(67577) מבוא למערכות לומדות | תרגיל 4

שם: נמרוד בר גיורא | ת"ז: 207090622

חלק תאורטי

למידות PAC

שאלה 1

1. For $\mathcal{A}$ some learning algorithm, $\mathcal{D}$ a probability distribution over $\mathcal{X}$ and the 0-1 loss function (i.e misclassification), prove the following are equivalent:

   (a) $\forall \varepsilon, \delta > 0 \quad \exists m(\varepsilon, \delta) \quad \text{s.t} \quad \forall m \geq m(\varepsilon, \delta) \quad \mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(\mathcal{A}(S)) \leq \varepsilon] \geq 1-\delta$
   (b) $\lim_{m\to\infty} \mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] = 0$

הוכחה: $(b \Rightarrow a)$ נניח ש־ $\lim\limits_{m\to\infty} \mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right)\right] = 0$.
יהיו $\varepsilon, \delta > 0$. מאי־שיוויון מרקוב נובע שלכל $m \in \mathbb{N}$ מתקיים:

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right) > \varepsilon\right] \leqslant \frac{\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right)\right]}{\varepsilon}$$

מהגדרת הגבול נובע שקיים $m\left(\varepsilon,\delta\right) \in \mathbb{N}$ כך שלכל $m \geqslant m\left(\varepsilon,\delta\right)$ מתקיים ש־

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right)\right] < \varepsilon\cdot\delta$$

ולכן לכל $m \geqslant m\left(\varepsilon,\delta\right)$ מתקיים:

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right) > \varepsilon\right] < \frac{\varepsilon\cdot\delta}{\varepsilon} = \delta \quad \overset{\text{המאורע המשלים}}{\Longrightarrow} \quad \mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right) \leqslant \varepsilon\right] \geqslant 1-\delta$$

$(a \Rightarrow b)$ נניח שלכל $\varepsilon,\delta > 0$ קיים $m\left(\varepsilon,\delta\right) \in \mathbb{N}$ כך שלכל $m \geqslant m\left(\varepsilon,\delta\right)$ מתקיים ש־ $\mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right) \leqslant \varepsilon\right] \geqslant 1-\delta$.
יהיו $\varepsilon,\delta > 0$. נשתמש בנוסחת התוחלת השלמה על $\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_D\left(\mathcal{A}\left(S\right)\right)\right]$ ונקבל לכל $m \geqslant m\left(\varepsilon,\delta\right)$ ש־

$$\begin{aligned}\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right)\right] = {} & \mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right) \leqslant \varepsilon\right]\cdot\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right) \mid L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right) \leqslant \varepsilon\right] + \\ & + \mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right) > \varepsilon\right]\cdot\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right) \mid L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right) > \varepsilon\right]\end{aligned}$$

נחסום כל אחד מהביטויים לעיל בנפרד:
מכך ש־ $\mathbb{P}_{S\sim\mathcal{D}^m}$ היא פונקציית הסתברות נובע ש־

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[L_D\left(\mathcal{A}\left(S\right)\right) \leqslant \varepsilon\right] \leqslant 1$$

מההנחה של a (תוך שימוש במאורע המשלים) נובע ש־

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right) > \varepsilon\right] = 1 - \mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right) \leqslant \varepsilon\right] \overset{\overset{a}{\downarrow}}{\leqslant} 1-\left(1-\delta\right) = \delta$$

ממונוטוניות התוחלת, ומכך שזו תוחלת מותנית נובע ש־

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right) \mid L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right) \leqslant \varepsilon\right] \leqslant \mathbb{E}_{S\sim\mathcal{D}^m}\left[\varepsilon\right] = \varepsilon$$

לפי ההגדרה ־ $L_{\mathcal{D}} = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[h\left(\mathbf{x}\right) \neq f\left(\mathbf{x}\right)\right]$. לכן מכיוון ש־ $\mathbb{1}_{[h(\mathbf{x})\neq f(\mathbf{x})]} \leqslant 1$ אז ממונוטוניות התוחלת (מופעלת על $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}$ ואז על $\mathbb{E}_{S\sim\mathcal{D}^m}$):

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right) \mid L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right) > \varepsilon\right] \leqslant \mathbb{E}_{S\sim\mathcal{D}^m}\left[1\right] = 1$$

ולכן קיבלנו שלכל $m \geqslant m\left(\varepsilon,\delta\right)$ מתקיים ־

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right)\right] \leqslant 1\cdot\varepsilon + \delta\cdot 1 = \varepsilon + \delta$$

ומכאן לפי הגדרת הגבול ־ $\lim\limits_{m\to\infty}\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right)\right] = 0$. ■

שאלה 2

2. Let $\mathcal{X} := \mathbb{R}^2$, $\mathcal{Y} := \{0,1\}$ and let $\mathcal{H}$ be the class of concentric circles in the plane, i.e.,

$$\mathcal{H} := \{h_r : r \in \mathbb{R}_+\} \quad \text{where} \quad h_r(\mathbf{x}) = \mathbb{1}_{[\|\mathbf{x}\|_2 \leq r]}$$

Prove that $\mathcal{H}$ is PAC-learnable and its sample complexity is bounded by

$$m_{\mathcal{H}}(\varepsilon,\delta) \leq \frac{\log(1/\delta)}{\varepsilon}$$

<u>אלגוריתם למידה</u>: בהינתן סט דגימות $S = \left\{\left(\mathbf{x}_i, y_i\right)\right\}_{i=1}^m$ האלגוריתם $\mathcal{A}$ יחזיר את ההיפותזה $\mathcal{A}\left(S\right) = h_{\hat{r}} \in \mathcal{H}$ כאשר:

$$\hat{r} = \max_{i\in[m]:\, y_i=1}\left\|\mathbf{x}_i\right\|_2$$

כאשר מגדירים $\max\emptyset = 0$.

כלומר האלגוריתם יחזיר את המעגל סביב הראשית עם הרדיוס הגדול ביותר שמכיל את הדגימות שהלייבל שלהן הוא 1, מתוך סט הדגימות הנתון.

במילים אחרות ־ הרדיוס של המעגל יקבע לפי הדגימה $\mathbf{x}_i$ כך ש־ $y_i = 1$ ו־$\left\|\mathbf{x}_i\right\|_2$ מקסימלי.

הוכחת נכונות:

הוכחה: נראה שלכל מעגל סביב הראשית ברדיוס r, לכל הסתברות $\mathcal{D}$ מעל $\mathcal{X}$ ולכל $\varepsilon,\delta \in (0,1)$ ־ אם נדגום m דגימות i.i.d לפי ההתפלגות $\mathcal{D}$ אז עם הסתברות של לפחות $1-\delta$ המעגל סביב הראשית ברדיוס $\hat{r}$ שיוחזר ע"י האלגוריתם לעיל הוא בעל שגיאה של לכל היותר ε.

אז יהיו $\mathcal{D},\varepsilon,\delta$ כנ"ל. נשים לב תחילה שהרדיוס של המעגל שמוחזר ע"י האלגוריתם תמיד יהיה קטן/שווה לרדיוס האמיתי ־ $\hat{r} \leqslant r$, ולכן:

$$L_{\mathcal{D}}\left(h_{\hat{r}}\right) = \mathbb{P}\left(\left\|\mathbf{x}_i\right\|_2 \in \left(\hat{r}, r\right]\right)$$

נרצה למצוא מספר דגימות גדול מספיק m שיבטיח ש־ $\mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(h_{\hat{r}}\right) \leqslant \varepsilon\right] \geqslant 1-\delta$.

