מבוא למערכות לומדות – 67577 – תרגיל 3

אופק אבידן – 318879574

2. <u>חלק תיאורטי:</u>

<u>PAC Learnability 2.1</u>

.1

2.1 PAC Learnability

1. 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

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

When proving, *do not* use a VC-Dimension argument. Instead, prove the claim directly from the PAC learnability definition by showing a specific algorithm and analyzing its sample complexity. Derive the sample complexity explicitly.
Hint: Remember that for every $\varepsilon > 0$ it holds that $1 - \varepsilon \leq e^{-\varepsilon}$
Hint 2: The main idea is similar to section 1.2 in recitation 6

נגדיר אלגו׳ A אשר בהנתן סט אימון

$$S = \{(X_i, y_i)\}_{i=1}^m \subseteq \mathbb{R}^2 \times \{0,1\}$$

A יחזיר מעגל φ אשר יכיל את כל המופעים החיוביים. המעגל φ ש-A מחזיר יהיה המעגל בעל הרדיוס הקטן ביותר שמכיל את המופעים האלה.

נניח ריאלייזביליות ויהי ψ מעגל עם שגיאה 0. נסמן ב-$r(\psi)$ את רדיוס המעגל. יהי $\epsilon > 0$ ויהי $r \leq r(\psi)$ המקיים:

$$D_x(\{x: r \leq ||x|| r(\psi)\} = 0$$

נגדיר $E = \{X \in \mathbb{R}^2 \ \ s.t. \ \ r \leq ||x|| \leq r(\psi)\}$. נשים לב כי ההסתברות ש-$L_D(\varphi) \geq \epsilon$ חסומה מלעיל על ידי ההסתברות שאף ב-S שייכת ל-E ומתקיים:

$$\mathbb{P}(L_D(\varphi) \geq \epsilon) \leq (1-\epsilon)^m \leq e^{-\epsilon m} \leq \delta$$

בנוסף,

$$e^{-\epsilon m} \leq \delta \Rightarrow m(\epsilon, \delta) \leq \left(\frac{\log\left(\frac{1}{\delta}\right)}{\epsilon}\right)$$

2.2 ***VC-Dimension***

2. 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 $VC-dim(\mathcal{H}_1) \leq VC-dim(\mathcal{H}_2)$.

נגדיר $\mathrm{VC}-\dim(H_1) = d_1$. מההגדרה, זה אומר שקיימת קבוצה של d_1 נקודות שיכולה להינתץ על ידי H_1, ואין שום קבוצה בגודל $d_1 + 1$ נקודות שיכולות להינתץ על ידי H_1.

מכך ש-$H_1 \subseteq H_2$, כל פונקציית היפותזה ב-H_1 נמצאת גם ב-H_2. לכן אם קבוצה של נקודות מנותצת על ידי H_1, אז הקבוצה הזו מנותצת גם על ידי H_2 שכן כל הלייבלים שאפשריים עם H_1 אפשריים גם עם H_2. מכאן ניתן לומר כי המספר המקסימלי של נקודות ש-H_1 יכולה לנתץ (שהוא d_1) הוא החסם התחתון של מספר הנקודות ש-H_2 יכולה לנתץ. לכן, $\mathrm{VC}-\dim(H_2) \geq d_1 = VC-Dim(H_1)$ כנדרש.

3. Let $\mathcal{X} = \{0,1\}^d$ and $\mathcal{Y} = \{0,1\}$, and assume $d \geq 2$. Each sample $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ consists of an assignment to d boolean variables ($\mathbf{x}$) and a label (y). For each boolean variable x_k, $k \in [d]$, there are two literals: x_k and $\bar{x}_k = 1 - x_k$. The class $\mathcal{H}_{\text{con}}$ is defined by boolean conjunctions over any subset of these $2d$ literals. Compute the VC dimension of $\mathcal{H}_{\text{con}}$ and prove your answer.
   Hint: Let's start by understanding how Boolean Conjunctions class work. For example: let $d = 5$ and consider the hypothesis that labels $\mathbf{x}$ according to the following conjunction

For $\mathbf{x} = (0,1,1,1,1)$ the label would be 0, and for $\mathbf{x} = (1,1,0,0,0)$ the label would be 1. The conjunction over the empty set is defined as the 'all positive hypothesis' ($h^+(\mathbf{x}) = 1$ for all $\mathbf{x}$), and any conjunction that contains $x_k \wedge \bar{x}_k$ for some k gives the 'all negative hypothesis' ($h^-(\mathbf{x}) = 0$ for all $\mathbf{x}$).

