תרגיל קצר מספר 2 - מבוא למערכות לומדות

עומר שמחי - 316572593

22 באפריל 2021

1. עצי החלטה

נראה בסיס נתונים כל שהפעלה של $ID3$ עליו תניב עץ מעומק 3 ואילו בחירה אחרת של שאלות מפצלות, תינתן לנו עץ החלטה מעומק 2 בדיוק. להלן הטבלה המתארת את בסיס הנתונים:

ID	$Fever$	$Cough$	$Smell\ loss$	$corona$
1	F	T	F	F
2	F	T	F	F
3	T	T	T	T
4	T	F	F	F
5	T	T	T	T
6	T	T	T	T
7	T	T	F	T

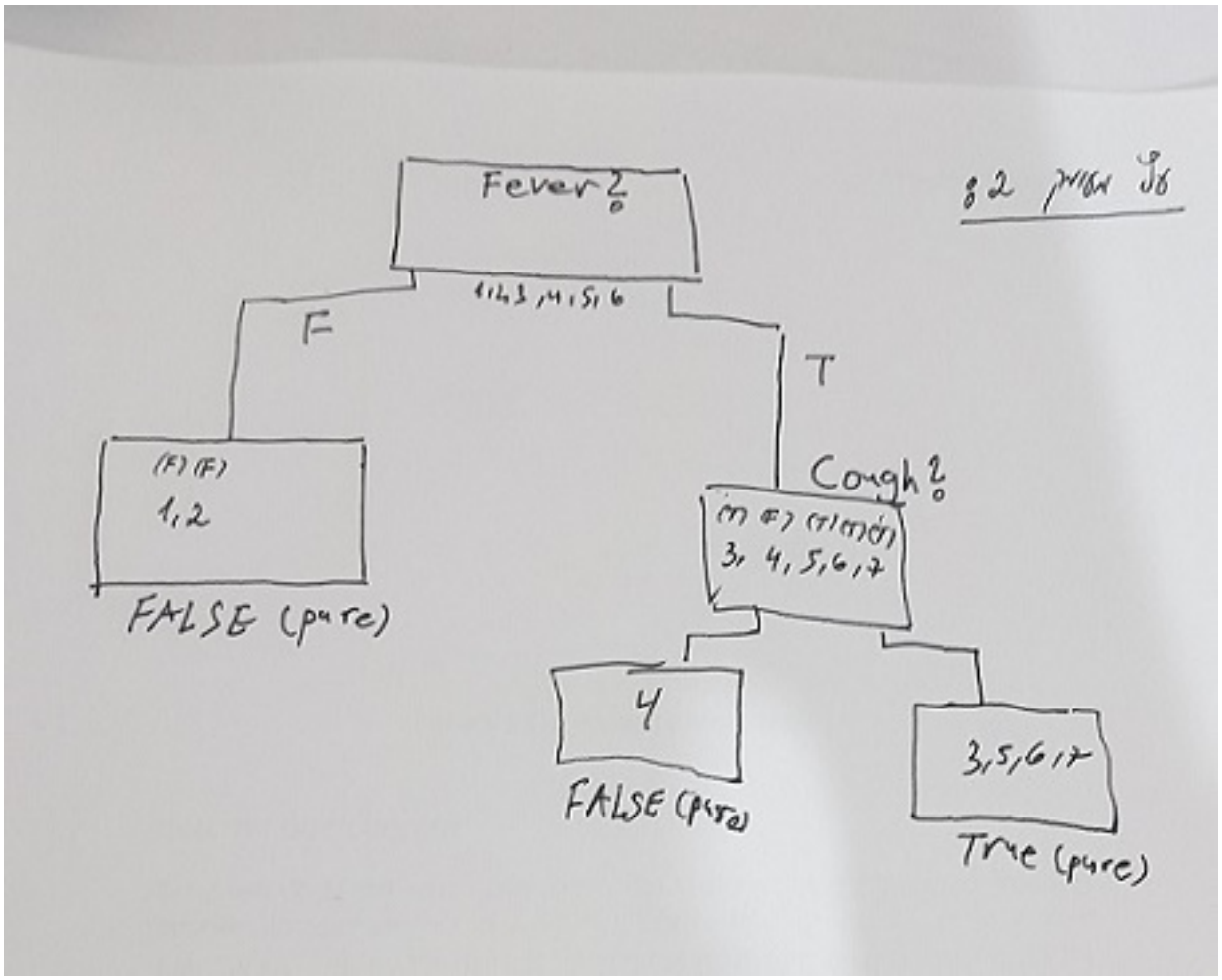
כעת נפריד לשני עצי ההחלטות:

1. **עץ ראשון - בחירת שאלות ללא $ID3$:**

   (א) שאלה על הצומת הראשונה - $Fever?$

   (ב) שאלה על הצומת שאיננה $pure$ - $Cough?$.

   נקבל לפי נתונים הטבלה את העץ הבא:



2. **עץ שני - בחירת שאלות לפי האלגוריתם הגרידי** $ID3$**:**

נציג את החישוב המלא (לפי נתונים הטבלה) ואז את העץ המתקבל.

$attribute$	$\frac{\lvert v_{a=T}\rvert}{\lvert v\rvert}$	$\frac{\lvert v_{a=F}\rvert}{\lvert v\rvert}$	$H\left(v_{a=T}\right)$	$H\left(v_{a=F}\right)$	$IG\left(v,a\right)-H\left(v\right)$
$Fever$	$\frac{4}{7}$	$\frac{3}{7}$	$H\left(\frac{3}{4}\right)$	$H\left(\frac{1}{3}\right)$	$-\frac{4}{7}H\left(\frac{3}{4}\right)-\frac{3}{7}H\left(\frac{1}{3}\right)$
$Smell\ loss$	$\frac{3}{7}$	$\frac{4}{7}$	0	$H\left(\frac{14}{4}\right)$	$-\frac{4}{7}H\left(\frac{3}{4}\right)=-0.472\,(*)$
$Cough$	$\frac{6}{7}$	$\frac{1}{7}$	$H\left(\frac{4}{6}\right)$	$H\left(1\right)=0$	$-\frac{6}{7}H\left(\frac{4}{6}\right)=-0.787\,(**)$

$$H\left(\frac{1}{4}\right)=-\frac{1}{4}\log\frac{1}{4}-\frac{3}{4}\log\frac{3}{4}=0.826\quad(*)$$

$$H\left(\frac{4}{6}\right)=-\frac{4}{6}\log\frac{4}{6}-\frac{2}{6}\log\frac{2}{6}=0.918\quad(**)$$

וכן:

$$H\left(\frac{1}{4}\right) = H\left(\frac{3}{4}\right)$$

היות שהאנטרופיה סמטרית סביב $x = \frac{1}{2}$. כלומר $IG(v,a) - H(v)$ של $Smell\ \ loss$ הוא הכי גדול ומכאן היות ש־$H(v)$ קבוע אצל כולם נסיק כי $IG_{max} = IG_{Smell\ \ loss}$ ולכן נפצל את הצומת הראשון לפי $Smell\ \ loss$.