נתבונן לצורך כך בהסתברות של המאורע המשלים ־ $\mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(h_{\hat{r}}\right) > \varepsilon\right]$. נשים לב שאם ההסתברות לדגום נקודה $\mathbf{x}_i \in \mathbb{R}^2$ שנמצאת בהפרש בין המעגל האמיתי (ברדיוס r) למעגל שהחזיר $\mathcal{A}$ (ברדיוס $\hat{r}$) היא גדולה מ־ε, כלומר אם $\mathbb{P}\left(\left\|\mathbf{x}_i\right\|_2 \in \left(\hat{r}, r\right]\right) > \varepsilon$, אז ההסתברות שבסט הדגימות S לא תהיה אף נקודה באיזור הזה היא לפחות $\left(1-\varepsilon\right)^m$. פורמלית:

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(h_{\hat{r}}\right) > \varepsilon\right] = \mathbb{P}_{S\sim\mathcal{D}^m}\left[\bigwedge_{i=1}^{m}\left\|\mathbf{x}_i\right\|_2 \notin \left(\hat{r}, r\right]\right] \overset{\text{i.i.d}}{=} \prod_{i=1}^{m}\mathbb{P}_{\mathbf{x}_i\sim\mathcal{D}}\left(\left\|\mathbf{x}_i\right\|_2 \notin \left(\hat{r}, r\right]\right) \leqslant \left(1-\varepsilon\right)^m \leqslant e^{-\varepsilon m}$$

מכאן נובע ש־

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(h_{\hat{r}}\right) \leqslant \varepsilon\right] = 1 - \mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(h_{\hat{r}}\right) > \varepsilon\right] \geqslant 1 - e^{-\varepsilon m}$$

לכן מספר הדגימות צריך לקיים:

$$1-e^{-\varepsilon m}\geqslant 1-\delta \implies e^{-\varepsilon m}\leqslant \delta \implies -\varepsilon m\leqslant \ln\left(\delta\right) \implies m\geqslant \frac{-\ln\left(\delta\right)}{\varepsilon}=\frac{\ln\left(\delta^{-1}\right)}{\varepsilon}=\frac{\ln\left(\frac{1}{\delta}\right)}{\varepsilon}$$

כלומר, לכל $m\geqslant \frac{\ln\left(\frac{1}{\delta}\right)}{\varepsilon}$ נקבל ש־

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(h_{\hat{r}}\right)\leqslant \varepsilon\right]\geqslant 1-\delta$$

ומכיוון ש־$\varepsilon,\delta,\mathcal{D}$ היו שרירותיים - $\mathcal{H}$ היא למידה־PAC וה־sample complexity שלה מקיים:

$$m_{\mathcal{H}}\left(\varepsilon,\delta\right)\leqslant \frac{\ln\left(\frac{1}{\delta}\right)}{\varepsilon}$$

$\blacksquare$

VC-Dimension

שאלה 3

3. Let $\mathcal{X}=\{0,1\}^n$ and $\mathcal{Y}=\{0,1\}$, for each $I\subseteq[n]$ define the parity function:

$$h_I(\mathbf{x})=\left(\sum_{i\in I}x_i\right)\bmod 2.$$

What is the VC-dimension of the class $\mathcal{H}_{parity}=\{h_I\mid I\subseteq[n]\}$? Prove your answer.

ה־VC-dimention של המחלקה $\mathcal{H}_{\text{parity}}=\{h_I\ :\ I\subseteq[n]\}$ הוא n.

הוכחה: נשים לב תחילה שלכל $i\subseteq[n]$ מתקיים לכל $\mathbf{x}\in\{0,1\}^n$ ש־

$$h_I\left(\mathbf{x}\right)=\left(\sum_{i\in I}x_i\right)\text{mod}2=\left\langle \mathbf{x}\mid v_I\right\rangle \text{mod}2$$

כאשר $v_I\in\{0,1\}^n$ הוא וקטור המוגדר ע"י:

$$\left[v_I\right]_i=\begin{cases}1 & i\in I\\ 0 & i\notin I\end{cases}$$

כלומר הוא וקטור שהקואורדינטה ה־i שלו מציינת האם האינדקס i שייך לקבוצת האינדקסים I.
לכן לכל $I\subseteq[n]$ מתקיים ש־ $h_I\in\mathcal{H}$ <u>אם ורק אם</u> קיים לה וקטור v_I כנ"ל.
נראה שהקבוצה $C=\{e_1,\ldots,e_n\}\subseteq\{0,1\}^n$ (כאשר לכל $i\in[n]$ הוקטור e_i הוא וקטור היחידה ה־i) מנותצת ע"י $\mathcal{H}$:
לכל n לייבלים $y_1,\ldots,y_n\in\{0,1\}$ עבור הוקטורים ב־C נגדיר

$$v_I=\begin{pmatrix}y_1\\ \vdots\\ y_n\end{pmatrix}$$

ונקבל לכל $i\in[n]$ ש־

$$h_I\left(e_i\right)=\left\langle e_i\mid v_I\right\rangle \text{mod}2=\left[v_I\right]_i\text{mod}2=y_i\text{mod}2\overset{y_i\in\{0,1\}}{\overset{\downarrow}{=}}y_i$$

כלומר $\mathcal{H}$ מנתצת את C הנ"ל, ולכן n $\text{VC-DIM}\left(\mathcal{H}\right) \geqslant n$.

תהי כעת קבוצה $C = \left(\mathbf{x}_1, \ldots, \mathbf{x}_{n+1}\right) \subseteq \{0,1\}^n$ בגודל $n+1$. נשים לב ש־$\{0,1\}^n \subseteq \mathbb{R}^n$, ולכן (כאשר חושבים על $\mathbb{R}^n$ כעל מרחב אוקלידי) C היא בהכרח קבוצה <u>תלויה ליניארית</u>. לכן קיים צירוף ליניארי לא־טריוויאלי מתאפס של איברי C:

$$\sum_{i=1}^{n+1} a_i \mathbf{x}_i = \bar{0}$$

נניח בה"כ ש־ $a_{n+1} \neq 0$ ונקבל ש־

$$\mathbf{x}_{n+1} = \sum_{i=1}^{n} b_i \mathbf{x}_i \quad \text{☼}$$

כאשר $b_i := -\frac{a_i}{a_{n+1}}$ לכל $i \in [n]$. נתבונן כעת בלייבלים הבאים עבור C:

$$\forall i \in [n+1]: \quad y_i = \begin{cases} b_i \text{mod} 2 & i \in [n] \\ \left(1 + \sum\limits_{j=1}^{n} \left(b_j \text{mod} 2\right)\right) \text{mod} 2 & i = n+1 \end{cases}$$

נניח בשלילה שקיים $v_I \in \{0,1\}^n$ כך שלכל $i \in [n+1]$ מתקיים:

$$y_i = \left\langle \mathbf{x}_i \mid v_I \right\rangle \text{mod} 2 = \begin{cases} b_i \text{mod} 2 & i \in [n] \\ \left(1 + \sum\limits_{j=1}^{n} \left(b_j \text{mod} 2\right)\right) \text{mod} 2 & i = n+1 \end{cases}$$