חשוב לציין שהייתי בשעת קבלה של גבי וראיתי את הפתרון שם

הפתרון באופן כללי יהיה להראות כי $VC-Dimension(H_{con}) \leq d$, לאחר מכן להראות כי $VC-Dimension(H_{con}) \geq d$ ולכן מטריכוטומיה נקבל $VC-Dimension(H_{con}) = d$

נראה תחילה כי $VC-Dimension(H_{con}) \leq d$:

נראה שלא קיימת קבוצה מגודל *d+1*, כאשר נסמנה *C*, כך ש-H_{con} תנתץ אותה. מההגדרה, זה יוכיח כי $VC-Dimension(H_{con}) \leq d$.

תהי קבוצה שרירותית של וקטורים $C = \{c_1, \dots, c_{d+1}\}$, דהיינו $|C| = d + 1$, כך שכל וקטור נמצא ב-$\{0,1\}^d$. מעקרון שובך היונים קיים וקטור מתוך הקבוצה c_t $(1 \leq t \leq d+1)$, כך שלכל $j \in \{1, \dots, d\}$ קיים $k \in \{1, \dots, d+1\} \backslash \{t\}$. כמו כן מתקיים $(c_k)_j = (c_t)_j$. דהיינו, לכל קואורדינטה בוקטור c_t קיים וקטור נוסף **שונה** מאותה הקבוצה, עם אותה קואורדינטה.

כעת, נוכל להראות שלא ניתן להשיג את ה*labeling* הבא:

$$y_i = \begin{matrix} 1 & i \neq t \\ 0 & i = t \end{matrix}$$

שכן, וכאן מגיע הטוויסט, נניח בשלילה שקיימת $h \in H_{con}$ פונקציה שמצליחה ליצור את הלייבלינג המדובר, דהיינו

$$h(x_i) = y_i = \begin{matrix} 1 & i \neq t \\ 0 & i = t \end{matrix}$$

כעת, אם זה נכון, אז ניתן לומר כי לכל $c \in C\backslash\{c_t\}$ שמקיימים את האילוץ, כי כדי לקבל 1 בכניסות (קואורדינטות) ששונות מ-t חייב להתקיים $h(x_j) = 1$ $\forall j \in \{1, \dots, d+1\}\backslash\{t\}$. אנו גם יודעים ממקודם כי $h(c_t) = 0$ ומכך קיים ליטרל אחד (**לפחות**) ב-h ש-c_t מפר אותו (אחרת היה 1). נסמן את הקואורדינטה הראשונה (הליטרל הראשון) שמפר את h ב-x. כעת, נשים לב כי קיים וקטור אחר $y \in C\backslash\{c_t\}$ **שונה** כך ש-$c_x = (c_t)_x$ ומכאן ש-$c(c_x) = 0$ בסתירה. למדנו כי אם לא קיימת h כזו, ומכך ש-h נבחרה באופן שרירותי, אזי לכל h לא ניתן להשיג את הלייבלינג y ומההגדרה של ניתוץ ו-$VCDim$ נקבל כי $VC - Dimension(H_{con}) \leq d$.

כעת, נראה את החלק השני בהוכחה, שהוא $VC - Dimension(H_{con}) \geq d$, הישארו עמנו!

נמצא את הקבוצה בגודל d שתענה על הניתוץ ואז $VC - Dimension(H_{con}) \geq d$ ונסיים. אך מי זו תהיה? נקרא לקבוצה הזו C, כך שכאמור, $|C| = d$ ו-H_{con} מנתצת אותה. נגדיר אותה לרגע להיות וקטורי הבסיס של $\mathbb{R}^d$, דהיינו, $C = \{e_1, \dots, e_d\}$. כעת לחלק המעניין: יהי וקטור שרירותי $x \in \{0,1\}^d$, ונוכיח כי $H_{con_c} = \{h_c = (h(e_1), \dots, h(e_d)) | h \in H_{con}\}$. נסמן את האינדקסים של קואורדינטות האפסים בוקטור x ב-$x_1, \dots, x_k$ (ואכן, $0 \leq k \leq d$). דהיינו x_1 הוא הכניסה הראשונה ששווה 0, x_2 השניה וכך הלאה...

כעת סוף סוף נבנה את h_{con} בצורה מעט מתוחכמת אך בצורה הבאה: $h \coloneqq v_{\tilde{x_1}} \wedge \dots \wedge v_{\tilde{x_k}}$. כעת נשים לב שבמידה והוקטור x הוא וקטור ללא כניסות עם אפסים $h = \emptyset$ ולכן באופן ריק מתקיים כי $h(v) = 1$ לכל $v \in \{0,1\}^d$.

