הטכניון – מכון טכנולוגי לישראל

הפקולטה למדעי המחשב

מבוא למערכות לומדות (236756)

סמסטר אביב תשפ"ב – 12 ביולי 2022

מרצה: ד"ר ניר רוזנפלד

מבחן מסכם מועד א' – 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).
- מותר להשתמש במחשבון.
- יש לכתוב בעט **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 – הקפידו להקיף 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.**

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: Feature selection and Classification models [20 נק']

נתון סֶט אימון עם $d \geq 3$ מאפיינים (features) ועם תיוגים בינאריים, משמע $\boldsymbol{x}_i \in \mathbb{R}^d, y_i \in \{-1,1\}$.

בסט האימון אין שתי דוגמאות זהות.

נתון שהקורלציה בין שני מאפיינים מסוימים A, B בסט האימון היא בדיוק -1.

משמע, לכל דוגמה $\boldsymbol{x}_i$ מתקיים $x_i[B] = ax_i[A] + b$ עבור $a < 0$ וְ-$b \in \mathbb{R}$.

לומדים שני מסווגים:

- בשלב הראשון: לומדים מסווג על סט האימון המקורי ומחשבים עליו את דיוק האימון.
- בשלב השני:
  - מסירים את המאפיין B ומקבלים סט אימון מעודכן (שבו $d - 1$ פיצ'רים).
  - מאמנים מסווג חדש על סט האימון המעודכן, ומחשבים עליו את דיוק האימון המעודכן.

עבור כל אלגוריתם למידה, סמנו האם דיוק האימון של המסווג החדש על סט האימון המעודכן זהה בהכרח לזה של המסווג המקורי על סט האימון המקורי.

הסבירו בקצרה את תשובותיכם (2-4 משפטים בכל סעיף).

הניחו שאין צעדים סטוכסטיים (אקראיים) בריצת האלגוריתמים ואין שגיאות נומריות (בפרט, בעיות קמורות מתכנסות לפיתרון האנליטי שלהן במדויק).

א. kNN עם $k = 1$ (דוגמה לא נחשבת שכנה של עצמה). דיוק האימון זהה בהכרח? כן / **לא**

הסבר: נניח למשל שהסקאלה של פיצ'ר B הרבה יותר גדולה מכל האחרות (במקרה קיצון: $a \to -\infty$). הפיצ'ר הזה "משתלט" על המרחקים האוקלידים. לאחר הסרתו נקבל משקלים שונים לגמרי ושכנויות שונות.

ב. Hard-SVM ליניארי לא הומוגני בהנחה שהדאטה המקורי פריד. דיוק האימון זהה בהכרח? **כן** / לא

הסבר: הדאטה המקורי פריד לינארית. הראינו **בתרגול** שפיצ'רים ת"ל לא מוסיפים capacity למודלים לינאריים. לכן הדאטה לאחר ההסרה גם הוא פריד לינארית בהכרח. בשני המקרים יתקבל דיוק 100%.

ג. Soft-SVM ליניארי לא הומוגני עם $\lambda = 1$ בהנחה שהדאטה המקורי פריד. דיוק האימון זהה בהכרח? כן / **לא**

הסבר: אמנם מבחינת Hinge ניתן להגיע במרחב החדש לאותו loss כמו של המפריד האופטימלי המקורי, אבל רכיב הרגולריזציה עשוי להשתנות. לכן ייתכן שהפיתרון האופטימלי ישתנה ונקבל מפריד שונה ודיוק שונה.

ד. ID3 המשתמש באנטרופיה ובונה עץ בעומק מירבי 4. דיוק האימון זהה בהכרח? **כן** / לא

הסבר: בכל שלב בריצת ID3, כאשר נבחר פיצול לפי סף של B במרחב המקורי, ניתן לבחור פיצול לפי סף שקול של A במרחב החדש. לאחר הפיצול יתקבלו צמתים זהים ולכן ה-IG של שני הפיצולים המדוברים זהה ומקסימלי (לפי אופן פעולת ID3 במרחב המקורי). לכן ID3 **בהכרח** יבחר את הפיצול השקול לפי סף של A ויתקבל עץ זהה ודיוק זהה.

שאלה 2: Kernel SVM [18 נק']

<u>תזכורת</u>: פונקציה מהווה קרנל חוקי אם ניתן לכתוב אותה בתור $K(u,v) = \langle \phi(u), \phi(v) \rangle = \sum_{i=1}^{n} \phi_i(u)\phi_i(v)$ עבור מיפוי כלשהו $\phi: \mathcal{X} \to \mathbb{R}^n$.

