מבוא למערכות לומדות (236756)
סמסטר חורף תשע"ה

מבחן סוף סמסטר – מועד א'

מרצה: ניר אילון

מתרגל: יובל דגן

הנחיות
1. במבחן זה 28 עמודים כולל עמוד זה.
2. משך המבחן שלוש שעות (180 דקות).
3. כל חומר עזר אסור לשימוש.
4. ניתן לרשום בעפרון או בעט בצבעים כחול או שחור.
5. כל התשובות יכתבו על טופס הבחינה, **ויש להחזירו בתום הבחינה**
6. יש לענות על כל השאלות.
7. **יש לענות אך ורק בתוך משבצות התשובה.**
8. אין חובה למלא את כל משבצת התשובה – לעיתים היא תהייה גדולה רק בהרבה מהנדרש.
9. יש להקיף או לסמן את האפשרות הנכונה, **ולא לבצע סימון כלשהוא על אפשרויות לא נכונות.**
10. אנא כתבו בכתב יד ברור וקריא. **תשובה בכתב יד שאינו קריא לא תיבדק.**
11. **נא לכתוב רק את מה שהתבקשתם—אין צורך בהסברים או פרטים נוספים.**

פינת האנגלית הטובה:

תכונה	**Feature**
תיוג	**Label**
כל המוסיף גורע!	**More is less!**

12. נא לא לתלוש עמודים מטופס הבחינה.

בהצלחה!

דף נוסחאות

1. $\binom{n}{k} \leq n^k$
2. שגיאת הכללה = true error = $L_{\mathcal{D}}^{01}$.
3. $e \approx 2.72$
4. תהא $\mathcal{H}$ מחלקת היפותזות של בעיית למידה כלשהי, ו- S קבוצת אימון שנבחרת באקראי. נסמן

$$\hat{h} = argmin_{h\in\mathcal{H}} L_D^{01}(h)$$
$$h^* = argmin_{h\in\mathcal{H}} L_S^{01}(h)$$

אזי, לכל $\delta > 0$: בהסתברות של לפחות $1-\delta$ מתקיים:

$$L_D^{01}(\hat{h}) \leq L_D^{01}(h^*) + O\left(\sqrt{\frac{VCDIM(\mathcal{H}) + \frac{1}{\log(\delta)}}{|S|}}\right)$$

שאלה 1

בכל אחד מהסעיפים הבאים עליכם לסמן "נכון" או "לא נכון". במקומות המיועדים, השיבו **בקצרה** על השאלות המילוליות.
שימו לב: ישנן תשובות נכון/לא נכון שדורשות הסבר מילולי, וכאלה שאינן דורשות הסבר כזה.

1. ידוע שקיים אלגוריתם הפותר ERM מעל מסווגים לינאריים ביחס לשגיאת 0/1 בזמן פולינומיאלי במימד ובמספר הדוגמאות.

☐ נכון

שם האלגוריתם:

☒ לא נכון NPC

2. שגיאת ה- hinge מהווה חסם עליון על שגיאת ה- 0/1.

☒ נכון

☐ לא נכון

3. השגיאה הריבועית מהווה חסם עליון על שגיאת ה- 0/1.

☐ נכון

הוכחה קצרה:

☐ לא נכון

דוגמא נגדית:

4. אם K_1 היא פונקצית גרעין (Kernel) וכן K_2 פונקצית גרעין, אז הפונקציה $K_1 + K_2/3$ היא פונקציית גרעין.

☒ נכון

נימוק:

☐ לא נכון

נימוק:

5. אלגוריתם ה-backpropagation הוא סוג של ירידת גראדיאנט (gradient descent) ולכן מוביל תמיד לפיתרון אופטימלי באימון כל רשת עיצבית המשתמשת בפונקציית אקטיבציה גזירה כגון סיגמויד (sigmoid).

☐ נכון

☒ לא נכון

נימוק: מינימום לוקאלי

6. שיטת בייז נאיבית (Naïve Bayes) מעל וקטור מאפיינים בינאריים נותן, כפלט, מפריד לינארי.

☐ נכון

☐ לא נכון

דוגמא נגדית:

7. שיטת k-nearest neighbor תמיד עובדת יותר טוב (מבחינת שגיאת הכללה) ככל ש- k יותר גדול (וכל שאר הפרמטרים נשארים קבועים).

☐ נכון

נימוק:

☒ לא נכון

דוגמא נגדית:

8. שיטת k-nearest neighbor תמיד עובדת יותר טוב (מבחינת שגיאת הכללה) ככל ש- k יקטן יותר (וכל שאר הפרמטרים נשארים קבועים).

☐ נכון

נימוק:

☒ לא נכון

דוגמא נגדית:

9. הגדלת מחלקת ההיפותזות עבור בעיית למידה נתונה תמיד מגדילה את הסיבוכיות החישובית של CONSISTENT.

