# תרגיל בית 5 – מבוא ללמידה חישובית

## מגיש: נתן בלוך

**Theory Questions.**

**שאלה 1. Suboptimality of ID3.**

**סעיף א.** נבצע את אלגוריתם ה-$ID3$ על מנת לקבל עץ החלטה עבור סט האימון הנתון, ונראה כי טעות האימון הינה לפחות $\frac{1}{4}$.
נגדיר את הקבוצות הבאות – $A=\{1,2,3\}$, כאשר כל אינדקס $i$ ב-$A$ מייצג את הפרדיקט $h_i(x) \ \coloneqq \ (x_i == i)$. כמו כן נגדיר את סט האימון –

$$S=\{((1,1,1),1),((1,0,0),1),((1,1,0),0),((0,0,1),0)\}$$

נתחיל בביצוע האלגוריתם $BuildTree(S,A)$ כפי שהוצג בהרצאה ובתרגול. ראשית, ב-$S$ דגימות עם $labels$ שונים, וכן $A$ אינה ריקה, ולכן נעבור לחישוב ה-$Gain(S,i)$ ומציאת ה- $argmax$ המתאים. ניזכר בהגדרה –

$$Gain(S,i)=C(\mathbb{P}_S(Y=1))-(\mathbb{P}_S[X=1]\cdot C(\mathbb{P}_S[Y=1|X=1])+\mathbb{P}_S[X=0]C(\mathbb{P}_S[Y=1|X=0]))$$

כאשר $C(a)$ הינה פונק' המבוססת על פונק' האנטרופיה, כך שהיא מקיימת $C(0)=C(1)=0$, וכן $C\left(\frac{1}{2}\right)$ מקסימלי.
כמו כן נבחין כי – $C(\mathbb{P}_S(Y=1))=C\left(\frac{1}{2}\right)$. ועבור $S,A$ נקבל ערכים של –

$$\boldsymbol{Gain(S,1)}=C\left(\frac{1}{2}\right)-\left(\frac{3}{4}\cdot C\left(\frac{2}{3}\right)+\frac{1}{4}\cdot\underbrace{C(0)}_{=0}\right)=C\left(\frac{1}{2}\right)-\frac{3}{4}\cdot C\left(\frac{2}{3}\right)>0$$

שכן $C\left(\frac{1}{2}\right)>C\left(\frac{2}{3}\right)$, ולכן בפרט $C\left(\frac{1}{2}\right)>C\left(\frac{2}{3}\right)>\frac{3}{4}C\left(\frac{2}{3}\right)$ שכן $C$ אי-שלילית לפי הגדרתה. כמו כן –

$$\boldsymbol{Gain(S,2)}=C\left(\frac{1}{2}\right)-\left(\frac{1}{2}\cdot C\left(\frac{1}{2}\right)+\frac{1}{2}\cdot C\left(\frac{1}{2}\right)\right)=C\left(\frac{1}{2}\right)-C\left(\frac{1}{2}\right)=0$$

$$\boldsymbol{Gain(S,3)}=C\left(\frac{1}{2}\right)-\left(\frac{1}{2}\cdot C\left(\frac{1}{2}\right)+\frac{1}{2}\cdot C\left(\frac{1}{2}\right)\right)=C\left(\frac{1}{2}\right)-C\left(\frac{1}{2}\right)=0$$

ולפיכך $argmax_{i\in A}Gain(S,i)=1$, ולכן נבחר את הפרדיקט $(x_1==0)$ להיות הפרדיקט בשורש העץ. נגדיר כעת –

$$S_1=\{(x,y)\in S: x_1=1\}=\{((1,1,1),1),((1,0,0),1),((1,1,0),0)\},\qquad S_0=\{(x,y)\in S:\ \ x_1=0\}=\{((0,0,1),0)\}$$

וכיוון ש- $S_0\neq\emptyset$ וכן $S_1\neq\emptyset$, נמשיך רקורסיבית עם $BuildTree(S_0,\{2,3\})$ ועם $BuildTree(S_1,\{2,3\})$.

כעת, עבור $BuildTree(S_0,\{2,3\})$, נבחין כי כל ה-$labels$ של דגימות ב-$S_0$ הם 0, ולפיכך נחזיר עלה עם $label$ של 0 –


כעת, עבור $BuildTree(S_1,\{2,3\})$, נבחין כי ב-$S_1$ דגימות עם $labels$ שונים, וכן $A$ אינה ריקה, ולכן נעבור לחישוב ה-$Gain(S,i)$ ומציאת ה- $argmax$ המתאים. כמו כן נבחין כי – $C\left(\mathbb{P}_{S_1}(Y=1)\right)=C\left(\frac{2}{3}\right)$. ועבור $S_1,A_1=\{2,3\}$ נקבל ערכים של –

$$\boldsymbol{Gain(S,2)}=C\left(\frac{2}{3}\right)-\left(\frac{1}{3}\cdot\underbrace{C(1)}_{=0}+\frac{2}{3}\cdot C\left(\frac{1}{2}\right)\right)=C\left(\frac{2}{3}\right)-\frac{2}{3}C\left(\frac{1}{2}\right)$$

$$\boldsymbol{Gain(S,3)}=C\left(\frac{2}{3}\right)-\left(\frac{2}{3}\cdot C\left(\frac{1}{2}\right)+\frac{1}{3}\cdot\underbrace{C(1)}_{=0}\right)=C\left(\frac{2}{3}\right)-\frac{2}{3}C\left(\frac{1}{2}\right)$$

**כלומר קיבלנו כי $\boldsymbol{Gain(S,2)=Gain(S,3)}$, ולכן במקרה זה הפרדיקט נבחר באופן אקראי. נחלק למקרים לפי הבחירה, ונראה כי הטעות על גבי סט האימון הינה לפחות $\frac{1}{4}$ בכל אחת מהבחירות, ולפיכך טעות האימון הינה לפחות $\frac{1}{4}$.**

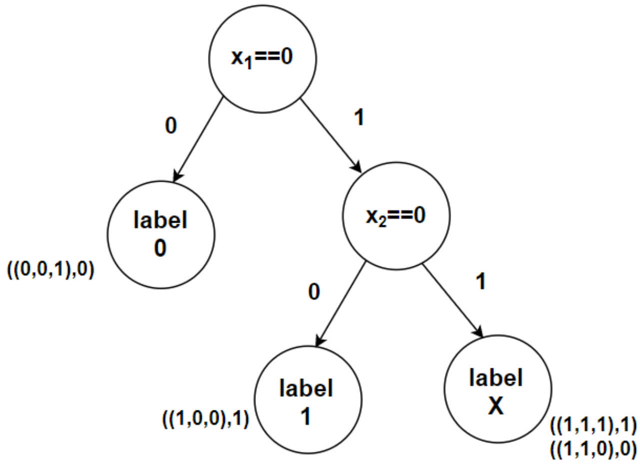
כמו כן, נזכור כי כאן עומק העץ חסום ע"י 2, ולכן בהנתן בחירה של פרדיקט ה-$labels$ ברמה הבאה ייקבעו על פי החלוקה לקבוצות שהפרדיקט משרה.

**במקרה בו נבחר בפרדיקט ($x_2 == 0$),** נבחין כי נקבל את החלוקה הבאה לקבוצות – $S_0' = \{(x,y) \in S_1 : x_2 = 0\} = \{((1,0,0),1)\}$, ולכן בעלה המתאים נקבע $label$ של 1, כך שיתאים לרוב(ובפרט במקרה זה לכל הדגימות ב-$S_0'$). בנוסף, נגדיר את הקבוצה השנייה המתאימה לפרדיקט –

$$S_1' = \{(x,y) \in S_1 : x_2 = 1\} = \{((1,1,1),1)\,,((1,1,0),0)\,\}$$

ובמקרה זה נבחין כי כל $label$ מופיע פעם אחת בדיוק, ולכן קביעת ה-$label$ תתבצע גם כן שרירותית, ונניח כי נבחר $X$.

ובכל מקרה, העץ יראה באופן הבא –



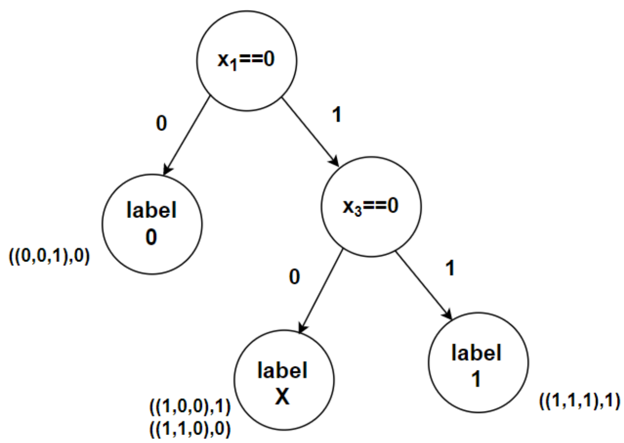

- ניתן להבחין כי העץ המתקבל על ידי אלגוריתם $ID3$ מסווג נכון את הדגימות $((0,0,1),0)$ ו-$((1,0,0),1)$.
- כמו כן, אם נבחר $label\ X = 1$, אזי העץ המתקבל יסווג נכון את הדגימה $((1,1,1),1)$ ולא יסווג נכון את הדגימה $((1,1,0),0)$, כלומר נקבל טעות אחת על סט האימון, ולכן ה-$error$ הינו $\frac{1}{4}$.
- כמו כן, אם נבחר $label\ X = 0$, אזי העץ המתקבל יסווג נכון את הדגימה $((1,1,0),0)$ ולא יסווג נכון את הדגימה $((1,1,1),0)$, כלומר נקבל טעות אחת על סט האימון, ולכן ה-$error$ הינו $\frac{1}{4}$.
- ובכל מקרה, נקבל טעות של $\frac{1}{4}$ על גבי סט האימון במקרה בו הפרדיקט הנבחר הינו $(x_2 == 0)$.

