הטכניון – מכון טכנולוגי לישראל

הפקולטה למדעי המחשב

מבוא למערכות לומדות (236756)

סמסטר אביב תשפ"ב – 23 בספטמבר 2022

מרצה: ד"ר ניר רוזנפלד

מבחן מסכם מועד ב' – פיתרון חלקי

שימו לב: הפתרונות המופיעים כאן הם חלקיים בלבד ומובאים בשביל לעזור לכם בתהליך הלמידה.

ייתכנו כאן חוסרים / ליקויים / טעויות של ממש.

הנחיות הבחינה:

- **משך הבחינה:** 3 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.
- קריאוּת:
  - 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 14 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.**

מבנה הבחינה:

- **חלק א' [76 נק']:** 3 שאלות פתוחות.
- **חלק ב' [24 נק']:** 4 שאלות סגורות (אמריקאיות) [כל אחת 6 נק'].

בהצלחה!

חלק א' – שאלות פתוחות [76 נק']

שאלה 1: Multi-Layer Perceptron (MLP) [26 נק']

נתון דאטה דו-ממדי עם סיווגים בינאריים (± 1).

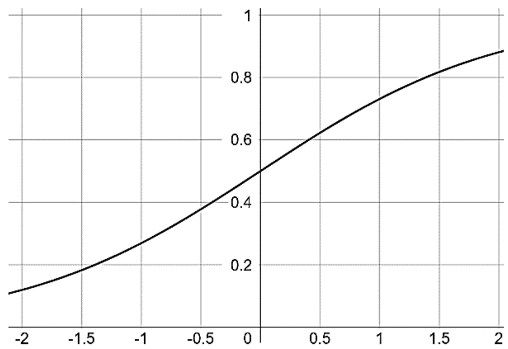
נבנה רשת MLP עם שתי שכבות ליניאריות בתור פונקציה $F: \mathbb{R}^2 \to (0,1)$ המוגדרת:

$$F(x_1, x_2) = \sigma(w_5 \cdot \mathrm{ReLU}(w_1 x_1 + w_2 x_2 + b_1) + w_6 \cdot \mathrm{ReLU}(w_3 x_1 + w_4 x_2 + b_2) + b_3),$$

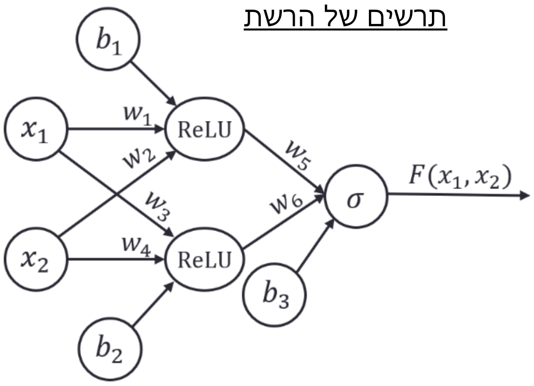
כאשר $w_1, \dots, w_6, b_1, b_2, b_3 \in \mathbb{R}$ הם פרמטרים סקלריים, פונקציית האקטיבציה היא $\mathrm{ReLU}(z) = \begin{cases} 0, & z \le 0 \\ z, & z > 0 \end{cases}$

והפלט של הרשת עובר דרך פונקציית הסיגמואיד: $\sigma(z) = \frac{1}{1+\exp\{-z\}}$.

פונקציית הסיגמואיד



תרשים של הרשת



נכין את הרשת לאימון.

נשים לב שהרשת מחזירה הסתברות ובסעיפים הבאים נשתמש ב-Negative-log-likelihood-loss המוגדר בתור:

$$\ell(\underbrace{x}_{\in\{0,1\}^2}, \underbrace{y}_{\in\{0,1\}}) = -y \ln(F(x_1, x_2)) - (1-y)\ln(1 - F(x_1, x_2)) .$$

א. [2 נק'] חשבו את הנגזרת החלקית $\frac{\partial \ell}{\partial F}$.

תשובה סופית (לרשותכם טיוטה בסוף הגיליון):

$$\frac{\partial \ell(x,y)}{\partial F} = -\frac{y}{F(x_1,x_2)} + \frac{1-y}{1-F(x_1,x_2)}$$

ב. [2 נק'] כִּתְבוּ פונקציה שמהווה subgradient לפונקציית ה-ReLU.