נשים לב כי מתקיים: $v = (h(e_1), \dots, h(e_d))$ וקטור שורה שכן לכל $j \in \{1, \dots, k\}$ אזי $h(e_j) = 1$ (כאשר $x_j = 1$), שכן $j \notin \{x_1, \dots, x_k\}$ ומכאן נקבל ש- $\tilde{x_j}$ לא יהיה חלק מהוקטור. בסה"כ $x \in H_{con_c}$ ומכאן ש-$VC - Dimension(H_{con}) \geq d$.

2.3 *Sample Complexity*

2.3 Sample Complexity

4. 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 $0 < \varepsilon_1 \le \varepsilon_2 < 1$, we have that $m_{\mathcal{H}}(\varepsilon_1,\delta) \ge m_{\mathcal{H}}(\varepsilon_2,\delta)$.
   - Similarly, show that given $\varepsilon \in (0,1)$, and given $0 < \delta_1 \le \delta_2 < 1$, we have that $m_{\mathcal{H}}(\varepsilon,\delta_1) \ge m_{\mathcal{H}}(\varepsilon,\delta_2)$.

נפרק את ההוכחה לשני חלקים:

חלק ראשון – מונוטוניות ביחס ל-ϵ

טענה: בהינתן $\delta \in (0,1)$ ו-$0 < \epsilon_1 \le \epsilon_2 < 1$, נקבל כי $m_H(\epsilon_1,\delta) \ge m_H(\epsilon_2,\delta)$.

מההגדרה, H היא *Pac learnable* אם, לכל $\epsilon > 0$ ולכל $\delta > 0$, קיים גודל מדגם m כך שעם הסתברות לפחות $1-\delta$, ההיפותזה הנלמדת h תשגה בלכל היותר ϵ.

כדי להשיג שגיאה קטנה יותר ϵ_1 כך ש-$\epsilon_2 > \epsilon_1$, באופן אינטואיטיבי, יהיה צורך בגודל מדגם גדול יותר מכיוון שהיפותזה מדויקת יותר (עם שגיאה קטנה יותר) דורשת בדרך כלל יותר נתונים ללמוד מהם.

באופן פורמלי, נניח בשלילה כי $m_H(\epsilon_1,\delta) < m_H(\epsilon_2,\delta)$. זה מרמז שדרושות פחות דוגמאות כדי להבטיח שגיאה של לכל היותר ϵ_1 ממה שצריך לשגיאה של לכל היותר ϵ_2, מה שסותר את האינטואיציה ואת העיקרון שהשגת שגיאה נמוכה יותר דורשת בדרך כלל יותר נתונים.

חלק שני – מונוטוניות ביחס ל-δ

טענה: בהינתן $\epsilon \in (0,1)$ ו-$0 < \delta_1 \le \delta_2 < 1$, נקבל $m_H(\epsilon,\delta_1) \ge m_H(\epsilon,\delta_2)$.

שוב, מההגדרה, H היא *Pac-learnable* אם, לכל $\epsilon > 0$ ולכל $\delta > 0$, קיים גודל מדגם m כך שעם הסתברות לפחות $1-\delta$, ההיפותזה הנלמדת h תשגה בלכל היותר ϵ.

כדי להגיע לרמת ביטחון גבוהה יותר (כלומר δ_1 קטן יותר כך ש-$\delta_1 < \delta_2$), בדרך כלל יהיה צורך בגודל מדגם גדול יותר מכיוון שביטחון גבוה יותר אומר שההיפותזה חייבת להיות מדויקת בוודאות רבה יותר.

באופן פורמלי, נניח בשלילה כי $m_H(\epsilon,\delta_1) < m_H(\epsilon,\delta_2)$. זה מרמז שדרושות פחות דגימות כדי להבטיח את הביטחון הגבוה ביותר (בהסתברות $1-\delta_1$). מה שסותר את העיקרון שהשגת ביטחון גבוה יותר דורשת יותר נתונים.

לכן, $m_H(\epsilon,\delta_1) \ge m_H(\epsilon,\delta_2)$ מוכיח ש-$m_H(\epsilon,\delta)$ היא לא-עולה ביחס ל-δ.