נתונות שתי פונקציות קרנל חוקיות $K_1, K_2: (\mathcal{X} \times \mathcal{X}) \to \mathbb{R}$ והמיפויים המתאימים להן $\phi, \psi: \mathcal{X} \to \mathbb{R}^n$. משמע, מתקיים:
$K_1(u,v) = \langle \phi(u), \phi(v) \rangle, \qquad K_2(u,v) = \langle \psi(u), \psi(v) \rangle$.

הוכיחו שהפונקציות הבאות מהוות קרנלים חוקיים (כפי שמוסבר בתזכורת לעיל).

בכל סעיף, הגדירו <u>בבירור</u> פונקציית מיפוי $\phi': \mathcal{X} \to \mathbb{R}^{n'}$ מתאימה והראו שמתקיים $K'(u,v) = \langle \phi'(u), \phi'(v) \rangle$ כנדרש.

א. [5 נק'] הפונקציה $K'(u,v) = \left(K_1(u,v)\right)^2$.

הוכחה:

נגדיר ϕ' עבור $n' = n^2$. באופן אינטואיטיבי נתאים כניסה לכל צירוף $i,j \in [n]$:

$$\phi'_{ij}(u) = \phi_i(u)\phi_j(u)$$

צריך להראות שמתקיים $K'(u,v) = \langle \phi'(u), \phi'(v) \rangle = \sum_{ij} \phi'_{ij}(u)\phi'_{ij}(v) = \left(K_1(u,v)\right)^2$.

ב. [5 נק'] הפונקציה $K'(u,v) = K_1(u,v) + 3 \cdot K_2(u,v) + 1$.

הוכחה:

נגדיר מיפוי בעזרת שרשור וקטורים: $\phi'(u) = \begin{bmatrix} \phi(u) \\ \sqrt{3}\psi(u) \\ 1 \end{bmatrix}$. צריך להוכיח את השוויון המבוקש.

ג. [8 נק'] אילו מהפעולות הבאות אמורות להפחית overfitting ב-Kernel SVM?
סמנו את כל התשובות המתאימות (השאלה אינה עוסקת במקרי קצה אלא במקרה הסביר).

a. לעבור מפונקציית הקרנל $K(\boldsymbol{u},\boldsymbol{v})=(\boldsymbol{u}^\top\boldsymbol{v})^p$ לפונקציית הקרנל $K(\boldsymbol{u},\boldsymbol{v})=(\boldsymbol{u}^\top\boldsymbol{v}+1)^p$, כאשר $p\in\mathbb{N}_{\ge 2}$.

b. להקטין את השונות σ^2 של קרנל RBF (שמוגדר בתור $K(\boldsymbol{u},\boldsymbol{v})=\exp\left\{-\frac{1}{2\sigma^2}\|\boldsymbol{u}-\boldsymbol{v}\|_2^2\right\}$).

c. לפתור (במדויק) את הבעיה ה-primal במקום לפתור (במדויק) את הבעיה ה-dual.

d. להגדיל את מקדם הרגולריזציה λ (בהתאמה: להקטין את C).

e. להגדיל את סֶט האימון (באופן $i.i.d.$ מאותה התפלגות של הדאטה המקורי).

f. להגדיל את סֶט המבחן (באופן $i.i.d.$ מאותה התפלגות של הדאטה המקורי).

הערות בדיקה: בסעיף ג' ירדו 2 נקודות על כל טענה מיותרת שסומנה או טענה נכונה שחסרה.

שאלה 3: VC-dimension [16 נק']

נגדיר את מחלקת ההיפותזות של מסווגים ליניאריים הומוגניים ב-d ממדים:

$$\mathcal{X} = \mathbb{R}^d,\ \ \mathcal{H}_{\text{lin}}^d = \{\boldsymbol{x} \mapsto \operatorname{sign}(\boldsymbol{w}^\top \boldsymbol{x}) \colon\ \boldsymbol{w} \in \mathbb{R}^d\}.$$

כדי למנוע טעויות, נקבע שלאורך כל השאלה $\operatorname{sign}(0) = 0$ (ואחרת הפונקציה מחזירה ± 1 בהתאם לסימן).

א. [7 נק'] הוכיחו שמתקיים $\operatorname{VCdim}(\mathcal{H}_{\text{lin}}^d) \geq d$.

הוכחה:

כמו בתרגול.

