הטכניון – מכון טכנולוגי לישראל

הפקולטה למדעי המחשב

מבוא למערכות לומדות (236756)

סמסטר אביב תשפ"ג – 16 ביולי 2023

מרצה: ד"ר ניר רוזנפלד

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.
- 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 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.**

בהצלחה!

שאלה 1: Feature Selection [8 נק']

סמנו את התשובות המתאימות (לפי ההוראות). אין צורך לכתוב הסברים. סימונים לא ברורים יובילו לפסילת התשובה.

נתון ה-dataset הבא.

sample/feat.	$x[1]$	$x[2]$	$x[3]$	y
#1	1	-1	1	-1
#2	-1	1	1	-1
#3	1	-1	1	-1
#4	-1	1	1	-1
#5	1	1	-1	$+1$
#6	1	1	1	$+1$

נפעיל sequential feature selection (כפי שהגדרנו בתרגיל הבית) בעזרת אלגוריתם בסיס $\mathcal{A}$ שפותר ERM ומוצא מסווג ליניארי לא-הומוגני עם שגיאת 0-1 מינימלית על הנתונים. כאן – נמצא שני פיצ'רים.

אילו פיצ'רים ייבחרו בתהליך Forward Selection?

a. $x[1], x[2]$.

b. $x[1], x[3]$.

c. $x[2], x[3]$.

d. מהנתונים הקיימים, ניתן לענות רק לגבי פיצ'ר יחיד.

e. מהנתונים הקיימים, לא ניתן לענות כלל.

אילו פיצ'רים ייבחרו בתהליך Backward Selection?

a. $x[1], x[2]$.

b. $x[1], x[3]$.

c. $x[2], x[3]$.

d. מהנתונים הקיימים, ניתן לענות רק לגבי פיצ'ר יחיד.

e. מהנתונים הקיימים, לא ניתן לענות כלל.

שאלה 2: Capacity של מחלקות שונות [21 נק']

נתונה התפלגות $\mathcal{D}$ על דאטה עם $d \geq 10$ פיצ'רים בינאריים ($\mathcal{X} = \{0,1\}^d$) ותיוגים בינאריים ($\mathcal{Y} = \{-1,+1\}$).

ידוע שההתפלגות $\mathcal{D}$ נותנת לכל דוגמה אפשרית $\mathbf{x} \in \mathcal{X}$ הסתברות כלשהי סופית גדולה ממש מאפס.

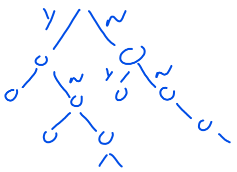
ידוע שהתיוג האמיתי של דוגמה כלשהי הוא חיובי <u>אם ורק אם</u> בדוגמה יש לפחות שני פיצ'רים כלשהם "דולקים" (שערכם 1).

$(1 - P(1) - P(0))$

לדוגמה: כאשר $d = 5$, לדוגמאות $(0,1,1,0,1)$ ו-$(1,1,0,0,0)$ יש בהכרח תיוג <u>חיובי</u>, ולדוגמה $(0,0,1,0,0)$ בהכרח תיוג <u>שלילי</u>.

עבור כל אחת מהמחלקות הבאות, נבדוק האם ניתן ליצור בעזרתה מסווג שיגיע לשגיאת **<u>הכללה</u>** אפס.

<u>שימו לב</u>: זו שאלה על מחלקות היפותזות, ולא על אלגוריתמי למידה ספציפיים.



א. <u>האם קיים עץ החלטה בינארי (צומת שאינו עלה יכול להתפצל רק לשני צמתים) שמגיע לשגיאת הכללה אפס?</u>

סמנו: כן / לא

אם קיים – מהו עומק העץ המינימלי הנדרש (בסדר גודל, למשל $\Theta(1), \Theta(\ln d), \Theta(d), \Theta(2^d)$ וכו')? הסבירו בקצרה.

אחרת – הסבירו באופן מפורט למה לא.

תשובה (לרשותכם דפי טיוטה בסוף הגיליון):

הרי מספיק לבדוק האם יש 2 פיצ'רים דלוקים (לפחות) בשביל להחליט

האם התיוג חיובי, ואז ניתן לבנות עץ כאשר השורש שואל לגבי

הפיצ'ר הראשון (דלוק/כבוי), הצומת ה-1 שואלת לגבי הפיצ'ר השני וכך הלאה

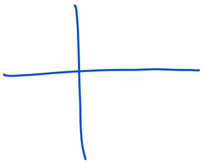
לכל היותר נרחיב עץ בעומק d (במידה וכל הפיצ'רים כבויים/כולם חוץ מאחד)

ולכן $O(d)$

דבר היחיד לשגיאות ההכללה להיות אפס (כי יש לנו ידע מקדים על הדאטא הכולל)

שימו לב: הסעיף הבא עוסק במסווגים ליניאריים ומתקיים בהם $\text{sign}(\mathbf{w}^\top\mathbf{x}) = \begin{cases} -1, & \mathbf{w}^\top\mathbf{x} < 0 \\ 0, & \mathbf{w}^\top\mathbf{x} = 0 \\ +1, & \mathbf{w}^\top\mathbf{x} > 0 \end{cases}$.

