

# תרגול 4 - K-NN ו Decision trees

Slides

PDF

Code

## תקציר התיאוריה

### בעיות סיווג

בעיות סיווג הם בעיות supervised learning שבהם labels (תוויות) מוגבלות לסט סופי של ערכים.

- בבעיות סיווג נהוג להתייחס לחזאי כאל מסווג (classifier) או discriminator (מקטלג).
- את הערכים השונים שאותם התוויות יכול לקבל מכנים **מחלקות**.
- את מספר המחלקות נסמן בקורס ב  $C$ .
- בעיות סיווג שבהם יש רק 2 מחלקות,  $C = 2$ , מכוונות בעיות **סיווג בינארי**.

• בסיווג בינארי, מקובל להשתמש באחת מהאופציות הבאות לסימון המחלקות:

- $y \in \{0, 1\}$
- $y \in \{-1, 1\}$

• בסיווג לא בינארי, מקובל להשתמש באחת מהאופציות הבאות לסימון המחלקות:

- $y \in \{1, 2, \dots, C\}$
- $y \in \{0, 1, \dots, C - 1\}$

דוגמאות:

- מערכת לזיהוי הונאות בכרטיסי אשראי (המקרה בו אדם לא מורשה משתמש בכרטיס או מספר אשראי של אדם אחר). במקרה זה  $x$  יכול להיות וקטור אשר מכיל את מאפייני העיסקה, כגון מחיר, שעה, ומיקום, ו  $y$  יקבל אחד משני ערכים:

- 0 - העסקה לגיטימית.
- 1 - חשד להונאה.

- מערכת לעיבוד כתב יד (OCR). במקרה זה  $x$  יכול להיות לדוגמה תמונה של אות ו  $y$  יהיה שווה למחלקה אשר מייצגת את האות בתמונה:

- a :1
- b :2
- c :3
- ...

### Misclassification rate

לרוב בבעיות סיווג לא תהיה משמעות למרחק בין החיזוי  $\hat{y}$  לערך האימיתי של  $y$ . לדוגמה, בניסיון לזהות את האות  $g$ , חיזוי של האות  $f$  (אשר מופיעה בצמוד ל  $g$  באלף בית) הוא לא בהכרח חיזוי טוב יותר מ  $q$  (אשר נמצא רחוק יותר).

לכן לרוב בבעיות סיווג נפוץ להשתמש ב misclassification rate כפונקציית מחיר. פונקציית מחיר זו הינה מסוג פונקציית risk אשר משתמשת ב zero-one-loss המוגדר באופן הבא:

$$l(\hat{y}, y) = I\{\hat{y} \neq y\}$$

פונקציית ה misclassification rate אם כן נראית כך:

$$R(h) = \mathbb{E} [I\{h(\mathbf{x}) \neq y\}]$$

החזאי האופטימאלי של פונקציית מחיר / risk זו הינו החזאי אשר מחזיר את ה  $y$  הכי סביר (הכי שכיח, ה mode) בהסתברות של  $y$  בהינתן  $\mathbf{x}$ :

$$h^*(\mathbf{x}) = \arg \max_y p(y|\mathbf{x} = \mathbf{x})$$

## (K-NN (K-Nearest Neighbours

K-NN הינו אלגוריתם דיסקרימינטיבי לפתרון בעיות סיווג. באלגוריתם זה החיזויים נעשים ישירות על פי המדגם באופן הבא:  
בהינתן  $\mathbf{x}$  מסויים:

1. נבחר את  $K$  הדגימות בעלות ה  $\mathbf{x}^{(i)}$  הקרובים ביותר ל  $\mathbf{x}$ . (לרוב נשתמש במרחק אוקלידי, אך ניתן גם לבחור פונקציית מחיר אחרות).

2. תוצאת החיזוי תהיה התווית השכיחה ביותר (majority vote) מבין  $K$  התוויות של הדגימות שנבחרו בשלב 1.

במקרה של שיוון:

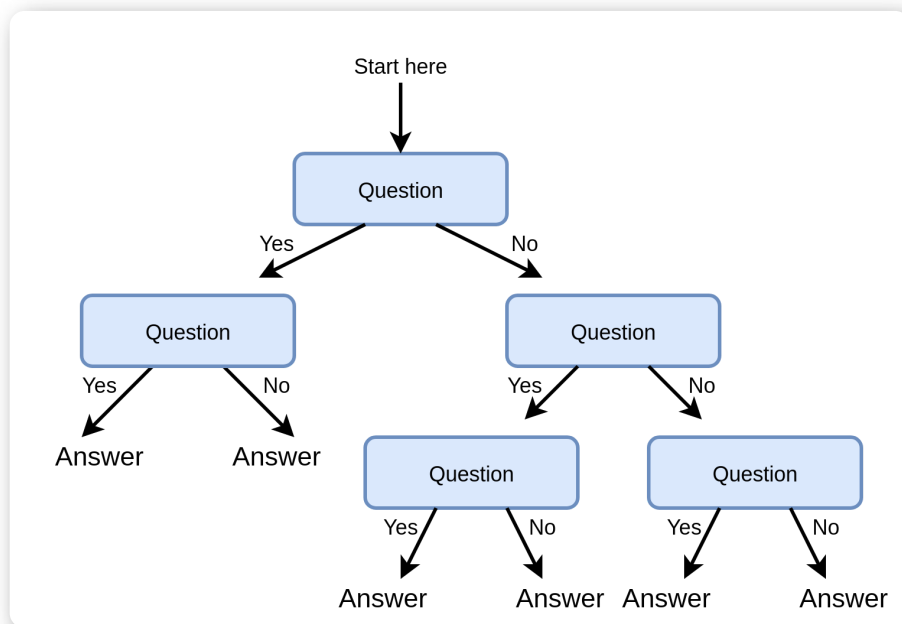
- במקרה של שיוויון בשלב 2, נשווה גם את המרחק הממוצע בין ה  $\mathbf{x}$ -ים השייכים לכל תווית. אנו נבחר בתווית בעלת המרחק הממוצע הקצר ביותר.
- במקרה של שיוויון גם בין המרחקים הממוצעים, נבחר אקראית.

## K-NN לבעיות רגרסיה

ניתן להשתמש באלגוריתם זה גם לפתרון בעיות רגרסיה אם כי פתרון זה יהיה לרוב פחות יעיל. בבעיות רגרסיה ניתן למצע על התוויות במקום לבחור את תווית השכיחה.

## Decision trees (עצי החלטה)

עצי החלטה הם כלי נפוץ (גם מחוץ לתחום של מערכות לומדות) לקבלת החלטות על סמך אוסף של עובדות.



טרמינולוגיה:

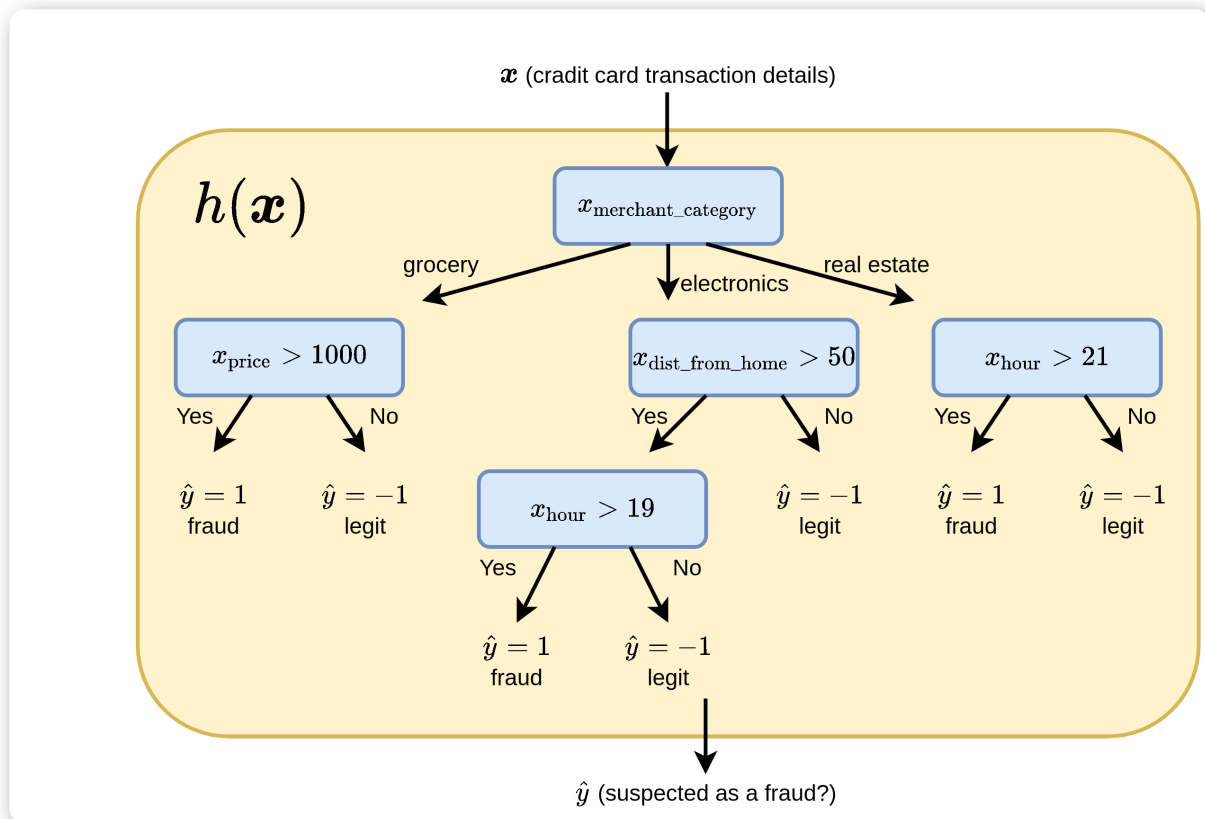
- **root (שורש)** - נקודת הכניסה לעץ.