ונקבל (תוך שימוש באריתמטיקה של mod) ש־

$$\begin{aligned} \left\langle \mathbf{x}_{n+1} \mid v_I \right\rangle \text{mod} 2 &\overset{\text{☼}}{=} \left\langle \sum_{i=1}^{n} b_i \mathbf{x}_i \mid v_I \right\rangle \text{mod} 2 = \left(\sum_{i=1}^{n} b_i \left\langle \mathbf{x}_i \mid v_I \right\rangle \right) \text{mod} 2 = \sum_{i=1}^{n} \left(\left(b_i \text{mod} 2\right) \left\langle \mathbf{x}_i \mid v_I \right\rangle \text{mod} 2 \right) = \\ &\underset{\text{מההנחה בשלילה}}{\underset{\uparrow}{=}} \sum_{i=1}^{n} \left(\left(b_i \text{mod} 2\right) \left(b_i \text{mod} 2\right) \right) = \sum_{i=1}^{n} \left(b_i \text{mod} 2\right)^2 \underset{\left(b_i \text{mod} 2\right) \in \{0,1\}}{\underset{\uparrow}{=}} \sum_{i=1}^{n} \left(b_i \text{mod} 2\right) = \\ &= \left(\sum_{i=1}^{n} \left(b_i \text{mod} 2\right) \right) \text{mod} 2 \neq \left(1 + \sum_{i=1}^{n} \left(b_i \text{mod} 2\right) \right) \text{mod} 2 \underset{\text{לפי הבחירה של הלייבלים}}{\underset{\uparrow}{=}} y_{n+1} \end{aligned}$$

וזו סתירה. לכן לא ניתן לנתץ תת־קבוצות בגודל $n+1$ של $\{0,1\}^n$, וזה אומר ש־ $\text{VC-DIM}\left(\mathcal{H}\right) < n+1$, ובסך הכל קיבלנו ש־

$$\text{VC-DIM}\left(\mathcal{H}\right) = n$$

■

שאלה 4

4. Given an integer k, let $([a_i,b_i])_{i=1}^{k}$ be any set of k intervals on $\mathbb{R}$ and define their union $A=\cup_{i=1}^{k}[a_i,b_i]$. The hypothesis class $\mathcal{H}_{k-intervals}$ includes the functions: $h_A(x):=\mathbb{1}_{[x\in A]}$, for all choices of k intervals. Find the VC-dimension of $\mathcal{H}_{k-intervals}$ and prove your answer. Show that if we let A be any finite union of intervals (i.e. k is unlimited), then the resulting class $\mathcal{H}_{intervals}$ has VC-dimension ∞.

ה־VC-dimention של המחלקה $\mathcal{H}_{k\text{-intervals}}$ הוא $2k$.

הוכחה: נראה שהקבוצה $C=\{x_1,\dots,x_{2k}\}\subseteq\mathbb{R}$ כאשר $x_i=i$ לכל $i\in[2k]$ (קבוצת כל המספרים הטבעיים עד $2k$, כולל) מנותצת ע"י $\mathcal{H}_{k\text{-intervals}}$:

לכל $2k$ לייבלים $y_1,\dots,y_{2k}$ עבור איברי C נוכל לבנות היפותזה $h_A\in\mathcal{H}_{k\text{-intervals}}$ כך ש־ $h_A(x_i)=y_i$ לכל $i\in[2k]$, באופן הבא:

אם $y_i=0$ לכל $i\in[2k]$ אז נגדיר את A להיות איחוד של k קטעים שמוכלים ב־ $(-\infty,0]$, ונקבל ש־ $A\notin x_i$ לכל $i\in[2k]$, כלומר ש־ $h_A(x_i)=0=y_i$ לכל i.

אחרת, הקבוצה A שעליה מבוססת h_A תהיה איחוד של קטעים המכסים רצפים של x_i־ים מתוך C שהלייבל שלהם הוא 1. ניתן לבנות אותה באופן הבא:

בכל שלב $\ell\in[k]$, אם קיים $i\in[2k]\setminus[b_{\ell-1}]$ כך ש־ $y_i=1$ נגדיר את הקטע $[a_\ell,b_\ell]$ ע"י־

$$a_\ell=\min\{i\in[2k]\setminus[b_{\ell-1}]\ :\ y_i=1\}$$
$$b_\ell=\max\{i\in[2k]\setminus[a_\ell]\ :\ y_j=1\ \forall\, j\in[i]\setminus[a_\ell]\ ,\ y_{i+1}=0\}$$

כאשר מגדירים $b_0=0$ (ולכן $[b_0]=\emptyset$), ו־ $y_{2k+1}=0$.

אם לא קיים i כנ"ל אז הבניה תעצר (ופורמלית נגדיר $[a_\ell,b_\ell]=\emptyset$). A מוגדרת ע"י:

$$A=\bigcup_{\ell=1}^{k}[a_\ell,b_\ell]$$

מכיוון שבמקרה הזה אנחנו מניחים שקיים $i\in[2k]$ כך ש־ $y_i=1$, אז מתקבל בבניה הזאת לפחות קטע $[a_\ell,b_\ell]$ אחד.

נשים לב שמספר הקטעים הוא לכל היותר k (זו הסיבה לכך ש־ $\ell\in[k]$) כי בין כל שני קטעים $[a_\ell,b_\ell],[a_{\ell+1},b_{\ell+1}]$ קיים לפי הבניה $i\in[a_{\ell+1}-1]\setminus[b_\ell]$ (לפחות אחד) כך ש־ $y_i=0$. לכן אם נניח בשלילה שניתן לבנות כך $k+1$ קטעים, נקבל שיש ב־C לפחות k נקודות x_i כך ש־ $y_i=0$, ובנוסף בכל אחד מהקטעים יש לפחות נקודה אחת x_i כך ש־ $y_i=1$, ומכאן נקבל שיש ב־C לפחות $2k+1$ נקודות, בסתירה להגדרתה.

עכשיו מהגדרת A נובע שלכל $x_i\in C$ מתקיים ש־ $x_i\in A$ אם ורק אם $y_i=1$, או באופן שקול $x_i\notin A$ אם ורק אם $y_i=0$, ולכן

$$h_A(x_i)=y_i$$

כלומר $\mathcal{H}_{k\text{-intervals}}$ מנתצת את C, ולכן $\text{VC-DIM}(\mathcal{H}_{k\text{-intervals}})\geqslant 2k$.

תהי כעת קבוצה $C=\{x_1,\dots,x_{2k+1}\}\subseteq\mathbb{R}$ כלשהי. נתבונן בלייבלים הבאים עבור הנקודות ב־C:

$$y_i=\begin{cases}1 & i=1\,(\text{mod}2)\\ 0 & i=0\,(\text{mod}2)\end{cases}$$
$$\overset{1}{y_1},\overset{0}{y_2},\overset{1}{y_3}\dots,\overset{0}{y_{2k}},\overset{1}{y_{2k+1}}$$

כלומר הלייבלים הם 0 לכל x_i עם אינדקס זוגי ו־1 לכל x_i עם אינדקס אי־זוגי. לכן יש ב־C בדיוק k נקודות עם לייבל 0 ו־$k+1$ נקודות עם לייבל 1.

