הטכניון – מכון טכנולוגי לישראל

הפקולטה למדעי המחשב

מבוא למערכות לומדות (236756)

סמסטר חורף תשפ"ג – 15 בפברואר 2023

מרצה: ד"ר יונתן בלינקוב

<u>מבחן מסכם מועד א'</u>

הנחיות הבחינה:

- **משך הבחינה:** שלוש 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.
- הוכחות והפרכות צריכות להיות פורמליות.
- 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 17 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!

חלק א' – שאלות פתוחות [94 נק']

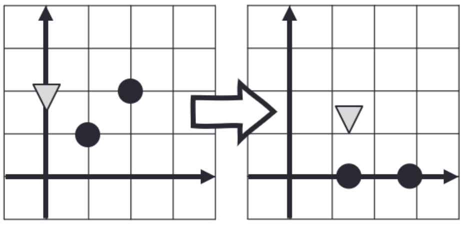
שאלה 1: רגישות של מסווגים לסיבובים [24 נק']

נתון סֵט אימון דו-ממדי עם תיוגים בינאריים, משמע לכל $i = 1, \dots, m$ מתקיים $\mathbf{x}_i \in \mathbb{R}^2, y_i \in \{-1,1\}$.

לומדים שני מסווגים:

- <u>בשלב הראשון</u>: לומדים מסווג על סט האימון המקורי ומחשבים עליו את דיוק האימון.
- <u>בשלב השני</u>:
  - מסובבים את כל הדאטה ב-$45°$ סביב ראשית הצירים (ראו דוגמה).
  - מאמנים מסווג <u>חדש</u> על סט האימון המעודכן,
    ומחשבים עליו את דיוק האימון המעודכן.

<u>סיבוב לדוגמה</u>



<u>תזכורת</u>: הסיבוב של הדאטה יכול להתבצע ע"י מיפוי $\mathbf{x}_i \mapsto \mathbf{Q}\mathbf{x}_i$, כאשר $\mathbf{Q} \in \mathbb{R}^{2\times 2}$ הינה מטריצת סיבוב ממשית, שהיא מטריצה אורתונורמלית המקיימת $\mathbf{Q}^{-1} = \mathbf{Q}^\top$.

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

הסבירו <u>בקצרה</u> את תשובותיכם (2-4 משפטים בכל סעיף).

הניחו שאין צעדים אקראיים או שגיאות נומריות בריצת האלגוריתמים (בעיות קמורות מתכנסות לפתרון האנליטי במדויק).

א. <u>kNN עם $k = 3$ (דוגמה לא נחשבת שכנה של עצמה).</u> דיוק האימון זהה בהכרח? **כן / לא**

הסבר: ____

____

____

____

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

הסבר: ____________________

____

____

____

ג. AdaBoost with decision stumps, maximum of $T = 10$ iterations. דיוק האימון זהה בהכרח? **כן / לא**

הסבר: ____________________

____

____

____

ד. Homogeneous logistic regression with L2 regularization (ללא הנחה שהדאטה המקורי פריד):

$$\operatorname{argmin}_{\mathbf{w}\in\mathbb{R}^2} \sum_i \ln(1+\exp(-y_i\mathbf{w}^\top\mathbf{x}_i)) + \lambda\|\mathbf{w}\|_2^2, \quad \text{for some } \infty > \lambda > 0$$

דיוק האימון זהה בהכרח? **כן / לא**

הסבר: ____________________

____

____

____

שאלה 2: VC-dimension [20 נק']

א. [6 נק'] הוכיחו שלכל מחלקת היפותזות סופית $\mathcal{H}$ מתקיים בהכרח $\mathrm{VCdim}(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$.

רמז: ניתן להוכיח זאת בשלילה ובקצרה.

הוכחה (לרשותכם דפי טיוטה בסוף הגיליון):

ב. [4 נק'] נגדיר מרחב דוגמאות סופי $\mathcal{X} = \{x_1, x_2\}$ (במרחב יש שתי דוגמאות במרחב כלשהו, לא נבטא במפורש את x_1, x_2)

ומחלקת היפותזות סופית $\mathcal{H} = \{h_1, h_2\}$, כאשר h_1, h_2 מוגדרות באופן הבא:

	h_1	h_2
x_1	$+1$	-1
x_2	-1	-1

(המשמעות של הטבלה היא, למשל, שמתקיים $h_1(x_1) = +1$.)

