הטכניון – מכון טכנולוגי לישראל
הפקולטה למדעי המחשב

מבוא למערכות לומדות (236756)

סמסטר אביב תשפ"א – 11 באוקטובר 2021

מרצה: ד"ר ניר רוזנפלד

מבחן מסכם מועד ב'

הנחיות הבחינה:

- **משך הבחינה:** 3 שעות.
- **חומר עזר:** המבחן בחומר סגור (ללא ספרים, מחברות, דפי נוסחאות).
- אין צורך במחשבון.
- מותר לכתוב בעט או בעיפרון, כל עוד הכתב קריא וברור.
- יש לכתוב את תשובותיכם **על גבי שאלון זה** בכתב יד קריא. תשובה בכתב יד שאינו קריא לא תיבדק.
- במבחן 13 עמודים ממוספרים סה"כ, כולל עמוד שער זה שמספרו 1 ושלושה עמודי טיוטה בסוף הגיליון.
- אין בחירה בין השאלות.
- נא לכתוב רק את המבוקש ולצרף הסברים קצרים עפ"י ההנחיות.
- **בתום המבחן יש להגיש את שאלון זה בלבד.**

מבנה הבחינה:

- **חלק א' [72 נק']:** 4 שאלות פתוחות [כל אחת 18 נק']
- **חלק ב' [28 נק']:** 7 שאלות סגורות (אמריקאיות) [כל אחת 4 נק']

בהצלחה!

חלק א' – שאלות פתוחות [72 נק']

שאלה 1: ההשפעה של נירמול על מסווגים שונים [18 נק']

נתון סט אימון (train set) בעל מאפיינים (פיצ'רים) רציפים לא מנורמלים ותיוגים בינאריים, משמע $\boldsymbol{x}_i \in \mathbb{R}^d, \boldsymbol{y}_i \in \{-1,1\}$.

- בשלב הראשון לומדים מסווג על סט האימון ובודקים את דיוק האימון על סט האימון. 100%
- כעת, מנרמלים כל מאפיין (feature) בעזרת min-max scaling, כך שכל הערכים באותו מאפיין יהיו **בין 0 ל-1**.
- בשלב השני מאמנים מסווג חדש על סט האימון המנורמל, ובודקים את דיוק האימון על סט האימון המנורמל. 100%

עבור כל אלגוריתם למידה, סמנו האם דיוק **האימון** של המסווג החדש לאחר הנירמול זהה בהכרח לזה של המסווג המקורי.
רק אם סימנתם שהדיוק **לא בהכרח זהה**, הסבירו בקצרה מדוע.

הניחו שאין צעדים סטוכסטיים (אקראיים) בריצת האלגוריתמים.

א. ID3 המשתמש באנטרופיה ובונה עץ בעומק מירבי 4 דיוק האימון: **זהה בהכרח** / **לא בהכרח זהה**

הסבר (אם "לא בהכרח"): ______

ב. AdaBoost with decision stumps דיוק האימון: **זהה בהכרח** / **לא בהכרח זהה**

הסבר (אם "לא בהכרח"): ______

ג. k-NN כאשר $k = 1$ (דוגמת אימון לא נחשבת שכנה של עצמה) דיוק האימון: **זהה בהכרח** / **לא בהכרח זהה**

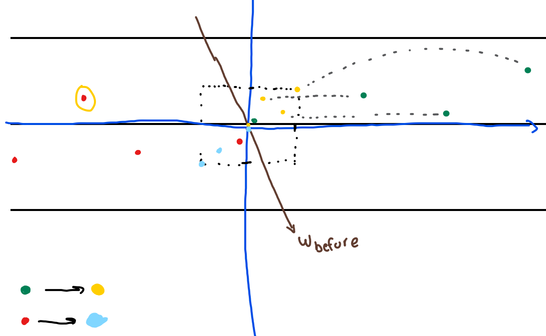
הסבר (אם "לא בהכרח"): נניח שכל הפיצ'רים חוץ מאחד מתפלגים בין 1 ל 0
ופיצ'ר אחד בעל outlier שנמצא ב ∞ ועוד נקודות שנמצאות במרחק
ε, כאשר נרמול minmax ונקבל ש הנקודות של הפיצ'ר הזה מתקרבות ושוליות
לכאורה לא (המשקל שלו

