Short HW2: Classification: Introduction

1. Decision Trees:

1.1. הדוגמה תלקח מעולם הספורט:

תכונת יעד: האם הקבוצה תנצח במשחק?

תכונות בינאריות:

1. האם הקבוצה משחקת באולמה הביתי?
2. האם הקבוצה מדורגת גבוהה יותר בטבלת הליגה?
3. האם הקבוצה ניצחה במשחקה האחרון?

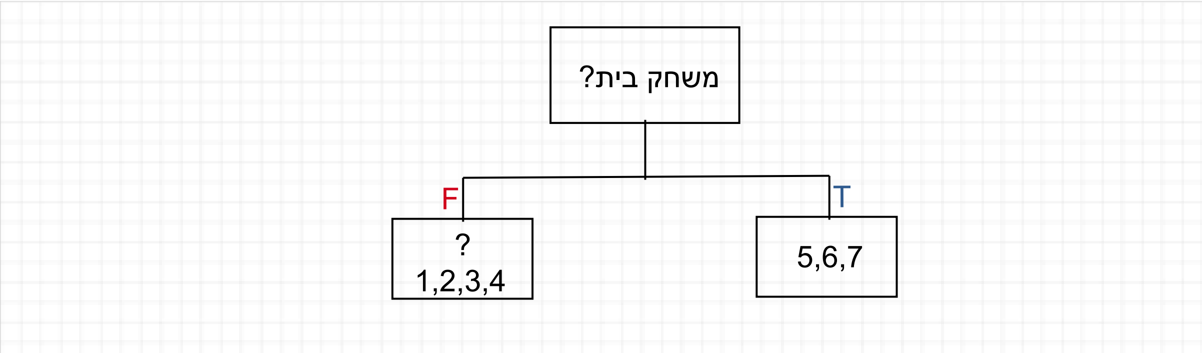
מאגר הנתונים:

מספר דוגמה	משחק בית	דירוג גבוהה	ניצחה משחק אחרון	ניצחה במשחק
1	F	F	F	F
2	F	F	T	F
3	F	T	F	F
4	F	T	T	T
5	T	F	F	T
6	T	F	T	T
7	T	T	T	T

1.2. נפעיל את אלגוריתם ID3 באופן ידני בדומה לתרגול, לכן תחילה נחשב את ה-Information Gain של כל תכונה על מנת למצוא את התכונה בעלת ה-Information Gain הגדולה ביותר:

תכונה	$\frac{\|v_{\alpha=T}\|}{\|v\|}$	$\frac{\|v_{\alpha=F}\|}{\|v\|}$	$H(v_{\alpha=T})$	$H(v_{\alpha=F})$	$IG(v,\alpha)-H(v)$
משחק בית	$\frac{3}{7}$	$\frac{4}{7}$	$H\left(\frac{3}{3}\right)$	$H\left(\frac{1}{4}\right)$	$-\frac{4}{7}H\left(\frac{1}{4}\right) \cong -0.46$
דירוג גבוהה	$\frac{3}{7}$	$\frac{4}{7}$	$H\left(\frac{2}{3}\right)$	$H\left(\frac{2}{4}\right)$	$-\frac{3}{7}H\left(\frac{3}{3}\right)-\frac{4}{7}H\left(\frac{2}{4}\right) \cong -0.97$
ניצחה משחק אחרון	$\frac{4}{7}$	$\frac{3}{7}$	$H\left(\frac{3}{4}\right)$	$H\left(\frac{1}{3}\right)$	$-\frac{4}{7}H\left(\frac{3}{4}\right)-\frac{3}{7}H\left(\frac{1}{3}\right) \cong -0.86$

נבחר את התכונה בעלת ה-Information Gain הגדול ביותר, כלומר החלוקה הראשונה בעץ תהיה לפי האם הקבוצה משחקת באולמה הביתי.