משמע, margin של אפס בהכרח גורר שגיאת סיווג.



ב. האם קיים מפריד ליניארי לא הומוגני שמגיע לשגיאת הכללה אפס? **סמנו: כן / לא**

אם קיים – הציעו וקטור $\mathbf{w} \in \mathbb{R}^d$ וסקלר $b \in \mathbb{R}$ שמגיעים לשגיאה אפס (במקרה כזה לא נדרש הסבר נוסף).

אחרת – הסבירו באופן מפורט למה לא.

תשובה:

לא, לדוגמא:

$(0,1,0,\dots,0) = \text{OFF}$

$(1,0,0,\dots,0) = \text{OFF}$

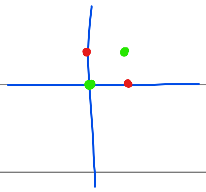
$(1,1,0,\dots,0) = \text{ON}$

$(0,0,1,1,0,\dots,0) = \text{ON}$

נשים לב כי הדגמון הנה כבר אינו פריד לינארית

ב 2 המימדים הראשונים (ובפרט אינו פריד)

ניתן לדמיין זאת כך:



וכו אם יש מימדים מוסתרים

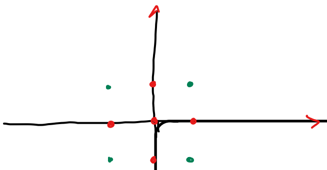
ג. האם ניתן ליצור מסווג 1-Nearest-Neighbor (עם מטריקת ℓ_2 האוקלידית) שמגיע לשגיאת הכללה אפס?

סמנו: כן / לא

(שימו לב, "ליצור מסווג 1NN" דורש למעשה ליצור אוסף מתאים של דוגמאות אימון.)

אם ניתן – מה הגודל המינימלי של סט אימון שנדרש ליצירת מסווג כזה (בסדר גודל, למשל $\Theta(1), \Theta(\ln d)$ וכו')?

הסבירו בקצרה.



אחרת – הסבירו באופן מפורט למה לא.

תשובה:

כן, אבל רק אם יש לנו "שובר שוויון" כלשהו עבור

החלטות במרחק זהה (אם שוויון - OFF)

O האימון יכיל את ראשית הצירים OFF $(0,0,\dots,0)$

ואת כל הנק' עם 2 ביטים דלוקים בדיוק $\binom{d}{2} = \frac{d(d-1)}{2}$ וסה"כ $O(d^2)$

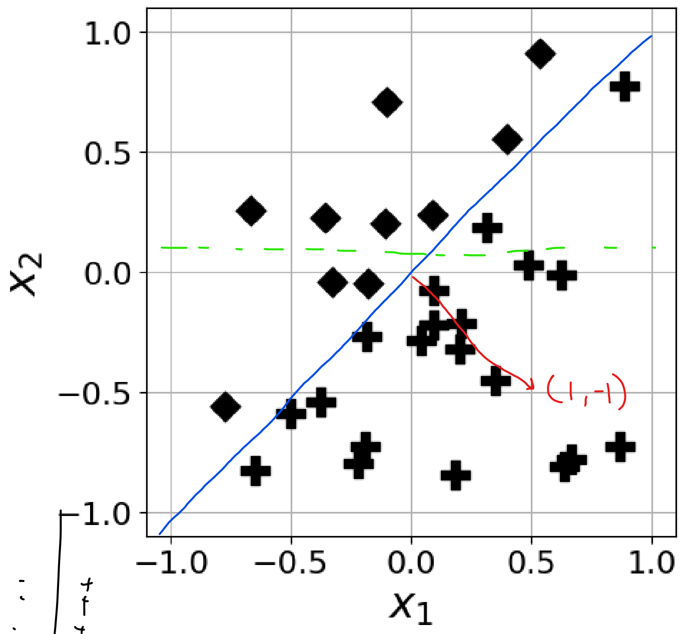
אם אין שובר שוויון: O האימון צריך להכיל את כל הנק' האפשריות

שמתויגות ב OFF ואת כל הנק' עם 2 ביטים כלומר $1+d+\binom{d}{2} = O(d^2)$

שאלה 3: AdaBoost ופרספטרון [21 נק']

משמאל מופיע אלגוריתם AdaBoost.

מימין סט אימון עם $m = 30$ דוגמאות בדו-ממד ($\mathcal{X} = \mathcal{R}^2$) עם תיוגים בינאריים ($+1$ מסומן ב-✚ ו-(-1) מסומן ב-◆).



