**חלק יבש**

1. הטענה נכונה. הוכחה:

ראשית, נוכיח שתי טענות עזר:

טענה א': בהינתן דיוק שונה על שני עצים, הם נבדלים במבנה שלהם.

בהינתן קבלת דיוק שונה, עבור שני מסווגים ID3, בהכרח דוגמת מבחן שסווגה בצורה אחת בעץ א', סווגה אחרת בעץ ב'. כלומר, במהלך ביצוע החיזוי נלקח מסלול שונה בעץ א' מאשר בעץ ב'. אחרת, היינו מגיעים לאותו עלה ועל כן העץ שונה.

טענה ב': נרמול Minmax משמר סדר.

בה"כ נניח כי נראה כי מתקיים: .

נסמן , .

נניח בשלילה כי נרמול Minmax משנה את דיוק מסווג ID3.

מטענה א', לאחר הנרמול בהכרח התקבל עץ שונה, נביט בצומת הראשון שבמהלך הבניה נבחר שונה בין שני העצים. כלומר, בעץ א' נבחר קריטריון א', ובעץ ב' נבחר קריטריון ב'. כלומר, בעץ ב' קריטריון א' היה בעל IG נמוך משל קריטריון ב'. כלומר קריטריון ב' הפריד לשתי קבוצות בגדלים וקריטריון א' הפריד לשתי קבוצות בגדלים .  
לפי טענה ב', הנרמול אינו משנה סדר ולכן קריטריון ב' היה מפריד בעץ א' גם לשתי קבוצות בגדלים , דבר אשר היה מניב IG גדול יותר מאשר בקריטריון א'. בסתירה לבניית מסווגי ID3.

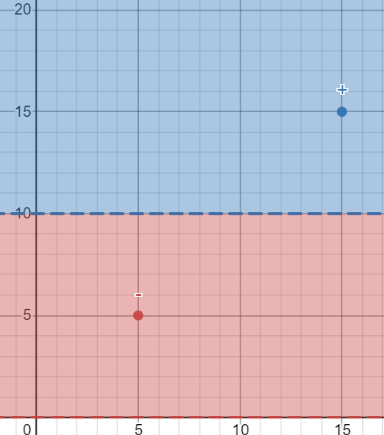
ולכן הטענה נכונה.

1. סעיף א'

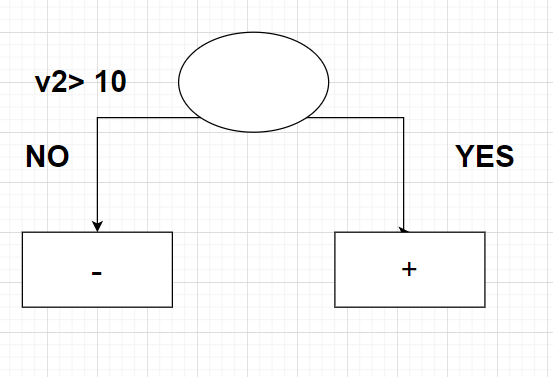
מסווג מטרה:

קבוצת האימון:

אשר מתוארים בגרף, תחום כחול – ערכי '+' במסווג המטרה, תחום אדום '-'.



העץ הנלמד ע"י ID3:



לכל דוגמת מבחן ID3 יסווג נכון. עבור דוגמת המבחן נראה כי KNN יטעה:

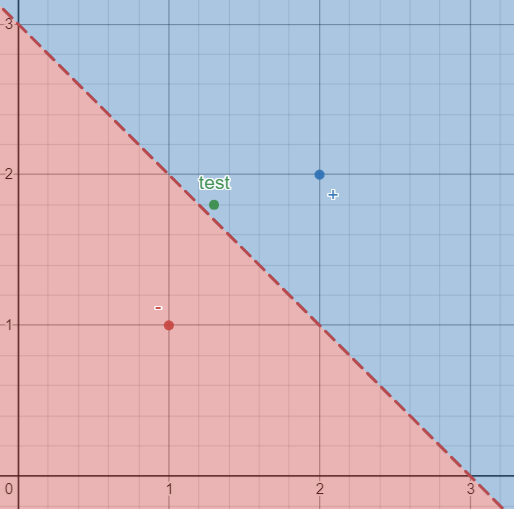
עבור : דוגמת האימון הקרובה במרחק אוקלידי היא '+' ולכן טועה.

עבור : יש שתי דוגמאות אימון, ולכן ערך הv1 המקסימלי יהיה בדוגמת ה'+' ולכן שוב יטעה.