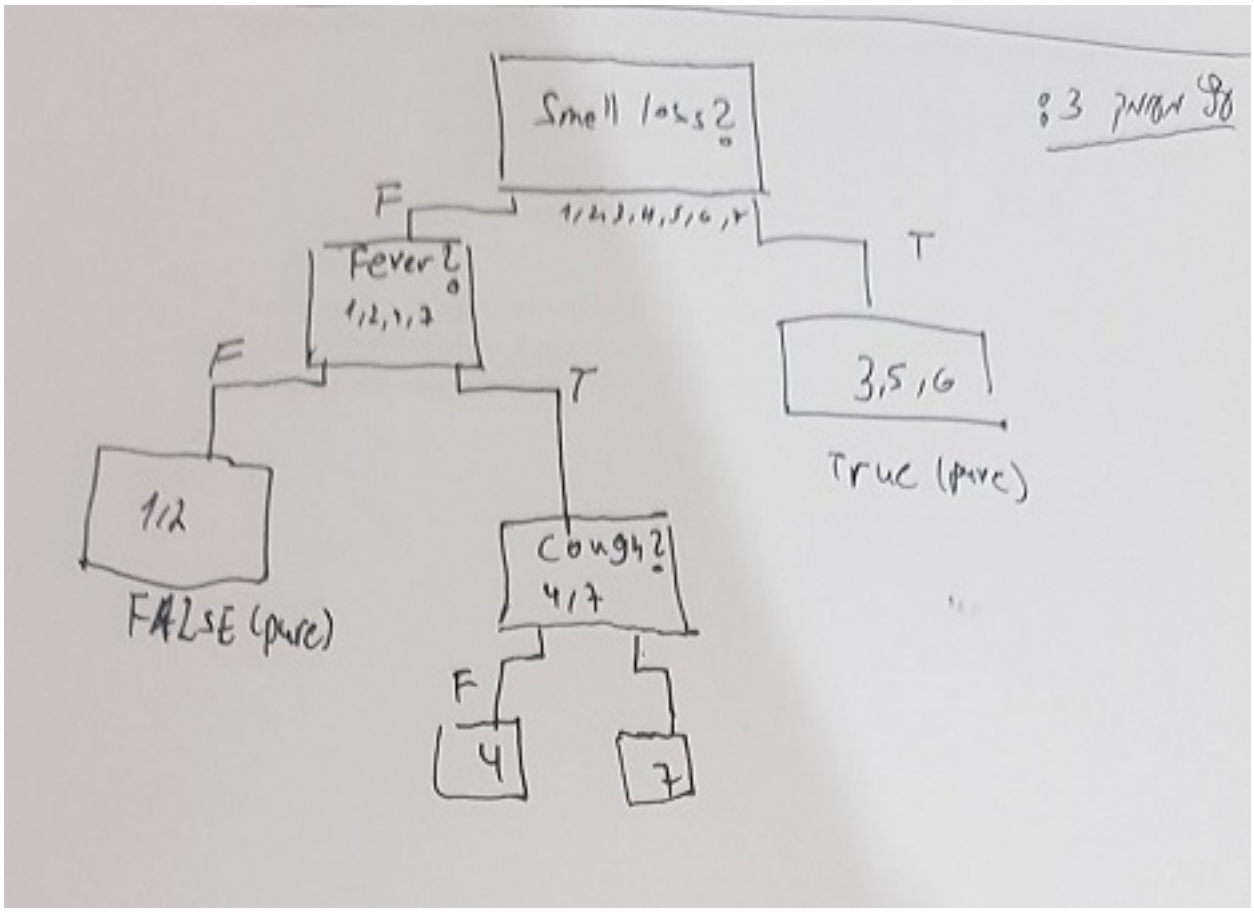
באופן דומה נמשיך לשלב הבא (בהתאם לפיצול שקיבלנו בעץ ־ מוצג בעץ הסופי למטה. יתר על כן, למעשה מתקבל אותו חישוב שהיה בתרגול).

$attribute$	$\frac{\lvert v_{a=T}\rvert}{\lvert v\rvert}$	$\frac{\lvert v_{a=F}\rvert}{\lvert v\rvert}$	$H(v_{a=T})$	$H(v_{a=F})$	$IG(v,a) - H(v)$
$Fever$	$\frac{1}{2}$	$\frac{1}{2}$	$H\left(\frac{2}{4}\right) = H\left(\frac{1}{2}\right)$	0	$-\frac{1}{2}H\left(\frac{1}{2}\right) = -\frac{1}{2}$
$Cough$	$\frac{3}{4}$	$\frac{1}{4}$	$H\left(\frac{2}{6}\right) = H\left(\frac{1}{3}\right)$	0	$-\frac{3}{4}H\left(\frac{1}{3}\right) = -0.689\,(*)$

כאשר:

$$H\left(\frac{1}{3}\right) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = 0.918 \quad (*)$$

כלומר $IG(v,a) - H(v)$ של $Fever$ הוא הכי גדול ומכאן היות ש־$H(v)$ קבוע אצל כולם נסיק כי $IG_{max} = IG_{Fever}$ ולכן נפצל בפיצול השני לפי $Fever$.



