הטכניון – מכון טכנולוגי לישראל

הפקולטה למדעי המחשב

מבוא למערכות לומדות (236756)

סמסטר חורף תשפ"ג – 15 בפברואר 2023

מרצה: ד"ר יונתן בלינקוב

מבחן מסכם מועד א'

הנחיות הבחינה:

- **משך הבחינה:** שלוש שעות.
- **חומר עזר:** המבחן בחומר סגור (ללא ספרים, מחברות, דפי נוסחאות).
- מחשבון: מותר.
- כלי כתיבה: עט בלבד.
- יש לכתוב את התשובות **על גבי שאלון זה**.
- מותר לענות בעברית או באנגלית.
- הוכחות והפרכות צריכות להיות פורמליות.
- קריאות:
  - תשובה בכתב יד לא קריא – **לא תיבדק**.
  - בשאלות רב-ברירה – הקיפו את התשובות בבירור. סימונים לא ברורים יביאו לפסילת התשובה.
  - לא יתקבלו ערעורים בנושא.
- במבחן 17 עמודים ממוספרים סה"כ, כולל עמוד שער זה שמספרו 1 ושלושה עמודי טיוטה בסוף הגיליון.
- נא לכתוב רק את המבוקש ולצרף הסברים קצרים עפ"י ההנחיות.
- **בתום המבחן יש להגיש את שאלון זה בלבד.**

בהצלחה!

חלק א' – שאלות פתוחות [94 נק']

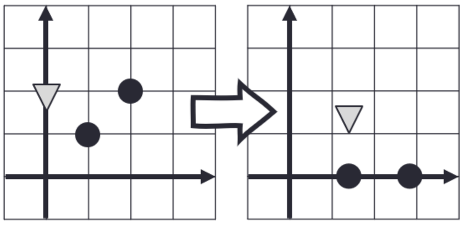
שאלה 1: רגישות של מסווגים לסיבובים [24 נק']

נתון סֶט אימון דו-ממדי עם תיוגים בינאריים, משמע לכל $i = 1, \dots, m$ מתקיים $\mathbf{x}_i \in \mathbb{R}^2, y_i \in \{-1,1\}$.

לומדים שני מסווגים:

- בשלב הראשון: לומדים מסווג על סט האימון המקורי ומחשבים עליו את דיוק האימון.
- בשלב השני:
  - מסובבים את כל הדאטה ב-$45°$ סביב ראשית הצירים (ראו דוגמה).
  - מאמנים מסווג חדש על סט האימון המעודכן,
    ומחשבים עליו את דיוק האימון המעודכן.

סיבוב לדוגמה



תזכורת: הסיבוב של הדאטה יכול להתבצע ע"י מיפוי $\mathbf{x}_i \mapsto \mathbf{Q}\mathbf{x}_i$, כאשר $\mathbf{Q} \in \mathbb{R}^{2\times 2}$ הינה מטריצת סיבוב ממשית, שהיא מטריצה אורתונורמלית המקיימת $\mathbf{Q}^{-1} = \mathbf{Q}^{\top}$.

עבור כל אלגוריתם למידה, סמנו האם דיוק האימון של המסווג החדש על סט האימון המעודכן זהה בהכרח לזה של המסווג המקורי על סט האימון המקורי.

הסבירו בקצרה את תשובותיכם (2-4 משפטים בכל סעיף).

הניחו שאין צעדים אקראיים או שגיאות נומריות בריצת האלגוריתמים (בעיות קמורות מתכנסות לפתרון האנליטי במדויק).

