תרגול 4 - K-NN ו

Decision trees

PDF

תקציר התיאוריה

בעיות סיווג

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

- **בבעיות סיווג נהוג להתייחס לחזאי כאל מסווג (classifier) או discriminator (מקטלג).**
- **מחלקות - הערכים השונים שהתוויות מקבלות.**
- **את מספר המחלקות נסמן ב- C.**
- **סיווג בינארי - בעיות שיש בהן 2 מחלקות, $C=2$.**

- **בסיווג בינארי, מקובל לסמן את המחלקות באופנים הבאים:**
  - $y \in \{0, 1\}$.
  - $y \in \{-1, 1\}$.

- **בסיווג לא בינארי, מקובל להשתמש באחד מהאופציות הבאות לסימון המחלקות:**
  - $y \in \{1, 2, \ldots, C\}$.
  - $y \in \{0, 1, \ldots, C-1\}$.

דוגמאות

- **מערכת לזיהוי הונאות בכרטיסי אשראי:**

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.**

דוגמאות

- **מערכת לעיבוד כתב יד (OCR):**

במקרה זה x יכול להיות לדוגמא תמונה של אות ו y יהיה שווה למחלקה אשר מייצגת את האות בתמונה:

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

Misclassification rate

- לרוב בבעיות סיווג לא תהיה משמעות ל״מרחק״ בין החיזוי $\hat{y}$ לערך האמיתי של y.

- לדוגמא, בניסיון לזהות את האות a, חיזוי של האות b הוא לא בהכרח חיזוי טוב יותר מ s (למרות ש-b קרובה יותר ל-a באלפבית).

- לכן, נפוץ להשתמש ב- misclassification rate כפונ' המחיר

Misclassification rate

- פונקציית ה- loss המתאימה:

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

- ה- Risk עבור חזאי h:

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

החזאי האופטימאלי של הינו זה אשר מחזיר את ה y הכי סביר:

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

(K-NN (K-Nearest Neighbours

K-NN הינו אלגוריתם דיסקרימינטיבי לפתרון בעיות סיווג.

באלגוריתם זה החיזויים נעשים ישירות על פי המדגם באופן הבא:

בהינתן x מסויים:

1. **נבחר את K הדגימות בעלות ה $x^{(i)}$ הקרובים ביותר ל x.**

   (לרוב נשתמש במרחק אוקלידי, אך ניתן גם לבחור פונקציות מחיר אחרות).

2. **החיזוי יהיה התווית השכיחה ביותר מבין K התוויות של הדגימות שנבחרו בשלב 1.**

- במקרה של שוויון בשלב 2, נשווה את המרחק הממוצע בין ה x-ים השייכים לכל תווית ונבחר בתווית בעלת המרחק הממוצע הקצר ביותר.

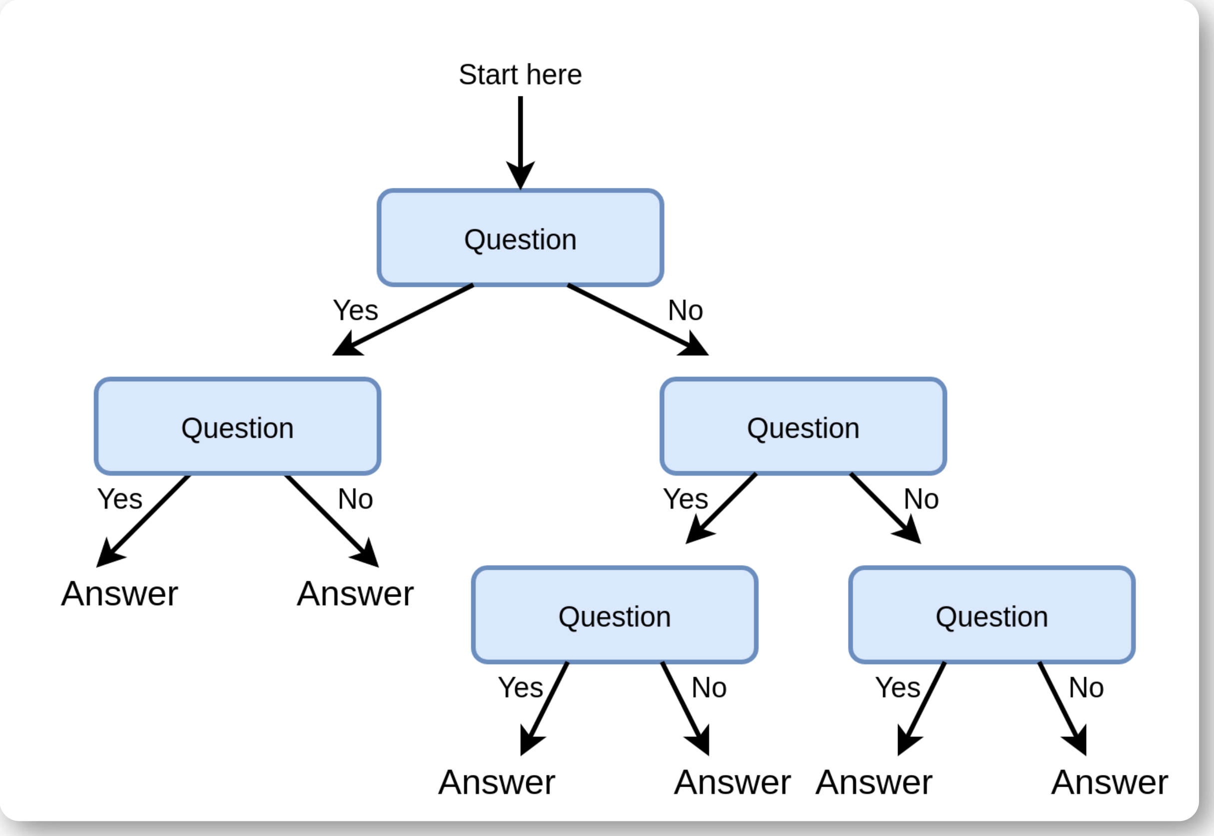
- במקרה של שוויון גם בין המרחקים הממוצעים, נבחר אקראית.

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.