☐ **נכון**

נימוק:

☐ **לא נכון**

דוגמא נגדית:

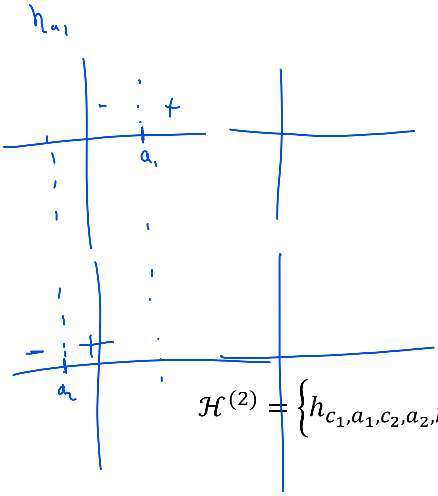
10. הקטנת מחלקת ההיפותזות עבור בעייה נתונה (כלומר, הוצאת היפותזות מהמחלקה) יכולה רק להקטין את גודל קבוצת האימון הדרושה לצורך השגת שגיאת הכללה ε באמצעות אלגוריתם CONSISTENT (במיקרה ה-Realizable).

☐ **נכון**

נימוק:

☐ **לא נכון**

דוגמא נגדית:



שאלה 2

בשאלה הנוכחית, $\mathcal{X} = \mathbb{R}$ ו- $\mathcal{Y} = \{-1,1\}$. נגדיר לכל $a \in \mathbb{R}$ פונקציה

$$h_a(x) = \begin{cases} +1, & x \geq a \\ -1, & x < a \end{cases}$$

נגדיר את המחלקה

$$\mathcal{H}^{(2)} = \left\{ h_{c_1,a_1,c_2,a_2,b}(x) = \begin{cases} 1, & c_1 h_{a_1}(x) + c_2 h_{a_2}(x) \geq b \\ -1, & otherwise \end{cases} : c_1, c_2, a_1, a_2, b \in \mathbb{R} \right\}$$

$-h_2 + h_4 \geq$

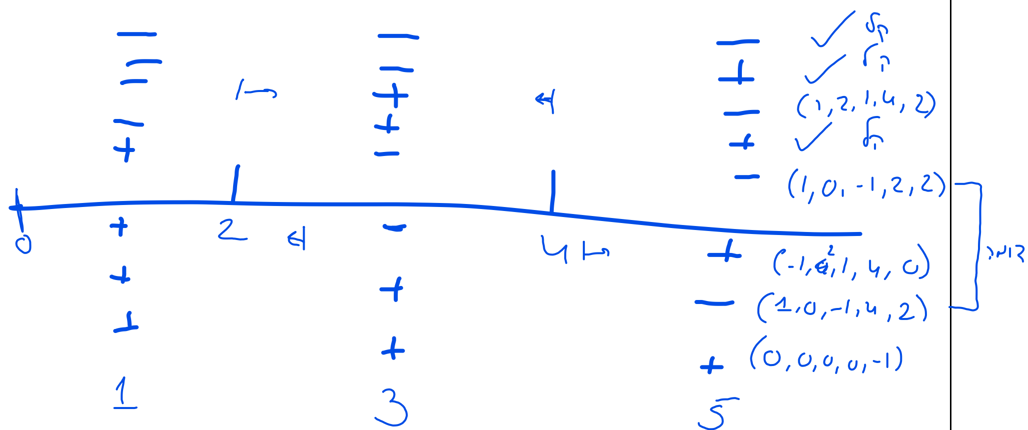
1. מהו $VCDIM(\mathcal{H}^{(2)})$? הוכיחו את תשובתכם.
דגשים לפתרון: נסמן את התשובה ב- D. בהוכחה עליכם להראות קבוצת D נקודות ולשכנע שניתן לנתץ אותה (ניתן להיעזר בסימטריות. במהלך ההוכחה אין צורך למצוא את c_1, a_1, c_2, a_2, b מפורשות, אלא להסביר איך מוצאים אותם). הסבירו מדוע לא ניתן לנתץ קבוצות בגודל D+1.

פתרון:

$VCdim(H^{(2)}) = 3$

ניתן להפריד ל 3 חלקים ולתת לכל חלק ערך חיובי/שלילי
כלומר $VCdim=3$ ולא נצליח להחזיק חנה' שנין ולנתק אותן לפי

$c_1 h_{\alpha_1}(x) + c_2 h_{\alpha_2}(x) \geq b$



$(c_1, \alpha_1, c_2, \alpha_2, b)$

$(1,2,1,4,2)$

$(1,0,-1,2,2)$

$(-1,2,1,4,0)$

$(1,0,-1,4,2)$

$(0,0,0,0,-1)$

$h_2 - h_4(x) \geq 2$

$x=1: (-1)-(-1)=0$

$x=3: 1-(-1)=2$

$5: 1-1=0$

$\alpha_1=2 \quad \alpha_2=4$

$c_1=1 \quad c_2=-1$

$\alpha_1=0 \quad \alpha_2=4$

$c_1=1 \quad c_2=-1$

$h_0(x) - 2h_4(x) \geq 2$

$x=1: 1+2=3$

$x=3: 1+2=3$

$x=5: 1-2=-1$

$\alpha_1=2 \quad \alpha_2=4 \quad b=0$

$c_1=-1 \quad c_2=1$

$-h_2(x)+h_4(x) \geq 0$

$x=1: -(-1)+(-1)=0$

$3: -1+(-1)=-2$

$5: -1+1=0$

2. עבור $k \in \mathbb{N}$ כלשהו, נגדיר את המחלקה:

$$\mathcal{H}_{+}^{(k)} = \left\{ h_{c_1,a_1,\ldots,c_k,a_k,b}(x) = \begin{cases} 1, & c_1 h_{a_1}(x) + \cdots + c_k h_{a_k}(x) \geq b \\ -1, & otherwise \end{cases} : a_1, \ldots, a_k, b \in \mathbb{R}, c_1, \ldots, c_k \geq 0 \right\}$$

שימו לב שכאן $c_1, \ldots, c_k$ הם אי שליליים.

מהו מימד ה- VC של $\mathcal{H}_{+}^{(k)}$? הוכיחו (אותם הדגשים של הסעיף הקודם תקפים גם כאן).

פתרון: VCdim = 1

3. נגדיר $\mathcal{H} = \{s \cdot h_a(x): s \in \{\pm 1\}, a \in \mathbb{R}\}$.
עבור משימת סיווג כלשהי, נלקחו קבוצת אימון וקבוצת מבחן באקראי, המסומנות S_{train}, S_{test}. השתמשו ב- Adaboost עם T איטרציות ולומד חלש שבהינתן התפלגות $\mathcal{D}$, מחזיר את $argmin_{h\in\mathcal{H}} L_{\mathcal{D}}^{01}(h)$. נלמדה היפותזה $\hat{h}$, והתקיים ש- $L_{S_{test}}^{01}(\hat{h})$ גבוה מדי. הוצעו כמה הצעות לשיפור:

א. הגדלת S_{train}
ב. הקטנת S_{train}
ג. הגדלת T
ד. הקטנת T
ה. שינוי הלומד החלש כך שיחזיר את $argmin_{h\in\mathcal{H}^{(2)}} L_{\mathcal{D}}^{01}(h)$ (כלומר החלפת $\mathcal{H}$ ב- $\mathcal{H}^{(2)}$)

אילו מההצעות שהוצעו הן בעלות פוטנציאל להוריד את $L_{S_{test}}^{01}(\hat{h})$ בצורה משמעותית כאשר הן מבוצעות לבדן, בכל אחד מהמקרים הבאים?

i. $L_{S_{train}}^{01}(\hat{h})$ קטן מאוד, ו- $L_{S_{test}}^{01}(\hat{h})$ גדול. overfitting

סמנו את כל התשובות הנכונות:
א ☑ ב ☒ ג ☒ ד ☑ ה ☒

ii. $L_{S_{train}}^{01}(\hat{h})$ ו- $L_{S_{test}}^{01}(\hat{h})$ שניהם גדולים ובעלי ערך כמעט זהה. under

סמנו את כל התשובות הנכונות:
א ☒ ב ☒ ג ☑ ד ☒ ה ☑

שאלה 3

1. נתונים $x_1, \dots, x_m$ משתנים מקריים בלתי תלויים מהתפלגות $Poisson(\lambda)$. התפלגות זו מוגדרת עבור פרמטר $\lambda > 0$ מעל הערכים הטבעיים $\{0,1,2,\dots\}$, כאשר הסיכוי לערך x הוא

$$\frac{\lambda^x e^{-\lambda}}{x!}$$

חשבו משערך MLE ל- λ (ביטוי של $x_1, \dots, x_m$), והראו את צעדי החישוב:

תשובה סופית:

$$\hat{\lambda} = \left[\qquad\qquad\qquad \right]$$

חישוב:

2. נתונה בעיית פרדיקציה שבה $\mathcal{X} = \{0,1,2,\ldots\}^2$ ו- $\mathcal{Y} = \{0,1\}$. הוחלט לפתור אותה באמצעות שיטת בייז נאיבית, תחת ההנחה שלכל $i = 1,2$: $x[i]$ מתפלג $Poisson(\lambda_0)$ בהינתן ש $y = 0$, ואחרת $x[i]$ מתפלג $Poisson(\lambda_1)$.

a. תחת ההנחה בשאלה זו, חשבו את ההסתברות הבאה (כפונקציה של λ_0, λ_1):

$$\Pr[(x[1] = 5 \wedge x[2] = 3\ )|\ \ y = 1] = \left[\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\right]$$

b. נתונה קבוצת האימון הבאה $(x_1 = (0,2), y_1 = 0), (x_2 = (2,4), y_2 = 1)$
איזו פרדיקציה מתאימה ל- $x = (2,2)$ אם משתמשים בשיטת MLE לשיערוך λ_0, λ_1, ובנוסף מניחים ש- $\Pr(y = 1) = 0.5$? הראו את החישובים בפתרון.

פתרון:

שאלה 4

הערה: במהלך השאלה אין צורך להתייחס להסתברות השגיאה, δ. ניתן להחשיבה כקבוע (O(1)).