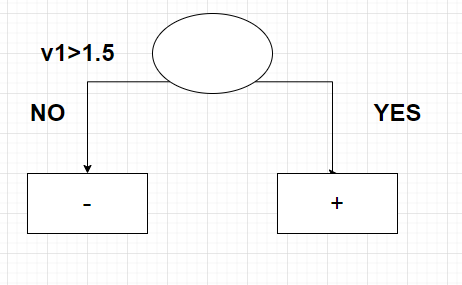
סעיף ב'  
מסווג מטרה:

קבוצת האימון היא:

אשר מתוארים בגרף, תחום כחול – ערכי '+' במסווג המטרה, תחום אדום '-'.



העץ הנלמד ע"י ID3:



עבור דוגמת המבחן , ID3 יטעה ויסווג אותה כ'-'.

המסווג הנלמד ע"י KNN עם הוא בדיוק מסווג המטרה (מסימטריות דוגמאות האימון).

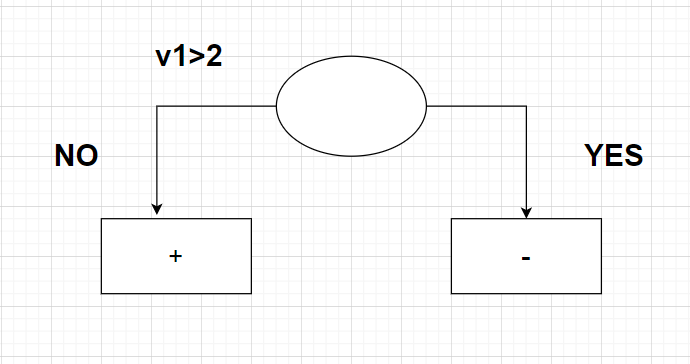
סעיף ג'

מסווג מטרה:

קבוצת האימון היא:



העץ הנלמד ע"י ID3:



נסתכל על דוגמת מבחן (x=(5,0), +).

עבור KNN, עם , הסיווג יקבע לפי ולכן KNN יסווג '-' באופן שגוי .

עבור ID3, ולכן תסווג כ'-' באופן שגוי.

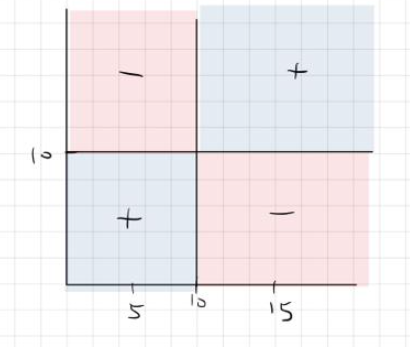
סעיף ד'

מסווג מטרה:

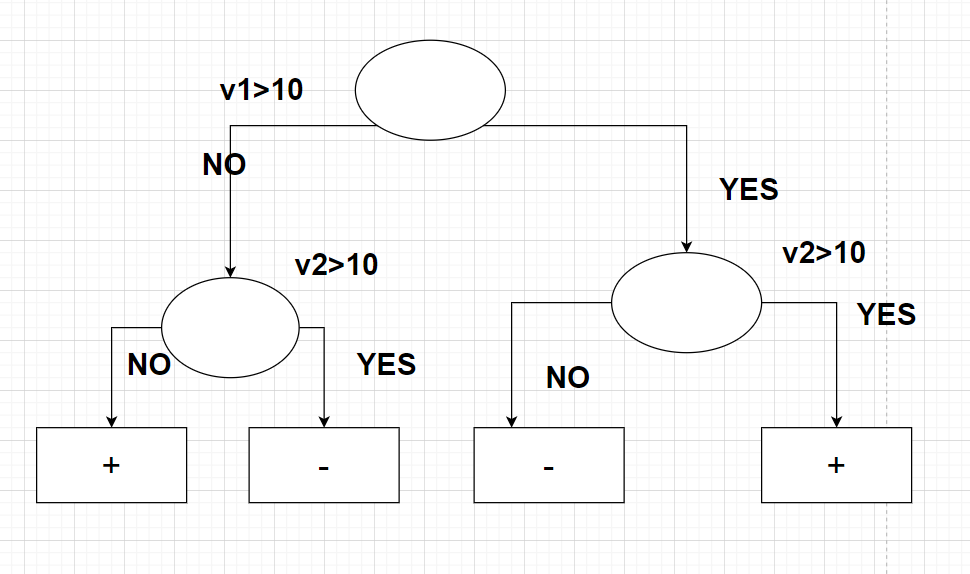
מסווג מטרה מתואר בציור, בצבע אדום '-' ובצבע כחול '+'

קבוצת האימון היא:

כמסומן בגרף:



העץ הנלמד ע"י ID3:



KNN למד את מסווג המטרה*, עבור , ציור* decision boundaries יהיה בדיוק המסווג. וכן עבור ID3 כל דוגמאות המבחן יסווגו באופן נכון.

1. א. בקבוצת האימון ישנן 5 דוגמאות עם סיווג 1, ו5 דוגמאות עם סיווג 0. לכן ללא תלות בקלט מסווג majority classifier יבחר בסיווג 1. ערך הדיוק יהיה 0.5.

