IML ex4

.1

צ"ל:

$$\forall \varepsilon,\delta>0 \quad \exists m(\varepsilon,\delta) \text{ s.t. } \forall m\geq m(\varepsilon,\delta)$$
$$P_{S\sim D^m}\left[L_D(A(S))\leq \varepsilon\right]\geq 1-\delta$$
$$\Longleftrightarrow$$
$$\lim_{m\to\infty} E_{S\to D^m}\left[L_D(A(S))\right]=0$$

הוכחה:

$\Downarrow$: $\forall \varepsilon,\delta>0$, $\forall m\geq m(\varepsilon,\delta)$ מתקיים ש $\cancel{m(\varepsilon,\delta)=\frac{1}{\delta+\varepsilon}}$

$$P_{S\sim D^m}(L_D(A(S))\leq\varepsilon)\geq 1-\delta \to P_{S\sim D^m}(L_D(A(S))>\varepsilon)\leq\delta$$

$$E_{S\sim D^m}(L_D(A(S)))=\int_0^\infty P(L_D(A(S))>t)\,dt=$$

$$=\int_0^\infty P(L_D(A(S))>t)\,dt\leq \delta+\int_0^\varepsilon P(L_D(A(S))>t)\,dt\leq \boxed{\delta+\varepsilon}$$

לכן $0\xleftarrow{m\to\infty}0\leq E_{S\sim D^m}(L_D(A_S))\leq \delta+\varepsilon\leq\frac{1}{m}\xrightarrow{m\to\infty}0$

$L_D\geq 0$

וממשפט הסנדוויץ' $E_{S\sim D^m}(L_D(A_S))\xrightarrow{m\to\infty}0$

$\Uparrow$:

$$\lim_{m\to\infty} E_{S\to D^m}\left[L_D(A(S))\right]=0$$

לפי מרקוב, לכל $a>0$: $\frac{E_{S\to D^m}\left[L_D(A(S))\right]}{a}\geq P(L_D(A(S))\geq a)$

יהיו $\varepsilon,\delta>0$. אז:

$$P(L_D(A(S))\geq\varepsilon)\leq\frac{E_{S\to D^m}\left[L_D(A(S))\right]}{\varepsilon}\xrightarrow{m\to\infty}0$$

לכן קיים m גדול מספיק כך ש נמצא בקטע $[0,\bar\delta]$

לכל $\bar\delta\in\mathbb{R}$ ובפרט $\bar\delta=\delta$. לכן קיים $m(\varepsilon,\delta)$ גדול מספיק כך ש:

$$P(L_D(A(S))\geq\varepsilon)\leq\delta \Rightarrow P(L_D(A(S))\leq\varepsilon)\geq 1-\delta$$

$\mathcal{X}=\mathbb{R}^2$ $\mathcal{Y}=\{0,1\}$ $\mathcal{H}=\{h_r : r\in\mathbb{R}_+\}$ 2

כאשר $h_r(\bar{x})=\mathbb{1}_{[\|\bar{x}\|_2\le r]}$

$m_{\mathcal{H}}(\epsilon,\delta)\le\frac{\log(\frac{1}{\delta})}{\epsilon}$ צ"ל

ושהמחלקה היא PAC למידה.

הוכחה:

נתחיל בסימונים:

$$A: D^m \to (\mathcal{X}\to\mathcal{Y})$$

$$A(S)=h_{r_S}(\bar{x}),\ r_S=\max\{|\bar{x}| \mid x\in S\}$$ כאשר:

יהיו $\epsilon,\delta>0$, ותהי D התפלגות מעל $\mathcal{X}$.

$$r_\epsilon=\max\{r\in\mathbb{R}_+ \mid D[\bar{x} \mid r_{real}\ge\bar{x}\ge r_\epsilon]\ge\epsilon\}$$ כאשר

כעת אם S מכיל $\bar{x}$ כך ש $r_\epsilon\le|\bar{x}|\le r_{real}$

אזי ש- $L_D(A(S))\le\epsilon$ ולכן

$$P_{S\sim D^m}[L_D(A(S))>\epsilon]=P_{S\sim D^m}\left[\bigwedge_{i=1}^m\{|\bar{x}_i|\notin[r_\epsilon,r_{real}]\}\right]=$$

$$=[1-P(|\bar{x}_i|\in[r_\epsilon,r_{real}])]^m\le\exp(-mP(|\bar{x}_i|\in[r_\epsilon,r_{real}]))$$

$$P(|\bar{x}_i|\in[r_\epsilon,r_{real}])\ge\epsilon$$ לפי ההגדרה

2. המשך. ולכן:

$$P_{S\sim D^m}[L_D(A(S))>\epsilon] \le e^{-\epsilon m}$$

ז"א

$$P_{S\sim D^m}[L_D(A(S))\le\epsilon] \ge 1-e^{-\epsilon m}$$

ולכן עבור

$$m_H(\epsilon,\delta)=\frac{\log(\frac{1}{\delta})}{\epsilon}$$

נקבל ש:

$$P_{S\sim D^m}[L_D(A(S))\le\epsilon] \ge 1-\delta$$

ומאחר ש m_H מינימלי המקיים תכונה

כן נקבל שהמחלקה PAC למידה

עם

$$\boxed{m_H(\epsilon,\delta) \le \frac{\log(\frac{1}{\delta})}{\epsilon}}$$

מ.ש.ל

$\cancel{e^{-\epsilon m} \le \delta}$

$\cancel{-\epsilon m \le \ln\delta}$

$\cancel{-m \le \frac{\ln\delta}{\epsilon}}$

$\cancel{m \ge \frac{\ln(\frac{1}{\delta})}{\epsilon}}$

.3 $\mathcal{X}=\{0,1\}^n \quad \mathcal{Y}=\{0,1\}$

לכל $I\subseteq[n]$ נגדיר $h_I(\bar{x})=(\sum_{i\in I}x_i) \bmod 2$.

