<u>עבודת הגשה מספר 4</u>

<u>קורס</u>: מבוא למערכות לומדות (67577)

<u>מגיש</u>: דור מסיקה, <u>ת.ז</u>: 318391877

11 במאי 2022

החלק התיאורטי

PAC Learnabillity

.1

עבור אלגוריתם למידה A, התפלגות D מעל $\mathcal{X}$ ופונקציית $loss_{0-1}$ $(misclassification)$, נוכיח שהבאים הינם שקולים:
א)

$$\forall\varepsilon,\delta>0,\quad \exists m\left(\varepsilon,\delta\right)\quad s.t\ \ \forall m\geq m\left(\varepsilon,\delta\right)\quad P_{S\sim D^{m}}\left[L_{D}\left(A\left(S\right)\right)\leq\varepsilon\right]\geq1-\delta$$

ב)

$$\underset{m\rightarrow\infty}{lim}\mathbb{E}_{S\sim D^{m}}\left[L_{D}\left(A\left(S\right)\right)\right]=0$$

<u>**פתרון:**</u> נוכיח את הטענה באמצעות הכלה דו כיוונית.
$(a\Longleftarrow b)$ מתקיים כי $\underset{m\rightarrow\infty}{lim}\mathbb{E}_{S\sim D^{m}}\left[L_{D}\left(A\left(S\right)\right)\right]=0$. נשתמש באי שיוויון מרקוב, ונקבל כי לכל $\varepsilon>0$

$$P_{S\sim D^{m}}\left(L_{D}\left(A\left(S\right)\right)\leq\varepsilon\right)=1-P_{S\sim D^{m}}\left(L_{D}\left(A\left(S\right)\right)\geq\varepsilon\right)\geq1-\frac{\mathbb{E}_{S\sim D^{m}}\left[L_{D}\left(A\left(S\right)\right)\right]}{\varepsilon}$$

ומכיוון שמתקיים כי $\underset{m\rightarrow\infty}{lim}\frac{\mathbb{E}_{S\sim D^{m}}\left[L_{D}\left(A\left(S\right)\right)\right]}{\varepsilon}=0$, מהנתון ומאריתמטיקה של גבולות, נקבל כי

$$\underset{m\rightarrow\infty}{lim}P_{S\sim D^{m}}\left(L_{D}\left(A\left(S\right)\right)\leq\varepsilon\right)\geq1$$

כלומר, קיים m' גדול מספיק כך ש־$P_{S\sim D^{m}}\left(L_{D}\left(A\left(S\right)\right)\leq\varepsilon\right)\geq1$ לכל $m\geq m'$, ולכן לכל $\delta>0$ ו־$\varepsilon>0$ נוכל למצוא m' גדול מספיק התלוי בהם, נסמנו $m\left(\varepsilon,\delta\right)$ כך שלכל m הגדול ממנו יתקיים כי

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

כפי שרצינו להוכיח.
$(b\Longleftarrow a)$ מתקיים כי $\forall\varepsilon,\delta>0,\quad \exists m\left(\varepsilon,\delta\right)\quad s.t\ \ \forall m\geq m\left(\varepsilon,\delta\right)\quad P_{S\sim D^{m}}\left[L_{D}\left(A\left(S\right)\right)\leq\varepsilon\right]\geq1-\delta$, ונראה שלכל $\varepsilon,\delta>0$ ולכל $m\geq m\left(\varepsilon,\delta\right)$ מתקיים כי $\mathbb{E}_{S\sim D^{m}}\left[L_{D}\left(A\left(S\right)\right)\right]\leq\varepsilon+\delta$.
יהיו $\varepsilon,\delta>0$ ו־$m\left(\varepsilon,\delta\right)$ משתנה התלוי בהם. נשים לב שמנוסחאת התוחלת נקבל

$$\mathbb{E}_{S\sim D^{m}}\left[L_{D}\left(A\left(S\right)\right)\right]\leq P_{S\sim D^{m}}\left[L_{D}\left(A\left(S\right)\right)\leq\varepsilon\right]\cdot\varepsilon+P_{S\sim D^{m}}\left[L_{D}\left(A\left(S\right)\right)>\varepsilon\right]\cdot1\leq$$

מההנחה מתקיים כי

$$P_{S\sim D^{m}}\left[L_{D}\left(A\left(S\right)\right)>\varepsilon\right]<\delta$$

ולכן בסך הכל נקבל עבור ביטוי זה

$$\leq \varepsilon \cdot 1 + P_{S\sim D^m}\left[L_D\left(A\left(S\right)\right) > \varepsilon\right] \leq \varepsilon + \delta$$

ולכן כפי שנראה בסעיפים הבאים נובע כי

$$\underset{m\to\infty}{lim}\mathbb{E}_{S\sim D^m}\left[L_D\left(A\left(S\right)\right)\right] \leq 0 + 0$$

אבל מאי שליליות התוחלת נקבל סך הכל

$$\underset{m\to\infty}{lim}\mathbb{E}_{S\sim D^m}\left[L_D\left(A\left(S\right)\right)\right] = 0$$

כפי שרצינו להוכיח.

.2

יהיו $\mathcal{X} := \mathbb{R}^2$**,** $\mathcal{Y} := \{0,1\}$ **ו־**$\mathcal{H}$ **מחלקה של עיגולים במרחב, כלומר** $\mathcal{H} := \left\{h_r : r \in \mathbb{R}_+,\ h_r\left(x\right) = 1_{\left[||x||_2 \leq r\right]}\right\}$**.**
נוכיח כי $\mathcal{H}$ **הינה למידה** PAC**, ושה־**$sample\ complexity$ **שלה חסום על ידי** $m_{\mathcal{H}(\varepsilon,\delta)} \leq \frac{log\left(\frac{1}{\delta}\right)}{\varepsilon}$**.**

<u>**פתרון:**</u> נוכיח זאת ישירות מהגדרת ה־$PAC\ learnability$ על ידי כך שנראה אלגוריתם ספציפי וניתוח ה־$sample\ complexity$, בדומה לאופן שראינו בתרגול עם האינטרוול.
בנוסף, נזכור שלכל $\varepsilon > 0$ מתקיים כי $1-\varepsilon \leq e^{-\varepsilon}$.

