תרגיל בית 3

ספי עזמי saf.azmi@campus.technion.ac.il 204511414

אופיר מנור ofir.manor@campu.technion.ac.il 316084623

חלק ב'

שאלה 1)

נוכיח באינדוקציה שעבור דאטה עם תכונות רציפות ותיוגים בינאריים אלגוריתם ID3, בונה את אותו עץ החלטה עם הדאטה בצורתו המקורי ומנורמל דרך minmax, ולכן הנרמול לא משפיע על דיוק המסווג על קבוצת האימון או על קבוצת הבוחן.

למה: עבור קבוצת דאטה עם תכונות רציפות $V = \{v_1, v_2, \dots, v_n\}$ ותיוגים בינאריים $L = \{0,1\}$ ותכונה f נקבל שהפיצול שייבחר בשיטת אוטו-דיסקרטיזציה יחזיר שתי תת-קבוצות זהות וIG זהה עבור התכונה המקורית, ועבור התכונה אחרי נרמול minmax

הוכחה: נסמן ב $V_{rise} = v_{i_1}, v_{i_2}, \dots, v_{i_n}$ סידור עולה של נק' הדאטה ב V לפי סדר עולה, ואת $T = \{t_1, t_2, \dots, t_{n-1}\}$ קבוצת נקודות הגבול כך ש $t_j = \frac{1}{2} \cdot (v_{i_j} + v_{i_{j+1}})$. לפי ID3 יבחר t_j כך שפיצול ל$V_1 = \{v_{i_1}, \dots, v_{i_j}\}$ ו $V_2 = \{v_{i_{j+1}}, \dots, v_{i_n}\}$ יחזיר את הIG המקסימלי. נסמן $v'_{i_j} = \frac{(v_{i_j} - v_{i_1})}{(v_{i_n} - v_{i_1})}$ את הנרמול minmax של נק' הדאטה v_{i_j}. הנרמול אינו משנה את סדר הנק' דאטה ולכן נסמן את הסידור העולה של נק' הדאטה אחרי הנרמול ב $V'_{rise} = v'_{i_1}, v'_{i_2}, \dots, v'_{i_n}$ ואת $T' = \{t'_1, t'_2, \dots, t'_{n-1}\}$ כאשר

$t'_j = \frac{1}{2} \cdot \left(v'_{i_j} + v'_{i_{j+1}}\right) = \frac{t_j - v_{i_1}}{v_{i_n} - v_{i_1}}$ וכלן עבור כל t'_j הפיצול שמתקיים עבורו הוא $V'_1 = V_1, V'_2 = V_2$ ולכן הערך של IG עבור פיצול הדאטה לפי t'_j שווה שלערך עבור הפיצול לפי t_j. מכיוון שעבור הדאטה המקורית נבחר פיצול לפי t_j מכיוון שערך ה IG היה מקסימלי, אזי עבור הדאטה המנורמל יבחר t'_j מכיוון שערכי ה IG שווים. לכן הפיצול והערך IG המתקבלים בID3 יהיו זהים ■

בסיס: עבור כלל הדאטה המקורי, בעץ ההחלטה המקורי נבחר התכונה f_i והפיצול t_j אשר ממקסם את הIG ומהלמה נקבל שלכל תכונה f עבור הדאטה המנורמל נקבל את הפיצול וIG זהים לדאטה המקורי, לכן עבור הדאטה המנורמל ID3 יפצל לפי f_i וt_j

הנחה: נניח שID3 לפי דאטה מנורמל בחר את אותם תכונות ופיצולים עבור n הפיצולים הראשונים.

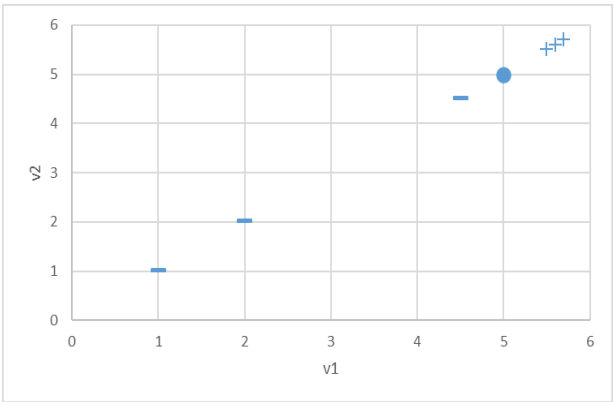
צעד: זהה לבסיס. ■

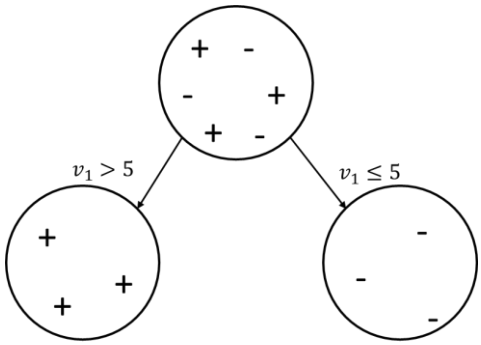
שאלה 2)