ד. Log. regression הומוגני כשהדאטה המקורי פריד ליניארית הומוגנית

דיוק האימון: **זהה בהכרח / לא בהכרח זהה**

הסבר (אם "לא בהכרח"): הדאטה יכול להפוך להיות לא פריד

ליניארית והומוגנית לאחר הנרמול minmax



ה. Hard-SVM לא הומוגני כשהדאטה המקורי פריד ליניארית דיוק האימון: **זהה בהכרח / לא בהכרח זהה**

הסבר (אם "לא בהכרח"): הדאטה יישאר פריד [illegible] לאחר הנרמול

ו. Soft-SVM לא הומוגני כאשר $\lambda = 1$ דיוק האימון: **זהה בהכרח / לא בהכרח זהה**

הסבר (אם "לא בהכרח"): ועם 'סגולות' להתכנס ל-1 - מתפזרים לנורמה לקטון ובכך נשיג פתרון ''יותר''

ומה

שאלה 2: ההשפעה של מיפויים על פרידות ליניארית [18 נק']

נתון אוסף דוגמאות עם מאפיינים דו-ממדיים רציפים $\boldsymbol{x} \in \mathbb{R}^2$ ותיוגים בינאריים $y \in \{-1, +1\}$.

ידוע כי אוסף הדוגמאות פריד ליניארית ע"י וקטור נתון $\boldsymbol{w} \in \mathbb{R}^2$ ורכיב bias נתון $b \in \mathbb{R}$.

כעת נסתכל על פונקציות מיפוי $\phi: \mathbb{R}^2 \to \mathbb{R}^k$ ונבחן את השפעתן על אוסף הדוגמאות. k יכול להיות גדול/קטן/שווה ל-2.

עבור כל אחת מפונקציות המיפוי הבאות, סמנו האם אוסף הדוגמאות אחרי המיפוי עדיין פריד ליניארית בהכרח.

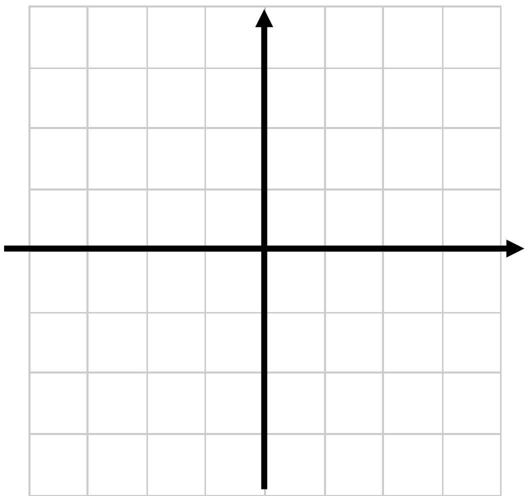
אם כן: השתמשו ב-$\boldsymbol{w}, b$ כדי להציע וקטור חדש $\boldsymbol{w}' \in \mathbb{R}^k$ ורכיב bias חדש $b' \in \mathbb{R}$ המפרידים את האוסף אחרי המיפוי.

אם לא: ציירו אוסף דוגמאות מתוייגות במרחב המקורי $\mathbb{R}^2$ שהמיפוי המוצע הופך ללא פריד ליניארית.

א. [6 נק'] ϕ היא הורדת ממד ליניארית בעזרת PCA לממד יחיד.

פורמלית, תהא $\mathbf{U} \in \mathbb{R}^{2\times 1}$ מטריצת ההטלה של PCA, אזי המיפוי הינו $\phi(\boldsymbol{x}) = \mathbf{U}^\top \boldsymbol{x}$ ומתקיים $k = 1$.