תשובה סופית (לרשותכם טיוטה בסוף הגיליון):

$$\mathrm{ReLU}'(z) = \begin{cases} 0, & z \le 0 \\ 1, & z > 0 \end{cases}$$

לשם הפשטות, נגדיר שלושה סימוני עזר:

$$F(x_1,x_2)=\sigma(\underbrace{w_5\cdot \mathrm{ReLU}(\overbrace{w_1x_1+w_2x_2+b_1}^{\triangleq a_1})+w_6\cdot \mathrm{ReLU}(\overbrace{w_3x_1+w_4x_2+b_2}^{\triangleq a_2})+b_3}_{\triangleq a_3}).$$

לשימושכם בהמשך, להלן כמה נגזרות חלקיות מהשכבה הראשונה:

$\frac{\partial a_3}{\partial w_4}=w_6\cdot \mathrm{ReLU}'(a_2)\cdot x_2$	$\frac{\partial a_3}{\partial w_3}=w_6\cdot \mathrm{ReLU}'(a_2)\cdot x_1$	$\frac{\partial a_3}{\partial w_2}=w_5\cdot \mathrm{ReLU}'(a_1)\cdot x_2$	$\frac{\partial a_3}{\partial w_1}=w_5\cdot \mathrm{ReLU}'(a_1)\cdot x_1$
		$\frac{\partial a_3}{\partial b_2}=w_6\cdot \mathrm{ReLU}'(a_2)$	$\frac{\partial a_3}{\partial b_1}=w_5\cdot \mathrm{ReLU}'(a_1)$

ומהשכבה השנייה:

$\frac{\partial a_3}{\partial b_3}=1$	$\frac{\partial a_3}{\partial w_6}=\mathrm{ReLU}(a_2)$	$\frac{\partial a_3}{\partial w_5}=\mathrm{ReLU}(a_1)$

ג. [2 נק'] חשבו את הנגזרת החלקית $\frac{\partial \ell}{\partial a_3}$ (שימו לב שכבר חישבנו את $\frac{\partial \ell(x,y)}{\partial F}$).

<u>תכונת עזר</u>: הנגזרת של הסיגמואיד היא $\frac{\mathrm{d}}{\mathrm{dz}}\sigma(z)=\sigma(z)\big(1-\sigma(z)\big)$.

תשובה סופית:

$$\frac{\partial \ell(x,y)}{\partial a_3}=-y+\sigma(a_3)$$

שני הסעיפים הבאים מדגימים **בעיה** שיכולה לקרות בזמן אימון עם ReLU.

ד. [7 נק'] נניח שהפרמטרים מאותחלים באופן הבא: $w_1=\cdots=w_6=0,\quad b_1=b_2=b_3=-1$.

נחשב את <u>ערכי</u> הפרמטרים אחרי צעד gradient descent <u>יחיד</u> לפי דוגמה <u>נתונה</u> (x,y) עם גודל צעד $\eta=1$.

מלאו את התשובות הסופיות בטבלאות.

<u>שימו לב</u>: התשובות יכולות להיות תלויות ב-(x,y) אבל לא ב-a_1,a_2,a_3.

מותר להשאיר ביטויים כמו $\sigma(c)$ כאשר $c\in\mathbb{R}$ מספר קבוע <u>מפורש</u>, מבלי לחשב את ערכם במחשבון.

<u>First layer</u>

Parameter	Value
w_1	0
w_2	0
w_3	0
w_4	0
b_1	-1
b_2	-1

<u>Second layer</u>

Parameter	Value
w_5	0
w_6	0
b_3	$-1-(\sigma(-1)-y)$

ה. [7 נק'] מה יקרה אחרי $T\geq 2$ צעדי גרדיינט (לפי <u>אותה</u> דוגמה (x,y) ואותו η)? ענו בקצרה ובאופן איכותי (qualitative).

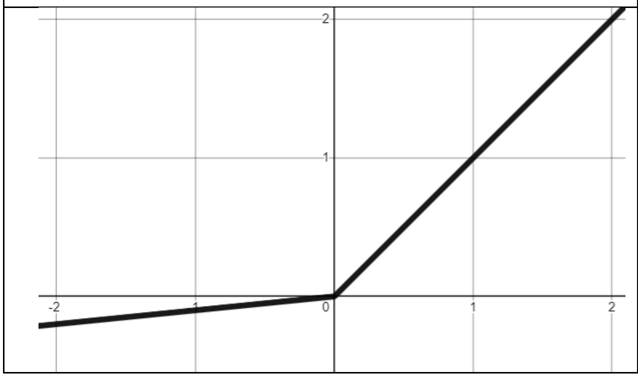
תשובה סופית (לרשותכם טיוטה בסוף הגיליון):

הכל יישאר ללא שינוי חוץ מ-b_3, כי כל הגרדיינטים האחרים תמיד יהיו אפסים.

