הטכניון – מכון טכנולוגי לישראל

הפקולטה למדעי המחשב

מבוא למערכות לומדות (236756)

סמסטר חורף תשפ"ב – 07 במרץ 2022

מרצה: ד"ר יונתן בלינקוב

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).
- אין צורך במחשבון.
- מותר לכתוב בעט **בלבד**.
- מותר לענות בעברית או באנגלית.
- יש לכתוב את התשובות **על גבי שאלון זה** בכתב יד קריא. תשובה בכתב יד לא קריא – לא תיבדק.
- במבחן 16 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']:** 4 שאלות פתוחות.
- **חלק ב' [24 נק']:** 4 שאלות סגורות (אמריקאיות) [כל אחת 6 נק'].

בהצלחה!

חלק א' – שאלות פתוחות [76 נק']

שאלה 1 – Linear regression & Optimization [14 נק']

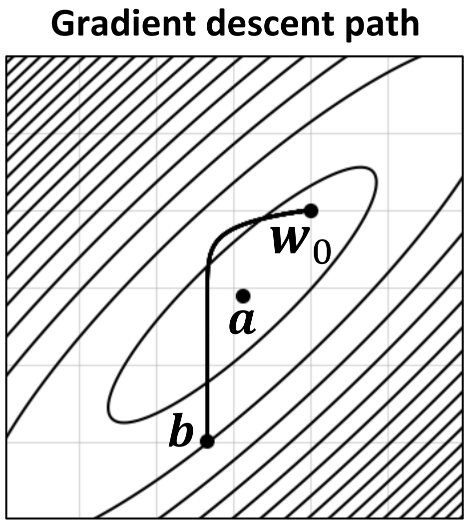
נתונה בעיית רגרסיה ליניארית דו-ממדית (בעיה בשני פרמטרים): $\text{argmin}_{\boldsymbol{w}\in\mathbb{R}^2}\frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{w}^\top\boldsymbol{x}_i-y_i)^2$.

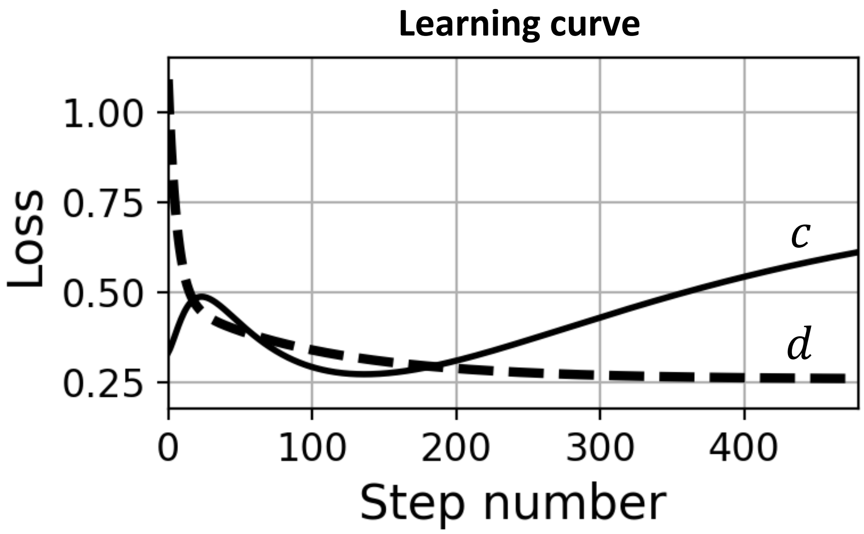
אוספים דאטה S ומחלקים אותו לסט אימון ולסט ואלידציה.

מתחילים מווקטור $\boldsymbol{w}_0=\underline{\boldsymbol{0}}$ ופותרים את הבעיה (עבור סט האימון) בעזרת gradient descent (לא SGD) עם גודל צעד η.

<u>בתרשים השמאלי</u>: המסלול המלא שנוצר מאימון עם GD החל מ-$\boldsymbol{w}_0$ (המסלול מתואר ע"י עקומה במרחב $\mathbb{R}^2$, ומראה את כל הפתרונות $\boldsymbol{w}_0=\underline{\boldsymbol{0}},\boldsymbol{w}_1,\boldsymbol{w}_2,\dots,\boldsymbol{w}_{480}=\boldsymbol{b}$). קווי המתאר (ה-level sets) מתארים loss landscape שהמינימום שלו הוא בנקודה a. עליכם להבין האם מדובר ב-loss על ה-training או ה-validation.

<u>בתרשים הימני</u>: מופיע גרף ההתכנסות המראה את ה-training loss ואת ה-validation loss.