(עצי החלטה) Decision trees

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

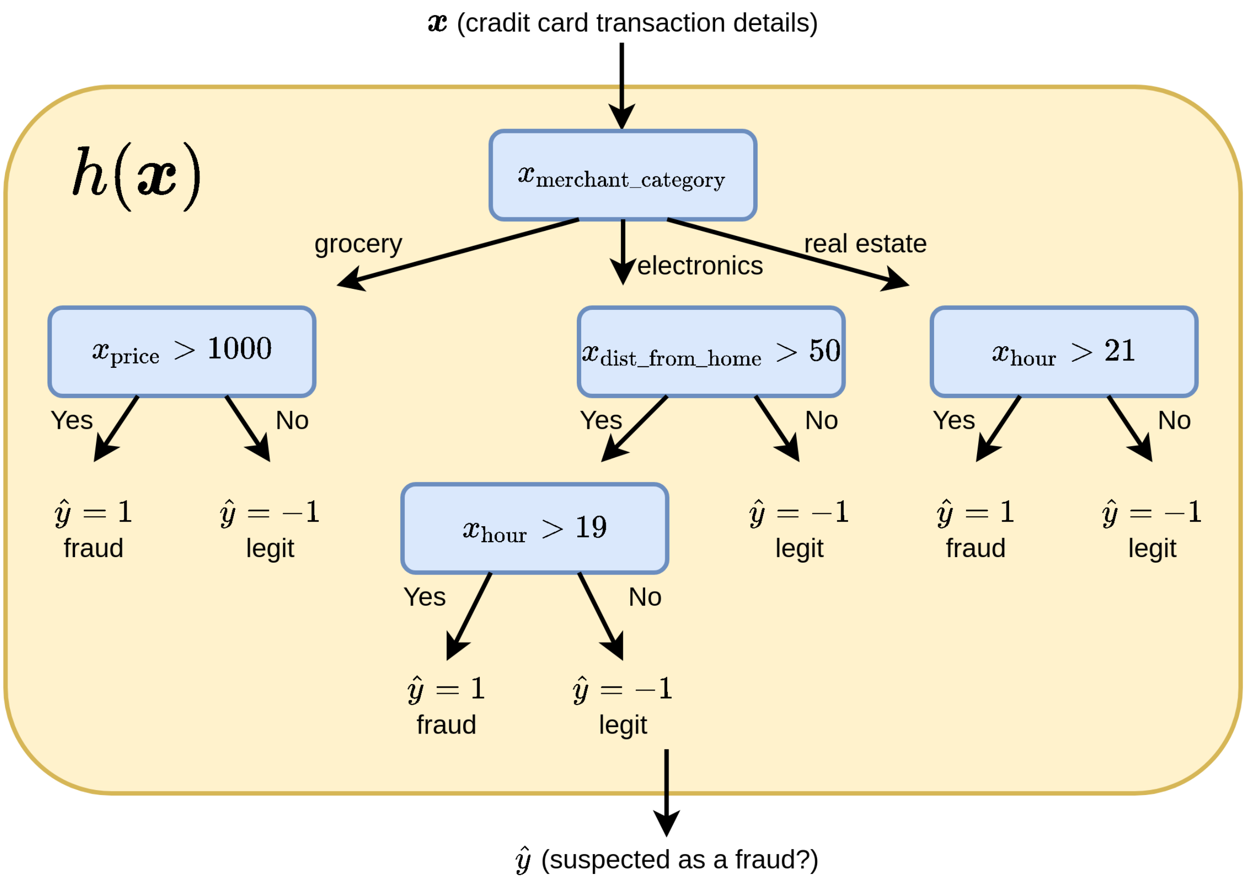


טרמינולוגיה:

- **root (שורש)** - נקודת הכניסה לעץ.
- **node (צומת)** - נקודות ההחלטה / 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 משתנה דיסקרטי אשר מקבל סט סופי של ערכים, נוכל גם לפצל לפי הערכים האפשריים של x_i


$\boldsymbol{x}$ (cradit card transaction details)
$h(\boldsymbol{x})$
$x_{\text{merchant_category}}$
grocery
electronics
real estate
$x_{\text{price}} > 1000$
$x_{\text{dist_from_home}} > 50$
$x_{\text{hour}} > 21$
Yes
No
Yes
No
Yes
No
$\hat{y} = 1$
fraud
$\hat{y} = -1$
legit
$x_{\text{hour}} > 19$
$\hat{y} = -1$
legit
$\hat{y} = 1$
fraud
$\hat{y} = -1$
legit
Yes
No
$\hat{y} = 1$
fraud
$\hat{y} = -1$
legit
$\hat{y}$ (suspected as a fraud?)

היתרונות של השימוש בעץ החלטה כחזאי:

1. פשוט למימוש (אוסף של תנאים else .. if ..).
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 מסויים.

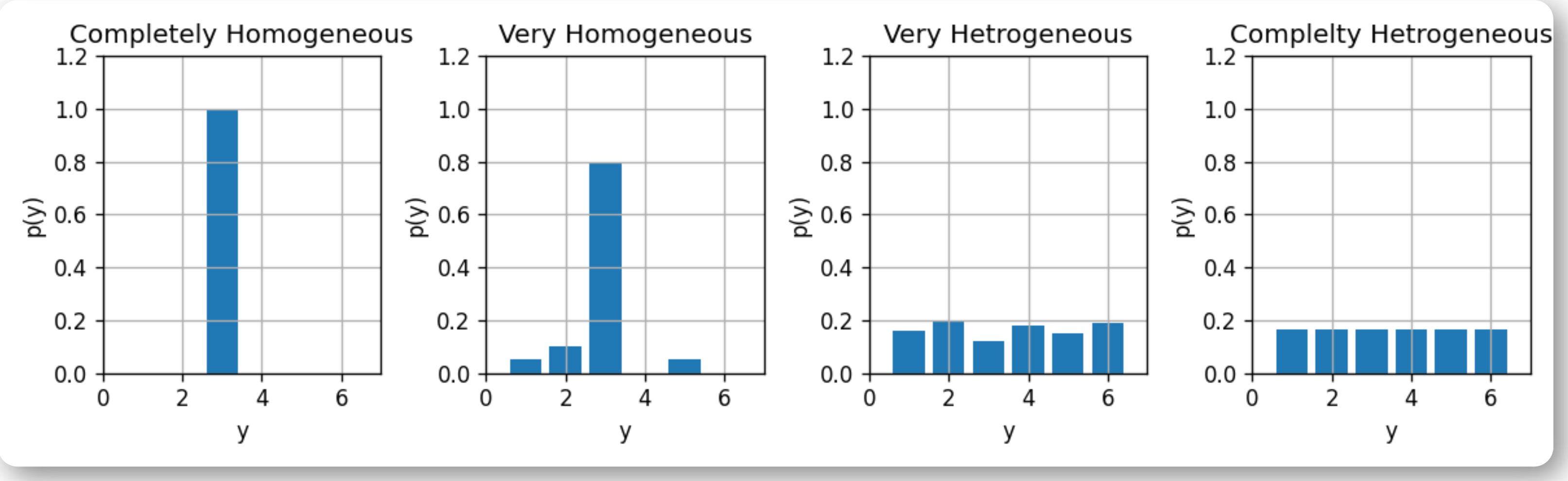
בניית עץ החלטה לסיווג

כיצד נוכל לבנות עץ החלטה על סמך מאגר של ריאליזציות שיש ברשותנו?

מדדים לחוסר ההומוגניות של פילוג

נתון:

- משתנה אקראי דיסקרטי y אשר מקבל את הערכים $y \in \{1, 2, \ldots, C\}$
- פונקציית הסתברות $p(y)$
- נגדיר כמה מדדים אשר בוחנים 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)

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

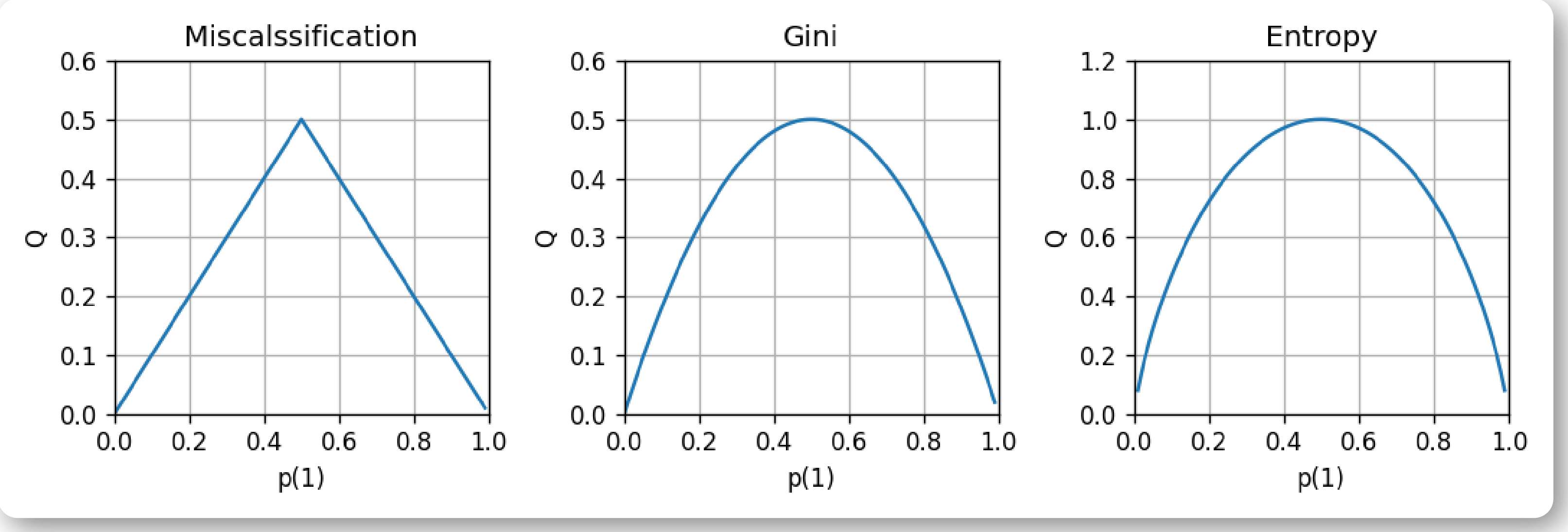
- אינדקס Gini:

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

- אנטרופיה:

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

- מדדים אלו שווים ל-0 עבור פילוגים **אחידים** (הומוגניים) וגדלים ככל שהפילוג הולך ונעשה אחיד.
- נראה את ההתנהגות של המדדים לעיל עבור משתנה אקראי בינארי:



חוסר הומוגניות ממוצעת של עץ - הציון של עץ החלטה

1. נעביר את הדגימות מהמדגם דרך העץ ונפצל אותם על פי העלים. 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. נחשב את הפילוג האמפירי של התויות שהגיעו לכל עלה:

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

3. נחשב את חוסר ההומוגניות של כל עלה: $Q(\hat{p}_j)$

4. הציון הכולל של העץ הינו הממוצע המשוקלל של חוסר ההומוגניות של העלים ביחס למספר הדגימות בכל עלה:

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

שלב ראשון - בניה של עץ מלא

- נרצה שהפילוג בעלים של העץ יהיה כמה שיותר הומוגני, כלומר שמדד חוסר ההומוגניות יהיה נמוך.
  - בעצם נרצה עץ מאוד ״החלטי״
  - בכדי להימנע מ overfitting, נרצה לעשות זאת על ידי שימוש בכמה שפחות צמתים (nodes).
- כדי למצוא את הפתרון האופטיאמלי יש לעבור על כל העצים האפשריים, אך חיפוש זה יכול לקחת זמן רב

- **פתרון: נחפש פתרון שאינו בהכרח האופטימאלי על ידי בניה של העץ בצורה חמדנית (greedy).**

- **נתחיל מה root ונוסיף nodes כך שבכל שלב נבחר את ה node אשר מניב את העץ עם מדד החוסר הומוגניות הנמוך ביותר.**

- **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.**

- **במקרים שבהם יש במדגם שתי דגימות עם אותו ה x אך y שונה, לא ניתן להגיע למדד חוסר הומוגניות 0.**

שלב שני - pruning (גיזום)

- כדי להקטין את ה- overfitting של העץ ניתן להשתמש ב- validation set על מנת לבצע pruning של העץ באופן הבא:
    - עבור כל אחד מהעלים:
    - לכל node, נבדוק 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

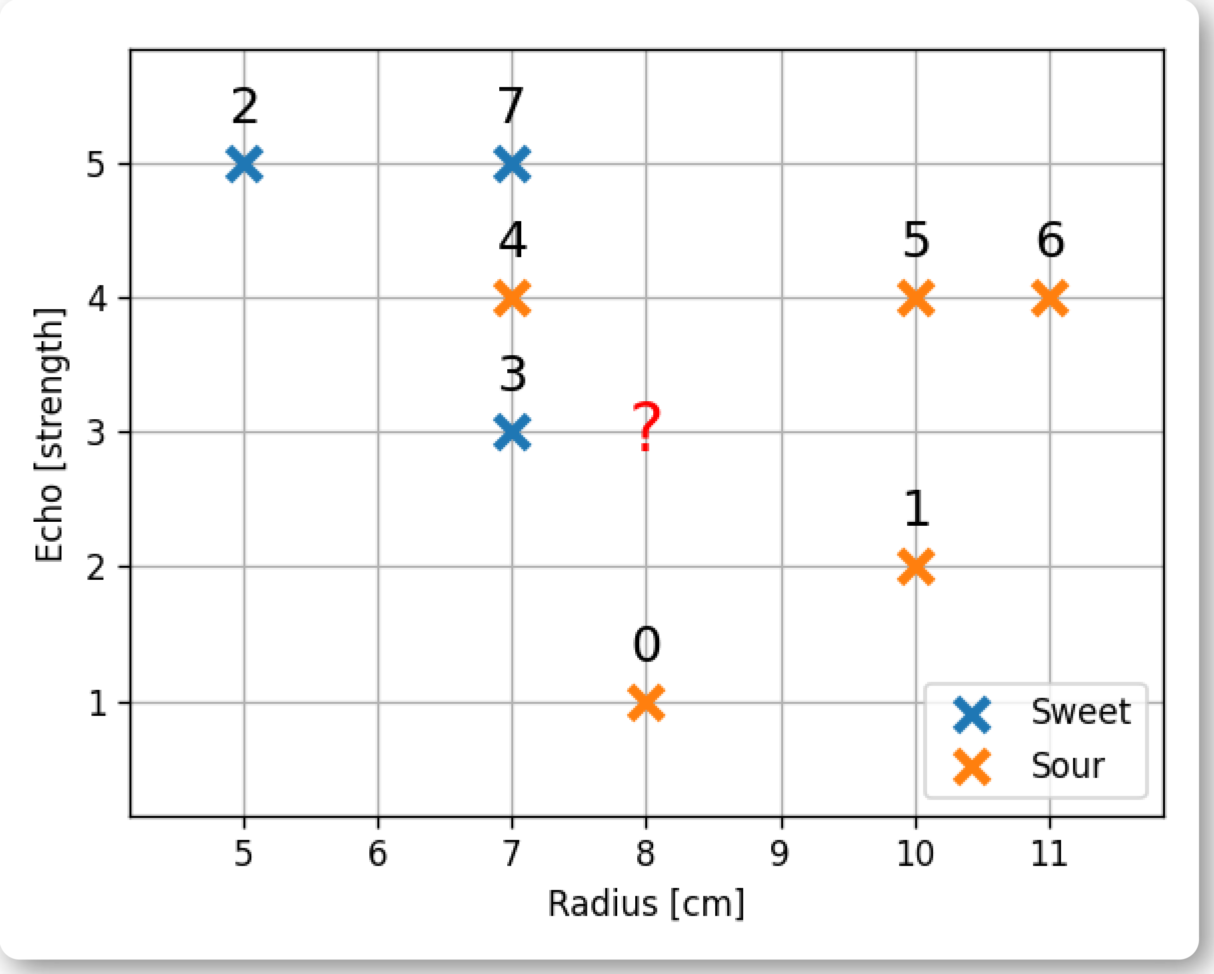
ניתן להשתמש בעצים גם לפתרון בעיות רגרסיה. במקרה של רגרסיה עם פונקציית מחיר של MSE, הבניה של העץ תהיה זהה מלבד שני הבדלים:

1. תוצאת החיזוי בעלה מסויים תהיה הערך הממוצע של התוויות באותו עלה. (במקום הערך השכיח)
2. את מדד חוסר ההומוגניות נחליף בשגיאה הריבועית של חיזוי 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
7	7	5	1



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

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

- **אם K קטן אנו למעשה מתאימים לכל נקודה ב-training set איזור החלטה משלה.**

- **במצב זה החזאי יתן חיזוי מושלם על המדגם, אך איזורי החלטה אלו, אשר תלויים במיקומים המקריים של נקודות בודדות, לא בהכרח ייצגו את האופי של הפילוג האמיתי. זהו המקרה של overfitting.**

- **כאשר K יהיה גדול מאוד אנו למעשה נמצע על איזורים מאוד גדולים ולכן החזאי יתעלם מהשינויים העדינים בפילוג של הנקודות, ויתייחס רק למגמה המאד כללית. זה המצב של underfitting.**

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

- **במקרה הקיצוני שבו K שווה לגודל ה dataset כל חיזוי יתבצע על סמך כל הנקודות במדגם ולכן יהיה שווה תמיד לתווית השכיחה ביותר במדגם.**

- **במקרה זה החיזוי יהיה תמיד -1, כלומר שהאבטיח אינו מתוק ללא כל תלות בהד וברדיוס.**

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 המחיר.

