

Decision tree

1. This question demonstrates that ID3 does not always yield an optimal tree. Consider the following dataset:

ID	Feature A	Feature B	Feature C	Target
1	1	1	1	+1
2	1	0	0	+1
3	1	1	0	-1
4	0	0	1	-1

1.1. Run ID3 manually up to depth 2. In the case of a leaf with non-zero entropy predict the majority label in that leaf. In case of a tie, predict “-1”. The nodes and edges in the decision tree should be labeled appropriately. You can use your preferred graphics software to draw the tree, or you can hand draw it (neatly). Calculate the training error on the given dataset and state which tuples were misclassified.

נקודות:

- אם האנטרופיה לא שווה לאפס נבחר לנבא לפי הלייבל הנפוץ בעלה.
- במקרה של שוויון נבחר את הלייבל -1.

ID3:

נשתמש בנוסחאות הבאות :

$$\operatorname{argmax}_a IG(v = \{1,2,3,4\}, a) = \operatorname{argmax}_a (H(v) - \frac{|v_{a=T}|}{|v|} H(v_{a=T}) - \frac{|v_{a=F}|}{|v|} H(v_{a=F}))$$

$$H(v) = -p_v \log_2 p_v - (1 - p_v) \log_2 (1 - p_v)$$

$$p_v = \frac{|\{x, y\} \in v \mid y = 1\}|}{|v|}$$

נכין טבלה כמו שראינו בתרגול בשביל למצוא את הפיצ'ר שיתן את ה information gain המקסימלי בכל פיצול וכך נקבל אנטרופיה קטנה בכל פיצול.

$$H(v) = \frac{1}{2} : \text{האנטרופיה ההתחלתית היא}$$

Attribute	$\frac{ v_{a=1} }{ v }$	$\frac{ v_{a=0} }{ v }$	$H(v_{a=1})$ אנטרופיה אחרי פיצול	$H(v_{a=0})$ אנטרופיה אחרי פיצול	$IG(v, a) - H(v)$
Feature A	$\frac{3}{4}$	$\frac{1}{4}$	$H(\frac{2}{3})$	$H(0)$	$-\frac{3}{4}H(\frac{2}{3}) - \frac{1}{4}H(0) = -0.6885$
Feature B	$\frac{1}{2}$	$\frac{1}{2}$	$H(\frac{1}{2})$	$H(\frac{1}{2})$	$-\frac{1}{2}H(\frac{1}{2}) - \frac{1}{2}H(\frac{1}{2}) = -1$
Feature C	$\frac{1}{2}$	$\frac{1}{2}$	$H(\frac{1}{2})$	$H(\frac{1}{2})$	$-\frac{1}{2}H(\frac{1}{2}) - \frac{1}{2}H(\frac{1}{2}) = -1$

חישובים :

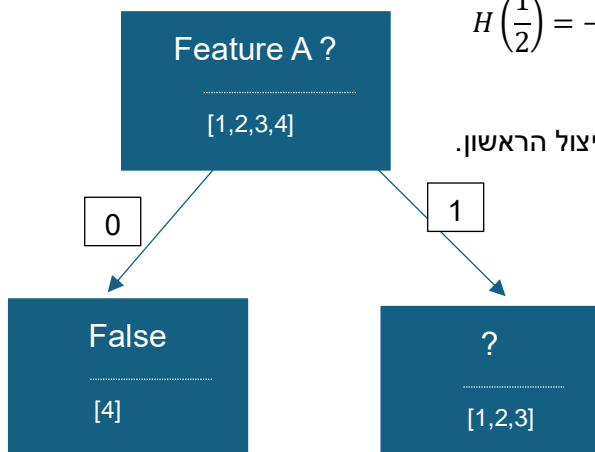
$$H\left(\frac{2}{3}\right) = -\frac{2}{3} \log \frac{2}{3} - \frac{1}{3} \log \frac{1}{3} = 0.918$$

$$H\left(\frac{1}{2}\right) = -\frac{1}{2} \log \frac{1}{2} - \frac{1}{2} \log \frac{1}{2} = 1$$

ניתן לראות ש-IG הגדול הוא של feature A ולכן נבחר לפצל על פיו בפיצול הראשון.

