצאה הפיצול עבור הממוצעים המנורמלים יהיה זהה, ולכן אותה כמות דוגמאות תעבור לכל בן של אותה הצומת, ערך האנטרופיה יהיה זהה עבור כל התכונות ולכן האנטרופיה המינימלית תהיה עבור אותן תכונות ואותם ערכי Threshold מנורמלים, ולכן כל צומת תתפצל באופן זהה כך ש:

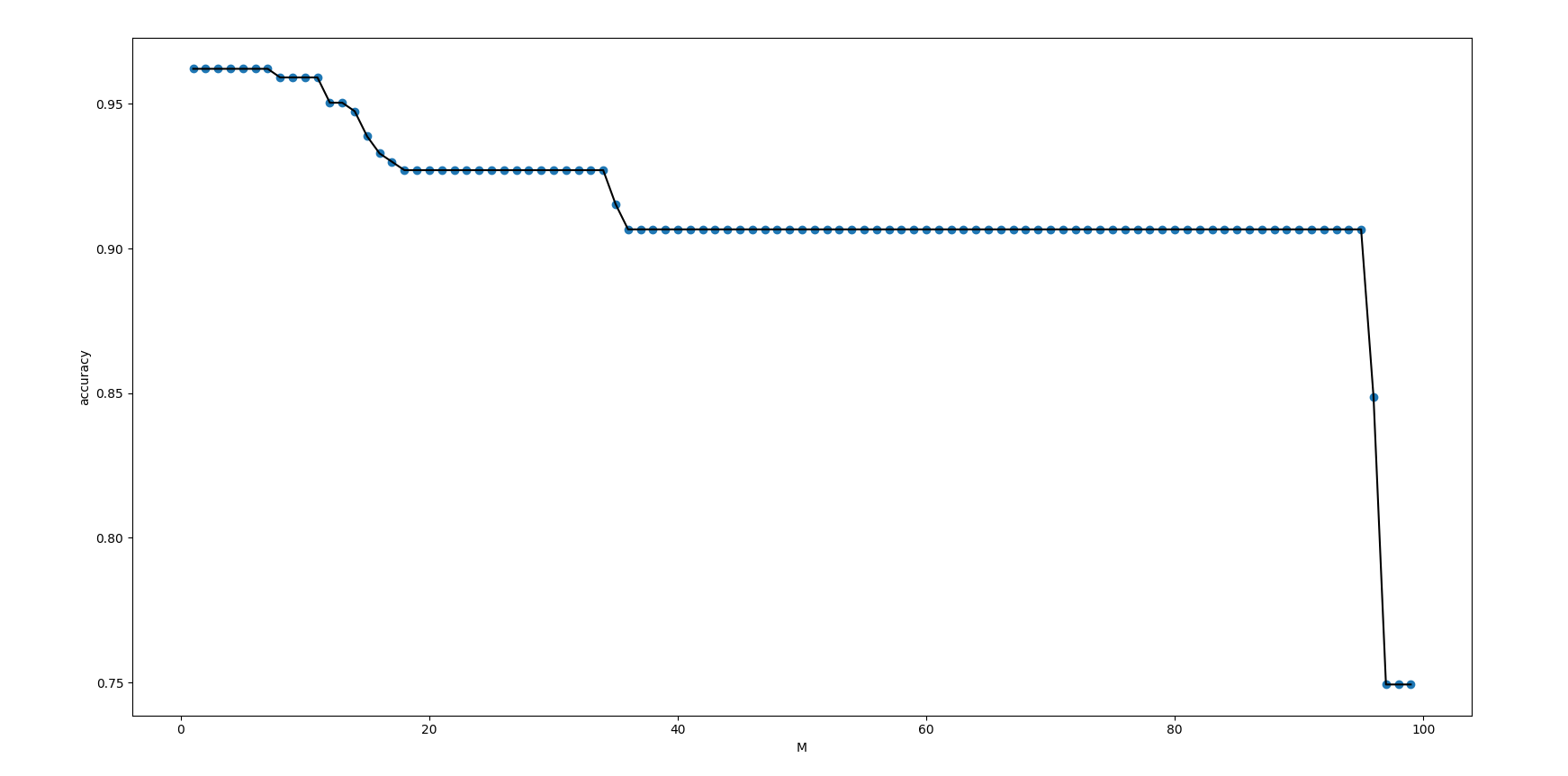
2. בעץ המנורמל, דוגמאות מנורמלות יעברו באותו מסלול.

