מערכות לומדות תרגיל 4

שיר שבח 322407701

12 במאי 2022

חלק תיאורטי

2.1 PAC Learnability

1. עבור אלגוריתם למידה $\mathcal{A}$ כלשהו, התפלגות $\mathcal{D}$ מעל $\mathcal{X}$ ופונקציית $0-1\ loss$ (misclassification), הוכיחו שהבאים שווים:

$$(a)\ \forall\varepsilon,\delta>0\ \ \exists m(\varepsilon,\delta)\ \ \textbf{s.t}\ \ \forall m\geq m(\varepsilon,\delta)\ \ \mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right)\leq\varepsilon\right]\geq1-\delta$$

$$(b)\ \lim_{m\to\infty}\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(A\left(S\right)\right)\right]=0$$

$\underline{(a)\Rightarrow(b):}$

יהי $\varepsilon>0$. נרצה להוכיח כי $\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(A\left(S\right)\right)\right]<\varepsilon$.
עפ"י הגדרת תוחלת עבור מ"מ רציפים:

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(A\left(S\right)\right)\right]=\int_0^1 t\cdot\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}\left(A\left(S\right)\right)=t\right)dt$$

נבחר כי מתקיים משפט (a) עבור $\frac{\varepsilon}{2}$ ו־$\delta=\frac{\varepsilon}{2}$. לכן עבור $m_0>m\left(\frac{\varepsilon}{2},\frac{\varepsilon}{2}\right)$ מתקיים:

$$
\begin{aligned}
\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(A\left(S\right)\right)\right] &= \int\limits_0^1 t\cdot\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}\left(A\left(S\right)\right)=t\right)dt\\
&= \int\limits_0^{\frac{\varepsilon}{2}} t\cdot\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}\left(A\left(S\right)\right)=t\right)dt+\int\limits_{\frac{\varepsilon}{2}}^1 t\cdot\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}\left(A\left(S\right)\right)=t\right)dt\\
&\le \int\limits_0^{\frac{\varepsilon}{2}} \frac{\varepsilon}{2}\cdot\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}\left(A\left(S\right)\right)=t\right)dt+\int\limits_{\frac{\varepsilon}{2}}^1 1\cdot\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}\left(A\left(S\right)\right)=t\right)dt\\
&= \frac{\varepsilon}{2}\cdot\int\limits_0^{\frac{\varepsilon}{2}} \mathbb{P}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}\left(A\left(S\right)\right)=t\right)dt+\int\limits_{\frac{\varepsilon}{2}}^1 \mathbb{P}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}\left(A\left(S\right)\right)=t\right)dt\\
&= \frac{\varepsilon}{2}\cdot\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}\left(A\left(S\right)\right)\le\frac{\varepsilon}{2}\right)+\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}\left(A\left(S\right)\right)>\frac{\varepsilon}{2}\right)\\
&= \frac{\varepsilon}{2}\cdot\left(1-\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}\left(A\left(S\right)\right)>\frac{\varepsilon}{2}\right)\right)+\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}\left(A\left(S\right)\right)>\frac{\varepsilon}{2}\right)\\
&= \frac{\varepsilon}{2}\cdot 1-\frac{\varepsilon}{2}\cdot\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}\left(A\left(S\right)\right)>\frac{\varepsilon}{2}\right)+\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}\left(A\left(S\right)\right)>\frac{\varepsilon}{2}\right)\\
&= \frac{\varepsilon}{2}\cdot 1+\left(1-\frac{\varepsilon}{2}\right)\cdot\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}\left(A\left(S\right)\right)>\frac{\varepsilon}{2}\right)\\
&< \frac{\varepsilon}{2}\cdot 1+\left(1-\frac{\varepsilon}{2}\right)\cdot\frac{\varepsilon}{2}\\
&= \varepsilon-\frac{\varepsilon^2}{4}<\varepsilon
\end{aligned}
$$

$\underline{:(b)\Rightarrow(a)}$

יהי $\varepsilon,\delta>0$.