<u>ענו</u>: מתקיים $\boxed{} = \mathrm{VCdim}(\mathcal{H})$. <u>הוכיחו</u> את תשובתכם.

הוכחה:

רמז: גם בסעיפים הבאים ניתן להגדיר מרחב דוגמאות סופי $\mathcal{X} = \{x_1, \dots, x_n\}$ ואוסף סופי $\mathcal{H}$ של היפותזות $h: \mathcal{X} \to \{-1, +1\}$.

על ההיפותזות להיות **שונות** (distinct) במובן שאין שתי היפותזות ב-$\mathcal{H}$ שמחזירות פלט זהה על כֹּל הדוגמאות ב-$\mathcal{X}$.

בשני הסעיפים אין צורך להוכיח את ה-VC-dimension של המחלקות שתציעו.

ג. [5 נק'] הראו מחלקת היפותזות $\mathcal{H}$ עבורה $|\mathcal{H}| = 4$ וגם $\text{VCdim}(\mathcal{H}) = 1$.

תשובה (לרשותכם דפי טיוטה בסוף הגיליון):

ד. [5 נק'] הראו מחלקות היפותזות $\mathcal{H}_1, \mathcal{H}_2$ עבורן $\text{VCdim}(\mathcal{H}_1) = \text{VCdim}(\mathcal{H}_2) = 1$ וגם $\text{VCdim}(\mathcal{H}_1 \cup \mathcal{H}_2) = 3$.

תשובה:

שאלה 3: רגרסיה ליניארית [25 נק']

בשאלה זו מרחב הדוגמאות הוא $\mathcal{X}=\mathbb{R}^d$ ומרחב התיוגים הוא $\mathcal{Y}=\mathbb{R}$.

נתון מדגם $S=\{(\mathbf{x}_\mathrm{i},y_i)\}_{i=1}^m$.

רגרסיה ליניארית המשלבת רגולריזציה מסוג L1 ומסוג L2 נקראת ElasticNet:

$$\mathbf{w}^\star \triangleq \operatorname{argmin}_{\mathbf{w}\in\mathbb{R}^d}\left(\underbrace{\sum_{i=1}^m(\mathbf{w}^\top\mathbf{x}_i-y_i)^2}_{\mathcal{L}(\mathbf{w};S)}+\underbrace{\lambda\alpha\|\mathbf{w}\|_2^2+\lambda(1-\alpha)\|\mathbf{w}\|_1}_{R_{\lambda,\alpha}(\mathbf{w})}\right),$$

עבור $\lambda\geq 0,\alpha\in[0,1]$ כלשהם.

א. [5 נק'] בטור הימני יש שלוש השָׂמוֹת של λ,α, שהופכות את הבעיה לעיל לבעיות רגרסיה מוּכָּרות. בטור השמאלי, יש חמש שיטות גנרטיביות. במהלך הקורס ראינו שקילות בין שתי בעיות מימין לשתי שיטות משמאל (תחת הנחות מסוימות על הדאטה). את השקילות השלישית עליכם להסיק לבד.

מלאו את המשבצות הריקות של שלוש הבעיות בטור הימני במספרים של שלוש השיטות השקולות להן בטור השמאלי.

a. הבעיה לעיל כאשר $\lambda=0$ ☐

b. הבעיה לעיל כאשר $\lambda>0,\alpha=0$ ☐

c. הבעיה לעיל כאשר $\lambda>0,\alpha=1$ ☐

i. שערוך MLE

ii. שערוך Naïve Bayes

iii. שערוך MAP עם Gaussian prior

iv. שערוך MAP עם Binomial prior

v. שערוך MAP עם Laplace prior

מעתה נניח $\boldsymbol{\lambda > 0, \alpha \in (0,1)}$ בלבד.

