שאלה 6: MDP

100	0	0
-50	0	S

נתון ה MDP בציור: הagent נמצא בS. אם יגיע לאחד ממשבצות שבשורה הימנית המשחק יסתיים. עבור כל נסיון של מהלך ממשבצת למשבצת סמוכה העלות היא 5. ההנחה היא שעבור כל תזוזה יש 75% סיכוי להצליח ו10% סיכוי להזרק לשני שני הצדדים ו5% סיכוי להזרק אחורה. אם ה agent נתקע בקיר, הוא נשאר באותה המשבצת.

עליכם לבצע שתי איטרציות של Value Iteration.

יש לרשום את ה utility אחרי האיטרציה הראשונה של כל משבצת ריקה

יש לרשום את ה utility אחרי האיטרציה השנייה של כל משבצת ריקה וכן לרשום מהי ה policy האופטימאלי (בצורת חץ) לאחר שתי האיטרציות בכל משבצת ריקה.

שאלה 6: MDP

100	0	0
-50	0	S

נתון ה MDP בציור: הagent נמצא בS. אם יגיע לאחד ממשבצות שבשורה הימנית המשחק יסתיים. עבור כל נסיון של מהלך ממשבצת למשבצת סמוכה העלות היא 5. ההנחה היא שעבור כל תזוזה יש 75% סיכוי להצליח ו10% סיכוי להזרק לשני שני הצדדים ו5% סיכוי להזרק אחורה. אם ה agent נתקע בקיר, הוא נשאר באותה המשבצת.

עליכם לבצע שתי איטרציות של Value Iteration.

יש לרשום את ה utility אחרי האיטרציה הראשונה של כל משבצת ריקה

100	70	-5
-50	-7.5	-5

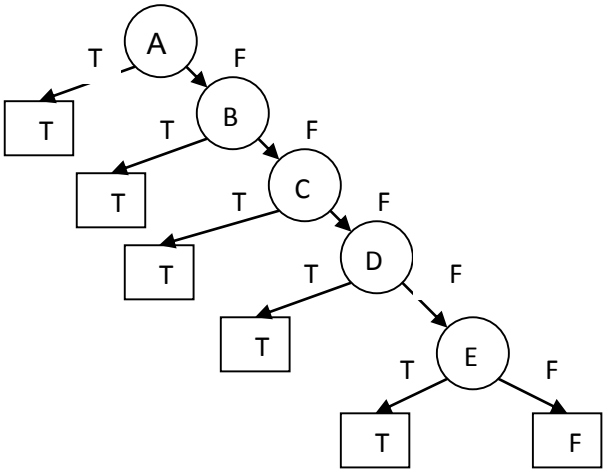
יש לרשום את ה utility אחרי האיטרציה השנייה של כל משבצת ריקה וכן לרשום מהי ה policy האופטימאלי (בצורת חץ) לאחר שתי האיטרציות בכל משבצת ריקה.

100	76	46.25
-50	41.625	-10.125

עץ החלטה

א. צייר עץ החלטה (decision tree) המתאר את הנוסחה הבוליאנית הבאה: (6)

$$A \vee B \vee C \vee D \vee E$$



נתונה טבלת האימון הבאה:

A	B	C	Target
T	T	F	F
F	T	F	T
F	T	T	T
F	F	T	F
T	F	F	F

א. צייר עץ החלטה לקבוצת הדוגמאות. אם קיימות מספר תכונות עם אותו gain יש לבחור לפי סדר הא"ב. יש לציין בקודקודי העץ את ערכי ה-gain הרלוונטיים.

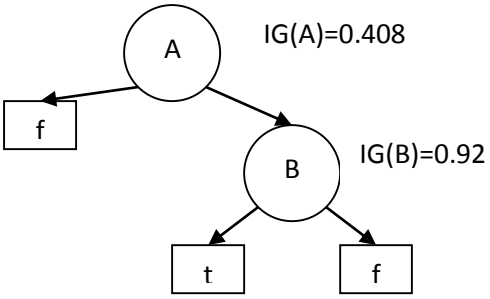
ב. מהי הפונקציה הבוליאנית המיוצגת בעץ שמצאת

<u>תשובה:</u>

א.