Initialize $D^{(1)} = \left(\frac{1}{m}, \dots, \frac{1}{m}\right)$

For $t = 1, \dots, T$:

$h_t = \mathcal{A}\left(S, D^{(t)}\right)$

$\epsilon_t = \sum_i D_i^{(t)} \cdot \mathbf{1}_{h_t(x_i) \neq y_i}$

$\alpha_t = \frac{1}{2}\log\left(\frac{1}{\epsilon_t} - 1\right)$

$D_i^{(t+1)} = \frac{1}{Z_t} D_i^{(t)} \exp\left(-\alpha_t y_i h_t(x_i)\right)$

$h_s(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$

א. [5 נק'] כאשר ניתן למקבל הרצות נפרדות של $\mathcal{A}$ (עד T הרצות במקביל), מה פקטור השיפור שניתן לקבל בזמן ההרצה של AdaBoost? הסבירו בקצרה את תשובתכם.

תשובה:

$e^{-\alpha_t y_i h_t(x_i)}$ פקטור השיפור -

כאשר $\alpha_t = \frac{1}{2}\log\left(\frac{1}{\epsilon_t} - 1\right)$ ועוד $\epsilon_t = \sum_i D_i^{(1)} \cdot \mathbb{1}_{h_t(x_i) \neq y_i} \le \frac{1}{2}$

ϵ_t לא יכול להיות יותר מחצי, כי אנו בוחרים h_t כך ששגיאתו

תהיה מינימלית ותמיד נוכל לבחור פחות מחצי (אם יותר מחצי אז

להפוך את המסווג ונקבל פחות מחצי)

במקביל אין שיפור מכיוון שבכל אלגוריתם שנבנה על הצעדים הקודמים (boosting)

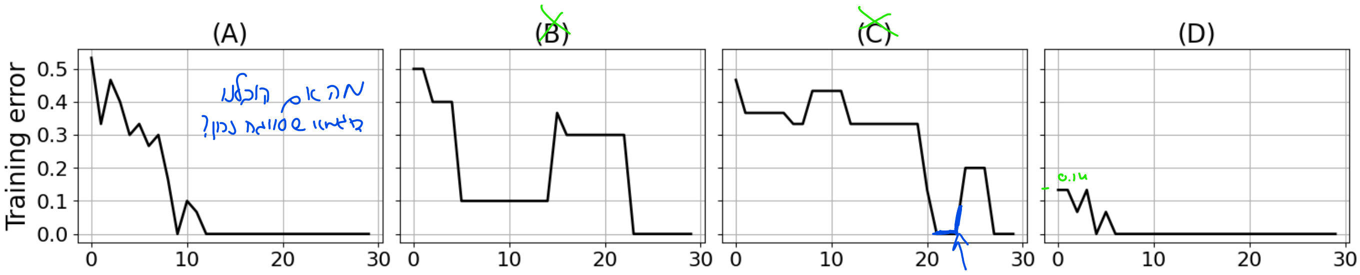
ולא bagging

ב. [5 נק'] מצאו $\mathbf{w} \in \mathbb{R}^2$ שיוצר מפריד ליניארי הומוגני עם שגיאת אימון 0 על הדאטה (משמע, $\text{sign}(\mathbf{w}^\top \mathbf{x}_i) = y_i$).

כתבו את שני הערכים במסגרת: $\mathbf{w} = [10^{10}, -10^{10}]$.

(ניתן להבין את התשובה על סמך התרשים; בבדיקת התשובה יינתן מרווח שגיאה סביר בערכים המספריים עצמם.)

ג. [11 נק'] לפניכם ארבע עקומות התכנסות שונות של שגיאת ה-0-1 על הנתונים שבתרשים הקודם.



(i)

(ii)

ציר y בתרשימים מראים את שגיאת האימון על <u>כל</u> 30 דוגמאות האימון.

נתאים **שניים** מהתרשימים לאלגוריתמי הלמידה הבאים:

i. אלגוריתם פרספטרון עם גודל צעד $\eta = 1$ ואתחול $\mathbf{w}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

ציר x בתרשים המתאים מראה 30 איטרציות של האלגוריתם (בכל איטרציה האלגוריתם רואה דוגמה בודדת).

הקיפו את האות של התרשים ששייך לאלגוריתם זה. **(A)** **(B)** **(C)** **(D)**

הסבירו בקצרה את תשובתכם.