ב. [12 נק'] בסעיף זה נניח שיש $d = 2$ פיצ'רים והם **זהים** זה לזה, כלומר $\forall i \in [m]:\ \mathrm{x}_i[1] = \mathrm{x}_i[2]$ (וכאמור, $\lambda > 0, \alpha \in (0,1)$).

(a) לכל $\mathbf{w} = \begin{bmatrix} \mathrm{w}[1] \\ \mathrm{w}[2] \end{bmatrix} \in \mathbb{R}^2$ נגדיר את $\widetilde{\mathbf{w}} = \begin{bmatrix} \mathrm{w}[2] \\ \mathrm{w}[1] \end{bmatrix}$.

הוכיחו: תחת ההנחות, לכל $\mathbf{w} \in \mathbb{R}^2$ מתקיים $\mathcal{L}(\mathbf{w}; S) + R_{\lambda,\alpha}(\mathbf{w}) = \mathcal{L}(\widetilde{\mathbf{w}}; S) + R_{\lambda,\alpha}(\widetilde{\mathbf{w}})$.

הוכחה:

להלן טענת עזר שניתן להשתמש בה במידת הצורך בהמשך השאלה.

<u>טענת עזר</u>: פונקציית הרגולריזציה $R_{\lambda,\alpha}$ היא **קמורה במובן החזק** (strictly convex). משמע, מתקיים באי-שיוויון חזק:

$$\forall \mathbf{w}_1 \neq \mathbf{w}_2 \in \mathbb{R}^d, \forall \beta \in (0,1):\ \ \beta R_{\lambda,\alpha}(\mathbf{w}_1) + (1-\beta) R_{\lambda,\alpha}(\mathbf{w}_2) > R_{\lambda,\alpha}(\beta \mathbf{w}_1 + (1-\beta)\mathbf{w}_2).$$

(b) <u>הוכיחו</u>: תחת ההנחות (2 פיצ'רים זהים, $\lambda > 0, \alpha \in (0,1)$), שני מקדמי הפתרון האופטימלי זהים.

משמע, מתקיים: $\mathrm{w}^\star[1] = \mathrm{w}^\star[2]$.

הוכחה:

להלן משפט. **הבינו אותו היטב.** בפרט, השתכנעו שהוא מכליל את שהוכחנו בסעיף הקודם.

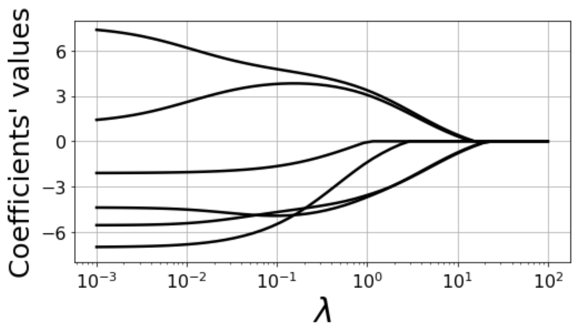
<u>משפט</u>[1]: תחת הנחות (לא מגבילות במיוחד) על הנתונים, כל שני מקדמים של פתרון ה-Elastic Net מקיימים:

$$\forall a,b\in[d]:\ \left(|\mathbf{w}^{\star}[a]-\mathbf{w}^{\star}[b]|\le\tfrac{1}{\lambda\alpha}\sqrt{\left(1-\rho_{a,b}\right)\cdot c}\right),$$

כאשר $\rho_{a,b}\triangleq\frac{\mathrm{Cov}(\mathbf{x}[a],\mathbf{x}[b])}{\sigma_a\sigma_b}\in[-1,1]$ הוא מקדם המִתְאָם (קורלציה) של פִּירסון בין הפיצ'רים a,b (על פני כל הדוג') ו-$c>0$ הוא קבוע כלשהו.