- **node (צומת)** - נקודות החלטה / פיצול של העץ - השאלות.
- **leaves (עלים)** - הקצוות של העץ - התשובות.
- **branch (ענף)** - חלק מתוך העץ המלא (תת-עץ).

נוכל להשתמש בעצי החלטה שכאלה לבניית חזאים. הדרך הנפוצה לגדיר את השאלות על הענפים של העץ הינם על ידי תנאים על רכיב **יחיד** של  $x$ . ספציפית:

- לרוב נשתמש בתנאי מהצורה  $x_i > a$ , כאשר יש לבחור את  $i$  ו  $a$ .
- כאשר  $x_i$  הוא מתשנה דיסקרטי אשר מקבל סט קטן של ערכים נוכל גם לפצל לפי הערכים האפשריים של  $x_i$ .

לדוגמא



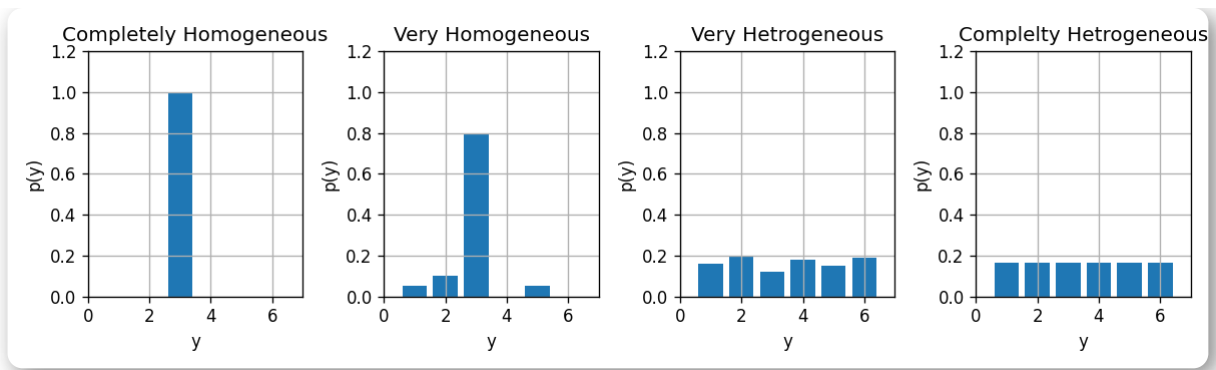
היתרונות של השימוש בעץ החלטה כחזאי:

1. פשוט למימוש (אוסף של תנאי if .. else ..).
2. מתאים לעבודה עם משתנים קטגוריים (משתנים בדדים אשר מקבלים אחד מסט מצומצם של ערכים).
3. Explainable - ניתן להבין בדיוק מה היו השיקולים שלפיהם התקבל חיזוי מסוים.

## בניית עץ החלטה לסיווג

### מדדים לחוסר ההומוגניות של פילוג

בהינתן משתנה אקראי דיסקרטי  $y$  אשר מקבל אחד מ  $C$  ערכים  $y \in \{1, 2, \dots, C\}$  ועם פונקציית הסתברות  $p(y)$ , נגדיר כמה מדדים אשר מודדים עד כמה הפילוג של  $y$  רחוק מלהיות פילוג אשר מייצר דגימות הומוגניות (זאת אומרת פילוג שהוא פונקציית דלתא):



• שגיאת הסיווג (אשר המתקבלת בעבור חיזוי של הערך הכי סביר)

$$Q(p) = 1 - \max_{y \in \{1, \dots, C\}} p(y)$$

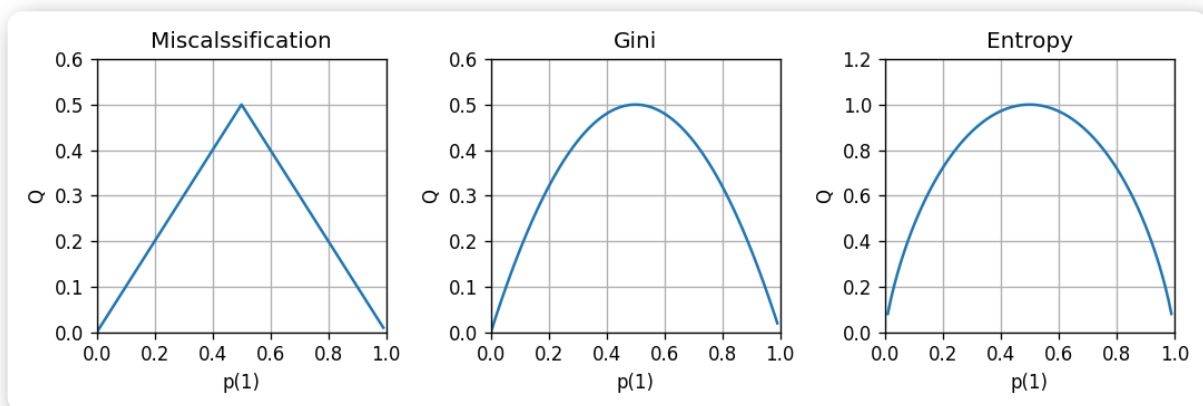
• אינדקס Gini:

$$Q(p) = \sum_{y \in \{1, \dots, C\}} p(y)(1 - p(y))$$

• אנטרופיה:

$$Q(p)(= H(p)) = \sum_{y \in \{1, \dots, C\}} -p(y) \log_2 p(y)$$

מדדים אלו שווים ל-0 בעבור פילוגים אשר מייצרים דגימות הומוגניות והם גדלים ככל שהפילוג הולך ונעשה אחיד. השרטוטים הבאים מראים את ההתנהגות של המדדים האלה במקרה של משתנה אקראי בינארי:



## חוסר הומוגניות ממוצעת של עץ

בהינתן מדגם מסוים ומדד חוסר הומוגניות מסוים נגדיר את הציון של עץ החלטה נתון באופן הבא:

1. נעביר את הדגימות מהמדגם דרך העץ ונפצל אותם על פי העלה שאליו הם הגיעו. נסמן את האינדקסים של הדגימות שהגיעו לעלה ה- $j$  ב- $\mathcal{I}_j$ . נסמן את כמות הדגימות שהגיעו לעלה ה- $j$  ב- $N_j$ .

2. לכל עלה נחשב את הפילוג האמפירי של התויות שהגיעו אליו באופן הבא:

$$\hat{p}_{j,y} = \frac{1}{N_j} \sum_{i \in \mathcal{I}_j} I\{y_i = y\}$$

( $p_{j,y}$  הוא פשוט השכיחות של הערך  $y$  מבין התויות בעלה ה- $j$ )

3. בעזרת הפילוג האמפירי נחשב את חוסר הומוגניות של כל עלה:

$$Q(\hat{p}_j)$$

4. הציון הכולל של העץ יהיה הממוצע המושכלל של חוסר ההומוגניות של העלים ביחס לכמות הדגימות שהגיעה לכל עלה:

$$Q_{\text{total}} = \sum_j \frac{N_j}{N} Q(\hat{p}_j)$$

## שלב ראשון - בניה של עץ מלא

בכדי לקבל סיווג כמה שיותר טוב נרצה שהפילוג בעלים של העץ יהיו כמה שיותר הומוגניים, זאת אומרת שמדד חוסר ההומוגניות יהיה 0. בכדי להימנע כמה שיותר מ overfitting נרצה לעשות זאת על ידי שימוש בכמה שפחות nodes.