א. kNN עם $k = 3$ (דוגמה לא נחשבת שכנה של עצמה). דיוק האימון זהה בהכרח? **כן** / **לא**

הסבר: כן כי המרחק הוא נורמה 2 אשר אין משמעות לסיבוב, המרחקים נשארים זהים

ב. Hard-SVM ליניארי לא הומוגני בהנחה שהדאטה המקורי פריד. דיוק האימון זהה בהכרח? **כן** / **לא**

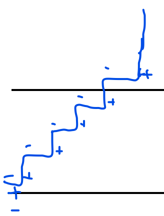
הסבר: הדאטה כאשר פריד עם לאחר סיבוב כפיתע

ולכן הדיוק יהיה 100% ב-2 המקרים

ג. AdaBoost with decision stumps, maximum of $T = 10$ iterations. דיוק האימון זהה בהכרח? **כן** / **לא**

הסבר: אם הדאטה פרדלינארית ע"י הישר $y=x$ למשל, כבש אופציונלית עם 10

איטרציות יכולים לא 100% דיוק אימון



לאחר הסיבוב הדאטה יהיה פריד ע"י הישר $y=0$

שנוכל לנסות אחרי איטרציה אחת $\Rightarrow$ 100% דיוק

ד. Homogeneous logistic regression with L2 regularization (ללא הנחה שהדאטה המקורי פריד):

$$\text{argmin}_{\mathbf{w}\in\mathbb{R}^2} \sum_i \ln(1+\exp(-y_i\mathbf{w}^\top\mathbf{x}_i)) + \lambda\|\mathbf{w}\|_2^2\,,\quad \text{for some } \infty > \lambda > 0$$

$w_i = (Qw)^T$ $\quad Qx_i$ $\quad \|Qw\|^2 = \|w\|^2$

דיוק האימון זהה בהכרח? **כן** / **לא**

הסבר: עניין, כל הוא ש לזוה Qw ונקבל את אותה

הנורמה.

שאלה 2: VC-dimension [20 נק']

א. [6 נק'] הוכיחו שלכל מחלקת היפותזות **סופית** $\mathcal{H}$ מתקיים בהכרח $\mathrm{VCdim}(\mathcal{H}) \le \log_2(|\mathcal{H}|)$.

רמז: ניתן להוכיח זאת בשלילה ובקצרה.

$\log_2(4)=2$

$\log_2(2)=1$

$\log_2(1)=0$

הוכחה (לרשותכם דפי טיוטה בסוף הגיליון):

נניח כי $VCdim(H) > \log_2(H)$

$h_1(x) = sign(x)$ $H=\{h_1\}$ [illegible] $VCdim(H)=VCdim\{h_1\}=0$

כי הרי כל נק' (יחידה) שנבחר, אם הלייבל + סימן נכון, אבל הלייבל נגדי שלילי

נכון, ולהפך. לכן הנחת $VCdim(H) > \log_2(1)=0$ וזו סתירה,

קיימת קב' $C > \log_2(H)$ של נק' שH מנתצת אותה

מספר התיוגים האפשריים הינו $Num_{labels} > 2^{|C|} = 2^{\log_2(H)} = |H|$

לכל תיוג תהיה תלויה לפחות היפותזה אחת $|H| \ge Num_{labels} > |H|$

סתירה ולכן $VCdim(H) \le \log_2(H)$

ב. [4 נק'] נגדיר מרחב דוגמאות סופי $\mathcal{X} = \{x_1, x_2\}$ (במרחב יש שתי דוגמאות במרחב כלשהו, לא נבטא במפורש את x_1, x_2)

ומחלקת היפותזות סופית $\mathcal{H} = \{h_1, h_2\}$, כאשר h_1, h_2 מוגדרות באופן הבא:

$\exists C \quad \forall y \quad \exists h \quad \forall i \;\; h(x_i) = y_i$

	h_1	h_2
x_1	$+1$	-1
x_2	-1	-1

$+\ -\qquad +\ -\qquad -\ +\qquad -\ -$

(המשמעות של הטבלה היא, למשל, שמתקיים $h_1(x_1) = +1$.)

$VCdim(H) \le \log_2(H) = 1$

$+ \quad + \quad - \quad h_1(x_1) = +1 \quad h$

ענו: מתקיים $\text{VCdim}(\mathcal{H}) = \boxed{1}$. הוכיחו את תשובתכם.