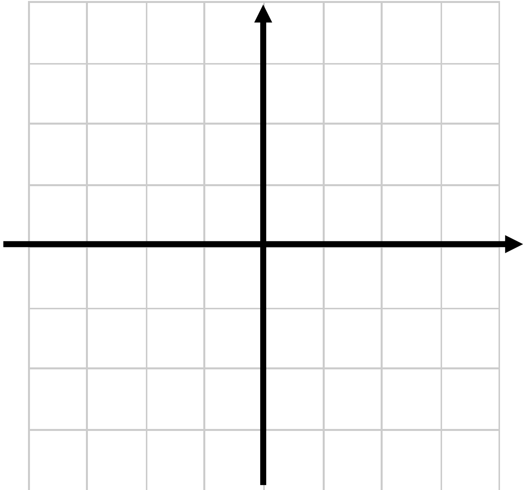
אוסף הדוגמאות אחרי המיפוי: **בהכרח / לא בהכרח** פריד ליניארית (סמנו).

אם סימנתם "בהכרח":	**אם סימנתם "לא בהכרח" (דוגמה במרחב המקורי):**
$\mathbb{R} \ni w' =$ $\mathbb{R} \ni b' =$	

ב. [6 נק'] ϕ מורידה ליניארית לממד יחיד בעזרת PCA ואז מחזירה את הקלט לדו-ממד בעזרת המטריצה "ההופכית".

פורמלית, תהא $\mathbf{U} \in \mathbb{R}^{2\times 1}$ מטריצת ההטלה של PCA, אזי המיפוי הינו $\phi(\boldsymbol{x}) = \mathbf{U}\mathbf{U}^{\top}\boldsymbol{x}$ ומתקיים $k = 2$.

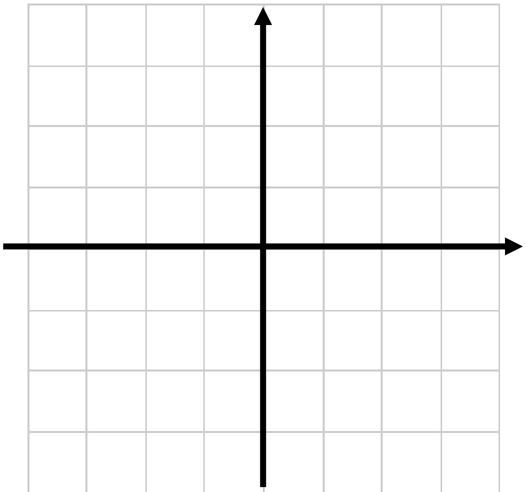
אוסף הדוגמאות אחרי המיפוי: **בהכרח / לא בהכרח** פריד ליניארית (סמנו).

אם סימנתם "בהכרח":	**אם סימנתם "לא בהכרח" (דוגמה במרחב <u>המקורי</u>):**
$\mathbb{R}^2 \ni \boldsymbol{w}' =$ $\mathbb{R} \ni b' =$	

ג. [6 נק'] ϕ היא מיפוי פולינומיאלי ממעלה 3 (כל המונומים <u>עד</u> מעלה 3)

פורמלית, המיפוי הינו $\phi(\boldsymbol{x}) = [1,\ x_1,\ x_2,\ x_1x_2,\ x_1^2,\ x_2^2,\ x_1^2x_2,\ x_1x_2^2,\ x_1^3,\ x_2^3]$ ומתקיים $k = 10$.

אוסף הדוגמאות אחרי המיפוי: **בהכרח / לא בהכרח** פריד ליניארית (סמנו).

אם סימנתם "בהכרח":	**אם סימנתם "לא בהכרח" (דוגמה במרחב <u>המקורי</u>):**
$\mathbb{R}^{10} \ni \boldsymbol{w}' = [0, w_1, w_2, 0, 0, \ldots, 0]$ $\mathbb{R} \ni b' = b$	

שאלה 3: אופטימיזציה [18 נק']

רגרסיה ליניארית המשלבת רגולריזציה מסוג L1 ורגולריזציה מסוג L2 נקראת ElasticNet.