נראה ש $VC\text{-}dim(H_n)=n$.

$$\bar{e}_i=\begin{pmatrix}0\\0\\\vdots\\1\\\vdots\\0\end{pmatrix} \text{ אינדקס } i$$

נתבונן ב $C=\{\bar{e}_i\}_{i=1}^n$

~~xxxx~~ יהי סדר מבוקש $\bar{y}\in\{0,1\}^n$.

אזי נגדיר $I=\{i \mid y_i=1\}$ ונקבל

לכל $i\in[n]$: $h_I(\bar{x}_i)=y_i$

לכן H_n shatters C.

אניח שקיימת C כך ש $|C|>n$ וגם H_n shatters C

מ- קיים ש $H_n=\{h_I \mid I\subseteq[n]\}\Rightarrow|H_n|=2^n$

נניח ש H_n shat. C. אזי הצמצום של H_n

ל C הוא בגודל $2^{|C|}$.

נשים לב ש $|H_n|=2^n<2^{|C|}$.

סתירה.

מ.ש.ל.

4. יהי $k\in\mathbb{N}$,

$$H_{k\text{-intervals}} = \left\{ h_A(x) = \mathbb{1}_{x\in A} \;\middle|\; A=\bigcup_{i=1}^{k} [a_i,b_i] ,\ a_i<b_i \in\mathbb{R} \right\}$$

אז נראה $VC\text{-}dim(H_{k\text{-intervals}}) = 2k$

נתבונן ב $C=(1,2,\dots,2k)$

יהי פלט $y\in\{0,1\}^{2k}$. נגדיר

$$[a_i,b_i] = \begin{cases} [i-\frac{1}{2}, i+\frac{1}{2}] & y_i=1 \\ [-1,0] & \text{אחרת} \end{cases}$$

ונקבל ש $h_A(x_i)=y_i$ לכל $x_i\in C$.

נניח בשלילה שקיימת $C\subseteq\mathbb{R}$ כך ש $|C|\geq 2k$.

אזי יש ב C לפחות $2k+1$ נקודות,

נסמן $x_1<\dots<x_{2k+1}$. נתבונן בפלט הבא:

$$y_i = \begin{cases} 1 & i \text{ אי-זוגי} \\ 0 & i \text{ זוגי} \end{cases}$$

נניח שקיימים k קטעים שמרכיבים את A כך ש $h_A(x_i)=y_i$ $\forall i$

בפרט נובע שלכל i אי-זוגי $x_i\in[a_j,b_j]$ קטע כלשהו. ויש לנו $k+1$ אינדקסים אי-זוגיים.

נניח בשלילה שיש שני אינדקסים אי-זוגיים i, k עוקבים יש קטע יחיד שמכיל את שניהם. אזי

נשים לב $x_i<x_{i+1}<x_k$ ולכן $y_{i+1}=0$ סתירה. לכן

יש לפחות $k+1$ קטעים, סתירה. לכן $VCdim(H_{k\text{-int}})=2k$

.5 H מחלקת היפותזות, H למידה PAC.

ט. יהי $\delta\in(0,1)$, ויהיו $0<\epsilon_1\leq\epsilon_2<1$.

צ"ל: $m_H(\epsilon_1,\delta)\geq m_H(\epsilon_2,\delta)$

מלמידות-PAC של H נסיק שלכל $m\geq m_H(\epsilon_2,\delta)$

מתקיים $P_{S\sim D^m}[L_D(h_S)\leq\epsilon_2]\geq 1-\delta$

נניח בשלילה ש $m_H(\epsilon_1,\delta)\underset{(*)}{<}m_H(\epsilon_2,\delta)$

נזכר ש $P_{S\sim D^{m_H(\epsilon_2,\delta)}}[L_D(h_S)\leq\epsilon_2]\geq 1-\delta$

וגם

$P_{S\sim D^{m_H(\epsilon_1,\delta)}}[L_D(h_S)\leq\epsilon_1]\geq 1-\delta$

ממינימליות ה- sample-complexity נסיק ש:

$$m_H(\epsilon_1,\delta)\leq m_H(\epsilon_2,\delta)\ \wedge\ m_H(\epsilon_2,\delta)\leq m_H(\epsilon_1,\delta)$$

ולכן הם שווים, בסתירה ל $(*)$.

נ. יהי $\epsilon\in(0,1)$ ויהיו $0<\delta_1\leq\delta_2<1$.

צ"ל $m_H(\epsilon,\delta_1)\geq m_H(\epsilon,\delta_2)$.