הסבר:
סביר ביותר שהיא B מכיוון שלפעמים נראה לה(שת)חוות בגלל
שאנו בוחרים נק' שמסווגות נכון, אבל גם (אפילו את A) השגיאה יכולה לעלות תוך כדי האימון
אם למשל אנו בוחרים נק' בהלה שסווגה לא נכון ~~ב~~ ב- דבר שעלול לגרום
לסיבוב המסווג לא כדי פגיעה בסיום האימון
C לא ונפסל כי מגיע לשגיאה 0 אחרי הרבה (אולי עוזרים משגיאה 0)
B יותר סביר מ D בגלל השגיאה הראשונית (~15%) ונעשית לא 0)

ii. אלגוריתם AdaBoost כאשר $\mathcal{A}$ הוא אלגוריתם ERM ללמידת Decision Stump (עץ החלטה בעומק 1) ביחס לדוגמאות ולמשקלים הנתונים.

ציר x בתרשים המתאים מראה 30 איטרציות של האלגוריתם (כפי שמופיע בתחילת השאלה).

הקיפו את האות של התרשים ששייך לאלגוריתם זה. **(A)** **(B)** **(C)** **(D)**

הסבירו בקצרה את תשובתכם.

הסבר:
B,C לא מתאימים לAdaBoost בגלל ה השתנויות
A הכי סביר כי יש Stump ששה רק 5 אותות על ההתחלה
והיא זה שיבחר ! ~15% מתוך 30 זה ~5 שיבצו בול.
D פחות סביר בגלל השגיאה ההתחלתית של הדוגמ הספציפי התחון.

שאלה 4: מסווגים ליניאריים ואופטימיזציה [25 נק']

בכל השאלה נתון מדגם $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$ (כאשר $\mathcal{X} = \mathbb{R}^d$ וְ-$\mathcal{Y} = \{-1, +1\}$) ונניח שהוא פריד ליניארית הומוגנית.

נגדיר בעיית למידה קמורה בעזרת ה-exponential loss: $\underset{\mathbf{w}\in\mathbb{R}^d}{\operatorname{argmin}} \underbrace{\sum_{i=1}^{m} \exp(-y_i \mathbf{w}^\top \mathbf{x}_i)}_{\triangleq \mathcal{L}(\mathbf{w})}$.

$((x_i[0], x_i[1], \ldots, x_i[\ ]), \pm 1)^m$

א. [4 נק'] השתמשו במטריצת ההסיאן (שהיא $\nabla^2_{\mathbf{w}}\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{m} \exp(-y_i \mathbf{w}^\top \mathbf{x}_i)\, \mathbf{x}_i \mathbf{x}_i^\top$) כדי להוכיח שהבעיה קמורה.

הוכחה:

$\nabla_w^2 \mathcal{L}(w) = \sum_{i=1}^{m} e^{-y_i w^T x_i} \cdot (x_i x_i^T)$

כידוע $e^{\square} > 0$ עבור כל ערך $C \in \mathbb{R}$!

$x_i x$

$x_i x_i^T = \begin{bmatrix} x_i[0]x_i[0] & \cdots & x_i[0]x_i[d-1] \\ \vdots & \ddots & \vdots \\ x_i[d-1]x_i[0] & \cdots & x_i[d-1]x_i[d-1] \end{bmatrix}$

נשים לב כי זו מ' סימטרית

ובפרט PSD ולכן קמורה

לסיכום בסה"כ יש לנו $\left(\sum (\text{קבוע חיובי})(\text{קמורה})\right)$ דבר הנותן לנו תוצאה קמורה.

ב. [8 נק'] למרות הקמירות, הוכיחו שתחת הנחת הפרידות, לבעיה <u>אין מינימום</u>.

הוכחה (לרשותכם דפי טיוטה בסוף הגיליון):

מכיוון הבעיה פריד ליניארית הומוגנית, נקח וקטור ω

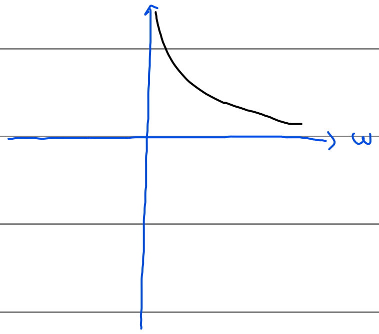
הפותר את הבעיה ונניח כי הוא המינימלי $\forall i \; y_i w^T x_i > 0$

אז נקח $\omega_\alpha = \alpha\omega$ כאשר $\alpha > 1$ ונקבל כי $\forall i \; y_i \alpha w^T x_i = \alpha(y_i w^T x_i) > y_i w^T x_i > 0$

<u>אך</u> $\omega_\alpha > \omega$ ולכן $e^{-y_i w^T x_i} > e^{-y_i w_\alpha^T x_i}$ $\mathcal{L}(\omega) = \sum_{i=1}^{m} e^{-y_i w^T x_i} > \sum_{i=1}^{m} e^{-y_i w_\alpha^T x_i} = \mathcal{L}(w_\alpha)$

כלומר סתירה לכך ש ω המינימלי, משמע אין מינימום לבעיה למרות הקמירות.

