הטכניון – מכון טכנולוגי לישראל

הפקולטה למדעי המחשב

מבוא למערכות לומדות (236756)

סמסטר חורף תשפ"ד – 12 באפריל 2024

מרצה: ד"ר יונתן בלינקוב

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.
- קריאות:
  - סימונים לא ברורים בשאלות רב-ברירה ו/או תשובות מילוליות בכתב יד לא קריא יובילו לפסילת התשובה.
  - לא יתקבלו ערעורים בנושא.
- במבחן 14 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.**
- **לזכאים להערכה חלופית מתאפשרת בחירה בין שאלות 3 ו-4.**

בהצלחה!

שאלה 1: Maximum Likelihood Estimation [12 נק']

נתונות m דגימות חד-ממדיות $S = \{x_1, x_2, \dots, x_m\}$ שנדגמו i.i.d. ממשתנה מקרי $X \sim \mathrm{Geo}(p)$ עבור $p \in (0,1]$ לא ידוע.

נזכיר שמשתנה גאומטרי מאפיין את מספר ה"ניסויי ברנולי" (הטלות מטבע) בסיכוי p עד להשגת "הצלחה" אחת ומאופיין ע"י פונקציית ההסתברות הבאה:

$$\Pr(X = x_i | p) = (1-p)^{x_i - 1} p.$$

שערכו (estimate) את הפרמטר p באמצעות Maximum Likelihood Estimation.

משמע, עליכם לפתח ביטוי עבור $\hat{p} = \mathrm{argmax}_p\, L(S;p) = \mathrm{argmax}_p \Pr(x_1, x_2, \dots, x_m; p)$.

תשובה:

ln מונוטוני

$$\hat{p} = \mathrm{argmax}_p L(S;p) = \mathrm{argmax}_p \ln L(S;p) = \mathrm{argmax}_p \ln \Pr(x_1, x_2, \dots, x_m; p) =$$

$$= \mathrm{argmax}_p \ln \prod_{i=1}^{m} \Pr(x_i; p) = \mathrm{argmax}_p \sum_{i=1}^{m} \ln\left[(1-p)^{x_i - 1} \cdot p\right] =$$

$$= \mathrm{argmax}_p \sum_{i=1}^{m} (x_i - 1)\ln(1-p) + \ln(p)$$

$$\nabla_p = \sum_{i=1}^{m}\left(-\frac{x_i - 1}{1-p} + \frac{1}{p}\right) = \frac{m}{p} + \frac{\sum 1 - x_i}{(1-p)} = 0 \Rightarrow \frac{m}{p} + \frac{m}{1-p} = \frac{\sum x_i}{(1-p)} \Rightarrow \frac{m(1-p) + mp}{p(1-p)} = \frac{\sum x_i}{(1-p)}$$

$$\Rightarrow \frac{m}{p\cancel{(1-p)}} = \frac{\sum (x_i)}{\cancel{1-p}} \Rightarrow \frac{1}{p} = \frac{1}{m}\sum x_i \Rightarrow \bar{X} = \frac{1}{p} \Rightarrow \boxed{p = \frac{1}{\bar{X}}}$$

שאלה 2: VC-dimension of Decision Trees [24 נק']

א. [4 נק'] להלן ההגדרה של "ניתוץ". השמטנו מההגדרה את הַכַּמָּתִים.