עפ"י אי שיוויון מרקוב:

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

אם נסמן $a=\varepsilon$ ו־ $\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(A\left(S\right)\right)\right]\le\varepsilon\cdot\delta$, אזי יתקיים:

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

וקיבלנו

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

שזה שקול ל־

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right)\leq\varepsilon\right]\geq1-\delta$$

2. יהי $\mathcal{X}:=\mathbb{R}^2,\mathcal{Y}:=\{0,1\}$ ויהי $\mathcal{H}$ מחלקה של מעגלים קונצנטריים במישור:

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

הוכיחו כי $\mathcal{H}$ היא למידה־PAC וה־sample complexity חסום ע"י

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

בעת הוכחה, אל תשתמשו בטיעון VC-Dimension. במקום זאת הוכיחו את הטענה ישירות מהגדרת למידות PAC על ידי הצגת אלגוריתם ספציפי וניתוח מורכבות המדגם שלו.

האלגוריתם A שניקח הוא אלגוריתם שחוזה את r באופן הבא:

$$\hat{r}=\frac{1}{2}\cdot\max_{x_i,x_j\in\mathcal{X}}\left\{d\left(x_i,x_j\right)\right\}\ \ \text{s.t}\ \ y_i=y_j=1$$

נסמן את איזור החיזוי שלנו (מעגל $h_{\hat{r}}$) ב־$\hat{I}$, ואת האיזור האמיתי, (h_r) ב־I.
איזור הטעות הוא $I\backslash\hat{I}$. לכן

$$L_D\left(h_s\right)=\mathbb{P}\left(x\in I\backslash\hat{I}\right)$$

נרצה לחשב את $\mathbb{P}\left\{L_D\left(h_s\right)\leq\varepsilon\right\}$. ראשית נחשב את $\mathbb{P}\left\{L_D\left(h_s\right)>\varepsilon\right\}$.

$$\mathbb{P}\left\{L_D\left(h_s\right)>\varepsilon\right\}=\mathbb{P}\left(\bigwedge_{i=1}^{m}x_i\notin I\backslash\hat{I}\right)\overset{iid}{=}\prod_{i=1}^{m}\mathbb{P}\left(x_i\notin I\backslash\hat{I}\right)\leq\left(1-\varepsilon\right)^m\leq e^{-\varepsilon m}$$

לכן:

$$\mathbb{P}\left\{L_D\left(h_s\right)\leq\varepsilon\right\}=1-e^{-\varepsilon m}$$

אנו נרצה ש־

$$1-e^{-\varepsilon m} \geq 1-\delta$$
$$e^{\varepsilon m} \geq \frac{1}{\delta}$$
$$\varepsilon m \geq \log\left(\frac{1}{\delta}\right)$$
$$m \geq \frac{\log\left(\frac{1}{\delta}\right)}{\varepsilon}$$

VC-Dimension 2.2

3. יהי $\mathcal{X}=\{0,1\}^n$ ו־ $\mathcal{Y}=\{0,1\}$, לכל $I\subseteq[n]$ הגדירו את הפונקציה הזוגית:

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

מה מימד VC של המחלקה $\mathcal{H}_{parity}=\{h_I \mid I\subseteq[n]\}$? הוכיחו את תשובתכם.

נשים לב כי לכל מימד שקטן מ־n, ניתן לנתץ את h. מספיק להראות עבור דוגמה אחת, ניקח $\{0\}^n,...,\{0\}^1$, אזי לכל $i, I\subseteq[n]$:

$$h_I\left(\{0\}^i\right)=0$$

בנוסף, משום שגודל ההיפותזה הוא $|\mathcal{H}|=2^n$, זאת קבוצה סופית. ולכן $\text{VCdim}(\mathcal{H})\leq\log_2(|\mathcal{H}|)$ ולכן $\text{VCdim}(\mathcal{H})\leq$ $\log_2(2^n)=n$ מה שגורר כי $\text{VCdim}(\mathcal{H})=n$.

4. בהינתן מספר שלם k, יהי $([a_i,b_i])_{i=1}^k$ להיות קבוצה כלשהי של קטעים על $\mathbb{R}$ ומוגדר האיחוד $A=\bigcup_{i=1}^k[a_i,b_i]$. מחלקת ההיפותזות $\mathcal{H}_{k-intervals}$ כוללת את הפונקציות: $h_A(x):=\mathbb{1}_{[x\in A]}$, לכל בחירות של k מקטעים. מצאו את מימד VC של $\mathcal{H}_{k-intervals}$ והוכיחו את תשובתכם. הראו כי אם אנו נותנים ל־A להיות כל איחוד סופי של מקטעים (כלומר k לא מוגבל), אז המחלקה המתקבלת $\mathcal{H}_{intervals}$ יש מימד VC ∞.