מכיוון שבכדי למצוא את הפתרון האופטימלי יש לעבור על כל העצים האפשריים, אנו נחפש פיתרון שאינו בהכרח האופטימלי על ידי בניה של העץ בצורה חמדנית (greedy). אנו נתחיל מה root ונוסיף nodes כך שבכל שלב נבחר את ה node אשר מניב את העץ עם מדד החוסר ההומוגניות הנמוך ביותר. אנו נעשה זאת על ידי מעבר על כל האופציות האפשריות לבחור את ה node. ממשיכים לפצל את העלים של העץ כל עוד מדד חוסר ההומוגניות יורד.

מלבד במקרים שבהם יש במדגם שתי דגימות עם אותו ה  $x$  אך  $y$  שונה, יהיה ניתן להגיע למדד חוסר ההומוגניות 0. זאת אומרת שבכל עלה יש דגימות עם תוויות מסוג אחד בלבד. באותם מקרים שבהם לא ניתן להגיע לתוויות בודדת לכל עלה, נבחר את תוצאת החיזוי להיות הערך הכי שכיח באותו עלה.

## שלב שני - pruning (גיזום)

בכדי להקטין את מידת ה overfitting של העץ ניתן להשתמש ב validation set על מנת לבצע pruning של העץ באופן הבא:

מתחילים מכל אחד מהעלים ובודקים לכל node, האם הסרה שלו משפרת או לא משנה את ביצועי העץ על ה validation set. במידה וזה אכן המצב מסירים אותו. ממשיכים כך עד שאין עוד מה להסיר.

## Regression Tree

ניתן להשתמש בעצים גם לפתרון בעיות רגרסיה. במקרה של רגרסיה עם פונקציית מחיר של MSE, הבניה של העץ תהיה זהה מלבד שני הבדלים:

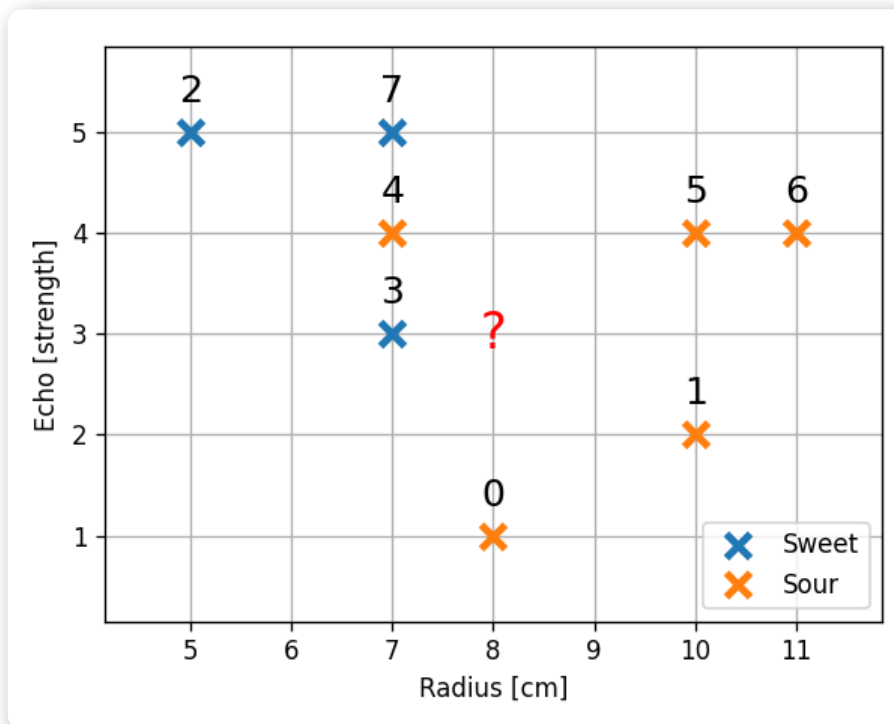
1. תוצאת החיזוי בעלה מסויים תהיה הערך הממוצע של התוויות באותו עלה. (במקום הערך השכיח)
2. את מדד חוסר ההומוגניות נחליף בשגיאה הריבועית של החיזוי של העץ.

## תרגיל 4.1

סטודנט נבון ניגש לבחור אבטיחים בסופרמרקט. ידוע כי זוהי רק תחילתה של עונת האבטיחים וקיים מספר לא מבוטל של אבטיחי בוסר. הסטודנט שם לב כי ניתן לאפיין את האבטיחים ע"פ ההד בהקשה וע"פ קוטר האבטיח. הסטודנט החליט למפות את ניסיון העבר שלו:

	Radius	Echo	Sweetness
0	8	1	-1
1	10	2	-1
2	5	5	1
3	7	3	1
4	7	4	-1
5	10	4	-1
6	11	4	-1

	Radius	Echo	Sweetness
	7	7	5
			1



הסטודנט מחזיק בידו אבטיח בעל בעוצמה 3 ורדיוס 8 ס"מ. על מנת לחזות האם האבטיח בידו מתוק או חמוץ נבנה חזאי בעזרת K-NN.

**(1)** (ללא קשר לבעיה בשאלה) כיצד הפרמטר  $K$  משפיע על השיגה שאנו צופים לקבל באלגוריתם ה K-NN? ממה תנבע השיגה כאשר  $K$  יהיה קטן וממה תנבע השיגה כאשר  $K$  יהיה גדול?

**(2)** בעבור המדגם הנתון, מה קורה במקרה שבו  $K = 8$ ? האם שיגה זו היא שיגה  $overfitting$  או  $underfitting$ ?

**(3)** השתמשו ב  $leave-one-out$  cross validation על מנת לקבוע את ה  $K$  האופטימאלי מבין הערכים 1,3,5,7. השתמשו ב  $missclassification$  rate כפונקציית המחר.

**(4)** השתמשו ב  $K$  שמצאתם בכדי לחשב את החיזוי הסופי.

## פתרון 4.1

**(1)**

כפי שראינו בהרצאה, כאשר  $K$  יהיה קטן אנו למעשה מתאימים לכל נקודה ב  $train$  set איזור החלטה משלה. במצב זה החזאי יתן חיזוי מושלם על המדגם, אך איזורי החלטה אלו, אשר תלויים במקומים המקריים של נקודות בודדות, לא בהכרח ייצגו את האופי של הפילוג האמיתי. זהו בדיוק המקרה של  $overfitting$ .

כאשר  $K$  יהיה גדול מאד אנו למעשה נמצע על איזורים מאד גדולים ולכן החזאי יתעלם מהשינויים העדינים בפילוג של הנקודות, וייתחס רק למגמה המאד כללית. זה בדיוק המצב של  $underfitting$ .