IG(A)=I(2/5,3/5)-(2/5*I(0/2,2/2)+3/5*I(2/3,1/3))=0.96-(2/5*0+3/5*0.92)=0.96-0.552=0.408

IG(B)=I(2/3,1/3)-(2/3*I(2/2,0/2)+1/3(0/1,1/1))=0.92



ב.(-A^B)

<u>שאלה 2</u>: <u>עצי החלטה</u>

הנח כי נתונות לך הדוגמאות הבאות. F1, F2 F3 יכולים לקבל את הערכים a,b,c,d.

	F1	F2	F3	Output
ex1	b	d	c	+
ex2	c	d	b	-
ex3	c	a	c	+
ex4	b	a	b	-
ex5	a	b	a	+
ex6	a	b	a	-
ex7	d	c	c	+

א) מה ה- Gain ש- information gain יחשב לכל אחד מה- features - F1, F2 ו- F3?

מה ה- feature שייבחר?

תשובה:

נחשב את ה- Gain הכללי- I(4/7,3/7) = 0.9852

IG(F1) = I(4/7,3/7) – (2/7 * I(1/2,1/2) +2/7 * I(1/2,1/2) + 2/7 * I(1/2,1/2) + 1/7 * I(1,0)) = 0.9852 – 6/7 = 0.1280

IG(F2) = I(4/7,3/7) – (2/7 * I(1/2,1/2) + 2/7 * I(1/2,1/2) + 1/7 * I(1,0) + 2/7 * I(1/2,1/2)) = 0.1280

IG(F3) = I(4/7,3/7) – (2/7 * I(1/2,1/2) + 2/7 * I(0,1) + 3/7 * I(1,0) + 0) = 0.6994

ולכן F3 ייבחר.

ב) בלי קשר לתשובתך ב- א), הנח כי F1 נבחר להיות בקודקוד השורש.

הראה כיצד יראו הקריאות הרקורסיביות של אלגוריתם ה- Decision Tree [רמה אחת]. הראה מה הפרמטרים שישלחו (אבל אל תבצע את החישובים).

תשובה:

עבור v1=a DTL({aba+,aba-},{F2,F3},+)

עבור v2=b DTL({bdc+,bab-},{F2,F3},+)

עבור v3=c DTL({cdb-,cac+},{F2,F3},+)

עבור v4=d DTL({dcc+},{F2,F3},+)

ג) בלי קשר לתשובתך ב- א) ו ב), הנח כי F3 נבחר להיות בקודקוד השורש.

הראה כיצד יראו הקריאות הרקורסיביות של אלגוריתם ה- Decision Tree [רמה אחת]. הראה מה הפרמטרים שישלחו (אבל אל תבצע את החישובים).

תשובה:

עבור v1=a DTL({aba+,aba-},{F1,F2},+)

עבור v2=b DTL({cdb-,bab-},{F1,F2},+)

עבור v3=c DTL({bdc+,cac+,dcc+},{F1,F2},+)

עבור v4=d DTL({},{F1,F2},+)

נתונה טבלה ובה נתוני אימון. מוגדרים שלושה מאפיינים F1,F2 ו-F3. לכל מאפיין ישנם שלושה ערכים אפשריים: a,b או c.

א. עליכם לבנות עץ החלטה המייצג את מאפיין הפלט output על סמך האלגוריתם שנלמד בשיעור. אם קיימות מספר תכונות עם אותו gain f1 עדיף על f2 ו-f2 עדיף על f3. ערך ה-default כאשר אין דוגמא הוא false.
ב. ציינו בקודקודי העץ הרלוונטיים את ה-gain של התכונה שבחרתם.
ג. כתבו את הנוסחה הלוגית המתארת את העץ שקבלתם.