נרצה להוכיח כי מימד VC עבור מחלקת היפותזות $\mathcal{H}_{k-intervals}$ הוא $2k$.

לכן נרצה להוכיח כי לכל $d<2k$, קיים צמצום C, $|C|=d$, כך ש $\mathcal{H}_{k-intervals}$ מנתצת את C.

עבור דגימה $\{x_1,x_2,..,x_{2k}\}$, נעטוף במקטע אחד אינדקסים שמתויגים ברצף כ־1. לדוגמה, עבור $k=3$, $|X|=6$, $X=$ $\{0,1,1,0,0,1\}$ אז $A=[1,2]\cup[5]$. נשים לב כי גודל הקבוצות באיחוד הוא לכל היותר k.

משום שבמחלקת ההיפותזות שלנו $\mathcal{H}_{k-intervals}$ זהו איחוד של k מקטעים, קיבלנו כי $C=h_A(x)\in\mathcal{H}_{k-intervals}$, כלומר $\mathcal{H}_{k-intervals}$ מנתצת את C.

נרצה להוכיח כי לכל $|C|=2k+1$, $\mathcal{H}_{k-intervals}$ לא מנתצת את C.

עבור דגימה $\{x_1,...,x_{2k},x_{2k+1}\}$, אם מספר הנקודות המתוייגות כ־1 גדול מ־k, אזי משובך היונים קיים מקטע שבו יש 2 נקודות. אם ביניהם נשים נקודה המתוייגת ל־0, הם לא יוכלו להיות באותו מקטע, ובכך אין פונקציה ב־ $\mathcal{H}_{k-intervals}$ שיכולה לתייג נכון דגימה זו. ולכן $\mathcal{H}_{k-intervals}$ לא מנתצת את C.

עבור המחלקת $\mathcal{H}_{intervals}$, נשים לב כי לכל k, קבוצה בגודל $2k$ מנותצת ע"י פונקציה ב־$\mathcal{H}_{k-intervals}$. קל וחומר עבור מחלקת היפוטזות $\mathcal{H}_{intervals}$. ולכן $\mathcal{H}_{intervals}$ מנתצת קבוצה מכל גודל סופי. לכן מימד VC שלה הוא ∞.

2.3 Monotonicity

5. יהי $\mathcal{H}$ מחלקת היפותזות עבור סיווג בינארי. נניח כי $\mathcal{H}$ למידה PAC וסיבוכיות המודל ניתן ע"י $m_{\mathcal{H}}(\cdot,\cdot)$. הראו כי $m_{\mathcal{H}}$ מונוטוני יורד בכל אחד מהפרמטרים שלו. זאת אומרת:

- הראו כי בהינתן $\delta \in (0,1)$, ובהינתן $0 < \varepsilon_1 \le \varepsilon_2 < 1$, חייב כי $m_{\mathcal{H}}(\varepsilon_1,\delta) \ge m_{\mathcal{H}}(\varepsilon_2,\delta)$.

- בדומה, הראו כי בהינתן $\varepsilon \in (0,1)$ ובהינתן $0 < \delta_1 \le \delta_2 < 1$, חייב כי $m_{\mathcal{H}}(\varepsilon,\delta_1) \ge m_{\mathcal{H}}(\varepsilon,\delta_2)$.

עבור $\mathcal{H}$ למידה PAC אנו יודעים כי מתקיים:

$$m_{\mathcal{H}}(\varepsilon,\delta) \sim \frac{VCdim(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)}{\varepsilon}$$

בהינתן $\delta \in (0,1)$, ו $0 < \varepsilon_1 \le \varepsilon_2 < 1$, מתקיים:

$$m_{\mathcal{H}}(\varepsilon_1,\delta) \sim \frac{VCdim(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)}{\varepsilon_1} \ge \frac{VCdim(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)}{\varepsilon_2} \sim m_{\mathcal{H}}(\varepsilon_2,\delta)$$

ובהינתן $\varepsilon \in (0,1)$ ו $0 < \delta_1 \le \delta_2 < 1$

$$m_{\mathcal{H}}(\varepsilon,\delta_1) \sim \frac{VCdim(\mathcal{H}) + \log\left(\frac{1}{\delta_1}\right)}{\varepsilon} \ge \frac{VCdim(\mathcal{H}) + \log\left(\frac{1}{\delta_2}\right)}{\varepsilon} \sim m_{\mathcal{H}}(\varepsilon,\delta_2)$$

כנדרש.