1. <u>אלגוריתם הלמידה</u>: בהינתן $S = \{(x_i, y_i)\}_{i=1}^m$ אלגוריתם הלמידה יחזיר את העיגול $\hat{C}$ שנותן את ההתאמה ההדוקה ביותר לדגימות החיוביות, כלומר העיגול בעל השטח הקטן ביותר שיכלול בתוכו את כל הדגימות עם לייבל 1 ולא יכלול בתוכו את כל הדגימות עם הלייבל 0. כמובן שאם לא קיימים כאלה נחזיר $\emptyset$. באופן פורמלי, A יחזיר $h_{\hat{r}}$ כך ש־$\hat{r}$ מגדיר את רדיוס המעגל האופטימלי, שאותו נרצה להביא למינימום. באופן פורמלי, האלגוריתם A יחזיר $h_{\hat{r}}$ כך ש־

$$\hat{r} = \underset{i:y_i=1}{max}||x_i||_2$$

2. <u>הוכחה למידות PAC</u>: נראה שלכל עיגול נכון C, לכל הסתברות D מעל $\mathcal{X}$ולכל $\varepsilon, \delta \in (0,1)$, אם נדגום באופן אחיד ובלתי תלוי דגימות מ־D עם הסתברות של לפחות $1-\delta$, העיגול $\hat{C}$ שיוחזר על ידי האלגוריתם שהגדרנו לעיל יהיה בעל טעות של לכל היותר ε.

משלב זה, כאשר נתאר מרחב בין שני עיגולים C ו־C'', זו תהיה הטבעת מרחב שחיה בין שפת העיגול הפנימי לשפת העיגול החיצוני.
נתחיל עם ההבחנה הבאה: העיגול המתאים ההדוק ביותר $\hat{C}$ תמיד מוכל בעיגול הנכון C. לכן, הטעות של האלגוריתם יכולה לבוא אך ורק מדגימות בעלות לייבל חיובי שנפלו בשטח של $C \setminus \hat{C}$, כלומר מתקיים כי $L_D\left(h_S\right) = \mathbb{P}\left(x \in C \setminus \hat{C}\right)$.
באופן אינטואיטיבי, נרצה לטעון שבהינתן דאטה סט גדול מספיק, לא סביר שההסתברות תחת D של דגימות חיוביות ב־$C \setminus \hat{C}$ הוא גבוה, ולכן שגיאת ההכללה לא אמורה להיות גבוהה מידי.
לכן, נרצה למצוא ערך של m שיבטיח שבהסתברות $1-\delta$ שגיאת ההכללה תהיה קטנה מ־ε, לכל $\varepsilon, \delta \in (0,1)$.
נרצה לחשב את ההסתברות לטעות, שזו כפי שראינו מתקיימת עבור דגימה $x_i \in [\hat{r}, r)$. נקבל כי

$$P_{S\sim D^m}\left[L_D\left(S\right) \leq \varepsilon\right] = 1 - P_{S\sim D^m}\left[L_D\left(h_S\right) > \varepsilon\right]$$

בשלב זה נשים לב כי אם ההסתברות של דגימה מ־$[\hat{r}, r)$ היא ε, אזי S לא יכלול דגימה מ־$[\hat{r}, r)$ בהסתברות של $(1-\varepsilon)^m$.
ככל ש־S יותר גדול, כך נהיה פחות סביר שחלק גדול מ־D לא יהיה ניתן לביטוי ב־S, ובאופן פורמלי

$$P_{S\sim D^m}\left[L_D^-\left(h_S\right) > \varepsilon\right] = P_{S\sim D^m}\left[\overset{m}{\underset{i=1}{\cap}} x_i \notin [\hat{r}, r)\right] = \overset{m}{\underset{i=1}{\Pi}} P_{x_i\sim D}\left(x_i \notin [\hat{r}, r)\right) \leq (1-\varepsilon)^m \leq e^{(-\varepsilon m)}$$

ולכן נקבל בסך הכל כי

$$P_{S\sim D^m}\left[L_D\left(S\right) \leq \varepsilon\right] \geq 1 - e^{(-\varepsilon m)} \geq 1 - \delta$$

ומכך נקבל כי כמות הדגימות הדרישה היא

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

ולכן בסך הכל נסיק כי מחלקת ההיפותזות של העיגול הינה למידה PAC, עם $sample\ complexity$ מהצורה

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

כפי שרצינו להוכיח.

$VC - Dimension$

3.

יהיו $\mathcal{X} := \{0,1\}^n$ **ו־**$\mathcal{Y} := \{0,1\}$**, ולכל** $I \subseteq [n]$ **נגדיר את פונקציית הזוגיות** $h_I(x) = \left(\sum_{i\in I} x_i\right) mod 2$.
נראה מהו ה־$VC - dimension$ **של מחלקת ההיפותזות** $\mathcal{H}_{parity} = \{h_I : I \subseteq [n]\}$.