נניח שיש דוגמת מבחן וצומת כלשהי עם פיצ'רf וThreshold T. ערך הפיצ'ר עבור דוגמת המבחן הוא ונניח בה"כ , כלומר הדוגמא תעבור לצומת הבן שמתאימה לערכים קטנים מהThreshold. מתכונת שימור הסדר של הנרמול שהראנו קודם, מתקיים גם ולכן גם הדוגמא המנורמלת תעבור לאותו בן של אותה צומת בעץ המנורמל. זה נכון עבור כל פיצול בעץ ולכן בסופו של דבר דוגמת המבחן תעבור באותו מסלול, תגיע לאותו עלה ותסווג באופן זהה.

הראנו שהעץ שיבנה יהיה זהה רק עם ערכים מנורמלים של Threshold ומצד שני דוגמאות המבחן יגיעו לאותם עלים ולכן הסיווגים יהיו זהים במקרה של נרמול ולא תהיה השפעה על התוצאה עבור קבוצת המבחן (כל עוד גם היא מנורמלת באופן זהה)

# שאלה 3

3.1. הגיזום מנסה למנוע את תופעת התאמת היתר (overfitting) בה האלגוריתם מאומן על קבוצת אימון אך לא יודע להכליל ולקבל תוצאות מספיק טובות על קבוצות אחרות. החשיבות שלו הוא שהוא מוריד עלים שהתקבלו ככל הנראה מדוגמאות שהן "רעש" ולא מייצגות את העולם באופן כללי.

3.3 מצורף הגרף, אמנם נתבקשנו לבחור לפחות 5 ערכי M שונים אבל בגלל נדיבות יתר שלי תקבלו אפילו 100 :  


למעשה התוצאה הטובה ביותר היא עבור M=1, והיא יורדת ככל שמעלים את M (בחלק מהנקודות היא לא משתנה כי אין עלה שמתפצל לפי אותה כמות דגימות).  
כלומר ללא גיזום מוקדם כלל התקבלה התוצאה הטובה ביותר. התוצאה הממוצעת בין הfolds על הvalidation set היא 96 אחוז.

3.4 מכיוון שהתוצאה הטובה ביותר היא ללא גיזום בכלל בסעיף זה קיבלנו שוב תוצאה של ואין שיפור בעץ.