**(2)**

במקרה הקיצוני שבו  $K$  שווה לגודל ה  $dataset$  כל חיזוי יתבצע על סמך כל הנקודות במדגם ולכן יהיה שווה תמיד לתוית השכיחה ביותר במדגם. במקרה זה החיזוי יהיה תמיד 1-, זאת אומרת שאנו נחזה שהאבטיח הוא חמוץ ללא כל תלות בהד

### 3

בכדי לקבוע את הערך האופטימאלי של  $K$  מתוך הערכים הנתונים בעזרת K-fold cross validation עלינו לחשב את ציון ה validation לכל ערך של  $K$  ולכל fold (שימוש בנקודה אחרת כ validation set). לאחר מכאן נמצע על ה folds השונים על מנת לקבל את הציון של כל  $K$ . נרכז בטבלה את החיזוי המשוערך לכל fold ולכל  $K$ :

point	Correct label	K=1 prediction	K=3 prediction	K=5 prediction	K=7 prediction
0	-1	✓ -1 (nn=[1])	✓ -1 (nn=[1 3 4])	✓ -1 (nn=[1 3 4 5 7])	✓ -1 (nn=[1 3 4 5 7 6 2])
1	-1	✓ -1 (nn=[5])	✓ -1 (nn=[5 0 6])	✓ -1 (nn=[5 0 6 3 4])	✓ -1 (nn=[5 0 6 3 4 7 2])
2	1	✓ 1 (nn=[7])	✓ 1 (nn=[7 4 3])	✗ -1 (nn=[7 4 3 0 5])	✗ -1 (nn=[7 4 3 0 5 1 6])
3	1	✗ -1 (nn=[4])	✗ -1 (nn=[4 7 0])	✗ -1 (nn=[4 7 0 2 1])	✗ -1 (nn=[4 7 0 2 1 5 6])
4	-1	✗ 1 (nn=[3])	✗ 1 (nn=[3 7 2])	✗ 1 (nn=[3 7 2 5 0])	✓ -1 (nn=[3 7 2 5 0 1 6])
5	-1	✓ -1 (nn=[6])	✓ -1 (nn=[6 1 4])	✓ -1 (nn=[6 1 4 3 7])	✓ -1 (nn=[6 1 4 3 7 0 2])
6	-1	✓ -1 (nn=[5])	✓ -1 (nn=[5 1 4])	✓ -1 (nn=[5 1 4 3 7])	✓ -1 (nn=[5 1 4 3 7 0 2])
7	1	✗ -1 (nn=[4])	✓ 1 (nn=[4 2 3])	✗ -1 (nn=[4 2 3 5 0])	✗ -1 (nn=[4 2 3 5 0 6 1])
Avg. score		3/8	2/8	4/8	3/8

ניתן לראות כי בעבור  $K = 3$  אנו מקבלים את השגיאה הממוצעת הקטנה ביותר  $2/8$  לכן נקבע את  $K$  לערך זה.

### 4

נבדוק את בשלות האבטיח שהסטודנט מחזיק בידו על סמך המדגם כולו עם  $K = 3$ . שלושת הנקודות הקרובות ביותר לנקודה  $(8, 3)$  הינם הנקודות 3, 0 ו 4. מכיוון ששתיים מהן עם תווית של -1 אנו נחזה שאבטיח זה הוא בוסר.

## שאלה 4.2 - בניית עץ החלטה

בנה עץ החלטה המבוסס על קריטריון האנטרופיה, אשר בהינתן נתוני צבע שער, גובה, משקל, והשימוש בקרם הגנה, חוזה האם עתיד האדם להכוות מהשמש היוקדת. סט דוגמאות הלימוד לצורך בניית העץ מוצג בטבלה הבאה:

Hair	Height	Weight	Lotion	Result (Label)
blonde	average	light	no	sunburned
blonde	tall	average	yes	none
brown	short	average	yes	none
blonde	short	average	no	sunburned
red	average	heavy	no	sunburned

Hair	Height	Weight	Lotion	Result (Label)
brown	tall	heavy	no	none
brown	average	heavy	no	none
blonde	short	light	yes	none

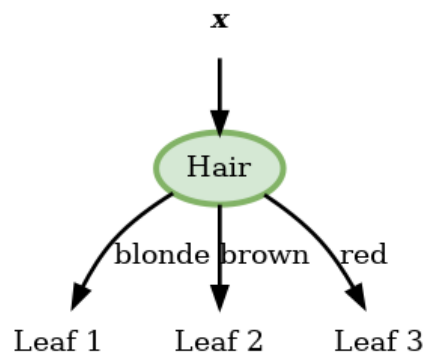
## פתרון 4.2

נפעל על פי האלגוריתם ונתחיל מה root ונתחיל להוסיף nodes:



במקרה זה יש לנו 4 nodes אפשריים (בעבור כל שדה של  $x$ ). נחשב את האנטרופיה הממוצעת של כל אחד מהם ונבחר את המינימאלי.

### Hair



נעביר את המדגם דרך העץ ונרכז בטבלה הבאה את התוויות ואנטרופיה המתקבלות בכל עלה:

	Leaf ( $j$ )	$N_j$	$\hat{p}_j$	$H(\hat{p}_j)$
Blonde	1	4	$\{\frac{2}{4}, \frac{2}{4}\}$	$-\frac{1}{2} \log \frac{1}{2} - \frac{1}{2} \log \frac{1}{2} = 1$
Brown	2	3	$\{\frac{0}{3}, \frac{3}{3}\}$	$-0 \log(0) - 1 \log(1) = 0$
Red	3	1	$\{\frac{1}{1}, \frac{0}{1}\}$	$-1 \log(1) - 0 \log(0) = 0$

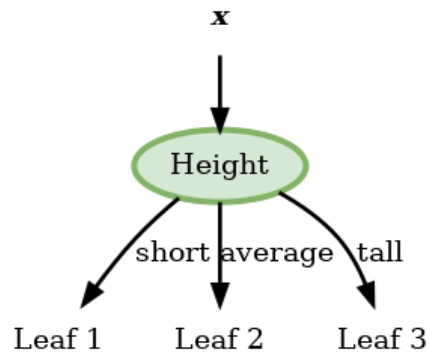
נחשב את הממוצע הממושקל של האנטרופיה על שלושת העלים:



$$Q_{\text{total}} = \sum_j \frac{N_j}{N} Q(\hat{p}_j) = \frac{4}{8} \cdot 1 + \frac{3}{8} \cdot 0 + \frac{1}{8} \cdot 0 = \frac{1}{2}$$

נמשיך לשדה הבא.

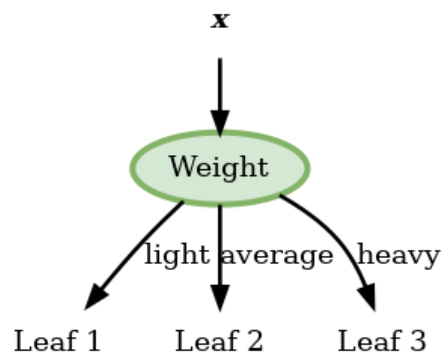
## Height



	Leaf ( $j$ )	$N_j$	$\hat{p}_j$	$H(\hat{p}_j)$
Short	1	3	$\{\frac{1}{3}, \frac{2}{3}\}$	$-\frac{1}{3} \log \frac{1}{3} - \frac{2}{3} \log \frac{2}{3} = 0.918$
Average	2	3	$\{\frac{2}{3}, \frac{1}{3}\}$	$-\frac{2}{3} \log \frac{2}{3} - \frac{1}{3} \log \frac{1}{3} = 0.918$
Tall	3	2	$\{\frac{0}{2}, \frac{2}{2}\}$	$-0 \log(0) - 1 \log(1) = 0$

$$Q_{\text{total}} = \sum_j \frac{N_j}{N} Q(\hat{p}_j) = \frac{3}{8} \cdot 0.918 + \frac{3}{8} \cdot 0.918 + \frac{2}{8} \cdot 0 = 0.69$$

## Weight

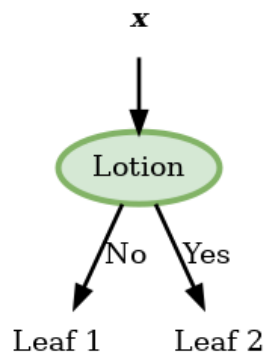


	Leaf ( $j$ )	$N_j$	$\hat{p}_j$	$H(\hat{p}_j)$
Light	1	2	$\{\frac{1}{2}, \frac{1}{2}\}$	$-\frac{1}{2} \log \frac{1}{2} - \frac{1}{2} \log \frac{1}{2} = 1$

	Leaf ( $j$ )	$N_j$	$\hat{p}_j$	$H(\hat{p}_j)$
Average	2	3	$\{\frac{1}{3}, \frac{2}{3}\}$	$-\frac{1}{3} \log \frac{1}{3} - \frac{2}{3} \log \frac{2}{3} = 0.918$
Heavy	3	3	$\{\frac{1}{3}, \frac{2}{3}\}$	$-\frac{1}{3} \log \frac{1}{3} - \frac{2}{3} \log \frac{2}{3} = 0.918$

$$Q_{\text{total}} = \sum_j \frac{N_j}{N} Q(\hat{p}_j) = \frac{2}{8} \cdot 1 + \frac{3}{8} \cdot 0.918 + \frac{3}{8} \cdot 0.918 = 0.9385$$