נניח בשלילה שקיימת $h_A \in \mathcal{H}_{k\text{-intervals}}$ כך ש־ $h_A(x_i) = y_i$ לכל $i \in [2k+1]$.
מכיוון שיש $k+1$ נקודות ב־C עם לייבל 1, ב־A יש k קטעים, וכל נקודה עם לייבל 1 חייבת להיות מוכלת בקטע כלשהו ב־A אז (מעיקרון שובך היונים) קיים קטע $[a_\ell, b_\ell] \subseteq A$ וקיימות $x_i, x_j \in C$ **שונות** (בה"כ $x_i < x_j$) כך ש־

$$x_i, x_j \in [a_\ell, b_\ell] \quad \wedge \quad y_i = y_j = 1$$

מבחירת הלייבלים נובע שקיימת נקודה $x_d \in C$ כך ש־ $x_i < x_d < x_j$ וגם $y_d = 0$, ולכן:

$$y_d \in [a_\ell, b_\ell] \subseteq A$$

כלומר מתקיים ש־ $h_A(x_d) = 1 \neq y_d$, וזו סתירה.
לכן כל קבוצה בגודל $2k+1$ לא ניתנת לניתוץ ע"י $\mathcal{H}_{k\text{-intervals}}$, כלומר $\text{VC-DIM}(\mathcal{H}_{k\text{-intervals}}) < 2k$ ובסה"כ קיבלנו ש־

$$\text{VC-DIM}(\mathcal{H}_{k\text{-intervals}}) = 2k$$

■

מימד ה־VC של מחלקת ההיפותזות $\mathcal{H}_{\text{intervals}}$ (שבה אין חסם על מספר הקטעים ב־A) הוא ∞ כי:
לכל $n \in \mathbb{N}$ הקבוצה $C = \{x_1, \ldots, x_{2n}\} \subseteq \mathbb{R}$ כאשר $x_i = i$ לכל $i \in [2n]$ (קבוצת כל המספרים הטבעיים עד $2n$, כולל) **ניתנת לניתוץ** ע"י $\mathcal{H}_{\text{intervals}}$, כי לכל סדרת לייבלים $y_1, \ldots, y_{2n}$ עבור הנקודות ב־C נוכל לבנות את A כאיחוד של קטעים **בדיוק** כמו שעשינו בחלק הראשון של ההוכחה לעיל. נצטרך להשתמש ב־n קטעים לכל היותר בבניה הזאת, וזה אפשרי מכיוון שאין חסם על מספר הקטעים ב־A במחלקת ההיפותזות הזו.
מכיוון שלכל $n \in \mathbb{N}$ קיימת קבוצה $C \subseteq \mathbb{R}$ בגודל $|C| = 2n$ הניתנת לניתוץ ע"י $\mathcal{H}_{\text{intervals}}$, אז

$$\forall n \in \mathbb{N}: \quad \text{VC-DIM}(\mathcal{H}_{\text{intervals}}) \geqslant 2n \quad \Longrightarrow \quad \text{VC-DIM}(\mathcal{H}_{\text{intervals}}) = \infty$$

מונוטוניות

שאלה 5

> 5. Let $\mathcal{H}$ be a hypothesis class for a binary classification task. Suppose that $\mathcal{H}$ is PAC learnable and its sample complexity is given by $m_{\mathcal{H}}(\cdot,\cdot)$. Show that $m_{\mathcal{H}}$ is monotonically non-increasing in each of its parameters. That is:
>   - Show that given $\delta \in (0,1)$, and given $0 < \varepsilon_1 \leq \varepsilon_2 < 1$, we have that $m_{\mathcal{H}}(\varepsilon_1, \delta) \geq m_{\mathcal{H}}(\varepsilon_2, \delta)$.
>   - Similarly, show that given $\varepsilon \in (0,1)$, and given $0 < \delta_1 \leq \delta_2 < 1$, we have that $m_{\mathcal{H}}(\varepsilon, \delta_1) \geq m_{\mathcal{H}}(\varepsilon, \delta_2)$.

- **הוכחה:** נתונה $\mathcal{H}$ למידה PAC. יהי $\mathcal{A}$ אלגוריתם שלומד אותה, יהי $\delta \in (0,1)$ ויהיו $0 < \varepsilon_1 \leqslant \varepsilon_2 < 1$.
לפי ההגדרה של למידות־PAC ושל ה־sample complexity מתקיים לכל $m \in \mathbb{N}$:

$$\text{I)} \quad m \geqslant m_{\mathcal{H}}(\varepsilon_1, \delta) \quad \Longleftrightarrow \quad \mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(\mathcal{A}(S)) \leqslant \varepsilon_1] \geqslant 1-\delta$$
$$\text{II)} \quad m \geqslant m_{\mathcal{H}}(\varepsilon_2, \delta) \quad \Longleftrightarrow \quad \mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(\mathcal{A}(S)) \leqslant \varepsilon_2] \geqslant 1-\delta$$

ומכך ש־ $\varepsilon_1 \leqslant \varepsilon_2$ נובע ש־ $\{L_{\mathcal{D}}(\mathcal{A}(S)) \leqslant \varepsilon_1\} \subseteq \{L_{\mathcal{D}}(\mathcal{A}(S)) \leqslant \varepsilon_2\}$.
לכן עבור $m = m_{\mathcal{H}}(\varepsilon_1, \delta)$ מתקיים לפי I:

$$1-\delta \leqslant \mathbb{P}_{S\sim\mathcal{D}^{m_{\mathcal{H}}(\varepsilon_1,\delta)}}[L_{\mathcal{D}}(\mathcal{A}(S)) \leqslant \varepsilon_1] \leqslant \mathbb{P}_{S\sim\mathcal{D}^{m_{\mathcal{H}}(\varepsilon_1,\delta)}}[L_{\mathcal{D}}(\mathcal{A}(S)) \leqslant \varepsilon_2]$$

ומכאן לפי II -

$$m_{\mathcal{H}}\left(\varepsilon_1,\delta\right) \geqslant m_{\mathcal{H}}\left(\varepsilon_2,\delta\right)$$

■

- **הוכחה:** נתונה $\mathcal{H}$ למידה PAC. יהי $\mathcal{A}$ אלגוריתם שלומד אותה, יהי $\varepsilon \in (0,1)$ ויהיו $0 < \delta_1 \leqslant \delta_2 \leqslant 1$.