ו. [6 נק'] אילו מפונקציות האקטיבציה הבאות ימנעו את הבעיה שהדגמנו בסעיפים הקודמים (עבור אתחול זהה)?
סמנו את <u>כל</u> האפשרויות המתאימות.

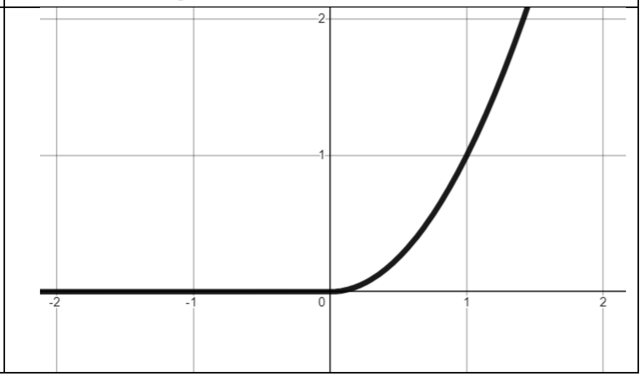
(A)

Leaky ReLU: $\max\{0.1x, x\}$



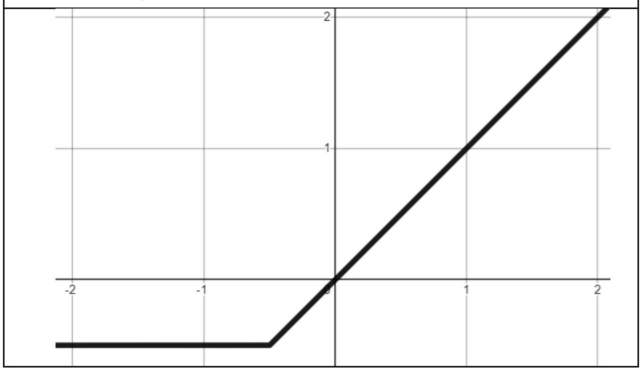
(B)

Squared ReLU: $(\max\{0,x\})^2$



(C)

1/2-Shifted ReLU: $\max\{-0.5, x\}$



שאלה 2: מסווגים ליניאריים [20 נק']

נתון דאטה d-ממדי $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ עם סיווגים בינאריים (± 1).

א. [10 נק'] נשתמש ב-Hinge loss ונגדיר את הבעיה <u>הקמורה</u> (ללא רגולריזציה): $\underset{\boldsymbol{w}\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1}{m}\sum_{i=1}^m \max\{0, 1 - y_i \boldsymbol{w}^\top \boldsymbol{x}_i\}$.

i. עבור דוגמה כלשהי $(\boldsymbol{x}_i, y_i)$, האם הפונקציה $\max\{0, 1 - y_i \boldsymbol{w}^\top \boldsymbol{x}_i\}$ חוסמת מלמעלה את ה-0-1 loss? נמקו בקצרה.

תשובה והסבר קצר:

כן.

ii. **נתון:** הדאטה פריד ליניארית הומוגנית.

הוכיחו \ הפריכו: לבעיה שהוגדרה קיים פיתרון <u>אופטימלי</u> כלשהו $\boldsymbol{w}^* \in \mathbb{R}^d$ עם נורמה <u>סופית</u> (משמע $\|\boldsymbol{w}^\star\|_2 < \infty$).

תשובה:

לפי הנתונים קיים $\boldsymbol{w}$ מפריד "נכון", משמע $\forall i \in [m]$ מתקיים $z_i \triangleq y_i \boldsymbol{w}^\top \boldsymbol{x}_i > 0$.

יש אוסף סופי של נקודות אימון ולכן ניתן להגדיר $\alpha \triangleq \frac{1}{\min_{i\in[m]} y_i \boldsymbol{w}^\top \boldsymbol{x}_i}$.