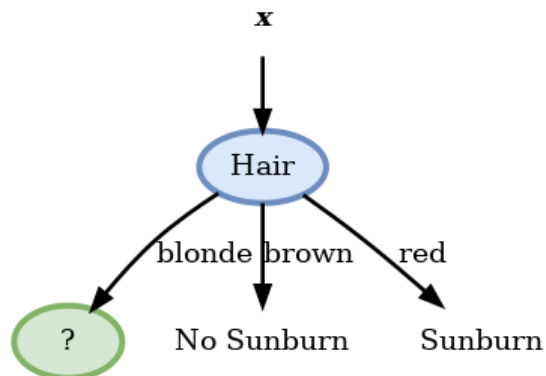
## Lotion



	Leaf ( $j$ )	$N_j$	$\hat{p}_j$	$H(\hat{p}_j)$
No	1	5	$\{\frac{3}{5}, \frac{2}{5}\}$	$-\frac{3}{5} \log \frac{3}{5} - \frac{2}{5} \log \frac{2}{5} = 0.97$
Yes	2	3	$\{\frac{0}{3}, \frac{3}{3}\}$	$-0 \log(0) - 1 \log(1) = 0$

$$Q_{\text{total}} = \sum_j \frac{N_j}{N} Q(\hat{p}_j) = \frac{5}{8} \cdot 0.97 + \frac{3}{8} \cdot 0 = 0.606$$

מכאן שהמאפיין האופטימלי לפיצול הראשון (על פי קריטריון האנטרופיה) הוא **Hair**. לכן נבחר בו להיות ה node הראשון. נשים לב כי בעבור node זה שני הפיצולים של brown ו red כבר הומוגניים לגמרי (מכילים רק סוג אחד של תוויות) ולכן לא נמשיך לפצל אותם ונרשום את החיזוי המקבל בכל עלה:

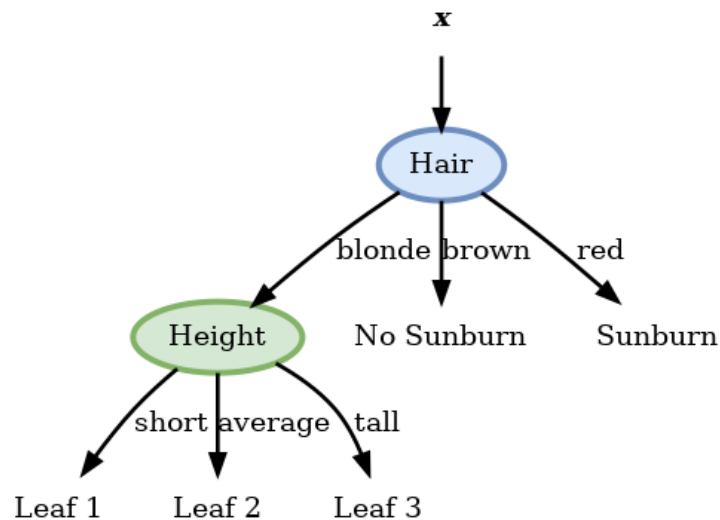


נמשיך כעת באופן דומה לבחור את ה node בעבור הענף של blond. מכיוון שאין טעם לבדוק שוב את ה node של hair נשאר לנו לבדוק רק את שלושת האופציות הנותרות. לשם הנוחות נרכז את הדגימות האשר מגיעות לענף זה:

Height	Weight	Lotion	Result
average	light	no	sunburned
tall	average	yes	none
short	average	no	sunburned
short	light	yes	none

## Height

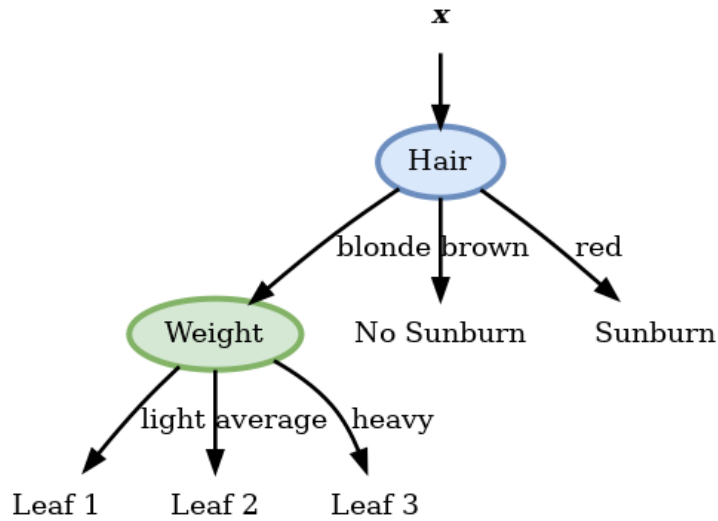
בכדי לבחור את ה node האופטימאלי נוכל להתעלם מכל מה שקורה בענפים אחרים ולחשב רק את האנטרופיה המתקבלת בענף הזה (של ה blond).



	Leaf (j)	$N_j$	$\hat{p}_j$	$H(\hat{p}_j)$
Short	1	2	$\{\frac{1}{2}, \frac{1}{2}\}$	$-\frac{1}{2} \log \frac{1}{2} - \frac{1}{2} \log \frac{1}{2} = 1$
Average	2	1	$\{\frac{0}{1}, \frac{1}{1}\}$	$-0 \log(0) - 1 \log(1) = 0$
Tall	3	1	$\{\frac{0}{1}, \frac{1}{1}\}$	$-1 \log(1) - 0 \log(0) = 0$

$$Q_{\text{blond}} = \sum_j \frac{N_j}{N} Q(\hat{p}_j) = \frac{2}{8} \cdot 1 + \frac{1}{8} \cdot 0 + \frac{1}{8} \cdot 0 = 0.25$$

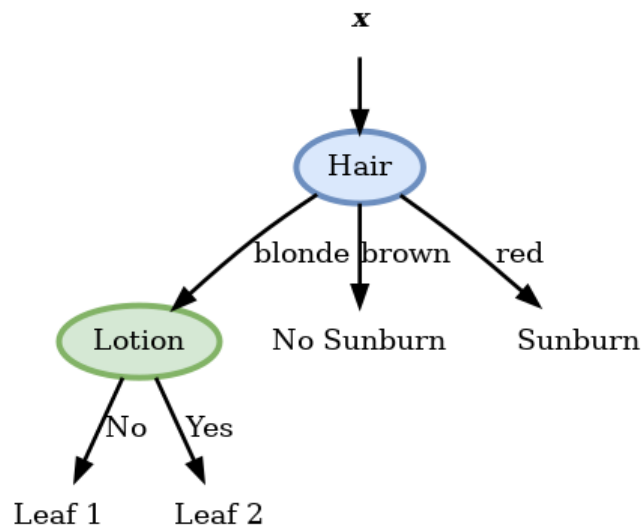
## Weight



	Leaf ( $j$ )	$N_j$	$\hat{p}_j$	$H(\hat{p}_j)$
Light	1	2	$\{\frac{1}{2}, \frac{1}{2}\}$	$-\frac{1}{2} \log \frac{1}{2} - \frac{1}{2} \log \frac{1}{2} = 1$
Average	2	2	$\{\frac{1}{2}, \frac{1}{2}\}$	$-\frac{1}{2} \log \frac{1}{2} - \frac{1}{2} \log \frac{1}{2} = 1$
Heavy	3	0		

$$Q_{\text{total}} = \sum_j \frac{N_j}{N} Q(\hat{p}_j) = \frac{2}{8} \cdot 1 + \frac{2}{8} \cdot 1 = 0.5$$

## Lotion

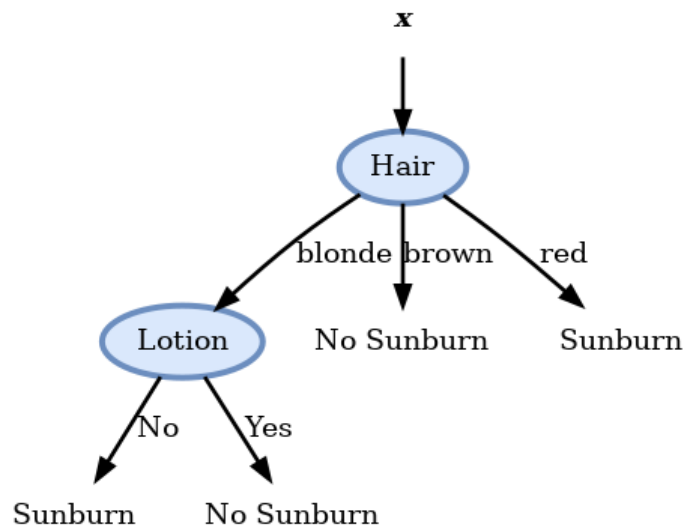


Leaf ( $j$ )	$N_j$	$\hat{p}_j$	$H(\hat{p}_j)$
--------------	-------	-------------	----------------

	Leaf ( $j$ )	$N_j$	$\hat{p}_j$	$H(\hat{p}_j)$
No	1	2	$\{\frac{2}{2}, \frac{0}{2}\}$	$-1 \log(1) - 0 \log(0) = 0$
Yes	2	2	$\{\frac{0}{2}, \frac{2}{2}\}$	$-0 \log(0) - 1 \log(1) = 0$

$$Q_{\text{total}} = \sum_j \frac{N_j}{N} Q(\hat{p}_j) = \frac{2}{8} \cdot 0 + \frac{2}{8} \cdot 0 = 0$$

פיצול זה נותן אנטרופיה 0 ולכן הוא הפיצול האופטימאלי ואנו נבחר בו. עץ ההחלטה הסופי יראה אם כן:



עץ זה ממיין באופן מושלם את המדגם.