**במקרה בו נבחר בפרדיקט ($x_3 == 0$),** נבחין כי נקבל את החלוקה הבאה לקבוצות – $S_1'' = \{(x,y) \in S_1 : x_3 = 1\} = \{((1,1,1),1)\}$, ולכן בעלה המתאים נקבע $label$ של 1, כך שיתאים לרוב(ובפרט במקרה זה לכל הדגימות ב-$S_1''$). בנוסף, נגדיר את הקבוצה השנייה המתאימה לפרדיקט –

$$S_0'' = \{(x,y) \in S_1 : x_3 = 0\} = \{((1,0,0),1)\,,((1,1,0),0)\,\}$$

ובמקרה זה נבחין כי כל $label$ מופיע פעם אחת בדיוק, ולכן קביעת ה-$label$ תתבצע גם כן שרירותית, ונניח כי נבחר $X$.

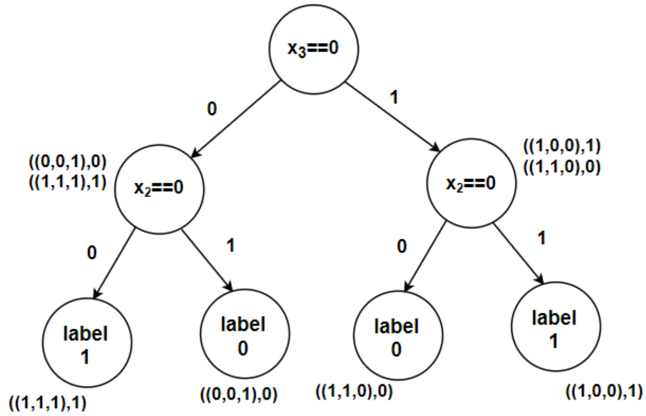
ובכל מקרה, העץ יראה באופן הבא –




- ניתן להבחין כי העץ המתקבל על ידי אלגוריתם $ID3$ מסווג נכון את הדגימות $((0,0,1),0)$ ו-$((1,1,1),1)$.
- כמו כן, אם נבחר $label\ X = 1$, אזי העץ המתקבל יסווג נכון את הדגימה $((1,0,0),1)$ ולא יסווג נכון את הדגימה $((1,1,0),0)$, כלומר נקבל טעות אחת על סט האימון, ולכן ה-$error$ הינו $\frac{1}{4}$.
- כמו כן, אם נבחר $label\ X = 0$, אזי העץ המתקבל יסווג נכון את הדגימה $((1,1,0),0)$, ולא יסווג נכון את הדגימה $((1,0,0),1)$, כלומר נקבל טעות אחת על סט האימון, ולכן ה-$error$ הינו $\frac{1}{4}$.
- ובכל מקרה, נקבל טעות של $\frac{1}{4}$ על גבי סט האימון במקרה בו הפרדיקט הנבחר הינו $(x_3 == 0)$.

**ובסה"כ, בכל מקרה אפשרי, מקבלים טעות של $\frac{1}{4}$ על גבי סט האימון, כנדרש.**

**סעיף ב. נמצא עץ החלטה בו נקבל סיווג מושלם של הנקודות בסט האימון, ובפרט נקבל שגיאה של אפס על גבי סט האימון.**




ואכן ניתן להבחין כי סיווג כל אחת מהנקודות סט האימון הינו נכון, שכן כל אחת מהדגימות מגיעה לעלה ייחודי, וכן ה-$label$ של אותו עלה הותאם לערך ה-$y$ של אותו דגימה המגיעה לעלה זה.
בנוסף, העומק של עץ ההחלטה הנ"ל הינו 2, כנדרש.

**שאלה 2. AdaBoost.** נתונות $x_1, \dots, x_m \in \mathbb{R}^d$, וכן $y_1, \dots, y_m \in \{-1,1\}$ ה-$labels$ המתאימים. נבצע את אלגוריתם $AdaBoost$ כפי שנלמד בתרגול, ונניח כי אנו באיטרציה ה-$t$, וכן נניח כי $\varepsilon_t > 0$.

**סעיף א. לא להגשה.**

**סעיף ב.** נדרש להוכיח כי מתקיים – $\Pr_{x\sim D_{t+1}}[h_t(x) \neq y)] = \frac{1}{2}$.

**פיתרון.** נכתוב את ההסתברות הרצויה בצורה המפורשת –

$$\Pr_{x\sim D_{t+1}}[h_t(x) \neq y)] = \sum_{\substack{i=1 \\ s.t. \\ h_t(x_i)\neq y_i}}^{m} D_{t+1}(x_i)$$

ראינו בתרגול את הטענה הבאה –

$$\sum_{i=1}^{m} D_{t+1}(x_i) y_i h_t(x_i) = 0$$

ונשתמש בה ע"י פיצול לפי האינדיקטור $\mathbb{I}[y_i = h_t(x_i)]$, ונקבל –

$$0 = \sum_{i=1}^{m} D_{t+1}(x_i) y_i h_t(x_i) = \sum_{\substack{i=1 \\ s.t. \\ h_t(x_i)\neq y_i}}^{m} D_{t+1}(x_i) y_i h_t(x_i) + \sum_{\substack{i=1 \\ s.t. \\ h_t(x_i)= y_i}}^{m} D_{t+1}(x_i) y_i h_t(x_i)$$

וכאשר $h_t(x_i) \neq y_i$, בהכרח $y_i h_t(x_i) = -1$, וכן כאשר $h_t(x_i) = y_i$, בהכרח $y_i h_t(x_i) = 1$, ולכן נקבל –

$$0 = \sum_{i=1}^{m} D_{t+1}(x_i) y_i h_t(x_i) = \sum_{\substack{i=1 \\ s.t. \\ h_t(x_i)\neq y_i}}^{m} D_{t+1}(x_i) \underbrace{y_i h_t(x_i)}_{=-1} + \sum_{\substack{i=1 \\ s.t. \\ h_t(x_i)= y_i}}^{m} D_{t+1}(x_i) \underbrace{y_i h_t(x_i)}_{=1} = -\sum_{\substack{i=1 \\ s.t. \\ h_t(x_i)\neq y_i}}^{m} D_{t+1}(x_i) + \sum_{\substack{i=1 \\ s.t. \\ h_t(x_i)= y_i}}^{m} D_{t+1}(x_i)$$

ומכאן ע"י העברת אגפים נקבל –

$$(*) \quad \sum_{\substack{i=1 \\ s.t. \\ h_t(x_i)\neq y_i}}^{m} D_{t+1}(x_i) = \sum_{\substack{i=1 \\ s.t. \\ h_t(x_i)= y_i}}^{m} D_{t+1}(x_i)$$

כמו כן, $D_{t+1}$ הינה **התפלגות** על הקבוצה $\{x_1, \dots, x_m\}$, ולכן בהכרח –

$$1 = \sum_{i=1}^{m} D_{t+1}(x_i) = \sum_{\substack{i=1 \\ s.t. \\ h_t(x_i)\neq y_i}}^{m} D_{t+1}(x_i) + \sum_{\substack{i=1 \\ s.t. \\ h_t(x_i)= y_i}}^{m} D_{t+1}(x_i) \underbrace{=}_{(*)} \sum_{\substack{i=1 \\ s.t. \\ h_t(x_i)\neq y_i}}^{m} D_{t+1}(x_i) + \sum_{\substack{i=1 \\ s.t. \\ h_t(x_i)\neq y_i}}^{m} D_{t+1}(x_i) = 2 \sum_{\substack{i=1 \\ s.t. \\ h_t(x_i)\neq y_i}}^{m} D_{t+1}(x_i)$$

ולכן –

$$\sum_{\substack{i=1 \\ s.t. \\ h_t(x_i)\neq y_i}}^{m} D_{t+1}(x_i) = \frac{1}{2}$$

כנדרש, שכן למעשה הראנו כי –

$$\Pr_{x\sim D_{t+1}}[h_t(x) \neq y)] = \sum_{\substack{i=1 \\ s.t. \\ h_t(x_i)\neq y_i}}^{m} D_{t+1}(x_i) = \frac{1}{2}$$

■

**סעיף ג.** נדרש להראות כי אלגוריתם $AdaBoost$ לא יחזיר את אותה היפותזה פעמיים ברצף, כלומר נדרש להראות כי $h_t \neq h_{t+1}$.