לפי ההגדרה של למידות-PAC מתקיים לכל $m \in \mathbb{N}$:

$$\begin{aligned} \text{I)} \quad m \geqslant m_{\mathcal{H}}\left(\varepsilon,\delta_1\right) \quad &\Longleftrightarrow \quad \mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right) \leqslant \varepsilon\right] \geqslant 1-\delta_1 \\ \text{II)} \quad m \geqslant m_{\mathcal{H}}\left(\varepsilon,\delta_2\right) \quad &\Longleftrightarrow \quad \mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right) \leqslant \varepsilon\right] \geqslant 1-\delta_2 \end{aligned}$$

מכך ש- $\delta_1 \leqslant \delta_2$ נובע ש- $1-\delta_2 \geqslant 1-\delta_1$ ולכן עבור $m = m_{\mathcal{H}}\left(\varepsilon,\delta_1\right)$ מתקיים לפי I:

$$\mathbb{P}_{S\sim\mathcal{D}^{m_{\mathcal{H}}(\varepsilon,\delta_1)}}\left[L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right) \leqslant \varepsilon\right] \geqslant 1-\delta_1 \geqslant 1-\delta_2$$

ומכאן לפי II -

$$m_{\mathcal{H}}\left(\varepsilon,\delta_1\right) \geqslant m_{\mathcal{H}}\left(\varepsilon,\delta_2\right)$$

■

שאלה 6

6. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two classes for binary classification, such that $\mathcal{H}_1 \subseteq \mathcal{H}_2$. Show that $VC-dim(\mathcal{H}_1) \leq VC-dim(\mathcal{H}_2)$.

הוכחה: נניח בשלילה ש- $\text{VC-DIM}\left(\mathcal{H}_2\right) > \text{VC-DIM}\left(\mathcal{H}_1\right)$.

נסמן ב-C_1 את הקבוצה בגודל המקסימלי המנותצת ע"י $\mathcal{H}_1$, וב-C_2 את הקבוצה בגודל המקסימלי המנותצת ע"י $\mathcal{H}_2$. אזי:

$$\left|C_1\right| = \text{VC-DIM}\left(\mathcal{H}_1\right) > \text{VC-DIM}\left(\mathcal{H}_2\right) = \left|C_2\right|$$

נסמן $C_1 = \{x_1,\ldots,x_m\}$ אז לפי ההגדרה של צמצום של מחלקת היפתזות:

$$\begin{aligned} \mathcal{H}_1\Big|_{C_1} &= \left\{\left(h\left(x_1\right),\ldots,h\left(x_m\right)\right) \ : \ h \in \mathcal{H}_1\right\} \\ \mathcal{H}_2\Big|_{C_1} &= \left\{\left(h\left(x_1\right),\ldots,h\left(x_m\right)\right) \ : \ h \in \mathcal{H}_2\right\} \end{aligned}$$

ולכן $\mathcal{H}_1\Big|_{C_1} \subseteq \mathcal{H}_2\Big|_{C_1}$. אבל מכך ש-$\mathcal{H}_1$ מנתצת את C_1 נובע ש-

$$2^{\left|C_1\right|} = \left|\mathcal{H}_1\Big|_{C_1}\right| \leqslant \left|\mathcal{H}_2\Big|_{C_1}\right| \leqslant 2^{\left|C_1\right|} \quad \Longrightarrow \quad \left|\mathcal{H}_2\Big|_{C_1}\right| = 2^{\left|C_1\right|}$$

ולכן $\mathcal{H}_2$ מתנצת את C_1, וזו סתירה לכך ש-C_2 היא הקבוצה המקסימלית המנותצת על ידה (כי ראינו ש- $\left|C_1\right| > \left|C_2\right|$).

לכן $\text{VC-DIM}\left(\mathcal{H}_1\right) \leqslant \text{VC-DIM}\left(\mathcal{H}_2\right)$.

■

Agnostic-PAC

שאלה 7

7. Prove that if $\mathcal{H}$ has the uniform convergence property with function $m_{\mathcal{H}}^{UC}:(0,1)^2 \to \mathbb{N}$, then $\mathcal{H}$ is Agnostic-PAC learnable with sample complexity $m_{\mathcal{H}}(\varepsilon,\delta) \leq m^{UC}(\varepsilon/2,\delta)$.

הוכחה: תהי $\mathcal{H}$ מחלקת היפותזות שמקיימת את תכונת ההתכנסות בהחלט עם $m_{\mathcal{H}}^{UC}:(0,1)^2 \to \mathbb{N}$,
ביחס לפונקציית loss כלשהי $\ell:\mathcal{H}\times(\mathcal{X}\times\mathcal{Y})\to[0,\infty)$.
יהיו $\varepsilon,\delta\in(0,1)$, ותהי $\mathcal{D}$ פונקציית התפלגות על $\mathcal{X}\times\mathcal{Y}$. אז לפי ההגדרה לכל $m\geqslant m_{\mathcal{H}}^{UC}\left(\frac{\varepsilon}{2},\delta\right)$ מתקיים ש־

$$\mathcal{D}^m\left(\left\{S\in(\mathcal{X}\times\mathcal{Y})^m \mid S \text{ is } \frac{\varepsilon}{2}\text{-representative}\right\}\right)\geqslant 1-\delta$$

נבחר את אלגוריתם הלמידה $\mathcal{A}_{\mathrm{ERM}}$, כלומר את:

$$\mathcal{A}_{\mathrm{ERM}}: S\mapsto h \quad \text{s.t.:} \quad h\in\left\{\underset{h\in\mathcal{H}}{\operatorname{argmin}} L_S(h)\right\}$$

כעת לפי למה 4.5.2 מספר הקורס (שראינו גם בהרצאה), אם סט דגימות S הוא $\frac{\varepsilon}{2}$־מייצג את $\mathcal{D},\mathcal{H},\ell$ אז $\mathcal{A}_{\mathrm{ERM}}(S):=h_s$ מקיים ש־

$$L_{\mathcal{D}}(h_s)\leqslant \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h)+\varepsilon$$

ולכן:

$$\left\{S\in(\mathcal{X}\times\mathcal{Y})^m \mid S \text{ is } \frac{\varepsilon}{2}\text{-representative}\right\}\subseteq\left\{S\in(\mathcal{X}\times\mathcal{Y})^m \mid L_{\mathcal{D}}(h_s)\leqslant \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h)+\varepsilon\right\}$$

ולכן ממונוטוניות פונקציית ההתפלגות נובע שמתקיים לכל $m\geqslant m_{\mathcal{H}}^{UC}\left(\frac{\varepsilon}{2},\delta\right)$:

$$1-\delta\leqslant\mathcal{D}^m\left(\left\{S\in(\mathcal{X}\times\mathcal{Y})^m \mid S \text{ is } \frac{\varepsilon}{2}\text{-representative}\right\}\right)\leqslant\mathcal{D}^m\left(\left\{S\in(\mathcal{X}\times\mathcal{Y})^m \mid L_{\mathcal{D}}(h_s)\leqslant \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h)+\varepsilon\right\}\right)$$

ובפרט זה מתקיים עבור $m=m_{\mathcal{H}}^{UC}\left(\frac{\varepsilon}{2},\delta\right)$.
כלומר הראינו שלכל $\varepsilon,\delta\in(0,1)$, ולכל $\mathcal{D}$ פונקציית התפלגות על $\mathcal{X}\times\mathcal{Y}$, קיים אלגוריתם למידה ־ $\mathcal{A}_{\mathrm{ERM}}$ ־ המקיים לכל $m\geqslant m_{\mathcal{H}}^{UC}\left(\frac{\varepsilon}{2},\delta\right)$:

$$\mathcal{D}^m\left(\left\{S\in(\mathcal{X}\times\mathcal{Y})^m \mid L_{\mathcal{D}}(\mathcal{A}_{\mathrm{ERM}}(S))\leqslant \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h)+\varepsilon\right\}\right)\geqslant 1-\delta$$