בעיית האופטימיזציה היא (עבור $\lambda_1, \lambda_2 > 0$):

$$\mathrm{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \underbrace{\left(\|\mathbf{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2 + \lambda_1\|\boldsymbol{w}\|_1 + \lambda_2\|\boldsymbol{w}\|_2^2\right)}_{\triangleq p(\boldsymbol{w})}$$

בשאלה זו נבחן את הקמירות של בעיה זו.

> **תזכורת:** תהא C קבוצה קמורה.
> הפונקציה $f: C \to \mathbb{R}$ נקראת פונקציה קמורה אם מתקיים
> $$\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in C,\ \ \forall t \in [0,1]:\quad tf(\boldsymbol{x}_1) + (1-t)f(\boldsymbol{x}_2) \geq f(t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_2)$$

א. [6 נק'] הוכיחו לפי הגדרה שפונקציית הערך המוחלט $|a|$ (עבור סקלאר $a \in \mathbb{R}$) היא פונקציה קמורה.

הוכחה:

~~$t f(x$~~

יהיו $a,b \in C$; $t \in [0,1]$ כלשהו ; $f(x)=|x|$:

$$tf(a)+(1-t)f(b) = t|a|+(1-t)|b| \underset{\substack{t\in[0,1]\\ 1-t\in[0,1]}}{=} |ta|+|(1-t)b| \underset{\text{א"ש המשולש}}{\geq} |ta+(1-t)b| = f(ta+(1-t)b)$$

כלומר הפו' $|a|$ קמורה לפי הגדרה

ב. [6 נק'] הוכיחו שפונקציית המטרה $p(\boldsymbol{w})$ קמורה ב-$\boldsymbol{w}$.

באפשרותכם להשתמש בתכונות שנלמדו בהרצאה או בתרגול (אך עליכם לכתוב אותן במפורש).

כמו כן, תוכלו להשתמש בכך שהראיתם בתרגיל בית שהפונקציה $\|\mathbf{A}\boldsymbol{w}+\boldsymbol{b}\|_2^2$ קמורה ב-$\boldsymbol{w}$ לכל $\mathbf{A}, \boldsymbol{b}$.

הוכחה:

$$\underbrace{\|Xw+b\|_2^2}_{\text{קמורה לפי התרגיל}} + \underbrace{\lambda\|w\|_1}_{\text{מכפלה של סקלר חיובי בפו' קמורה (סעיף א')}} + \lambda_2\underbrace{\|w\|_2^2}_{\nabla_w^2\|w\|_2^2 = I} \underset{\text{קמורה כסכום של פו' קמורות}}{=} P(w)$$

PSD ולכן קמורה

כלומר פו' קמורה X סקלר חיובי

ג. [6 נק'] כתבו וקטור שמהווה subgradient לפי $\boldsymbol{w}$ לפונקציית המטרה $p(\boldsymbol{w})$.

תשובה סופית (לרשותכם עמודי טיוטה בסוף השאלון):

$\nabla_{\boldsymbol{w}}(\|\mathbf{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2+\lambda_1\|\boldsymbol{w}\|_1+\lambda_2\|\boldsymbol{w}\|_2^2)=$

$2X^T(Xw-y)+2\lambda_2 w+\lambda_1 h(w)$

$h(w)=\begin{cases}-1 & w<0\\ 0 & w=0\\ 1 & w>0\end{cases}$

שאלה 4: פונקציות Kernel [18 נק']

תזכורת: פונקציה $K: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ היא קרנל (kernel) חוקי <u>אמ"מ</u>
ניתן למצוא פונקציית מיפוי $\phi: \mathbb{R}^d \to \mathbb{R}^p$ (עבור $p \in \mathbb{N}$) עבורה מתקיים התנאי $\forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d: K(\boldsymbol{u}, \boldsymbol{v}) = \phi(\boldsymbol{u})^\top \phi(\boldsymbol{v})$.