נגדיר מפריד חדש $\boldsymbol{w}_\alpha \triangleq \alpha \boldsymbol{w}$.

כעת $\forall i \in [m]$ מתקיים $\mathcal{L}(\boldsymbol{w}_\alpha) = \frac{1}{m}\sum_{i=1}^m \max\left\{0, 1 - \underbrace{\alpha y_i \boldsymbol{w}^\top \boldsymbol{x}_i}_{\geq 1}\right\} = 0$.

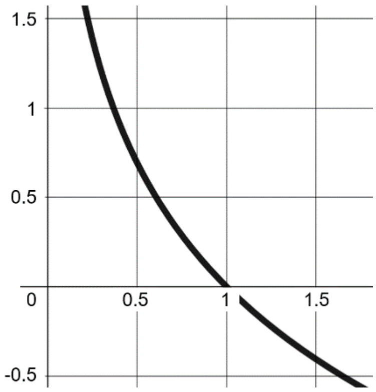
ב. [10 נק'] נשתמש ב-Log. loss ונגדיר את הבעיה <u>הקמורה</u> (ללא רגולריזציה): $\underset{\boldsymbol{w}\in\mathbb{R}^d}{\operatorname{argmin}}\left\{\frac{1}{m}\sum_{i=1}^{m}-\ln\left(\frac{1}{1+\exp\{-y_i\boldsymbol{w}^\top\boldsymbol{x}_i\}}\right)\right\}$.

i. עבור דוגמה כלשהי $(\boldsymbol{x}_i,y_i)$, האם הפונקציה $-\ln\left(\frac{1}{1+\exp\{-y_i\boldsymbol{w}^\top\boldsymbol{x}_i\}}\right)$ חוסמת מלמעלה את ה-loss 0-1? נמקו בקצרה.

תשובה והסבר קצר:

לא. למשל עבור $-y_i\boldsymbol{w}^\top\boldsymbol{x}_i=\epsilon\to 0$ מקבלים $-\ln\frac{1}{2}\approx 0.7<1$.

<u>לשימושכם:</u>
תרשים של הפונקציה
$-\ln(x)$



ii. **נתון:** הדאטה פריד ליניארית הומוגנית.

הוכיחו \ הפריכו: לבעיה שהוגדרה קיים פיתרון <u>אופטימלי</u> כלשהו $\boldsymbol{w}^*\in\mathbb{R}^d$ עם נורמה <u>סופית</u> (משמע $\|\boldsymbol{w}^*\|_2<\infty$).

תשובה:

לפי הנתונים קיים $\boldsymbol{w}$ מפריד "נכון", משמע $\forall i\in[m]$ מתקיים $z_i\triangleq y_i\boldsymbol{w}^\top\boldsymbol{x}_i>0$.

ויהי מפריד חדש $\boldsymbol{w}_\alpha\triangleq\alpha\boldsymbol{w}$ עבור $\alpha>1$.

ככל שנגדיל את α נקבל loss נמוך יותר כי $\frac{1}{1+\exp\{\underbrace{-\alpha z_i}_{<0}\}}$ מונוטוני עולה ב-α ו-$(-\ln x)$ מונוטוני יורד ב-x.

שאלה 3: Kernel SVM [30 נק']

א. [12 נק'] נתון דאטה d-ממדי $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ עם סיווגים בינאריים (± 1). נתונה פונקציית Kernel כלשהי $K: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$.

צוות מחקר פתר שתי בעיות אופטימיזציה שלמדנו:

(i) Dual Linear SVM לפי ה-raw features. נסמן את וקטור המשתנים הדואליים שנלמדו בתור $\boldsymbol{\alpha} \in \mathbb{R}_+^m$.

(ii) Dual Kernel SVM לפי פונקציית הקרנל K. נסמן את וקטור המשתנים הדואליים שנלמדו בתור $\boldsymbol{\alpha}' \in \mathbb{R}_+^m$.

נתון שבשני המקרים נמצאו פתרונות שמשתמשים בְּ-$\lceil \log m \rceil$ וקטורים בתור support vectors
(משמע, בכל אחד מהפתרונות $\boldsymbol{\alpha}, \boldsymbol{\alpha}'$ יש בדיוק $\lceil \log m \rceil$ כניסות שאינן 0).