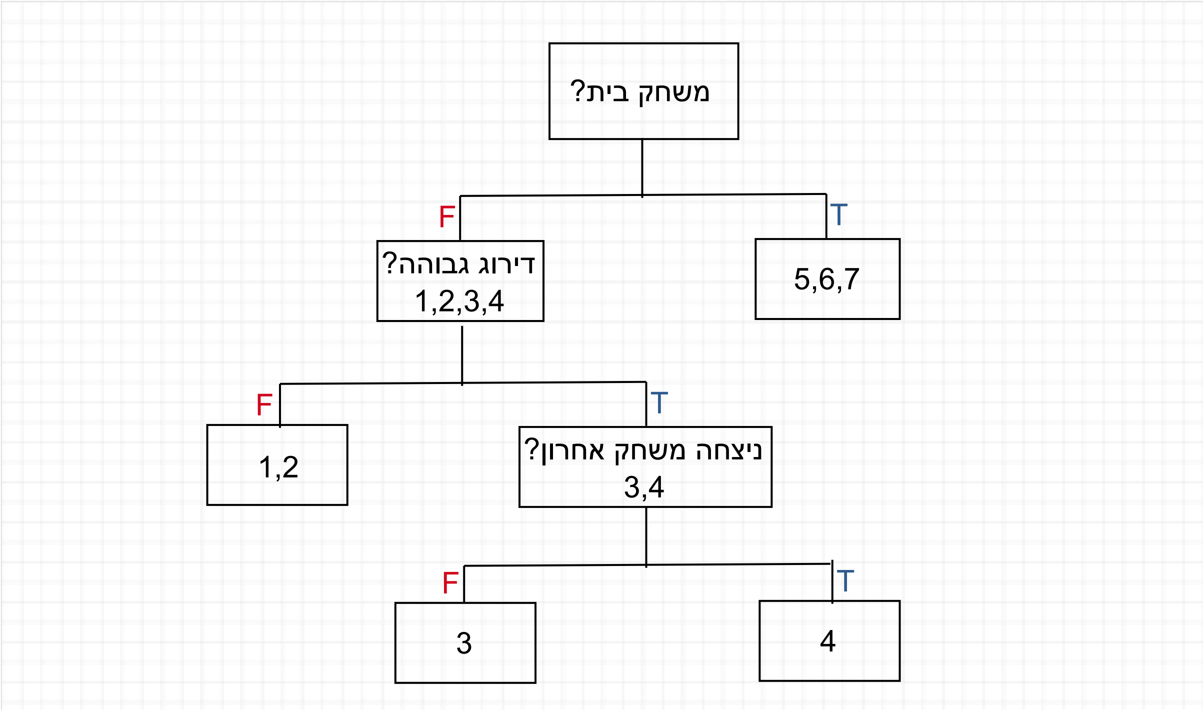


את הדוגמאות שסימנו ב-F עבור משחק בית עדיין לא תייגנו באופן סופי, כעת נרצה לתייג גם אותן ולכן נרצה למצוא את התכונה הבאה לפיה נחלק את התכונות כאשר התכונות הנותרות הן האם הקבוצה מדורגת גבוהה יותר מיריבתה והאם הקבוצה ניצחה במשחק האחרון שלה:

מספר דוגמה	דירוג גבוהה	ניצחה משחק אחרון	ניצחה במשחק
1	F	F	F
2	F	T	F
3	T	F	F
4	T	T	T