טבלת הנתונים:

#	f1	f2	f3	output
1	b	b	a	t
2	a	a	a	f
3	b	a	a	f
4	b	c	a	t
5	c	b	a	f
6	a	c	b	f
7	b	a	c	t
8	a	a	c	f
9	c	b	c	f

$$I(\frac{p}{p+n},\frac{n}{p+n}) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

$$remainder(A) = \sum_{i=1}^{v}\frac{p_i+n_i}{p+n} I(\frac{p_i}{p_i+n_i},\frac{n_i}{p_i+n_i})$$

$$IG(A) = I(\frac{p}{p+n},\frac{n}{p+n}) - remainder(A)$$

I(3/9,6/9)=0.92

Reminder(f1)=3/9*0+4/9*I(3/4,1/4)+2/9*0=4/9*0.81=**0.36**

Reminder(f2)=4/9*I(1/4,3/4)+3/9*I (1/3,2/3)+2/9*I(1/2,1/2)=4/9*0.81+3/9*0.92+2/9*1

=**0.889**

Reminder(f3)=5/9*I(2/5,3/5)+1/9*I(0,1)+3/9*I(1/3,2/3)=5/9*0.97+0+3/9*0.92=**0.845**

IG(f1)=0.92-0.36=**0.56**

#	f1	f2	f3	output
1	B	b	a	t
3	B	a	a	f
4	B	c	a	t
7	B	a	c	t

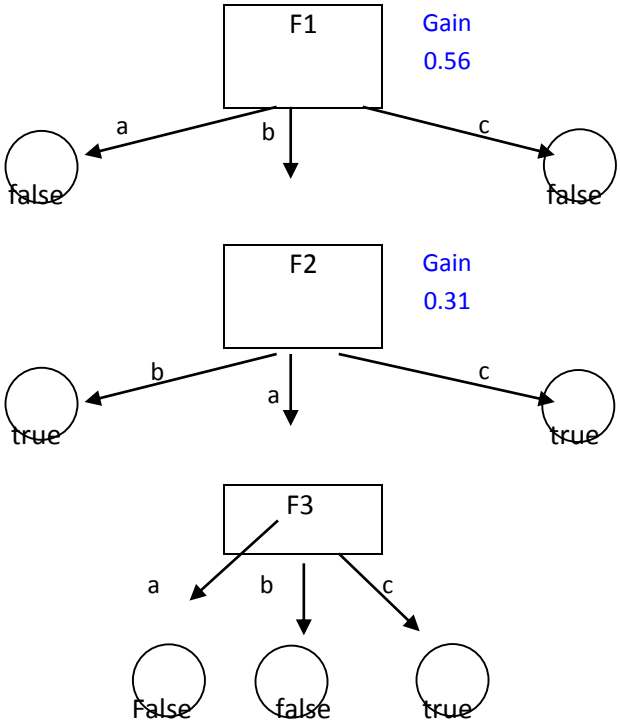
I(3/4,1/4)=0.81

Remainder(f2)=2/4*I(1/2,1/2)+1/4*I(1,0)+1/4*I(1,0)=0.5+0+0=**0.5**

Remainder(f3)=3/4*I(2/3,1/3)+1/4*I(1,0)=3/4*0.92=**0.69**

IG(f2)=0.81-0.5=0.31

#	f1	f2	f3	output
3	B	a	a	f
7	B	a	c	t



Bollean Function:

f1=b^{(f2=b)V(f2=c)V[(f2=a)^(f3=c)]}

2. (תשס"ג ב') נתונה סוכנות הכרויות עם database גדול של אנשים. על מנת לספק שירות יעיל אתם מעוניינים לבנות מודל לכל לקוח חדש בהתבסס על קבוצה מייצגת קטנה, אותה הלקוח כבר קטלג (ה-training set). ברגע שהמודל נלמד ניתן לסנן עבורו את ה-database ולשלוח רק את המועמדים הרלוונטיים.

נתונה קבוצת הדוגמאות הבאות ומטרתכם ללמוד את הפונקציה Candidate המחזירה yes / no, המציין את פוטנציאל המועמד ללקוח.

i	*Sex*	*Status*	*Age*	*Education*	*Location*	*Looks*	*Candidate*
1	M	Single	<30	HS	North	Good	No
2	F	Single	30-40	BA	Center	Good	Yes
3	F	Single	>40	MA	South	Great	No
4	M	Divorced	30-40	MA	South	Great	No
5	F	Single	<30	MA	North	Great	No
6	F	Single	30-40	HS	Center	Good	No
7	F	Single	30-40	BA	South	Good	Yes
8	F	Single	>40	HS	Center	Great	No
9	F	Divorced	30-40	MA	South	Great	Yes
10	F	Divorced	<30	HS	South	Good	No
11	F	Single	>40	BA	Center	Good	No
12	F	Divorced	<30	MA	Center	Great	No

ציירו עץ החלטה לקבוצת הדוגמאות. אם קיימות מספר תכונות עם אותו gain יש לבחור לפי סדר הא"ב. יש לציין בקודקודי העץ את ערכי ה-gain הרלוונטיים.