- כדי 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 אחת 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

- קיבלנו את השגיאה הממוצעת הקטנה ביותר עבור $K=3$.
- לכן נקבע את K לערך זה.

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

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

שאלה 4.2 – בניית עץ החלטה

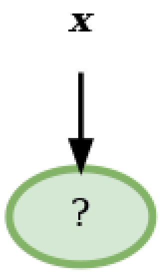
בנה עץ החלטה המבוסס על קריטריון האנטרופיה, אשר בהינתן נתוני צבע שער, גובה, משקל, והשימוש בקרם הגנה, חוזה האם עתיד האדם להכוות מהשמש היוקדת.

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
brown	tall	heavy	no	none
brown	average	heavy	no	none
blonde	short	light	yes	none

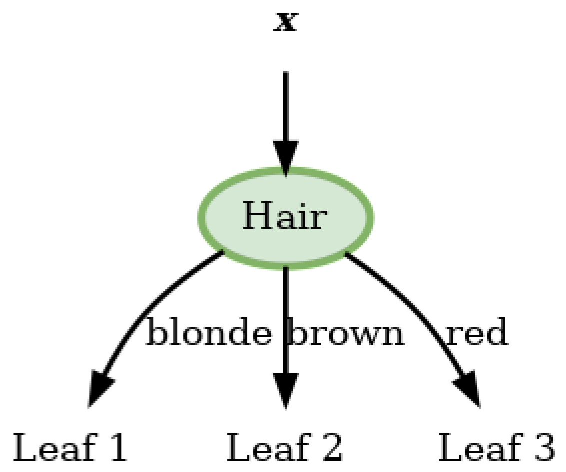
פתרון 4.2

- נפעל על פי האלגוריתם ונתחיל מה root ונתחיל להוסיף nodes:



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

Hair

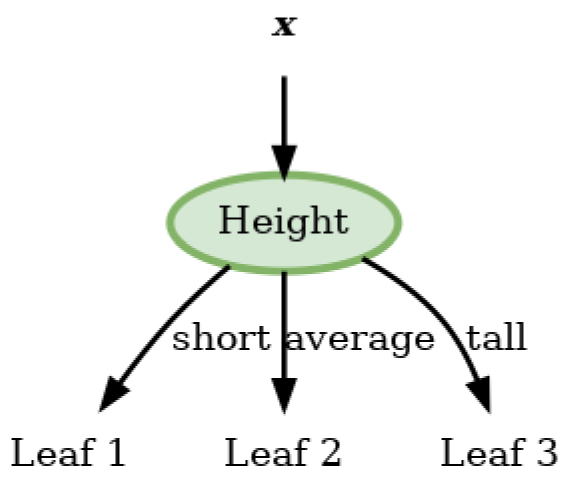


	Leaf (j)	N_j	$\hat{p}_j$	$H(\hat{p}_j)$
Blonde	**1**	**4**	$\{\frac{2}{4}, \frac{2}{4}\}$	$-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$
Brown	**2**	**3**	$\{\frac{0}{3}, \frac{3}{3}\}$	$-0\log(0) - 1\log(1) = 0$
Red	**3**	**1**	$\{\frac{1}{1}, \frac{0}{1}\}$	$-1\log(1) - 0\log(0) = 0$

נחשב את הממוצע הממושקל של האנטופיה על שלושת העלים:

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

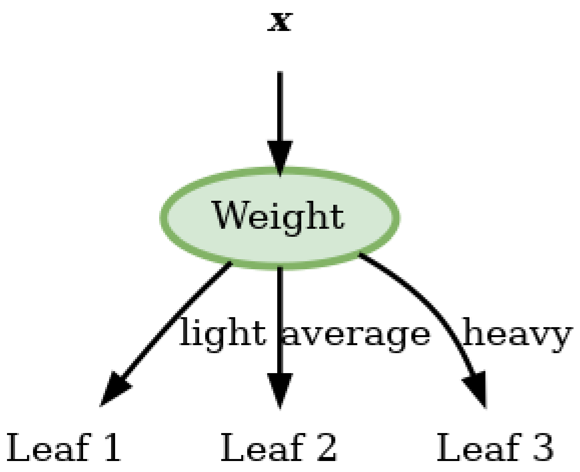
Height



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

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

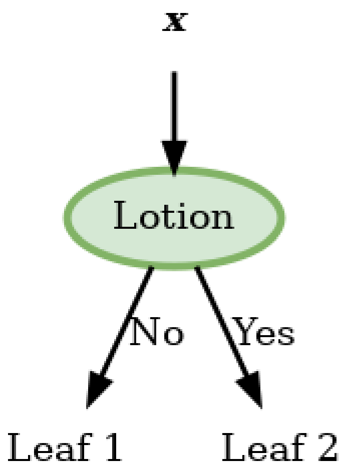
Weight



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



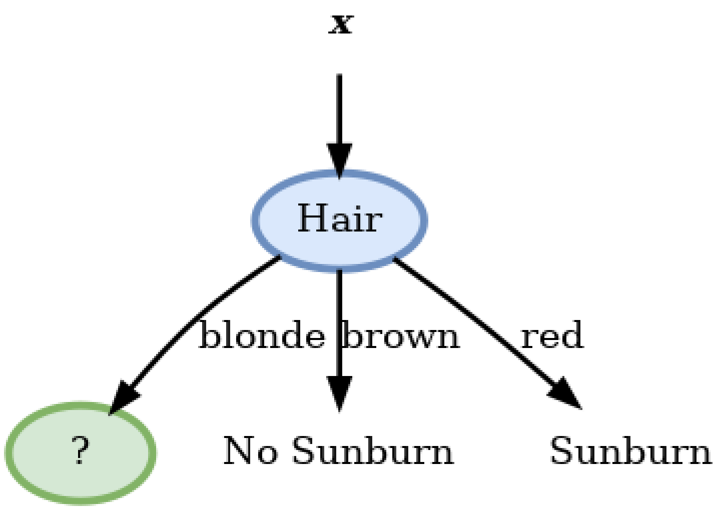
	Leaf (j)	N_j	$\hat{p}_j$	$H(\hat{p}_j)$
No	**1**	**5**	$\{\frac{3}{5}, \frac{2}{5}\}$	$-\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.97$
Yes	**2**	**3**	$\{\frac{0}{3}, \frac{3}{3}\}$	$-0\log(0) - 1\log(1) = 0$

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

- המאפיין האופטימלי לפיצול הראשון הוא **Hair**.
  - לכן נבחר בו להיות ה node הראשון.

- נשים לב כי בעבור node זה שני הפילוגים של brown ו red כבר הומוגניים לגמרי (מכילים רק סוג אחד של תוויות) ולכן לא נמשיך לפצל אותם.