6. יהי $\mathcal{H}_1$ ו־$\mathcal{H}_2$ שתי מחלקות המסווגות בינארי, כך ש־$\mathcal{H}_1 \subseteq \mathcal{H}_2$. הראו כי $\mathbf{VCdim}(\mathcal{H}_1) \le \mathbf{VCdim}(\mathcal{H}_2)$.

נניח בשלילה כי $\mathcal{H}_1 \subseteq \mathcal{H}_2$ אולם $\mathrm{VCdim}(\mathcal{H}_1) > \mathrm{VCdim}(\mathcal{H}_2)$. אזי קיים צמצום C כך ש $\mathcal{H}_1$ מנתצת את C אולם $\mathcal{H}_2$ לא. סתירה לכך שלכל $h \in \mathcal{H}_1$ חייב כי $h \in \mathcal{H}_2$.

2.4 Agnostic-PAC

7. הוכיחו כי אם $\mathcal{H}$ יש תכונת התכנסות אחידה עם פונקציה $m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$, אז $\mathcal{H}$ היא למידה Agnostic-PAC עם סיבוכיות מודל $m_{\mathcal{H}}(\varepsilon,\delta) \le m^{UC}\left(\frac{\varepsilon}{2},\delta\right)$.

משום של־$\mathcal{H}$ יש תכונת UC עם פונקציית $m_{\mathcal{H}}^{UC}$, יהי S סט דגימות $\frac{\varepsilon}{2}$-מייצג (מגודל $m > m^{UC}\left(\frac{\varepsilon}{2},\delta\right)$) עבור ההיפותזה $\mathcal{H}$, התפלגות $\mathcal{D}$, ופונקציית loss ℓ. נסמן $h_S = \mathrm{ERM}_{\mathcal{H}}(S)$.
לכן, בהסתברות של לפחות $1-\delta$ מתקיים:

$$L_D(h_S) \le L_{\mathcal{D}}(h) + \frac{\varepsilon}{2}$$

לכן, לכל $h \in \mathcal{H}$:

$$L_D\left(h_S\right) \overset{\clubsuit}{\leq} L_S\left(h_S\right)+\frac{\varepsilon}{2} \overset{\spadesuit}{\leq} \min_{h\in\mathcal{H}} L_S\left(h\right)+\frac{\varepsilon}{2} \overset{\clubsuit}{\leq} \min_{h\in\mathcal{H}} L_D\left(h\right)+\frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\min_{h\in\mathcal{H}} L_D\left(h\right)+\varepsilon$$

$\clubsuit$ - S הוא ε־מייצג לכן לפי ההגדרה:

$$\left|L_S\left(h\right)-L_{\mathcal{D}}\left(h\right)\right| \leq \frac{\varepsilon}{2} \Leftrightarrow -\frac{\varepsilon}{2} \leq L_S\left(h\right)-L_{\mathcal{D}}\left(h\right) \leq \frac{\varepsilon}{2}$$

$\spadesuit$ - $h_S=\mathrm{ERM}_{\mathcal{H}}\left(S\right)$.

לכן קיבלנו כי $\mathcal{H}$ היא Agnostic-PAC עם $m_{\mathcal{H}}\left(\varepsilon,\delta\right) \leq m^{UC}\left(\frac{\varepsilon}{2},\delta\right)$.