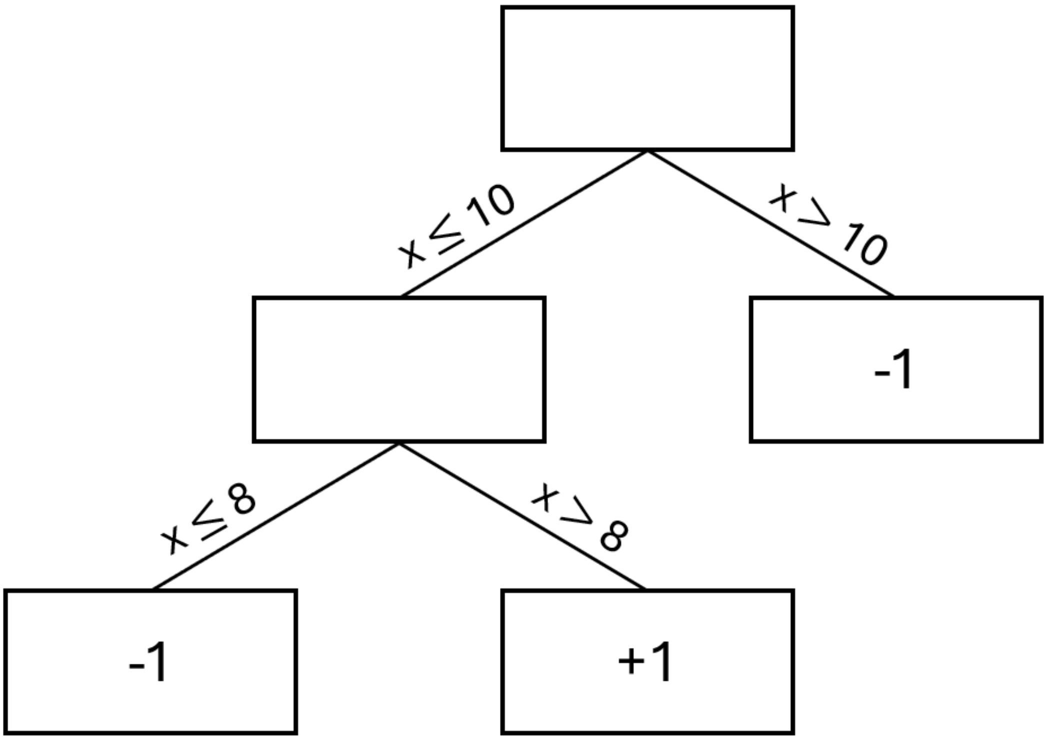
השלימו את שלושת הכמתים החסרים. בכל מקום כתבו <u>בבירור</u> האם חסר בהגדרה ∀ או ∃.

$$\mathcal{H}\text{ shatters } C \iff \underbrace{\forall}_{\text{השלימו}} y_1,\dots,y_{|C|}\in\mathcal{Y}:\ \underbrace{\exists}_{\text{השלימו}} h\in\mathcal{H}:\ \underbrace{\forall}_{\text{השלימו}} x_i\in C:\ h(x_i)=y_i$$

מעתה והלאה בשאלה זו מרחב הדוגמאות הוא $\mathcal{X}=\mathbb{R}$ ומרחב התיוגים הוא $\mathcal{Y}=\{-1,+1\}$.

מחלקת ההיפותזות $\mathcal{H}$ עליה נשאל הינה עצי החלטה מעומק $L\geq 1$.

<u>תזכורת</u>: עצי ההחלטה בקורס הינם עצים בינאריים בהם הצמתים הפנימיים מתפצלים לפי thresholds. העומק של העץ לא כולל את השורש אבל כן את העלים. לדוגמה, להלן מוצג עץ החלטה "חוקי" מעומק 2.



ב. [10 נק'] עבור עומק כללי $L \ge 1$, הוכיחו שהחסם התחתון הוא $\mathrm{VCdim}(\mathcal{H}) \ge 2^L$.

הוכחה פורמלית:

יהי $y_1,\dots,y_{2^{|L|}} \in \mathcal{Y}$ תיוג כלשהו כאשר הם ממוינים לפי ערכי ה x_i שלהם כך ש $x_1 < x_2 < \dots < x_{2^{|L|}}$

בכל רמת עומק נפריד (כמו בחיפוש בינארי/חיפוש בינארי) לפי הערך האמצעי בתת העץ שלו (ימינה/רמה ראשונה) $y_1,\dots,y_{2^{|L|-1}}$ החלקים שמאלה

ו $y_{2^{|L|-1}+1},\dots,y_{2^{|L|}}$ החלקים ימינה) וכך הלאה עד שבסוף אנו מקבלים עלה עם כל תיוג בעלה משל עצמו!

לשם כך עבור כל y_i ~~=1~~ נגדיר את הסיווג להיות תואם בהתאם

נשים לב כי יש לנו לכל היותר $2^{|L|}$ נק' שמוגדרות בעץ אחד בעץ (.........) ולכן עומק העץ

הדרוש הוא $|L|$ כלומר קיים עץ כך שלכל הנק' קיימת הפרדה ב $\mathcal{H}$

שמשיגה את כל התיוגים האפשריים $\Leftarrow$ $\mathcal{H}$ מנתצת את $2^{|L|}$ $\Leftarrow$ $\mathrm{VCdim}(\mathcal{H}) \ge 2^{|L|}$

ג. [10 נק'] עבור עומק כללי $L \geq 1$, הוכיחו שהחסם העליון הוא $\text{VCdim}(\mathcal{H}) \leq 2^L$.