א. [4 נק'] התאימו בין הפתרונות $\boldsymbol{a},\boldsymbol{b}$ שבתרשים השמאלי לבין הפיתרון האופטימלי על סט האימון $\boldsymbol{w}^*_{\text{train}}$ והפיתרון האופטימלי על סט הואלידציה $\boldsymbol{w}^*_{\text{val}}$. התאימו בין העקומות c,d לבין ה-training loss וה-validation loss. מלאו את המקומות הריקים באותיות a,b,c,d כנדרש.

$\boldsymbol{w}^*_{\text{train}}$ matches $\underbrace{\quad b \quad}_{\text{השלימו}}$, $\boldsymbol{w}^*_{\text{val}}$ matches $\underbrace{\quad a \quad}_{\text{השלימו}}$,

training loss matches $\underbrace{\quad d \quad}_{\text{השלימו}}$, validation loss matches $\underbrace{\quad c \quad}_{\text{השלימו}}$.

שימו לב כיצד שני ה-losses (ובפרט ה-training loss) לא יורדים מתחת 0.25.

ב. [5 נק'] אילו מהפתרונות הבאים עשויים לשפר את ה-loss ה-training בסוף האימון (ביחס לתוצאות המוצגות לעיל)?

סמנו את כָּל התשובות המתאימות.

a. הוספת רגולריזציית ℓ^2.

b. שימוש במדיניות early stopping (עצירת ה-GD לפני התכנסות, לפי קריטריון כלשהו).

c. אימון עם SGD (עם batch_size=1) במקום GD, במשך מספר צעדים זהה (480).

d. מיפוי של שני ה-*features* המקוריים ל-*feature mapping* פולינומיאלי.

e. סיבוב מערכת הצירים של ה-features המקוריים (ב-dataset S כולו) ב-$45°$ סביב ראשית הצירים ($\boldsymbol{w}_0 = \underline{\mathbf{0}}$).

ג. [5 נק'] אילו מהפתרונות הבאים עשויים לשפר את ה-loss ה-validation בסוף האימון (ביחס לתוצאות המוצגות לעיל)?

סמנו את כָּל התשובות המתאימות.

a. הוספת רגולריזציית ℓ^2.

b. שימוש במדיניות *early stopping* (עצירת ה-*GD* לפני התכנסות, לפי קריטריון כלשהו).

c. אימון עם *SGD* (עם *batch_size=1*) במקום *GD*, במשך מספר צעדים זהה (480).

d. מיפוי של שני ה-*features* המקוריים ל-*feature mapping* פולינומיאלי.

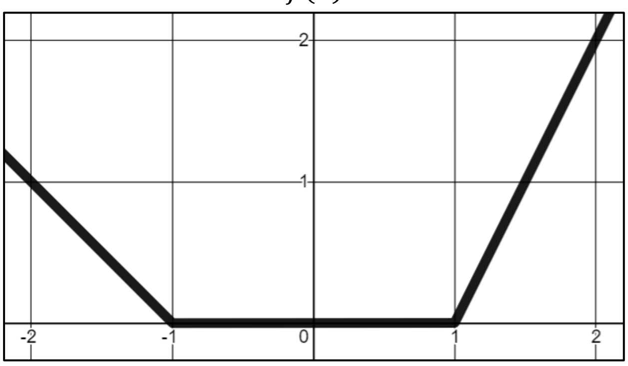
e. סיבוב מערכת הצירים של ה-features המקוריים (ב-dataset S כולו) ב-$45°$ סביב ראשית הצירים ($\boldsymbol{w}_0 = \underline{\mathbf{0}}$).

הערות בדיקה: בסעיפים ב'-ג', ירדה נקודה על כל טענה מיותרת שסומנה או טענה נכונה שחסרה.

שאלה 2 – Deep learning [20 נק']

נתונות שתי הפונקציות הרציפות $f, g: \mathbb{R} \to \mathbb{R}$ שבתרשימים הבאים:

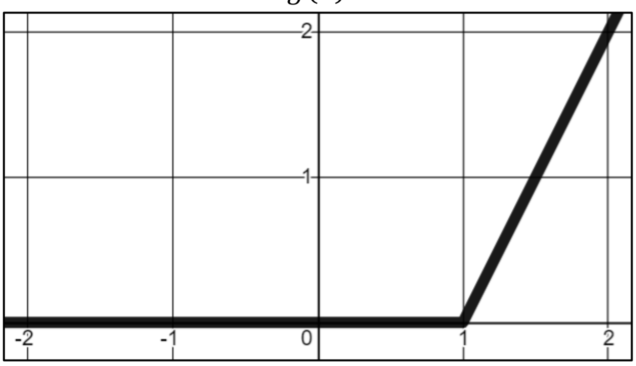
$f(x)$



בתחום $[-1, 1]$ הפונקציה היא אפס.