<u>פתרון:</u> ראשית, נשים לב שגודל מחלקת ההיפותזות הוא 2^n, שכן קיימת פונקציה h_I עבור כל תת קבוצה I ב־n, ואנו יודעים שקיימות 2^n תתי קבוצות כאלו.
נרצה להראות כי ה־$VC - dimension$ של $\mathcal{H}_{pairity}$הוא n. לשם כך, נצטרך להראות ש־$\mathcal{H}$ מנטצת קבוצה בגודל n, ושהיא לא מנטצת כל קבוצה בגודל $n+1$.
1. נקח את קבוצת וקטורי הבסיס הסטנדרטי $C = \{e_1, ..., e_n\}$. עבור כל קבוצת לייבלים $(y_1, ..., y_n) \in \{0,1\}^n$ נשים לב שנוכל לקבל אותה באופן הבא:
עבור כל לייבל j שערכו 1, ההיפותזה $h_{\{j\}}$ תקבל אותו, שכן היא סוכמת את כל הכניסות ה־j של וקטורי הבסיס הסטנדרטי, וברור שסכום זה שווה ל־1.
באופן כללי, לקבל קבוצת ליבלים עם ערכי 1 במקומות שונים, ההיפותזה תהיה זו שעבורה I הינה קבוצת הערכים שהינם בדיוק אלו שערכם אחד בכניסות המתאימות.
לכן הראינו כי C מנוטצת על ידי מחלקת ההיפותזות $\mathcal{H}$, כלומר מתקיים כי $VC - Dim(\mathcal{H}) \geq n$.
2. מצד שני, ראינו בתרגול הוכחה לטענה כי עבור מחלקת היפותזות $\mathcal{H}$ בגודל סופי, מתקיים כי $VC - Dim(\mathcal{H}) \leq log_2|\mathcal{H}|$.
מכיוון שמתקיים במקרה זה כי $|\mathcal{H}| = 2^n$, הרי שנקבל מטענה זו כי $VC - Dim(\mathcal{H}) \leq n$, ולכן קיבלנו בסך הכל כי

$$VC - Dim(\mathcal{H}) = n$$

כפי שרצינו להוכיח.

4.

בהינתן מספר k**, יהי** $([a_i, b_i])_{i=1}^k$ **קבוצה של** k **אינטרוולים על** $\mathbb{R}$**, ונגדיר את האיחוד שלהם** $A = \bigcup_{i=1}^k [a_i, b_i]$.
מחלקת ההיפותזות $H_{k-intervals}$ **כוללת את הפונקציות** $h_A(x) = 1_{[x\in A]}$ **לכל בחירות של** k **אינטרוולים.**
נמצא את ה־$VC - dimension$ **של** $\mathcal{H}_{k-intervals}$ **ונוכיח את התשובה.**
בנוסף, נראה כי אם ניתן ל־A **להיות איחוד סופי כלשהו של אינטרוולים (כלומר** k **אינו מוגבל), אזי הקלאס המתקבל** $\mathcal{H}_{intervals}$ **הנו בעל** $VC - dimension = \infty$.

<u>פתרון:</u> נרצה להראות כי ה־$VC - Dimension$ של $\mathcal{H}_{k-intervals}$ הוא $2k$. לשם כך, נצטרך להראות ש־$\mathcal{H}$ מנטצת קבוצה בגודל $2k$, ושהיא לא מנטצת כל קבוצה בגודל $2k$.
1. נקח C קבוצה כלשהי של דגימות בגודל $2k$ המסודרת בסדר עולה, ונראה שכל קבוצת לייבלים $(y_1, ..., y_{2k}) \in \{0,1\}^{2k}$ נוכל לקבל באופן הבא:
ראשית, נשים לב שבמקרה המורכב ביותר הלייבלים יהיו מסודרים בזיגזג, כלומר 0,1,0,1..1,0,1. עבור k ערכי ה־1 נצטרך איטרוול שיתפוס כל דגימה בנפרד, וביניהם עבור דגימות שיקבלו לייבל 0 יהיו קטעים ללא אינטרוול קיים.
שנית, קל לשים לב שבכל מקרה אחר, בו יש שני לייבלים סמוכים שערכם זהה, אם מדובר בערך 1 נוכל להגדיר אינטרוול אחד שיתפוס את שניהם, או או מדובר בערך 0 פשוט לא יהיה שום אינטרוול בקטע זה, ולכן נוכל בצורה

קלה יותר לקבל את קבוצת הלייבלים, ובמקרים מסוימים אפילו לא יהיה צורך בכל k האינטרוולים כדי לקבל את קבוצת הלייבלים. לכן נקבל כי $VC-Dim\left(\mathcal{H}_{k-intervals}\right)\geq 2k$.

2. נראה כי ה־$VC-dimension$ של $\mathcal{H}_{k-interval}$ הוא לכל היותר $2k$. באופן אינטואיטיבי, גודל הלייבלים שלא נוכל כבר לקבל צריך לכלול $k+1$ ערכי 1 עם רווחים של לייבלים בעלי ערך 0 ביניהם, שכן במקרה זה נצטרך $k+1$ אינטרוולים זרים.
לכן, נניח בשלילה שקיימת קבוצת לייבלים בגודל $2k+1$ המנוטצת על ידי $\mathcal{H}_{k-intervals}$, כלומר, נוכל לקבל את קבוצת הלייבלים $(1,0,1...,0,1)$. זאת כמובן בסתירה שאין ברשותנו $k+1$ אינטרוולים זרים, שכן במקרה זה קיימות לנו $k+1$ דגימות שנצטרך לקבל, כשבין כל אחת לשנייה דגימה שנצטרך לא לקבל. לכן קיבלנו בסך הכל כי $VC-Dim\left(\mathcal{H}_{k-interval}\right)=2k$.

בנוסף לכל, נוכל כעת לשים לב שאם k כלל אינו מוגבל עבורנו, נוכל עבור כל קבוצת לייבלים y להתאים אינטרוול עבור כל לייבל שערכו 1 ולדאוג שלא יותאמו אינטרוולים לדגימות שערכן 0, זאת בצורה הכי ישירה שיש, ולכן במקרה זה $\mathcal{H}_{k-interval}$ תוכל לנטץ כל גודל של קבוצת דגימות C. כלומר במצב זה יתקיים כי $VC-dimension\left(\mathcal{H}_{k-interval}\right)=\infty$.

$Monotonicity$

5.

תהי $\mathcal{H}$ מחלקת היפותזות עבור קלסיפיקציה בינארית. נניח כי $\mathcal{H}$ הינה למידה PAC וכי ה־$sample\ complexity$ שלה נתון על ידי $m_{\mathcal{H}(\cdot,\cdot)}$.