## תרגיל 4.3

נתון המדגם הבא של ערכי תצפית של  $\mathbf{x} = [x_1, x_2, x_3]^T$  ושל תוויות  $y$ :

	$x_1$	$x_2$	$x_3$	$y$
1	1	1	-1	1
2	1	-1	-1	1
3	-1	-1	-1	1
4	-1	-1	-1	-1
5	1	1	1	-1

נרצה לבנות עץ החלטה על מנת לחזות את  $y$  על סמך  $\mathbf{x}$ . נרצה להשתמש במדד חוסר הומוגניות חדש מסוג Squared Root Gini אשר מוגדר באופן הבא:

$$Q(p) = \sum_y \sqrt{p(y)(1-p(y))}$$

- (1) בנו עץ מלא על סמך קריטריון זה. כמה nodes יש בעץ שמצאתם?
- (2) חשבו את הציון (score) של עץ זה תחת פונקציית המחיר של misclassification rate. האם ניתן להגיע לסיווג מושלם במקרה זה?
- (3) האם בעבור מקרה זה ניתן לבנות עץ אשר מגיע לאותו ציון כמו העץ שמצאתם בסעיף 1 אך עם פחות nodes? אם כן, הציעו סיבה אפשרית למה האלגוריתם בו השתמשתם בסעיף הקודם לא מצא את העץ הזה.

## 4.3 פתרון

(1)

נתחיל מה root ונבדוק את שלושת ה nodes האפשריים תחת מדד השגיאה החדש:



$x_1$

	Leaf ( $j$ )	$N_j$	$\hat{p}_j$	$Q(\hat{p}_j)$
1	1	3	$\{\frac{2}{3}, \frac{1}{3}\}$	$2\sqrt{\frac{2}{3} \cdot \frac{1}{3}} = 0.94$
-1	2	2	$\{\frac{1}{2}, \frac{1}{2}\}$	$2\sqrt{\frac{1}{2} \cdot \frac{1}{2}} = 1$

$$Q_{\text{total}} = \frac{3}{5} \cdot 0.94 + \frac{2}{5} \cdot 1 = 0.96$$

$x_2$

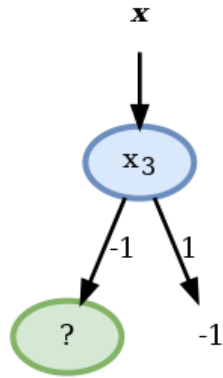
נשים לב ש node זה נותן חלוקה דומה של התוויות לזו של  $x_1$  ולכן נקבל את אותו ערך המדד של 0.96.

$x_3$

	Leaf ( $j$ )	$N_j$	$\hat{p}_j$	$Q(\hat{p}_j)$
-1	1	4	$\{\frac{3}{4}, \frac{1}{4}\}$	$2\sqrt{\frac{3}{4} \cdot \frac{1}{4}} = 0.86$
1	2	1	$\{\frac{0}{1}, \frac{1}{1}\}$	$2\sqrt{0 \cdot 1} = 0$

$$Q_{\text{total}} = \frac{4}{5} \cdot 0.86 + \frac{1}{5} \cdot 0 = 0.7$$

לכן נבחר את ה node הראשון להיות התנאי על  $x_3$ . מכיוון שהענף של  $x_3 = 1$  כבר הומוגני לא נחלק אותו יותר:



הדגימות הרלוונטיות כרגע הם:

	$x_1$	$x_2$	$y$
1	1	1	1
2	1	-1	1
3	-1	-1	1
4	-1	-1	-1

נבדוק את שני השדות שנותרו:

$x_1$

	Leaf ( $j$ )	$N_j$	$\hat{p}_j$	$Q(\hat{p}_j)$
1	1	2	$\{\frac{2}{2}, \frac{0}{2}\}$	$2\sqrt{1 \cdot 0} = 0$
-1	2	2	$\{\frac{1}{2}, \frac{1}{2}\}$	$2\sqrt{\frac{1}{2} \cdot \frac{1}{2}} = 1$

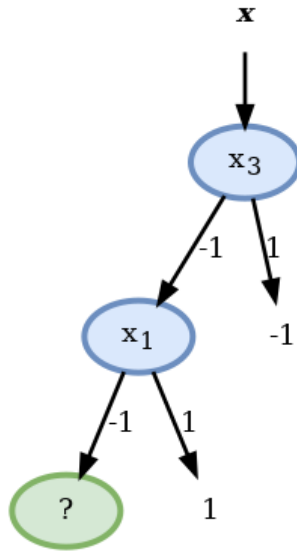
$$Q_{\text{branch}} = \frac{2}{5} \cdot 0 + \frac{2}{5} \cdot 1 = 0.4$$

$x_2$

	Leaf ( $j$ )	$N_j$	$\hat{p}_j$	$Q(\hat{p}_j)$
1	1	1	$\{\frac{1}{1}, \frac{0}{1}\}$	$2\sqrt{1 \cdot 0} = 0$
-1	2	3	$\{\frac{2}{3}, \frac{1}{3}\}$	$2\sqrt{\frac{2}{3} \cdot \frac{1}{3}} = 0.94$

$$Q_{\text{branch}} = \frac{1}{5} \cdot 0 + \frac{3}{5} \cdot 0.94 = 0.57$$

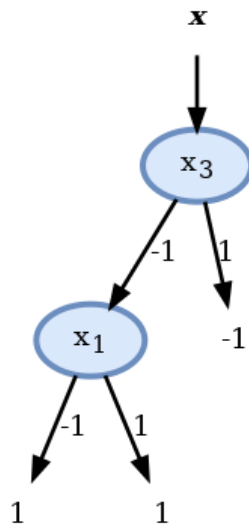
ה node האופטימאלי כאן הוא הפיצול לפי  $x_1$  ונשים לב שהענף של  $x_1 = 1$  כבר הומוגני:



בעבור הענף של  $x_1 = -1$  הדגימות הרלוונטיות הם:

	$x_2$	$y$
3	-1	1
4	-1	-1

במקרה זה למרות שלא הגענו לפילוג הומוגני לא נוכל לפצל יותר את הענף כי הערכים של  $x_2$  זהים בעבור שני הדגימות ולכן לא ניתן להבחין ביניהם. במקרה זה נחבר את החיזוי באופן שרירותי להיות 1 ונסיים את הבניה של העץ:



בעץ שמצאנו ישנם 2 nodes.

**(2)**

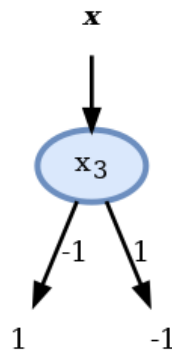


בעבור העלים אשר הפילוג של התגיות בהם הינו הומוגני החיזוי יהיה מושלם. שגיאות חיזוי יתקבלו רק בעלה של  $x_3 = 1 = -1$  אשר לא הצליח להגיע לפילוג הומוגני. מכיוון שבחרנו (באופן שרירותי) שהחיזוי בעלה זה יהיה 1 הדגימה היחידה אשר תסווג לא נכון היא דגימה 4. מכאן שהחזאי שבנינו יעשה על המדגם שגיאה אחת מתוך 5, זאת אומרת misclassification rate של  $1/5 = 0.2$ .