נניח בשלילה ש $m_H(\epsilon,\delta_1)<m_H(\epsilon,\delta_2)$

מלמידות PAC של H נסיק שלכל $m\geq m_H(\epsilon,\delta_1)$

$P_{S\sim D^m}[L_D(h_S)\leq\epsilon]\geq 1-\delta_1\geq 1-\delta_2$

ובפרט ל $m=m_H(\epsilon,\delta_2)$.

נזכר גם שלכל $m\geq m_H(\epsilon,\delta_2)$ מתקיים

$P_{S\sim D^m}[L_D(h_S)\leq\epsilon]\geq 1-\delta_2$

ממינימליות ה- sample-comp. נסיק

כמו מקודם $m_H(\epsilon,\delta_1)=m_H(\epsilon,\delta_2)$

בסתירה להנחת השלילה.

מ.ש.ל

נניח שהפונ' ב h_1 ו h_2 הן מהצורה $h: X \to \{0,1\}$.

6. $H_1 \subseteq H_2$. צ"ל $VC\text{-}dim(H_1) \le VC\text{-}dim(H_2)$.

נסמן $d_{1,2} = VC\text{-}dim(H_{1,2})$.

מהגדרת ה $VC\text{-}dim$ נובע שקיימת
C בגודל d_1 כך ש ~~H_1~~
H_1 shatters C

צ"ל שהצמצום של H_1 ל C הוא בגודל $2^{|C|}$.

צ"ל ש $\left|\{h_C \mid h \in H_1\}\right| = 2^{|C|}$

כאשר h_C הוא צמצום H ל- C.

מאחר ש $H_1 \subseteq H_2$ נובע ש

$$\{h_C \mid h \in H_1\} \subseteq \{h_C \mid h \in H_2 \supseteq H_1\} \subseteq 2^C$$

ולכן $\left|\{h_C \mid h \in H_2\}\right| = 2^{|C|}$

ולכן גם H_2 shatters C

ולכן מהגדרת ה $VC\text{-}dim$

$$\boxed{d_2 \ge d_1}$$

.7 H מתכנסת במ"ש עם $m_H^{UC}:(0,1)^2\to\mathbb{N}$.

צ"ל: H למידה-PAC עם $m_H(\epsilon,\delta)\le m_H^{UC}(\frac{\epsilon}{2},\delta)$.

ע"פ נתון שלכל $0<\epsilon,\delta<1$ ולכל $m\ge m_H^{UC}(\epsilon,\delta)$

$$D^m(\{S\in(X\times Y)^m \mid \forall h\in H\ |L_S(h)-L_D(h)|<\epsilon\})\ge 1-\delta$$

ולכן עבור $m\ge m_H^{UC}(\frac{\epsilon}{2},\delta)$ מתקיים

$$D^m(\{S\in(X\times Y)^m \mid \forall h\in H\ |L_S(h)-L_D(h)|<\frac{\epsilon}{2}\})\ge 1-\delta$$

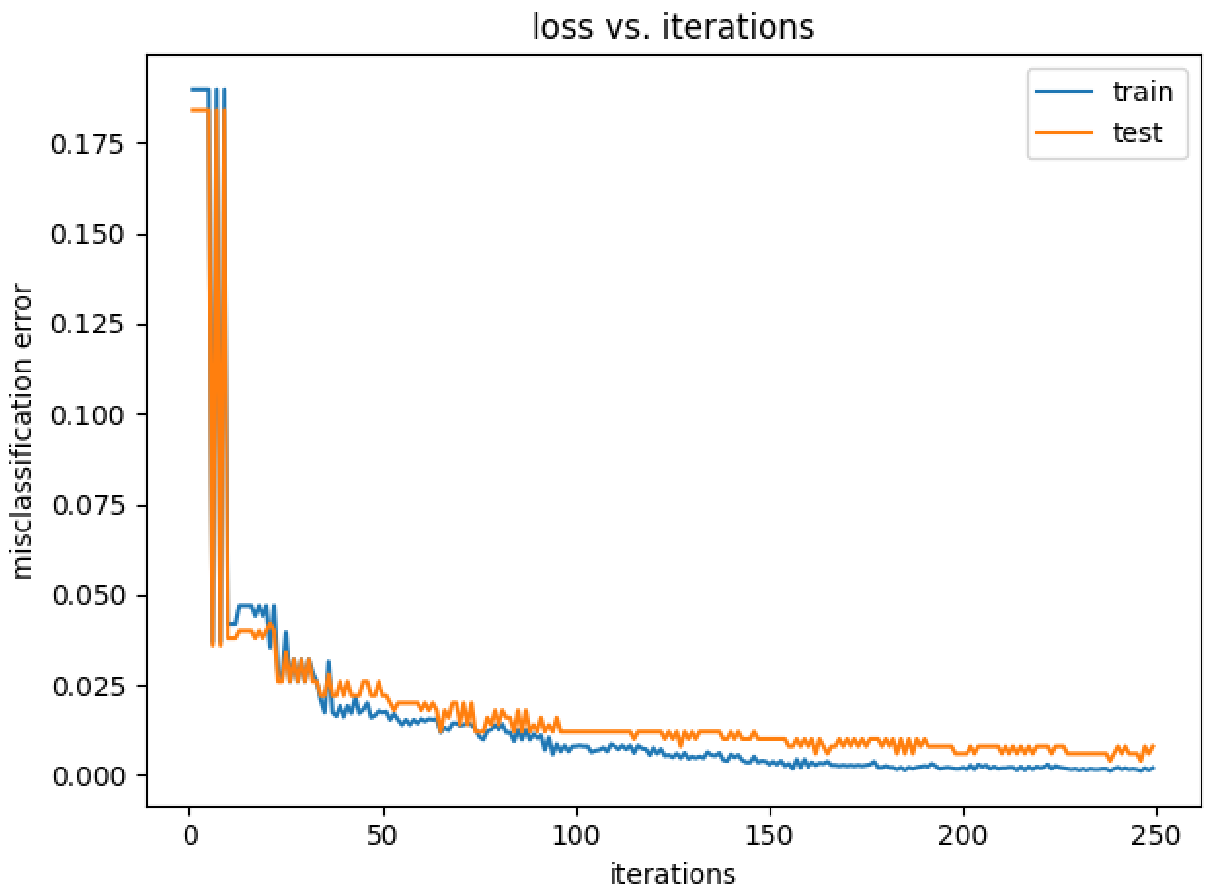
ולכן לכל $m\ge m_H^{UC}(\frac{\epsilon}{2},\delta)$ מתקיים: (כי המרחק (בנורמה 2) הוא $\frac{\epsilon}{2}$ ולכן ההפרש הוא לכל היותר ϵ)