ולכן לפי ההגדרה $\mathcal{H}$ היא Agnostic-PAC למידה עם sample complexity המקיים ש־ $m_{\mathcal{H}}(\varepsilon,\delta)\leqslant m_{\mathcal{H}}^{UC}\left(\frac{\varepsilon}{2},\delta\right)$, כנדרש. ■

שאלה 8

8. Let $\mathcal{H}$ be a hypothesis class over a domain $\mathcal{Z}=\mathcal{X}\times\{\pm1\}$, and consider the 0-1 loss function. Assume that there exists a function $m_{\mathcal{H}}$, for which it holds that for every distribution $\mathcal{D}$ over $\mathcal{Z}$ there is an algorithm $\mathcal{A}$ with the following property: when running $\mathcal{A}$ on $m\geq m_{\mathcal{H}}$ i.i.d. examples drawn from $\mathcal{D}$, it is guaranteed to return, with probability at least $1-\delta$, a hypothesis $h_S:\mathcal{X}\to\{\pm1\}$ with $L_{\mathcal{D}}(h_S)\leq\min_{h\in\mathcal{H}}L_{\mathcal{D}}(h)+\varepsilon$. Is $\mathcal{H}$ agnostic PAC learnable? Prove or show a counter example.

<u>פתרון:</u>

מחלקת היפותזות $\mathcal{H}$ שמקיימת את התנאי הזה היא לא בהכרח למידה־agnostic-PAC עם פונקציית ה־loss ־ ℓ_{0-1}.

הסיבה לכך היא שהתנאי הזה מאפשר לאלגוריתם הלמידה $\mathcal{A}$ "להכיר" את ההתפלגות $\mathcal{D}$ על $\mathcal{Z}$, ובפרט מאפשר לו לחשב את $L_{\mathcal{D}}(h)$ לכל $h \in \mathcal{H}$, דבר שאינו אפשרי ב־framework של agnostic PAC. דוגמא נגדית:

נגדיר $\mathcal{X} = \mathbb{R}$, וניקח את $\mathcal{H}$ להיות מחלקת כל הפונקציות מ־$\mathcal{X}$ ל־ $\{\pm 1\}$ (Unstructured Hypotheses Class), כלומר $\mathcal{H} = \mathbb{R}^{\{\pm 1\}}$.

$\mathcal{H}$ מקיימת את התנאי הנתון:

נגדיר $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ ע"י $m_{\mathcal{H}}(\varepsilon, \delta) = 1$ לכל $\varepsilon, \delta \in (0,1)$.

לכל התפלגות $\mathcal{D}$ על $\mathcal{Z}$ ניקח אלגוריתם $\mathcal{A} : \mathcal{Z}^m \to \mathcal{H}$ המוגדר לכל $\mathcal{Z}^m \supseteq S = \{(x_i, y_i)\}_{i=1}^{m}$ ע"י:

$$\mathcal{A}(S) = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \{L_{\mathcal{D}}(h)\}$$

נקבל שלכל $\varepsilon, \delta \in (0,1)$, כאשר מפעילים את $\mathcal{A}$ על $S \in \mathcal{Z}^m$ עבור $m \geqslant m_{\mathcal{H}}(\varepsilon, \delta)$ הוא מחזיר בהסתברות $1 - \delta \leqslant 1$ היפותזה $h_S : \mathcal{X} \to \{\pm 1\}$ כך ש־

$$L_{\mathcal{D}}(h_S) \overset{\text{מהגדרת } \mathcal{A}}{\overset{\downarrow}{=}} \min_{h \in \mathcal{H}} \{L_{\mathcal{D}}(h)\} \leqslant \min_{h \in \mathcal{H}} \{L_{\mathcal{D}}(h)\} + \varepsilon$$

מצד שני, ראינו בהרצאה ש־$\mathcal{H}$ היא לא למידה־PAC.

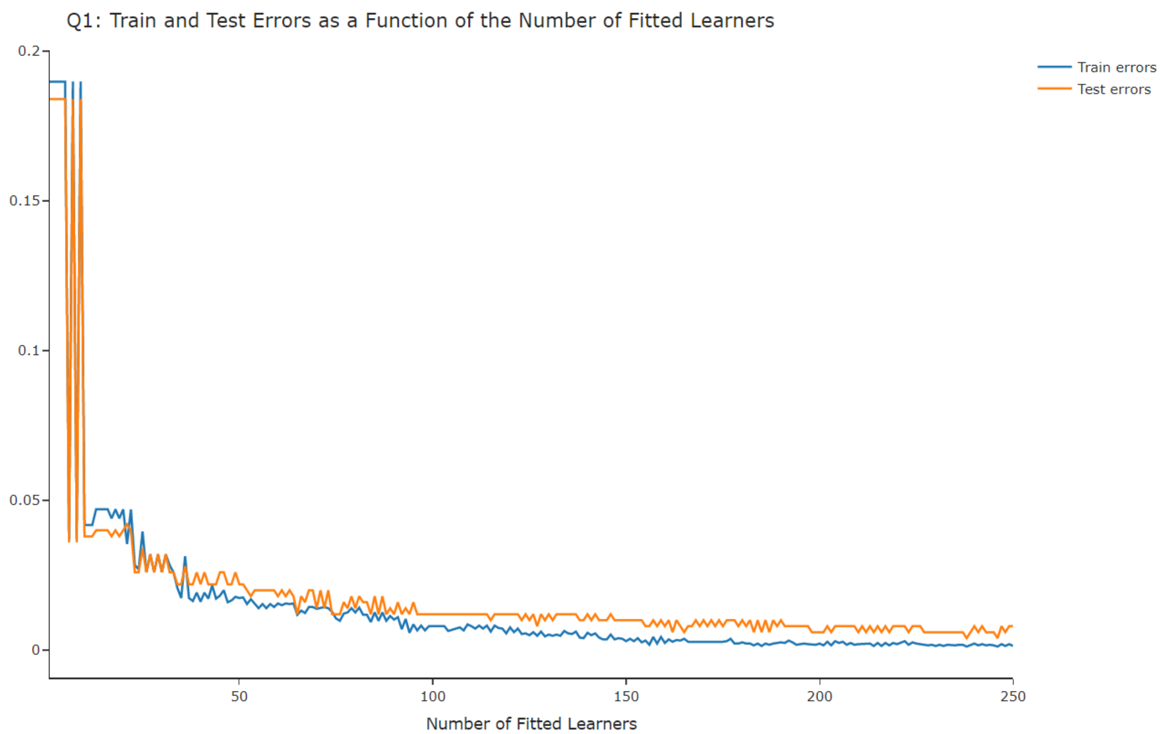
ראינו בנוסף שאם מחלקת היפותזות היא למידה־agnostic-PAC עם פונקציית ה־loss ־ ℓ_{0-1} אז היא גם למידה PAC.

לכן מכך ש־$\mathcal{H}$ לא למידה PAC נובע שהיא **לא** למידה ־agnostic-PAC עם ℓ_{0-1}.