בתחום $(1, \infty)$ השיפוע הוא $\mathbf{2}$ ובתחום $(-\infty, -1)$ הוא $\mathbf{-1}$.

$g(x)$

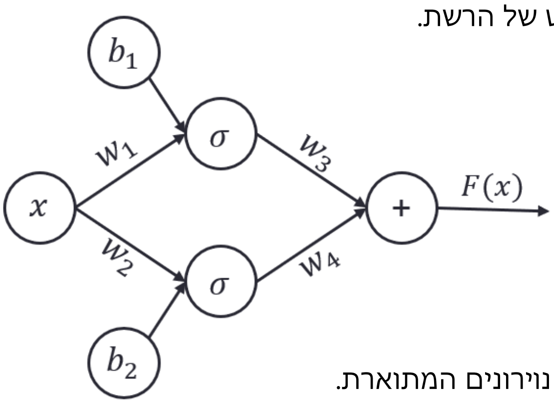


בתחום $(-\infty, 1]$ הפונקציה היא אפס.

בתחום $(1, \infty)$ השיפוע של $\boldsymbol{g}$ הוא $\mathbf{2}$.

נרצה ללמוד את הפונקציות f, g בעזרת רשת הנוירונים הבאה: $F(x) = w_3 \cdot \sigma(w_1 \cdot x + b_1) + w_4 \cdot \sigma(w_2 \cdot x + b_2)$.

כאשר $w_1, w_2, w_3, w_4, b_1, b_2 \in \mathbb{R}$ הם פרמטרים סקלריים ו-$x \in \mathbb{R}$ הוא הקלט של הרשת.



פונקציית **האקטיבציה** σ יכולה להיות אחת מהשתיים:

1. סיגמואיד, משמע $\sigma(z) = \frac{1}{1+e^{-z}}$.
2. ReLU, משמע $\sigma(z) = \max\{0, z\}$.

בשני הסעיפים הבאים נראה שניתן לממש את הפונקציות f, g בעזרת רשת הנוירונים המתוארת.

א. [4 נק'] נבחר $w_2 = b_2 = w_4 = 0$. כתבו את ערכי w_1, w_3, b_1 ואת הבחירה של σ שמקיימים: $\forall x \in \mathbb{R}: \ F(x) = g(x)$.

תשובה סופית (לרשותכם דפי טיוטה בסוף הגיליון):

First layer: $w_1 = \underbrace{\ 1\ }_{\text{השלימו}}$; $b_1 = \underbrace{\ -1\ }_{\text{השלימו}}$; **Second layer:** $w_3 = \underbrace{\ 2\ }_{\text{השלימו}}$.

Activation: Sigmoid or **ReLU** (circle your choice).

ב. [5 נק'] כתבו את ערכי $w_1, w_2, w_3, w_4, b_1, b_2$ ואת הבחירה של σ שמקיימים: $\forall x \in \mathbb{R}: \ F(x) = f(x)$.

תשובה סופית:

First layer: $w_1 = \underbrace{\ 1\ }_{\text{השלימו}}$; $w_2 = \underbrace{\ -1\ }_{\text{השלימו}}$; $b_1 = \underbrace{\ -1\ }_{\text{השלימו}}$; $b_2 = \underbrace{\ -1\ }_{\text{השלימו}}$;

Second layer: $w_3 = \underbrace{\ 2\ }_{\text{השלימו}}$; $w_4 = \underbrace{\ 1\ }_{\text{השלימו}}$.

Activation: Sigmoid or **ReLU** (circle your choice).

הערה: סעיף ג' לא תלוי בסעיפים הקודמים.

ג. [4 נק'] הציעו פונקציית loss שמתאימה לבעיית הרגרסיה שהוגדרה, כך שמזעור של $\sum_{i=1}^{m}\ell\big(f(x_i),\ F(x_i)\big)$ יביא ללמידת פרמטרים מתאימים שיקיימו $F(x)=f(x)$.

Answer: $\underbrace{\boldsymbol{\ell}(a,b)}_{\ell:\mathbb{R}\times\mathbb{R}\to\mathbb{R}_{\geq 0}} = \underbrace{\qquad (a-b)^2 \qquad}_{\text{השלימו}}$.

הערה: סעיף ד' תלוי בסעיף ב' רק דרך הבחירה של פונקציית האקטיבציה.

בסעיף הבא נבחן את הקמירות של הבעיה שנוצרה.