נסתכל על הפיצרים שנשארו לנו :

ID	Feature B	Feature C	Target
1	1	1	+1
2	0	0	+1
3	1	0	-1

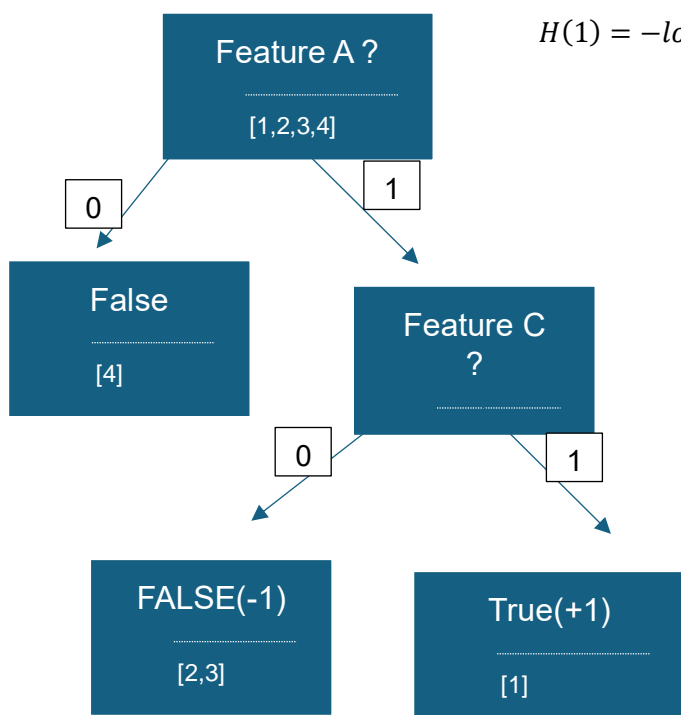


Attribute	$\frac{ v_{a=1} }{ v }$	$\frac{ v_{a=0} }{ v }$	$H(v_{a=1})$ אנטרופיה אחר פיצול	$H(v_{a=0})$ אנטרופיה אחר פיצול	$IG(v, a) - H(v)$
Feature B	$\frac{2}{3}$	$\frac{1}{3}$	$H\left(\frac{1}{2}\right)$	$H(1)$	$-\frac{2}{3}H\left(\frac{1}{2}\right) - \frac{1}{3}H(1) = -\frac{2}{3}$
Feature C	$\frac{1}{3}$	$\frac{2}{3}$	$H(1)$	$H\left(\frac{1}{2}\right)$	$-\frac{1}{3}H(1) - \frac{2}{3}H\left(\frac{1}{2}\right) = -\frac{2}{3}$

$$H(1) = -\log 1 - \log 1 = 0$$

קיבלנו ש-IG של שני הפיצרים שווה, לכן נבחר אחד מהם- נבחר את פיצ'ר C

התבקשנו לבנות עד עומק 2 ולכן אין צורך להמשיך.