א. [6 נק'] בסעיף זה הניחו מרחב דוגמאות דו-ממדי, משמע $d = 2$.
נוכיח שהפונקציה $K_0(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u}^\top \boldsymbol{v})^2$ הינה קרנל חוקי (זהו קרנל פולינומיאלי ממעלה 2).
כִּתבו במפורש פונקציית מיפוי $\phi: \mathbb{R}^2 \to \mathbb{R}^3$ שמקיימת את התנאי הנדרש עם K_0.

$(u_1, u_2)\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$

$(u_1v_1 + u_2v_2)^2$

$u_1^2v_1^2 + 2u_1v_1u_2v_2 + u_2^2v_2^2$

$$\phi(\boldsymbol{u}) = \begin{bmatrix} u_1^2 & , & \sqrt{2}\,u_1u_2 & , & u_2^2 \end{bmatrix}^\top$$

בשני הסעיפים הבאים, תהיינה ψ, ψ' שתי פונקציות מיפוי ממרחב הדוגמאות למרחב דו-ממדי, משמע $\psi, \psi': \mathbb{R}^d \to \mathbb{R}^2$.
יהיו K_1, K_2 הקרנלים המוגדרים ע"י ψ, ψ', משמע מתקיים $K_1(\boldsymbol{u}, \boldsymbol{v}) = \psi(\boldsymbol{u})^\top \psi(\boldsymbol{v})$ וגם $K_2(\boldsymbol{u}, \boldsymbol{v}) = \psi'(\boldsymbol{u})^\top \psi'(\boldsymbol{v})$.

$\psi(u)^T\psi(v) + \psi'(u)^T\psi'(v)$

ב. [6 נק'] נוכיח שהפונקציה $K_3(\boldsymbol{u}, \boldsymbol{v}) = \underbrace{K_1(\boldsymbol{u}, \boldsymbol{v})}_{\in\mathbb{R}} + \underbrace{K_2(\boldsymbol{u}, \boldsymbol{v})}_{\in\mathbb{R}}$ הינה קרנל חוקי.

$(u_1'\ u_2')\begin{pmatrix} v_1' \\ v_2' \end{pmatrix} = u_1'v_1' + u_2'v_2' + u_1''v_2'' + u_2''v_2''$

כִּתבו במפורש פונקציית מיפוי $\phi: \mathbb{R}^d \to \mathbb{R}^p$ שמקיימת את התנאי הנדרש עם K_3 (עליכם להסיק את p בעצמכם).

$$\phi(\boldsymbol{u}) = \begin{bmatrix} \psi_1(u), & \psi_2(u), & \psi'_1(u), & \psi'_2(u) \end{bmatrix}^\top$$

$\phi(u)^T\phi(v) = \psi(u)^T\psi(v)$

$\psi_i(u)$ — הרכיב i של $\psi(u)$

ג. [6 נק'] נוכיח שהפונקציה $K_4(\boldsymbol{u}, \boldsymbol{v}) = \underbrace{K_1(\boldsymbol{u}, \boldsymbol{v})}_{\in\mathbb{R}} \cdot \underbrace{K_2(\boldsymbol{u}, \boldsymbol{v})}_{\in\mathbb{R}}$ הינה קרנל חוקי.

$(u_1'\ u_2')\begin{pmatrix} v_1' \\ v_2' \end{pmatrix} \cdot (u_1''\ u_2'')\begin{pmatrix} v_1'' \\ v_2'' \end{pmatrix}$

כִּתבו במפורש פונקציית מיפוי $\phi: \mathbb{R}^d \to \mathbb{R}^4$ שמקיימת את התנאי הנדרש עם K_4.

$$\phi(\boldsymbol{u}) = \begin{bmatrix} \psi_1(u)\psi'_1(u), & \psi_1(u)\psi'_2(u), & \psi_2(u)\psi'_1(u), & \psi_2(u)\psi'_2(u) \end{bmatrix}^\top$$

$(u_1'v_1' + u_2'v_2')(u_1''v_1'' + u_2''v_2'') = u_1'u_1''v_1'v_1'' + u_1'u_2''v_1'v_2'' + u_2'u_1''v_2'v_1'' + u_2'u_2''v_2'v_2''$

חלק ב' – שאלות אמריקאיות [28 נק']

סמנו את התשובות המתאימות (לפי ההוראות). בחלק זה אין צורך לכתוב הסברים.