סה"כ קיבלנו שעדיף לשאול שאלות לפיצול על פיצ'רים **לא לפי** $ID3$ במקרה הנ"ל, כי מקבלים בצורה זו עץ החלטות עם עומק קטן יותר.

<u>תשובה עבור 1.4:</u> עבור הגבלה על הגובה $max\_depth = 2$ נקבל **שגיאה אמפירית**:

$$Empirical\_Error = \frac{1}{7} \cdot \sum_{i=1}^{7} 1_{\{y_i \neq h(x_i)\}} = \frac{1}{7}$$

2. ספרביליות

2.1

1. kNN עם $k = 1$.

   פתרון - רק מערך הנתונים (C).

2. kNN עם $k = 3$.

   פתרון - רק מערך הנתונים $(A) + (C)$.

3. kNN עם $k = m - 1$.

   פתרון - אף אחד לא ייתן שגיאת אימון 0.

4. SVM ליניארי.

   פתרון - (A) בלבד.

5. עץ החלטות ללא קריטריון עצירה.

   פתרון - על $(A) + (B) + (C)$.

6. עץ החלטות עם לכל היותר 2 רמות.

   פתרון - אף אחד לא ייתן שגיאת אימון 0.

7. עץ החלטות עם לכל היותר 4 רמות.

   פתרון - שגיאת אימון 0 על (B).

2.2

1. kNN עם $k = 1$.

   פתרון - **התשובה לא תשתנה**, המרחק בין הנקודות נשאר אותו דבר ולכן kNN יתנהג באותו אופן.

2. kNN עם $k = 3$.

   פתרון - **התשובה לא תשתנה**, המרחק בין הנקודות נשאר אותו דבר ולכן kNN יתנהג באותו אופן.

3. kNN עם $k = m - 1$.

   פתרון -**התשובה לא תשתנה**, המרחק בין הנקודות נשאר אותו דבר ולכן kNN יתנהג באותו אופן.

4. SVM ליניארי.

פתרון - **התשובה לא תשתנה.** הכפלה במטריצת סיבוב לא משנה את המיקומים היחסיים ואת המרחקים בין הנקודות ולכן עדיין רק ב־(A) יש מפריד יחיד.

5. עץ החלטות ללא קריטריון עצירה.

פתרון - **התשובה לא תשתנה.** אין הגבלה על גובה העץ ולכן נוכל לבנות תמיד עץ החלטות עם שגיאת אימון 0 (כלומר תמיד נוכל לפצל עד שגיאת אימון 0).

6. עץ החלטות עם לכל היותר 2 רמות.

פתרון - **התשובה יכולה להשתנות.** עבור מערך הנתונים (A), נשים לב כי קיים לו מפריד ליניארי (בזווית פחות או יותר $180° - 45° = 135°$) ולכן אם נזיז (ע"י הכפלה במטריצת סיבוב מתאימה) למשל ב־$-45°$ נקבל מפריד יחיד שיהפוך את הדטה לספרבילי לחלוטין (מפריד שמקביל לציר ה־y). מכאן של־(A) יהיה עץ החלטות שדרגתו לא עולה על 2 וזה משנה את התשובה מסעיף 6 של 2.1.

7. עץ החלטות עם לכל היותר 4 רמות.

פתרון - **התשובה יכולה תשתנה.** אותו נימוק כמו בסעיף הקודם (סעיף 6) עבור (A) (כלומר כעת נצליח גם עבור (A)).