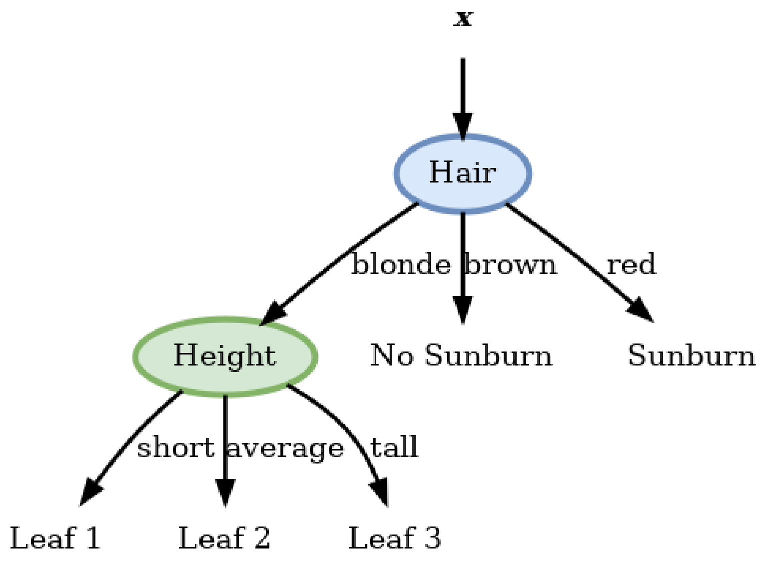
- נרשום את החיזוי המקבל בכל עלה:



- נמשיך כעת באופן דומה לבחור את ה node בעבור הענף של blonde.
- 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

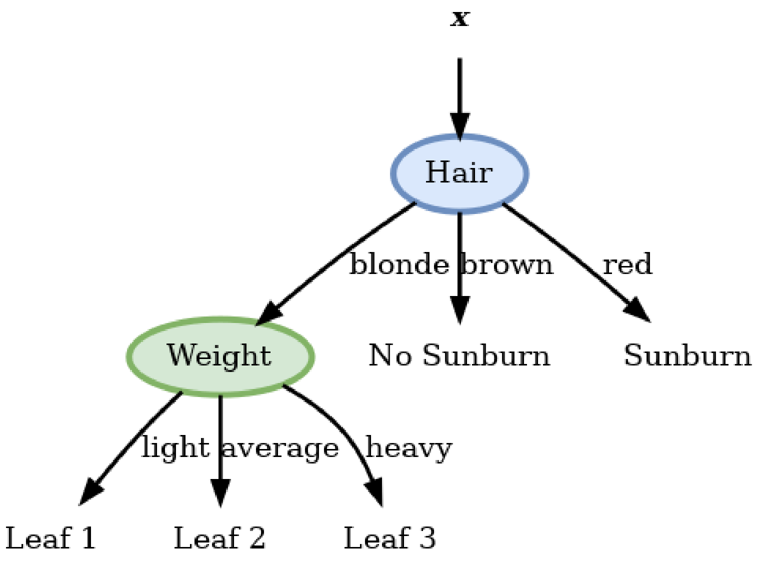
Height



	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**	$\{\frac{0}{1}, \frac{1}{1}\}$	$-1\log(1) - 0\log(0) = 0$

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

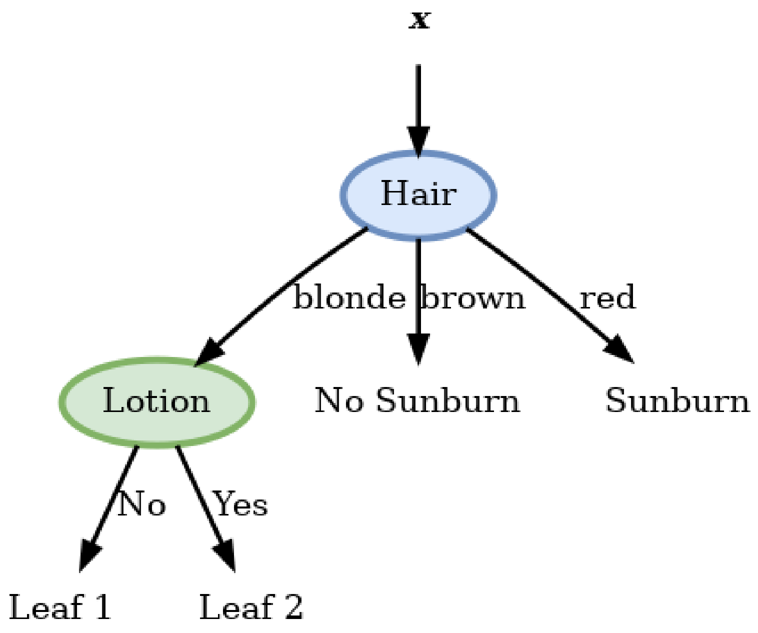
Weight



	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**		

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

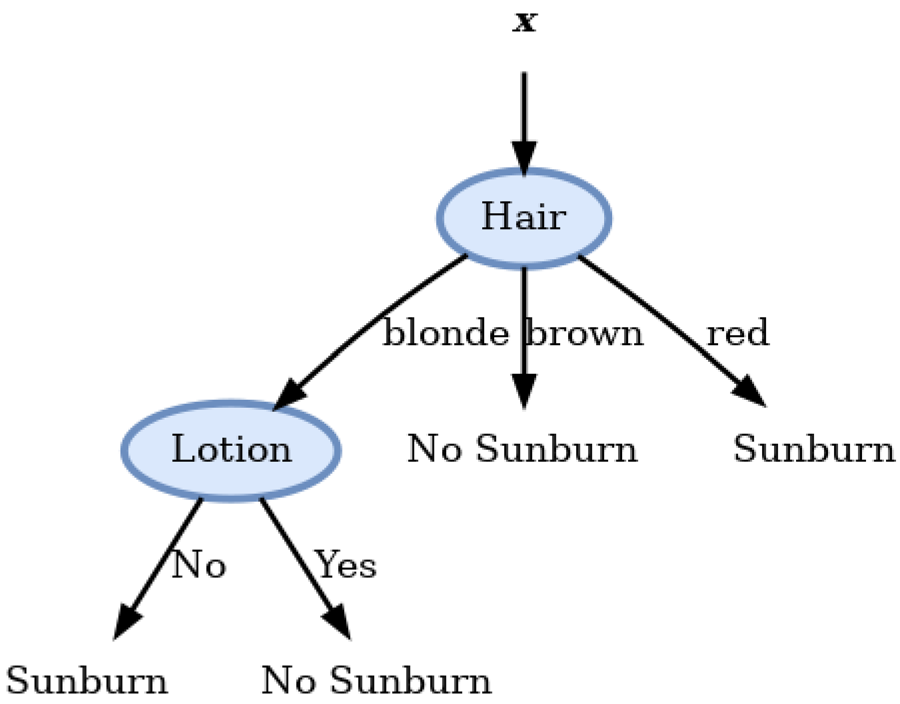
Lotion



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

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

- עץ ההחלטה הסופי יראה אם כן:



- עץ זה ממיין באופן מושלם את המדגם.

תרגיל 4.3

נתון המדגם הבא של ערכי תצפית של $\mathrm{x} = [\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3]^\top$ ותוויות y :

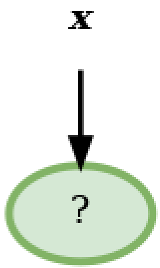
	x_1	x_2	x_3	y
1	1	1	-1	1
2	1	-1	-1	1
3	-1	-1	-1	1
4	-1	-1	-1	-1
5	1	1	1	-1

נרצה לבנות עץ החלטה על מנת לחזות את y על סמך x.
נרצה להשתמש במדד חוסר הומגניות חדש מסוג Square 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?