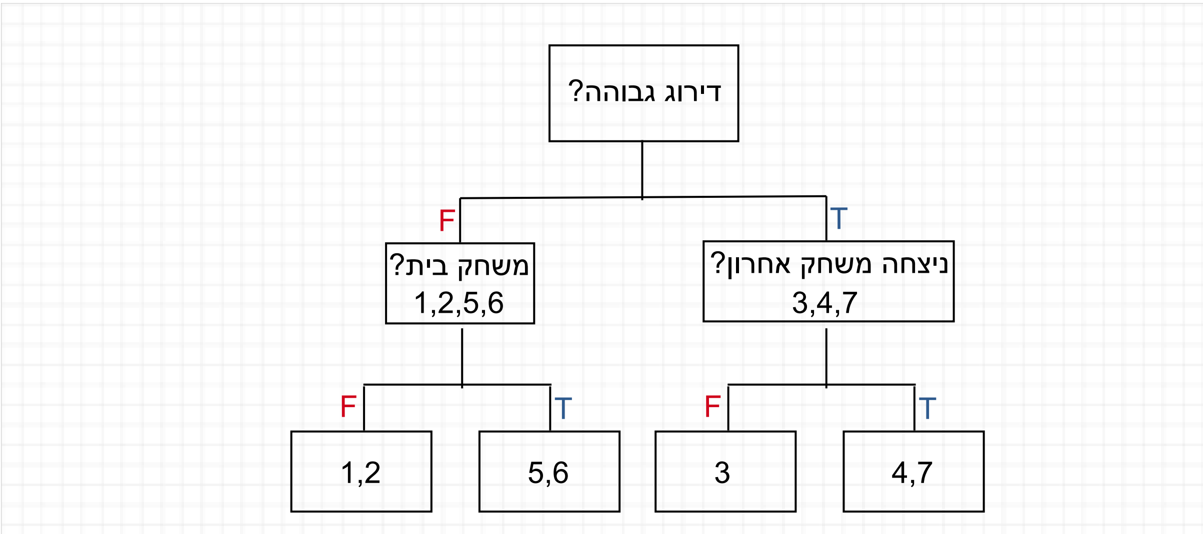
תכונה	$\frac{\lvert v_{\alpha=T}\rvert}{\lvert v\rvert}$	$\frac{\lvert v_{\alpha=F}\rvert}{\lvert v\rvert}$	$H(v_{\alpha=T})$	$H(v_{\alpha=F})$	$IG(v,\alpha)-H(v)$
דירוג גבוה	$\frac{2}{4}$	$\frac{2}{4}$	$H\left(\frac{1}{2}\right)=1$	$H\left(\frac{0}{2}\right)=0$	$-\frac{2}{4}=-0.5$
ניצחה משחק אחרון	$\frac{2}{4}$	$\frac{2}{4}$	$H\left(\frac{1}{2}\right)=1$	$H\left(\frac{0}{2}\right)=0$	$-\frac{2}{4}=-0.5$

כעת שתי התכונות מניבות את אותו ה-Information Gain ולכן נבצע בחירה שרירותית לחלק את הצומת לפי האם הקבוצה מדורגת יותר גבוה מהקבוצה השנייה, כלומר החלוקה האחרונה תתבצע לפי האם הקבוצה ניצחה במשחקה האחרון או לא.



כלומר מצאנו עץ בעומק 3 שמתאים לנתונים.

1.3. כעת נראה עץ בעומק 2 שמתאים לנתונים הנתונים:

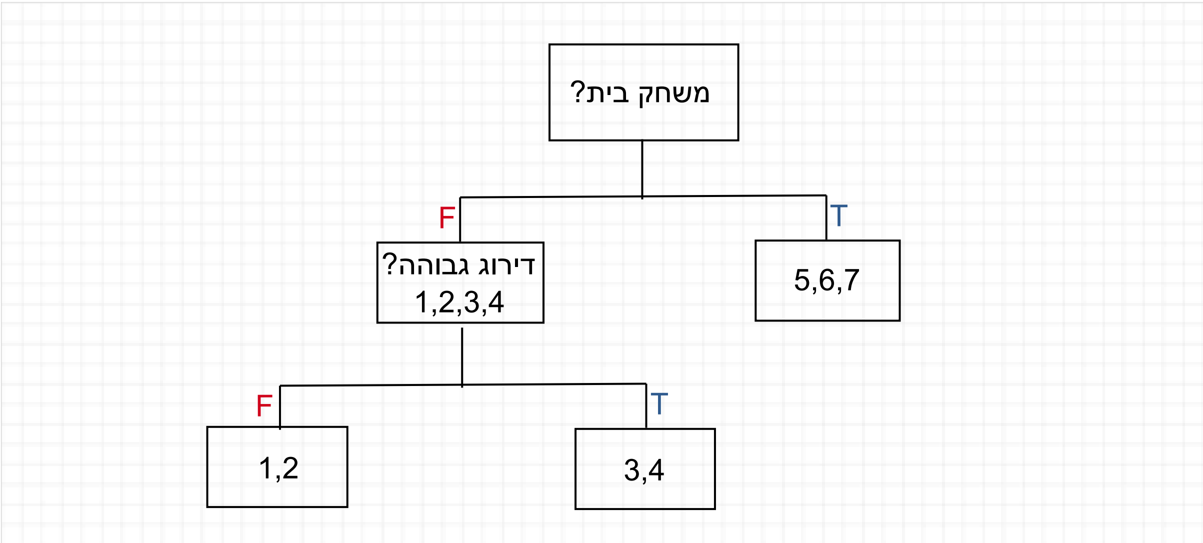


ואכן נבחין כי העץ בעומק 2 ומתאים לנתונים באופן מדויק, נראה כי השגיאה האמפירית היא 0:

$$empirical\ error = \frac{1}{m} \cdot \sum_{i=1}^{m} \mathbb{1}\{Dataset\_ID(i) \neq Tree\_Label(i)\} = \frac{1}{m} \cdot \sum 0 = 0$$