כפי שציינו קודם, מכיוון שלדגימות 3 ו 4 יש את אותו  $x$  אך  $y$  שונה לא ניתן להפריד ביניהם ותמיד על אחד מהם החיזוי יהיה לא נכון. לכן הציון של 0.2 הוא הציון המינימאלי שאותו ניתן לקבל על המדגם הזה.

### (3)

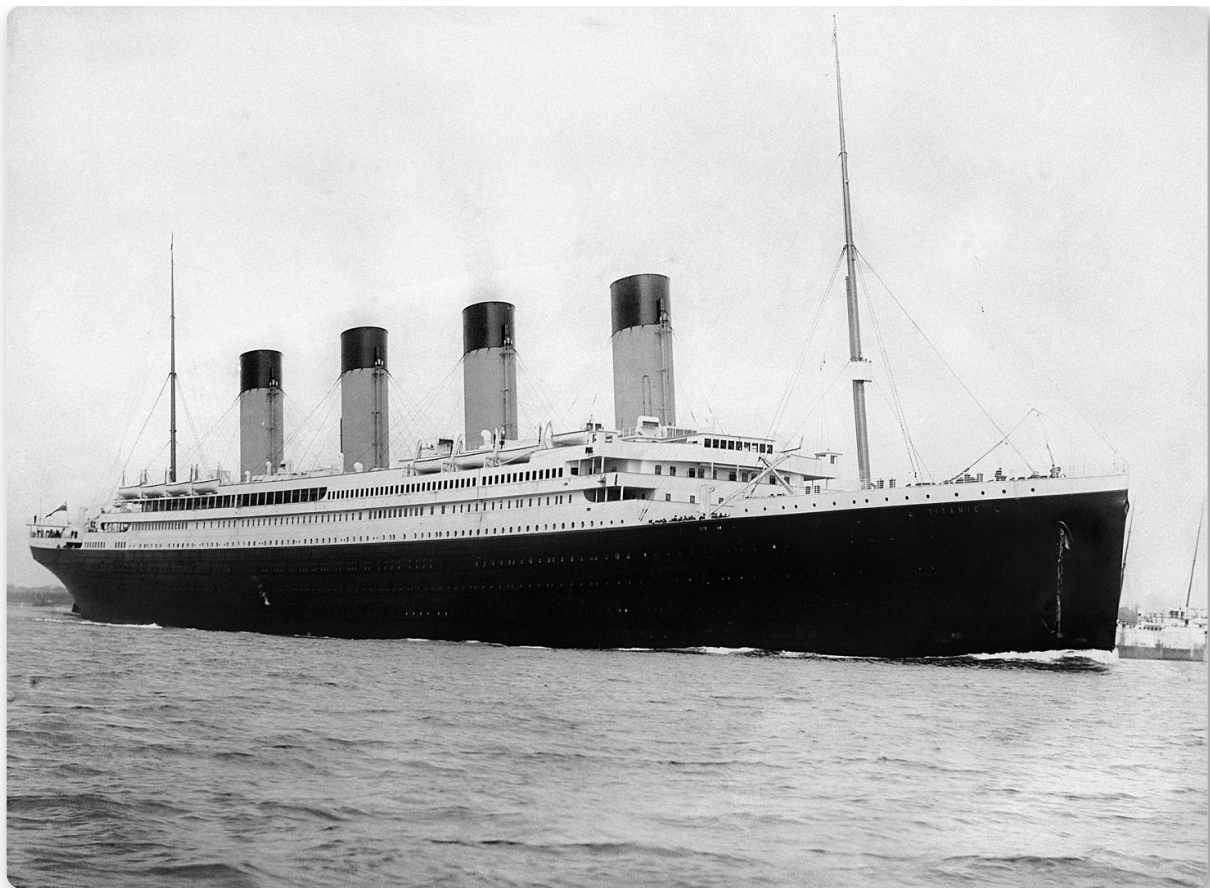
נשים לב שלמעשה ה node השני בעץ לא עושה כלום משום שללא תלות בערך של  $x_2$  הוא חוזר 1 ולכן ניתן באותה המידה להשתמש גם בעץ הבא ולקבל את אותו החיזוי:



הסיבה שהאלגוריתם לא התכנס לפיתרון זה הינה שבבנייה של העץ ניסינו למזער את מדד ה squared root gini הממוצע ולא את שגיאת החיזוי ומכיוון שאלו שתי בעיות שונות, גם הפתרונות שלהם יכולים להיות שונים.

## חלק מעשי - הטיטניק

Code



אחד ה-Datasets הנפוצים למשחקים פשוטים והדגמות של מערכות לומדות הוא רשימת הנוסעים של ספינת הטיטניק. רשימה זו מכילה פרטים שונים על כל אחד מהנוסעים יחד עם אינדיקטור של איזה מהנוסעים שרד. בעיית supervised learning שניתן להגדיר על מדגם זה הינה הבעיה של לנסות ולחזות איזה מהנוסעים שרד ואיזה לא על סמך הפרטים של כל נוסע. את המדגם המקורי ניתן למצוא [פה](#), אנו נעבוד עם גרסה מעט יותר נקיה שלו, שאותה ניתן למצוא [פה](#).

נציג את 10 השורות הראשונות במדגם:

boat	embarked	cabin	fare	ticket	parch	sibsp	age	sex	name	survived	pclass
2	S	B5	211.338	24160	0	0	29	female	Allen, Miss. Elisabeth Walton	1	1 0
nan	S	C22 C26	151.55	113781	2	1	2	female	Allison, Miss. Helen Loraine	0	1 1
nan	S	C22 C26	151.55	113781	2	1	30	male	Allison, Mr. Hudson Joshua Creighton	0	1 2
nan	S	C22 C26	151.55	113781	2	1	25	female	Allison, Mrs. Hudson J C (Bessie Waldo Daniels	0	1 3

boat	embarked	cabin	fare	ticket	parch	sibsp	age	sex	name	survived	pclass
3	S	E12	26.55	19952	0	0	48	male	Anderson, Mr. Harry	1	1 4
10	S	D7	77.9583	13502	0	1	63	female	Andrews, Miss. Kornelia Theodosia	1	1 5
nan	S	A36	0	112050	0	0	39	male	Andrews, Mr. Thomas Jr	0	1 6
D	S	C101	51.4792	11769	0	2	53	female	Appleton, Mrs. Edward Dale (Charlotte Lamson)	1	1 7
nan	C	nan	49.5042	PC 17609	0	0	71	male	Artagaveytia, Mr. Ramon	0	1 8
nan	C	C62 C64	227.525	PC 17757	0	1	47	male	Astor, Col. John Jacob	0	1 9

במדגם הנקי יש 999 רשומות.

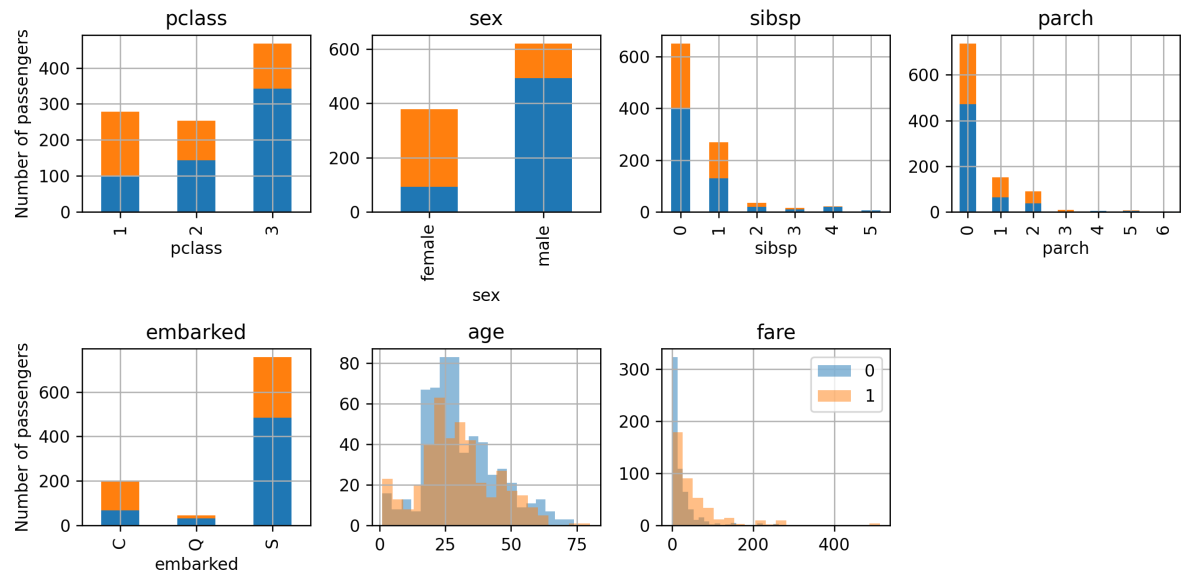
## השדות

בתרגול זה לא נשתמש בכל השדות אלא רק בשדות הבאים:

- **pclass**: מחלקת הנוסע: 1, 2 או 3
- **sex**: מין הנוסע
- **age**: גיל הנוסע
- **sibsp**: מס' של אחים ובני זוג של כל נוסע על האונייה
- **parch**: מס' של ילדים או הורים של כל נוסע על האונייה
- **fare**: המחיר שהנוסע שילם על הכרטיס
- **embarked**: הנמל בו עלה הנוסע על האונייה (C = Cherbourg; Q = Queenstown; S = Southampton)
- **survived**: התיוג, האם הנוסע שרד או לא

## התרשמות ראשונית בעזרת גרפים

נציג את הפילוג של כל אחד מהשדות בעבור האנשים ושרדו והאנשים שלא:



מתוך הגרפיים ניתן לראות כי אכן ישנם מאפיינים שיוכלו לסייע לנו לשפר את החיזוי שלנו. לדוגמא, ניתן לראות כי לנשים היה סיכוי גבוה בהרבה לשרוד לגברים וכך גם לנוסעים במחלקה הראשונה.

## הגדרת הבעיה

נסמן:

- $x$  : הוקטור האקראי אשר מכיל את כל פרטי הנוסע.
- $y$  : המשתנה האקראי של האם הנוסע שרד או לו.

נרצה למצוא חזאי (מסווג) טוב ככל האפשר תחת פונקציית המחיר של  $\text{miscallssification rate}$ .

אנו נעשה זאת בעזרת  $\text{decision tree}$ .

## חלוקת ה dataset

- נחלק את המדגם ל  $\text{train set } 80\%$  ו  $\text{test set } 20\%$ .
- נחלק את ה  $\text{train set}$  פעם נוספת ל  $\text{train set } 75\%$  ו  $\text{validation set } 25\%$ .

## בניית עץ בעל שלוש רמות

נבנה את העץ על פי קריטריון Gini. נתחיל מה  $\text{root}$  ונוסיף בכל פעם את ה  $\text{node}$  שמזער את המדד. בעבור ה  $\text{node}$  הראשון:



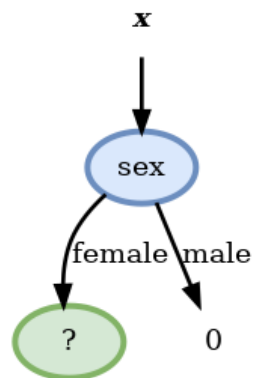
האופציות הם:

Score before split: 0.492

Scores:

- pclass: 0.436
- sex: 0.360 <-
- sibsp: 0.479
- parch: 0.473
- embarked: 0.460
- age >= 9: 0.488
- fare >= 15.7417: 0.448

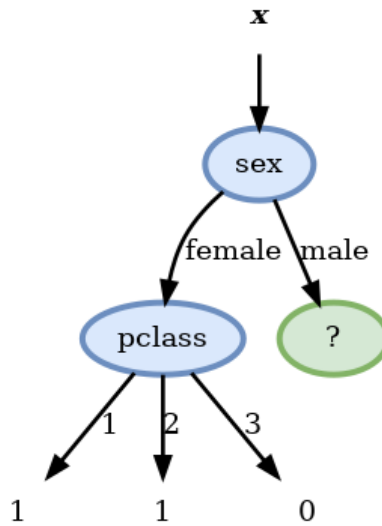
ולכן נבחר למיין לפי המגדר. נמשיך באופן זהה לכל שאר ה nodes.



Score before split: 0.146

Scores:

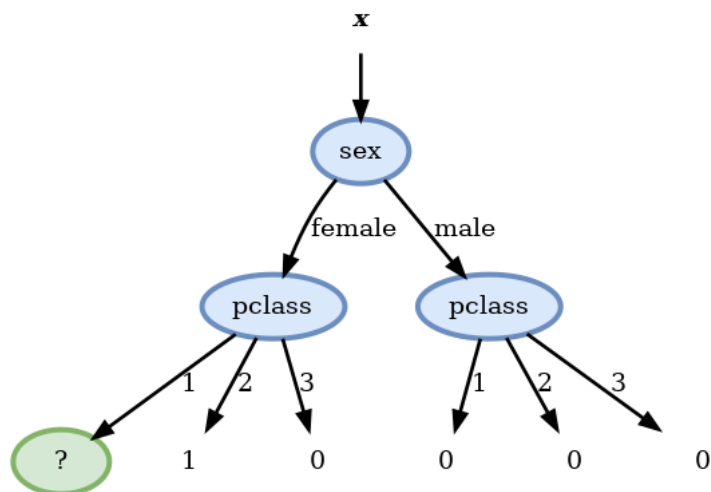
- pclass: 0.109 <-
- sex: 0.146
- sibsp: 0.140
- parch: 0.143
- embarked: 0.130
- age >= 48: 0.142
- fare >= 10.5: 0.126



Score before split: 0.214

Scores:

- pclass: 0.202 <-
- sex: 0.214
- sibsp: 0.212
- parch: 0.209
- embarked: 0.205
- age >= 10: 0.207
- fare >= 26.2875: 0.205



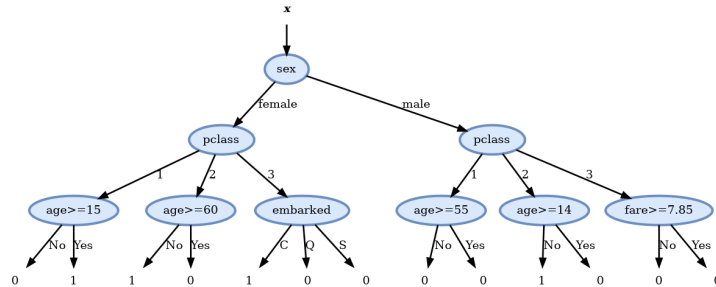
Score before split: 0.010

Scores:

- pclass: 0.010
- sex: 0.010
- sibsp: 0.009
- parch: 0.008

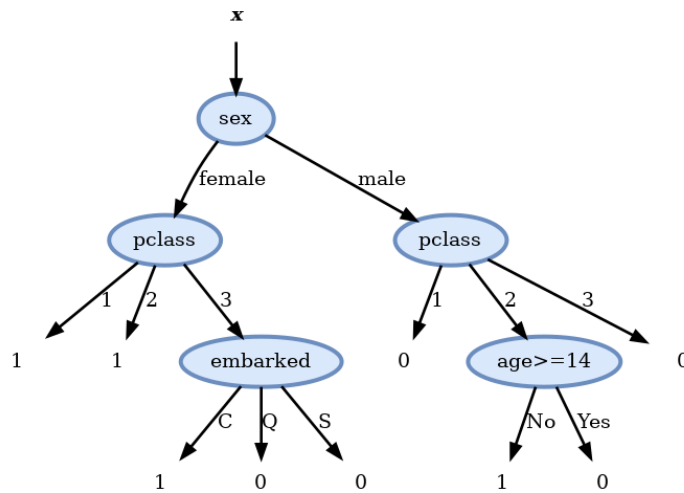
- embarked: 0.009
- age >= 15: 0.007 <-
- fare >= 151.55: 0.009

נמשיך עד שנמלא את כל השייבה השלישית ונקבל



## Pruning

לאחר חישוב העץ המלא נשתמש ב validation set על מנת להסיר את הענפים שלא משפרים (או פוגעים) בציון על ה validation set. בדיקה זו מראה שיש ארבעה node שלא תורמים לשיפור התוצאה ולכן נסיר אותם ונקבל את העץ הסופי הבא:



## ביצועים

נחשב את הציון (misclassification rate) המתקבל על ה test set:

- הציון על ה test set הינו: 0.205

זאת אומרת שיש לנו סיכוי של 80% לחזות נכונה האם אדם מסוים שרד או לא.