רופאת ילדים מעוניינת לנבא את הסיכוי להופעת אפילפסיה עד גיל 12, בהינתן התיק הרפואי של הילד/ה בגיל 4. כל תיק כזה כולל d מאפיינים בינאריים (מספרים ב- $\{0,1\}$) שהרופאה סבורה שקשורים שמחלת האפילפסיה.

1. בשלב הראשון, הרופאה מנסה לבצע למידת PAC באמצעות ERM. לשם כך היא מחליטה על מחלקת ההיפותזות $\mathcal{H}_k = \{h: \{0,1\}^d \to \{0,1\}: h \text{ depends on } \leq k \text{ input coordinates}\}$. במילים – זו מחלקת הפונקציות הבינאריות שהפלט שלהן תלוי לכל היותר ב- k משתנים. לדוגמא, הפונקציה $x \mapsto x[1] \vee x[2] \vee x[7]$ שייכת ל $\mathcal{H}_3$ בעוד שהפונקציה הבאה לא שייכת ל $\mathcal{H}_5$:
   $x \mapsto (x[3] \wedge x[7] \wedge x[9] \wedge x[11]) \vee (x[4] \wedge x[1] \wedge x[13])$
   רישמו חסם עליון טוב ככל שתוכלו למספר הדוגמאות שלהן תיזדקק הרופאה כדי למצוא היפותזה ששגיאת ההכללה שלה לכל היותר ε מעל שגיאת ההכללה האופטימלית במחלקה. יש לבטא את החסם כפונקציה של d, k, ε. (מותר לכתוב חסם אסימפטוטי למשל (...)O).

תשובה (אין צורך להסביר):

2. לרופאה לא היו מספיק דוגמאות כדי לבצע את הרעיון שבסעיף הקודם, עבור k, ε שנראו לה מתאימים. מתמחה א' הציע לרופאה לוותר על $d/2$ מאפיינים כלשהם. מתמחה ב' הציע לרופאה להחליף את k ב- k/2. איזו מבין שתי ההצעות צפויה להוריד בצורה יותר משמעותית את כמות הדוגמאות הדרושות?

תשובה: ☐ מתמחה א' ☐ מתמחה ב'
הסבר קצר:

3. נסמן ב- $\mathcal{H}_k'$ את קבוצת כל הפונקציות הבינאריות שמקבלות k משתנים בינאריים, כלומר, קבוצת הפונקציות מ- $\{0,1\}^k$ ל- $\{0,1\}$.

הרופאה החליטה לא להקשיב למתמחים, ובמקום זאת ללכת בדרך דו שלבית.

- שלב א': להפעיל שיטת feature selection לבחירת k מאפיינים.
- שלב ב': ללמוד באמצעות ERM פונקציה בינארית מעל $\mathcal{H}_k'$.

לשם בחירת המאפיינים, היא השתמשה בפונקציה בשם voodoo_selector הפועלת כך:

- קלט: $S, k, magic$.
  - S היא קבוצה של בדיוק 666 דוגמאות מתויגות
  - k הוא מספר המאפיינים הדרושים
  - $magic$ הוא פרמטר מספרי בתחום $\{1,2,\dots,k^3\}$ המשפיע על הפלט של voodoo_selector בצורה מסתורית.
- פלט: k אינדקסים של מאפיינים טובים (בתקווה).

הרופאה כתבה קוד שמנסה את כל הערכים האפשריים של $magic$. עזרו לה להשלים את הקוד, וענו על השאלה שמופיעה מיד אחריו.

הערה: המשיכו להניח שהרופאה שואפת לקבל שגיאת הכללה של ε מעל שגיאת ההכללה של הפיתרון האופטימלי ב- $\mathcal{H}_k$ (בהנחה ש- voodoo_selector טוב).

\# Input: k, ε

\# Output: A hypothesis

מותר לכתוב ביטויים אסימפטוטיים, למשל $\Theta(\dots)$. אין חובה למלא את כל המשבצות הריקות.

[ ]

$m = [\quad]$

[ ]

obtain sample $S = ((x_1, y_1), \dots, (x_{m+666}, y_{m+666}))$

for $magic = 1 \dots k^3$:

$(f[1], \dots, f[k]) = voodoo\_selector\big(\big((x_{m+1}, y_{m+1}), \dots, (x_{m+666}, y_{m+666})\big), k, magic\big)$

define $x_i' = \big(x_i\big[f[1]\big], \dots, x_i\big[f[k]\big]\big)$ for all $i = 1 \dots m$

$$h = ERM\left(\mathcal{H}_k', \left(\left(x'_{[\quad]}, y_{[\quad]}\right), \dots, \left(x'_{[\quad]}, y_{[\quad]}\right)\right)\right)$$

$$err = L^{01}\left(h, \left(\left(x'_{[\quad]}, y_{[\quad]}\right), \dots, \left(x'_{[\quad]}, y_{[\quad]}\right)\right)\right)$$

[ ]

end for

return [ ]