**פיתרון.** נניח בשלילה כי $h_t = h_{t+1}$. לפי ההגדרה שראינו בהרצאה, לומד חלש($WL(S,D)$) מחזיר היפותזה $h$ כך ש- $e_{S,D}(h) \leq \frac{1}{2} - \gamma$, עבור $\gamma > 0$.
כמו כן, נבחין כי $h_{t+1}$ מתקבל על ידי לומד חלש עם הפרמטרים $WL(S, D_{t+1})$, ולכן מתקיים –

$$\frac{1}{2} \underbrace{>}_{\substack{as\\ \gamma>0}} \frac{1}{2} - \gamma \geq e_{S,D_{t+1}}(h) = \sum_{i=1}^{m} D_{t+1}(x_i)\mathbb{I}[y_i \neq h_{t+1}(x_i)] =$$

$$= \sum_{\substack{1\leq i\leq m\\ s.t.\\ y_i\neq h_{t+1}(x_i)}} D_{t+1}(x_i)\underbrace{\mathbb{I}[y_i \neq h_{t+1}(x_i)]}_{=1} + \sum_{\substack{1\leq i\leq m\\ s.t.\\ y_i= h_{t+1}(x_i)}} D_{t+1}(x_i)\underbrace{\mathbb{I}[y_i \neq h_{t+1}(x_i)]}_{=0} =$$

$$= \sum_{\substack{1\leq i\leq m\\ s.t.\\ y_i\neq h_{t+1}(x_i)}} 1 \cdot D_{t+1}(x_i) + \underbrace{\sum_{\substack{1\leq i\leq m\\ s.t.\\ y_i= h_{t+1}(x_i)}} 0 \cdot D_{t+1}(x_i)}_{=0} = \sum_{\substack{1\leq i\leq m\\ s.t.\\ y_i\neq h_{t+1}(x_i)}} D_{t+1}(x_i)$$

$$= \Pr_{x\sim D_{t+1}}[h_{t+1}(x) \neq y)] \underbrace{=}_{h_{t+1}=h_t} \Pr_{x\sim D_{t+1}}[h_t(x) \neq y)]$$

כלומר בסה"כ הראנו כי – $\Pr_{x\sim D_{t+1}}[h_t(x) \neq y)] < \frac{1}{2}$, וזה בסתירה כאמור לסעיף הקודם. לפיכך, נסיק כי הנחת השלילה שגויה, כלומר ההיפותזות אינן זהות, ואכן $h_{t+1} \neq h_t$, כנדרש בסעיף זה.

**סעיף ד.** נדרש להראות כי בחירת משקלים באופן $w_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$ מביאה למינימום את הפונק' $Z_t$.

**פיתרון.** לפי ההגדרה, מתקיים –

$$Z_t = \sum_{i=1}^{m} D_t(x_i)e^{-yi\cdot w_t\cdot h_t(x_i)} = \sum_{\substack{1\leq i\leq m\\ s.t.\ \ y_i=w_t(x_i)}} D_t(x_i)e^{-w_t\cdot\overbrace{yi\cdot h_t(x_i)}^{=1}} + \sum_{\substack{1\leq i\leq m\\ s.t.\ \ y_i\neq w_t(x_i)}} D_t(x_i)e^{-w_t\cdot\overbrace{yi\cdot h_t(x_i)}^{=-1}} =$$

$$= \sum_{\substack{1\leq i\leq m\\ s.t.\ \ y_i=w_t(x_i)}} D_t(x_i)e^{-w_t} + \sum_{\substack{1\leq i\leq m\\ s.t.\ \ y_i\neq w_t(x_i)}} D_t(x_i)e^{w_t} = e^{-w_t}\underbrace{\left(\sum_{\substack{1\leq i\leq m\\ s.t.\ \ y_i=w_t(x_i)}} D_t(x_i)\right)}_{=1-\varepsilon_t} + e^{w_t}\underbrace{\left(\sum_{\substack{1\leq i\leq m\\ s.t.\ \ y_i\neq w_t(x_i)}} D_t(x_i)\right)}_{\varepsilon_t} =$$

$$= e^{-w_t}(1-\varepsilon_t) + e^{w_t}(\varepsilon_t)$$

כלומר קיבלנו כי – $Z_t = e^{-w_t}(1-\varepsilon_t) + e^{w_t}(\varepsilon_t)$, וזו כאשר נסתכל על פונק' זו כפונק של $w_t$(במשתנה יחיד), נקבל –

$$Z_t(w_t) = e^{-w_t}(1-\varepsilon_t) + e^{w_t}(\varepsilon_t)$$

ונרצה להראות כי המינימום של $Z_t(w_t)$ מתקבל עבור $w_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$, ולשם כך נגזור ונשווה לאפס על מנת למצוא נקודת קיצון –

$$Z_t'(w_t) = -e^{-w_t}(1-\varepsilon_t) + e^{w_t}(\varepsilon_t) = 0$$

$$\Longrightarrow e^{-w_t}(1-\varepsilon_t) = e^{w_t}(\varepsilon_t)$$

$$[multiply\ by\ e^{w_t}] \Longrightarrow \underbrace{e^{w_t}e^{-w_t}}_{=1}(1-\varepsilon_t) = e^{w_t}e^{w_t}(\varepsilon_t)$$

$$\Longrightarrow \left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) = e^{2w_t} \quad \Longrightarrow \quad ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) = 2w_t \quad \Longrightarrow \quad w_t = \frac{1}{2}ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$$

וזו נקודת הקיצון של הפונק' $Z_t(w_t)$, וניתן להראות כי זו נקודת מינימום ע"י כך שנראה כי $Z_t''(w_t) > 0$, ואכן ע"י גזירה פעם נוספת נקבל –

$$Z_t''(w_t) = e^{-w_t}(1-\varepsilon_t) + e^{w_t}(\varepsilon_t) > 0$$

כאשר הנכונות של כך נובעת מכך ש- $0 \leq \varepsilon_t \leq 1$, שכן מוגדר ע"י התפלגות – $\varepsilon_t := \Pr_{x\sim D_t}[h_t(x) \neq y)]$, וכן מכך שלפחות אחד מבין $\varepsilon_t, 1-\varepsilon_t$ הוא גדול ממש מאפס, ומכך שהפונק $e^x > 0$ לכל איקס, אכן מתקבל כי בהכרח $Z_t''(w_t) > 0$, ולכן ניתן להסיק כי מדובר במינימום של הפונק' $Z_t$, כנדרש.

**שאלה 3. Sufficient Condition for Weak Learnability.**

נתון סט אימון $S=\{(x_1,y_1),\dots,(x_n,y_n)\}$, ונניח כי קיים $\gamma>0$, ומסווגים $h_1,\dots,h_k\in\mathcal{H}$, ומקדמים $a_1,\dots,a_k\geq 0$ כך ש-$\sum_{i=1}^{k}a_k=1$, וכן –

$$y_i\sum_{j=1}^{k}a_jh_j(x_i)\ \geq\gamma$$

לכל $(x_i,y_i)\in S$.

**סעיף א.** נדרש להראות כי לכל התפלגות $D$ מעל $S$, קיים $1\leq j\leq k$ כך ש – $\Pr_{i\sim D}[h_j(x_i)\neq y_i]\leq\frac{1}{2}-\frac{\gamma}{2}$.

**פיתרון.** תהי התפלגות $D$ מעל $S$. נניח בשלילה כי לכל $1\leq j\leq k$ מתקיים – $\Pr_{i\sim D}[h_j(x_i)\neq y_i]>\frac{1}{2}-\frac{\gamma}{2}$.