א. נגדיר סיווג מטרה $f(v_1, v_2,) = \begin{cases} 0 & \mid v_1 \le 5 \\ 1 & \mid otherwise \end{cases}$ וקבוצת דאטה

$$D = \{((4.5,4.5),0),((5.5,5.5),1),((5.6,5.6),1),((5.65,5.65),1),((1,1),0),((2,2),0)\}$$

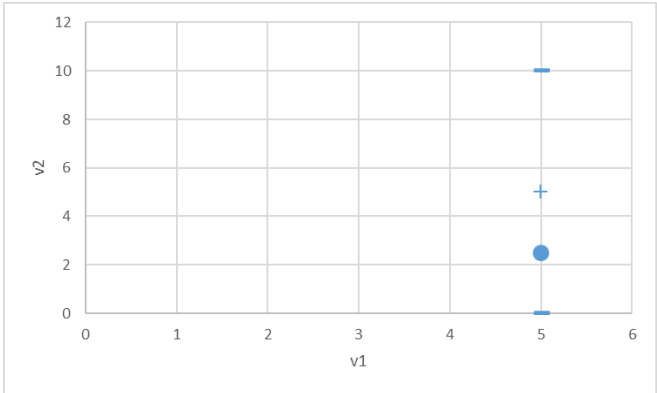
נראה שעבור ID3 יתקבל f ועבור כל $k \in \{1,3,5\}$ KNN יטעה על $d = ((5,5),0)$

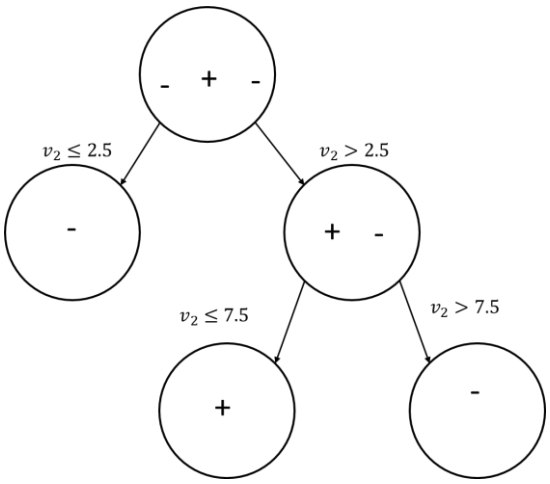




ב. נגדיר סיווג מטרה $f(v_1, v_2) = \begin{cases} 1 \mid \sqrt{(v_2-5)^2} \leq 2.5 \\ 0 \mid \quad otherwise \end{cases}$ עם דאטה

$D = \{((5,0),0),((5,5),1),((10,5),0)\}$ נקבל שעבור $K = 1$ KNN ייצג את f ועבור הדוגמה $d = ((5,2.5),1)$ ID3 (כפי שנלמד בהרצאה) **יטעה**.