בזמן מבחן (לאחר האימון) כשמקבלים דוגמה חדשה לסיווג $\boldsymbol{x} \in \mathbb{R}^d$, כללי ההחלטה של המודלים הינם:

Linear SVM

$$h_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \operatorname{sign}\left(\sum_{i=1}^m \alpha_i y_i \boldsymbol{x}_i^\top \boldsymbol{x}\right)$$

Kernel SVM

$$h_{\boldsymbol{\alpha}'}(\boldsymbol{x}) = \operatorname{sign}\left(\sum_{i=1}^m \alpha'_i y_i K(\boldsymbol{x}_i, \boldsymbol{x})\right)$$

(i) בזמן המבחן, מה סיבוכיות <u>המקום</u> המינימלית שנדרשת עבור כלל ההחלטה של Linear SVM? סמנו והסבירו בקצרה.

a. $\mathcal{O}(d)$ c. $\mathcal{O}(\log(m)\cdot d)$ e. $\mathcal{O}(m^2)$

b. $\mathcal{O}(m)$ d. $\mathcal{O}(m \cdot d)$ f. $\mathcal{O}(d^2)$

הסבר <u>תמציתי</u>:

מדובר במסווג לינארי. צריך לשמור רק וקטור מפריד יחיד. מתקיים $h_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \operatorname{sign}\left(\underbrace{\left(\sum_{i=1}^m \alpha_i y_i \boldsymbol{x}_i^\top\right)}_{\triangleq \boldsymbol{w}^\top} \boldsymbol{x}\right)$

הערות בדיקה: למי שסימנו את תשובה (c) ירד ניקוד חלקי בלבד.

(ii) בזמן המבחן, מה סיבוכיות <u>המקום</u> המינימלית שנדרשת עבור כלל ההחלטה של Kernel SVM (ללא הנחות על הקרנל)?

a. $\mathcal{O}(d)$ c. $\mathcal{O}(\log(m)\cdot d)$ e. $\mathcal{O}(m^2)$

b. $\mathcal{O}(m)$ d. $\mathcal{O}(m \cdot d)$ f. $\mathcal{O}(d^2)$

הסבר <u>תמציתי</u>:

למדנו ש-RBF-Kernel מוגדר בתור: $K(\boldsymbol{u},\boldsymbol{v}) = \exp\left\{-\frac{1}{2\sigma^2}\|\boldsymbol{u}-\boldsymbol{v}\|_2^2\right\}$ עבור היפרפרמטר $\sigma^2 > 0$.

ב. [4 נק'] כעת, נבין את ההתנהגות של כלל ההחלטה של RBF-Kernel SVM בגבול $\sigma^2 \to \infty$.

ניתן להניח:

- וקטור המקדמים $\boldsymbol{\alpha}' \in \mathbb{R}_+^m$ הדואליים חסום. משמע, קיים $0 < c_1 < \infty$ כך שמתקיים $\|\boldsymbol{\alpha}'\|_2 \le c_1$.
- הדוגמאות בהתפלגות חסומות. משמע, קיים $0 < c_2 < \infty$ כך שֶׁ-$\forall \boldsymbol{x} \in \mathcal{X}$ מתקיים $\|\boldsymbol{x}\|_2 \le c_2$.

חשבו את הגבול $\lim_{\sigma^2\to\infty} \text{sign}(\sum_{i=1}^m \alpha_i' y_i K(\boldsymbol{x}_i,\boldsymbol{x})) = \lim_{\sigma^2\to\infty} h_{\alpha'}(\boldsymbol{x})$.

רמז: כאן הגבול מקיים $\lim_{\sigma^2\to\infty} \text{sign}(\sum_{i=1}^m \alpha_i' y_i K(\boldsymbol{x}_i,\boldsymbol{x})) = \text{sign}\left(\lim_{\sigma^2\to\infty} (\sum_{i=1}^m \alpha_i' y_i K(\boldsymbol{x}_i,\boldsymbol{x}))\right)$.

תשובה:

$$= \text{sign}\left(\sum_{i=1}^m \alpha_i' y_i\right) = \mathbb{I}\left[\left(\sum\nolimits_{y_i=+1} \alpha_i'\right) > \left(\sum\nolimits_{y_i=-1} \alpha_i'\right)\right]$$

ומקבלים כלל החלטה קבוע שתלוי במקדמים.