הוכחה:

$VCdim(H) \ge 1$: $C = \{x_1\} \Leftarrow y_1 = 1$: $h_1(x_1) = 1 = y_1$

$y_1 = -1$: $h_2(x_1) = -1 = y_1$

כמו כן:

$VCdim(H) < 2$: $C = \{x_1, x_2\} \Leftarrow y_1 = -1, y_2 = +1$

$h_1(x_1) = 1 \ne y_1$; $h_2(x_2) = -1 \ne y_2$

כלומר לכל C, קיים תיוג כך שכל היפותזה ב H תטעה לפחות על פוני' אחת.

$VCdim(H) \le 1$: לפי א' $VCdim(H) \le \log_2(|H|) = \log_2(2) = 1$

<u>רמז</u>: גם בסעיפים הבאים ניתן להגדיר מרחב דוגמאות סופי $\mathcal{X} = \{x_1, \dots, x_n\}$ ואוסף סופי $\mathcal{H}$ של היפותזות $h: \mathcal{X} \to \{-1, +1\}$.

על ההיפותזות להיות **<u>שונות</u>** (distinct) במובן שאין שתי היפותזות ב-$\mathcal{H}$ שמחזירות פלט זהה על <u>כל</u> הדוגמאות ב-$\mathcal{X}$.

בשני הסעיפים אין צורך להוכיח את ה-VC-dimension של המחלקות שתציעו.

ג. [5 נק'] הראו מחלקת היפותזות $\mathcal{H}$ עבורה $|\mathcal{H}| = 4$ וגם $\text{VCdim}(\mathcal{H}) = 1$.

תשובה (לרשותכם דפי טיוטה בסוף הגיליון):

$\text{Sign}(x)$ $\quad H = \{\text{Sign}(x), -\text{Sign}(x), \text{Sign}(2x), -\text{Sign}(2x)\}$

$-\text{Sign}(x)$

$\exists C \; \forall y \; \exists h \; \forall i$

$h_1 \quad h_2 \quad h_3 \quad h_4$

	x_1	x_2	x_3	
~~h_1~~	~~0~~	~~1~~	~~0~~	
h_2	0	1	1	✓
h_3	1	0	1	✓
h_4	1	1	0	✓
h_1	1	1	1	✓

$+$ $-$ x_1 x_2 x_3

$x_1 \; x_2 \; x_3$ [illegible]

ד. [5 נק'] הראו מחלקות היפותזות $\mathcal{H}_1, \mathcal{H}_2$ עבורן $\text{VCdim}(\mathcal{H}_1) = \text{VCdim}(\mathcal{H}_2) = 1$ וגם $\text{VCdim}(\mathcal{H}_1 \cup \mathcal{H}_2) = 3$.

תשובה:

$H_1 =$

	x_1	x_2	x_3
h_1	0	1	1
h_2	1	0	1
h_3	1	1	0
h_4	1	1	~~0~~ 1

$H_2 =$

	x_1	x_2	x_3
h_1	1	0	0
h_2	0	1	0
h_3	0	0	1
h_4	0	0	0

(V.C) [illegible] $\quad (1,1)$ [illegible]

$H_1 \cup H_2 =$

	x_1	x_2	x_3
h_1''	0	0	0
h_2''	0	0	1
h_3''	0	1	0
h_4''	0	1	1
h_5''	1	0	0
h_6''	1	0	1
h_7''	1	1	0
h_8''	1	1	1

[illegible] 3 נק' x_1, x_2, x_3 [illegible]

$\text{VCdim}(H_1 \cup H_2) \geq 3$

$\text{VCdim}(H) \leq \log_2(8) = 3$

$\text{VCdim} = 3$

שאלה 3: רגרסיה ליניארית [25 נק']

בשאלה זו מרחב הדוגמאות הוא $\mathcal{X} = \mathbb{R}^d$ ומרחב התיוגים הוא $\mathcal{Y} = \mathbb{R}$.

נתון מדגם $S = \{(\mathbf{x}_{\mathrm{i}}, y_i)\}_{i=1}^m$.