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**	$\{\frac{2}{3},\frac{1}{3}\}$	$2\sqrt{\frac{2}{3}\frac{1}{3}}=0.94$
-1	**2**	**2**	$\{\frac{1}{2},\frac{1}{2}\}$	$2\sqrt{\frac{1}{2}\frac{1}{2}}=1$

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

x_2

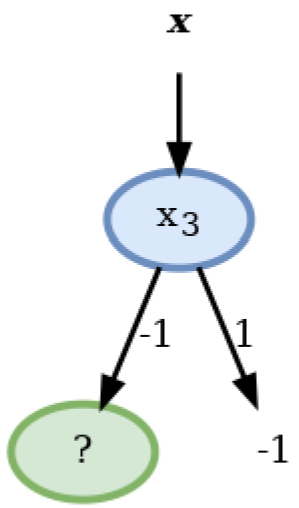
נשים לב ש node זה נותן חלוקה דומה של התוויות לזו של x_1 **ולכן נקבל את אותו ערך המדד של** 0.96**.**

x_3

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

	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**	$\{\frac{1}{2}, \frac{1}{2}\}$	$2\sqrt{\frac{1}{2}\frac{1}{2}} = 1$

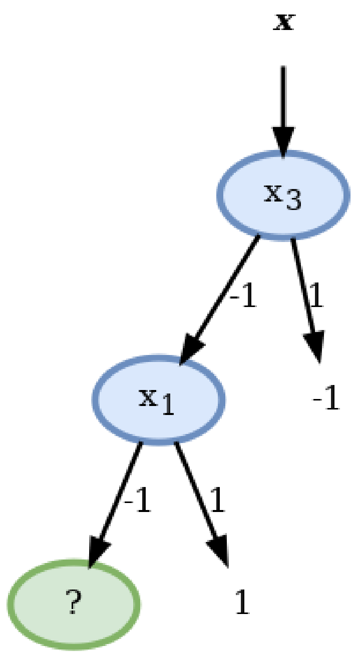
$$Q_{\text{branch}} = \frac{2}{5} \cdot 0 + \frac{2}{5} \cdot 1 = 0.4$$

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**	$\{\frac{2}{3}, \frac{1}{3}\}$	$2\sqrt{\frac{2}{3}\frac{1}{3}} = 0.94$

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

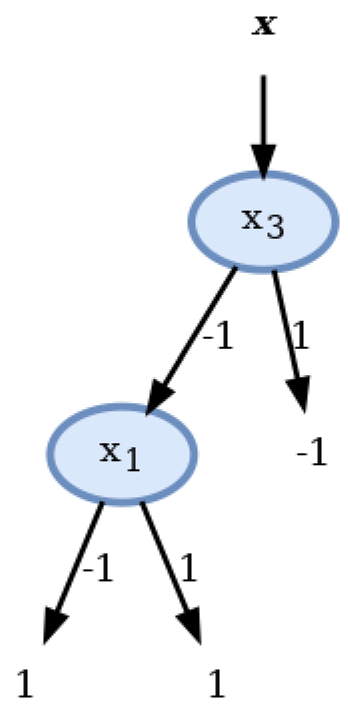
ה node האופטימאלי כאן הוא הפיצול לפי x_1 ונשים לב שהענף של $x_1 = 1$ כבר הומוגני:



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

	x_2	y
3	**-1**	**1**
4	**-1**	**-1**

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



בעץ שמצאנו ישנם 2 nodes.

2) חשבו את הציון (score) של עץ זה תחת פונקציית המחיר של misclassification rate. האם ניתן להגיע לסיווג מושלם במקרה זה?

- **בעבור העלים אשר הפילוג של התגיות בהם הינו הומוגני החיזוי יהיה מושלם.**

- **שגיאות חיזוי יתקבלו רק בעלה של $\mathrm{x}_3 = \mathrm{x}_1 = -1$ 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 $1/5 = 0.2$.**

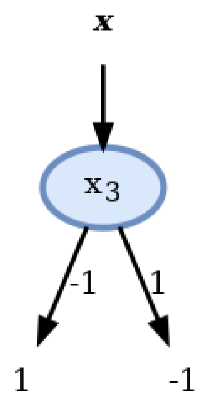
- **כפי שציינו קודם, מכיוון שלדגימות 3 ו 4 יש את אותו x אך y שונה, לא ניתן להפריד בניהם ותמיד על אחד מהם החיזוי יהיה לא נכון.**

- **לכן הציון של 0.2 הוא הציון המינימאלי (הטוב ביותר) שאותו ניתן לקבל על המדגם הזה.**

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 אפשריות למה האלגוריתם בו השתמשתם בסעיף הקודם לא מצא את העץ הזה.

- **נשים לב שלמעשה ה- node השני בעץ לא עושה כלום:**
  - **ללא תלות בערך של x_2 הוא חוזה 1**

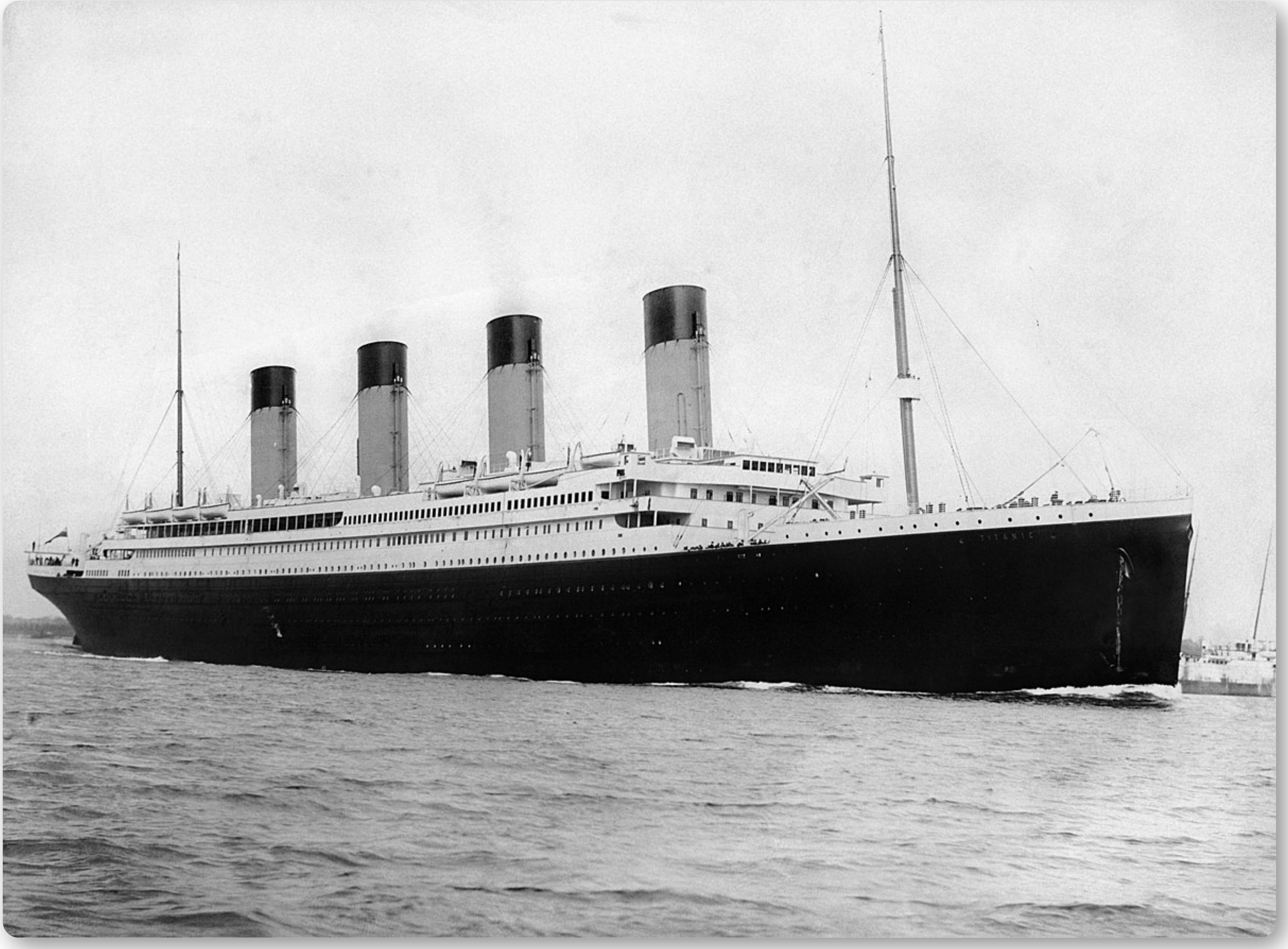
- **לכן, באותה המידה ניתן להשתמש גם בעץ הבא ולקבל את אותו החיזוי:**



מדוע האלגוריתם לא התכנס לפתרון זה?