נסתכל על הנתון $y_i\sum_{j=1}^{k}a_jh_j(x_i)\ \geq\gamma$ המתקיים לכל $1\leq i\leq n$, וניקח תוחלת על פני ההתפלגות $D$ ונקבל –

$$\gamma=\mathbb{E}_D[\gamma]\leq\mathbb{E}_D\left[y_i\sum_{j=1}^{k}a_jh_j(x_i)\right]=\sum_{i=1}^{n}\left(\mathbb{P}[\mathrm{i}]\cdot y_i\sum_{j=1}^{k}a_jh_j(x_i)\right)=\sum_{i=1}^{n}\left(\mathbb{P}[\mathrm{i}]\sum_{j=1}^{k}y_ia_jh_j(x_i)\right)=$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{k}\mathbb{P}[\mathrm{i}]y_ia_jh_j(x_i)=\sum_{j=1}^{k}\sum_{i=1}^{n}\mathbb{P}[\mathrm{i}]y_ia_jh_j(x_i)=\sum_{j=1}^{k}\left(\sum_{\substack{i=1\\ s.t.\\ y_i=h_j(x_i)}}^{n}\mathbb{P}[\mathrm{i}]a_j\underbrace{y_ih_j(x_i)}_{=1}+\sum_{\substack{i=1\\ s.t.\\ y_i\neq h_j(x_i)}}^{n}\mathbb{P}[\mathrm{i}]a_j\underbrace{y_ih_j(x_i)}_{=-1}\right)=$$

$$=\sum_{j=1}^{k}\left(\sum_{\substack{i=1\\ s.t.\\ y_i=h_j(x_i)}}^{n}a_j\mathbb{P}[\mathrm{i}]-\sum_{\substack{i=1\\ s.t.\\ y_i\neq h_j(x_i)}}^{n}a_j\mathbb{P}[\mathrm{i}]\right)=\sum_{j=1}^{k}\sum_{\substack{i=1\\ s.t.\\ y_i=h_j(x_i)}}^{n}a_j\mathbb{P}[\mathrm{i}]-\sum_{j=1}^{k}\sum_{\substack{i=1\\ s.t.\\ y_i\neq h_j(x_i)}}^{n}a_j\mathbb{P}[\mathrm{i}]=$$

$$=\sum_{\substack{i=1\\ s.t.\\ y_i=h_j(x_i)}}^{n}\sum_{j=1}^{k}a_j\mathbb{P}[\mathrm{i}]-\sum_{\substack{i=1\\ s.t.\\ y_i\neq h_j(x_i)}}^{n}\sum_{j=1}^{k}a_j\mathbb{P}[\mathrm{i}]=\sum_{\substack{i=1\\ s.t.\\ y_i=h_j(x_i)}}^{n}\mathbb{P}[\mathrm{i}]\underbrace{\sum_{j=1}^{k}a_j}_{=1}-\sum_{\substack{i=1\\ s.t.\\ y_i\neq h_j(x_i)}}^{n}\mathbb{P}[\mathrm{i}]\underbrace{\sum_{j=1}^{k}a_j}_{=1}=$$

$$=\sum_{\substack{i=1\\ s.t.\\ y_i=h_j(x_i)}}^{n}\mathbb{P}[\mathrm{i}]-\sum_{\substack{i=1\\ s.t.\\ y_i\neq h_j(x_i)}}^{n}\mathbb{P}[\mathrm{i}]=\Pr_{i\sim D}\left[h_j(x_i)=y_i\right]-\Pr_{i\sim D}\left[h_j(x_i)\neq y_i\right]$$

כלומר קיבלנו כי – $\gamma\ \leq\ \Pr_{i\sim D}\left[h_j(x_i)=y_i\right]-\Pr_{i\sim D}\left[h_j(x_i)\neq y_i\right]$, ומכאן נסיק כי –

$$\Pr_{i\sim D}\left[h_j(x_i)=y_i\right]\geq\gamma+\Pr_{i\sim D}\left[h_j(x_i)\neq y_i\right]\underbrace{>}_{\substack{assumption\\ by\ contradiction}}\gamma+\frac{1}{2}-\frac{\gamma}{2}=\boldsymbol{\frac{1}{2}+\frac{\gamma}{2}}$$

**כלומר קיבלנו כי** $\Pr_{i\sim D}\left[h_j(x_i)=y_i\right]>\frac{1}{2}+\frac{\gamma}{2}$, וכן $\Pr_{i\sim D}[h_j(x_i)\neq y_i]>\frac{1}{2}-\frac{\gamma}{2}$, אך נבחין גם כי מתקיים –

$$1=\Pr_{i\sim D}\left[h_j(x_i)\neq y_i\right]+\Pr_{i\sim D}\left[h_j(x_i)=y_i\right]>\frac{1}{2}-\frac{\gamma}{2}+\frac{1}{2}+\frac{\gamma}{2}=1$$

וזו סתירה באמור. על כן, נסיק כי הנחת השלילה שגויה, כלומר אכן קיים $1\leq j\leq k$ כך ש – $\Pr_{i\sim D}[h_j(x_i)\neq y_i]\leq\frac{1}{2}-\frac{\gamma}{2}$, כנדרש בסעיף זה.

**סעיף ב.** נגדיר $k=4d-1$, וכן נגדיר $2d$ היפותזות באופן הבא –

$$h_{j,a}(x)=\begin{cases}1 & x\geq a_j\\ -1 & else\end{cases}\qquad,\qquad h_{j,\beta}(x)=\begin{cases}1 & x\leq\beta_j\\ -1 & else\end{cases}$$

נגדיר גם $2d-1$ היפותזות קבועות, באופן הבא –

$$h_{-\infty}(x)=\begin{cases}1 & x\leq-\infty\\ -1 & else\end{cases}$$

וכן נסמן את $d$ ההיפותזות הראשונות $h_{j,a}(x)$ כ-$h_j$, עבור $j=1,\dots,d$, וכן $d$ ההיפותזות הבאות $h_{j,\beta}(x)$ כ-$h_j$, עבור $j=d+1,\dots,2d$, וכן $2d-1$ ההיפותזות הנותרות כ-$h_j$, עבור $j=2d+1,\dots,4d-1$.

בנוסף, נגדיר $a_j = \frac{1}{4d-1}$ לכל $j$, ולפיכך מתקיים – $\sum_{j=1}^{k=4d-1} a_j = 1$. בנוסף, נגדיר גם – $\gamma = \frac{1}{4d-1}$.
כעת נוכיח את הטענה הנדרשת, כי לכל $1 \le i \le n$, מתקיים האי-שוויון –

$$y_i \sum_{j=1}^{k} a_j h_j(x_i) \ge \gamma$$

ונוכיח זאת ע"י חלוקה למקרים לפי ערכים אפשריים של $y_i$.

במקרה בו $y_i = 1$, אזי ניתן להסיק כי $x_i \in [a_1, b_1] \times \ldots \times [a_n, b_n]$, ולכן –

$$y_i \sum_{j=1}^{k} a_j h_j(x_i) \underbrace{=}_{\substack{y_i=1 \\ a_j=\frac{1}{4d-1}}} \frac{1}{4d-1} \sum_{j=1}^{4d-1} h_j(x_i) = \frac{1}{4d-1}\left[\sum_{j=1}^{d} \underbrace{h_j(x_i)}_{\substack{=1\ as \\ x_i \ge a_r \\ for\ all\ r}} + \sum_{j=d+1}^{2d} \underbrace{h_j(x_i)}_{\substack{=1\ as \\ x_i \le b_r \\ for\ all\ r}} + \sum_{j=2d+1}^{4d-1} \underbrace{h_j(x_i)}_{\substack{=-1 \\ as\ const. \\ hypo.}}\right] =$$

$$= \frac{1}{4d-1}[d + d - (2d-1)] = \frac{1}{4d-1} = \gamma$$

כנדרש במקרה זה, שכן הראנו למעשה שיוויון אך זהו גם אי-שיוויון חלש בפרט.

במקרה בו $y_i = -1$, אזי ניתן להסיק כי קיים $1 \le r \le n$ עבורו $x_i \notin [a_r, b_r]$, כלומר מתקיים אחד מבין השניים – $x_i < a_r$ או $x_i > b_r$. ואכן –

$$y_i \sum_{j=1}^{k} a_j h_j(x_i) = -\frac{1}{4d-1} \sum_{j=1}^{4d-1} h_j(x_i) = -\frac{1}{4d-1}\left[\sum_{\substack{j=1 \\ j \ne r}}^{d} h_j(x_i) + h_r(x_i) + \sum_{\substack{j=d+1 \\ j \ne d+r}}^{2d} h_j(x_i) + h_{d+r}(x_i) + \sum_{j=2d+1}^{4d-1} \underbrace{h_j(x_i)}_{\substack{=-1 \\ as\ const. \\ hypo.}}\right] =$$

$$= -\frac{1}{4d-1}[h_r(x_i) + h_{d+r}(x_i)] - \frac{1}{4d-1}\left[\sum_{\substack{j=1 \\ j \ne r}}^{d} h_j(x_i) + \sum_{\substack{j=d+1 \\ j \ne d+r}}^{2d} h_j(x_i) - (2d-1)\right] =$$

$$= -\frac{1}{4d-1}[h_r(x_i) + h_{d+r}(x_i)] - \frac{1}{4d-1}\sum_{\substack{j=1 \\ j \ne r}}^{d} h_j(x_i) - \frac{1}{4d-1}\sum_{\substack{j=d+1 \\ j \ne d+r}}^{2d} h_j(x_i) + \frac{1}{4d-1}(2d-1) \underbrace{\ge}_{(*)}$$

$$\ge -\frac{1}{4d-1}[h_r(x_i) + h_{d+r}(x_i)] - \frac{1}{4d-1}(d-1) - \frac{1}{4d-1}(d-1) + \frac{1}{4d-1}(2d-1) =$$

$$= \underbrace{-\frac{1}{4d-1}\underbrace{[h_r(x_i) + h_{d+r}(x_i)]}_{\le 0\ by\ (**)}}_{therefore\ \ge 0} + \frac{1}{4d-1} \ge \frac{1}{4d-1} = \gamma$$