ID	Feature B	Target
2	0	+1
3	1	-1

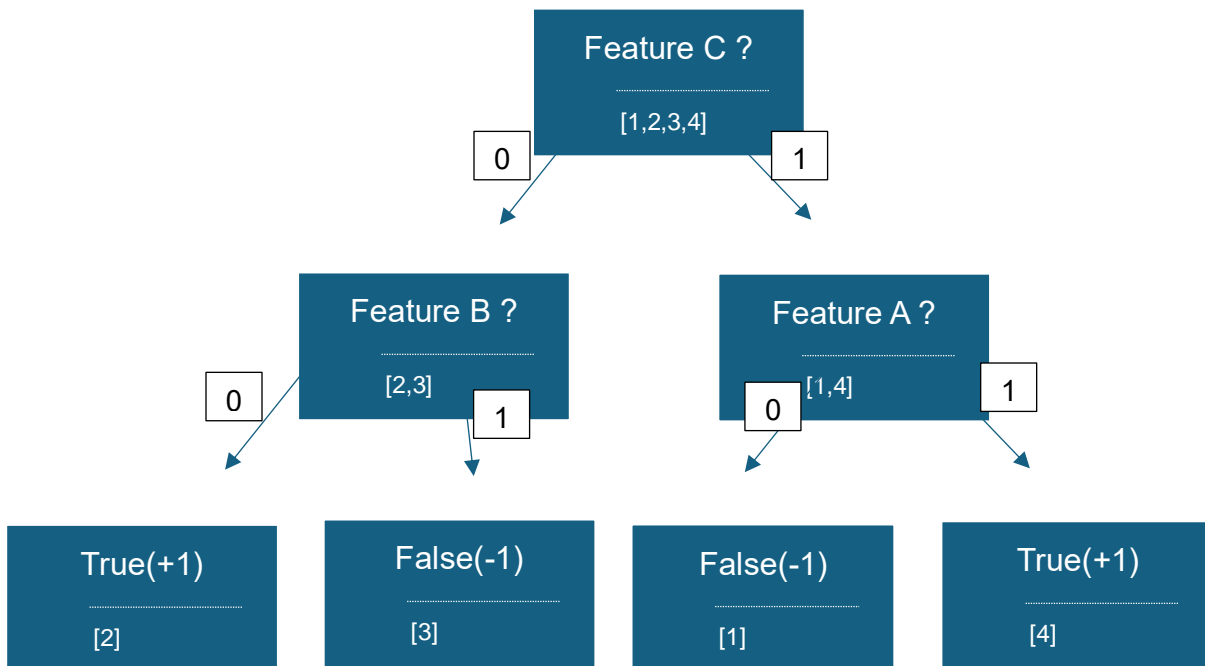
האנטרופיה לא שווה לאפס ולכן נבא לפי הלייבל הכי נפוץ, מכיוון שכל לייבל מופיע פעם אחת, כלומר בכמות שווה, נבחר את הלייבל להיות -1- כמו שהתבקשנו בנתונים לשאלה.

נחשב את השגיאה האמפירית :

$$Empirical_{Error} = \frac{1}{m} \sum_{i=1}^m I[y_i \neq h(x_i)] = \frac{1}{4} \sum_{i=1}^4 I[y_i \neq h(x_i)] = \frac{1}{4}$$

data של סוג נכון הוא זה עם ID=2.

1.2. Find (manually) a decision tree of depth 2 that fits the training data perfectly.



ניתן לראות שהעץ בעומק 2 שקיבלנו תואם בצורה מושלמת את ה training data בשאלה.

2. We now wish to prove that the information gain, i.e., $IG(v, a) = H(v) - \frac{|v_{a=T}|}{|v|} H(v_{a=T}) - \frac{|v_{a=F}|}{|v|} H(v_{a=F})$,

is non-negative for any set of examples v and a binary attribute a .

Recall our definition of $v_{a=T} \triangleq \{(x_i, y_i) \in v \mid x_i[a] = T\}$ and $v_{a=F} \triangleq \{(x_i, y_i) \in v \mid x_i[a] = F\}$.

See the following dialogue, where we (really) asked [ChatGPT](#) to prove this (we shortened it a little):

Question: could you prove that the information gain in id3 cannot be negative?

Answer: Yes, I can prove this using mathematical reasoning.

In the ID3 algorithm, the information gain (IG) of an attribute A with respect to a dataset D is defined as: $IG(A) = Entropy(D) - \sum(|D_v| / |D| * Entropy(D_v))$ where $Entropy(D)$ is the entropy of the original dataset D, $|D_v|$ is the number of instances in the subset of D that have a specific value v for the attribute A, and $Entropy(D_v)$ is the entropy of that subset.

Entropy is a measure of the randomness or unpredictability of the target variable, and it is calculated as: $Entropy(D) = - \sum(p(c) * \log_2(p(c)))$ where $p(c)$ is the proportion of instances in D that belong to class c.

Now, let's assume that the IG for a given attribute A is negative, i.e., $IG(A) < 0$.

