הטכניון – מכון טכנולוגי לישראל

הפקולטה למדעי המחשב

מבוא למערכות לומדות (236756)

סמסטר אביב תשפ"ב – 23 בספטמבר 2022

מרצה: ד"ר ניר רוזנפלד

<u>מבחן מסכם מועד ב'</u>

הנחיות הבחינה:

- **משך הבחינה:** 3 שעות.
- **חומר עזר:** המבחן בחומר סגור (ללא ספרים, מחברות, דפי נוסחאות).
- מחשבון: מותר.
- כלי כתיבה: עט <u>בלבד</u>.
- 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 <u>בבירור</u>. 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).

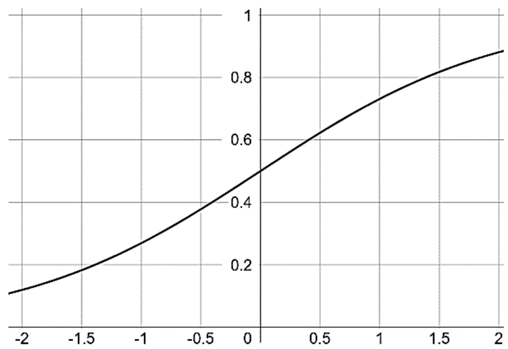
נבנה רשת MLP עם שתי שכבות ליניאריות בתור פונקציה $F: \mathbb{R}^2 \to (0,1)$ המוגדרת:

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

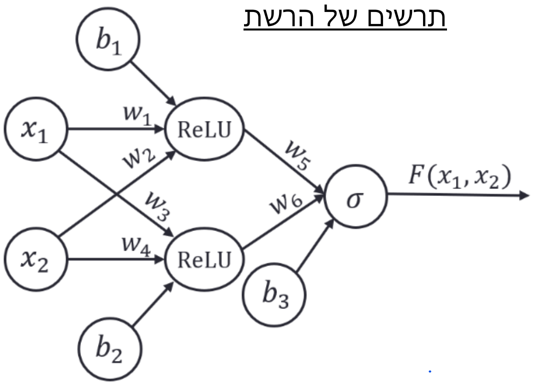
כאשר $w_1, \ldots, w_6, b_1, b_2, b_3 \in \mathbb{R}$ הם פרמטרים סקלריים, פונקציית האקטיבציה היא $\mathrm{ReLU}(z) = \begin{cases} 0, & z \le 0 \\ z, & z > 0 \end{cases}$

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

<u>פונקציית הסיגמואיד</u>



<u>תרשים של הרשת</u>



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 $\frac{\partial \ell}{\partial F}$.

$\ln(1-x) = \frac{-1}{}$

תשובה סופית (לרשותכם טיוטה בסוף הגיליון):

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

ב. [2 נק'] כִּתְבוּ פונקציה שמהווה <u>subgradient</u> 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:

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

לשימושכם בהמשך, להלן כמה נגזרות חלקיות מהשכבה הראשונה:

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

ומהשכבה השנייה:

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

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:

$$\frac{\partial \ell(x,y)}{\partial a_3}=\frac{\partial \ell}{\partial F}\cdot\frac{\partial F}{\partial a_3}=\left(-\frac{y}{\sigma(a_3)}+\frac{(1-y)}{1-\sigma(a_3)}\right)\cdot\sigma(a_3)\big(1-\sigma(a_3)\big)=$$

$$-y(1-\sigma(a_3))+(1-y)\sigma(a_3)=$$

$$=-y+\cancel{y\sigma(a_3)}+\sigma(a_3)-\cancel{y\sigma(a_3)}=\sigma(a_3)-y$$

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.