<u>גרף $e^{-\omega}$:</u>



<u>טענה (אין צורך להוכיח)</u>: מתקיים $\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}) = -\sum_{i=1}^{m}\exp(-y_i\mathbf{w}^\top\mathbf{x}_i)\, y_i\mathbf{x}_i$.

<u>נסמן</u>: יהיו $\mathbf{w}(1),\ldots,\mathbf{w}(t)\in\mathbb{R}^d$ הפתרונות המתקבלים מהרצת GD (מלא, לא סטוכסטי) על ה-exponential loss החל מהווקטור $\mathbf{w}(0)=\mathbf{0}_d$, עם גודל צעד $\eta>0$.

ג. [5 נק'] בסעיף זה <u>בלבד</u> נניח שיש רק דוגמה אחת $(\mathbf{x},y)$ במדגם, שהתיוג שלה הוא $y=1$ ושהיא מנורמלת ($\|\mathbf{x}\|=1$).
כמו כן, נניח שגודל הצעד בהרצת GD הוא $\eta=1$.
קל לראות שתחת תנאי הסעיף מתקיים: $\mathbf{w}(t) = \underbrace{\|\mathbf{w}(t)\|}_{\in\mathbb{R}}\underbrace{\mathbf{x}}_{\in\mathbb{R}^d}$ (ניתן להשתמש בכך מבלי להוכיח זאת).
<u>הוכיחו</u> שתחת תנאי הסעיף מתקיים: $\|\mathbf{w}(t)\| = \|\mathbf{w}(t-1)\| + \mathrm{e}^{-\|\mathbf{w}(t-1)\|}$ (אין צורך להשתמש באינדוקציה).

הוכחה:

$\cancel{\|w(t-1)\| + e^{\overbrace{-\|w(t-1)\|}^{N_{t-1}}} = \|w(t-2)\| + e^{N_{t-1}} + e^{N_{t-2}} = \ldots}$

$\cancel{= 0 + e^{N_{t-1}} + e^{N_{t-2}} + \ldots + e^{N_1} = e^{-\|w(t-1)\|} + \ldots + e^{-\|w(1-1)\|} = \sum_{m=0}^{t-1} e^{-\|w(m)\|} =}$

$\cancel{= (e^{-\|w(0)\|},\ e^{-\|w(2)\|} + e^{-\|w(1)\|},\ \ldots) = (1,\ 1+}$

$w_t = w_{t-1} - e^{-w_{t-1}^T x}\cdot x = \|w_{t-1}\|x - e^{-\|w_{t-1}\|x^T x}\cdot x = \left(\|w_{t-1}\| - e^{-\|w_{t-1}\|x^T x}\right)x \;/\cdot x^T$

(צעד גרדיאנט)

$(x^T x = I)$

$\frac{w_t}{x} = \|w_{t-1}\| - e^{-\|w_{t-1}\|}$

$\|w_t\| = \|w_{t-1}\| - e^{-\|w_{t-1}\|}$

תחת תנאי הסעיף האחרון ובהמשך לנוסחה הרקורסיבית שהוכחתם, ניתן להראות שמתקיים $\|\mathbf{w}(t)\| \approx \ln(t)$.

כעת נדון בתופעה כללית יותר (לדאטה פריד ליניארית עם $m \geq 1$ דוגמאות).

<u>נסמן</u>: $\hat{\mathbf{w}} = \operatorname{argmax}_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{\|\mathbf{w}\|}$ s.t. $\left(\left(\min_{i\in[m]} |\mathbf{w}^\top \mathbf{x}_i| = 1\right) \text{ and } (\forall i \in [m]:\ y_i \mathbf{w}^\top \mathbf{x}_i > 0)\right)$.

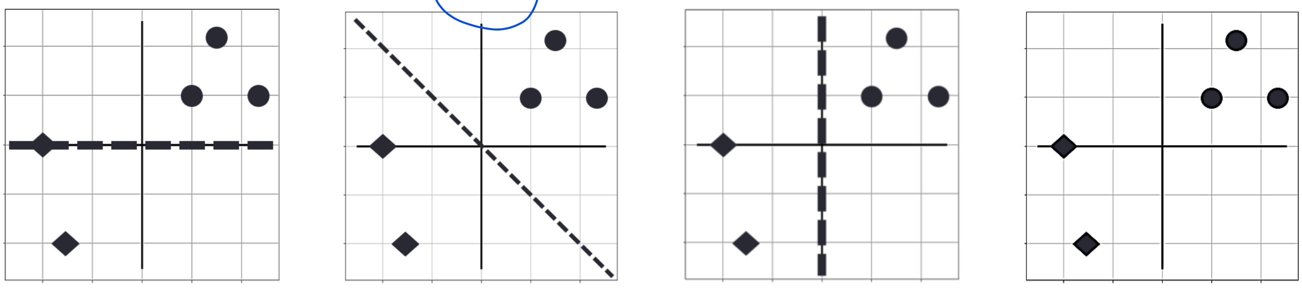
<u>משפט</u>: לכל $t \geq 3$, הפתרונות $\mathbf{w}(1), \dots, \mathbf{w}(t) \in \mathbb{R}^d$ שמתקבלים מהרצת GD על ה-exponential loss הם מהצורה:

$$\mathbf{w}(t) = \underbrace{\ln(t)}_{\in\mathbb{R}} \cdot \underbrace{\hat{\mathbf{w}}}_{\in\mathbb{R}^d} + \underbrace{\tilde{\mathbf{r}}(t)}_{\in\mathbb{R}^d}$$