Then we have: $Entropy(D) - \sum(|D_v| / |D| * Entropy(D_v)) < 0$

Rearranging the terms, we get: $Entropy(D) < \sum(|D_v| / |D| * Entropy(D_v))$

Since both $Entropy(D)$ and $Entropy(D_v)$ are non-negative, the right-hand side of the inequality is also non-negative.

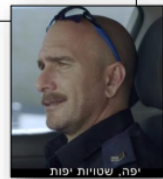
Therefore, we have: $Entropy(D) \geq 0$.

This contradicts the assumption that $IG(A) < 0$, because $Entropy(D) - \sum(|D_v| / |D| * Entropy(D_v))$ is the value of $IG(A)$.

Thus, we can conclude that the information gain in the ID3 algorithm is always non-negative and cannot be negative.

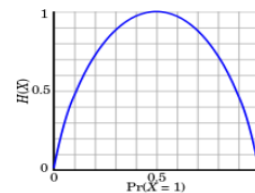
- 2.1. Attached is a quote from "ארץ נהדרת" that best describes the answer above.

Briefly explain what the problem with the above "proof" is.



$$H(v) = H\left(\frac{1}{|v|} \{(x, y) \in v \mid y = 1\}\right) = H(p_v) = -p_v \log_2 p_v - (1 - p_v) \log_2 (1 - p_v)$$

ראשית נשים לב שהאנטרופיה לא יכולה להיות שלילית כי p_v הוא שבר, מספר בין 0 ל-1 \log על מספר בין 0 ל-1 הוא מספר שלילי ולכן התוצאה של האנטרופיה היא מספר בין 0 ל-1, גם ראינו בתרגול את הפונקציה של האנטרופיה שהיא לא יכולה



להיות שלילית.

(וגם מבחינה משמעות של אנטרופיה שהיא מודדת את האי ודאות, אין משמעות לערכים שליליים).
Chat GPT ניסה להוכיח על ידי הנחה בשלילה, כלומר שאם IG הוא שלילי אז מגיעים לסתירה, אבל לא הגיע לשום סתירה.
בהתחלה הוא הניח בשלילה ש:

Now, let's assume that the IG for a given attribute A is negative, i.e., $IG(A) < 0$.

Then we have: $Entropy(D) - \sum(|D_v| / |D| * Entropy(D_v)) < 0$

Rearranging the terms, we get: $Entropy(D) < \sum(|D_v| / |D| * Entropy(D_v))$

לאחר מכן הוא ציין שהאנטרופיה לא יכולה להיות שלילית וזה באמת נכון (הסברנו למעלה) ולכן צד שמאל של המשוואה גדול שווה לאפס

Since both $Entropy(D)$ and $Entropy(D_v)$ are non-negative, the right-hand side of the inequality is also non-negative.

Therefore, we have: $Entropy(D) \geq 0$.

אבל זאת לא סתירה, כי גם אם מניחים זאת האי שווין ממשיך להתקיים.
אבל למרות זאת הוא החליט שזאת סתירה,

This contradicts the assumption that $IG(A) < 0$, because $Entropy(D) - \sum([|D_v| / |D| * Entropy(D_v)])$ is the value of $IG(A)$.

ולכן,



2.2. Let us really prove this.

Let v be an arbitrary set of examples and a be a binary attribute (feature).

Assume the classes are also binary.

The entropy is thus: $H(v) = H\left(\frac{1}{|v|}\{(x, y) \in v \mid y = 1\}\right) \triangleq H(p_v) = -p_v \log_2 p_v - (1 - p_v) \log_2 (1 - p_v)$.

(Notice that like in the tutorial, we use H with a set v or with a probability p_v , interchangeably.)

Prove: $IG(v, a) \geq 0$.

You can use the following property without proving it.