א. [4 נק'] מה מבין ההיפר-פרמטרים הבאים עשוי להשפיע על מספר המאפיינים (פיצ'רים) שהמודל הסופי ישתמש בהם?
סמנו את **כל** התשובות הנכונות.

a. k באלגוריתם k-NN
b. מקדם הרגולריזציה λ ברגרסיה ליניארית עם רגולריזציית L1
c. העומק המרבי *max_depth* באלגוריתם ID3
d. מספר האיטרציות המקסימלי T באלגוריתם AdaBoost עם Decision stump כמסווג חלש

ב. [4 נק'] מבין הטענות הבאות הקשורות ל-Kernel-SVM, סמנו את הטענה **השגויה**.

a. בחירת סוג הקרנל עשויה להשפיע על מידת ה-overfitting של המודל
b. מספר המשתנים בבעיית האופטימיזציה של Kernel-SVM הוא כמספר המאפיינים d (במקרה ההומוגני)
c. קיימים אוספי נתונים שאינם פרידים לינארית ש-Kernel-SVM יכול להפריד בצורה מושלמת
d. טריק ה-Kernel מתאפשר כי ב-learning objective של Kernel-SVM הדוגמאות מופיעות רק במכפלות פנימיות

ג. [4 נק'] נתון דאטה עם מאפיינים רציפים ותיוגים רציפים ($\boldsymbol{x}_i \in \mathbb{R}^d,\ y_i \in \mathbb{R}$).
פותרים בעיית רגרסיה עם דרגות שונות למיפויים פולינומיאליים (עד מעלה p) ומקדמי רגולריזציה λ שונים.
השורה הראשונה בטבלה שלפניכם מתארת את ביצועי האימון והמבחן של רגרסיה ליניארית ($p = 1$) ללא רגולריזציה.

השורות האחרות מתארות את הביצועים של ריצות שונות על אותו דאטה, אך עם ערכי λ, p שונים.
חלק מהשורות מתארות תוצאות אפשריות וחלק מתארות תוצאות בלתי אפשריות.
לכל אחת מארבע השורות, סמנו האם היא אפשרית או לא.

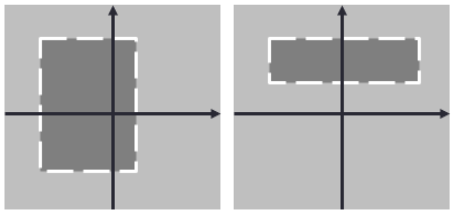
L2 Regular. strength	Polynomial Degree	Train MSE	Test MSE	האם אפשרית?
$\lambda = 0$	$p = 1$ (linear)	20	30	אפשרית (נתון)
$\lambda = 1$	$p = 1$	22	4	**כן / לא**
$\lambda = 1$	$p = 1$	4	20	**כן / לא**
$\lambda = 0$	$p = 2$	22	24	**כן / לא**
$\lambda = 0$	$p = 9$	4	32	**כן / לא**

ד. [4 נק'] בשיטת multinomial logistic regression, מטרת ה-Cross-entropy loss הינה:
(סמנו את התשובה הנכונה)

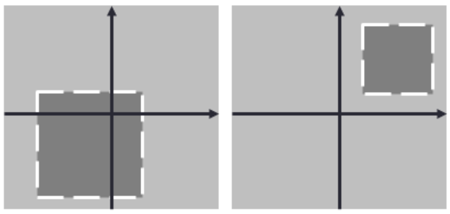
a. לאפשר מִקבול (parallelization) של הלמידה
b. לאפשר רגרסיה פולינומיאלית ממעלה 2
c. לקרב את הניבוי ההסתברותי של התיוג הנכון ל-1 ושל התיוגים האחרים ל-0
d. לוודא שהאלגוריתם יעצור אחרי מספר epochs סופי
e. לנרמל את הפלט של פונקציות ה-score של כל מחלקה באופן שיצרו התפלגות