- **בבניה של העץ ניסינו למזער את מדד ה squared root gini הממוצע ולא את שגיאת החיזוי. מכיוון שאלו שתי בעיות שונות, גם הפתרונות שלהן יכולים להיות שונים.**

חלק מעשי - הטיטניק



- **נשתמש ב- dataset שמבוסס על רשימת הנוסעים של ספינת הטיטניק.**
- **רשימה זו מכילה פרטים שונים על כל אחד מהנוסעים יחד עם אינדיקטור של איזה מהנוסעים שרד.**
- **ניתן להגדיר על סמך מדגם זה את בעיית supervised learning הבאה:**
  - **לחזות מי מהנוסעים שרד ומי לא על סמך פרטי הנוסע.**

- **את המדגם המקורי ניתן למצוא פה.**
- **אנו נעבוד עם גרסא יותר נקיה שלו שניתן למצוא פה.**

נציג את 10 השורות הראשונות במדגם:

body	boat	embarked	cabin	fare	ticket	parch	sibsp	age	sex	name	survived	pclass	
nan	2	S	B5	211.338	24160	0	0	29	female	Allen, Miss. Elisabeth Walton	1	1	0
nan	nan	S	C22 C26	151.55	113781	2	1	2	female	Allison, Miss. Helen Loraine	0	1	1
135	nan	S	C22 C26	151.55	113781	2	1	30	male	Allison, Mr. Hudson Joshua Creighton	0	1	2
nan	nan	S	C22 C26	151.55	113781	2	1	25	female	Allison, Mrs. Hudson J C (Bessie Waldo (Daniels	0	1	3
nan	3	S	E12	26.55	19952	0	0	48	male	Anderson, Mr. Harry	1	1	4
nan	10	S	D7	77.9583	13502	0	1	63	female	Andrews, Miss. Kornelia Theodosia	1	1	5
nan	nan	S	A36	0	112050	0	0	39	male	Andrews, Mr. Thomas Jr	0	1	6
nan	D	S	C101	51.4792	11769	0	2	53	female	Appleton, Mrs. Edward Dale (Charlotte (Lamson	1	1	7
22	nan	C	nan	49.5042	PC 17609	0	0	71	male	Artagaveytia, Mr. Ramon	0	1	8

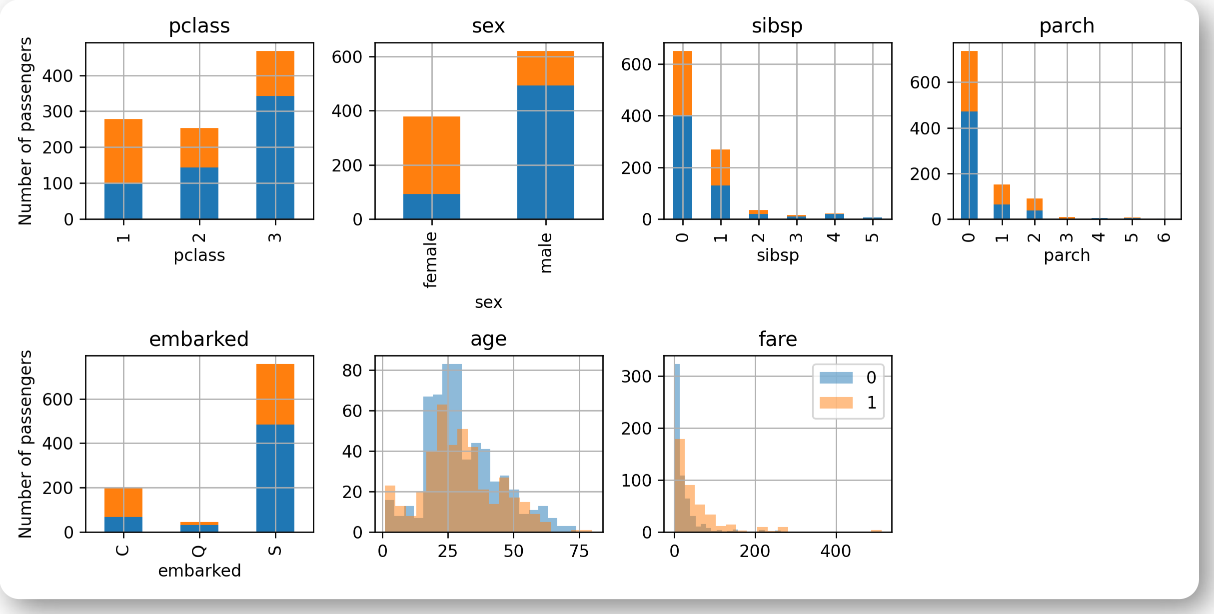
השדות

בתרגול נשתמש 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 האוניה
- **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 ששרדו ובעבור אלו שלא:



- ניתן לראות כי אכן ישנם מאפיינים שיוכלו לסייע לשנו לשפר את החיזוי שלנו.

- לדוגמא: לנשים היה סיכוי גבוהה בהרבה לשרוד מאשר גברים 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 לא.**

- **נרצה למצוא חזאי (מסווג) שיהיה טוב כל האפשר תחת פונקציית המחיר miscalssification rate.**

- **נעשה זאת בעזרת עץ החלטה**

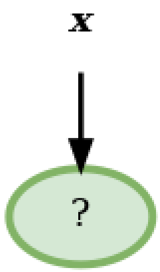
חלוקת ה dataset

- נחלק את המדגם ל 80% train set ו 20% test set.
- נחלק את ה train set פעם נוספת ל 75% train set ו 25% validation set.

בניית עץ בעל שלוש רמות

- נבנה את העץ על פי קריטריון Gini.
- נתחיל מה root ונוסיף בכל פעם את ה node שממזער את המדד.

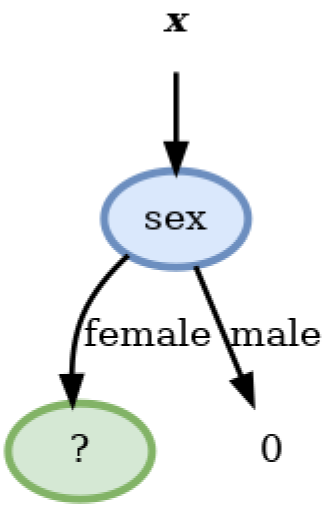
- בעבור ה- node הראשון:



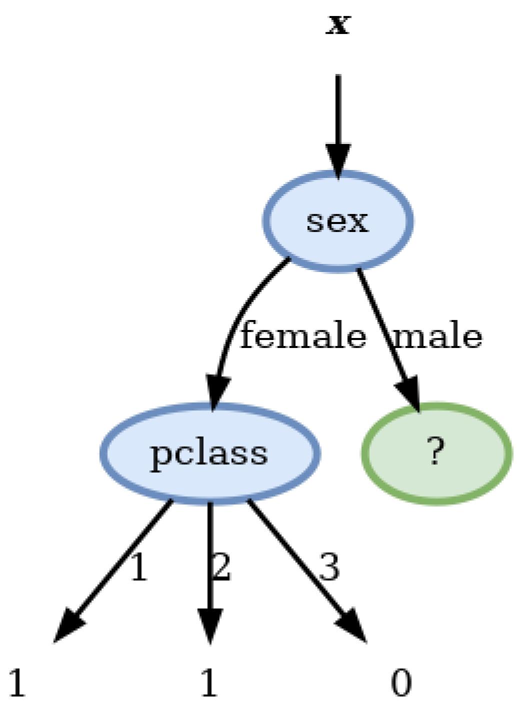
```
Score before split: 0.492
Scores:
- pclass: 0.436
- sex: 0.360   <-
- sibsp: 0.479
- parch: 0.473
- embarked: 0.460
- age >= 9: 0.488
- fare >= 15.7417: 0.448
```

* לכן נבחר למיין לפי המין.