**כנדרש,** כאשר המעבר המסומן ב-$(*)$, נובע מכך ש –

$$\sum_{\substack{j=1 \\ j \ne r}}^{d} \underbrace{h_j(x_i)}_{\le 1} \le d-1 \quad \underbrace{\Longrightarrow}_{\substack{multiply\ by \\ a\ negative\ num.}} \quad -\frac{1}{4d-1}\sum_{\substack{j=1 \\ j \ne r}}^{d} h_j(x_i) \ge \left(-\frac{1}{4d-1}\right)\cdot(d-1)$$

וכן – $\sum_{\substack{j=d+1 \\ j \ne d+r}}^{2d} \underbrace{h_j(x_i)}_{\le 1} \le d-1 \quad \underbrace{\Longrightarrow}_{\substack{multiply\ by \\ a\ negative\ num.}} \quad -\frac{1}{4d-1}\sum_{\substack{j=d+1 \\ j \ne d+r}}^{2d} h_j(x_i) \ge \left(-\frac{1}{4d-1}\right)\cdot(d-1)$

ובנוסף, המעבר המסומן ב-$(**)$ קיים לפיה נובע מהנחה, $1 \le r \le n$ עבורו $x_i \notin [a_r, b_r]$, כלומר מתקיים אחד מבין השניים – $x_i < a_r$ או $x_i > b_r$, נקבל כי בהכרח – $h_r(x_i), h_{d+r}(x_i) \le 1$, ולכן מכך ש- $h_{d+r}(x_i) = -1$ **או** $h_r(x_i) = -1$ ולכן בהכרח $h_r(x_i) + h_{d+r}(x_i) \le 0$, ולפיכך הביטוי כולו שבמעבר האחרון הינו חיובי, ובפרט חסום מלמטה ע"י אפס, ומכאן נכונות המעבר.

**ובסה"כ, הראנו כי אכן, לכל $1 \le i \le n$, מתקיים האי-שיוויון הנדרש – $y_i \sum_{j=1}^{k} a_j h_j(x_i) \ge \gamma$, כנדרש בסעיף זה.** ■

**שאלה 4. Linear Regression with dependent Variables.**

נתונה בעיית האופטימזציה הבאה –

$$argmin_w \|w^2\| \qquad s.t. \quad Xw = y$$

**סעיף א. לא להגשה.**

**סעיף ב.** נראה כי הפיתרון האופטימלי לבעיית אופטימיזציה זו $w^*$, מקיים – $w^* = X^T a$, כאשר $a \in \mathbb{R}^n$.

**פיתרון.** נשתמש בכופלי לגרנז' ונסתכל על הלגרנז'יאן. נכתוב את האילוצים בצורה הבאה –

$$Xw = y \qquad \Leftrightarrow \quad \forall i \in \{1, \dots, n\} \ \ x_i w = y_i \ \Leftrightarrow \quad \forall i \in \{1, \dots, n\} \ \ x_i w - y_i = 0$$

כאשר $x_i$ הן שורות המטריצה $X$. הלגרנג'יאן הינו –

$$\mathcal{L}(w, \beta) = \|w^2\| + \sum_{i=1}^{n} \beta_i (x_i w - y_i) = \sum_{i=1}^{d} w_i + \sum_{i=1}^{n} \beta_i (x_i w - y_i)$$

נבחן את הגרדיאנט ונקבל –

$$\frac{\partial \mathcal{L}(w, \beta)}{\partial w_k} = 2w_k + \sum_{i=1}^{n} \beta_i x_i^{(k)}$$

ולפיכך –

$$\nabla_w \mathcal{L}(w, \beta) = 2w + \sum_{i=1}^{n} \beta_i x_i$$

הלגרנג'יאן $\mathcal{L}(w, \beta)$ הינו פונק' קמורה ולפיכך המינימום הגלובאלי מתקבל בנקודה בה הגרדיאנט $\nabla_w \mathcal{L}(w, \beta) = 0$, ולכן קיימים $\beta = (\beta_1, \dots, \beta_n)$, כך שהם כופלי הלגרנז' והם מקיימים $\nabla_w \mathcal{L}(w, \beta) = 0$, ומכאן –

$$\nabla_w \mathcal{L}(w, \beta) = 2w + \sum_{i=1}^{n} \beta_i x_i = 0 \qquad \Rightarrow w = -\frac{1}{2} \sum_{i=1}^{n} \beta_i x_i = -\frac{1}{2} X^T \beta = X^T \left(-\frac{1}{2}\beta\right) \underbrace{=}_{\substack{when \\ a = -\frac{1}{2}\beta}} X^T a$$

כנדרש בטענה למעשה, שכן הוקטור הרצוי $a \in \mathbb{R}^n$ הינו למעשה הוקטור בו הקורדינטות הן $-\frac{1}{2}\beta_i$, כאשר $\beta_i$ הם למעשה כופלי הלגרנז'. ■

**סעיף ג.** נראה כי ניתן לחשב את $x^T w^*$ לכל $x \in \mathbb{R}^d$ ע"י שימוש במכפלות פנימיות בין וקטורים ב-$\mathbb{R}^d$ בלבד.

**פיתרון.** יהי $x \in \mathbb{R}^d$ וקטור כלשהו.
ראינו בסעיף ב' כי קיים וקטור $a \in \mathbb{R}^n$ כך ש-$w^* = X^T a$, וכעת אם נסמן את שורות המטריצה $X$ ב-$x_1, \dots x_n$, מתקיים –

$$x^T w^* = x^T X^T a = (Xx)^T a \underbrace{=}_{scalar} a^T (Xx) = (a_1 \quad \cdots \quad a_n) \cdot \begin{pmatrix} x_1 x \\ \vdots \\ x_n x \end{pmatrix} = \sum_{i=1}^{n} a_i x_i x \underbrace{=}_{\substack{kernel \\ trick.}} \sum_{i=1}^{n} a_i K_S(x_i, x)$$

כנדרש הראנו כי ניתן לחשב את $x^T w^*$ ע"י שימוש במכפלות פנימיות בין וקטורים ב-$\mathbb{R}^d$ בלבד, וכן ע"י שימוש בקרנל טריק.

**שאלה 5. Perceptron Lower Bound.**

נדרש להראות כי לכל $0 < \gamma < 1$, קיימים מספר $d > 0$, וקטור $w^* \in \mathbb{R}^d$ ורצף של דוגמאות $(x_1, y_1), \dots, (x_m, y_m)$ כך שמתקיימים שלושה תנאים –

$$(a)\quad \|x_i\| = 1$$
$$(b)\quad \frac{y_i x_i w^*}{\|w^*\|} \geq \gamma$$
$$(c)\ Perceptron\ makes\ \left\lfloor \frac{1}{\gamma^2} \right\rfloor\ mistakes\ on\ the\ sequence.$$

**פיתרון.** נשתמש ברמז הנתון ונבחר $m = d = \left\lfloor \frac{1}{\gamma^2} \right\rfloor$ וכן נגדיר את $\{x_i\}_{i=1}^m$ להיות וקטורי הבסיס הסטנדרטי של $\mathbb{R}^m$. כמו כן, נגדיר את ה-$true\ labels$ –

$$\forall i \in \{1, \dots, m\}\quad define\quad \boldsymbol{y_i = -1}$$

**וכעת נטען כי ע"י אלגוריתם ה-$\boldsymbol{perceptron}$ מתקבל הוקטור $\boldsymbol{w^* = (-1, -1, \dots, -1)}$ ומתקיימים כל התנאים הנדרשים.**

הטענה המרכזית ממנה ינבע מבנה הוקטור $w^*$ הינה טענה אינדוקטיבית לפיה –

**טענה.** לאחר $0 \leq t \leq m = \left\lfloor \frac{1}{\gamma^2} \right\rfloor$ איטרציות של אלגוריתם ה-$perceptron$ מתקיים $w_t = (-1, \dots -1\ ,\ \underbrace{0}_{\substack{index\\t}}\ ,\ 0 \dots)$, כאשר ספירת האינדקסים מ-0.

**הוכחה. מקרה הבסיס** הינו $t = 0$, ובו אכן מתקיימת מהגדרת האלגוריתם – $w_0 = (0, \dots, 0)$, שכן כך מאותחל וקטור זה.

**צעד האינדוקציה.** נניח כי לאחר $t$ צעדים מתקיים – $w_t = (-1, \dots, -1,\ \underbrace{0}_{\substack{index\\t}}\ , 0 \dots 0)$, ונסתכל על האיטרציה הבאה של האלגוריתם.