8. יהי $\mathcal{H}$ מחלקת היפוטזות מעל תחום $\mathcal{Z}=\mathcal{X}\times\{\pm1\}$, ועבור פונקציית loss $0-1$. הניחו כי קיימת פונקציה $m_{\mathcal{H}}$, שעבורה היא מחזיקה שלכל התפלגות $\mathcal{D}$ מעל $\mathcal{Z}$ קיים אלגוריתם $\mathcal{A}$ עם התכונות הבאות: כאשר מריצים את $\mathcal{A}$ על $m>m_{\mathcal{H}}$ דוגמאות iid המתפלגות מ־$\mathcal{D}$, מובטח שיחזור, עם הסתברות של לפחות $1-\delta$, היפוטזה $h_S:\mathcal{X}\to\{\pm1\}$ עם $L_{\mathcal{D}}\left(h_S\right)\leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}\left(h\right)+\varepsilon$. האם $\mathcal{H}$ למידת Agnostic-PAC? הוכיחו או הראו דוגמה מספרית.

נוכיח כי H אינה למידה agnostic PAC.

ההגדרה של למידות agnostic PAC:

יהיו $\varepsilon,\delta\in(0,1)$, נאמר כי אלגוריתם למידה A הוא למיד agnostic PAC עם בטיחות δ ודיוק ε ביחס לפונקציית $\ell\ loss$, מחלקת היפוטזות $\mathcal{H}$ והתפלגות D:

$$D^m\left(\left\{S \mid L_D\left(h_S\right) \leq \min_{h'\in\mathcal{H}} L_D\left(h'\right)+\varepsilon\right\}\right) \geq 1-\delta$$

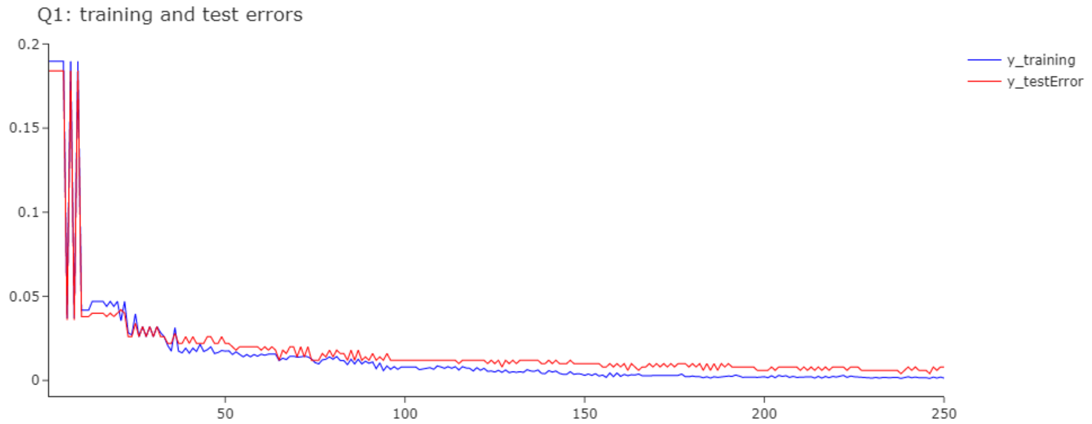
נשים לב כי לפי הגדרה אלגוריתם A אינו תלוי ב־D. לעומת זאת, A הנתון בשאלה תלוי ב־D.

דוגמה מספרית: ניקח H מחלקת היפוטזות שאינה למידה PAC, לדוגמה H היא כל הפונקציות הבינאריות מעל הממשיים. אם H למידה agnostic PAC אזי קיים אלגוריתם A עם סיבוכיות מקום $m_H\left(\varepsilon,\delta\right)$, כך שלכל התפלגות D $\mathbb{P}\left(A\left(S\right) \leq \min_{h\in H} L_D\left(h\right)+\varepsilon\right) \geq 1-\delta$.