הוכחה פורמלית:

כלומר נב"ש: $VCdim(\mathcal{H}) < 2^L+1$, יהי סידור נק' כלשהו $x_1, x_2, \dots, x_{2^L+1}$ בה"כ ממויינות מהקטן

לגדול כך: $x_1 < x_2 < \dots < x_{2^L+1}$ ונתייג אותן כך: $\begin{cases} y_i = +1, & i \% 2 = 0 \\ y_i = -1, & i \% 2 = 1 \end{cases}$

ונניח כי h' מסווגת את כולן נכון, כלומר $h'(x_i) = y_i$ עבור $\forall i$ $2^{L}+1$ נק'

ברמת התחתונה יש לנו 2^{L} עלים אז לפי עקרון שובך היונים קיים עלה שבו לפחות 2 דוגמאות

לפי תכונה על העלה שבררנו מוגדים כי הדוגמאות עוקבות בסידור כלומר x_i, x_{i+1} וגם

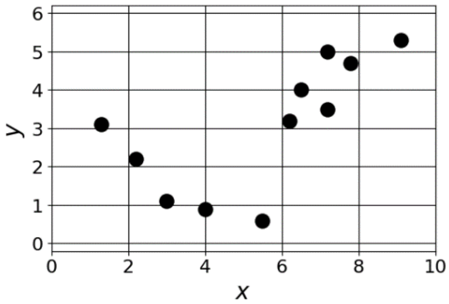
כלומר $y_i = y_{i+1}$ בסתירה לכך ש $y_i \neq y_{i+1}$ מתפרת h' ולכן עבור כל סידור x_i מצאנו

תיוג y_i כך שאף המסווג במחלקה לא מסווגת את כל הנק' נכון $\Rightarrow$ לא מנותצת $\Rightarrow VCdim(\mathcal{H}) \leq 2^L$

שאלה 3: Linear Regression [32 נק']

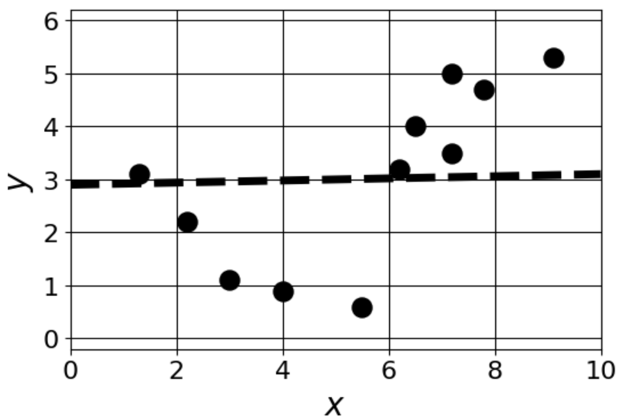
☐ לזכאים להערכה חלופית בלבד (כפי שהוגדרו באתר הקורס): סמנו את התיבה הזו אם ברצונכם **לדלג** על שאלה זו. המשקל של שאלות 1,2,4 יתפזר באופן יחסי על פני 100 נקודות. ניתן לדלג רק על שאלה אחת מתוך שאלות 3,4.

בתרשים שלפניכם מופיע מדגם אימון חד-ממדי $\mathcal{X} = \mathbb{R}, \mathcal{Y} = \mathbb{R}$.

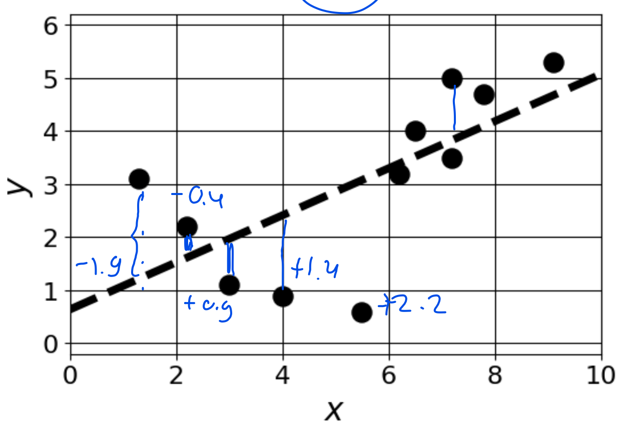


א. [2 נק'] לומדים רגרסיה לינארית לא הומוגנית (עם Least squares רגיל) על המדגם הנתון. הקיפו את האות של התרשים שמתאים לקו הרגרסיה שנלמד.