תזכורת: הפונקציה $g:\mathbb{R}\to\mathbb{R}$ נקראת פונקציה <u>קמורה</u> אם מתקיים

$$\forall z_1,z_2\in\mathbb{R},\ \ \forall t\in[0,1]:\quad tg(z_1)+(1-t)g(z_2)\geq g(t\,z_1+(1-t)z_2)$$

ד. [7 נק'] בסעיף זה נניח שוב $w_2=b_2=w_4=0$.

הוכיחו / הפריכו: הפונקציה $\ell\big(a,\ F(x)\big)$ קמורה ביחס לפרמטר w_1 (בהינתן כל בחירה של b_1,w_3,a,x).

ניתן לענות לפי הגדרת הקמירות או לפי מאפיינים שלמדנו (אך יש לציין אותם במפורש).

<u>שימו לב</u>: עליכם להשתמש בבחירה של σ מסעיף ב' ובבחירה של ℓ מסעיף ג'. אין להציב ערכים מסעיף א'.

תשובה (לרשותכם דפי טיוטה בסוף הגיליון):

נכתוב את הפונקציה במפורש: $\ell\big(a,\ F(x)\big)=(a-w_3\max(0,w_1x+b_1))^2$

הפונקציה **אינה** קמורה ב-w_1.

ניתן לראות זאת ע"י הפירוק (עוזר לאינטואיציה):

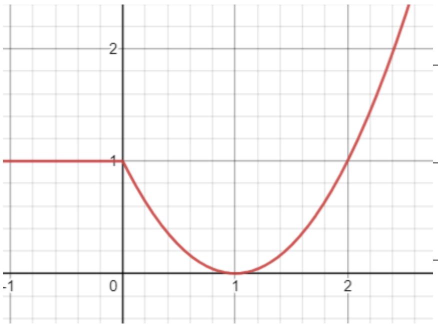
$$\ell\big(a,F(x)\big)=\underbrace{a^2}_{\text{קבועה}}-aw_3\underbrace{\max(0,w_1x+b_1)}_{\text{קמורה}}+\underbrace{w_3^2(\max(0,w_1x+b_1))^2}_{\text{קמורה}}$$

וברור שהפונקציה $-aw_3\max(0,w_1x+b_1)$ לא תמיד קמורה (בכפוף לסימון של $-aw_3$).

למשל, כאשר $b_1=0$ וגם $w_3=a=x=1$, מקבלים:

$$(1-\max(0,w_1))^2=\begin{cases}1, & w_1<0\\ (1-w_1)^2, & w_1\geq 0\end{cases}$$

שאינה פונקציה קמורה (ראו בתרשים).



שאלה 3 – Naïve Bayes [18 נק']

בשאלה זו נראה ש-Gaussian Naïve Bayes הינו מסווג לינארי.

נתון דאטה $S=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^m$ עם דוגמאות $\boldsymbol{x}_i\in\mathbb{R}^d$ ותיוגים $y_i\in\{0,1\}$.

לפי מידול Gaussian Naïve Bayes, נניח שההתפלגות המותנה של כל כניסה $k=1,\dots,d$ הינה:

$$(\mathrm{X}[k]\mid Y=y)\sim\mathcal{N}(\mu_y[k],\sigma[k]^2),$$

כאשר $\mathrm{X}[k]$ הוא המשתנה המקרי המתאים לכניסה ה-k ב-x, המסומנת $x[k]$.

תזכורת: פונקציית הצפיפות של התפלגות גאוסיאנית $\mathcal{N}(\mu,\sigma^2)$ נתונה ע"י: $P(z)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right)$.

א. [4 נק'] הוכיחו שמתקיים: $P(X=x\mid Y=1)=\left(\prod_{k=1}^d\frac{1}{\sigma[k]\sqrt{2\pi}}\right)\exp\left(-\sum_{k=1}^d\frac{1}{2\sigma[k]^2}(x[k]-\mu_1[k])^2\right)$.

הוכחה מנומקת:

לפי הנחת הנאיביות של NB: $P(X=x\mid Y=1)=\prod_k P(X[k]=x[k]\mid Y=1)$

לפי הנחת המודל הגאוסיאני:

$$\prod_k P(X[k]=x[k]\mid Y=1)=\prod_k\frac{1}{\sigma[k]\sqrt{2\pi}}\exp\left(\frac{1}{2\sigma[k]^2}(x[k]-\mu_1[k])^2\right)$$

משתמשים בכללים אלגברים פשוטים ומקבלים את מה שצריך להוכיח.