$$D^m(\{S\in(X\times Y)^m \mid \forall h\in H\ L_S(h)<L_D(h)+\epsilon\})\ge 1-\delta$$

ולכן H למידה-PAC עם $m_H(\epsilon,\delta)\le m_H^{UC}(\frac{\epsilon}{2},\delta)$.

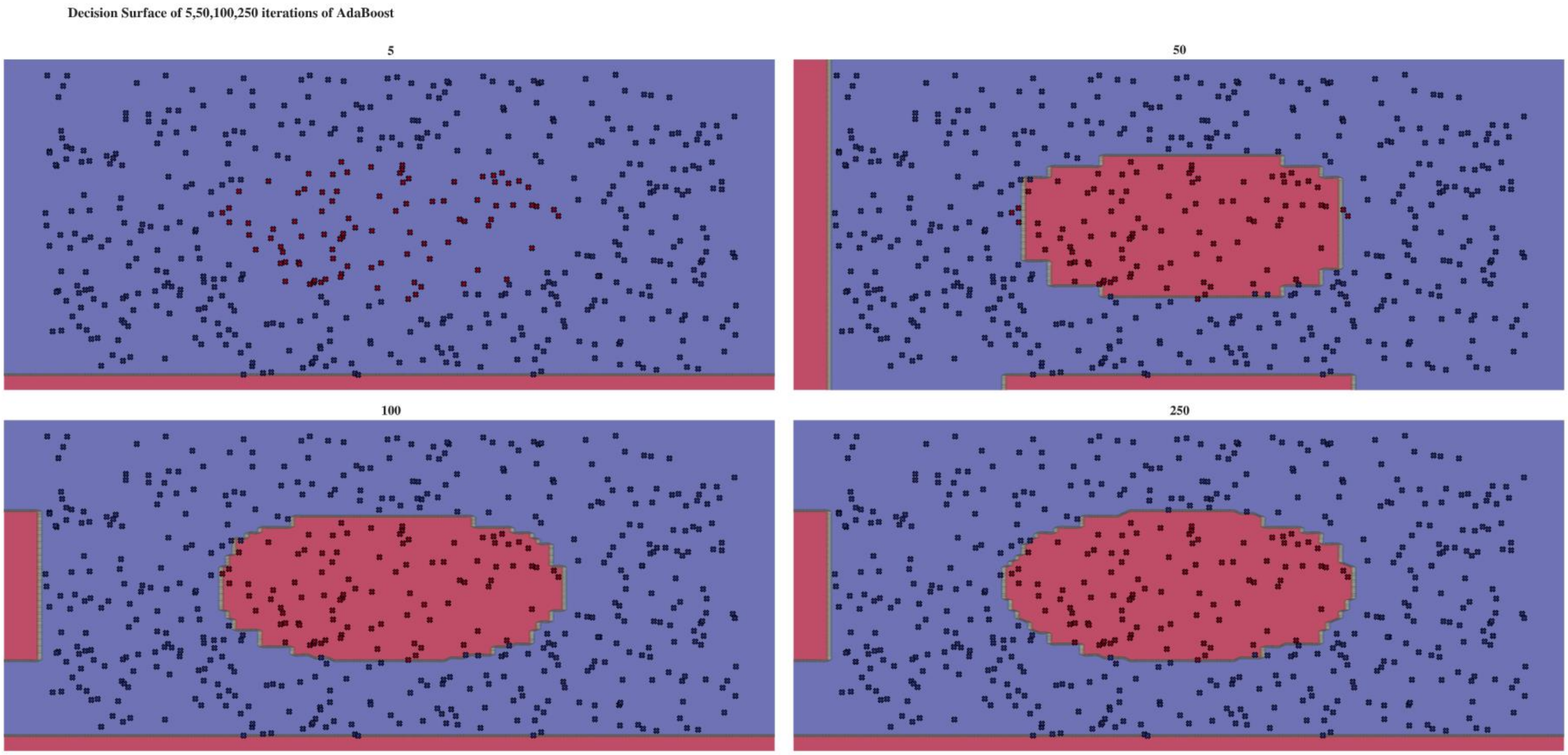
מערכות לומדות תרגיל 4 חלק מעשי:

שאלה 1



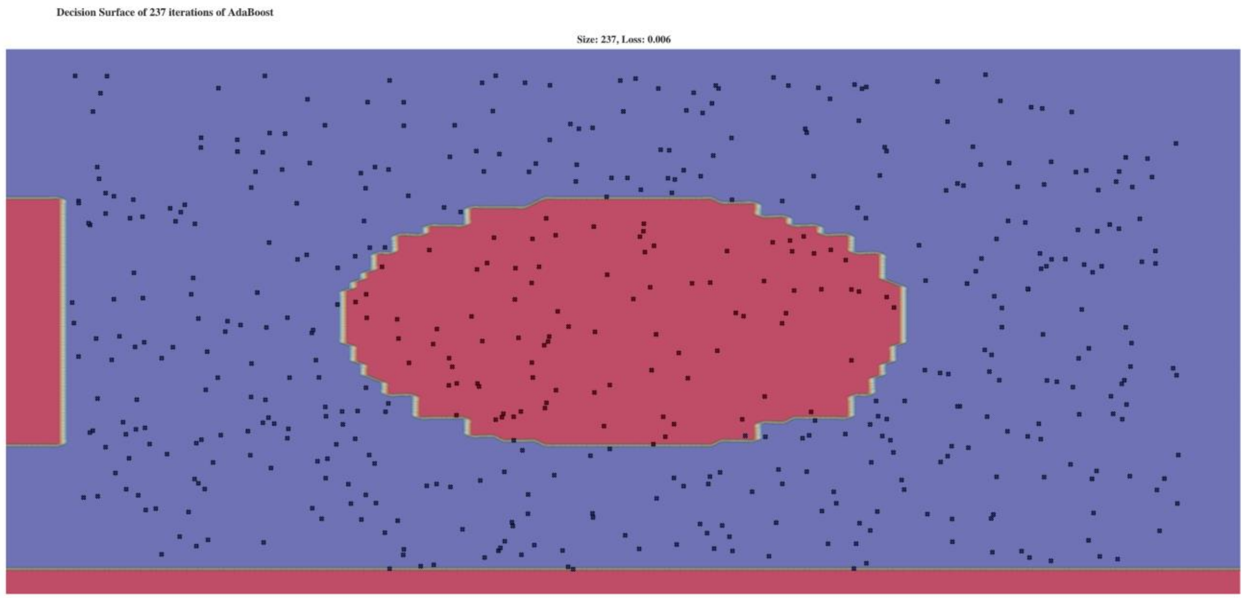
קיבלנו שעבור רעש מועט הtest והtrain מתנהגים דומה והשגיאה שלהם נהיית אפסית. אסיק שהמודל חזק ומשתפר עם עליית כמות האיטרציות. בנוסף המודל לא עושה overfit כי הtrain והtest משתפרים ביחד.

שאלה 2



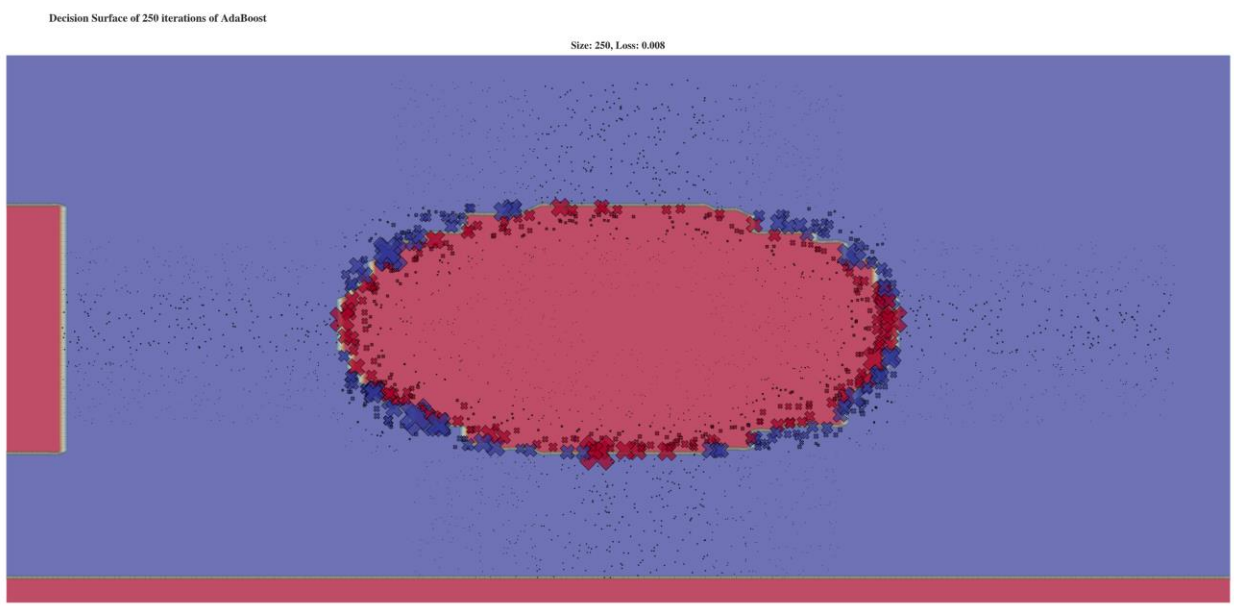
מגרפים אלו ניתן לראות שהביצועים הולכים ומשתפרים עם עליית מספר האיטרציות. המודל מצליח לחזות את הפונקצייה המקורית של המידע.