<u>פתרון</u>:

$I(3/12,9/12)=-(1/4)\log(1/4)-(3/4)\log(3/4)=0.5+0.311=0.811$

$Remainder(Age)=(4/12)*I(0/4,4/4) + (5/12)*I(3/5,2/5)+(3/12)*I(0/3,3/3)$

$= 0+(5/12)*[-0.6\log(0.6)-0.4\log(0.4)]+0=0.405$

$IG(Age)=I(3/12,9/12) - Remainder(Age) =0.811-0.405=0.41$

בצורה דומה נבדוק את שאר המאפיינים. המאפיין AGE נבחר כמאפיין בעל ה- IG הגדול ביותר.

נחלק את טבלת הדוגמאות על פי ערכי AGE האפשריים. אם גיל המתמודד קטן מ-30 או גדול מ-40 לכל הדוגמאות יש את אותו הערך ולכן נגדיר אותם כעלים. עתה נותרנו עם כל הדוגמאות בהן גיל המתמודד הוא בין 30 ל-40:

I	*Sex*	*Status*	*Age*	*Education*	*Location*	*Looks*	*Candidate*
2	F	Single	30-40	BA	Center	Good	Yes
4	M	Divorced	30-40	MA	South	Great	No
6	F	Single	30-40	HS	Center	Good	No
7	F	Single	30-40	BA	South	Good	Yes
9	F	Divorced	30-40	MA	South	Great	Yes

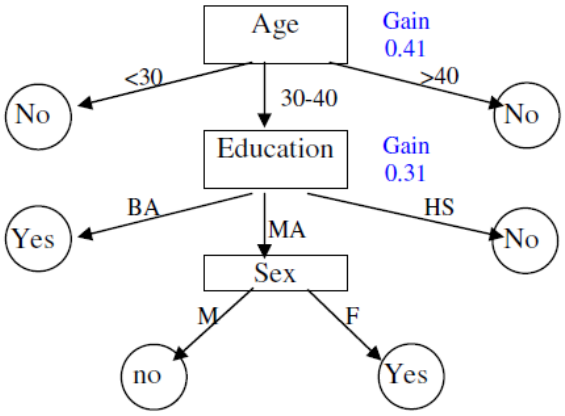
על מנת לבחור את המאפיין לתת העץ נחשב עבור כל מאפיין את IG:

$IG(Education)=I(3/5,2/5)-Reminder(Education)$

$I(3/5,2/5)=-0.6\log(0.6)-0.4\log(0.4)=0.442+0.529=0.971$

Remainder(Education)= (1/5)I(0/1,1/1)+(2/5)I(2/2,0/2)+(2/5)I(1/2,1/2)=0+0+0.4=0.4
IG(Education)=0.971-0.4=0.571

העץ המתקבל הוא:



3. (תשס"א א') נתונות הדוגמאות הבאות לעץ החלטה בנושא של האם להצביע בבחירות או לא. הפרמטרים הם מזג אוויר, מידת תמיכה, וסקרים.

החלטה	סקרים	מידת תמיכה	מזג אוויר
לא	תיקו	מועטה	סוער
כן	שלי מוביל	גדולה	בהיר
כן	שלי מוביל	גדולה	מעונן
לא	יריב מוביל	גדולה	סוער
כן	תיקו	מועטה	מעונן
לא	תיקו	גדולה	סוער
לא	שלי מוביל	מועטה	בהיר
כן	יריב מוביל	מועטה	בהיר
לא	שלי מוביל	מועטה	מעונן
כן	יריב מוביל	מועטה	מעונן

א. בנו עץ החלטה מתאים, במקרה שה-gain שווה בחרו עפ"י סדר הא"ב.
ב. ציינו בקודקודי העץ הרלוונטיים את ה-gain של התכונה שבחרתם.
ג. תארו נוסחה לוגית שיכולה לתאר את העץ שקבלתם.