(A)

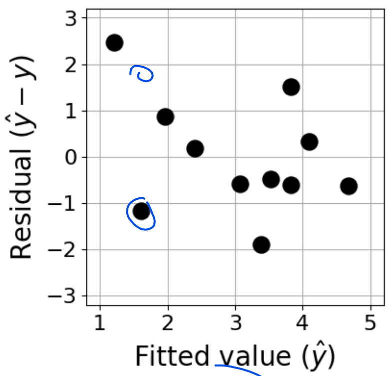


(B)

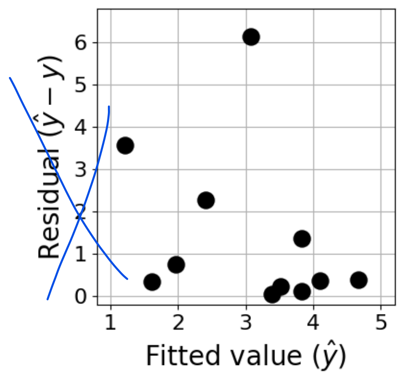


ב. [4 נק'] להלן תרשימי Residual analysis מהסוג שהוצג בהרצאה (שימו לב לצירי ה-y השונים). הקיפו את האות של התרשים היחיד שמתאים לנתונים ולקו הרגרסיה מהסעיף הקודם.

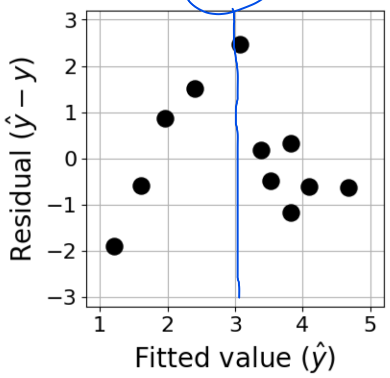
(A)



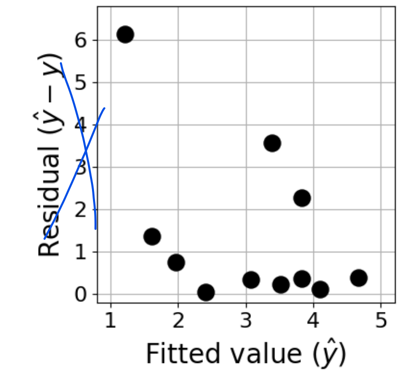
(B)



(C)



(D)

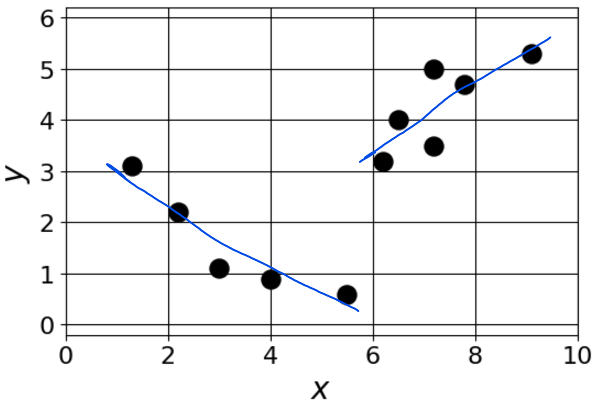


נגדיר מחלקת Thresholded linear regression המאפשרת התאמה של <u>שני</u> קווי רגרסיה נפרדים:

$$\mathcal{H}_1 = \{h_{w_1,b_1,w_2,b_2,\tau} \mid w_1, b_1, w_2, b_2, \tau \in \mathbb{R}\}, \quad \text{where} \quad h_{w_1,b_1,w_2,b_2,\tau}(x) = \begin{cases} w_1 x + b_1, & x < \tau \\ w_2 x + b_2, & x \geq \tau \end{cases}$$

כמו במקרה הלינארי, נגדיר היפותזה <u>אופטימלית</u> ב-$\mathcal{H}_1$ בתור היפותזה שמביאה למינימום את ה-MSE על הדאטה.