באיטרציה זו מבצעים ניחוש של $\hat{y}_t = sign(w_t x_t) = sign(w_t^t) = sign(0) = +1$, כאשר הנכונות של טיעון זה נובעת מכך שהוקטור $x_t$ הינו אחד מוקטורי הבסיס הסטנדרטי של $\mathbb{R}^m$, כלומר מהצורה $x_t = (0, \dots 0,\ \underbrace{1}_{\substack{index\\t}}\ , 0, \dots, 0)$, ולפיכך המכפלה $w_t x_t$ שווה למעשה לקואורדינטה ה-$t$ בוקטור $w_t$, שהינה כאמור אפס לפי הנחת האינדוקציה. כעת, בהמשך האיטרציה הנוכחית מתקיים כי – $\hat{y}_t = 1 \neq -1 = y_t$, ולפיכ מעדכנים את הוקטור כך –

$$w_{t+1} = w_t + y_t x_t = \left(-1, \dots, -1,\ \underbrace{0}_{\substack{index\\t}}\ , 0, \dots 0\right) - 1\left(0, \dots, 0,\ \underbrace{1}_{\substack{index\\t}}\ , 0, \dots 0\right) = \left(-1, \dots, -1,\ \underbrace{-1}_{\substack{index\\t}}\ , 0, \dots 0\right)$$

כרצוי בטענת האינדוקציה, שכן הראנו כי לאחר האיטרציה ה-$t+1$, מתקיים –

$$w_{t+1} = \left(-1, \dots, -1,\ \underbrace{-1}_{\substack{index\\t}}\ , 0, \dots 0\right) = \left(-1, \dots, -1, -1,\ \underbrace{0}_{\substack{index\\t+1}}\ , 0, \dots 0\right)$$

**מסקנה מטענה זו – לאחר $\boldsymbol{m}$ איטרציות, כלומר בסוף אלגוריתם ה-$\boldsymbol{perceptron}$ מתקיים –** $\boxed{\boldsymbol{w^* = (-1, -1, \dots, -1)}}$.

כעת נותר להראות כי עבור $d = \left\lfloor \frac{1}{\gamma^2} \right\rfloor$, הוקטור המתקבל מאלגוריתם ה-$perceptron$, שנסמן ב- $w^* = (-1, \dots, -1) \in \mathbb{R}^d$, ועבור רצף הדוגמאות של $(x_1, y_1), \dots, (x_m, y_m)$, כאשר $\{x_i\}_{i=1}^m$ וקטורי הבסיס הסטנדרטי של $\mathbb{R}^m$, והתיוגים המתאימים מקיימים – $y_i = -1$ לכל $i \in \{1, \dots, m\}$, מתקיימים התנאים הנדרשים בשאלה - $(a), (b), (c)$. ואכן –

$$(a)\ \ x_i\ = (0, \dots, 0, 1, 0, \dots, 0) \quad \Longrightarrow \quad \|x_i\| = 1 \quad as\ required.$$

$$(b)\quad \frac{y_i x_i \cdot w^*}{\|w^*\|} = \left[x_i \cdot w^* = w_i^* \quad and \quad \|w^*\| = \sqrt{\sum_{i=1}^{d} (-1)^2} = \sqrt{\sum_{i=1}^{d} 1} = \sqrt{d} = \sqrt{\frac{1}{\gamma^2}}\right] = \frac{\overbrace{y_i}^{=-1}\ \overbrace{w_i^*}^{=-1}}{\sqrt{\frac{1}{\gamma^2}}} = \frac{1}{\sqrt{\frac{1}{\gamma^2}}} = \sqrt{\gamma^2} = \gamma$$

הראנו בסעיף $(b)$ כי – $\frac{y_i x_i \cdot w^*}{\|w^*\|} = \gamma$, אך בפרט מתקיים גם אי-שיוויון חלש כנדרש. בנוסף,

$$(c)\ \ It\ was\ showed\ that\ w_t = \left(-1, \dots, -1,\ \overbrace{0}^{index\ t}\ , 0, \dots 0\right) \Rightarrow \forall t.\ \ \hat{y}_t = sign(w_t x_t) = sign(w_t^t) = sign(0) = +1 \neq -1 = y_t$$

ולפיכך, המסקנה המתקבלת היא כי בכל איטרציה של אלגוריתם ה-$perceptron$ , מקבלים כי $\hat{y}_t \neq y_t$, כלומר אלגוריתם ה-$perceptron$ מבצע שגיאה בכל איטרציה, ואם כך אלגוריתם ה-$perceptron$ אכן מבצע $\left\lfloor \frac{1}{\gamma^2} \right\rfloor$ שגיאות, כמספר האיטרציות שמבוצעות בו, שכזכור מספר האיטרציות הינו למעשה מספר הדוגמאות $m$.

בסה"כ הראנו כי לכל $0 < \gamma < 1$, קיימים פרמטרים $d, m, w^*, (x_1, y_1), \dots, (x_m, y_m)$ כך שמתקיימים שלושת התנאים הללו, כנדרש בשאלה. ■

**שאלה 6. Halving Algorithm.**

נסמן ב-$A_{Hal}$ את אלגוריתם ה-$Halving$ שנלמד בכיתה. יהי $d \geq 6$, וכן $\mathcal{X} = \{1,2,\dots d\}$, וכן $\mathcal{H} = \{h_{i,j} : 1 \leq i < j \leq d\}$, כך ש-

$$h_{i,j}(x) = \begin{cases} 1 & if \quad (x = i)\ or\ (x = j) \\ 0 & otherwise \end{cases}$$

נדרש להראות כי $M(A_{Hal}, \mathcal{H}) = 2$.

**פיתרון.** יהי רצף כלשהו של דגימות ותיוגים – $S = \left(\left(x_1, h^*(x_1)\right), \dots, \left(x_m, h^*(x_m)\right)\right)$. ראשית נבחין כי – $|\mathcal{H}| = \sum_{i=1}^{d-1} i = \frac{d(d-1)}{2}$ .

נסמן ב- $Majority(H', x')$ את התיוג שמתקבל ע"י רוב ההיפותזות ב-$H'$ עבור $x' \in \mathcal{X}$. נתחיל בביצוע אלגוריתם ה-$Halving$ כפי שנלמד בהרצאות, ע"י מעבר על רצף הדגימות והתיוגים ב-$S$. כעת, ייתכנו שני מצבים אפשריים.

**מקרה** $(\mathbf{1.1})$ בו האלגוריתם לא ביצע שגיאה באף אחת מן האיטרציות, כלומר לכל $1 \leq i \leq m$, מתקיים $Majority(\mathcal{H}, x_i) = h^*(x_i)$, ובפרט במצב זה מתקיים – $M_{A_{Hal}}(S) = 0$, במקרה זה מספר השגיאות הינו ומינימלי, ולא ניתן ללמוד על $M(A_{Hal}, \mathcal{H})$ .

**מקרה** $(\mathbf{1.2})$ בו האלגוריתם ביצע שגיאה(לפחות אחת) באיטרציה כלשהי,נסמן ב-$k_1$ את הדגימה הראשונה ב-$S$ , עבורה התקבלה שגיאה. כלומר לכל $1 \leq i < k_1$, מתקיים – $Majority(\mathcal{H}, x_i) = h^*(x_i)$, כלומר לא מתקבלות שגיאות באיטרציות אלו, ולפיכך בכל אחד מן השלבים הללו לא מתבצע שינוי ב-$\mathcal{H}$. אך עבור $k_1$, מתקיים – $Majority\left(\mathcal{H}, x_{k_1}\right) \neq h^*(x_{k_1})$.

כעת, מתקיים –
$$n_1 := \left|\left\{h_{i,j} \in \mathcal{H} : \quad h_{i,j}\left(x_{k_1}\right) = 1\right\}\right| = \left|\left\{h_{x_{k_1},j} \in \mathcal{H}: \ j > x_{k_1}\right\} \cup \left\{h_{i,x_{k_1}} \in \mathcal{H} : i < x_{k_1}\right\}\right|$$

$$\underbrace{=}_{\substack{disjoint \\ sets.}} \left|\left\{h_{x_{k_1},j} \in \mathcal{H}: \ j > x_{k_1}\right\}\right| + \left|\left\{h_{i,x_{k_1}} \in \mathcal{H} : i < x_{k_1}\right\}\right| = \left(d - x_{k_1}\right) + \left(x_{k_1} - 1\right) = d - 1$$

ולפיכך, $n_0 := \left|\left\{h_{i,j} \in \mathcal{H} : \quad h_{i,j}\left(x_{k_0}\right) = 0\right\}\right| = |\mathcal{H}| - n_1 = \frac{d(d-1)}{2} - (d-1) = (d-1)\left(\frac{d-2}{2}\right) = \frac{(d-1)(d-2)}{2}$, וכן –

$$n_1 = d - 1 < \frac{(d-1)(d-2)}{2} = n_0 \qquad \Leftrightarrow \quad d > 4$$

וביוון שלפי ההנחה $d \geq 6$, נוכל להסיק כי $n_0 > n_1$, ולפיכך – $Majority\left(\mathcal{H}, x_{k_1}\right) = 0$, ולפיכך $h^*\left(x_{k_1}\right) = 1$, שכן כידוע באיטרציה זו התקבלה השגיאה הראשונה ולכן – $Majority\left(\mathcal{H}, x_{k_1}\right) \neq h^*(x_{k_1})$. על כן, קבוצת ההיפותזות איתה האלגוריתם ממשיך הינה –

$$\boxed{\boldsymbol{\mathcal{H}' = \{h_{i,j} \in \mathcal{H} \ : \quad h_{i,j}\left(x_{k_1}\right) = 1\}}}$$

והיא מקיימת – $|\mathcal{H}'| = d - 1$ כפי שראינו בחישוב של $n_1$. **כמו כן, עד כה האלגוריתם ביצע שגיאה אחת בדיוק.**

כעת, נמשיך בלמידה לפי אלגוריתם ה-$Halving$ עבור $i > k_1$, ושוב ייתכנו שני מקרים אפשריים **בתת-מקרה** $(\mathbf{1.2})$.