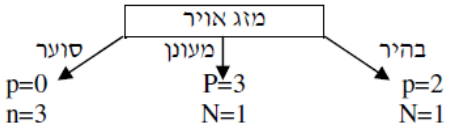
<u>פתרון</u>:

א+ב. בניית עץ החלטה לשאלה האם להצביע בבחירות:

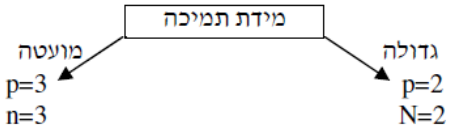
נחשב את האנטרופיה ההתחלתית –

$p=5, n=5 \rightarrow I(1/2, 1/2) = 1$

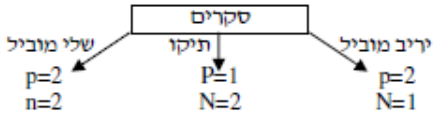
נבחר תכונה לשורש עץ ההחלטה:



$Gain(\text{מזג אויר}) = 1 - (0+4/10 * I(3/4,1/4) + 3/10 * I(2/3,1/3)) = 0.4$



Gain(מידת תמיכה)=0

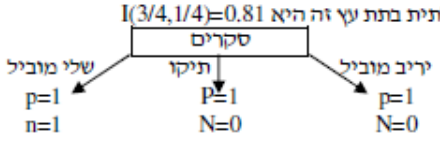


Gain(סקרים) =1 - (4/10 * I(1/2, 1/2) + 3/10 * I(1/3,2/3) + 3/10 * I(2/3,1/3))=0.05

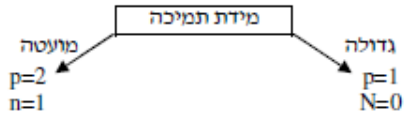
ה-Gain הכי גדול הוא של התכונה "מזג אויר" ולכן היא תהיה שורש העץ.

נבחר תכונה עבור תת העץ השמאלי בו מזג האויר הוא מעונן:

האנטרופיה ההתחלתית בתת עץ זה היא I(3/4,1/4)=0.81

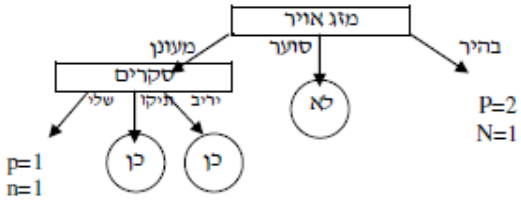


Gain(סקרים) =0.81 - (2/4*I(1/2,1/2) + 0)=0.31

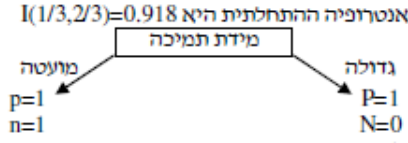


Gain(מידת תמיכה) =0.81 - (3/4*I(2/3,1/3)+0) =0.81-0.69=0.12

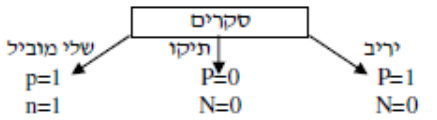
שורש תת העץ יהיה אם כן "סקרים"