חלק פרקטי

שאלה 1

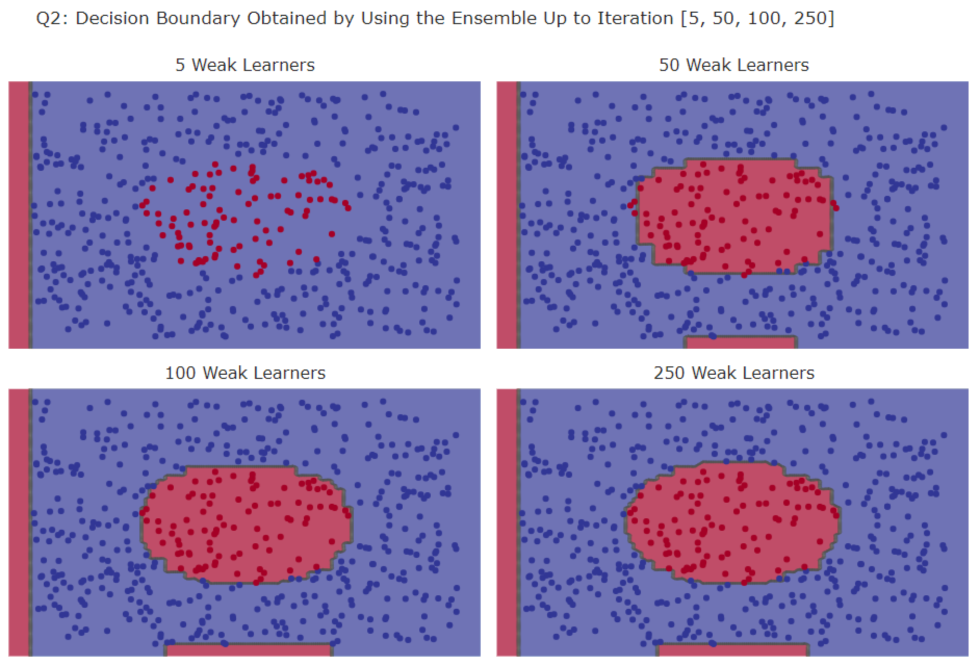
גרף המציג את השגיאה (misclassification error) של Adaboost על דגימות האימון (בכחול) ועל דגימות המבחן (בכתום) כפונקציה של מספר המודלים החלשים שנעשה בהם שימוש באלגוריתם:



רואים מהגרף שככל שמספר המודלים החלשים שמשתתפים באלגוריתם **גדול** יותר ־ השגיאה נהיית **קטנה** יותר. בנוסף, רואים שהשגיאה של המודל על דגימות האימון קטנה יותר מהשגיאה שלו על דגימות המבחן. עם זאת, הפער בשגיאה לא גדול כ"כ ולכן אפשר אולי להסיק שהאלגוריתם מכליל לא רע ולא מתאים את עצמו מדי לדאטא האימון (כלומר לא סובל מ־over fitting).

שאלה 2

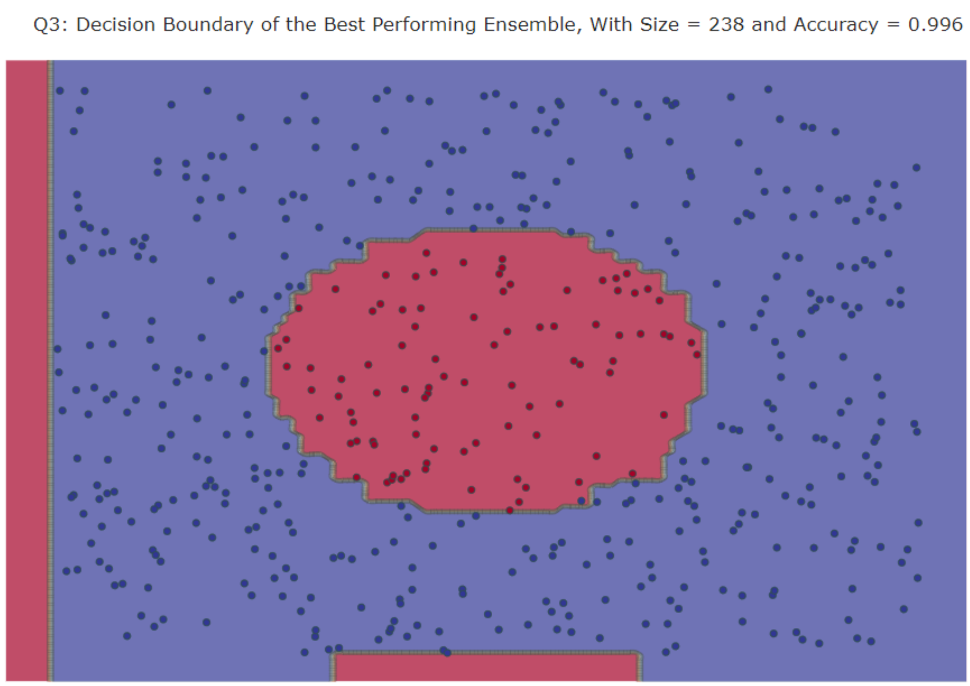
גרף המציג את ה־decision boundary של המודל עבור $5, 50, 100, 250$ לומדים חלשים שמשתתפים בו, כאשר הנקודות הן דגימות **המבחן**:



רואים מהגרף שככל שיותר לומדים חלשים משתתפים במודל, הוא נהיה אקספרסיבי יותר ומצליח לסווג טוב יותר את דגימות המבחן.

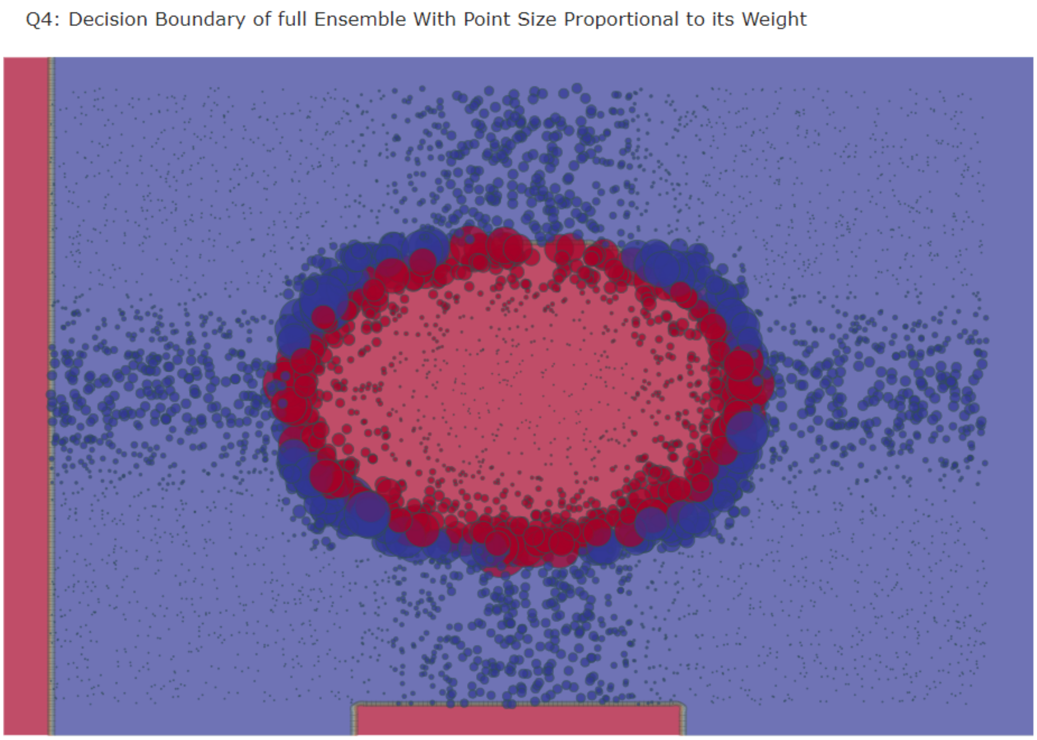
שאלה 3

גרף המציג את ה־decision boundary של המודל עבור האנסמבל בגודל האופטימלי (מתוך 250 האפשרויות):