$F(x_1,x_2)=\sigma(-1)=\sigma(b_3)$

$F(x_1,x_2)=\sigma$

<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	$0-0=0$
w_2	$0-0=0$
w_3	$0-0=0$
w_4	$0-0=0$
b_1	$-1-\eta\cdot 0=-1$
b_2	$-1-\eta\cdot 0=-1$

<u>Second layer</u>

Parameter	Value
w_5	$0-0=0$
w_6	$0-0=0$
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):

$w_1, \dots, w_6$ יישארו קבועים (0)

כנ"ל לגבי b_1, b_2

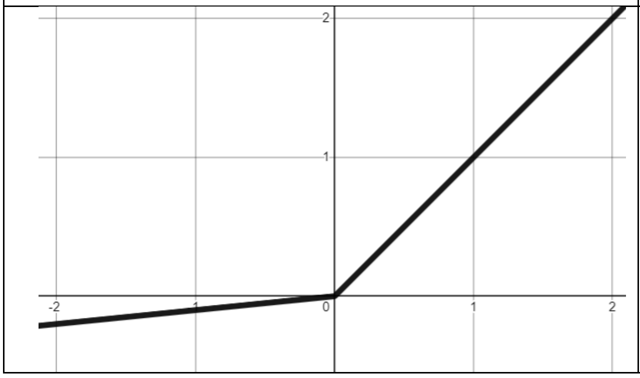
b_3 יישתנה ויהיה $(-1 - (\sigma(-1) - y)T)$

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

סמנו את <u>כָּל</u> האפשרויות המתאימות.

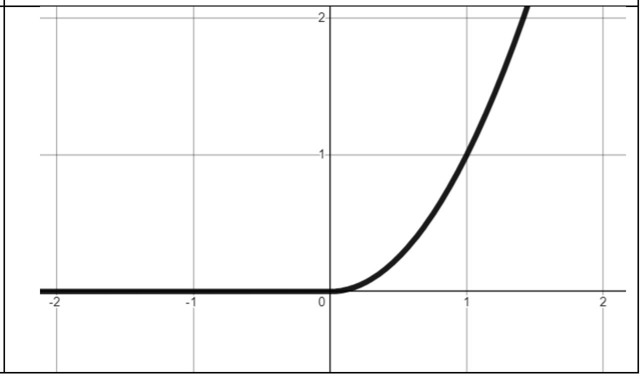
(A)

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



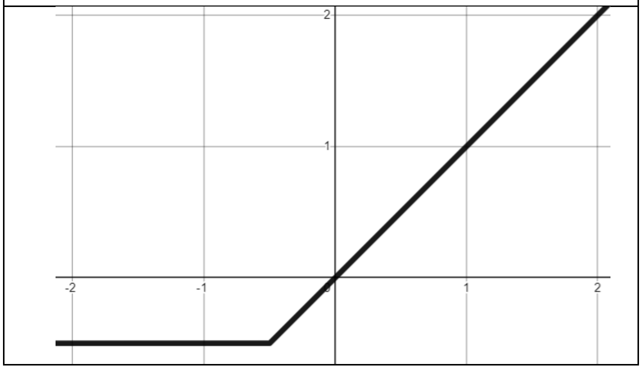
~~(B)~~

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



~~(C)~~

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

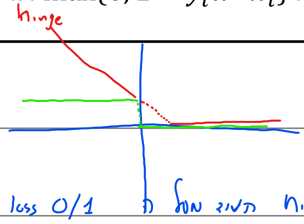


שאלה 2: מסווגים ליניאריים [20 נק']

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, ניתן לראות שעבור דוגמה לפחות, ה-hinge loss תמיד מעל ה loss 0/1

אם סיווגנו בטעות - ה-0-1 תביא לנו את הפסד 1 וה hinge תחזיר לנו את הסרק $1+\boxed{}$ (בכמה טעינו/התרחקנו)

אם צדקנו אך לא מספיק (בין 0 ל 1) אזי ה hinge עדיין מעניש ומתנו אך ה 0-1 תתאפס,

ואם צדקנו מספיק אזי שניהם אותו 0

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:

הדאטה פריד ליניארית הומוגנית אזי בהכרח קיים פתרון w כלשהו

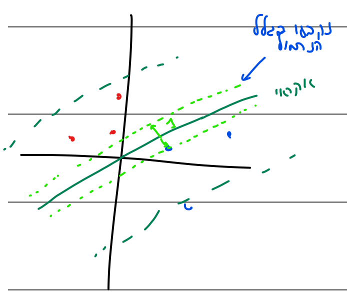
שמפריד אותו, כלומר לכל $1 \le i \le m$ מתקיים כי $y_i w^T x_i > 0$ (וקטורים ריקים וכו)