נראה כי $m_{\mathcal{H}}$ הינה מונוטונית יורדת בכל אחד מהפרמטרים שלה. במילים אחרות, נראה כי:

א) בהינתן $\delta\in(0,1)$ ו־$0<\varepsilon_1\leq\varepsilon_2<1$, מתקיים כי $m_{\mathcal{H}}\left(\varepsilon_1,\delta\right)\geq m_{\mathcal{H}}\left(\varepsilon_2,\delta\right)$.

ב) בהינתן $\varepsilon\in(0,1)$ ו־$0<\delta_1\leq\delta_2<1$, מתקיים כי $m_{\mathcal{H}}\left(\varepsilon,\delta_1\right)\geq m_{\mathcal{H}}\left(\varepsilon,\delta_2\right)$.

<u>**פתרון:**</u> **א.** נניח בשלילה כי $\varepsilon_1\leq\varepsilon_2$ אבל מתקיים כי $m_{\mathcal{H}}\left(\varepsilon_1,\delta\right)<m_{\mathcal{H}}\left(\varepsilon_2,\delta\right)$.
על פי ההגדרה, $m_{\mathcal{H}}(\varepsilon,\delta)$ מייצג את מספר הדגימות המינימלי הדרוש על מנת לקבל כי $P_{S\sim D^m}\left(L_D\left(A\left(S\right)\right)\leq\varepsilon\right)\geq 1-\delta$, ובמילים אחרות להבטיח בהסתברות $1-\delta$ שגיאת ההכללה תהיה קטנה מ־ε.
לכן לפי ההנחה עבור $\delta\in(0,1)$ נקבל כי

$$P_{S\sim D^m}\left(L_D\left(A\left(S\right)\right)\leq\varepsilon_2\right)\geq P_{S\sim D^m}\left(L_D\left(A\left(S\right)\right)\leq\varepsilon_1\right)\geq 1-\delta$$

כלומר השתמשנו בפחות דגימות על מנת לקבל חסם טוב יותר על על הסתברות שגיאה ההכללה, וזו סתירה.

ב. נניח בשלילה כי $\delta_1\leq\delta_2$ אבל מתקיים כי $m_{\mathcal{H}}\left(\varepsilon,\delta_1\right)<m_{\mathcal{H}}\left(\varepsilon,\delta_2\right)$.
על פי ההגדרה, $m_{\mathcal{H}}(\varepsilon,\delta)$ מייצג את מספר הדגימות המינימלי הדרוש על מנת לקבל כי $P_{S\sim D^m}\left(L_D\left(A\left(S\right)\right)\leq\varepsilon\right)\geq 1-\delta$, ובמילים אחרות להבטיח בהסתברות $1-\delta$ שגיאת ההכללה תהיה קטנה מ־ε.
לכן לפי ההנחה עבור $\varepsilon\in(0,1)$ נקבל כי

$$P_{S\sim D^m}\left(L_D\left(A\left(S\right)\right)\leq\varepsilon\right)\geq 1-\delta_1\geq 1-\delta_2$$

כלומר השתמשנו בפחות דגימות על מנת לקבל חסם טוב יותר על על הסתברות שגיאה ההכללה, וזו סתירה.

6.

תהיינה $\mathcal{H}_1$ ו־$\mathcal{H}_2$ שתי מחלקות עבור קלסיפיקציה בינארית, כך ש־$\mathcal{H}_1\subseteq\mathcal{H}_2$. נראה כי $VC-dim\left(\mathcal{H}_1\right)\leq VC-dim\left(\mathcal{H}_2\right)$.

<u>**פתרון:**</u> נסמן $VC-dim\left(\mathcal{H}_1\right)=k_1$ ו־$VC-dim\left(\mathcal{H}_2\right)=k_2$, ונניח בשלילה כי $k_1>k_2$. מההגדרה של $VC-Dimension$, קיים קבוצת דגימות $C=\{c_1,...,c_{k_1}\}$ כך ש$\mathcal{H}_1$מנטצת את C אבל $\mathcal{H}_2$ לא, כלומר קיימת היפותזה $h\in\mathcal{H}_1$ כך שבהינתן קבוצת הדגימות C יכולה לקבל כל וקטור לייבלים בינארי אפשרי בגודל k_1.
בנוסף, נתון כי $\mathcal{H}_1\subseteq\mathcal{H}_2$, כלומר מתקיים כי $h\in\mathcal{H}_2$, כלומר גם $\mathcal{H}_2$ מנטצת את C, וזו סתירה.

.7

נוכיח כי אם $\mathcal{H}$ בעלת התכנסות במידה שווה עם הפונקציה $m_{\mathcal{H}}^{UC}:(0,1)^2\rightarrow\mathbb{N}$, אזי $\mathcal{H}$ הינה $Agnostic-PAC$ למידה עם $sample\ complexity$ $m_{\mathcal{H}}(\varepsilon,\delta)\leq m^{UC}\left(\frac{\varepsilon}{2},\delta\right)$.

<u>פתרון:</u> מתכונת ההתכנסות במידה שווה של H עם $m_{\mathcal{H}}^{UC}$ מתקיים כי לכל $\varepsilon,\delta\in(0,1)$ ולכל פונקציית התפלגות D מעל $\mathcal{X}$ כי אם $S=\{(x_i,y_i)\}_{i=1}^{m}$ סט דגימות בגודל $m\geq m_{\mathcal{H}}^{UC}$ הנדגמות באופן אחיד ובלתי תלוי מ־D מתקיים כי $D^m(\{S\in Z^m: S\ is\ \varepsilon\ representative\})\geq1-\delta$, כלומר מהגדרת $\varepsilon\ representative$ לכל $h\in\mathcal{H}$ $|L_{S_m}(h)-L_D(h)|\leq\varepsilon$.
מכיוון שהטענה נכונה לכל $\varepsilon\in(0,1)$, נקח $\varepsilon‘=\frac{\varepsilon}{2}\leq\varepsilon$ ונרצה להראות כי על פי הגדרת למידות $Agnostic\ PAC$ כי

$$D^m\left(S\in Z^m:\exists h\in\mathcal{H},\ L_{S_m}(h)\leq\underset{h\in\mathcal{H}}{min}L_D(h)+\varepsilon\right)\geq1-\delta$$