ג. [8 נק'] עבור dataset עם 6 פיצ'רים, אימנו ElasticNet עם ערכי $\alpha\in\{0.01,0.5,0.99\}$ וערכי $\lambda>0$ שונים. כל עקומה בגרפים מראה את ערכיו של מקדם <u>יחיד</u> (מתוך 6) בפתרון שנוצר מצירוף של α (ברמת הגרף) ו-λ (בציר x).
<u>נתון</u>: ישנם שני זוגות של פיצ'רים עם קורלציה גבוהה (בכל זוג, קורלציה גבוהה בין שני הפיצ'רים של אותו זוג).

ליד כֹּל גרף סמנו את ערך α המתאים לו ביותר והסבירו <u>בקצרה</u> את בחירותיכם. היעזרו במשפט לעיל בהסברים.

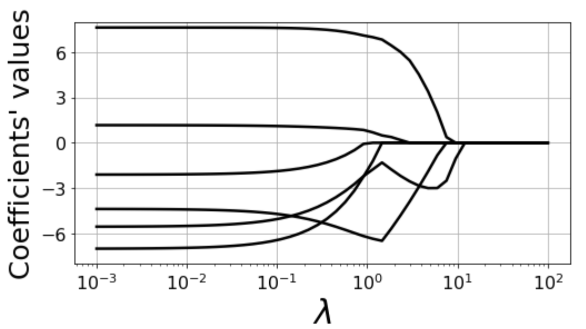


לאיזה ערך α מתאים הגרף? 0.01 / 0.5 / 0.99

הסבר: ______________________________

____

____

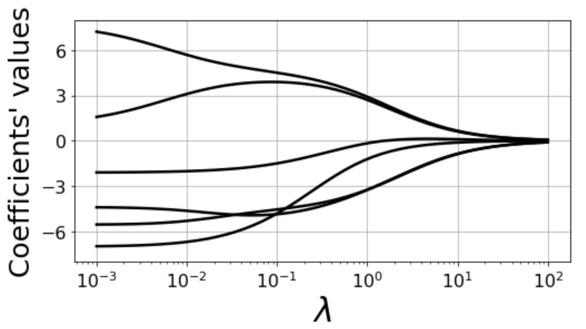


לאיזה ערך α מתאים הגרף? 0.01 / 0.5 / 0.99

הסבר: ______________________________

____

____



לאיזה ערך α מתאים הגרף? 0.01 / 0.5 / 0.99

הסבר: ______________________________

____

____

[1] גרסה פשטנית של משפט מהמאמר [Zou and Hastie. 2005] "Regularization and variable selection via the elastic net"

שאלה 4: מסווגים ליניאריים [25 נק']

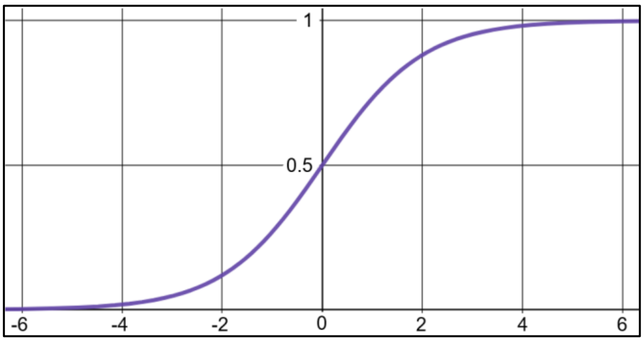
תזכורת:

כלל החלטה $h: \mathbb{R}^d \to \{-1,+1\}$ נקרא ליניארי אם ורק אם קיימים $\mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}$ כך ש-$\forall \mathbf{x} \in \mathbb{R}^d: \ h(\mathbf{x}) = \operatorname{sgn}(\mathbf{w}^\top \mathbf{x} + b)$.