**מקרה** $(\mathbf{1.2.1})$ הינו כאשר לא התבצעו עוד שגיאות ע"י האלגוריתם, כלומר לכל $i > k_1$, מתקיים– $Majority(\mathcal{H}', x_i) = h^*(x_i)$.

ובפרט במקרה זה, מתקיים – $M_{A_{Hal}}(S) = 1$, שכן התבצעה שגיאה אחת בדיוק באיטרציה עבור $\left(x_{k_1}, h^*\left(x_{k_1}\right)\right)$, ולפי שאר ההנחות של תת-מקרה זה, לא התבצעו שגיאות נוספות.

**מקרה** $(\mathbf{1.2.2})$ הינו כאשר התבצעה עוד שגיאה ע"י האלגוריתם, בשלב ה-$k_2$. בדומה, לכל $k_1 < i < k_2$, לא התבצעו שגיאות ע"י האלגוריתם, כלומר התקיים – $Majority(\mathcal{H}', x_i) = h^*(x_i)$ בכל אחת מן האיטרציות הללו, וכן $\mathcal{H}'$ לא השתנתה.

בשלב ה-$k_2$, האלגוריתם ביצע שגיאה, כלומר לכן $Majority\left(\mathcal{H}', x_{k_2}\right) \neq h^*(x_{k_2})$. נחלק לשני מקרים לפי $x_{k_1} \stackrel{?}{=} x_{k_2}$.

אם $x_{k_1} = x_{k_2}$, אזי לכל $h_{i,j} \in \mathcal{H}'$, מתקיים – $h_{i,j}\left(x_{k_2}\right) = 1$, כאשר הנכונות של כך נובעת מבניית $\mathcal{H}'$ לפי האלגוריתם, שכן אם

$$for\ h_{i,j} \in \mathcal{H}' \quad \Rightarrow \quad \left(\ i = x_{k_1}\ \ and\ j > x_{k_1}\ \right)\ or\ \left(\ j = x_{k_1}\ \ and\ i < x_{k_1}\right) \ \Rightarrow\ h_{i,j}\left(x_{k_2}\right) \underbrace{=}_{x_{k_1} = x_{k_2}} h_{i,j}\left(x_{k_1}\right) \underbrace{=}_{\substack{def \\ of\ h_{i,j}}} 1$$

ולפיכך, מתקיים – $Majority\left(\mathcal{H}', x_{k_2}\right) = 1$, ולכן $h^*\left(x_{k_2}\right) = 0$, ולכן קבוצת ההיפותזות איתה האלגוריתם ממשיך הינה –

$$\mathcal{H}'' = \left\{h_{i,j} \in \mathcal{H}' : \quad h_{i,j}\left(x_{k_2}\right) = 0\right\} \underbrace{=}_{x_{k_1} = x_{k_2}} \left\{h_{i,j} \in \mathcal{H}' : \quad h_{i,j}\left(x_{k_1}\right) = 0\right\} \underbrace{=}_{def.\ \mathcal{H}'} \emptyset$$

וזה בסתירה באמור לכך שקיימת $h^* \in \mathcal{H}$, כך ש – $h^*(x_i) = y_i$, כלומר בסתירה להנחת ה- $realizability$ הקיימת תחת אלגוריתם זה.

**על כן, נסיק כי בהכרח $\boldsymbol{x_{k_1} \neq x_{k_2}}$, שכן אחרת זו סתירה ל-$\boldsymbol{realizability}$ .**

ניזכר כי –

$$\mathcal{H}' = \{h_{i,j} \in \mathcal{H} \ : \ h_{i,j}(x_{k_1}) = 1\} = \{h_{x_{k_1},j} \in \mathcal{H}: \ j > x_{k_1}\} \cup \{h_{i,x_{k_1}} \in \mathcal{H} : i < x_{k_1}\}$$

**במקרה בו $x_{k_1} < x_{k_2}$,** מתקיים עבור ההיפותזות ב-$\mathcal{H}'$ –

$$for\ all\ i < x_{k_1} \quad : \quad h_{i,x_{k_1}}(x_{k_2}) = 0$$

$$for\ j > x_{k_1}: \quad h_{x_{k_1},j}(x_{k_2}) = \ 1 \quad if \quad j = x_{k_2} \quad , \ \ and \quad h_{x_{k_1},j}(x_{k_2}) = 0 \quad if \ \ j \neq x_{k_2}$$

ולפיכך, קל לראות כי – $Majority(\mathcal{H}', x_{k_2}) = 0$ במקרה זה, ולכן – $h^*(x_{k_2}) = 1$, שכן כאמור $Majority(\mathcal{H}', x_{k_2}) \neq h^*(x_{k_2})$, ולפיכך במקרה זה, קבוצת ההיפותזות איתה האלגוריתם ממשיך לאחר השגיאה השנייה ב-$k_2$ הינה –

$$\mathcal{H}'' = \{h_{i,j} \in \mathcal{H}' : \quad h_{i,j}(x_{k_2}) = 1\} = \{h_{x_{k_1},x_{k_2}}\}$$

ונבחין אם כך כי - $|\mathcal{H}''| = 1$, כלומר קבוצת ההיפותזות הנותרת הינה מגודל 1,ולכן למעשה האלגוריתם מצא את $h^*$, כנדרש, ולפיכך האלגוריתם הסתיים ובמקרה זה ביצע 2 שגיאות בדיוק, ולכן במקרה זה $M_{A_{Hal}}(S) = 2$.

**במקרה בו $x_{k_1} > x_{k_2}$,** מתקיים עבור ההיפותזות ב-$\mathcal{H}'$ –

$$for\ all\ j > x_{k_1} \quad : \quad h_{x_{k_1},j}(x_{k_2}) = 0$$

$$for\ i < x_{k_1} \quad : \quad h_{i,x_{k_1}}(x_{k_2}) = 1 \quad if \ \ i = x_{k_2} \quad , \ \ and \quad h_{i,x_{k_1}}(x_{k_2}) = 0 \quad if \ \ i \neq x_{k_2}$$

ולפיכך, קל לראות כי – $Majority(\mathcal{H}', x_{k_2}) = 0$ במקרה זה, ולכן – $h^*(x_{k_2}) = 1$, שכן כאמור $Majority(\mathcal{H}', x_{k_2}) \neq h^*(x_{k_2})$, ולפיכך במקרה זה, קבוצת ההיפותזות איתה האלגוריתם ממשיך לאחר השגיאה השנייה ב-$k_2$ הינה –

$$\mathcal{H}'' = \{h_{i,j} \in \mathcal{H}' : \quad h_{i,j}(x_{k_2}) = 1\} = \{h_{x_{k_2},x_{k_1}}\}$$

ונבחין אם כך כי - $|\mathcal{H}''| = 1$, כלומר קבוצת ההיפותזות הנותרת הינה מגודל 1,ולכן למעשה האלגוריתם מצא את $h^*$, כנדרש, ולפיכך האלגוריתם הסתיים ובמקרה זה ביצע 2 שגיאות בדיוק, ולכן במקרה זה $M_{A_{Hal}}(S) = 2$.