משמע מושגת שגיאה: $\min_{w_1,b_1,w_2,b_2,\tau\in\mathbb{R}} \frac{1}{m}\sum_{i=1}^{m}\left(h_{w_1,b_1,w_2,b_2,\tau}(x_i) - y_i\right)^2$.



ג. [2 נק'] ציירו על גבי התרשים שמשמאל היפותזה אופטימלית ב-$\mathcal{H}_1$ (זאת שאלת הבנה, לא נדרשת בציור רמת דיוק של מחשב).

ד. [8 נק'] ברשותכם אלגוריתם קופסה שחורה LS המקבל אוסף שרירותי של דוגמאות חד-ממדיות והתיוגים שלהן ופותר את בעיית ה-Least squares הלינארית הלא-הומוגנית הרגילה שלמדנו בקורס.
עבור דאטה חד-ממדי כללי (ולא רק למדגם הנתון בסעיפים הקודמים), הציעו אלגוריתם למידה המוצא פרמטרים w_1, b_1, w_2, b_2, τ שמייצגים היפותזה אופטימלית ב-$\mathcal{H}_1$. אין צורך להוכיח את נכונות האלגוריתם.

תשובה:

כאשר $x_1 < \dots < x_k$

לקבוע את $T = x_1, \dots, x_k$ לפי הסדר וכל פעם להריץ LS על

שני החלקים בנפרד ($x < T$, $T \le x$) ולסכם טעויות ולשמור, לבסוף להחזיר את

הערך T שהוביל לשגיאות LS מינימליות ולהחזיר את ההיפותזה שלו ביחד עם T זה

כעת, ניצור מחלקה דומה המכילה רק את הפונקציות הרציפות מ-$\mathcal{H}_1$. נעשה זאת ע"י הגדרת $\mathcal{H}_2$ עם פרמטריזציה שונה:

$$\mathcal{H}_2 = \left\{h'_{u_1,c_1 u_2,c_2} \,\middle|\, u_1, c_1 u_2, c_2 \in \mathbb{R}\right\}, \quad \text{where} \quad h'_{u_1,c_1 u_2,c_2}(x) = u_1 \max\{0, x - c_1\} + u_2 x + c_2$$

ה. [8 נק'] הוכיחו (מתמטית) שמתקיים $\mathcal{H}_2 \subseteq \mathcal{H}_1$.

$H_1 = \{h_{w_1,b_1,w_2,b_2,T} \mid w_1,b_1,w_2,b_2,T \in \mathbb{R}\}, w \quad h_?(x) = \begin{cases} w_1 x + b_1, & x<T \\ w_2 x + b_2, & x \geq T \end{cases}$

הוכחה:

$$h'_{(..)}(x) = u_1 \max\{0, x - c_1\} + u_2 x + c_2 = \begin{cases} u_2 x + c_2 & x < c_1 \\ u_1 x - u_1 c_1 + u_2 x + c_2 & x \geq c_1 \end{cases} =$$

$$= \begin{cases} u_2 x + c_2, & x < c_1 \\ (u_1+u_2)x + c_2 - u_1 c_1, & x \geq c_1 \end{cases}$$

יהי $h'_{u_1,..,c_2} \in H_2$, אזי

ונבחר $T = c_1$, $b_1 = c_2$, $w_1 = u_2$, $b_2 = c_2 - u_1 c_1$, $w_2 = u_1 + u_2$

כלומר

$$= \begin{cases} w_1 x + b_1, & x < T \\ w_2 x + b_2, & x \geq T \end{cases} \Rightarrow h'_{u_1,\dots,c_2} \in H_1 \Rightarrow H_2 \subseteq H_1$$

ו. [8 נק'] נתון דאטה חד-ממדי עם תיוגים רציפים $(\mathbf{x}_i \in \mathbb{R},\ y_i \in \mathbb{R})$ מחולק באופן אקראי לסט אימון ולסט מבחן.
פותרים את בעיית הרגרסיה על סט האימון ע"י מציאת היפותזות אופטימליות ב-$\mathcal{H}_1$ וב-$\mathcal{H}_2$ (ביחס ל-Training MSE).
כל שורה בטבלה מתארת תוצאות מתהליך הלמידה המתואר על דאטה נתון כלשהו.
חלק מהשורות מתארות תוצאות אפשריות וחלק מתארות תוצאות בלתי אפשריות.
לכל אחת מארבע השורות, סמנו האם היא אפשרית או לא.