ראשית, נקבל עבור $\varepsilon‘\ representative$ כי $L_D(h_S)\leq\underset{h\in\mathcal{H}}{min}L_D(h)+\varepsilon‘$ עבור $h_S=\underset{h\in\mathcal{H}}{argmin}L_S(h)$, ולכן מההתכנסות במידה שווה מתקיים

$$D^m\left(S\in Z^m:L_D(h_S)\leq\underset{h\in\mathcal{H}}{min}L_D(h)+\varepsilon\right)\geq1-\delta$$

ועל כל קיבלנו כי עבור h כנ"ל $\mathcal{H}$ הינה למידה $Agnostic-PAC$ עם $sample\ complexity$ של $m_{\mathcal{H}}(\varepsilon,\delta)\leq m^{UC}\left(\varepsilon‘=\frac{\varepsilon}{2},\delta\right)$, כנדרש.

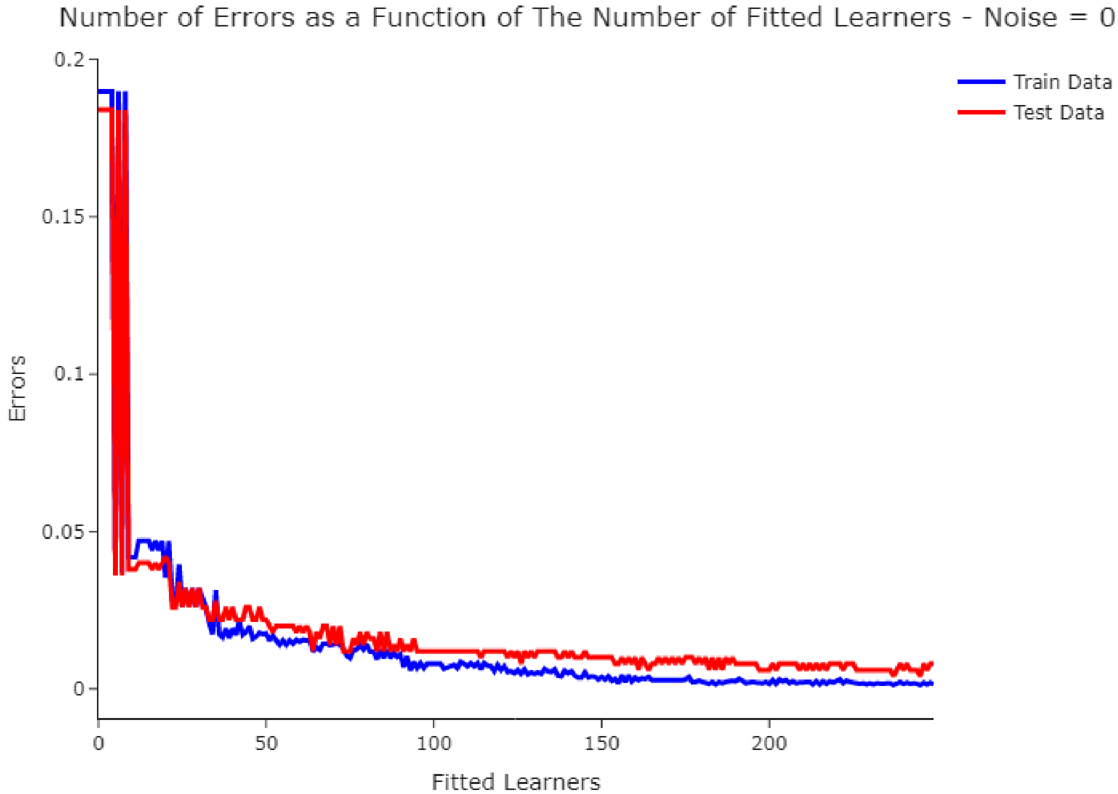
.8

תהי $\mathcal{H}$ מחלקה היפותזות מעל $\mathcal{Z}=\mathcal{X}\times\left\{\begin{matrix}+1\\-\end{matrix}\right\}$, עם פונקציית $loss_{0-1}$, ונניח שקיימת פונקציה $m_{\mathcal{H}}$ כך שלכל התפלגות D מעל $\mathcal{Z}$ קיים אלגוריתם A עם התכונה הבאה:
כאשר מריצים את A על $m\geq m_{\mathcal{H}}$ דגימות המתקבלות מ־D באופן אחיד ובלתי תלוי, מובטח שתוחזר בהסתברות לפחות $1-\delta$ היפותזה $h_s:\mathcal{X}\rightarrow\left\{\begin{matrix}+1\\-\end{matrix}\right\}$ עם $L_D(h_s)\leq min_{h\in\mathcal{H}}L_D(h)+\varepsilon$. נקבע האם $\mathcal{H}$ הינה למידה $Agnostic-PAC$ באמצעות הוכחה או דוגמה נגדית.