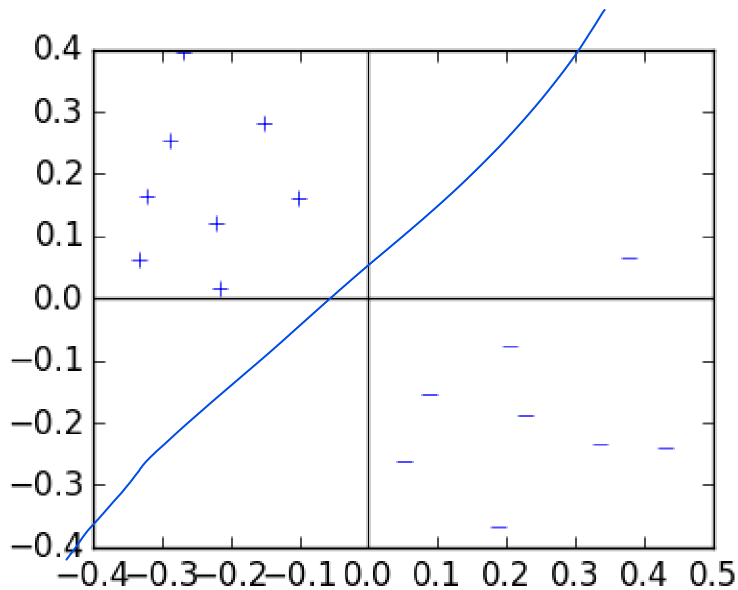
נא לכתוב הצדקה לאיתחול של m:

שאלה 5

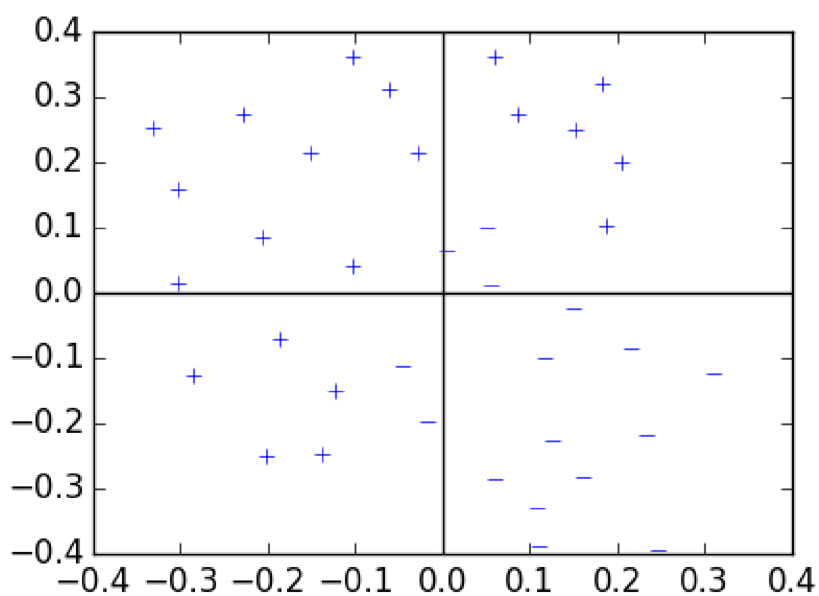
ליוסי היו שתי בעיות סיווג בינארי שונות, שעבור שתיהן $\mathcal{X} = \mathbb{R}^2$.

בכל אחת מהבעיות, הוא אסף דוגמאות למידה באקראי.

בעייה א' עוסקת בחיזוי העדפת כריכים של סטודנטים. הדוגמאות נראות כך:



בעייה ב' עוסקת בחיזוי העדפת קורסים של סטודנטים. הדוגמאות נראות כך:



ניזכר בהגדרות הבאות:

$$L_S^{hinge}(w) = \frac{1}{|S|}\sum_{(x,y)\in S} \max\{0,1-\langle w,x\rangle y\}$$

$$L_{S,\lambda}^{hinge}(w) = \frac{\lambda}{2}\left||w|\right|^2 + L_S^{hinge}$$

$$L_S^{01}(w) = \frac{1}{|S|}\sum_{(x,y)\in S} \begin{cases} 1, & sign(\langle w,x\rangle) \neq y \\ 0, & otherwise \end{cases}$$

בכל אחת משתי הבעיות, יוסי חילק את הדוגמאות לשתי קבוצות שוות בגודלן: קבוצת אימון S_{train} וקבוצת מבחן S_{test}. הוא ניסה למצוא $w \in \mathbb{R}^2$ שממזער את $L_{S_{train},\lambda}^{hinge}$.