$H_2 \subseteq H_1$

Optimal hypothesis of $\mathcal{H}_1$		Optimal hypothesis of $\mathcal{H}_2$		
Training MSE	**Test MSE**	**Training MSE**	**Test MSE**	**האם אפשרית?**
17	55	25	33	לא / כן
17	35	15	37	לא / כן
17	35	42	51	לא / כן
0	11	0	9	לא / כן

שאלה 4: נושאים שונים בפרספטרון וב-SVM [32 נק']

☐ לזכאים להערכה חלופית בלבד (כפי שהוגדרו באתר הקורס): סמנו את התיבה הזו אם ברצונכם **לדלג** על שאלה זו. המשקל של שאלות 1,2,3 יתפזר באופן יחסי על פני 100 נקודות. ניתן לדלג רק על שאלה אחת מתוך שאלות 3,4.

לפניכם אלגוריתם הפרספטרון (ההומוגני) המקבל כקלט מדגם אימון $S=\{(\mathbf{x}_i,y_i)\}_{i=1}^m$, כאשר $\mathcal{X}=\mathbb{R}^d$ וְ-$\mathcal{Y}=\{-1,+1\}$.

```
w = 0_d
errorFound = True

for epoch=1 to EPOCHS_NUM:
  errorFound = False

  for i=1 to m:
    ŷ_i = sign(wᵀx_i)

    if y_i != ŷ_i:        y_i wᵀx_i < 1
      w = w + y_i x_i
      errorFound = True

  if not errorFound:
    break
```

ראינו בתרגול שפרספטרון שקול ל-SGD על $\sum_i \max\{0, -y_i\, \mathbf{w}^\top \mathbf{x}_i\}$. כעת נראה תכונה דומה ל-Margin Perceptron.

נגדיר $\mathcal{L}(\mathbf{w}) = \sum_i \ell(\mathbf{w}; \mathbf{x}_i, y_i)$ כאשר $\ell(\mathbf{w}; \mathbf{x}_i, y_i) = \max\{0,1 - y_i\, \mathbf{w}^\top \mathbf{x}_i\}$. נזכיר שהפונקציה $\ell(\mathbf{w}; \mathbf{x}_i, y_i)$ קמורה ב-$\mathbf{w}$.

א. [5 נק'] הציעו subgradient של ℓ לפי $\mathbf{w}$, משמע, $\nabla_{\mathbf{w}} \ell(\mathbf{w}; \mathbf{x}_i, y_i)$ (כתבו רק תשובה סופית).

תשובה:

$$\nabla_w \begin{cases} 1 - y_i w^T x_i & ; y_i w^T x_i < 1 \\ 0 & ; y_i w^T x_i \ge 1 \end{cases} = \begin{cases} -y_i x_i & ; y_i w^T x_i < 1 \\ 0_d & , y_i w^T x_i \ge 1 \end{cases}$$

ב. [9 נק'] הציעו שינוי נקודתי לאלגוריתם הפרספטרון (ציינו את מספרי השורות שמתעדכנות ואת השינויים הנדרשים) כך שהאלגוריתם המעודכן יהיה שקול ל-SGD על $\mathcal{L}(\mathbf{w})$ שהגדרנו. יש להסביר בקצרה כיצד השקילות מתקיימת.

תשובה:

שורה 10: $y_i w^T x_i < 1$

$$w = w + \overset{1}{\eta} y_i x_i = \begin{cases} w + y_i x_i & ; y_i w^T x_i < 1 \\ w + 0 & ; y_i w^T x_i \ge 1 \end{cases} =$$

$$= w + \begin{cases} y_i x_i & ; y_i w^T x_i < 1 \\ 0 & ; y_i w^T x_i \ge 1 \end{cases} = w - \begin{cases} -y_i x_i & , y_i w^T x_i < 1 \\ 0_d & , y_i w^T x_i \ge 1 \end{cases} = \underbrace{w - \eta \nabla_w \ell(\,)}_{SGD}$$

<u>שימו לב:</u> ניתן לענות על הסעיפים הבאים גם מבלי לפתור את הסעיפים הקודמים.