<u>פתרון:</u> **$\mathcal{H}$ איננה למידה $Agnostic\ PAC$, ונציג לכך דוגמה נגדית.**
ראשית, נשים לב שההבדל בתיאור אל מול הגדרת למידות $Agnostic\ PAC$ כפי שראינו היא שבמקרה זה קיים אלגוריתם A לכל התפלגות D, לעומת ההגדרה שבה קיים אלגוריתם לומד A לכל התפלגות D מעל $\mathcal{Z}$.
כעת, תהי $\mathcal{H}$ מחלקת היפותזות של פונקציות המתאימות לכל מספר טבעי לייבל ב־$\{-1,1\}$.
מצד אחד, נוכל לראות כי כל קבוצת דגימות C בגודל סופי שניקח איננה מנוטצת על ידי $\mathcal{H}$, ולכן ישירות מההגדרה שראינו בהרצאה מתקיים כי $\mathcal{H}$ איננה למידה $Agnostic-PAC$.
מצד שני, בהינתן פונקציה $m_{\mathcal{H}}$ והתפלגות שרירותית D מעל $\mathcal{Z}$, נוכל לקחת אלגוריתם שיעבור על כל הדגימות ויחזיר את הליבלים הנכונים. במקרה זה השגיאה תהיה מינימלית לכל ε,δ, כלומר התנאים הנ"ל מתקיימים, אף על פי שראינו ש־$\mathcal{H}$ כנ"ל איננה למידה $Agnostic\ PAC$, ולכן קיבלנו את מה שרצינו להוכיח.

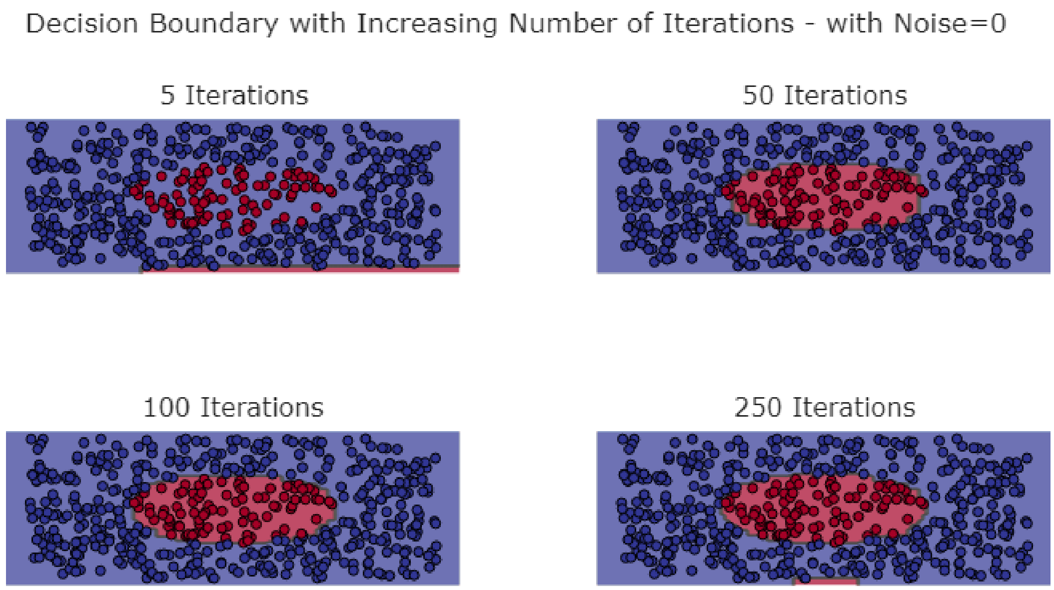
החלק המעשי

.1



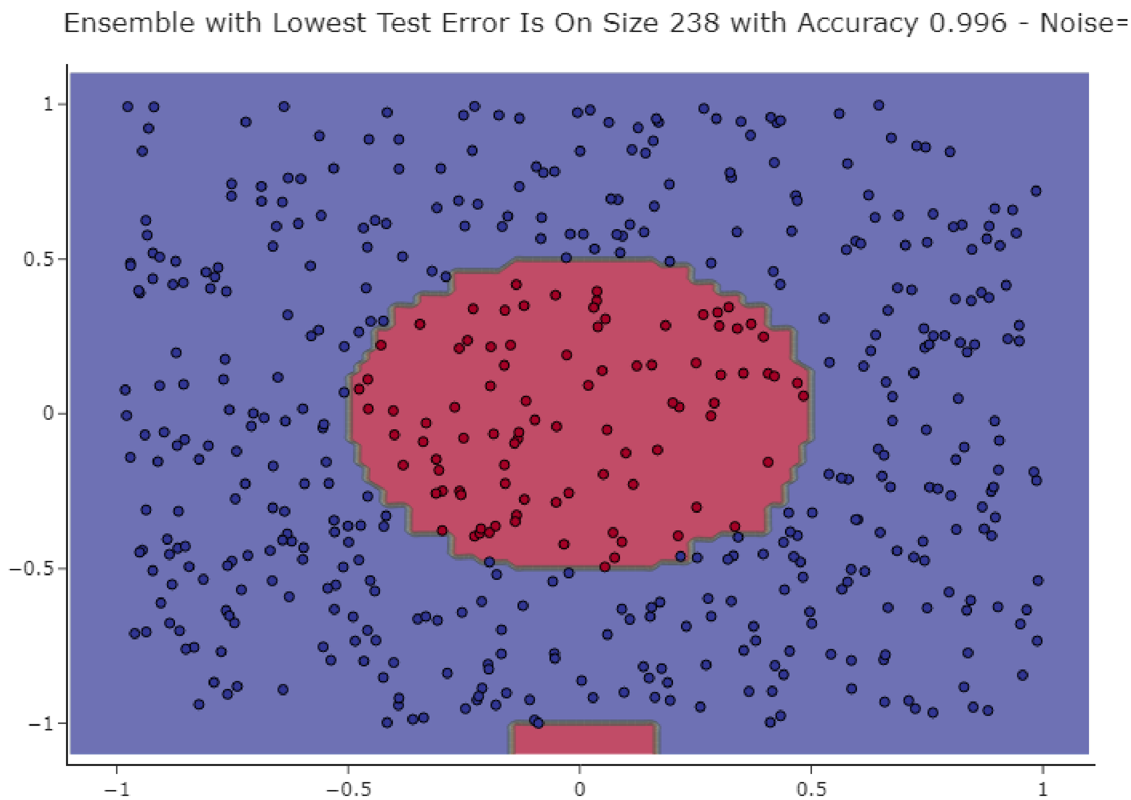
נוכל לראות שככל שנגדיל את מספר ה־$Weak\ learners$ שאנו משתמשים בהם ב־$Adaboost$, כך נקטין את השגיאה הן על ה־$Test$ והן על ה־$Train$,
כאשר לא כדאי להסתמך על כמות נמוכה מאוד של לומדים.