ב. עבור הfold הראשון (5 הנקודות השמאליות), תהיה שגיאה אחת. ולכן ערך הדיוק יהיה

עבור הfold השני, תהיה שגיאה אחת. ולכן ערך הדיוק יהיה .

לכן בסך הכל ערך הדיוק יהיה הממוצע .

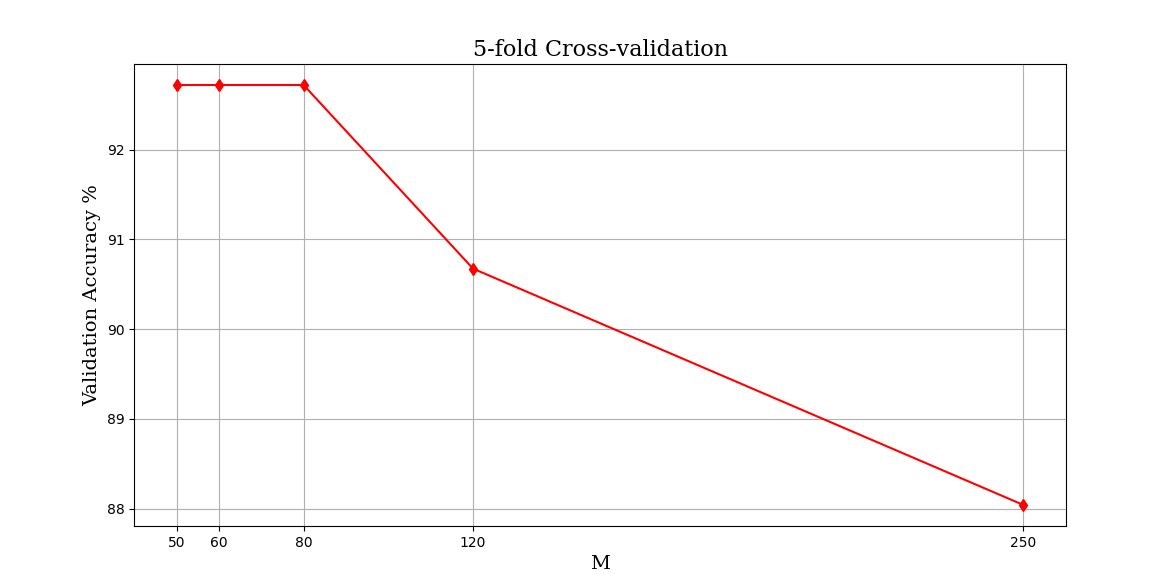
**חלק רטוב**

1. b. קיבלנו דיוק של 96.46%.

6.

a. באופן כללי גיזום משפר זמני ריצה, תורם להכללה טובה יותר על המידע. ללא גיזום העץ שייבנה יהיה מותאם באופן מושלם למידע של קבוצת האימון ועלול לגרום לטעויות בקבוצת המבחן. אנו מעוניינים להכליל על מידע האימון ולכן נרצה למנוע את תופעת הoverfitting.

c.



ניתן לראות בגרף כי כאשר ישנה ירידה בערך הדיוק, בגלל underfitting.

את התוצאה הטובה ביותר קיבלנו עבור עם ערך דיוק של 92.72%. תוצאה זאת תהיה פחות טובה מאשר תוצאת הטסט (כפי שנראה בהמשך) בגלל הבחירה הרנדומית של קבוצת הולידציה וגודלה הקטן יותר של קבוצת האימון. במהלך הולידציה אנחנו לא מגיעים להכללה טובה יותר של המידע כמו שנגיע לאחר אימון מלא.

d. עבור קיבלנו דיוק של 97.35%. אכן הגיזום שיפר את הביצועים ביחס להרצה ללא גיזום.

7. בשלב האימון נאחסן את הדוגמאות ואת הסיווג שלהן, ובשלב הסיווג נמדוד את הדמיון בין האובייקט הבלתי מסווג לכל דוגמאות האימון ששמרנו. הסיווג יקבע לפי החלטת הרוב של K הדוגמאות הכי קרובות.  
יתרונות – שלב האימון מתבצע בזמן אפסי, מימוש פשוט למדי.  
חסרונות – שלב הסיווג לוקח זמן רב (כגודל קבוצת האימון), רגישות למדידות בחישוב על פי מרחק.

8.

a. ניתן להביט בשאלה זאת, כבחירה של האם מאפיין כלשהו נכלל או לא נכלל בקבוצת המאפיינים הסופית, כלומר כל מאפיין הוא בעל 2 אופציות. קבוצת המאפיינים היא בגודל D, ולכן מספר תתי הקבוצות - .

b. כמו קודם, ניתן להביט על הבעיה כבחירה של b מאפיינים, מתוך הקבוצה הכוללת של בגודל D, ולכן התשובה היא