נגדיר: $\operatorname{sgn}(z) = \begin{cases} -1, & z \le 0 \\ +1, & z > 0 \end{cases}$

(כך שלא מתקבל אפס לשום קלט)

תזכורת: $\sigma(z) = \frac{1}{1+\exp\{-z\}}$



ובפרט מתקיים:

$\sigma(0) = 0.5, \quad \sigma(1) \approx 0.73, \quad \sigma(2) \approx 0.88, \quad \sigma(3) \approx 0.95$

$\sigma(4) \approx 0.98, \quad \sigma(5) \approx 0.99, \quad \sigma(6) \approx 1$

א. [5 נק'] הוכיחו שכלל ההחלטה של logistic regression (עבור $\mathbf{w}_1 \in \mathbb{R}^d, b_1 \in \mathbb{R}$ כלליים) הוא ליניארי:

$$h_1(\mathbf{x}) = \begin{cases} +1, & \sigma(\mathbf{w}_1^\top \mathbf{x} + b_1) > 0.5 \\ -1, & \text{אחרת} \end{cases}.$$

הוכחה:

להלן טענת עזר שניתן להשתמש בה במידת הצורך בהמשך התרגיל.

<u>טענת עזר</u>: כלל החלטה ליניארי h צריך לקיים $\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d:\ \underbrace{h(\mathbf{x}_1) = h(\mathbf{x}_2)}_{\text{if}} \Longrightarrow \underbrace{h\left(\frac{1}{2}\mathbf{x}_1 + \frac{1}{2}\mathbf{x}_2\right) = h(\mathbf{x}_1)}_{\text{then}}$.

ב. [10 נק'] עבור $\mathbf{w}_3 \in \mathbb{R}^d, b_3 \in \mathbb{R}$ כלליים כלשהם.

<u>הפריכו</u>: כלל ההחלטה הבא ליניארי: $h_3(\mathbf{x}) = \begin{cases} +1, & \sigma(\mathbf{w}_3^\top \sigma(\mathbf{x}) + b_3) > 0.5 \\ -1, & \text{אחרת} \end{cases}$,

כאשר הפעולה $\sigma(\mathbf{x})$ מחזירה וקטור $[\sigma(x[1]), \sigma(x[2]), \dots, \sigma(x[d])]^\top$.

<u>רמז</u>: כדאי להשתמש בטענת העזר. ניתן לבחור $\mathbf{w}_3, b_3$ ספציפיים כדי להפריך.

הפרכה (לרשותכם דפי טיוטה בסוף הגיליון):

ג. [10 נק'] נתונים $u_2 \in \mathbb{R}$, $\mathbf{w}_2 \in \mathbb{R}^d$ כלליים כלשהם ומתקיים $u_2 > 1$.

הוכיחו/הפריכו: כלל ההחלטה הבא ליניארי: $h_2(\mathbf{x}) = \begin{cases} +1, & \sigma(u_2\sigma(\mathbf{w}_2^\top \mathbf{x}) - 1) > 0.5 \\ -1, & \text{אחרת} \end{cases}$.

תשובה:

חלק ב' – שאלה אמריקאית [6 נק']

בשאלה הבאה סמנו את התשובות המתאימות (לפי ההוראות). בחלק זה אין צורך לכתוב הסברים.

א. בתרגילי הבית ראינו איך לבצע tuning על היפר-פרמטרים רבים בעזרת Grid search cross validation.
נתון אלגוריתם למידה $\mathcal{A}$ עם אוסף $\mathcal{P}$ של היפר-פרמטרים שדורשים tuning. נניח שלכל היפר-פרמטר יש בדיוק v ערכים שנבדקים. נסמן את מספר ה-folds בתהליך ה-CV בתור k ואת מספר דוגמאות האימון בתור m.

$x_1 \; \lambda_2 \ldots \lambda_{10}$ 1   5   100

סמנו את **<u>כל</u>** הטענות הנכונות בהכרח ביחס לריצה של תהליך ה-Grid search CV:

a. מספר הקריאות ל-$\mathcal{A}$ גדל בצורה ליניארית ב-$|\mathcal{P}|$.

b. מספר הקריאות ל-$\mathcal{A}$ גדל בצורה ליניארית ב-v.

c. מספר הקריאות ל-$\mathcal{A}$ גדל בצורה ליניארית ב-k.

d. סיבוכיות זמן הריצה של קריאה בודדת ל-$\mathcal{A}$ גדלה בצורה ליניארית ב-m.

$A = P \cdot |V| \cdot K$

$m(|k-1)\, V^{|P|}$

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):