.2



גם כאן אנו רואים שהגדלת כמות האיטרציות שאנו מבצעים, המובילה לשימוש ביותר $weak\ learners$, תשפר את ה־$Decision\ Boundry$, בפרט על ה־$Test\ Set$ ולכן נראה שלא מדובר מדובר ב־$Over\ fitting$.
נשים את הדגש על כך שמדובר בדגימות ללא רעש, מה שלא מתרחש במציאות, ולכן יותר פשוט ל־$Adaboost$ להשתמש ב־$weak\ learners$ ולקבל תוצאה כזו טובה.

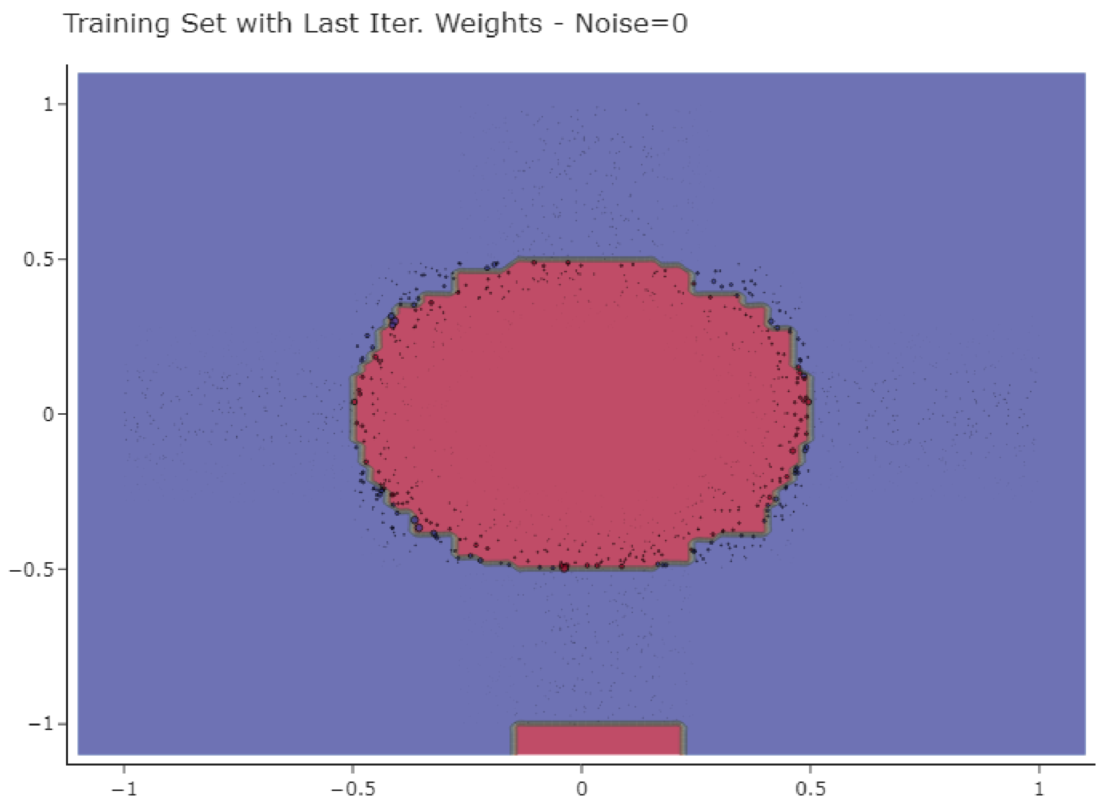
.3



ניתן לראות מגרף זה שעבור שימוש ב־238 *weak learners* נקבל את הדיוק הטוב ביותר ביותר, או במילים השגיאות השגיאה תהיה הקטנה ביותר שהיא 0.004.

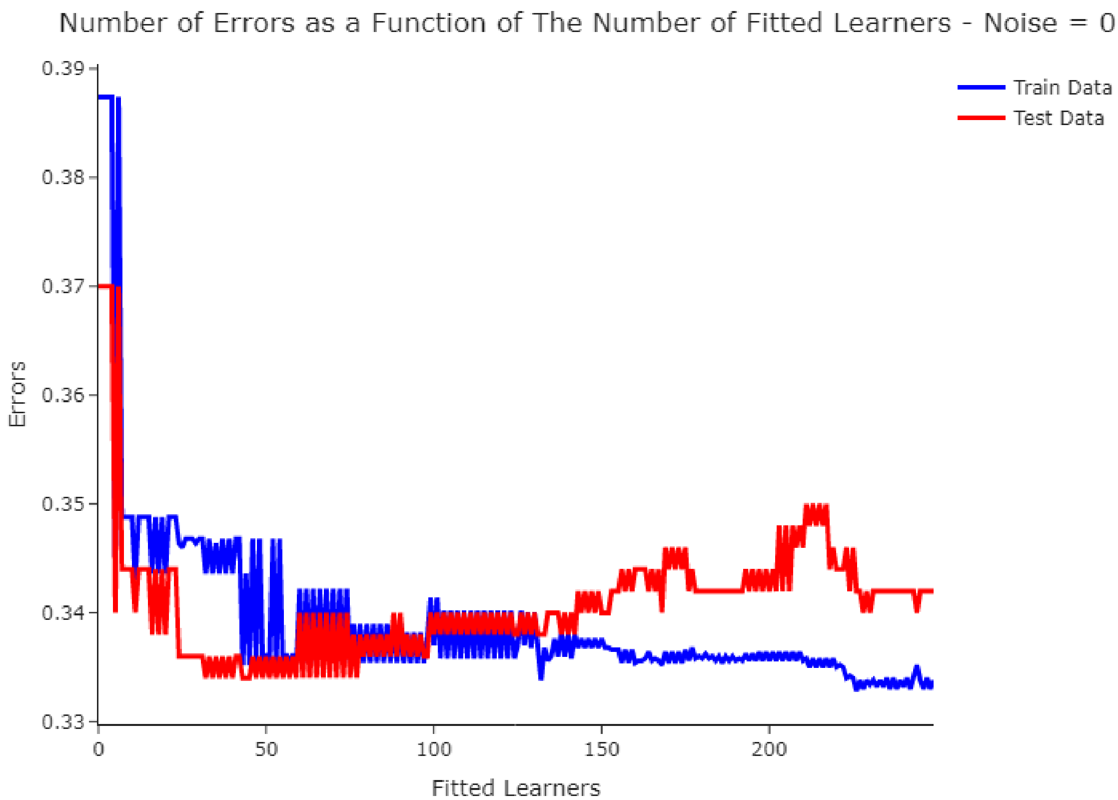
במקרה זה כפי שניתן, כאשר אין רעש, *Adaboost* מצליח לאמוד בצורה טובה מאוד את התפלגות ה־*Test data*.