לשם כך, השתמש ב- sub gradient descent, עם גודל צעד η. כלומר הוא ביצע:

- $w^{(0)} \leftarrow random \in \mathbb{R}^2$
- $for\ t = 1, \ldots, 250$
  - $\mathrm{w}^{(t)} \leftarrow w^{(t-1)} - \eta \nabla L_{S_{train},\lambda}(w^{(t-1)})$

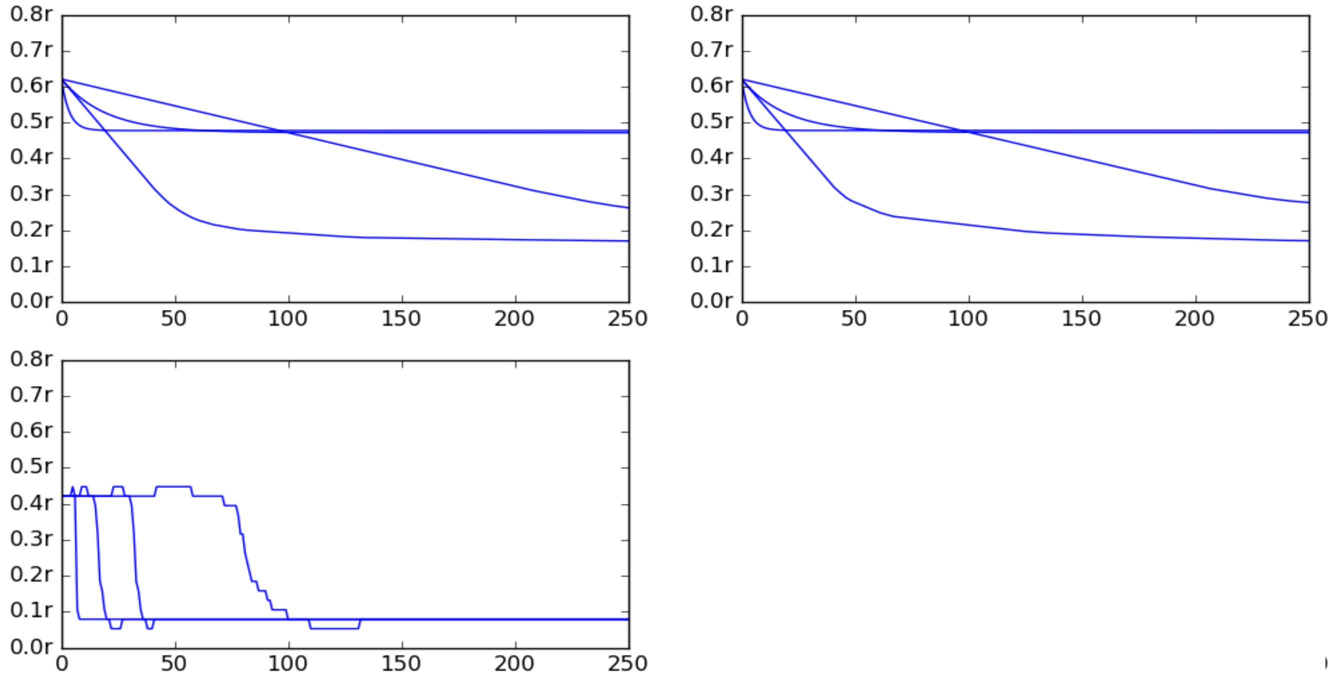
לכל אחת משתי בעיות הסיווג ביצע 4 הרצות, כך שבכל הרצה הוא השתמש בפרמטרים שונים.

הפרמטרים בהם השתמש:

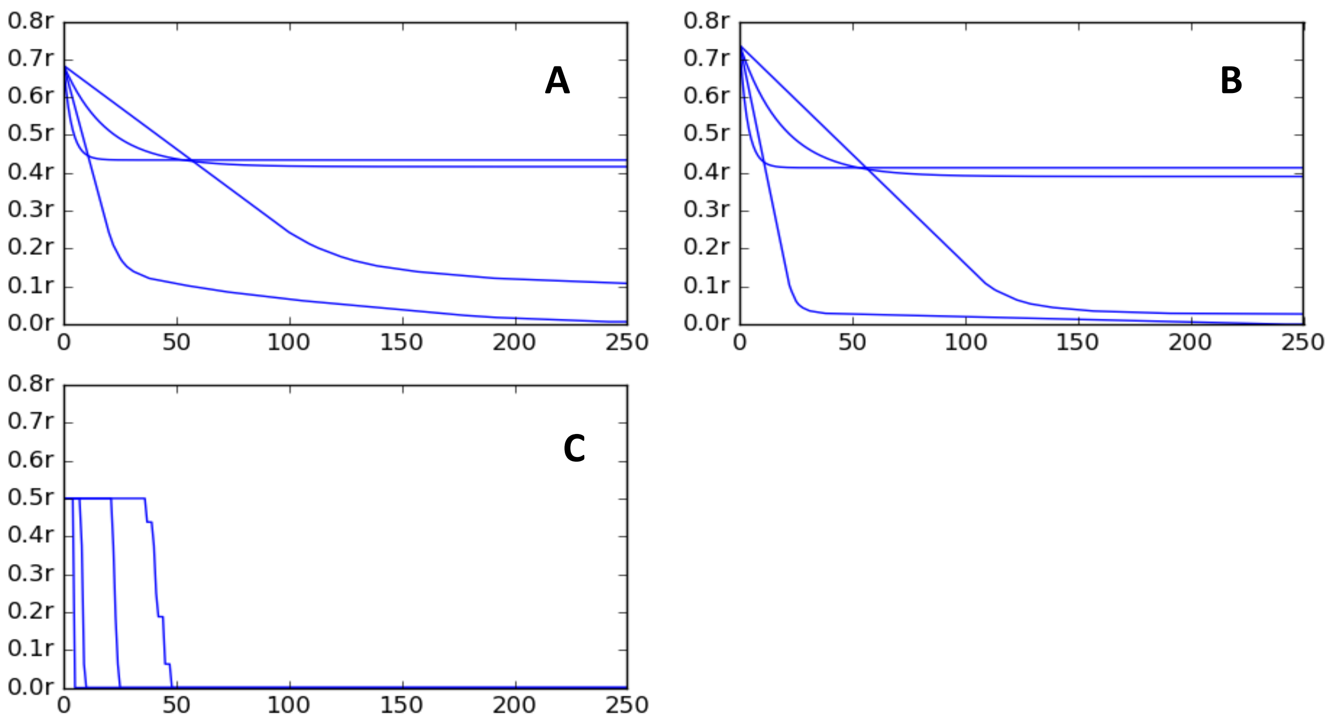
- $\lambda = 0, \eta = 0.1$
- $\lambda = 0.5, \eta = 0.1$
- $\lambda = 0, \eta = 0.5$
- $\lambda = 0.5, \eta = 0.5$