ג. [7 נק'] נתונה התפלגות $\mathcal{D}$ כלשהי על דוגמאות d-ממדיות חסומות (נניח $\forall \boldsymbol{x} \in \mathcal{D}: \|\boldsymbol{x}\|_2 \le 1$) וסיווגים בינאריים ($\pm 1$) מתאימים. וידוע שההתפלגות מאוזנת כך שמתקיים $\Pr_{(x,y)\sim\mathcal{D}}[y=1] = \Pr_{(x,y)\sim\mathcal{D}}[y=-1] = \frac{1}{2}$.

דוגמים 200 דוגמאות אימון ומאמנים עליהן חמישה מודלים שונים. לפניכם טבלה עם תוצאות האימון וההכללה.

דיוק / מודל	(א)	(ב)	(ג)	(ד)	(ה)
אימון	53%	92%	89%	100%	100%
הכללה	50%	89%	50%	23%	84%

מבין חמשת המודלים שנלמדו, שניים הם מודלי RBF-Kernel SVM עם ערכי σ^2 קיצוניים מאוד: $[10^{-6}, 10^6]$.

אילו?

הנחה: לשם פשטות, הניחו שבשני המודלים האלה הווקטור הדואלי $\boldsymbol{\alpha}' \in \mathbb{R}_+^m$ שנלמד מקיים $\forall i: \alpha_i' \in [0.1, 10]$.

הערות: אנו עוסקים במקרה הסביר ולא במקרי קצה. מדובר בניתוח אנליטי, לכן הניחו שאין שגיאות נומריות.

i. איזו עמודה מתאימה למודל RBF עם $\sigma^2 = 10^6$? (א) (ב) (ג) (ד) (ה)

ii. איזו עמודה מתאימה למודל RBF עם $\sigma^2 = 10^{-6}$? (א) (ב) (ג) (ד) (ה)

הסעיף הבא בלתי תלוי בסעיפים הקודמים.

נתונה נקודה $\boldsymbol{w} \in \mathbb{R}^d$.

נגדיר את הפונקציה $K: \left(\mathbb{R}^d \times \mathbb{R}^d\right) \to \mathbb{R}$ בתור $K(\boldsymbol{u}, \boldsymbol{v}) = \frac{1}{2}\left(\|\boldsymbol{u} - \boldsymbol{w}\|^2 + \|\boldsymbol{v} - \boldsymbol{w}\|^2 - \|\boldsymbol{u} - \boldsymbol{v}\|^2\right)$.

ד. [7 נק'] הוכיחו שהפונקציה K מהווה קרנל חוקי.

עשו זאת ע"י הגדרה ברורה של פונקציית מיפוי $\phi: \mathbb{R}^d \to \mathbb{R}^p$ והוכחה שמתקיים $K(\boldsymbol{u}, \boldsymbol{v}) = \langle \phi(\boldsymbol{u}), \phi(\boldsymbol{v}) \rangle$.

רמז: מומלץ להגדיר מיפוי שמקיים $p = d$.

תשובה (לרשותכם טיוטה בסוף הגיליון):

המיפוי הוא $\phi(\boldsymbol{z}) = \boldsymbol{z} - \boldsymbol{w}$.

חלק ב' – שאלות רב-ברירה [24 נק']

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

הערות בדיקה: בסעיפים ב'-ד' ירדו 2 נקודות על כל טענה מיותרת שסומנה או טענה נכונה שחסרה.

א. [6 נק'] נתונים מרחב דוגמאות $\mathcal{X} \subseteq \mathbb{R}^d$ ומחלקת היפותזות כלשהי $\mathcal{H}$ מעל $\mathcal{X}$.

בנוסף, נתונה תת-קבוצה $\mathcal{X}' \subset \mathcal{X}$.

נגדיר את מחלקת ההיפותזות $\mathcal{Q}$ על ידי **צמצום תחום ההגדרה** של ההיפותזות ב-$\mathcal{H}$ לתת-הקבוצה $\mathcal{X}'$:

$$\mathcal{Q} = \{\, q_h \triangleq h|_{\mathcal{X}'} \mid h \in \mathcal{H} \,\}, \;\; \text{where} \;\; q_h(x) = \begin{cases} h(x), & x \in \mathcal{X}' \\ \text{undefined}, & x \notin \mathcal{X}' \end{cases}.$$

סמנו את הטענה הנכונה.

a. מתקיים בהכרח $VCdim(\mathcal{H}) \geq VCdim(\mathcal{Q})$ וייתכנו מקרים שבהם $VCdim(\mathcal{H}) > VCdim(\mathcal{Q})$.

b. מתקיים בהכרח $\mathrm{VCdim}(\mathcal{H}) \leq \mathrm{VCdim}(\mathcal{Q})$ וייתכנו מקרים שבהם $\mathrm{VCdim}(\mathcal{H}) < \mathrm{VCdim}(\mathcal{Q})$.

c. מתקיים בהכרח $\mathrm{VCdim}(\mathcal{H}) = \mathrm{VCdim}(\mathcal{Q})$.

d. כל הטענות הקודמות שגויות.

