



מבוא למערכות לומדות (236756)

סמסטר אביב תשפ"ב – 23 בספטמבר 2022

מרצה: ד"ר ניר רחנפלד

מבחן מסכם מועד ב'

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.
- 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 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']

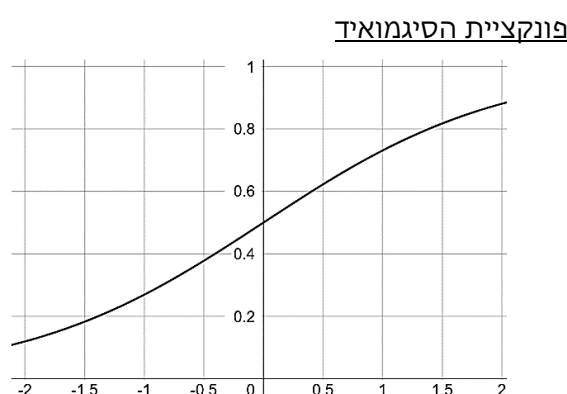
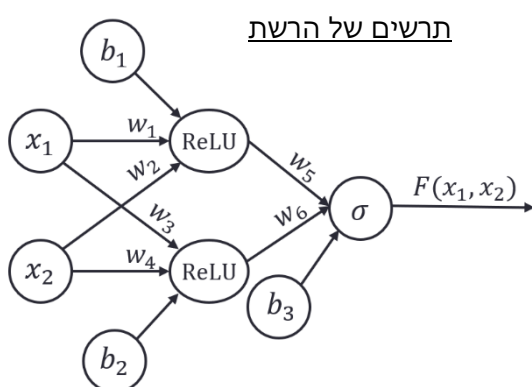
נתון דאטה דו-ממדי עם סיווגים בינאריים (± 1) .

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

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

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

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



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):

$$\frac{\partial \ell(x, y)}{\partial F} =$$

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

תשובה סופית (לרשותכם טיוטה בסוף הגיליון):

$$\text{ReLU}'(z) =$$

לשם הפשטות, נגדיר שלושה סימוני עזר:

$$F(x_1, x_2) = \sigma(\underbrace{w_5 \cdot \text{ReLU}(\underbrace{w_1x_1 + w_2x_2 + b_1}_{\triangleq a_1})}_{\triangleq a_3} + \underbrace{w_6 \cdot \text{ReLU}(\underbrace{w_3x_1 + w_4x_2 + b_2}_{\triangleq a_2})}_{\triangleq a_4} + b_3)$$

לשימושכם בהמשך, להלן כמה נגזרות חלקיות מהשכבה הראשונה:

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

$\frac{\partial a_3}{\partial w_5} = \text{ReLU}(a_1)$	$\frac{\partial a_3}{\partial w_6} = \text{ReLU}(a_2)$	$\frac{\partial a_3}{\partial b_3} = 1$	ומהשכבה השנייה:
--	--	---	-----------------

ג. [2 נק'] חשבו את הנגזרת החלקית $\frac{\partial \ell}{\partial a_3}$ (שימו לב שכבר חישבנו את $\frac{\partial \ell(x,y)}{\partial F}$).
 תכונת עזר: הנגזרת של הסיגמואיד היא $\frac{d}{dz} \sigma(z) = \sigma(z)(1 - \sigma(z))$.

תשובה סופית:

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

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

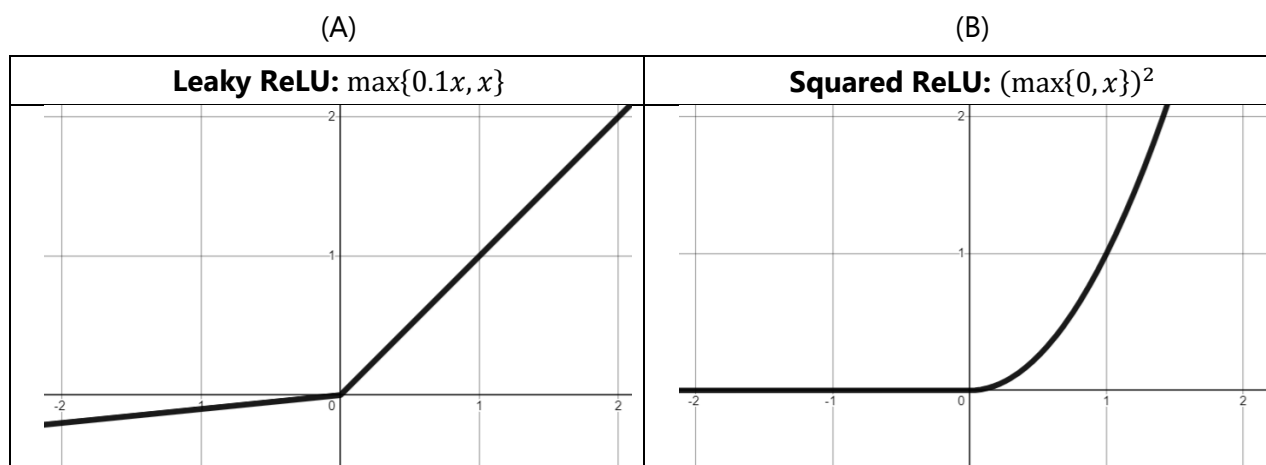
ד. [7 נק'] נניח שהפרמטרים מאותחלים באופן הבא: $w_1 = \dots = w_6 = 0, \quad b_1 = b_2 = b_3 = -1$.
 נחשב את ערכי הפרמטרים אחרי צעד gradient descent יחיד לפי דוגמה נתונה (x, y) עם גודל צעד $\eta = 1$.
 מלאו את התשובות הסופיות בטבלאות.
שימו לב: התשובות יכולות להיות תלויות ב- (x, y) אבל לא ב- a_1, a_2, a_3 .
 מותר להשאיר ביטויים כמו $\sigma(c)$ כאשר $c \in \mathbb{R}$ מספר קבוע מפורש, מבלי לחשב את ערכם במחשבון.

First layer		Second layer	
Parameter	Value	Parameter	Value
w_1		w_5	
w_2		w_6	
w_3		b_3	
w_4			
b_1			
b_2			

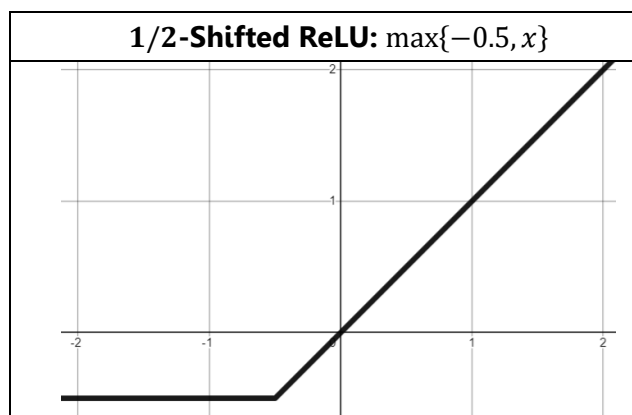
ה. [7 נק'] מה יקרה אחרי $T \geq 2$ צעדי גרדיינט (לפי אותה דוגמה (x, y) ואותו η)? ענו בקצרה ובאופן איכותי (qualitative).

תשובה סופית (לרשותכם טיוטה בסוף הגיליון):

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

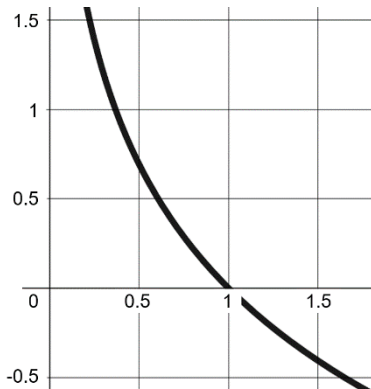


(C)



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

לשימושכם:
תרשים של הפונקציה
 $-\ln(x)$



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

[illegible]

שאלה 3: Kernel SVM [30 נק']

א. [12 נק'] נתון דאטה d -ממדי $\{(x_i, y_i)\}_{i=1}^m$ עם סיווגים בינאריים (± 1) . נתונה פונקציית Kernel כלשהי $K: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$. צוות מחקר פתר שתי בעיות אופטימיזציה שלמדנו:

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

נתון שבשני המקרים נמצאו פתרונות שמשתמשים ב- $\log m$ וקטורים בתור support vectors (משמע, בכל אחד מהפתרונות α, α' יש בדיוק $\log m$ כניסות שאינן 0).