יוסי הדפיס שני דוחות, כאשר אחד הדוחות מתייחס לבעיית חיזוי הכריכים, והאחר לבעיית חיזוי הקורסים. לא ידוע איזה דו"ח מתייחס לאיזו בעיה (הסברים עבור הדוחות יינתנו בהמשך).

דו"ח ראשון:



דו"ח שני:



נשים לב שכל דו"ח מכיל 3 תרשימים:

- אחד התרשימים מחשב את $L^{hinge}_{S_{train}}(w^{(t)})$ כפונקציה של t.
- תרשים אחר מחשב את $L^{hinge}_{S_{test}}(w^{(t)})$ כפונקציה של t.
- תרשים אחר מחשב את $L^{01}_{S_{train}}(w^{(t)})$ כפונקציה של t.

כל תרשים מכיל 4 גרפים. הגרפים השונים מתייחסים להרצות עם ערכים שונים לפרמטרים λ ו- η.

המספר r הנמצא בציר האנכי הוא קבוע מספרי חיובי כלשהו, הזהה בכל התרשימים

ענו על השאלות הבאות:

1. לאיזו בעיה מתייחס הדו"ח הראשון? נמקו.

☐ בעיית הכריכים ☑ בעיית הקורסים

נימוק: לא פריד לינארית - נגיע לשגיאה שונה 0
מפול (גם באימון וגם בטסט)

2. נביט בדו"ח השני. התאימו בין התרשימים, לפונקציות שהם מתייחסים אליהן, ונמקו:

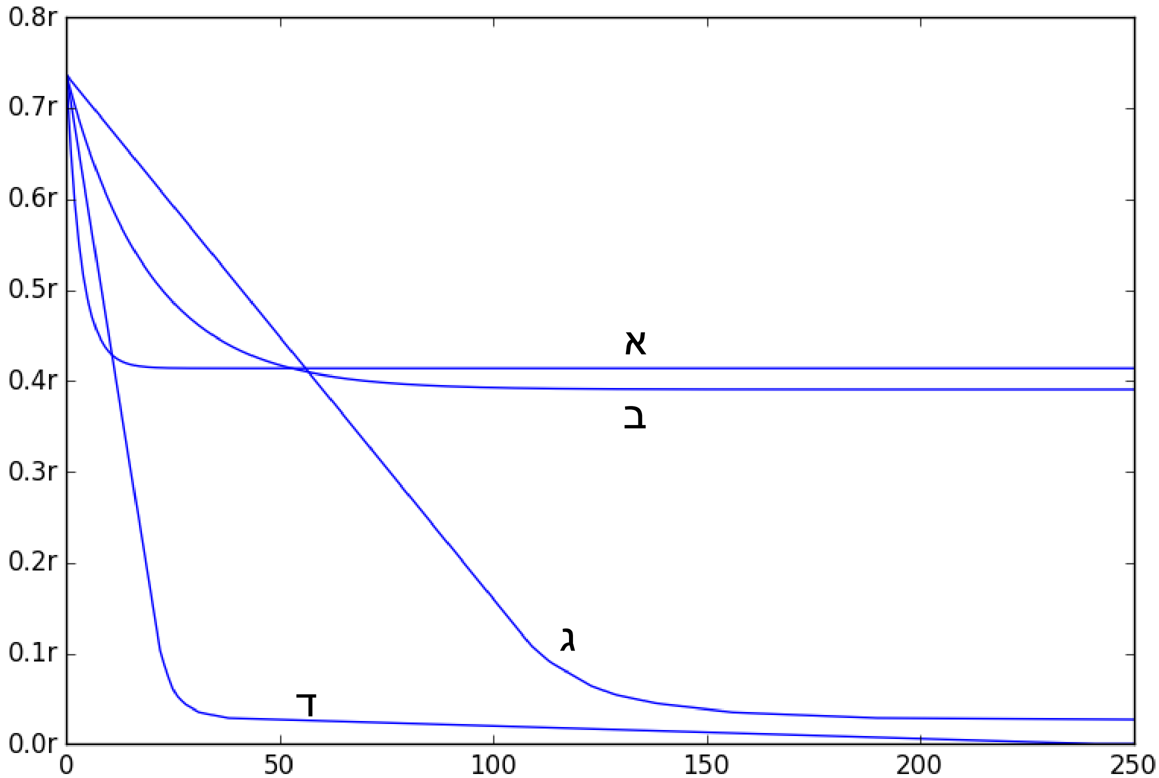
$L_{S_{train}}^{hinge}(w^{(t)})$ מתייחס לתרשים ☐ C ☑ B ☐ A