ב. [9 נק'] הוכיחו שממד ה-VC הוא בדיוק d על ידי כך שתוכיחו שמתקיים $\operatorname{VCdim}(\mathcal{H}_{\text{lin}}^d) < d+1$.

<u>רמז</u>: כל אוסף $\boldsymbol{x}_1, \dots, \boldsymbol{x}_d, \boldsymbol{x}_{d+1}$ של $d+1$ וקטורים כלשהם ב-$\mathbb{R}^d$ הינו תלוי ליניארית, ובהכרח אחד הווקטורים באוסף (בה"כ $\boldsymbol{x}_{d+1}$) מקיים $\boldsymbol{x}_{d+1} = \sum_{i=1}^d z_i \boldsymbol{x}_i$ עבור אוסף סקלרים $z_1, \dots, z_d \in \mathbb{R}$ שלפחות אחד מהם שונה מאפס.

הוכחה:

יהי אוסף $\boldsymbol{x}_1, \dots, \boldsymbol{x}_d, \boldsymbol{x}_{d+1}$ של $d+1$ וקטורים כלשהם ב-$\mathbb{R}^d$. לפי הרמז מתקיים $\boldsymbol{x}_{d+1} = \sum_{i=1}^d z_i \boldsymbol{x}_i$.

נגדיר תיוג
$$y_i = \begin{cases} \operatorname{sign}(z_i), & i \leq d, z_i \neq 0 \\ +1, & i \leq d, z_i = 0 \\ -1, & i = d+1 \end{cases}$$

נניח בשלילה שקיים $\boldsymbol{w} \in \mathbb{R}^{d+1}$ שמתייג נכונה את כל הנקודות. משמע, $\forall i \colon \operatorname{sign}(\boldsymbol{w}^\top \boldsymbol{x}_i) = y_i$.

לכן הפרדיקציה עבור $\boldsymbol{x}_{d+1}$ תהיה (נשתמש בכך שיש לפחות מקדם z_i שונה מאפס):

$$\operatorname{sign}(\boldsymbol{w}^\top \boldsymbol{x}_{d+1}) = \operatorname{sign}\left(\boldsymbol{w}^\top \sum_{i=1}^d z_i \boldsymbol{x}_i\right) = \operatorname{sign}\left(\sum_{i=1}^d \underbrace{z_i \boldsymbol{w}^\top \boldsymbol{x}_i}_{\substack{\geq 0 \\ \text{כי } \operatorname{sign}(\boldsymbol{w}^\top \boldsymbol{x}_i) = y_i = \operatorname{sign}(z_i)}}\right) = 1 \neq -1 = y_{d+1}$$

קיבלנו סתירה. קיים תיוג שלא ניתן ללמוד ולכן $\operatorname{VCdim}(\mathcal{H}_{\text{lin}}^d) < d+1$.

שאלה 4: Bagging and Boosting [22 נק']

משמאל מופיע אלגוריתם AdaBoost.

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

For t=1, …, T:

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

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

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

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

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

א. [5 נק'] שימו לב שבכל איטרציה האלגוריתם מעדכן את ההתפלגות באופן $D_i^{(t+1)} = \frac{1}{Z_t} D_i^{(t)} \exp\left(-\alpha_t y_i h_t(x_i)\right)$ כאשר $Z_t \triangleq \sum_{j=1}^{m} D_j^{(t)} \exp\left(-\alpha_t y_j h_t(x_j)\right)$ הוא גורם נירמול.

הוכיחו שמתקיים $D_i^{(t+1)} = c \cdot \exp\left(-y_i \sum_{k=1}^{t} \alpha_k h_k(x_i)\right)$ עבור קבוע כלשהו $c > 0$ (שתלוי ב-t).

הוכחה:

כמו בתרגול.

בסעיף הבא נוכיח שהבחירה של האלגוריתם ב-$\alpha_t = \frac{1}{2} \log\left(\frac{1}{\epsilon_t} - 1\right)$ הינה אופטימלית.

<u>תזכורת</u>: הגדרנו בתרגול את ההיפותזה ה-unthresholded ה"חזקה" באיטרציה t בתור: $h_s^{(t)} = \sum_{k=1}^{t} \alpha_k h_k(x_i)$,

והראינו שקיים קבוע חיובי C כך שערך ה-exp. loss של ההיפותזה הוא:

$$\mathcal{L}_{\exp}\left(h_s^{(t)}\right) \triangleq \frac{1}{m} \sum_{i=1}^{m} \exp\left\{-y_i \sum_{k=1}^{t} \alpha_k h_k(x_i)\right\} = \underbrace{C}_{>0} \cdot \left(e^{-\alpha_t} + \left(e^{\alpha_t} - e^{-\alpha_t}\right) \sum_i D_i^{(t)} 1_{h_t(x_i) \neq y_i}\right),$$

כאשר C אינו תלוי בבחירה של המקדם α_t שנבחר בזמן t.