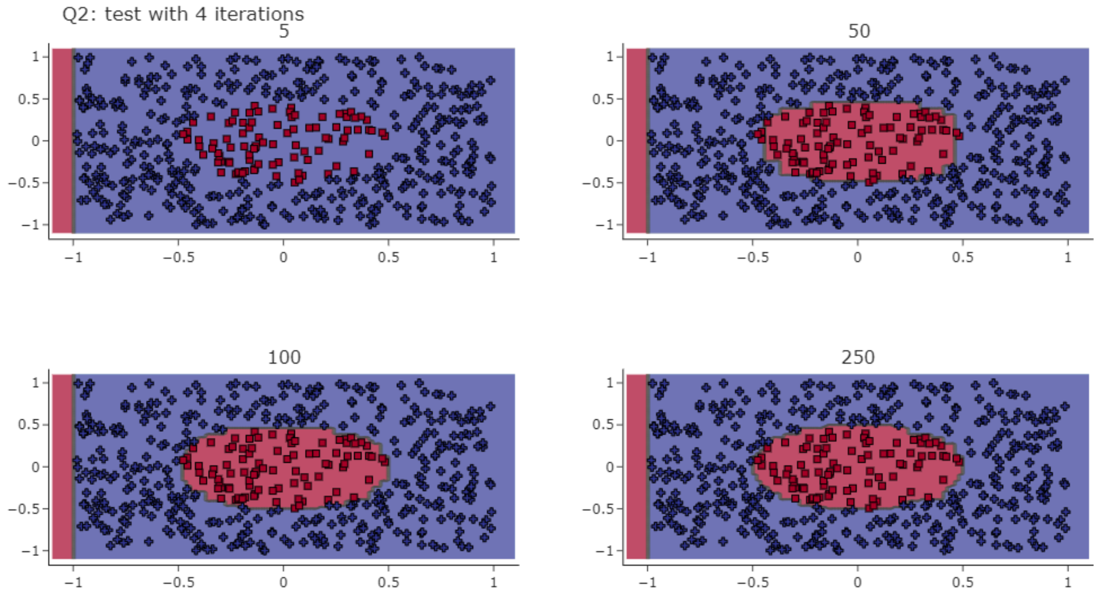
למדנו בתרגול כי H אינה למידה PAC אזי היא גם לא למידה agnostic PAC, סתירה.

מערכות לומדות חלק פרקטי

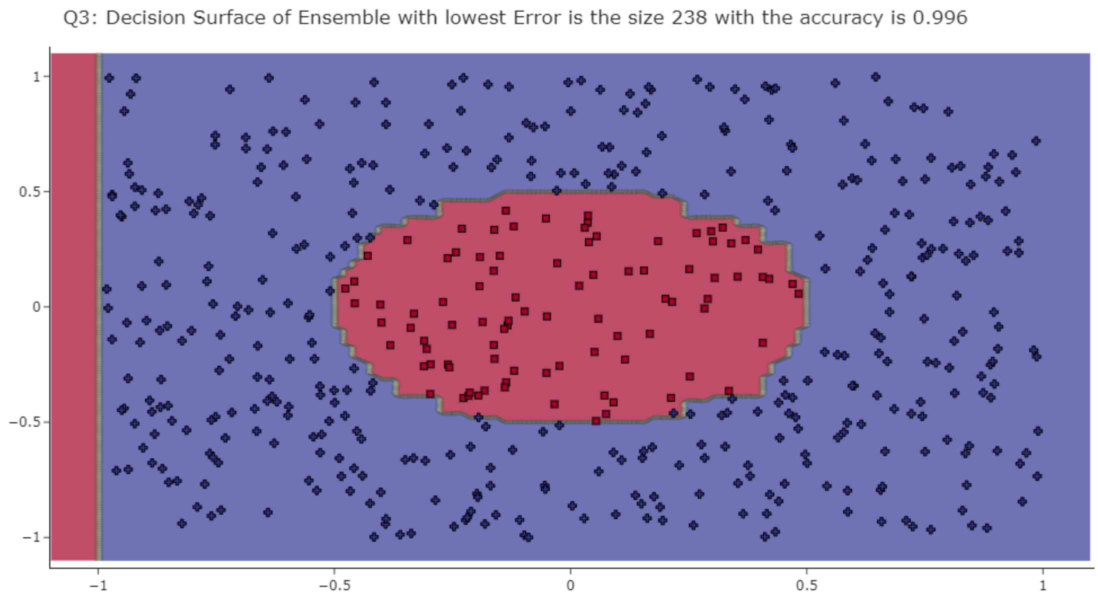
1. ניתן לראות לפי הגרף כי הטעות של הtrain אכן דומה לtest בכל איטרציה ואכן האלגוריתם עובד נכון.