רגרסיה ליניארית המשלבת רגולריזציה מסוג L1 ומסוג L2 נקראת ElasticNet:

$$\mathbf{w}^\star \triangleq \operatorname{argmin}_{\mathbf{w}\in\mathbb{R}^d}\left(\underbrace{\sum_{i=1}^m(\mathbf{w}^\top\mathbf{x}_i - y_i)^2}_{\mathcal{L}(\mathbf{w};S)} + \underbrace{\lambda\alpha\|\mathbf{w}\|_2^2 + \lambda(1-\alpha)\|\mathbf{w}\|_1}_{R_{\lambda,\alpha}(\mathbf{w})}\right),$$

עבור $\lambda \geq 0, \alpha \in [0,1]$ כלשהם.

א. [5 נק'] בטור הימני יש שלוש השָׂמוֹת של λ, α, שהופכות את הבעיה לעיל לבעיות רגרסיה מוּכָּרות. בטור השמאלי, יש חמש שיטות גנרטיביות. במהלך הקורס ראינו שקילות בין שתי בעיות מימין לשתי שיטות משמאל (תחת הנחות מסוימות על הדאטה). את השקילות השלישית עליכם להסיק לבד.

מלאו את המשבצות הריקות של שלוש הבעיות בטור הימני במספרים של שלוש השיטות השקולות להן בטור השמאלי.

a. הבעיה לעיל כאשר $\lambda = 0$	☐	i.	שערוך MLE
		ii.	שערוך Naïve Bayes
b. הבעיה לעיל כאשר $\lambda > 0, \alpha = 0$	☐	iii.	שערוך MAP עם Gaussian prior
		iv.	שערוך MAP עם Binomial prior
c. הבעיה לעיל כאשר $\lambda > 0, \alpha = 1$	☐	v.	שערוך MAP עם Laplace prior

מעתה נניח $\boldsymbol{\lambda > 0, \alpha \in (0,1)}$ בלבד.

ב. [12 נק'] בסעיף זה נניח שיש $d = 2$ פיצ'רים והם **זהים** זה לזה, כלומר $\forall i \in [m]\colon\ \mathrm{x}_i[1] = \mathrm{x}_i[2]$ (וכאמור, $\lambda > 0, \alpha \in (0,1)$).

(a) לכל $\mathbf{w} = \begin{bmatrix} \mathrm{w}[1] \\ \mathrm{w}[2] \end{bmatrix} \in \mathbb{R}^2$ נגדיר את $\widetilde{\mathbf{w}} = \begin{bmatrix} \mathrm{w}[2] \\ \mathrm{w}[1] \end{bmatrix}$.

הוכיחו: תחת ההנחות, לכל $\mathbf{w} \in \mathbb{R}^2$ מתקיים $\mathcal{L}(\mathbf{w}; S) + R_{\lambda,\alpha}(\mathbf{w}) = \mathcal{L}(\widetilde{\mathbf{w}}; S) + R_{\lambda,\alpha}(\widetilde{\mathbf{w}})$.

הוכחה:

להלן טענת עזר שניתן להשתמש בה במידת הצורך בהמשך השאלה.

טענת עזר: פונקציית הרגולריזציה $R_{\lambda,\alpha}$ היא **קמורה במובן החזק** (strictly convex). משמע, מתקיים באי-שיוויון חזק:

$$\forall \mathbf{w}_1 \neq \mathbf{w}_2 \in \mathbb{R}^d, \forall \beta \in (0,1):\ \ \beta R_{\lambda,\alpha}(\mathbf{w}_1) + (1-\beta) R_{\lambda,\alpha}(\mathbf{w}_2) > R_{\lambda,\alpha}(\beta \mathbf{w}_1 + (1-\beta)\mathbf{w}_2).$$

(b) הוכיחו: תחת ההנחות (2 פיצ'רים זהים, $\lambda > 0, \alpha \in (0,1)$), שני מקדמי הפתרון האופטימלי זהים.