Property: It holds $\forall \beta_1, \beta_2, \alpha \in [0, 1]$ that: $-\alpha H(\beta_1) - (1 - \alpha) H(\beta_2) \geq -H(\alpha \beta_1 + (1 - \alpha) \beta_2)$.

$$H(v) = H\left(\frac{1}{|v|}\{(x, y) \in v \mid y = 1\}\right) = H(p_v) = -p_v \log_2 p_v - (1 - p_v) \log_2 (1 - p_v)$$

Information gain:

$$IG(v, a) = H(v) - \frac{|v_{a=T}|}{|v|} H(v_{a=T}) - \frac{|v_{a=F}|}{|v|} H(v_{a=F}) \geq H(v) - H\left(\frac{|v_{a=T}|}{|v|} p_{(v_{a=T})} + \frac{|v_{a=F}|}{|v|} p_{(v_{a=F})}\right)$$

המעבר הזה הוא על פי הטענה בשאלה.
נסתכל על,

$$p_{v_{a=T}} = \frac{1}{|v_{a=T}|} |\{(x, y) \in v_{a=T} \mid y = 1\}|$$

$$p_{v_{a=F}} = \frac{1}{|v_{a=F}|} |\{(x, y) \in v_{a=F} \mid y = 1\}|$$

ונציב,

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

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

מתקיים :

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

ולכן נקבל (כי $H(v) = H\left(\frac{1}{|v|}\{(x, y) \in v \mid y = 1\}\right) = H(p_v)$)

$$H(v) - H(v) = 0$$

כלומר,

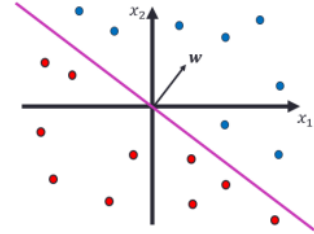
$$IG(v, a) \geq 0$$

Linear classification

In tutorial 02, we illustrated algebraic and geometric properties of linear classifiers. Specifically, we explained that a homogeneous linear classifier $h: \mathbb{R}^d \rightarrow \mathbb{R}$ is parameterized by a normal vector $\mathbf{w} \in \mathbb{R}^d$, such that:

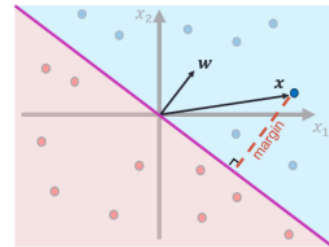
$$h(\mathbf{x}) = +1 \Leftrightarrow \mathbf{w}^T \mathbf{x} = \underbrace{\|\mathbf{w}\| \|\mathbf{x}\|}_{>0} \cos \angle(\mathbf{w}, \mathbf{x}) > 0 \Leftrightarrow \cos \angle(\mathbf{w}, \mathbf{x}) > 0.$$

Notice how the decision rule is invariant to the scale of \mathbf{w} , and only considers its direction.



Furthermore, we showed that the geometric (signed) margin of $\mathbf{x} \in \mathbb{R}^d$ is:

$$\frac{\mathbf{w}^T \mathbf{x}}{\|\mathbf{w}\|}$$



3. A nonhomogeneous linear classifier $h: \mathbb{R}^d \rightarrow \mathbb{R}$ is parameterized by a vector $\mathbf{w} \in \mathbb{R}^d$ and a bias $b \in \mathbb{R}$, such that:

$$h(\mathbf{x}) = +1 \Leftrightarrow \mathbf{w}^T \mathbf{x} > b.$$

Notice how b can almost be interpreted as the minimal margin required for a positive prediction (we say “almost” since $\mathbf{w}^T \mathbf{x}$ is not the margin, but rather $\frac{\mathbf{w}^T \mathbf{x}}{\|\mathbf{w}\|}$ is).

- 3.1. In the homogeneous case, we had a degree of freedom while choosing \mathbf{w} (due to the scale invariance).

What similar degree of freedom do we have while choosing \mathbf{w}, b in the non-homogeneous case?

That is, how can \mathbf{w}, b be changed without changing the classifier itself? Briefly explain your answer.

במקרה ההומוגני יש דרגת חופש כי נוכל לבחור את \mathbf{w} המתאים למסווג הלינארי ולכל $\alpha \in \mathbb{R}$, $\alpha \mathbf{w}$ תהיה תשובה למסווג זה.