נשים לב כי הנק' הכי קרובה ל w היא $\min_{i\in[m]}\{|y_i w^T x_i|\}$ ונסמנה ב C

נגדיר את w^* להיות $w^* = \frac{w}{C}$ (כעת ה margin מנורמל לפי הנק' הכי קרובה (והיא במרחק 1 מהמפריד)

$$L(w^*) = L\left(\frac{w}{C}\right) = \operatorname{argmin} \frac{1}{m}\sum \max\{0, 1 - \underbrace{y_i \frac{w^T}{C} \cdot x_i}_{\ge 1}\} = 0$$

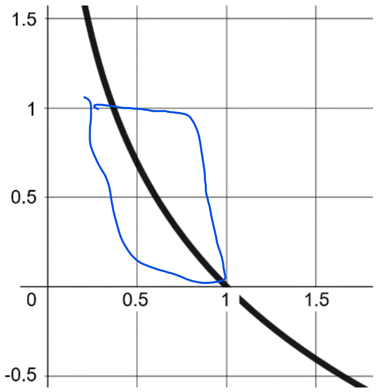
לכן קיים פתרון אופטימלי עם נורמה סופית



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-0-1 loss? נמקו בקצרה.

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



תשובה והסבר קצר:

ניקח נק' שמסווגת לא נכון כך ש $e^{-y_i w^T x_i} = e^{1/4}$

$-\ln\left(\frac{1}{1+e^{1/4}}\right) = -\ln(0.437) = 0.82$

כלומר נק' שמסווגת לא נכון נותנת 0.82 הפסד

ב 0-1 היה נותנת 1 ולכן הפסד המוגדר אינו

חוסמת מלמעלה את ה- 0-1 loss

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:

נזכיר בהרצאה: $-\ln\frac{1}{1+e^{()}} \to -\ln\left(\frac{1}{1+0}\right) \to 0$

נקודה: $-\ln\left(\frac{1}{1+e^{0}}\right) \to -\ln\frac{1}{2} \to 0.693$

קיים פתרון w כלשהו כך שלכל i: $y_i w^T x_i > 0$, יהי $f \to \infty$

יהי $w^* = f w$ עבור כל i: $-\ln\left(\frac{1}{1+e^{-y_i f w^T x_i}}\right) \xrightarrow{f\to\infty} 0$

כלומר ככל שנבחר f גדול יותר כך הפתרון יהיה אופטימלי יותר

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

(ii) Kernel SVM <u>Dual</u> 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 <u>support vectors</u>

(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>: מספיק הכפלת α וה-y-ים ושמירת הווקטור ונכפיל אותו רק ב-x שקיבלנו
את הכל

(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>: צריך לשמור את ה-x_i-ים ($\log m$ כאלה) ו~~צ~~
ולכל אחד, ~~צ~~ נדרש d מקום

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 $\sigma^2\to\infty$.

<u>ניתן להניח</u>:

- 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 $\lim_{\sigma^2\to\infty}h_{\alpha'}(\boldsymbol{x})=\lim_{\sigma^2\to\infty}\operatorname{sign}(\sum_{i=1}^m\alpha_i'y_iK(\boldsymbol{x}_i,\boldsymbol{x}))$.

<u>רמז</u>: כאן הגבול מקיים $\lim_{\sigma^2\to\infty}\operatorname{sign}(\sum_{i=1}^m\alpha_i'y_iK(\boldsymbol{x}_i,\boldsymbol{x}))=\operatorname{sign}\left(\lim_{\sigma^2\to\infty}(\sum_{i=1}^m\alpha_i'y_iK(\boldsymbol{x}_i,\boldsymbol{x}))\right)$.

תשובה:

$= \operatorname{sign}\left(\lim_{\sigma\to\infty}\left(\sum_{i=1}^m \alpha_i' y_i e^0\right)\right)$

$= \operatorname{sign}\left(\sum_{i=1}^m \alpha_i' y_i\right) = \mathbb{I}\left(\sum_{y_i=+1}\alpha_i' > \sum_{y_i=-1}\alpha_i'\right)$

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 ($\pm1$) 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.