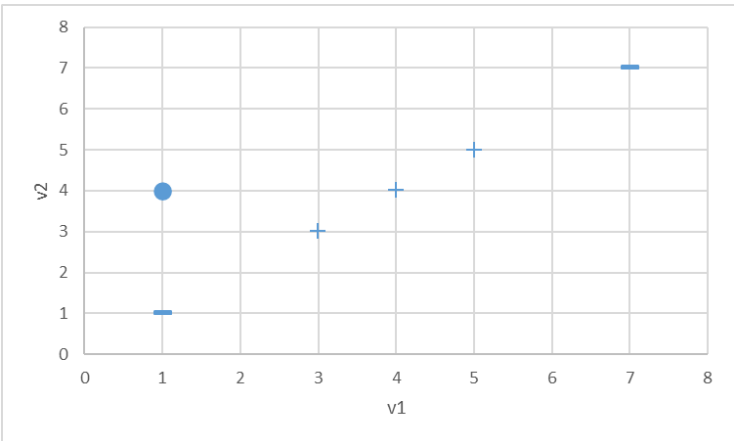


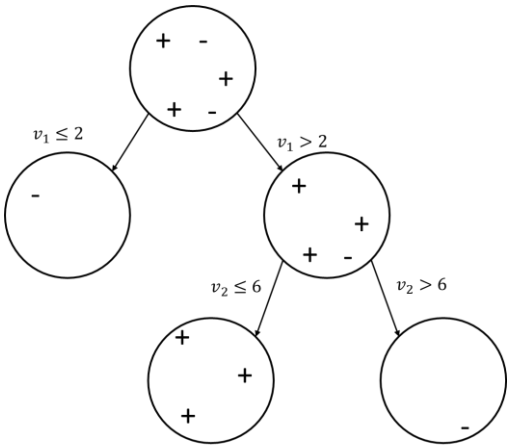


ג. נגדיר סיווג מטרה $f(v_1, v_2) = \begin{cases} 1 \mid 2.5 \leq v_1 \leq 6.5 \ and \ 2.5 \leq v_2 \leq 6.5 \\ 0 \mid \quad otherwise \end{cases}$ עם דאטה

$D = \{((1,1),0),((3,3),1),((4,4),1),((5,5),1),((7,7),0)\}$ כך שעבור $K = 1$ KNN

יטעה על $d_1 = ((1,4),0)$ וID3 יטעה על $d_2 = ((6.4,6.4),1)$





ד. נגדיר סיווג מטרה $f(v_1, v_2) = \begin{cases} 1 \mid & v_1 \geq 5 \\ 0 \mid & otherwise \end{cases}$ עם דאטה $D =$ $\{((2.5,1), 0), ((7.5,1), 1)\}$ ונקבל שעבור $K = 1$ KNN יקבל את f וש ID3 יקבל את f.

שאלה 3)

א. עבור הדאטה הנתון הMajority Classifier יחזיר תמיד 1 (מכיוון מספר הנק' הדאטה שהתיוג שלהם 1 ומספר הנק' שהתיוג שלהם 0 שווה). ישנם 10 דוגמאות סך הכל ומתוכם 5 התיוג שלהם 1. לכן הדיוק על קבוצת האימון היא $\frac{5}{10} = 50\%$

ב. נחלק מקרים:

1. עבור 5 הנק' השמאליות כקבות אימון נקבל שהMajority Classifier מחזיר תמיד 1, ועבור 5 הנק' הנותרות בתור קבוצת מבחן יש אחת שהתיוג שלה 1 והשאר 0, ולכן נקבל דיוק של $\frac{1}{5} = 20\%$.
2. עבור 5 הנק' הימניות כקבוצת אימון נקבל שהMajority Classifier מחזיר תמיד 0, ועבור 5 הנק' הנותרות בתור קבוצת מבחן יש אחת שהתיוג שלה 0 והשאר 1, לכן נקבלך דיוק של $\frac{1}{5} = 20\%$.

לכן נקבל שהדיוק הכללי הוא $\frac{\frac{1}{5}+\frac{1}{5}}{2} = \frac{1}{5} = 20\%$

חלק ג'

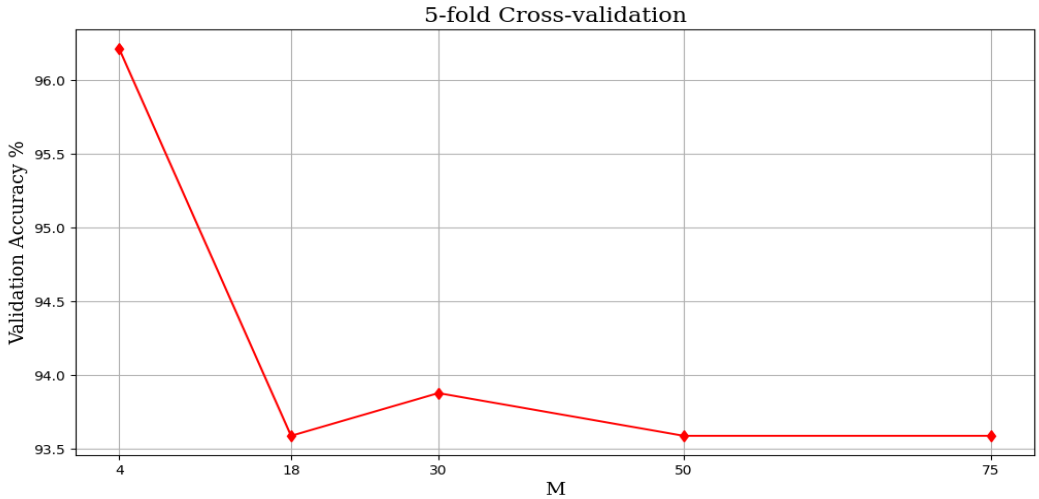
שאלה 5) סעיף b

בהרצת ID3 על קבוצת דאטה train ובדיקה על מול קבוצת מבחן train הגענו לדיוק של 94.46% משמע סיווג מדויק של 109 נקודות בדיקה מתוך 103.