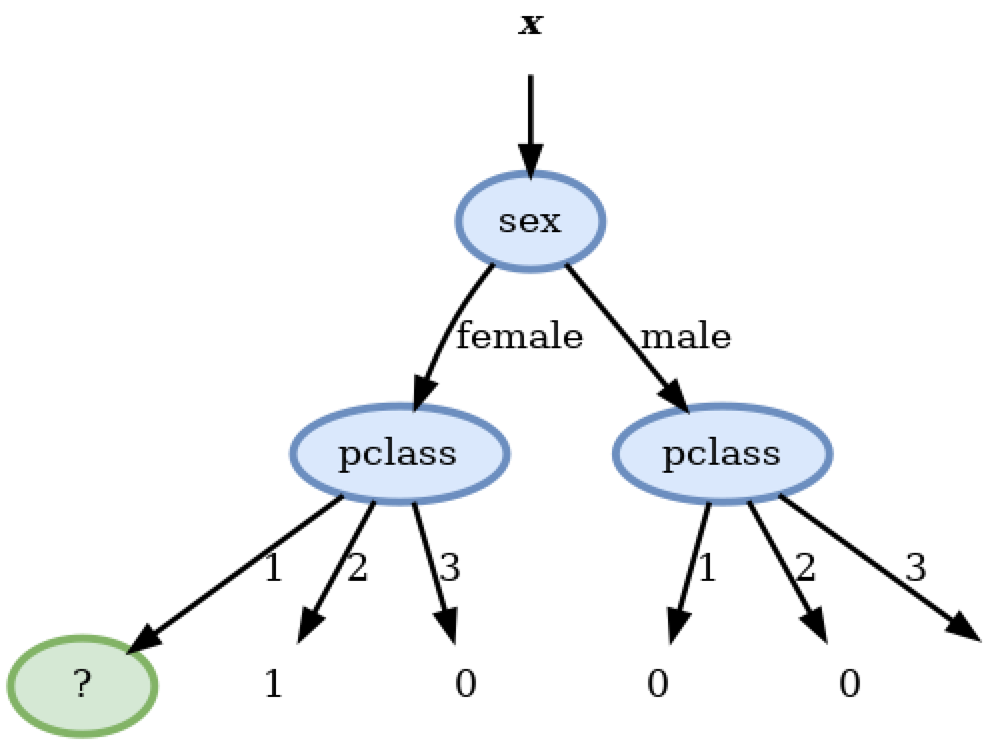
- נפעל באופן זהה לכל שאר ה nodes



```
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
```

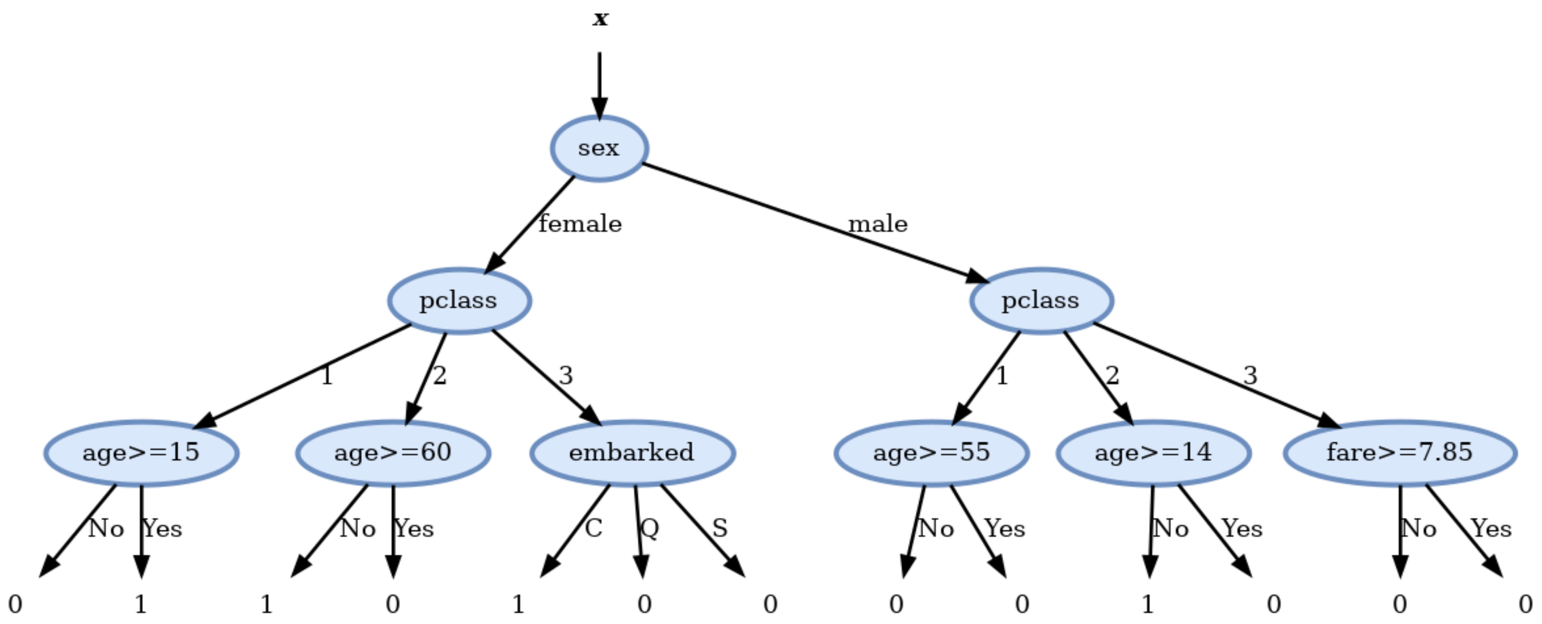

x
sex
female
male
pclass
?
1
2
3
1
1
0

```
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
```


x
sex
female
male
pclass
pclass
1
2
3
1
2
3
?
1
0
0
0
0

```
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
```

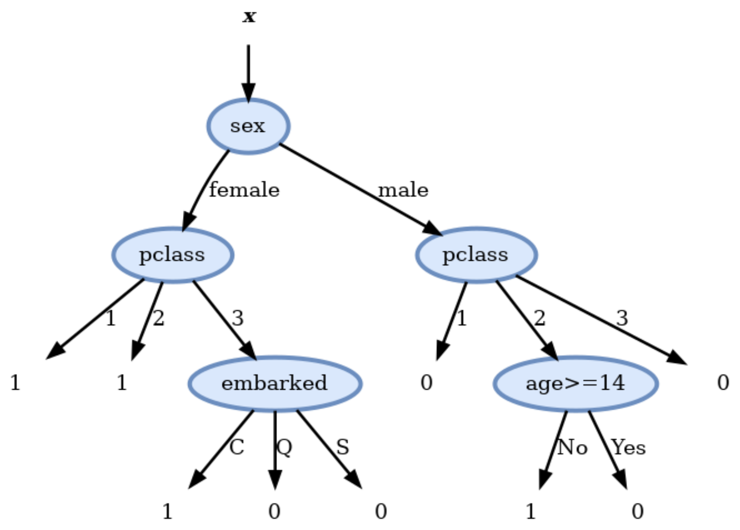
- נמשיך עד שנמלא את כל השכבה השלישית ונקבל:



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 nodes 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 מסויים שרד או לא.