ב. [5 נק'] הוכיחו שבהינתן h_t ו-$\left\{D_i^{(t)}\right\}_{i \in [m]}$, הבחירה בְּ-$\alpha_t = \frac{1}{2} \log\left(\frac{1}{\epsilon_t} - 1\right)$ ממזערת את $\mathcal{L}_{\exp}\left(h_s^{(t)}\right)$.

<u>הערה</u>: שימו לב ש-$\epsilon_t \triangleq \sum_i D_i^{(t)} \cdot \mathbf{1}_{h_t(x_i) \neq y_i}$ והוא קבוע בהינתן h_t ו-$\left\{D_i^{(t)}\right\}_{i \in [m]}$.

הוכחה:

כמו בתרגול.

הסעיף הבא לא קשור לאלגוריתם AdaBoost אלא לשיטת Ensemble שממשקלת היפותזות באופן כללי.

ג. [12 נק'] נתון סֶט אימון $\{(x_i, y_i)\}_{i=1}^{m}$ עם סיווגים בינאריים ($y_i \in \{-1,+1\}$).

צוות של 50 מומחים ומומחיות הגדיר ידנית 50 היפותזות בינאריות $h_1, \dots, h_{50}: \mathcal{X} \to \{-1,+1\}$.

המטרה שלנו היא למשקל אותן, משמע, ליצור היפותזה $h_\alpha(x) = \text{sign}\left(\sum_{k=1}^{50} \alpha_k h_k(x)\right)$.

כדי למנוע טעויות, נקבע שלאורך כל השאלה $\text{sign}(0) = 0$ (ואחרת הפונקציה מחזירה ± 1 בהתאם לסימן).

נתון: קיימים משקלים $\alpha_1^*, \dots, \alpha_{50}^* \in \mathbb{R}$ כך שהשגיאה האמפירית (שגיאת האימון) היא אפס.

כעת, נרצה ללמוד אוסף משקלים $\alpha_1, \dots, \alpha_{50} \in \mathbb{R}$.

(i) נסחו בעיית אופטימיזציה מתאימה **וקמורה** (ביחס למשקלים הנלמדים $\alpha_1, \dots, \alpha_{50}$) כך שהפיתרון שלה ממזער את השגיאה האמפירית.

תשובה (לרשותכם טיוטה בסוף הגיליון):

נסתכל על $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_{50}]^\top \in \mathbb{R}^{50}$ ועל $\boldsymbol{h}(x) = [h_1(x), \dots, h_{50}(x)]^\top \in \{\pm 1\}^{50}$ וֶקטורים.

$$\text{argmin}_{\boldsymbol{\alpha} \in \mathbb{R}^{50}} \sum_{i=1}^{m} \max\left\{0,\ \ 1 - y_i \sum_{k=1}^{50} \alpha_k h_k(x_i)\right\} = \text{argmin}_{\boldsymbol{\alpha} \in \mathbb{R}^{50}} \sum_{i=1}^{m} \max\{0,\ \ 1 - y_i \boldsymbol{\alpha}^\top \boldsymbol{h}(x_i)\}$$

(ii) נמקו בקצרה מדוע הבעיה שהגדרתם קמורה (מומלץ להשתמש בטענות מההרצאה ומהתרגול).

תשובה (לרשותכם טיוטה בסוף הגיליון):

נשים לב שקיבלנו בדיוק בעיית Soft-SVM בלי רגולריזציה במרחב החדש.

כפי שהראינו בתרגיל הבית – הפונקציה $1 - y_i \boldsymbol{\alpha}^\top \boldsymbol{h}(x_i)$ לינארית ב-$\boldsymbol{\alpha}$ ולכן קמורה ביחס אליו.

מקסימום בין שתי פונקציות קמורות הוא קמור.

ולסיום – סכום של פונקציות קמורות הוא קמור.

(iii) יהי אוסף המשקלים $\alpha_1, \dots, \alpha_{50} \in \mathbb{R}$ פיתרון (אופטימלי) של הבעיה הקמורה שהגדרתם לעיל.

הוכיחו שהשגיאה האמפירית של פיתרון זה היא מינימלית.