עבור וקטורים $\tilde{\mathbf{r}}(1), \dots, \tilde{\mathbf{r}}(t)$ לא ידועים שמקיימים $\|\tilde{\mathbf{r}}(t)\| = \mathcal{O}(\ln\ln(t))$.

ד. [8 נק'] לפניכם תרשימים של סט אימון פריד ליניארית עם 5 דוגמאות ובנוסף כמה גבולות החלטה (decision boundaries).

סט האימון | (a) | (b) | (c)



(מפריד אנכי) | (מפריד אופקי)

איזה תרשים מתאר את גבול ההחלטה שמתקבל ע"י $\lim_{t\to\infty} \frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|}$? (הניחו שאין שגיאות נומריות)

סמנו באופן <u>ברור</u> והסבירו בקצרה. **(a)** **(b)** **(c)** **(לא ניתן לדעת)**

הסבר:

$\lim_{t\to\infty} \frac{w(t)}{\|w(t)\|} = \lim_{t\to\infty} \frac{\ln(t)\cdot w + \tilde{r}(t)}{\|\ln(t)\cdot w + \tilde{r}(t)\|} = \frac{\ln(t)}{\|\ln(t)\|} = (1,1)$

$\|\tilde{r}(t)\| = O(\ln\ln(t))$

שאלה 5: רגרסיה ורגולריזציה [25 נק']

עבור $\boldsymbol{y}\in\mathbb{R}^m$, $\mathbf{X}\in\mathbb{R}^{m\times d}$ וְ-$\lambda>0$, נסמן את הפתרונות של שלוש בעיות רגרסיה ליניארית שלמדנו:

$$\hat{\mathbf{w}}_{\text{LS}}=\underset{\mathbf{w}\in\mathbb{R}^d}{\operatorname{argmin}}\|\mathbf{Xw}-\mathbf{y}\|_2^2\,,\qquad \hat{\mathbf{w}}_{\ell 2}=\underset{\mathbf{w}\in\mathbb{R}^d}{\operatorname{argmin}}\left(\|\mathbf{Xw}-\mathbf{y}\|_2^2+\lambda\|\mathbf{w}\|_2^2\right),\qquad \hat{\mathbf{w}}_{\ell 1}=\underset{\mathbf{w}\in\mathbb{R}^d}{\operatorname{argmin}}\left(\|\mathbf{Xw}-\mathbf{y}\|_2^2+\lambda\|\mathbf{w}\|_1\right)$$

<u>הנחה:</u> בשאלה זו נניח שהעמודות של $\mathbf{X}$ אורתונורמליות (משמע $\mathbf{X}^\top\mathbf{X}=\mathbf{I}_{d\times d}$).

<u>תזכורת:</u> הנגזרת של התבנית הריבועית היא $\nabla_{\mathbf{w}}\|\mathbf{Xw}-\mathbf{y}\|_2^2=2\mathbf{X}^\top(\mathbf{Xw}-\mathbf{y})$.

א. [2 נק'] תחת ההנחה (אורתונורמליות), הוכיחו שמתקיים $\hat{\mathbf{w}}_{\text{LS}}=\mathbf{X}^\top\mathbf{y}$.

תשובה:

$2X^T(Xw-y)=0 \ /\cdot\frac{1}{2}\cdot X$

$X^T(Xw-y)X=0$

$X^TX(Xw-y)=0$

$I(Xw-y)=0$

$Xw=y$

$\boxed{w_{LS}=X^Ty}$

ב. [3 נק'] תחת ההנחה, מצאו ביטוי סגור ל-$\hat{\mathbf{w}}_{\ell 2}$ כפונקציה של λ, $\hat{\mathbf{w}}_{\text{LS}}$. צרפו לתשובתכם פיתוח <u>קצר</u> מתאים.