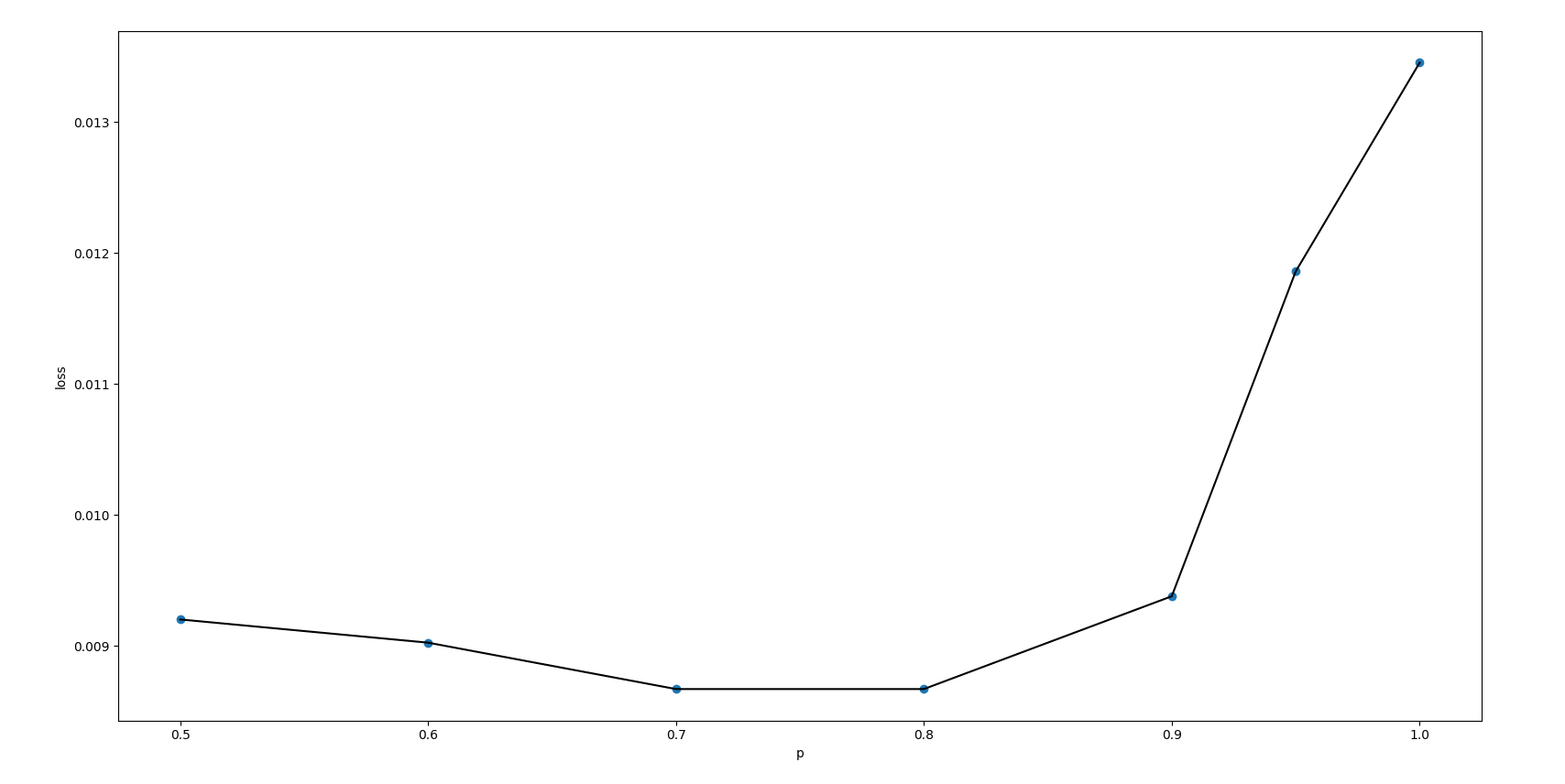
# שאלה 4

# 4.1 קיבלתי את ערך הloss הבא:

0.021238938053097345

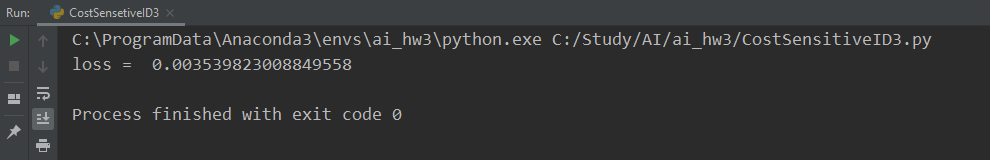
הוא נכון גם עבור "הגיזום הכי טוב" כי הגיזום הכי טוב הוא עבור M=1

4.2 נבצע החמרה בכל פיצול במקרים בהם רוב הדוגמאות הם של חולים. נגדיר פרמטר ובכל פעם שאנחנו מפצלים צומת, נבדוק אם החלק של הדגימות שערכן 1 (חולה) מכל הדגימות גדול מp, אם כן הצומת הרלוונטי יהפוך לעלה שנותן חיזוי של חולה. כך בעצם האלגוריתם מחמיר ונותן חיזוי של חולה לכל קבוצה בה יש רוב משמעותי של חולים ולא ממשיך לפצל אותה מתוך הנחה שיש הסתברות שהפיצול יתבסס על דגימות בודדות שנובעות מרעש.

4.3 ביצעתי תהליך cross validation עם k-fold=5 בדומה לשאלה הקודמת על הערך p שהוגדר. תוצאות הניסוי :

שהp הטוב ביותר הוא 0.7 או 0.8, בחרתי להשתמש ב0.8 ושיפררתי משמעותית את הloss.

קיבלנו:

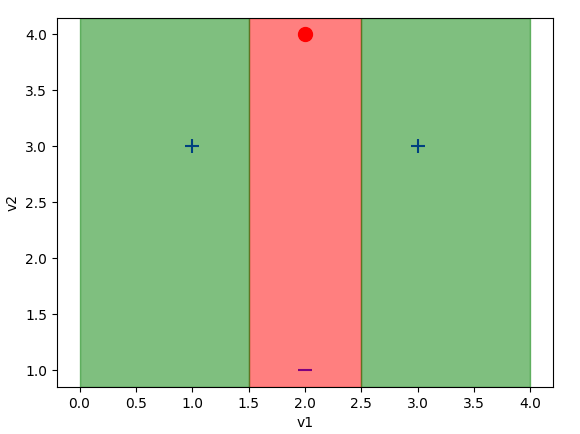
**

*loss = 0.003539823008849558*

# שאלה 5

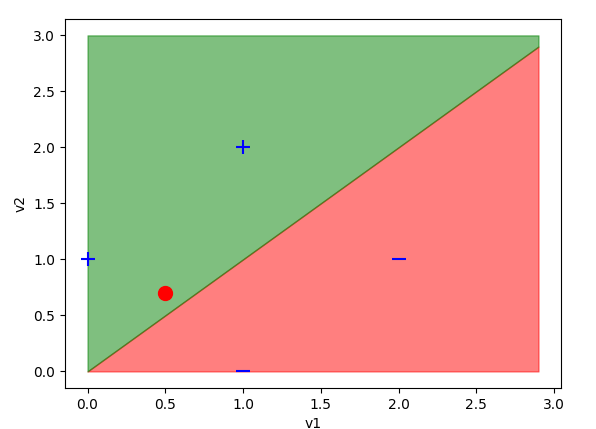
בכל הגרפים הבאים צבע הרקע מסמן את מסווג המטרה , כלומר הסיווג הנכון. רקע אדום מסמן דוגמא שלילית ורקע ירוק דוגמא חיובית.

## סעיף א'



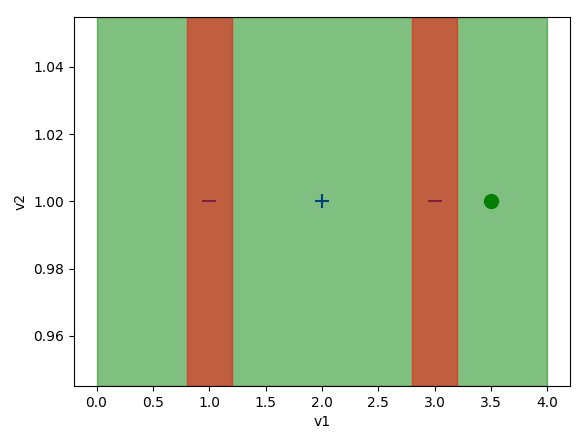
עבור דוגמת המבחן המסומנת בעיגול אדום, מסווג KNN יטעה עבור כל K כיוון שתמיד יהיו יותר פלוסים קרובים. לID3 יהיה קל ללמוד - פשוט 2 פיצולים לפי .

## סעיף ב'



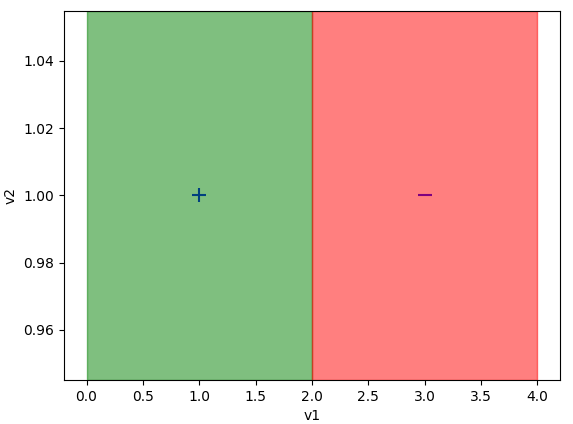
כאן KNN עם K=1 תמיד יהיה צודק כי דוגמאות באזור הירוק תמיד יהיו קרובות יותר לפלוס בעוד שדוגמאות באזור האדום יהיו קרובות יותר למינוס אבל ID3 לא ידע להפריד טוב לפי קווים אלכסוניים.

## סעיף ג'



כאן גם KNN וגם ID3 יטעו עבור דוגמת המבחן שמסומנת בעיגול ירוק. עבור KNN עם K=1 הכי קרוב הוא מינוס, וID3 יפצל את המרחב לפי דוגמת המינוס הימנית כך שכל מה שמימינה יקבל סיווג שלילי.

## סעיף ד'



כאן KNN עם K=1 תמיד יצדק כי משמאל לקו ההפרדה תמיד יהיה קרוב לדוגמא החיובית ולהיפך מצד ימין. ID3 מפצל לפי ממוצע ערכי הפיצ'ר לכן הוא ילמד פיצול אחד בדיוק בין 2 הדוגמאות.