ב. [6 נק'] סמנו את כׇּל הטענות הנכונות ביחס ל-**Feature selection**.

a. שיטות Wrapper (למשל Sequential feature selection) יש להפעיל לפני שלב ה-data imputation.

b. שיטות Wrapper (למשל Sequential feature selection) יש להפעיל לפני שלב ה-data normalization.

c. בבעיות סיווג: לפני האימון, ניתן להסיר כל פיצ'ר שיש קורלציה 0 בינו לבין ה-target variable, מבלי לפגוע בביצועים של אלגוריתמי למידה על סֵט האימון.

d. נתון עץ החלטה כלשהו בעומק L (מספר הקשתות המקסימלי מהשורש לעלה כלשהו).
כפי שלמדנו, כל צומת מְסַוֵּג לשתי אפשרויות בעזרת threshold על פיצ'ר אחד.
אזי, העץ כולו משתמש לכל היותר ב-$(2L-1)$ פיצ'רים.

e. מאמנים מסווג AdaBoost עם Decision stump כמסווג בסיס במשך T איטרציות.
אזי, המסווג ה"חזק" שמתקבל משתמש לכל היותר ב-T פיצ'רים.

ג. [6 נק'] נתונות שתי פונקציות קמורות $f,g\colon C\to\mathbb{R}$ המוגדרות מעל סט קמור C.

סמנו את כֹּל הטענות הנכונות בהכרח.

a. הפונקציה $h(z)=f(z)+g(z)$ הינה קמורה.

b. הפונקציה $h(z)=max\{f(z),g(z)\}$ הינה קמורה.

c. הפונקציה $h(z)=\min\{f(z),g(z)\}$ הינה קמורה.

d. הפונקציה $h(z)=f\big(g(z)\big)$ הינה קמורה.

e. הפונקציה $h(z)=af(z)+b$ הינה קמורה לכל $a,b\in\mathbb{R}$.

ד. [6 נק'] ניזכר בשתי פונקציות loss שלמדנו: $\ell_{\text{hinge}}(z)=\max\{0,1-z\},\ \ \ell_{\text{ramp}}(z)=\min\{1,\max\{0,1-z\}\}$.

מגדירים שתי בעיות סיווג ליניארי (עם דאטה זהה):

$$\underbrace{\operatorname{argmin}_{\boldsymbol{w}\in\mathbb{R}^d}\frac{1}{m}\sum_{i=1}^{m}\ell_{\text{ramp}}(y_i\boldsymbol{w}^\top\boldsymbol{x}_i)}_{\triangleq \mathrm{P}_{\text{ramp}}}\quad,\quad\underbrace{\operatorname{argmin}_{\boldsymbol{w}\in\mathbb{R}^d}\frac{1}{m}\sum_{i=1}^{m}\ell_{\text{hinge}}(y_i\boldsymbol{w}^\top\boldsymbol{x}_i)}_{\triangleq \mathrm{P}_{\text{hinge}}}.$$

סמנו את כֹּל הטענות הנכונות (השאלה עוסקת במקרה הסביר ולא במקרי קצה).

a. הבעיה P_{hinge} צפויה להיות יותר רגישה ל-*outliers* מאשר הבעיה P_{ramp}.

b. הבעיה P_{hinge} קמורה ואילו הבעיה P_{ramp} אינה קמורה.

c. עבור הבעיה P_{hinge}, נקודה בה הנגזרת מוגדרת ומתאפסת היא מינימום גלובאלי.

d. עבור הבעיה P_{ramp}, נקודה בה הנגזרת מוגדרת ומתאפסת היא מינימום גלובאלי.

e. ערך המינימום הגלובאלי של P_{hinge} הוא 0 $\Leftrightarrow$ ערך המינימום הגלובאלי של P_{ramp} הוא 0.