תשובה:

$\nabla_w \|Xw-y\|_2^2+\lambda\|w\|_2^2=2X^T(Xw-y)+2\lambda w=0 \ /:2\ /\cdot X$

$Xw-y+\lambda wX=0$

$w(X+\lambda X)-y=0 \Rightarrow wX(1+\lambda)=y \Rightarrow w=\frac{X^T\cdot y}{1+\lambda}=\boxed{\frac{w_{LS}}{1+\lambda}=w_{e2}}$

$\nabla_w \|Xw-y\|_2^2+\lambda\|w\|_1=2X^T(Xw-y)+\lambda=0$

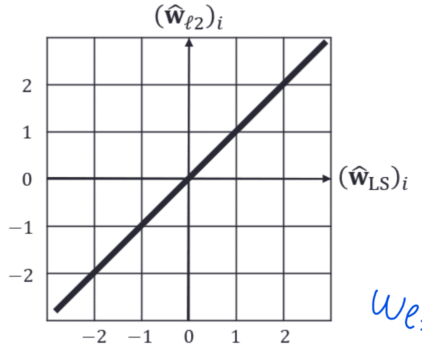
$=(Xw-y)=-\frac{1}{2}\lambda X \Rightarrow Xw=-\frac{1}{2}\lambda X+y$

$w_{e_1}=-\frac{1}{2}\lambda+X^Ty=\boxed{-\frac{1}{2}\lambda+w_{LS}=w_{e_1}}$

<u>נתונה טענה 1:</u> תחת ההנחה, מתקיים $(\hat{\mathbf{w}}_{\ell 1})_i = \operatorname{sign}((\hat{\mathbf{w}}_{\mathrm{LS}})_i) \cdot \max\left(0,\ |(\hat{\mathbf{w}}_{\mathrm{LS}})_i| - \frac{\lambda}{2}\right)$.

ג. [4 נק'] עבור כניסה i שרירותית וערך $\lambda = 2$, השתמשו בביטוי שמצאתם בסעיף הקודם ובטענה 1 וציירו <u>באופן ברור</u> על גבי התרשימים הבאים את העקומות של $(\hat{\mathbf{w}}_{\ell 2})_i$ וְ-$(\hat{\mathbf{w}}_{\ell 1})_i$ כפונקציה של $(\hat{\mathbf{w}}_{\mathrm{LS}})_i$.

דוגמה: לו היה מתקיים $(\hat{\mathbf{w}}_{\ell 2})_i = (\hat{\mathbf{w}}_{\mathrm{LS}})_i$, היה עליכם לצייר:

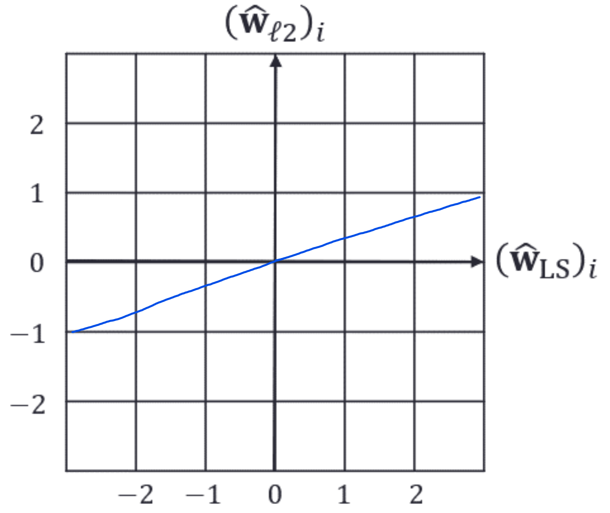
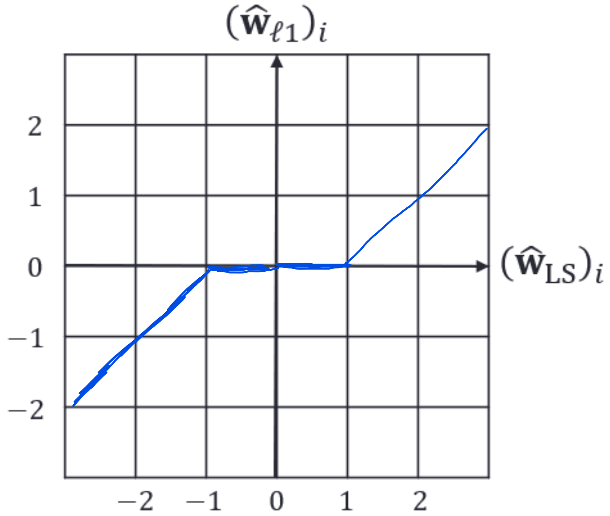


$$w_{\ell_2} = \frac{w_{LS}}{3}$$

$$w_{\ell_1} = sign(w_{LS}) \cdot max(0, w_{LS} - 1)$$

$(-1) \cdot$

תשובות (ציירו על גבי התרשימים):

$(\hat{\mathbf{w}}_{\ell 2})_i$ כפונקציה של $(\hat{\mathbf{w}}_{\mathrm{LS}})_i$ כאשר $\lambda = 2$	$(\hat{\mathbf{w}}_{\ell 1})_i$ כפונקציה של $(\hat{\mathbf{w}}_{\mathrm{LS}})_i$ כאשר $\lambda = 2$
	

ד. [8 נק'] הסבירו כיצד הסעיפים הקודמים ממחישים את התכונות של Ridge regression ו-Lasso וההבדלים ביניהם.