טענה (ללא הוכחה): בעזרת חוק בייס ונוסחת ההסתברות השלמה, ניתן להראות:

$$P(Y=1 \mid X=x)=\frac{1}{1+\frac{P(Y=0)P(X=x \mid Y=0)}{P(Y=1)P(X=x \mid Y=1)}}$$

טענה (ללא הוכחה): נסמן $p \triangleq P(Y=1)$. תחת הנחות השאלה ניתן להראות:

$$\frac{P(Y=0)P(X=x \mid Y=0)}{P(Y=1)P(X=x \mid Y=1)}=\frac{1-p}{p}\cdot\exp\left(\sum_{k=1}^{d}\left(\frac{\mu_0[k]-\mu_1[k]}{\sigma[k]^2}x[k]+\frac{\mu_1[k]^2-\mu_0[k]^2}{2\sigma[k]^2}\right)\right)$$

ב. [7 נק'] בעזרת האמור לעיל, הוכיחו שמתקיים: $P(Y=1 \mid X=x)=\frac{1}{1+\exp(\boldsymbol{w}^\top\boldsymbol{x}+b)}$.

בסוף ההוכחה, ציינו במפורש את ערכי $\boldsymbol{w}\in\mathbb{R}^d, \boldsymbol{b}\in\mathbb{R}$ המקיימים זאת.

הוכחה: לפי הסעיפים הקודמים והטענות:

$$P(Y=1 \mid X=x)=1+\frac{1-p}{p}\cdot\exp\left(\sum_{k=1}^{d}\left(\frac{\mu_0[k]-\mu_1[k]}{\sigma[k]^2}x[k]+\frac{\mu_1[k]^2-\mu_0[k]^2}{2\sigma[k]^2}\right)\right)$$

$$=\frac{1}{1+\exp\left(\ln\frac{1-p}{p}+\sum_{k=1}^{d}\frac{\mu_1[k]^2-\mu_0[k]^2}{2\sigma[k]^2}+\sum_{k=1}^{d}\left(\frac{\mu_0[k]-\mu_1[k]}{\sigma[k]^2}x[k]\right)\right)}$$

וקובעים $w[k]=\frac{\mu_0[k]-\mu_1[k]}{\sigma[k]^2}, \quad b=\ln\frac{1-p}{p}+\sum_{k=1}^{d}\frac{\mu_1[k]^2-\mu_0[k]^2}{2\sigma[k]^2}$

ג. [7 נק'] בהסתמך על האמור לעיל, הוכיחו כי מתקבל כלל החלטה ליניארי.

משמע, ההיפותזה $h(x)=\mathrm{argmax}_{y\in\{0,1\}}P(Y=y \mid X=x)$ מַשְׁרָה גבול החלטה ליניארי (decision boundary).

הוכחה תמציתית:

מתקיים $h(x)=1 \Leftrightarrow P(Y=1 \mid X=x)>P(Y=0 \mid X=x)$.

ע"י סדרה של פעולות הפיכות מתקבל: $P(Y=1 \mid X=x)>P(Y=0 \mid X=x)$

$$\frac{1}{1+\exp(\boldsymbol{w}^\top\boldsymbol{x}+b)}>1-\frac{1}{1+\exp(\boldsymbol{w}^\top\boldsymbol{x}+b)}$$

$$1>1+\exp(\boldsymbol{w}^\top\boldsymbol{x}+b)-1$$

$$1>\exp(\boldsymbol{w}^\top\boldsymbol{x}+b)\Leftrightarrow\boldsymbol{w}^\top\boldsymbol{x}+b<0$$

שזה כלל החלטה ליניארי.

שאלה 4 – SVM [24 נק']

היזכרו בבעיות ה-SVM במקרה ההומוגני (נניח שמתקיים $\lambda = 1$ בבעיה ה-Soft):

Soft SVM

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\operatorname{argmin}} \left(\frac{1}{m}\sum_{i=1}^{m}\max\{0, 1 - y_i \cdot \boldsymbol{w}^\top \boldsymbol{x}_i\} + \|\boldsymbol{w}\|_2^2\right)$$

Hard SVM

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\operatorname{argmin}} \|\boldsymbol{w}\|_2^2 \quad \text{s.t. } y_i \cdot \boldsymbol{w}^\top \boldsymbol{x}_i \ge 1,\ \forall i \in [m]$$

נתון דאטה d-ממדי $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ עם סיווגים בינאריים (± 1). ידוע שהדאטה פריד ליניארית ע"י מפריד הומוגני.
צוות מחקר פתר את שתי בעיות האופטימיזציה שלמעלה, וקיבל את $\boldsymbol{w}_{\text{hard}}, \boldsymbol{w}_{\text{soft}} \in \mathbb{R}^d$ כפתרונות.

א. [8 נק'] האם ניתן לומר שאחד מהמקרים $\|\boldsymbol{w}_{\text{hard}}\|_2 \le \|\boldsymbol{w}_{\text{soft}}\|_2$ או $\|\boldsymbol{w}_{\text{hard}}\|_2 \ge \|\boldsymbol{w}_{\text{soft}}\|_2$ מתקיים בהכרח?
אם כן, איזה מהם? בכל מקרה – הסבירו בקצרה.