והנ"ל מסכם את כל המקרים האפשריים בעת ביצוע אלגוריתם ה-$Halving$, **וראינו כי במקרה הגרוע ביותר, האלגוריתם מבצע 2 שגיאות**, כלומר ניתן אכן להסיק את הנדרש –

$$M(A_{Hal}, \mathcal{H}) = \sup_S M_{A_{Hal}}(S) = 2$$

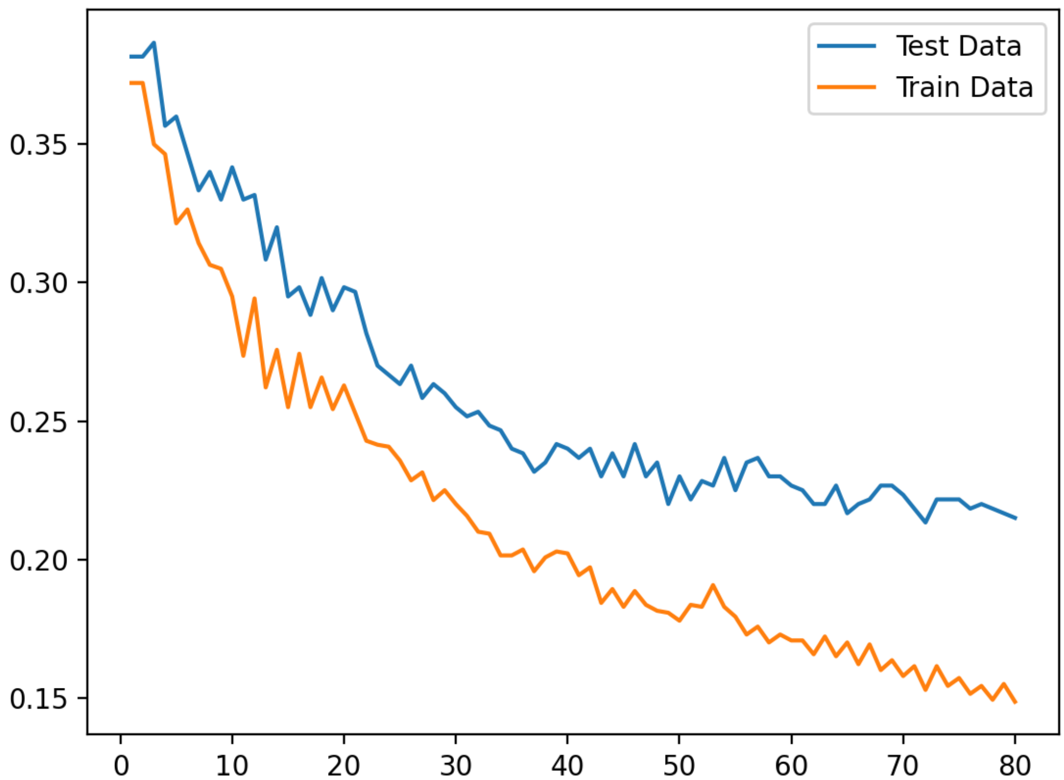
■

**Programming Assignment.**

**שאלה 1. AdaBoost.**

**סעיף א.**

הגרף המתקבל על ידי ריצת $AdaBoost$ תחת פרמטרים אלו, של הטעויות על גבי סט האימונים וסט הבדיקות לפי המסווג $sign(\sum_{j=1}^{t} a_j h_j(x_i))$ הינו –



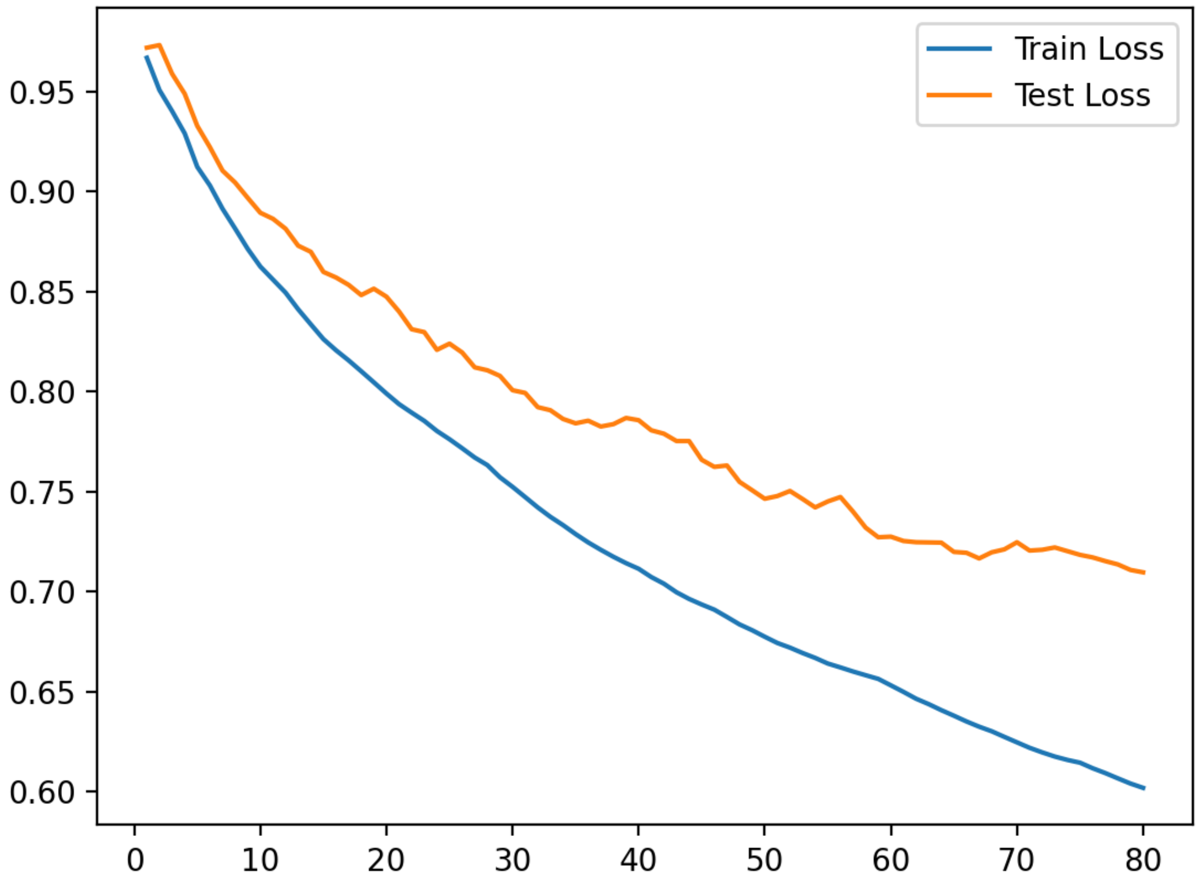

**סעיף ב.** נקבל את הפלט הבא –

```
WL - 1 is the hypothesis (1, 26, 0.5) and the word is bad
WL - 2 is the hypothesis (-1, 31, 0.5) and the word is many
WL - 3 is the hypothesis (-1, 22, 0.5) and the word is life
WL - 4 is the hypothesis (1, 311, 0.5) and the word is worst
WL - 5 is the hypothesis (-1, 37, 1.5) and the word is great
WL - 6 is the hypothesis (1, 372, 0.5) and the word is boring
WL - 7 is the hypothesis (-1, 282, 0.5) and the word is perfect
WL - 8 is the hypothesis (1, 292, 0.5) and the word is supposed
WL - 9 is the hypothesis (-1, 196, 0.5) and the word is performances
WL - 10 is the hypothesis (1, 88, 0.5) and the word is script
```

**שלושה לומדים חלשים שציפיתי כי יעזרו בצורה מיטבית לבצע סיווג של ביקורות הם $worst, great, boring$, שכן שימוש במילים אלו מגלה רבות על דעת כותב ה-$review$, שכן אלו מילים בעלות קונוטציה חיובית/שלילית חזקה מאוד, ומהם כאמור ניתן להבין האם החוות הדעת של היא חיובית/שלילית לגבי הסרט.**

**שלושה לומדים חלשים שציפיתי כי לא יעזרו(או יעזרו מעט לכל היותר) לבצע סיווג של ביקורות הם $performances, script, life$, שכן אלו מילים שאינן בעלות קונוטציה חיובית/שלילית, ולפיכך לא ניתן ללמוד מהם לגבי חיוביות/שליליות הביקורת שנכתבה. כמו כן, ייתכן כי האלגוריתם בחר בהם מכיוון שהם מילים המתארות בצורה קונקרטית(ספציפית) יותר את הסרט, למשל ע"י תיאור המשחק של השחקנים או תיאור מפורט של העלילה בסרט, ולפיכך למרות זאת, כן ניתן למצוא סיבות אפשריות לבחירה בהם שכן מילים אלו מתארות את הסרט למרות זאת, וכן ניתן ללמוד מהן על טיב הביקורת.**

**סעיף ג.**




זהו גרף הפונק' $\ell$ כפונק' של $T$, כאשר –

$$\ell = \frac{1}{m}\sum_{i=1}^{m} e^{-y_i \sum_{j=1}^{T} a_j h_j(x_i)}$$

וניתן לראות כי כאשר הפרמטר $T$ גדל, פונק' ה-$loss$ הן עבור ה-$Training\ data$ והן עבור ה-$Test\ data$, יורדות. הסבר אפשרי לכך הינו שכאמור מבצעים יותר איטרציות של $AdaBoost$ ולפיכך האלגוריתם מצליח ללמוד יותר מהמסווג האידאלי, ולכן מתקבלות שגיאות הולכות ופוחתות על גבי הנתונים. כמו כן, ניתן להבחין כי השגיאה על גבי סט ה-$Test$ גבוהה מהשגיאה המתקבלת על גבי סט ה-$Train$, והסבר אפשרי לכך הינו שהאלגוריתם למד על גבי סט האימון, ולכן כאמור הביצועים שלו על גבי סט האימון יהיו טובים יותר מאשר הביצועים על גבי סט נתונים חדש לגמרי. כמו כן, ניתן לראות כי השגיאה יורדת בקצב גבוה למדי כאשר $T$ גדל.