שאלה 3



בגרף זה ניתן לראות שעבור 237 איטרציות השגיאה היא 0.006! מאוד קטנה.

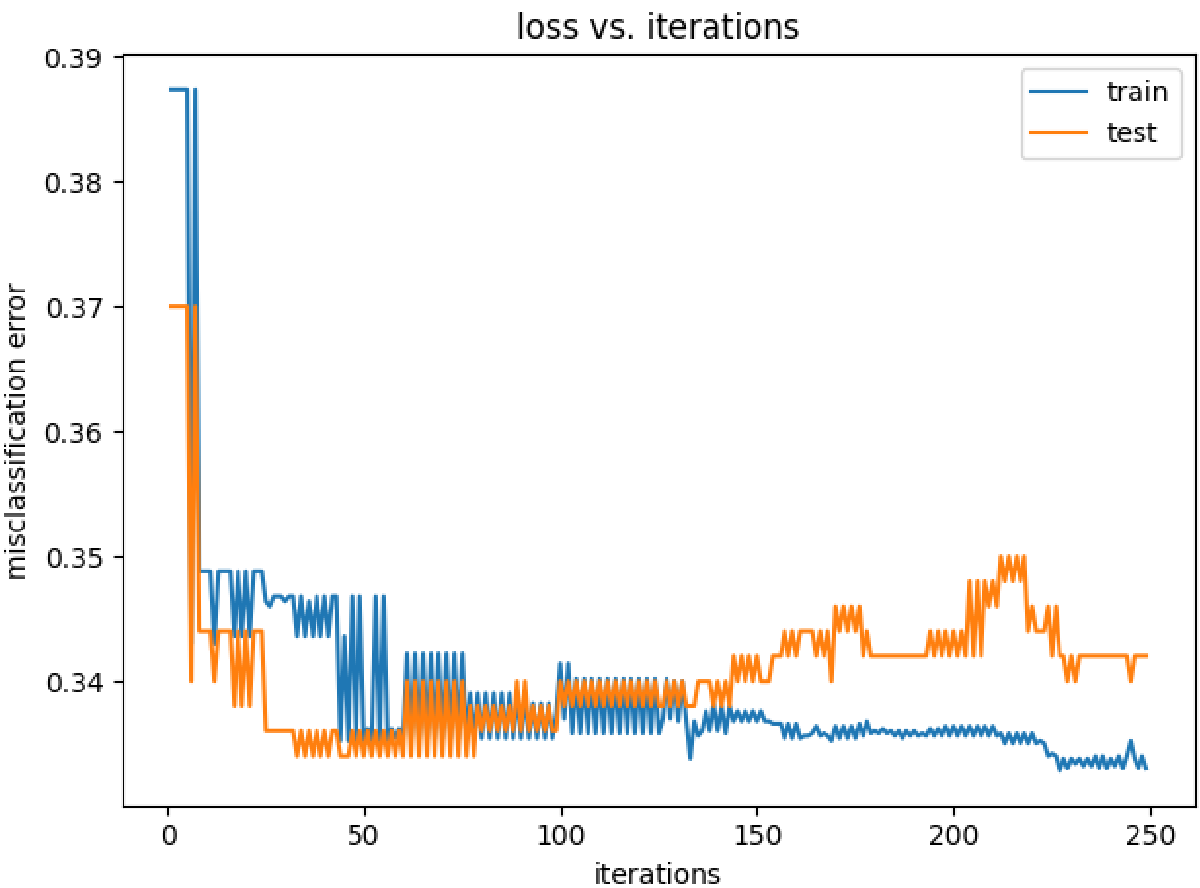
שאלה 4



ניתן לראות לפי גודל הנקודות שהנקודות הבעייתיות הן אלו שקרובות לקצה התחום האדום והכחול, ז"א נקודות שבאמת קשה להגיד לאיזה תחום הן משתייכות.