תשובה חלקית: בהכרח מתקיים $\|\boldsymbol{w}_{\text{hard}}\|_2 \ge \|\boldsymbol{w}_{\text{soft}}\|_2$.

ניתן לראות זאת בכמה דרכים. למשל, למדנו שכאשר $\lambda \to 0$, מתקיים $\boldsymbol{w}_{\text{hard}} \to \boldsymbol{w}_{\text{soft}}$.

לכן, כאשר משתמשים ב-$\lambda = 1$, הנורמה מקבלת **יותר** חשיבות מאשר ב-Hard, והבעיה תחזיר

פתרונות בנורמה נמוכה יותר.

ב. [8 נק'] נוסף עוד feature ממקור לא ידוע. הצוות פתר את בעיית ה-Hard-SVM המעודכנת וקיבל את $\boldsymbol{w}'_{\text{hard}} \in \mathbb{R}^{d+1}$.
האם ניתן לומר שאחד מהמקרים $\|\boldsymbol{w}_{\text{hard}}\|_2 \le \|\boldsymbol{w}'_{\text{hard}}\|_2$ או $\|\boldsymbol{w}_{\text{hard}}\|_2 \ge \|\boldsymbol{w}'_{\text{hard}}\|_2$ מתקיים בהכרח?
אם כן, איזה מהם? בכל מקרה – הסבירו בקצרה.

תשובה חלקית: בהכרח מתקיים $\|\boldsymbol{w}_{\text{hard}}\|_2 \ge \|\boldsymbol{w}'_{\text{hard}}\|_2$.

נניח בשלילה $\|\boldsymbol{w}_{\text{hard}}\|_2 < \|\boldsymbol{w}'_{\text{hard}}\|_2$. ניקח את $\boldsymbol{w}_{\text{hard}} \in \mathbb{R}^d$ וניצור $\boldsymbol{w}' = \begin{bmatrix}\boldsymbol{w}_{\text{hard}} \\ 0\end{bmatrix} \in \mathbb{R}^{d+1}$.

$$\forall i \in [m]:\ \ y_i \cdot \boldsymbol{w}_{\text{hard}}{}^\top \boldsymbol{x}_i \ge 1 \Longrightarrow y_i \cdot (\boldsymbol{w}')^\top \boldsymbol{x}_i \ge 1$$

ולכן $\boldsymbol{w}'$ הוא פיתרון חוקי לבעיה המעודכנת (נותן תמיד אפס לפיצ'ר החדש).

אבל, מתקיים $\|\boldsymbol{w}'_{\text{hard}}\|_2 < \|\boldsymbol{w}'\|_2 = \|\boldsymbol{w}_{\text{hard}}\|_2$, בסתירה לכך ש-$\boldsymbol{w}'_{\text{hard}}$ פיתרון אופטימלי

לבעיה המעודכנת.

ג. [8 נק'] יהי $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m}$ סט אימון כלשהו (לאו דווקא פריד ליניארית) עבורו $\boldsymbol{x}_i \in \mathbb{R}^d, y_i \in \{\pm 1\}$ ויהי $\boldsymbol{w}_{\text{soft}} \in \mathbb{R}^d \setminus \{\underline{\boldsymbol{0}}\}$ הפתרון של Soft-SVM על סט זה.

עליכם להוכיח **<u>אחת</u>** מבין שתי הטענות הבאות (השנייה מזכה בניקוד חלקי <u>בלבד</u>):

(i) [8 מתוך 8 נק'] לכל סט כזה ניתן להוסיף feature חדש (בלי לשנות את ה-features האחרים ואת התיוגים הנתונים), כך שמתקיים $\|\boldsymbol{w}_{\text{soft}}\|_2 > \|\boldsymbol{w}'_{\text{soft}}\|_2$, כאשר $\boldsymbol{w}'_{\text{soft}} \in \mathbb{R}^{d+1}$ אופטימלי עבור הבעיה המעודכנת (עם ה-feature החדש).