1.4. אם היינו מוסיפים מגבלה של עומק מקסימלי 2 עבור העץ, היינו מקבלים את אותו העץ אחרי שתי חלוקות, אבל בסוף החלוקה השנייה היינו מקבלים כי דוגמה 3 מתוייגת כ-T בניגוד לתיוג האמיתי שלה שהוא F.

כלומר היינו מקבלים את העץ הבא:



בשלב זה אלגוריתם ID3 היה עוצר ולכן היינו מקבלים שגיאה אמפירית שאינה 0:

$$empirical\ error = \frac{1}{m}\cdot\sum_{i=1}^{m}\mathbb{1}\{Dataset\_ID(i) \neq Tree\_Label(i)\} = \frac{1}{m}\cdot 1 \underbrace{=}_{m=7} \frac{1}{7}$$

2. Information Gain:

2.1. בהוכחה נטען כי:

$$Entropy(D) > 0$$
$$Sum([|Dv|\ /\ |D| * Entropy(Dv)]) > 0$$

ולכן לא ייתכן כי החיסור בניהם יתן מספר שלילי.
טענה זו לא סותרת את הנחת השלילה $(IG < 0)$ מפני שחיסור בין שני מספרים חיוביים לא בהכרח נותן תוצאה חיובית ויתכן כי ההפרש יהיה שלילי, כלומר יתכן כי מתקיים:

$$IG(A) = Entropy(D) - Sum([|Dv|\ /\ |D| * Entropy(Dv)]) < 0$$

ולכן זו אינה סתירה להנחת השלילה.

2.2. נשתמש בתכונה הנתונה בשאלה על מנת להוכיח כי $IG(v,a) \geqslant 0$:
לפי הגדרת IG נקבל כי:

$$IG(v,a) = H(v) - \frac{|v_{a=T}|}{|v|}H(v_{a=T}) - \frac{|v_{a=F}|}{|v|}H(v_{a=F})$$

$$\underbrace{=}_{\substack{\text{לפי הגדרת}\\ v_{a=T/F}}} H(v) - \underbrace{\frac{|v_{a=T}|}{|v|}}_{\alpha} H\underbrace{\left(\frac{|\{(x,y)\in v_{a=T} | y=1\}|}{|v_{a=T}|}\right)}_{\beta_1} - \underbrace{\frac{|v_{a=F}|}{|v|}}_{1-\alpha} H\underbrace{\left(\frac{|\{(x,y)\in v_{a=F} | y=1\}|}{|v_{a=F}|}\right)}_{\beta_2}$$

$$= H(v) + (-\alpha H(\beta_1) - (1-\alpha)H(\beta_2)) \underbrace{\geqslant}_{\text{לפי תכונת העזר}} H(v) + (-H(\alpha\beta_1 + (1-\alpha)\beta_2)) \quad (1)$$

כעת, ידוע כי הקבוצות $v_{a=T}, v_{a=F}$ הן קבוצות זרות ומשלימות ל-v, כלומר מתקיים $v_{a=T} \cup v_{a=F} = v$ אך $v_{a=T} \cap v_{a=F} = \varnothing$ ולכן מתקיים גם:

$$H(v) = H\left(\frac{|v_{a=T}| + |v_{a=F}|}{|v|}\right) = H\left(\frac{|\{(x,y) \in v_{a=T} | y = 1\}| + |\{(x,y) \in v_{a=F} | y = 1\}|}{|v|}\right)$$

$$\underbrace{=}_{|v|=|v_{a=T}|+|v_{a=F}|} H\left(\frac{|\{(x,y) \in v_{a=T} | y = 1\}|}{|v_{a=T}| + |v_{a=F}|} + \frac{|\{(x,y) \in v_{a=F} | y = 1\}|}{|v_{a=T}| + |v_{a=F}|}\right)$$