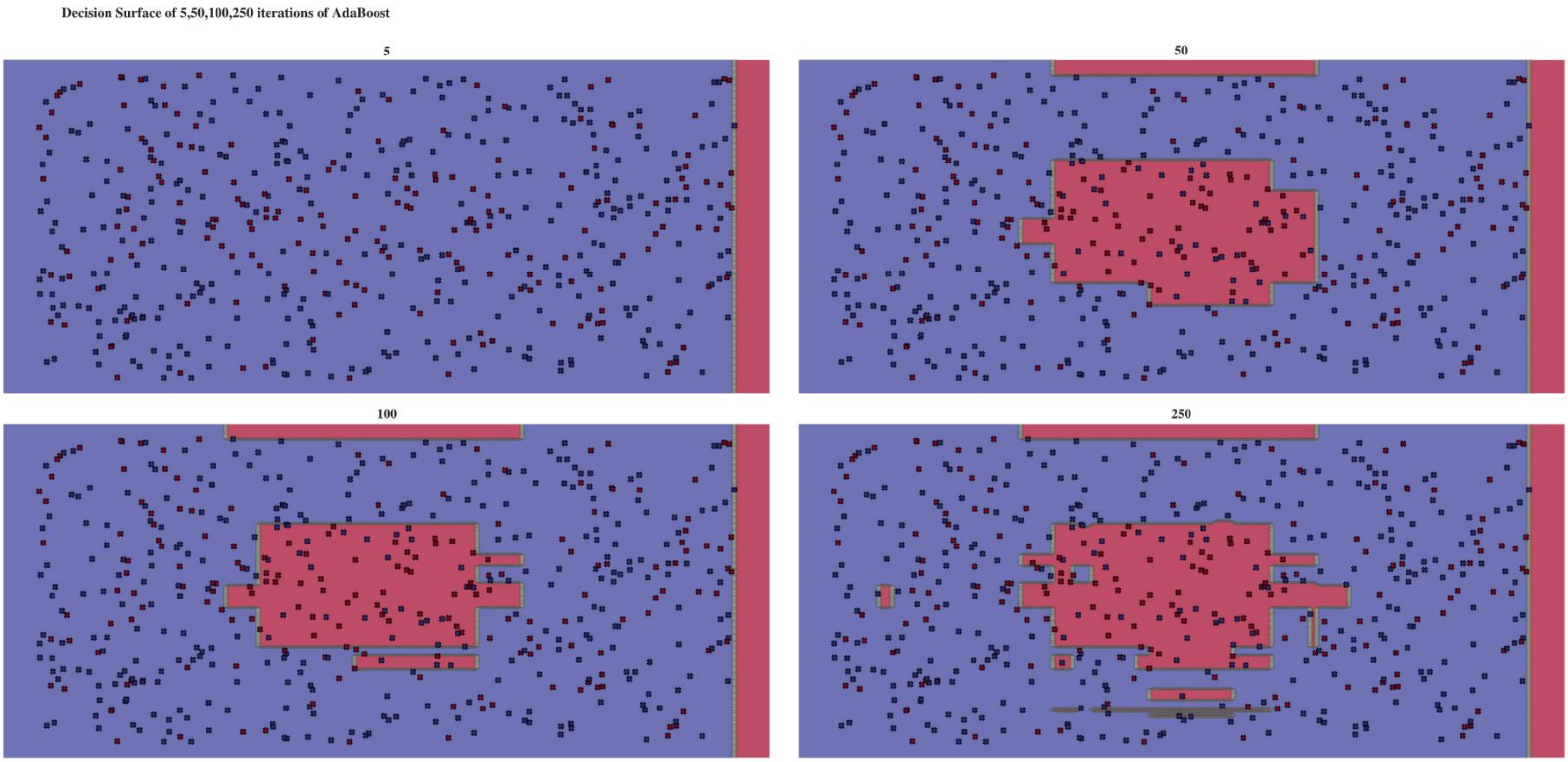
שאלה 5

1



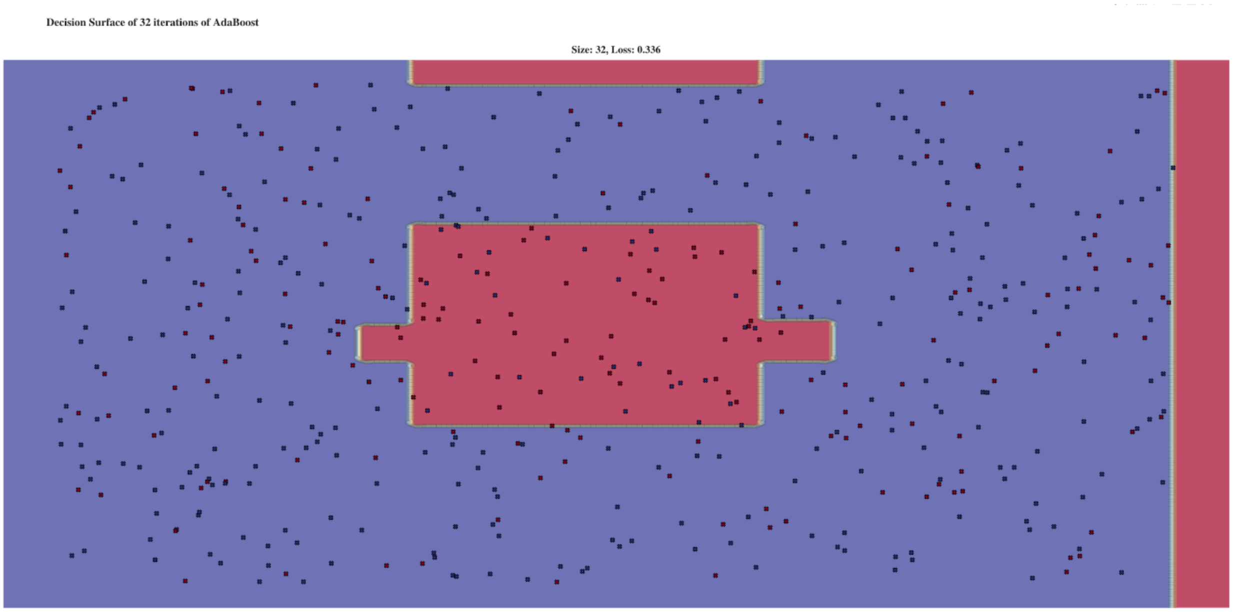
במקרה זה ניתן לראות שהרעש גורם לכך שהעלאה נוספת של מספר האיטרציות גורמת לoverfit במסגרתו הtrain משתפר והtest נעשה גרוע יותר. הביאס יורד אבל הואריאנס עולה עבור ערכים גדולים מ~37.