נבחר תכונה לתת העץ הימני. האנטרופיה ההתחלתית היא I(1/3,2/3)=0.918

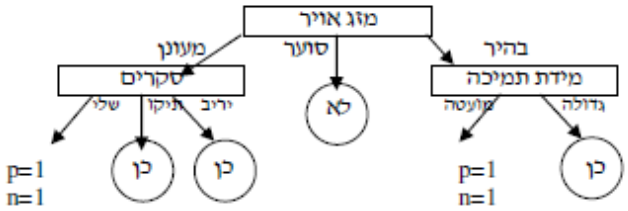


Gain(מידת תמיכה) =0.918 - (2/3+0) = 0.251

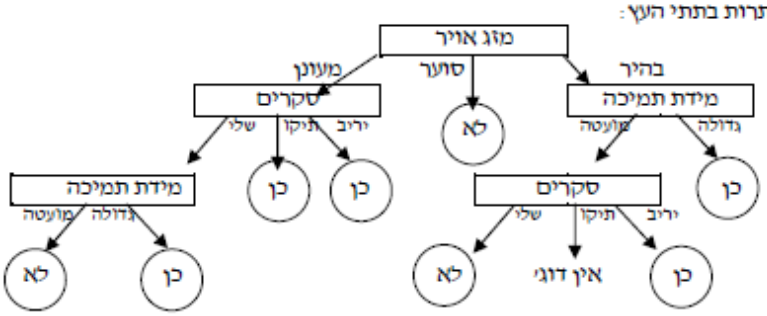


Gain(סקרים) =0.918 - (2/3+0)=0.251

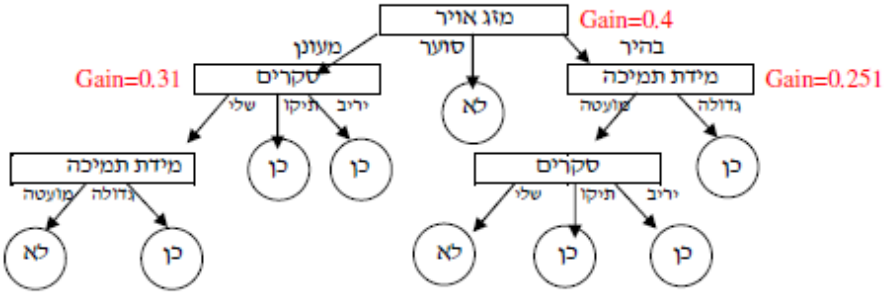
שורש תת העץ יהיה אם כן "מידת תמיכה", שכן במקרה שה- Gain שווה יש לבחור לפי סדר הא"ב. עד כה העץ הוא:



נציב את התכונות הנותרות בתתי העץ:



היות ואין דוגמאות עבור סקרים=תיקו ולא ניתן לנו ערך ברירת מחדל, נבחר עפ"י הרוב "כן", והעץ המתקבל עם ערכי ה-Gain:



ג. נסמן ב- W את מזג האוויר, T את מידת התמיכה ו- S את הסקרים. הנוסחא שמתארת את העץ היא:

((W= בהיר) ^ (T= גדולה)) ∨ ((W= בהיר) ^ (T= מועטה) ^ ~(S= שלי))

((W= מעונן) ^ ~(S= שלי)) ∨ ((W= מעונן) ^ (T= גדולה))

4. ענו נכון / לא נכון, ונמקו בקצרה:

א. *ניתן לייצג כל פונקציה בוליאנית עם k משתנים (attributes) ע"י עצי החלטה עם k קודקודים פנימיים (כולל השורש).*

ב. *אם לא פצלנו את עץ ההחלטה עפ"י התכונה המפרידה ביותר בכל שלב, נקבל פתרון שלא יתאים לכל הדוגמאות של הקלט.*

ג. *במידה והייתה שגיאה בדוגמאות, גם עץ ההחלטה שיתקבל יהיה שגוי.*

פתרון: לא נוצר בהכרח אותו עץ. בעץ החדש אי אפשר להשתמש במאפיין A ולכן יווצרו שלושה תתי עצים זהים. עץ זהה יוצר אם למאפיין A אין משמעות כלומר IG(A)=0 (ויש מספיק דוגמאות).

שאלה

בטבלה שלהלן מופיעות רשומות של 12 פציינטים. לכל אחד מהם יש נתונים עבור המאפיינים הבאים: מין, גיל (האם מעל 60), האם הוא סכרתי, האם יש לו לחץ דם גבוה, האם ה- EKG שלו חורג מהנורמה, וסיווג (classification) - האם הפציינט סובל מהפרעה בקצב הלב?

פציינט	מין	מעל 60	סכרתי	לחץ דם גבוה	EKG	הפרעה בקצב הלב
1	ז	+	+	-	-	-
2	ז	-	-	+	+	+
3	ז	-	+	+	-	-
4	ז	+	-	-	+	+
5	ז	+	+	+	-	+
6	ז	-	+	+	-	+
7	נ	-	-	+	-	-
8	נ	+	+	+	+	+
9	נ	-	+	-	+	+
10	נ	+	-	-	-	-
11	נ	+	+	-	-	-
12	נ	+	-	+	+	+

א. חשבו את האנטרופיה : H(HasArrhythmia|Sex=Female)

(השתמשו בנוסחה המופיעה בראש עמוד 704 בספר הלימוד).

$$H(\text{HasArrhythmia}|\text{Sex} = \text{Female}) = -(0.5\log_2 0.5 + 0.5\log_2 0.5) = 1$$

ב. מה היא התכונה שתיבחר להיות בשורש עץ ההחלטה?