10^6

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

השפעה נמוכה למרחק

השפעה גבוהה לקירבה ונוון ממש לחיבור,

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

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

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

<u>הערות</u>: אנו עוסקים במקרה הסביר ולא במקרי קצה. מדובר בניתוח <u>אנליטי</u>, 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 $K: (\mathbb{R}^d \times \mathbb{R}^d) \to \mathbb{R}$ בתור $K(\boldsymbol{u}, \boldsymbol{v}) = \frac{1}{2}(\|\boldsymbol{u}-\boldsymbol{w}\|^2 + \|\boldsymbol{v}-\boldsymbol{w}\|^2 - \|\boldsymbol{u}-\boldsymbol{v}\|^2)$.

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):

$\cancel{u^2} - 2uw + w^2$ $\quad \cancel{+v^2} - 2vw + w^2$ $\quad 2w^2 - 2vw - 2uw + 2uv$

$\cancel{+u^2} \cancel{-u^2} + 2uv \cancel{-v^2}$

$K(u,v) = \frac{1}{2}\left((u-w)^2 + (v-w)^2 - (u-v)^2\right) =$

$= w^2 + uv - wv - wu =$

$= u(v-w) + w(w-v) =$

$= u(v-w) - w(v-w) =$

$= (u-w)(v-w)$

$\varphi(u) = (u-w)$

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

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

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

d. כל הטענות הקודמות שגויות.

ב. [6 נק'] סמנו את <u>כֹּל</u> הטענות הנכונות ביחס ל-**Feature selection**.

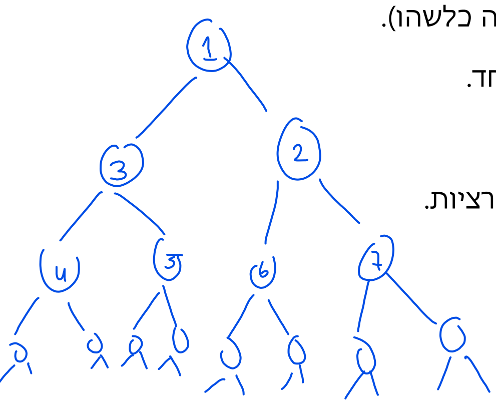
a. שיטות Wrapper (למשל Sequential feature selection) יש להפעיל <u>לפני</u> שלב ה-data imputation.

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

c. <u>בבעיות סיווג</u>: 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 מְסַוֵּג לשתי אפשרויות בעזרת threshold על פיצ'ר אחד. אזי, העץ כולו משתמש לכל היותר ב-$(2L-1)$ פיצ'רים.

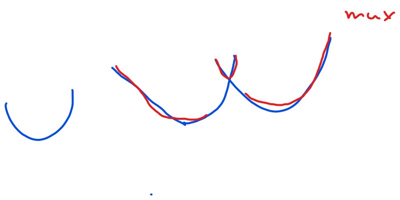
e. מאמנים מסווג AdaBoost עם Decision stump כמסווג בסיס במשך T איטרציות. אזי, המסווג ה"חזק" שמתקבל משתמש לכל היותר ב-T פיצ'רים.



10−1=9

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

סמנו את <u>כל</u> הטענות הנכונות בהכרח.



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

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

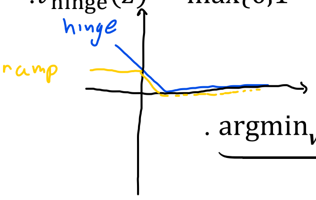
c. הפונקציה $h(z) = \min\{f(z), g(z)\}$ הינה קמורה.

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

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

בעצם הרמפ יותר מוגבל

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



מגדירים שתי בעיות סיווג ליניארי (עם דאטה זהה):

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

סמנו את <u>כל</u> הטענות הנכונות (השאלה עוסקת במקרה הסביר ולא במקרי קצה).

a. הבעיה $\mathrm{P}_{\text{hinge}}$ צפויה להיות יותר רגישה ל-outliers מאשר הבעיה P_{ramp}.

b. הבעיה $\mathrm{P}_{\text{hinge}}$ קמורה ואילו הבעיה P_{ramp} אינה קמורה.

c. עבור הבעיה $\mathrm{P}_{\text{hinge}}$, נקודה בה הנגזרת מוגדרת <u>ומתאפסת</u> היא מינימום גלובאלי.

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

e. <u>ערך</u> המינימום הגלובאלי של $\mathrm{P}_{\text{hinge}}$ הוא 0 $\Leftrightarrow$ <u>ערך</u> המינימום הגלובאלי של P_{ramp} הוא 0.

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):