באותו אופן, במקרה הלא הומוגני נוכל לבחור לכל \mathbf{w}, b , $\alpha \in \mathbb{R}$, כך ש $\alpha b, \alpha \mathbf{w}$ יהיו מתאימים לאותו מסווג לינארי.

$$h(\mathbf{x}) = +1 \Leftrightarrow \mathbf{w}^T \mathbf{x} > -b$$

כאשר נמצא את \mathbf{w} שמקיים את זה, אז המשוואה תשמר גם לכל $\alpha \in \mathbb{R}_{>0}$, ולכן גם $\alpha b, \alpha \mathbf{w}$ מקיימים:

$$h(\mathbf{x}) = +1 \Leftrightarrow \alpha \mathbf{w}^T \mathbf{x} > -\alpha b$$

(עבור α שלילי המשוואה תתהפך $h(\mathbf{x}) = +1 \Leftrightarrow \mathbf{w}^T \mathbf{x} < -b$ ולכן כלל ההחלטה ישתנה ולא יהיה כמו המסווג הנתון)

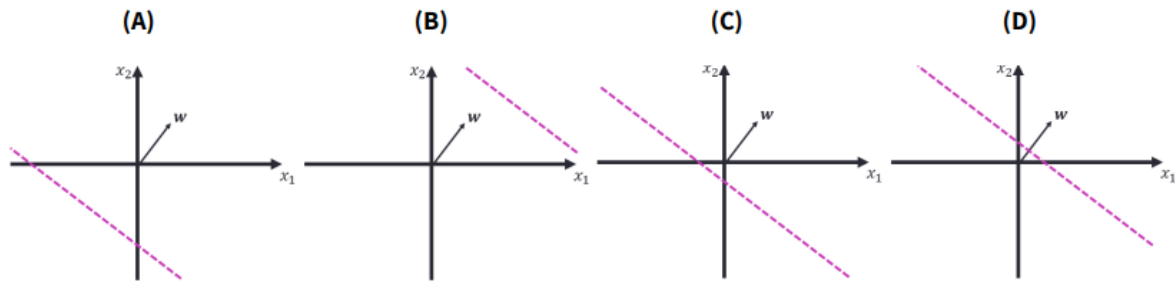
3.2. Following are 4 figures of decision boundaries induced by the same \mathbf{w} with different biases $b \in \mathbb{R}$.

The biases are ordered as: $b_1 > b_2 > 0 > b_3 > b_4$.

Which bias corresponds to which figure?

Write your answer in the following format (answers are made up): A is b_1 , B is b_2 , C is b_3 , D is b_4 .

No explanations are required.



A = b_1

B = b_4

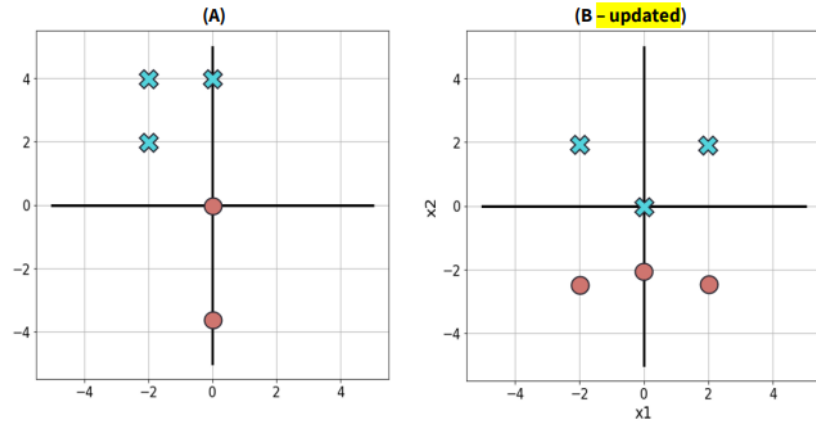
C = b_2

D = b_3

K Nearest Neighbors

4. Following are 2 training sets in the \mathbb{R}^2 feature space with 2 classes (blue/red).

Assume no dataset has two points in the exact same coordinates.



Following are 3 models.