<u>או</u>

(ii) [4 מתוך 8 נק'] לכל סט כזה ניתן להוסיף feature חדש (בלי לשנות את ה-features האחרים ואת התיוגים הנתונים), כך שקיים $\boldsymbol{w}' \in \mathbb{R}^{d+1}$ עבורו $\|\boldsymbol{w}_{\text{soft}}\|_2 > \|\boldsymbol{w}'\|_2$ וגם $\mathcal{L}(\boldsymbol{w}_{\text{soft}}) \geq \mathcal{L}'(\boldsymbol{w}')$, כאשר מגדירים

$$\mathcal{L}(\boldsymbol{w}) \triangleq \frac{1}{m}\sum_{i=1}^{m} \max\{0, 1 - y_i \cdot \boldsymbol{w}^\top \boldsymbol{x}_i\}, \quad \mathcal{L}'(\boldsymbol{w}) \triangleq \frac{1}{m}\sum_{i=1}^{m} \max\{0, 1 - y_i \cdot \boldsymbol{w}^\top \underbrace{\boldsymbol{x}'_i}_{\text{הדוגמאות המעודכנות עם ה–feature החדש}}\}.$$

הוכחה (יש לציין איזו טענה מוכיחים):

נוכיח את טענה (i). נסמן את ה-regularized loss בעזרת $\mathcal{L}_\lambda(\boldsymbol{w}) = \mathcal{L}(\boldsymbol{w}) + \underbrace{\lambda}_{=1} \|\boldsymbol{w}\|_2^2$.

נוסיף פיצ'ר חדש $x_i = \alpha y_i \in \{\pm\alpha\}$ עבור $\alpha > 0$. נגדיר וקטור משקלים $\boldsymbol{w}' = \left[0, \dots 0, \frac{1}{\alpha}\right]$.

מתקיים $\|\boldsymbol{w}'\|_2^2 = \frac{1}{\alpha^2} = \frac{1}{\alpha^2} + 0 = \mathcal{L}'_\lambda(\boldsymbol{w}') = \left(\frac{1}{\alpha}\right)^2 + \frac{1}{m}\sum_{i=1}^{m} \max\left\{0, 1 - \alpha\frac{1}{\alpha}\right\}$.

נבחר $\alpha > \frac{1}{\|\boldsymbol{w}_{\text{soft}}\|_2}$ כלשהו. מקבלים $\|\boldsymbol{w}_{\text{soft}}\|_2^2 > \frac{1}{\alpha^2} = \|\boldsymbol{w}'\|_2^2 = \mathcal{L}'_\lambda(\boldsymbol{w}')$.

מהאופטימליות של $\boldsymbol{w}'_{\text{soft}}$, מקבלים $\mathcal{L}'_\lambda(\boldsymbol{w}') \geq \mathcal{L}'_\lambda(\boldsymbol{w}'_{\text{soft}})$. ומההגדרה נובע $\mathcal{L}'_\lambda(\boldsymbol{w}'_{\text{soft}}) \geq \|\boldsymbol{w}'_{\text{soft}}\|_2^2$.

בסה"כ הראינו שמתקיים $\|\boldsymbol{w}_{\text{soft}}\|_2^2 > \mathcal{L}'_\lambda(\boldsymbol{w}') \geq \mathcal{L}'_\lambda(\boldsymbol{w}'_{\text{soft}}) \geq \|\boldsymbol{w}'_{\text{soft}}\|_2^2$. ■

נוכיח את טענה (ii). נבחר פיצ'ר k שעבורו $\boldsymbol{w}_{\text{soft}}[k] \neq 0$.

נשכפל את הפיצ'ר הזה כמו שהוא ונפצל את המשקל שלו: $\boldsymbol{w}'[k] = \boldsymbol{w}'[d+1] = \frac{1}{2}\boldsymbol{w}_{\text{soft}}[k]$.

קל להראות שמתקיים $\mathcal{L}(\boldsymbol{w}_{\text{soft}}) = \mathcal{L}'(\boldsymbol{w}')$.

מצד שני, הנורמה בהכרח <u>קטנה</u>, כי $\sqrt{\left(\frac{1}{2}a\right)^2 + \left(\frac{1}{2}a\right)^2} = \sqrt{\frac{1}{2}a^2} < \sqrt{a^2}$ לכל $a \neq 0$.

חלק ב' – שאלות רב-ברירה [24 נק']

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

הערות בדיקה: בסעיפים א'-ג', ירדו 2 נקודות על כל טענה מיותרת שסומנה או טענה נכונה שחסרה.

א. [6 נק'] היזכרו בבעיות רגרסיה ליניארית עם Least squares (LS), וסמנו את כֹּל התשובות הנכונות.

שימו לב: בסעיף זה, הגזירוּת מתייחסת להגדרת ה-gradients המסורתית, ולא ל-subgradients.

a. כאשר פותרים LS עם הפיצ'רים המקוריים, הבעיה קמורה ביחס ל-$\boldsymbol{w}$.

b. כאשר פותרים LS עם feature mapping שנקבע מראש (למשל פולינומיאלי), הבעיה קמורה ביחס ל-$\boldsymbol{w}$.

c. Ridge regression (ℓ^2) היא בעיה קמורה אך לא גזירה ביחס ל-$\boldsymbol{w}$.

d. Lasso (ℓ^1) היא בעיה קמורה אך לא גזירה ביחס ל-$\boldsymbol{w}$.