$L_{S_{test}}^{hinge}(w^{(t)})$ מתייחס לתרשים ☐ C ☐ B ☑ A

$L_{S_{train}}^{01}(w^{(t)})$ מתייחס לתרשים ☑ C ☐ B ☐ A

נימוק:

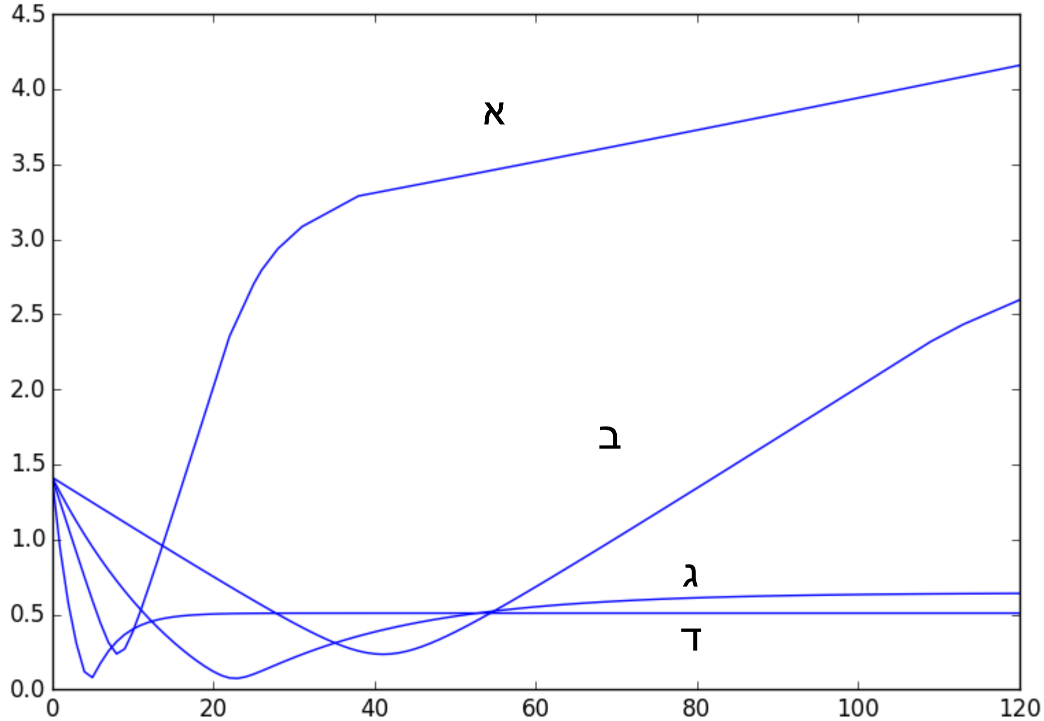
3. נביט בדו"ח השני, בתרשים B. לשם הנוחות, עותק מוגדל שלו נמצא כאן:



הגרפים השונים מסומנים באותיות א, ב, ג, ד. התאימו בין ערכי הפרמטרים השונים לגרפים השונים.

$\lambda = 0, \eta = 0.1$	א ☐	ב ☐	ג ☑	ד ☐
$\lambda = 0.5, \eta = 0.1$	א ☐	ב ☑	ג ☐	ד ☐
$\lambda = 0, \eta = 0.5$	א ☐	ב ☐	ג ☐	ד ☑
$\lambda = 0.5, \eta = 0.5$	א ☑	ב ☐	ג ☐	ד ☐

4. נמצא תרשים נוסף בדו"ח השני:



הוא מתאר את הנורמה של $w^{(t)}$ כתלות ב- t. התאימו בין ערכי הפרמטרים השונים לגרפים השונים שבתרשים זה:

$\lambda = 0, \eta = 0.1$	☐ א	☑ ב	☐ ג	☐ ד
$\lambda = 0.5, \eta = 0.1$	☐ א	☐ ב	☑ ג	☐ ד
$\lambda = 0, \eta = 0.5$	☑ א	☐ ב	☐ ג	☐ ד
$\lambda = 0.5, \eta = 0.5$	☐ א	☐ ב	☐ ג	☑ ד