ה. [4 נק'] בתרגול הגדרנו את מחלקת ההיפותזות $\mathcal{H}_{rect}$ של **מלבנים** מקבילים לצירים בדו-מימד והראינו שמתקיים $\mathrm{VCdim}(\mathcal{H}_{rect}) = 4$ (השטח שבתוך המלבן מסווג באופן חיובי, והשטח שמחוץ למלבן שלילי).
כעת נגדיר את מחלקת $\mathcal{H}_{sqr}$ של **ריבועים** (רוחב ואורך שווים) מקבילים לצירים בדו-מימד.

שתי היפותזות מתוך $\mathcal{H}_{rect}$



שתי היפותזות מתוך $\mathcal{H}_{sqr}$



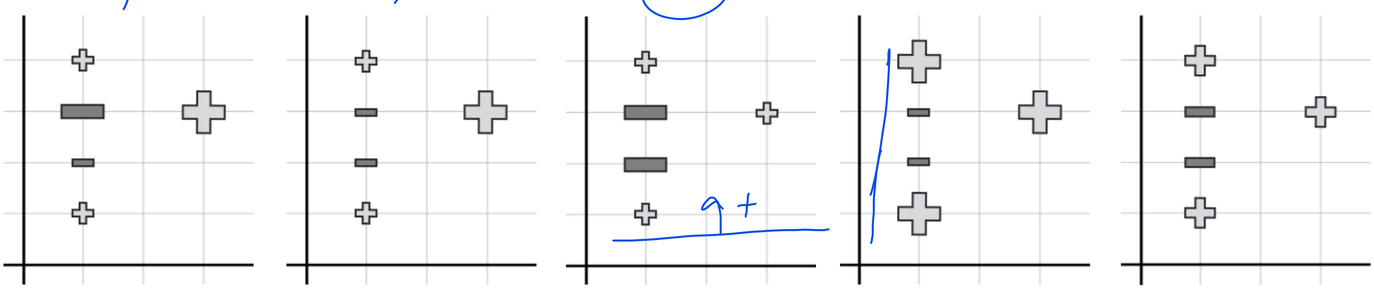
כתבו את ממד ה-VC של המחלקה החדשה (בין 1 ל-5).

$\mathrm{VCdim}(\mathcal{H}_{sqr}) = \boxed{3}$

ו. [4 נק'] על איזו הנחה מוותרים במעבר מההגדרה של PAC learnability לזו של agnostic PAC learnability?
סמנו את התשובה הנכונה.
a. הנחת ה-realizability
b. הנחת דיוק המבחן
c. ההנחה שהנתונים מפולגים זהה (identically distributed)
d. ההנחה שהנתונים בלתי תלויים (independent)
e. ההנחה שהדאטה פריד ליניארית (linear separability)

ז. [4 נק'] נתון דאטה עם תיוגים בינאריים ("+" או "-"). מריצים AdaBoost עם Decision stump כמסווג חלש.
גדלי הצורות בתרשימים מסמלים את ההסתברויות שהאלגוריתם מקצה לדוגמאות (הסתברות גבוהה = צורה גדולה).
רק אחד מהתרשימים הבאים מתאר התפלגות שניתן לקבל אחרי איטרציה אחת של AdaBoost.
הקיפו את האות שמתאימה לתרשיש זה.

הדאטה המקורי (a) (b) (c) (d)



מסגרת נוספת לשימושכם (יש לציין אם מדובר בטיוטה או בהמשך לתשובה אחרת):

מסגרת נוספת לשימושכם (יש לציין אם מדובר בטיוטה או בהמשך לתשובה אחרת):

מסגרת נוספת לשימושכם (יש לציין אם מדובר בטיוטה או בהמשך לתשובה אחרת):