הוכחה:

לפי הנתון – קיים $\boldsymbol{\alpha}^*$ עבורו לכל דוגמה מתקיים $\text{sign}\left(\boldsymbol{\alpha}^{*\top}\boldsymbol{h}(\boldsymbol{x_i})\right) = y_i \in \{\pm 1\}$.

ולכן $y_i\boldsymbol{\alpha}^{*\top}\boldsymbol{h}(\boldsymbol{x_i}) > 0$ לכל דוגמה.

מכאן מובטח שקיים פיתרון $\beta\boldsymbol{\alpha}^*$ עבור סקאלר כלשהו $\beta > 0$ גדול מספיק,

כך שמתקיים $\forall i: y_i(\beta\boldsymbol{\alpha}^*)^\top\boldsymbol{h}(\boldsymbol{x_i}) > 1$ ושהוא מאפס את ה-objective שהגדרנו.

מובטח שיימצא פיתרון שמאפס את ה-objective (כי קיימים פתרונות כאלה וה-objective חיובי).

בגלל שה-Hinge loss הוא surrogate עבור השגיאה האמפירית, גם השגיאה האמפירית תתאפס

עבור כל פיתרון אופטימלי של הבעיה שהוגדרה.

חלק ב' – שאלות אמריקאיות [24 נק']

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

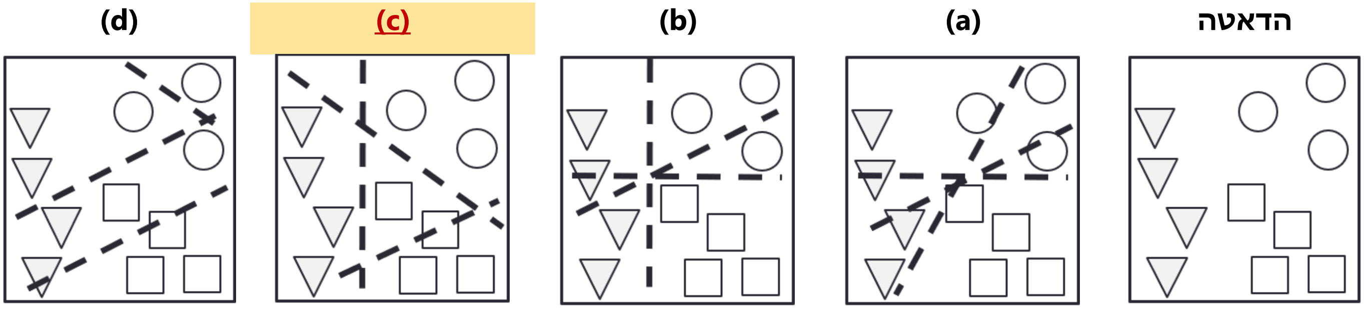
הערות בדיקה: בסעיפים ב'-ד' ירדו 2 נקודות על כל טענה מיותרת שסומנה או טענה נכונה שחסרה.

א. [6 נק'] נתון דאטה דו-ממדי עם שלושה תיוגים אפשריים – משולש, מעגל וריבוע.

מפעילים One-vs-All (OVA) עם פרספטרון בתור מסווג בסיס. לכל בעיה בינארית מריצים את הפרספטרון מספיק איטרציות כך שאם הבעיה פרידה ליניארית – הפרספטרון יתכנס (חישבו מה קורה אחרת).

בכל תרשים מצוירים שלושה גבולות החלטה שאמורים לתאר את הגבולות שהתקבלו ע"י מסווגי הבסיס (הבינאריים).

הקיפו את האות שמתאימה לתרשים היחיד שמתאר גבולות החלטה שיכולים להתקבל ע"י השיטה שתוארה לעיל.



ב. [6 נק'] נתון דאטה d-ממדי $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m}$ עם סיווגים בינאריים (± 1).

סמנו את כל הטענות הנכונות **בהכרח** ביחס **לשיטות להורדת ממד** לממד נמוך כלשהו.

הבהרה: זו שאלת חשיבה ולא שאלת טריוויה.

a. הדאטה המקורי פריד ליניארית ⟸ הדאטה פריד ליניארית לאחר הורדת ממד ליניארית עם PCA.

b. הדאטה המקורי פריד ליניארית ⟸ בהסתברות גבוהה, הדאטה פריד ליניארית לאחר הורדת ממד ליניארית עם Random Projections.

c. ~~הדאטה המקורי פריד ליניארית ⟸ הדאטה פריד ליניארית לאחר הורדת ממד עם Kernel-PCA (עם קרנל לא ליניארי).~~

d. הדאטה פריד ליניארית לאחר הורדת ממד ליניארית עם PCA ⟸ הדאטה המקורי פריד ליניארית.

e. PCA היא שיטה Supervised בעוד ש-Random Projections היא שיטה Unsupervised.