כעת נוכל להשתמש בפיתוח זה בביטוי מהנתון ולקבל:

$$H(\alpha\beta_1 + (1-\alpha)\beta_2) = H\left(\frac{|v_{a=T}|}{|v|} \cdot \frac{|\{(x,y) \in v_{a=T} | y = 1\}}{|v_{a=T}|} + \frac{|v_{a=F}|}{|v|} \cdot \frac{|\{(x,y) \in v_{a=F} | y = 1\}}{|v_{a=F}|}\right)$$

$$= H\left(\frac{|\{(x,y) \in v_{a=T} | y = 1\}| + |\{(x,y) \in v_{a=F} | y = 1\}|}{|v|}\right) = H(v)$$

וכעת אם נציב את הביטוי ב-(1) נקבל:

$$IG(v,a) \geqslant H(v) + (-H(\alpha\beta_1 + (1-\alpha)\beta_2)) = H(v) - H(v) = 0$$
$$\Longrightarrow IG(v,a) \geqslant 0$$

3. Separability:

3.1.

Model / Dataset	(A)	(B)	(C)
i.	כן	לא. קיימות שתי דוגמאות באיזור שבו $x \approx 3$ סביב ציר ה-y שהן בצבעים שונים - כלומר כל אחת מהן תסווג בצבע הלא נכון בשל הקרבה לצבע השני.	לא. אוסף הנתונים מורכב מזוגות סמוכים של דוגמאות בצבעים שונים. כל דוגמה מכל צמד תסווג באופן שגוי בגלל הקרבה לדוגמה מהסוג השני.
ii.	לא. יש סך הכל 4 נקודות כאשר מכל צבע ישנן 2. כלומר שלושת השכנים הקרובים של כל דוגמה הן דוגמה אחת מהצבע שלה ושתי דוגמאות מהצבע השני, לכן כל הנקודות יסווגו שגוי.	כן. (למרות דוגמה כחולה באיזור שבו $x \approx 3$ ו- $y \approx 0$ שלגביה ישנה התלבטות, היא קרובה יותר לשתי דוגמאות כחולות ולדוגמה אחת אדומה).	לא. קיימות דוגמאות (למשל הצמד העליון) שעבורן שתיים מתוך ה-3 השכנים הכי קרובים הם מהצבע השונה ולכן יסווגו בצורה שגויה.

iii.	כן $(y = x)$	לא. לא ניתן להעביר קו לינארי יחיד שיחלק את השדה באופן מדוייק, כל קו שכזה שננסה להעביר יחצה את המרחב לשתיים אך ישאיר שתי קבוצות של דוגמאות מסוגים שונים באותו הצד.	כן $(x = 0)$
iv.	כן	כן	כן

.3.2

i. עבור KNN עם m=1:

עבור מאגרי הנתונים A,B התשובה לא תשתנה ועבור מאגר C התשובה יכולה להשתנות.
עבור מאגר A, הכפלת התכונה x_1 בסקלר $\alpha > 0$ והשארת התכונה x_2 ללא שינוי תגרום לכך שהשכן הקרוב ביותר של כל דוגמה עדיין בצבע הנכון ולכן עדיין המודל מתאים במדוייק למאגר הנתונים.
עבור מאגר B, נקבל כי ה"מתיחה" של הציר שמודדד את x_1 לא תוכל לתקן את הסיווג השגוי של שתי הנקודות שנמצאות סביב $x = 3$ בסמוך לציר ה-y ולכן התשובה תשאר זהה.
עבור מאגר C, אם α יהיה גדול מספיק נקבל כי עבור כל זוג דוגמאות שלפני ההכפלה בסקלר היו הקרובות ביותר לא בהכרח יהיו הקרובות ביותר אלא דווקא כל שתי דוגמאות בגבהים שונים על ציר ה-y יהיו הקרובות ביותר ולכן נקבל כי המודל יכול להתאים באופן מדוייק למאגר C (כתלות בגודל של α).

ii. עבור KNN עם m=3:

עבור מאגרי הנתונים A התשובה לא תשנה ועבור המאגרים B,C כן עלולה להשתנות.
עבור מאגר A נקבל כי למרות ההכפלה של x_1 ב-α לא תשפיע על הסיווג, בגלל שבכל המאגר ישנן 4 דוגמאות בלבד ולכן ה-3 הקרובות ביותר לכל דוגמה ישארו זהות לכל α שנבחר, לכן התשובה עדיין תשאר שהסיווג שגוי - כלומר התשובה לא השתנתה.
עבור מאגר B, נקבל כי עבור $\alpha \rightarrow 0+$ ההשפעה של התכונה x_1 תבוטל ולכן הסיווג יערך רק לפי x_2, כלומר יתכן ודוגמאות מסויימות (לדוגמה הדוגמה האדומה סביב $x_1 \approx -4$ תהיה יותר קרובה לשלושת הדוגמאות הכחולות סביב $x_1 \approx 2.5$ ולכן תסווג בצורה שגויה, כלומר עבור ערכי α ששואפים ל-0 המודל לא בהכרח מתאים במדוייק למאגר נתונים זה.
עבור מאגר C, נקבל כי באותו אופן כמו עבור מודל i, עבור α מספיק גדולה נקבל כי הדוגמאות מאותו צבע יהיו יותר קרובות אחת לשנייה מאשר לזוג הסימטרי שלהן המשתקף סביב הציר של x_2, לכן התשובה עלולה להתשנות כתלות בערך של α.

iii. עבור שלושת המודלים (A,B,C) התשובה תשאר זהה ולא תשתנה עקב ההכפלה ב-α.
עבור מאגר A נקבל כי עדיין תהיה אפשרות להעביר קו ישר בעל משוואה הומוגנית שיחלק את הנתונים לשתי קבוצות באופן מדוייק ולכן התשובה לא תשתנה.
עבור מאגר B נקבל כי גם במקרה זה התשובה לא תשתנה, בגלל החלוקה של הדוגמאות ל-4 קבוצות מפורדות אשר מתחלקות בין הרביעים של מערכת הצירים לא קיים ערך של α שעבורו נוכל לקבל הפרדה בין שני סוגי הדוגמאות השונים לשתי קבוצות מופרדות לחלוטין ללא טעויות סיווג.
עבור מאגר C לכל α שנבחר עדיין נוכל לבחור בישר ההומגני $x = 0$ ולקבל הפרדה מוחלטות בין שני סוגי הדומגאות ולכן גם עבור מאגר זה התשובה תשאר זהה לסעיף 3.1

iv. עבור שלושת המודלים (A,B,C) התשובה תשאר זהה ולא תשתנה עקב ההכפלה ב-α.
עבור מאגר A נקבל כי נוכל להפריד בין שני סוגי הדוגמאות על ידי הפרדה לפי ערך כלשהו של x_2 ולכן נוכל לחלק את המאגר לשתי קבוצות אשר יתאימו לחלוקה האמיתית בצורה מדוייקת, כלומר התשובה יכולה להיות תלוייה רק ב-x_2 ולכן לא תושפע מהכפלת x_1 ב-α.
עבור מאגר B נקבל כי עדיין נוכל לחלק את הדוגמאות באופן מדוייק, לכל ערך של α שעבורו נכפול את התכונה x_1 נקבל כי נוכל להפריד את ארבעת הקבוצות של הדוגמאות באופן מדוייק לפי שתי החלוקות הבאות:
האם $x_1 > 0$ ולאחר מכן האם $x_2 > 0$, בגלל שידוע כי $\alpha > 0$ לא נוכל לבטל את החלוקה של הדוגמאות בין הצירים ולכן עדיין בעזרת שתי החלוקות שתוארו ניצור הפרדה מוחלטת בין ארבעת הקבוצות, כלומר עדיין נוכל לחלק את הדוגמאות במאגר הנתונים בצורה מדוייקת בעזרת המודל הנתון.
עבור מאגר C, בדומה ל-B נקבל כי נוכל להפריד את כל הדוגמאות ולתייגן במדוייק לפי האם $x_1 > 0$ או לא, הפרדה זו תחלק את המאגר לשתי קבוצות שמתאימות במדוייק לסוגן של הדוגמאות במאגר, ולכן גם הכפלה בכל קבוע $\alpha > 0$ לא תשנה את הסיווג של הנקודות בחלוקה, כלומר עדיין נוכל להתאים את הדוגמאות לסוגן בעזרת המודל התנתון.