ללא תלות בסעיפים הקודמים, נחזור לעסוק באלגוריתם הפרספטרון הרגיל כפי שמופיע בתחילת השאלה.

<u>משפט ההתכנסות של הפרספטרון</u>:

יהי דאטה פריד לינארית הומוגנית $S=\{(\mathbf{x}_i,y_i)\}_{i=1}^{m}$. נסמן את הנורמה המרבית של דוגמה ע"י $R \triangleq \max_i \|\mathbf{x}_i\|_2$ (סקלר).

יהי $\mathbf{w}_{\text{SVM}}$ פיתרון ה-Hard SVM על הדאטה, משמע: $\mathbf{w}_{\text{SVM}} \triangleq \operatorname{argmin}_{\mathbf{w}\in\mathbb{R}^d}\|\mathbf{w}\|_2 \ s.t.\ \forall i: y_i\mathbf{w}^\top\mathbf{x}_i \geq 1,$.

אזי, אלגוריתם הפרספטרון יעשה לכל היותר $(R\cdot\|\mathbf{w}_{\text{SVM}}\|_2)^2$ טעויות חיזוי בזמן האימון שלו.

ג. [9 נק'] נתון דאטה $S=\{(\mathbf{x}_i,y_i)\}_{i=1}^{100}$ פריד לינארית הומוגנית ש"הבטחת" המשפט עבורו הינה $(R\cdot\|\mathbf{w}_{\text{SVM}}\|_2)^2=10$.
יוצרים דאטה חדש ע"י הכפלת <u>כל</u> הדוגמאות ב-2, משמע $S'=\{(2\mathbf{x}_i,y_i)\}_{i=1}^{100}$.
מהי הבטחת המשפט לדאטה החדש? משמע, כמה טעויות חיזוי (לכל היותר) יעשה אלגוריתם הפרספטרון על הדאטה החדש? כתבו ערך <u>מספרי</u> מפורש והסבירו <u>בקצרה</u> את תשובתכם.

תשובה:

$W_{svm} = \operatorname{argmin} \|W\|_2 \ s.t\ \forall i: y_i \cancel{2} W^T 2x_i \geq 1$

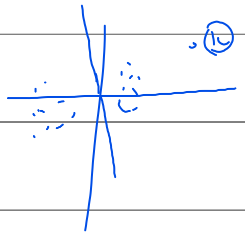
$y_i (2w^T) x_i \geq 1$

$W^* = \frac{1}{2} W_{svm}$ $\quad (2R \cdot \|\frac{1}{2} W_{svm}\|_2)^2 = (R\|W_{svm}\|)^2 = 10$

$R^* = 2R$

ד. [9 נק'] נתון דאטה $S=\{(\mathbf{x}_i,y_i)\}_{i=1}^{100}$ פריד לינארית הומוגנית ש"הבטחת" המשפט עבורו הינה $(R\cdot\|\mathbf{w}_{\text{SVM}}\|_2)^2=10$.
נסיר את הדוגמה האחרונה ונקבל דאטה חדש $S'=\{(\mathbf{x}_i,y_i)\}_{i=1}^{99}$. נסמן את החסם המעודכן בעזרת $(R'\cdot\|\mathbf{w}'_{\text{SVM}}\|_2)^2$.
בטענה שלפניכם, כתבו את היחס ($\geq,\leq,=$) המדויק <u>ביותר</u> כך שתהיה נכונה <u>בהכרח</u>: $10 \underbrace{\leq}_{\text{השלימו}} (R'\cdot\|\mathbf{w}'_{\text{SVM}}\|_2)^2$.

<u>הסבירו בקצרה</u>:



א) הורדנו תצפ' הכי רחוקה ⇐ $R'<R$ וקטן המרג'ין ⇐ $new<10$

ב) הורדנו נק' קרובה למ"ב ⇐ R נשאר זהה ⇐ נותרת רק נק' כי ניתן להגדיל מרג'ין ⇐ $new<10$

ג) הורדנו נק' לא תומכת ⇐ R זהה נותר זהה ⇐ ושאר שווה $new=10$

לא ניתן להיות המספר גדול יותר ע"י הורדת נק', רק ע"י הוספת נק' ניתן

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):