שאלה 4

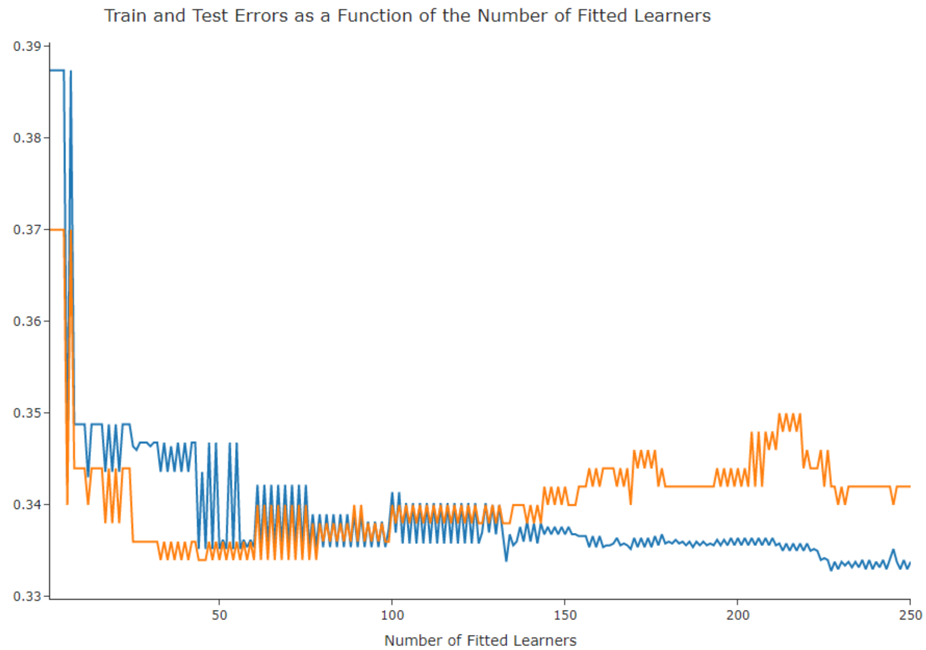
גרף המציב את ה־decision boundary של המודל המלא על דאטא **האימון**, כאשר גודל הנקודות פורפורציונלי להתפלגות שלהן בסוף ריצת ה־adaboost:



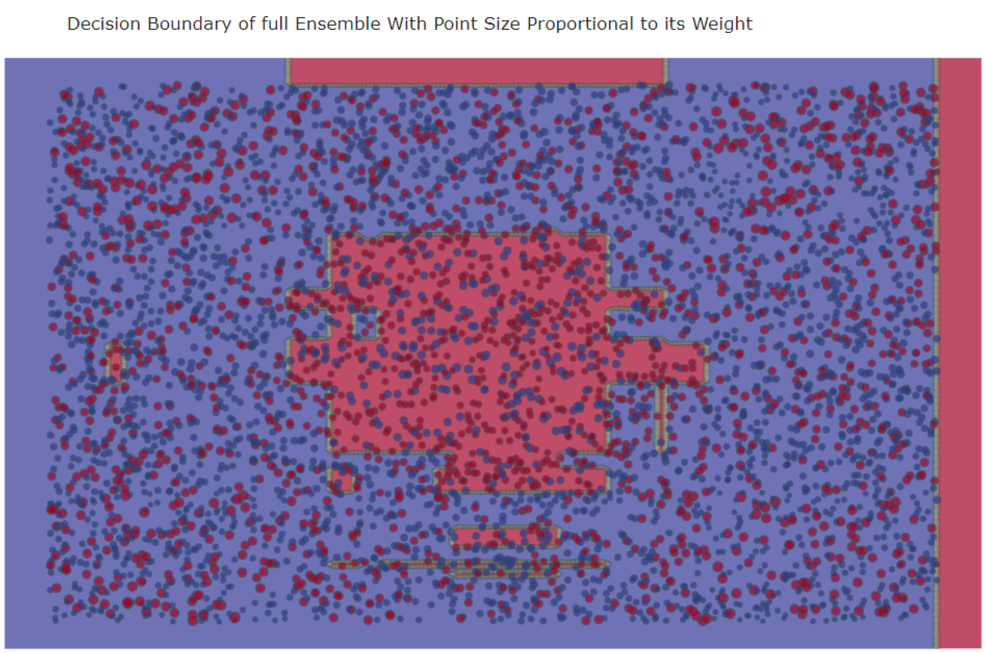
ניתן לראות שהנקודות עם המשקל הגדול ביותר בסוף ריצת האלגוריתם הן הנקודות שנמצאות על "הגבול" בין המחלקות. המשמעות של זה היא שמרבית הטעויות שהלומדים החלשים במודל עשו במהלך ה־fit נעשו על הנקודות הללו. אפשר להסיק מכך (וזו מסקנה הגיונית לחלוטין בלי קשר לגרף) שהנקודות שהכי "מאתגרות" את המודל הן הנקודות שנמצאות על הגבול בין המחלקות, ואילו הנקודות שהכי "קל" למודל לסווג הן הנקודות שנמצאות בסביבה הומוגנית יותר, כלומר רחוקות מנקודות עם לייבל ששונה מהלייבל שלהן.

שאלה 5

הגרפים המתאימים לשאלות 1 ו־4 עבור דאטא שנוצר עם רעש ברמה 0.4:



באופן דומה למקרה ללא הרעש, ככל שמספר הלומדים החלשים משתתפים במודל - השגיאה על סט האימון קטנה.
אבל במקרה הזה, בשלב מסויים ניתן לראות בגרף שהשגיאה על סט המבחן מתחילה לגדול. ניתן להסביר זאת בכך שהמודל מתחיל לעשות overfitting לדאטא וללמוד את הרעש.
ניתן לראות שהפער בין השגיאה של המודל על סט האימון והשגיאה שלו על סט המבחן גדל ככל שהוא עושה שימוש ביותר לומדים חלשים.
במונחי bias-variance אפשר לומר שככל שהמורכבות של המודל גדולה יותר, כלומר ככל שהוא עושה שימוש ב־ensemble גדול יותר, ה־variance שלו גדל בעוד שה־bias שלו קטן.



ניתן לראות שכאשר הדאטא נדגם עם רעש (ברמה די' גבוהה) המודל מתקשה יותר להפריד אותו.

מכיוון שיש נקודות רבות שהלומדים החלשים טועים עליהן, אז הגודל של רוב הנקודות בגרף השני הוא יחסית דומה.