משמע, מתקיים: $\mathrm{w}^\star[1] = \mathrm{w}^\star[2]$.

הוכחה:

להלן משפט. **הבינו אותו היטב.** בפרט, השתכנעו שהוא מכליל את שהוכחנו בסעיף הקודם.

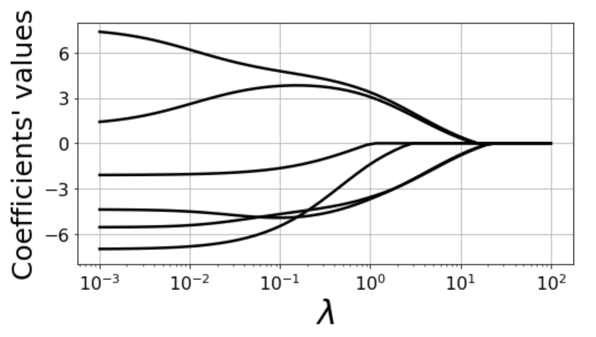
<u>משפט</u>[1]: תחת הנחות (לא מגבילות במיוחד) על הנתונים, כל שני מקדמים של פתרון ה-Elastic Net מקיימים:

$$\forall a,b\in[d]:\ \left(\left|\mathbf{w}^{\star}[a]-\mathbf{w}^{\star}[b]\right|\le\frac{1}{\lambda\alpha}\sqrt{\left(1-\rho_{a,b}\right)\cdot c}\right),$$

כאשר $\rho_{a,b}\triangleq\frac{\mathrm{Cov}(\mathbf{x}[a],\mathbf{x}[b])}{\sigma_a\sigma_b}\in[-1,1]$ הוא מקדם המְתְאָם (קורלציה) של פִּירסון בין הפיצ'רים a,b (על פני כל הדוג')
ו-$c>0$ הוא קבוע כלשהו.

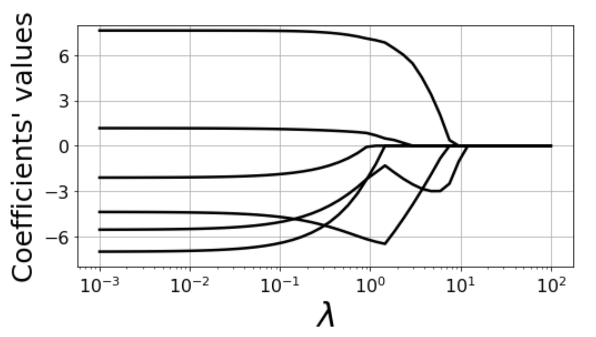
ג. [8 נק'] עבור dataset עם 6 פיצ'רים, אימנו ElasticNet עם ערכי $\alpha\in\{0.01,0.5,0.99\}$ וערכי $\lambda>0$ שונים.
כל עקומה בגרפים מראה את ערכיו של מקדם <u>יחיד</u> (מתוך 6) בפתרון שנוצר מצירוף של α (ברמת הגרף) ו-λ (בציר x).
<u>נתון</u>: ישנם שני זוגות של פיצ'רים עם קורלציה גבוהה (בכל זוג, קורלציה גבוהה בין שני הפיצ׳רים של אותו זוג).

ליד כׇּל גרף סמנו את ערך α המתאים לו ביותר והסבירו <u>בקצרה</u> את בחירותיכם. היעזרו במשפט לעיל בהסברים.



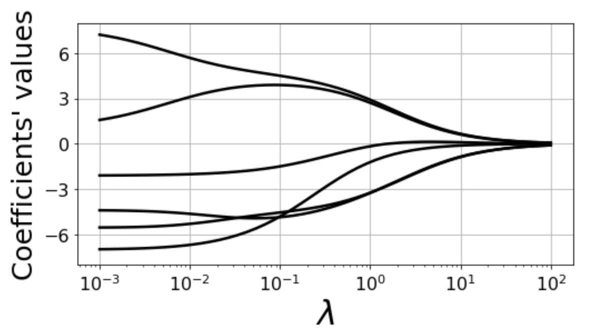
לאיזה ערך α מתאים הגרף? 0.01 / 0.5 / 0.99