שאלה 6)

b) גיזום עוזר להוציא נק' דאטה "רועשות", במיוחד באלה שתויגו לא נכון בדאטה מלהשפיע על עץ ההחלטה. כך הוא מונע תופעה של overfitting בזה שהוא כולל את הדוגמאות האלה בתוך עלים שמציינים תיוג שונה (וככל הנראה המתאים).

c) השתמשנו בערכי $M = [4,18,30,50,75]$



קיבלנו שהגיזום עם התוצאות הטובות ביותר הוא 4 והוא מקבל דיוק של 96.21% לפי K – fold cross validation עם $K = 5$ נראה בכך שעדיף גיזום עדין על גיזום גס מכיוון שהדוגמאות הרועשות הן יחסית "קרובות" לדוגמאות הנכונות, אבל שיש מספר דוגמאות יוצאי דופן

d) לאחר הרצת אלגוריתם ID3 *עם גיזום מוקדם בערך* $M = 4$ אנו מקבלים דיוק של 96.46% שהוא זהה לתוצאה שקיבלנו ללא גיזום. נשים גם לב שאם אנו משתמשים בערך $M = 50$ כפי שנרמז בשאלה אנו מקבלים דיוק של 97.35%, שהוא אכן יותר גבוה אך לא התשקף בבדיקה המקדימה והוא אכן משפר את הדיוק, אך לא בהרבה.

חלק' ד

שאלה 7)

אלגוריתם KNN מסווג נק דאטה חדשה על ידי כך שהוא מחשב את המרחק האוקלידי שלה מכל הנק' הדאטה שהתאמן עליהם כך שהערך של כך תכונה מסמן מיקום על ציר התכונה ומחזיר את המסווג של רוב K הקרובים ביותר.

יתרונות: אלגוריתם קל ליישום ומטפל בדאטה רועש על ידי כך שלוקח בחשבון K השכנים

חסרונות: סיווג עיתי ורגיש לבחירת תכונות

שאלה 8)

a) לפי הנלמד בתורת הקבוצות נקבל ש $|P(S)| = 2^{|S|} = 2^D$

b) לפי קומבינטוריקה נקבל שמספר תתי-קבוצות התכונות בגודל b שווה למספר הדרכים לבחור b איברים מתוך קבוצה של D איברים ולכן שווה $\binom{|S|}{b} = \binom{D}{b}$

שאלה 9)

a) יש לבדוק את הביצועים על קבוצת הוולידציה. מה שצריך לעשות זה לקחת את קבוצת האימון ולחלק אותה לקבוצת אימון מצומצמת יותר וקבוצת ולידציה, יש לאמן את האלגוריתם על קבוצת האימון המצומצמת ועם המאפיינים שנבחרו ולבדוק את הביצועים על קבוצת הוולידציה. בדיקה על קבוצת האימון מיותרת כי אליה התאמן העץ ועל קבוצת המבחן יוצרת מצב של תיאום יתר לקבוצת המבחן, ותפגע בתוצאות של בדיקות על קבוצות מבחן "אחרות" (לדוג' מטופל חדש שלא היה חלק מהקבוצות).

b) קיבלנו קבוצת מאפיינים בגודל 2, נשתנה שיפור של כ2.5% (78% עבור k=51, אל מול התוצאה של 75.50% על כל המאפיינים עבור k=51)

c) אנו מצבעים אלגוריתם חמדן פשוט אשר כל פעם מוסיף את המאפיין אשר מניב את הדיוק הגדול ביותר עד שאף מאפיין לא משפר את הדיוק. האלגוריתם מבצע את השלבים הבאים:

```
1) unused_features <- all features
2) best_accuarcy <- 0.0
3) model = KNN
4) while unused_features is not empty:
    a. accuracy_for_all_features <- [0]
    b. for feature in unused_features:
        i. best_features.append(feature)
        ii. accuracy_for_current_features <- [0]
        iii. for train_set, test_set from 5-fold validation of training data:
            1. model.train(train_set[best_features])
            2. accuracy<-accuracy(model,test_set)
            3. accuracy_for_current_features.append(accuracy)
        iv. avg_accuracy<-avg(accuracy_for_current_features)
        v. accuracy_for_all_features.append(avg_accuracy)
        vi. best_features.pop(feature)
    c. best_feature = max(accuracy_for_all_features)
    d. if accuracy(best_feature) > best_accuracy
        i. best_accuracy <- accuracy(best_feature)
    e. else
        i. break
    f. best_features.append(best_feature)
    g. unused_features.pop(best_feature)
5) return best_features
```