הניחו כי:

Cost(Sex)=Cost(Over60)=1

Cost(Diabetic)=3

Cost(HighPulse)=2

Cost(AbnormalEKG)=5

נשתמש בביטוי $A = \frac{Gain^2(S,A)}{Cost(A)}$ כדי לבחור מי יהיה בשורש עץ ההחלטה. בדוגמא עגלתי לשתי ספרות אחרונות.

gain(Sex)=0.98 - 0.5 * E(3, 3) - 0.5 * E(4, 2) = 0.02

A = 0.02^2 / 1.0 = 0

gain(over 60)=0.98 - 0.583 * E(4, 3) - 0.417 * E(3, 2) = 0

A = 0^2 / 1.0 = 0

gain(Diabetic)=0.98 - 0.583 * E(4, 3) - 0.417 * E(3, 2) = 0

A = 0^2 / 3.0 = 0

gain(HighPulse)=0.98 - 0.5 * E(5, 2) - 0.5 * E(2, 5) = 0.12

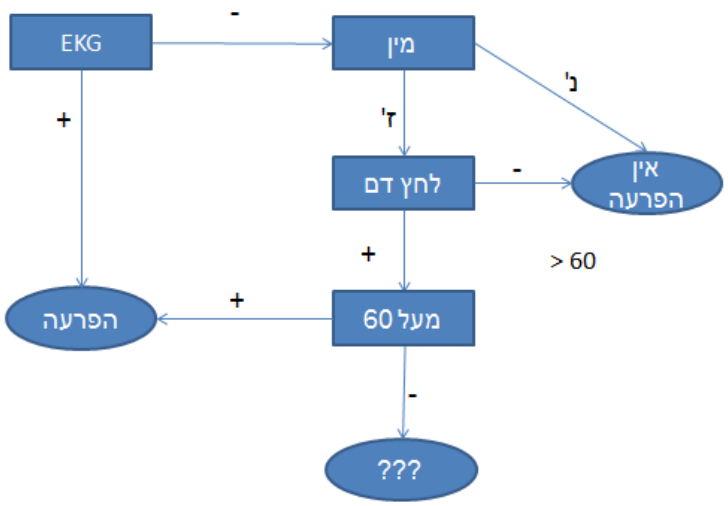
A = 0.12^2 / 2.0 = 0.01

gain(AbnormalEKG)=0.98 - 0.417 * E(5, 0) - 0.583 * E(2, 5) = 0.48

A = 0.48^2 / 5.0 = 0.05

EKG קיבל את התוצאות הטובות ביותר ולכן הוא יהיה בשורש עץ ההחלטה.

ג. בנו עץ החלטה לחיזוי קבלה לאוניברסיטה. פרטו את כל שלבי הבניה.



ד. נניח כי עבור קבוצה אחרת של פציינטים ידוע גילם המדוייק.

גילאי הדוגמאות החיוביות (positive examples) הם: {40,60,62,64,70,74,75,82}

וגילאי הדוגמאות השליליות (negative examples) הם {33,35,42,45,49,52,58,59,80}.

נניח שכל יתר התכונות ב-data set פחות "טובות" מתכונת הגיל , כך שנרצה לפצל את העץ

ב-Age=k על-ידי חלוקת קבוצת הדוגמאות לשתי תת-קבוצות: Age≥k ו-Age<k.

בהתבסס על תוספת אינפורמציה (information gain), איזו חלוקה נרצה לבחור ?

הדוגמאות השליליות מסומנות באדום.

33, 35, 40, 42, 45, 49, 52, 58, 59, 60, 62, 64, 70, 74, 75, 80, 82

מהשרטוט קל לראות שאם נשים את הגבול בין 59-60 נהיה זקוקים למספר הנמוך ביותר של ביטים כדי לסווג דוגמה חדשה, לכן הreminder יהיה הנמוך ביותר ולכן הgain יהיה מקסימלי.