- kNN with $k = 1$ (a training point is not considered a neighbor of itself) w.r.t. $d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|_2$
- kNN with $k = 3$ (a training point is not considered a neighbor of itself) w.r.t. $d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|_2$
- kNN with $k = 1$ (a training point is not considered a neighbor of itself) w.r.t. $d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|_1$

4.1. For each model above, write which datasets this model can perfectly fit (i.e., with 0 training error) and which datasets it cannot. Write your answers in a table, like in the example below. When you say a model cannot perfectly fit a certain dataset, explain why in 1-2 sentences (without drawings).

Model/dataset	(A)	(B)
<i>i.</i>	No כי למשל אם נסתכל על נקודה $(-2, 2)$, הנקודה הקרובה ביותר אליה היא $(0, 0)$ והיא ממחלקה אחרת ממנה, -1 במקום +1.	No כי למשל אם נסתכל על הנקודה $(0, -2)$ הנקודה הכי קרובה אליה היא $(0, 0)$ והיא ממחלקה אחרת ממנה
<i>ii.</i>	No כי למשל אם נסתכל על הנקודה $(0, -3.8)$ הנקודות הכי קרובות אליה היא נקודה מהמחלקה שלה ושתי נקודות מהמחלקה האחרת ולכן היא תסווג כנקודה מהמחלקה האחרת. זה נוצר כי יש שלוש נקודות כחולות ושתי נקודות אדומות ולכן כל נקודה שהיא אדומה תסווג כחולה.	Yes
<i>iii.</i>	yes	No כי למשל הנקודה ב $(0, 0)$ הכי קרובה לנקודה ב $(0, -1)$ שהיא ממחלקה אחרת. וגם הפוך.

4.2. Now assume that the data points from all datasets above are rotated by the same unknown angle θ

(around the origin). That is, each 2-dimensional data point $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ is mapped into $\begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$.

Without knowing the exact angle θ , answer for each of the models above:

- Might your answers for that model change?
- If not, briefly explain why.
- Otherwise, the answers for which datasets might change? Briefly explain why.

עבור i ו-iii :

שני המודלים האלו מבוססים על מרחק אוקלידי, ניתן לראות שהמרחק האוקלידי לא משתנה בין שתי נקודות כאשר מבצעים סיבוב מסביב לנקודה 0,0 לכן השכנים הקרובים שהיו קודם לכן ישארו גם אחרי הסיבוב. הסיבוב בעצם לא משתנה את המרחקים האוקלידיים כי הוא סימטרי.

עבור iii :

מדובר במרחק מנהטן, זהו מרחק שמתייחס למרחק על פי המרחק בצירים. במקרה הזה, הסיבוב כן משנה המרחק בין הנקודות על פי הצירים כי הנקודות משנות את המיקום שלהם. נסתכל על database A, ראינו שבלי הרוטציה הוא מתאים בצורה מושלמת למודל, אם נבצע סיבוב כך שהנקודה שנמצאת ב (-2,2) תהיה על ציר ה-y נקבל שהיא הנקודה הכי קרובה לנקודה ב (0,0) שלא משנה את מיקומה ולכן נקבל תוצאת סיווג לא נכונה, כלומר הנקודה האדומה (0,0) תקבל סיווג שהיא כחולה. עכשיו נסתכל על database B, התשובה כאן יכולה להשתנות בהתאם לזווית הסיבוב, כאשר הנקודות הכחולות יהיו על הצירים באמצעות הסיבוב (כלומר זווית של 45 מעלות), יוצר מצב שיש שתי נקודות שהכי קרובות ל (0,0) ולפי הפיאצה זה מקרה שלא מוגדר ולכן התשובה פה לא ידועה, במקרה והתשובה הייתה נבחרת ככחולה אז המודל היה מתאים לdatabase המוגדר. כל שאר הנקודות לא היו בעיה מקודם וגם לא עכשיו כי מרחק מנהטן לנקודה הכי קרובה עדיין נשאר מאותו מחלקה. ולכן נקבל גם במקרה הזה שהסיווג פה לא מתאים בצורה מושלמת.