.2



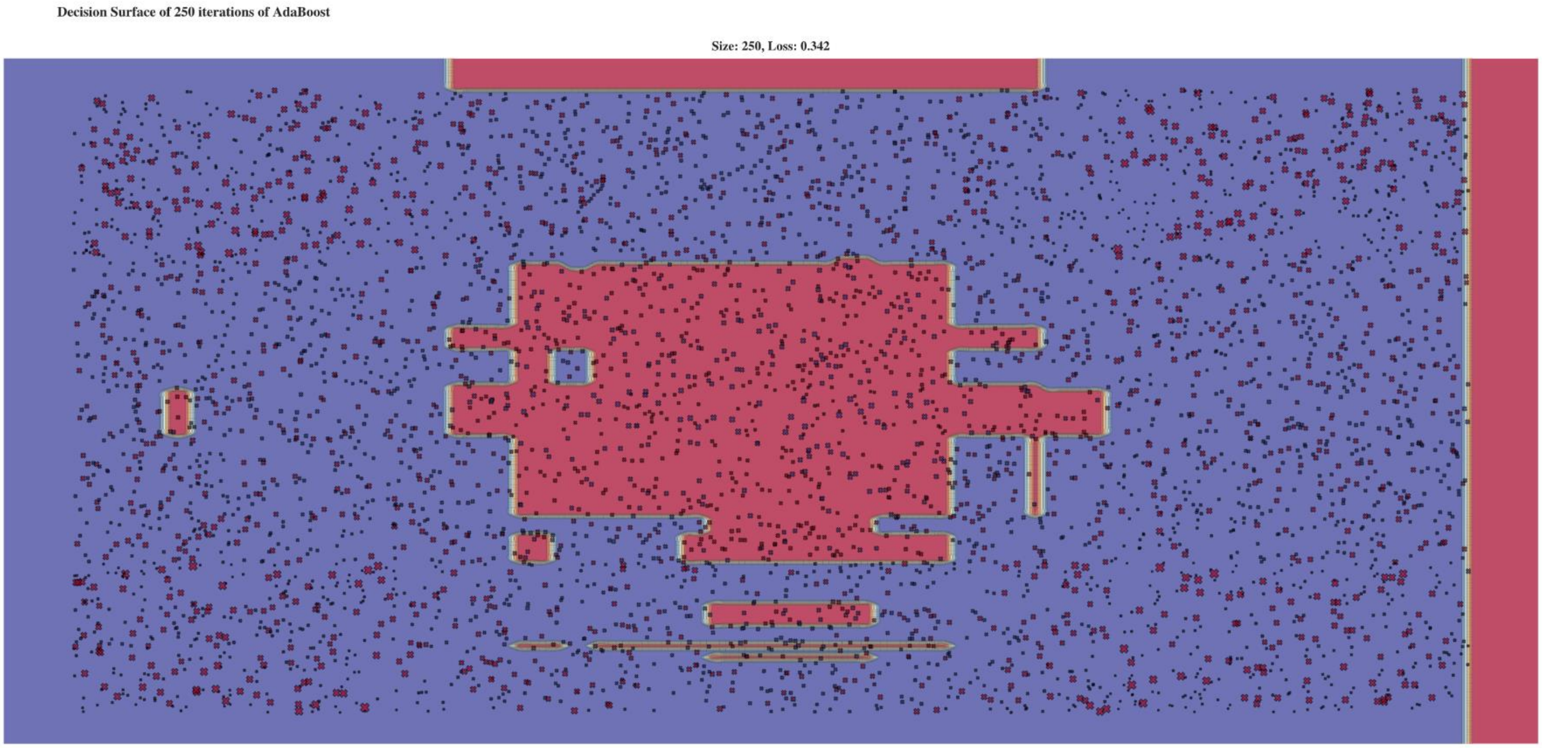
גם במקרה זה הגרפים מתארים יחסית טוב את ההתפלגות המקורית על אף שיש הרבה רעש. עבור מספר איטרציות גבוה מדי (250) יש מקטעים שמרמזים שהתבצע overfit.

.3



הפעם הפיט המיטבי הוא עבור 32 איטרציות והשגיאה היא 0.336. מכך שהאיטרציות נמוכות הפעם אסיק שמודל זה התאים יותר טוב מאחר שהואריאנס שלו נמוך יותר והוא פחות עושה overfit על המידע.

.4



בגרף זה ניתן לראות שהנקודות הבעייתיות הן אלו שנמצאות מחוץ לתחום ההגיוני עבורן, ז"א נקודות שנובעות מרעש.