הסבר:



לאיזה ערך α מתאים הגרף? 0.01 / 0.5 / 0.99

הסבר:



לאיזה ערך α מתאים הגרף? 0.01 / 0.5 / 0.99

הסבר:

[1] גרסה פשטנית של משפט מהמאמר [Zou and Hastie. 2005] “Regularization and variable selection via the elastic net”

שאלה 4: מסווגים ליניאריים [25 נק']

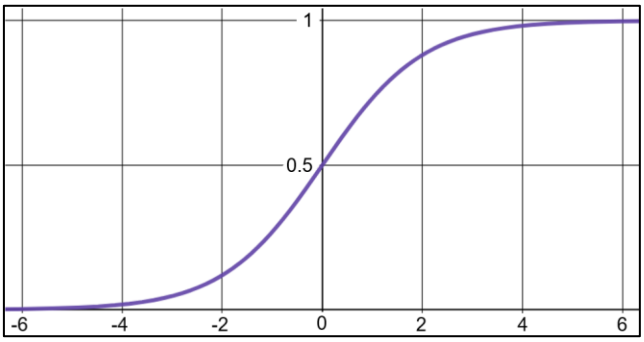
תזכורת:

כלל החלטה $h: \mathbb{R}^d \to \{-1,+1\}$ נקרא ליניארי אם ורק אם קיימים $\mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}$ כך ש-$h(\mathbf{x}) = \mathrm{sgn}(\mathbf{w}^\top \mathbf{x} + b)$ $\forall \mathbf{x} \in \mathbb{R}^d$:.

נגדיר: $\mathrm{sgn}(z) = \begin{cases} -1, & z \le 0 \\ +1, & z > 0 \end{cases}$

(כך שלא מתקבל אפס לשום קלט)

תזכורת: $\sigma(z) = \frac{1}{1+\exp\{-z\}}$



ובפרט מתקיים:

$\sigma(0) = 0.5, \quad \sigma(1) \approx 0.73, \quad \sigma(2) \approx 0.88, \quad \sigma(3) \approx 0.95$

$\sigma(4) \approx 0.98, \quad \sigma(5) \approx 0.99, \quad \sigma(6) \approx 1$

א. [5 נק'] הוכיחו שכלל ההחלטה של logistic regression (עבור $\mathbf{w}_1 \in \mathbb{R}^d, b_1 \in \mathbb{R}$ כלליים) הוא ליניארי:

$$h_1(\mathbf{x}) = \begin{cases} +1, & \sigma(\mathbf{w}_1^\top \mathbf{x} + b_1) > 0.5 \\ -1, & \text{אחרת} \end{cases}.$$

הוכחה:

$h_1(x) = \sigma(w_1^T x + b_1) = \frac{1}{1+e^{-(w_1^T x + b_1)}} > 0.5$

$\frac{1}{2} + \frac{1}{2} e^{-(w_1^T x + b_1)} < 1$

$e^{-(w_1^T x + b_1)} < 1 \quad / \cdot \ln$

$-(w_1^T x + b_1) < 0 \quad / \cdot (-1)$

$w_1^T x + b_1 > 0 \Rightarrow \quad h_1(x) = 1 = \mathrm{Sign}(w_1^T x + b_1)$

אחרת $h_1(x) = -1 = \mathrm{Sign}(w_1^T x + b_1)$

סה"כ $\exists w_1, b_1 : h_1(x) = \mathrm{Sign}(w_1^T x + b_1) \Leftrightarrow$ ליניארי

להלן טענת עזר שניתן להשתמש בה במידת הצורך בהמשך התרגיל.

טענת עזר: כלל החלטה ליניארי h צריך לקיים $\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d:\ \underbrace{h(\mathbf{x}_1) = h(\mathbf{x}_2)}_{\text{if}} \Longrightarrow \underbrace{h\left(\frac{1}{2}\mathbf{x}_1 + \frac{1}{2}\mathbf{x}_2\right) = h(\mathbf{x}_1)}_{\text{then}}$.