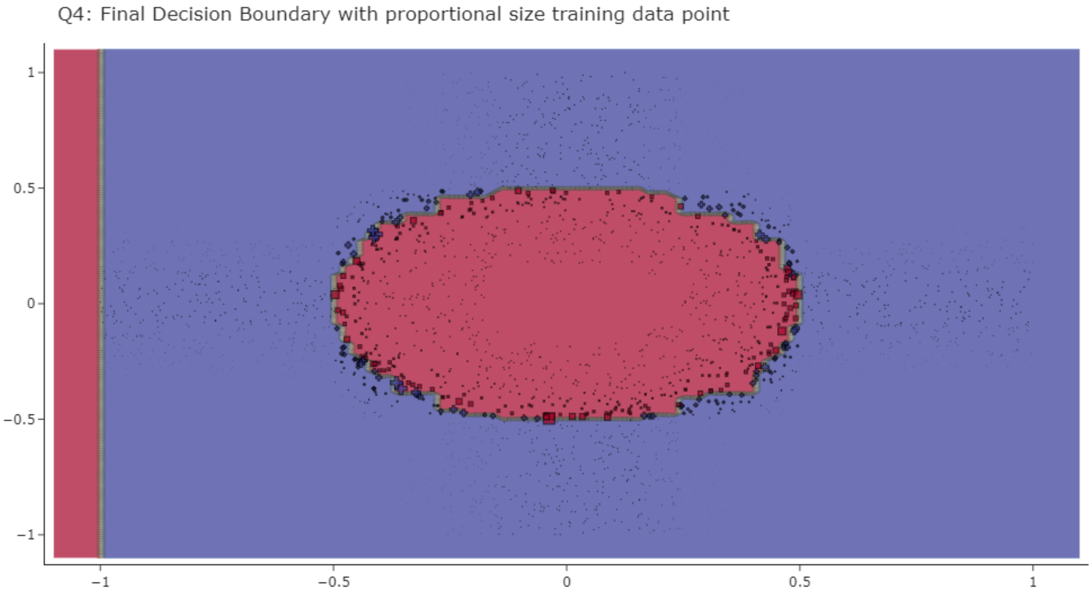
2. ניתן לראות לפי הגרפים שיצאו כי אכן ככל שמתקדמים באיטרציות החיזוי טוב יותר, אולם בין 50-100-250 השינויים מינוריים.



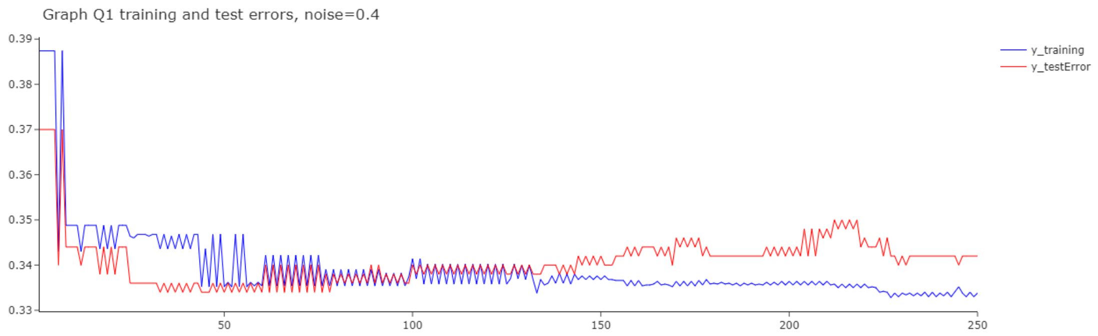
3.



4. ניתן לראות לפי הגרף שאכן הנקודות שיותר קשה לחזות אותן, משום שהן קרובות לthreshold, אכן המשקל שלהן גבוה יותר כי במהלך האלגוריתם הן נהפכו להיות קריטיות יותר. ואכן הנקודות הרחוקות מכלל ההחלטה קטנות יותר, ז"א המשקל שלהן קטן יותר, כי הן קלות יותר לחיזוי לפי האלגוריתם.



5. ניתן לראות בגרף הראשון כי בהתחלה לtrain יש טעות גדולה יותר מהtest, כי אכן הוא מתחשב בנקודות שהן רעש, אולם בהמשך, כשיש הרבה איטרציות לאגלוריתם adaboost, הוא כבר מתחיל לא להתייחס לנקודות רעש ויש טעות קטנה הרבה יותר.



כאן ניתן לראות כי משום שיש רעש, אכן החיזוי קשה יותר.