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

Kernel SVM

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

Linear SVM

$$h_{\alpha}(x) = \text{sign} \left(\sum_{i=1}^m \alpha_i y_i x_i^T x \right)$$

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

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

הסבר תמציתי:

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

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

הסבר תמציתי:

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

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

ניתן להניח:

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

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

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

תשובה:

ג. [7 נק'] נתונה התפלגות \mathcal{D} כלשהי על דוגמאות d -ממדיות חסומות (נניח $\|\mathbf{x}\|_2 \leq 1 \forall \mathbf{x} \in \mathcal{D}$) וסיווגים בינאריים (± 1) מתאימים. וידוע שההתפלגות מאחזת 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 מאד: $[10^{-6}, 10^6]$.

אילו? (יצאה הבהרה בזמן הבחינה שלצורך השאלה, ערכי σ^2 המדוברים שואפים לאינסוף ולאפס).

הנחה: לשם פשטות, הניחו שבשני המודלים האלה הווקטור הדואלי $\alpha' \in \mathbb{R}_+^m$ 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)

תשובה (לרשותכם טיוטה בסוף הגיליון):

חלק ב' – שאלות רב-ברירה [24 נק']

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

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

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

d. כל הטענות הקודמות שגויות.

ב. [6 נק'] סמנו את כל 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 סט האימון.

d. נתון עץ החלטה כלשהו בעומק L (מספר הקשתות המקסימלי מהשורש לעלה כלשהו).

כפי שלמדנו, כל צומת מסוג לשתי אפשרויות בעזרת threshold על פיצ'ר אחד.

אזי, העץ כולו משתמש לכל היותר ב- $(2L - 1)$ פיצ'רים.

e. מאמנים מסוג 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 $f, g: C \rightarrow \mathbb{R}$ המוגדרות מעל סט קמור C .

סמנו את כל הטענות הנכונות בהכרח.

- הפונקציה $h(z) = f(z) + g(z)$ הינה קמורה.
- הפונקציה $h(z) = \max\{f(z), g(z)\}$ הינה קמורה.
- הפונקציה $h(z) = \min\{f(z), g(z)\}$ הינה קמורה.
- הפונקציה $h(z) = f(g(z))$ הינה קמורה.
- 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):

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

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

- הבעיה P_{hinge} צפויה להיות יותר רגישה ל-outliers מאשר הבעיה P_{ramp} .
- הבעיה P_{hinge} קמורה ואילו הבעיה P_{ramp} אינה קמורה.
- עבור הבעיה P_{hinge} , נקודה בה הנגזרת מוגדרת ומתאפסת היא מינימום גלובאלי.
- עבור הבעיה P_{ramp} , נקודה בה הנגזרת מוגדרת ומתאפסת היא מינימום גלובאלי.
- ערך המינימום הגלובאלי של P_{hinge} הוא 0 \Leftrightarrow ערך המינימום הגלובאלי של P_{ramp} הוא 0.

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

A large rectangular box with rounded corners, containing 25 horizontal lines for writing. The box is empty, with the lines spaced evenly apart, providing a designated area for the student to provide additional answers or explanations.

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

A large rectangular box with rounded corners, containing 25 horizontal lines for writing. The box is empty, with the lines spaced evenly apart, providing a designated area for the student to provide additional answers or explanations.

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

A large rectangular box with rounded corners, containing 25 horizontal lines for writing. The lines are evenly spaced and extend across the width of the box. The box is intended for providing a second answer or further explanation.