הערות בדיקה: על אף שטענה c לא נכונה, הוחלט שלא להוריד ניקוד על סימון שלה בגלל שלא התעכבנו על זה בהרצאות.

ג. [6 נק'] נתונה בעיית רגרסיה לינארית (עבור $\lambda \geq 0,\ p \in \{0.5, 1, 2\}$):

$$\mathrm{argmin}_{\boldsymbol{w}} \left(\frac{1}{m}\sum_{i=1}^{m}(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2 + \lambda\|\boldsymbol{w}\|_p^p\right) = \mathrm{argmin}_{\boldsymbol{w}} \left(\frac{1}{m}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{w}\|_p^p\right).$$

סמנו את <u>כל</u> הטענות הנכונות.

<u>הערה</u>: בכל הסעיפים, שואלים על קמירות ביחס ל-$\boldsymbol{w}$.

<u>תזכורת</u>: $\|\boldsymbol{w}\|_p^p \triangleq \sum_{k=1}^{d}|w_k|^p$.

a. כאשר $\lambda = 0$: ההסיאן הוא $\frac{2}{m}\boldsymbol{X}^\top \boldsymbol{X}$ שהיא מטריצה PSD אבל לא בהכרח PD, ולכן הבעיה לא בהכרח קמורה.

b. <u>כאשר $\lambda > 0$ וגם $p = 2$: הבעיה קמורה בהכרח.</u>

c. <u>כאשר $\lambda > 0$ וגם $p = 1$: הבעיה קמורה בהכרח.</u>

d. כאשר $\lambda > 0$ וגם $p = 0.5$: הבעיה קמורה בהכרח.

e. כאשר $\lambda > 0$ וגם $p = 1$: הבעיה אינה גזירה ביחס ל-$\boldsymbol{w}$ ולכן לא ניתן להפעיל שיטות גרדיינט (או סאב-גרדיינט).

ד. [6 נק'] רוצים לאמן רשת נוירונים בתצורת Autoencoder.

נסמן את אוסף כל הפרמטרים של הרשת בתור Θ. לכל דוגמה $\boldsymbol{x}_i \in \mathbb{R}^d$ נסמן את הפלט של הרשת בתור $\hat{\boldsymbol{x}}_i \in \mathbb{R}^d$.

בהרצאה ראינו שניתן לאמן את הרשת ע"י פיתרון הבעיה הבאה: $\mathrm{argmin}_{\Theta}\left(\lambda\|\Theta\|_2^2 + \frac{1}{m}\sum_{i=1}^{m}\|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|_2^2\right)$.

אילו מבין הבעיות הבאות עשויות גם כן להתאים לאימון רשת Autoencoder (עבור $\lambda > 0$)?

סמנו את <u>כל</u> התשובות המתאימות.

a. $\mathrm{argmin}_{\Theta}\left(\lambda\|\Theta\|_2^2 + \frac{1}{m}\sum_{i=1}^{m}\|\boldsymbol{x}_i + \hat{\boldsymbol{x}}_i\|_2^2\right)$

b. $\mathrm{argmin}_{\Theta}\left(\lambda\|\Theta\|_2^2 + \frac{1}{m}\sum_{i=1}^{m} exp\{\|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|_2^2\}\right)$

c. $\mathrm{argmin}_{\Theta}\left(\lambda\|\Theta\|_2^2 - \frac{1}{m}\sum_{i=1}^{m}\|\boldsymbol{x}_i\|_2^2 + \frac{1}{m}\sum_{i=1}^{m}\|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|_2^2\right)$

d. $\mathrm{argmin}_{\Theta}\left(\lambda\|\Theta\|_2^2 + \frac{1}{m}\sum_{i=1}^{m}\max\{0, 1 - \mathrm{sign}(\boldsymbol{x}_i^\top \hat{\boldsymbol{x}}_i)\}\right)$

e. $\mathrm{argmin}_{\Theta}\left(-\lambda\|\Theta\|_2^2 + \frac{1}{m}\sum_{i=1}^{m}\frac{\boldsymbol{x}_i^\top \hat{\boldsymbol{x}}_i}{\|\boldsymbol{x}_i\|_2\|\hat{\boldsymbol{x}}_i\|_2}\right)$