ב. [6 נק'] הטענות הבאות עוסקות במודלים מסוג Linear Soft SVM, Perceptron, and Logistic Regression.

סמנו את כֹּל הטענות הנכונות.

a. Logistic Regression יכול ללמוד גם מפרידים לא ליניאריים בגלל שה-sigmoid מכניס non-linearity.

b. ניתן להכליל Logistic Regression לבעיות multiclass בעזרת פונקציית Softmax.

c. כל עוד הדאטה פריד ליניארית, Soft SVM ופרספטרון מחזירים את אותו המפריד.

d. בשלושת האלגוריתמים ניתן להשתמש ב-feature mapping כדי ללמוד מפרידים לא ליניאריים.

e. פרספטרון לומד בעזרת GD (לא stochastic), ואילו Soft SVM יש ללמוד באמצעות SGD.

ג. [6 נק'] נגדיר את פונקציית ה-squared loss הבאה: $\mathcal{L}(z) = (1-z)^2$.

סמנו את כֹּל הטענות הנכונות ביחס לפונקציה זו.

a. הפונקציה קמורה (*convex*) ביחס ל-z.

b. הנגזרת של הפונקציה היא $\frac{\partial}{\partial z}\mathcal{L} = 2z - 2$.

c. עבור בעיות סיווג ליניארי (ה-loss מחושב ע"י $z = y_i \boldsymbol{w}^\top \boldsymbol{x}_i$ והפרדיקציות ניתנות ע"י $(h(\boldsymbol{x}_i) = sgn(\boldsymbol{w}^\top \boldsymbol{x}_i)$, כאשר ה-training loss הוא 0, גם ה-training error הוא 0.

d. עבור בעיות סיווג ליניארי (ה-loss מחושב ע"י $z = y_i \boldsymbol{w}^\top \boldsymbol{x}_i$ והפרדיקציות ניתנות ע"י $(h(\boldsymbol{x}_i) = \mathrm{sgn}(\boldsymbol{w}^\top \boldsymbol{x}_i)$, כאשר ה-training error הוא 0, גם ה-training loss הוא 0.

ד. [6 נק'] נגדיר מחלקת היפותזות שמכלילה את המסווגים שמחזיר Adaboost לאחר T צעדים עם מחלקת היפותזות $\mathcal{B}$ בתור מסווגי בסיס. משמע:

$$\mathcal{H}_{\mathcal{B},T} = \left\{ h_{\text{strong}}(x) = \text{sign}(\textstyle\sum_{t=1}^{T} \alpha_t h_t(x)) \,\middle|\, \boldsymbol{\alpha} \in \mathbb{R}^T,\ h_t \in \mathcal{B} \right\}$$

לפניכם מספר טענות על ההשפעה של T ו-$\text{VCdim}(\mathcal{B})$ על $\text{VCdim}(\mathcal{H}_{\mathcal{B},T})$.

בחרו בטענה <u>היחידה</u> הנכונה (השאלה אינה עוסקת במקרי קצה אלא במקרה הסביר).

a. T גדל $\Leftarrow VCdim(\mathcal{H}_{\mathcal{B},T})$ גדל. $VCdim(\mathcal{B})$ גדל $\Leftarrow VCdim(\mathcal{H}_{\mathcal{B},T})$ גדל.

b. T קָטֵן $\Leftarrow \text{VCdim}(\mathcal{H}_{\mathcal{B},T})$ גָּדֵל. $\text{VCdim}(\mathcal{B})$ גָּדֵל $\Leftarrow \text{VCdim}(\mathcal{H}_{\mathcal{B},T})$ גָּדֵל.

c. T גָּדֵל $\Leftarrow \text{VCdim}(\mathcal{H}_{\mathcal{B},T})$ גָּדֵל. $\text{VCdim}(\mathcal{B})$ קָטֵן $\Leftarrow \text{VCdim}(\mathcal{H}_{\mathcal{B},T})$ גָּדֵל.

d. T קָטֵן $\Leftarrow \text{VCdim}(\mathcal{H}_{\mathcal{B},T})$ גָּדֵל. $\text{VCdim}(\mathcal{B})$ קָטֵן $\Leftarrow \text{VCdim}(\mathcal{H}_{\mathcal{B},T})$ גָּדֵל.

e. רק $\text{VCdim}(\mathcal{B})$ משפיע על $\text{VCdim}(\mathcal{H}_{\mathcal{B},T})$.

f. רק T משפיע על $\text{VCdim}(\mathcal{H}_{\mathcal{B},T})$.