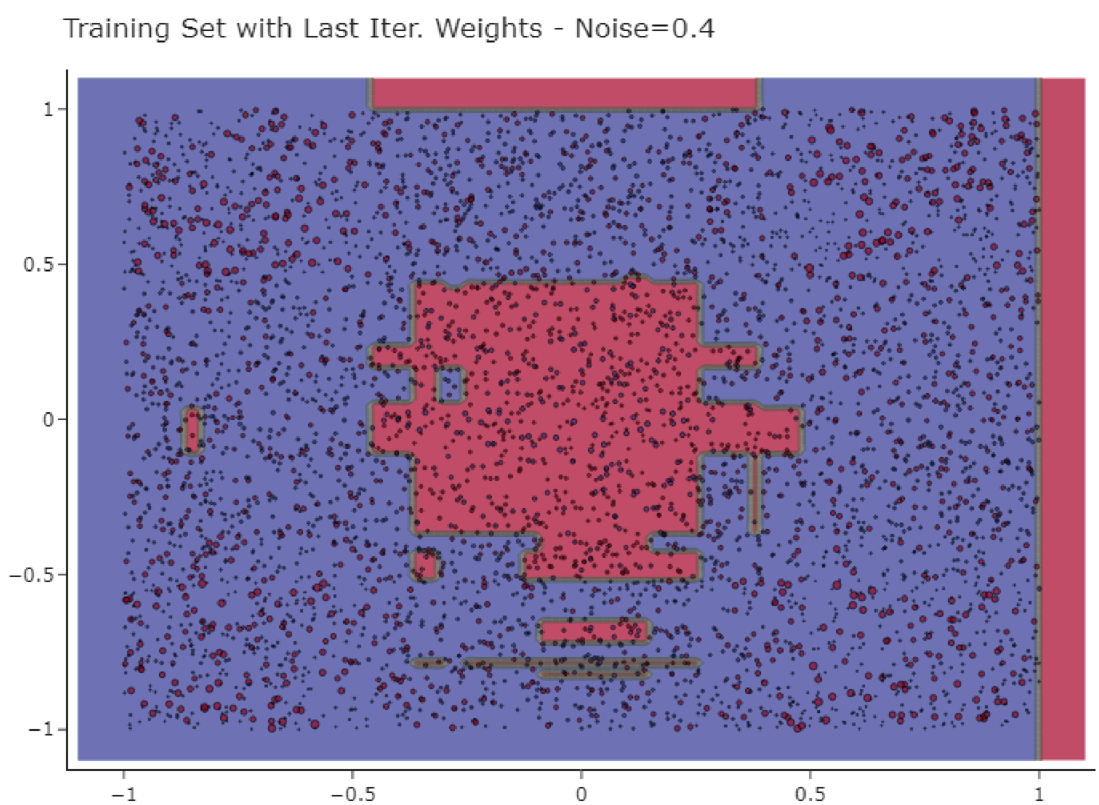
.4



כפי שראינו גם בתרגול ובהרצאה, האלגוריתם של $Adaboost$ פועל כך שהמשקל יגדל עבור דגימות האומד לא מצליח לסווג נכון, ולכן בגרף זה נוכל לראות דגימות אלה כגדולות ביותר.
על כן הדגימות שקלות עבור המודל הן אלו הקטנות יותר, שרובן בצורת $+$ בשטח הכחול הרחוק, ואלו שמאתגרות את האומד הן אלו שנמצאות בעיקר על שפת השטח האדום, ומוצאות כגדולות.

.5


Number of Errors as a Function of The Number of Fitted Learners - Noise = 0.
Errors
0.39
0.38
0.37
0.36
0.35
0.34
0.33
0
50
100
150
200
Fitted Learners
Train Data
Test Data


Training Set with Last Iter. Weights - Noise=0.4
1
0.5
0
−0.5
−1
−1
−0.5
0
0.5
1

ראשית, נשים לב ש־$Adaboost$ מתקשה מאוד במקרה זה לאמוד את התפלגות הדגימות, שכן בגרף הראשון אנו רואים שהשגיאה גדולה מאוד ביחס למקרה בו אין רעש, גם לאחר 250 איטרציות, אפילו עבור ה־$Train\ set$.
שנית, נוכל לראות כי כאשר אנו משתמשים במספר גדול של $weak\ learners$, קרי במקרה זה 250, האומד יתאים את עצמו ל־$Train\ set$ ויבצע $Overfitting$ עליו. בנוסף, כפי שנוכל לראות מהגרף השני, ככל שמספר זה יהיה גדול יותר כך יעלה ה־$variance$, שכן יהיה שינוי גדול עבור כל שינוי בדגימות, לעומת ירידה ב־$bias$ שכן ההתאמה ל־$Train\ set$ מתהדקת.