ב. [10 נק'] עבור $\mathbf{w}_3 \in \mathbb{R}^d, b_3 \in \mathbb{R}$ כלליים כלשהם.

הפריכו: כלל ההחלטה הבא ליניארי: $h_3(\mathbf{x}) = \begin{cases} +1, & \sigma(\mathbf{w}_3^\top \sigma(\mathbf{x}) + b_3) > 0.5 \\ -1, & \text{אחרת} \end{cases}$,

כאשר הפעולה $\sigma(\mathbf{x})$ מחזירה וקטור $[\sigma(x[1]), \sigma(x[2]), \dots, \sigma(x[d])]^\top$.

רמז: כדאי להשתמש בטענת העזר. ניתן לבחור $\mathbf{w}_3, b_3$ ספציפיים כדי להפריך.

הפרכה (לרשותכם דפי טיוטה בסוף הגיליון):

רעיון: נקח 2 נק' המסווגות ב (+1) ונניח שהוא הממוצע שלהן המתויגת ב -1

הדוגמה לא פריכה ליניאריות אך לאחר הפלט h_3 נראה כי הוא

כן, כלומר הסיווג של כולן יהיה נכון

$\sigma(6) \neq 1$ ולכן ננסה לבחור נק' רק שהיחלה תהיה 1

וגם הממוצע 1, אך הקלט לא יהיה 1

(בדו-מימד, ואחד מן המימדים יהיה ציר והשני לא

$+\ (3,14)$

$-\ (2,8)$ לא פריך

$+\ (1,2)$

$(0.88,1)\ -$ $(0.95,1)\ +$

$+\ (0.73, 0.88)$ פריך

סל

ג. [10 נק'] נתונים $u_2 \in \mathbb{R}, \mathbf{w}_2 \in \mathbb{R}^d$ כלליים כלשהם ומתקיים $u_2 > 1$.

ישר

הוכיחו/הפריכו: כלל ההחלטה הבא ליניארי: $h_2(\mathbf{x}) = \begin{cases} +1, & \sigma(u_2\sigma(\mathbf{w}_2^{\top}\mathbf{x}) - 1) > 0.5 \\ -1, & \text{אחרת} \end{cases}$.

תשובה:

* לוקחים מסווג $x^T w_2$ ועושים לו טרנספורמציה כלשהי (נותנים לכל ערך בין 0 ל1)

וכך שסווגה ~~כ~~ ב(+) תפלט עבור חצי בין 1 ל0.5

ב(−) ~~לא~~ ~~צכיון~~ 0.5 ל 0

בכל מכפלה ערך גדול מ0, וu_2 לא ישנה את הנוכחה לנו

כלומר

חלק ב' – שאלה אמריקאית [6 נק']

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

א. בתרגילי הבית ראינו איך לבצע tuning על היפר-פרמטרים רבים בעזרת Grid search cross validation.
נתון אלגוריתם למידה $\mathcal{A}$ עם אוסף $\mathcal{P}$ של היפר-פרמטרים שדורשים tuning. נניח שלכל היפר-פרמטר יש בדיוק v ערכים שנבדקים. נסמן את מספר ה-folds בתהליך ה-CV בתור k ואת מספר דוגמאות האימון בתור m.

סמנו את **<u>כֹּל</u>** הטענות הנכונות בהכרח ביחס לריצה של תהליך ה-Grid search CV:

a. מספר הקריאות ל-$\mathcal{A}$ גדל בצורה ליניארית ב-$|\mathcal{P}|$.

b. מספר הקריאות ל-$\mathcal{A}$ גדל בצורה ליניארית ב-v.

c. מספר הקריאות ל-$\mathcal{A}$ גדל בצורה ליניארית ב-k.

d. סיבוכיות זמן הריצה של קריאה בודדת ל-$\mathcal{A}$ גדלה בצורה ליניארית ב-m.

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

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

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