הסבר (בקצרה):

השמאלי מדמה Lasso בגלל התכונה של ה feature selection
כלומר הוא מאפס features שלא עברו threshold מסוים
(ניתן לראות זאת בכך שהם נמצאים על ציר ה x ורק ב $x>1$ ו-$x<-1$
הוא הגרף להיות ליניארי ב x לעומת זאת הימני מדמה Ridge בגלל
שאין גיזור שלפיהן וכל פיצ'ר יהיה משמעותי במקרה הזה כגדול ויו (ן)

כעת נוכיח את טענה 1 במקרה חד-ממדי פשוט, משמע: $d = 1$.

כלומר, עבור $\mathbf{X}, \boldsymbol{y} \in \mathbb{R}^m$ וְ-$\lambda > 0$, נגדיר:
$$\hat{w}_{\ell 1} = \underset{w \in \mathbb{R}}{\operatorname{argmin}} \underbrace{\left(\|\mathbf{X}w - \mathbf{y}\|_2^2 + \lambda |w|\right)}_{\triangleq \mathcal{L}(w)}, \qquad \hat{w}_{\text{LS}} = \underset{w \in \mathbb{R}}{\operatorname{argmin}} \|\mathbf{X}w - \mathbf{y}\|_2^2 .$$

<u>נמשיך להניח</u>: מתקיים $\mathbf{X}^\top\mathbf{X} = 1$ (ולכן עדיין מתקיים $\hat{w}_{\text{LS}} = \mathbf{X}^\top\mathbf{y}$).

לשימושכם תזכורת לתתי-נגזרות והכללה של תנאי האופטימליות למקרה של תת-נגזרת:

<u>תזכורת</u>: תהא $f: \mathbb{R} \to \mathbb{R}$ קמורה. נסמן ב-$\partial f(w)$ את קבוצת תתי-הנגזרות (sub-derivatives) שלה בנקודה $w \in \mathbb{R}$.

למשל, עבור $f(z) = |z|$, מתקיים $\partial f(-2) = \{-1\}$ וגם $\partial f(0) = [-1,1]$.

<u>תנאי אופטימליות (אין צורך להוכיח)</u>: תהא $f: \mathbb{R} \to \mathbb{R}$ קמורה. $f(w) = \min_{z \in \mathbb{R}} f(z)$ אמ"מ $0 \in \partial f(w)$.

ה. [8 נק'] הוכיחו שבמקרה החד-ממדי, תחת הנחת האורתונורמליות, מתקיים $\hat{w}_{\ell 1} = \operatorname{sign}(\hat{w}_{\text{LS}}) \cdot \max\left(0,\ |\hat{w}_{\text{LS}}| - \frac{\lambda}{2}\right)$.

הוכחה (לרשותכם דפי טיוטה בסוף הגיליון):

r(w) נגזרת של |w|, כשw גדול מ-0 +1, כשw קטן מ-0 -1.

$$W_{\ell_1} = \nabla_w \|Xw - y\|_2^2 + \lambda |w| = 2X^T(Xw - y) + \lambda r(w) = 0$$

$$\Rightarrow Xw - y = -\tfrac{1}{2}\lambda \cdot r(w) \cdot X \Rightarrow Xw = y - \tfrac{1}{2}\lambda X \cdot r(w) / X^T$$

$$\Rightarrow w = X^T y - \tfrac{1}{2}\lambda \cdot r(w) = \hat{w}_{LS} - \tfrac{\lambda}{2} \cdot r(w)$$

פיצול למקרים: $|w_{LS}| > \frac{\lambda}{2}$: $w = \operatorname{sign}(w_{LS}) \cdot \left(|\hat{w}_{LS}| - \frac{\lambda}{2}\right)$

ולכן כאשר $|w_{LS}| < \frac{\lambda}{2}$: נגדיר תת נגזרת כך להיות $r(0) = 0$

ונקבל $w = 0$

כלומר
$$\hat{w}_{\ell_1} = \begin{cases} 0, & |w_{LS}| < \frac{\lambda}{2} \\ \operatorname{sign}(\hat{w}_{LS}) \cdot \left(|w_{LS}| - \frac{\lambda}{2}\right), & \text{else} \end{cases}$$

שבהשוואה ערך ל: $\hat{w}_{\ell_1} = \operatorname{sign}(\hat{w}_{LS}) \cdot \max\left\{0, |w_{LS}| - \frac{\lambda}{2}\right\}$

(מקום נוסף בעמוד הבא)

המשך סעיף 5.ה':

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):