בסה״כ, הראינו כי סיבוכיות המדגם $m_H(\epsilon,\delta)$ היא מונוטונית יורדת ב-ϵ ו-δ:

- $m_H(\epsilon,\delta)$ היא לא-עולה ביחס לϵ
- $m_H(\epsilon,\delta)$ היא לא-עולה ביחס לδ

מאפיינים אלה תואמים את האינטואיציה שהשגת שגיאה קטנה יותר או ביטחון גבוה יותר דורשת יותר דוגמאות.

2.4 *Agnostic-PAC*

2.4 Agnostic-PAC

5. Let $\mathcal{H}$ be a hypothesis class over a domain $\mathcal{Z} = \mathcal{X} \times \{\pm 1\}$, 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 $h_S : \mathcal{X} \to \{\pm 1\}$ 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.

חשוב לציין שהייתי בשעת קבלה של גבי וראיתי את הפתרון שם

5. H לא בהכרח למידת אגנוסטיק-פאק. ניתן דוגמא נגדית תוך כדי שנעזר בתרגול – עבור מחלקת ההיפותזה הבאה: $H = \{X \to sign(\sin(\theta x)) \mid \theta \in \mathbb{R}^+\}$ כך ש-$X = \mathbb{R}\backslash\{0\}$

כעת ניעזר בתרגול של עמית – שם ראינו כי מחלקת היפותזות זו אינה למידת PAC שכן ה-vc *dimension* הוא אינסופי (שכן, היא מנתצת כל סט גודל לכל *labeling*).

לכן, העובר האלגוריתם A המחזיר תמיד $L_D(h)$ $h_D = \arg\min\_(h \in H)$ מתקיים $L_D(h_d) =$ $\min\_(h \in H) \quad L_D(h_d)$

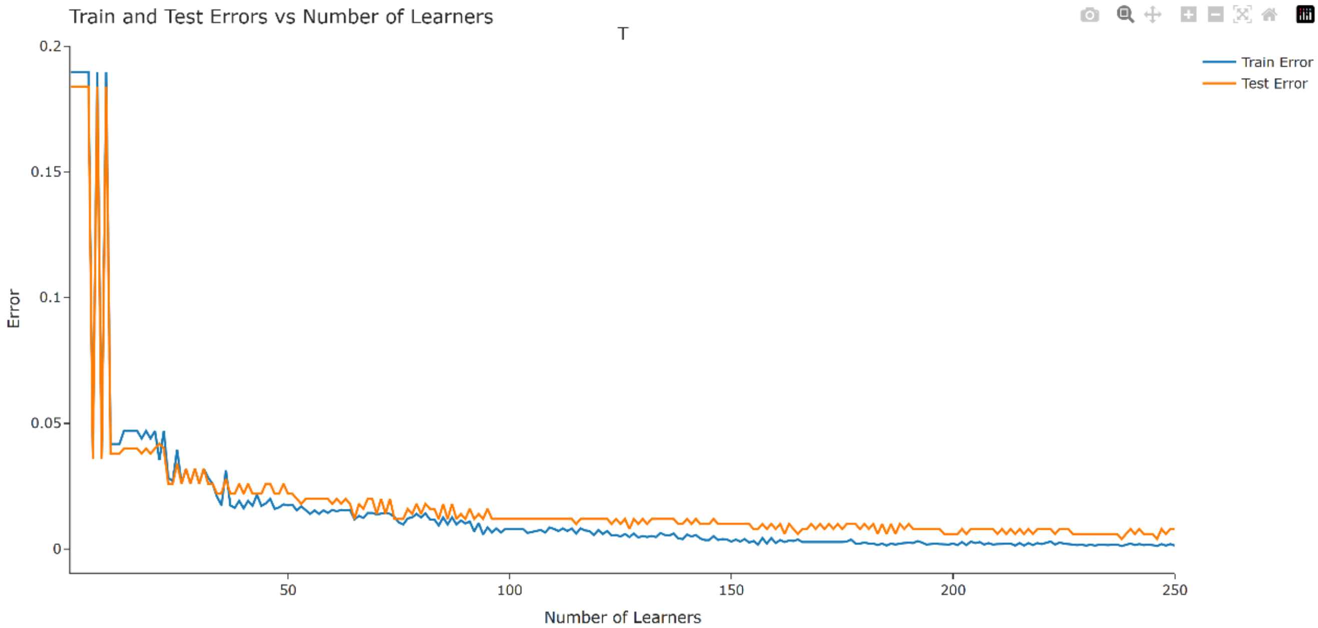
ומכך

$$\boldsymbol{P(L_D(h_d) = min_{h\in H}\, L_D(h_d) \leq min_{h\in H} L_d(h) + \epsilon) \geq 1-\delta}$$

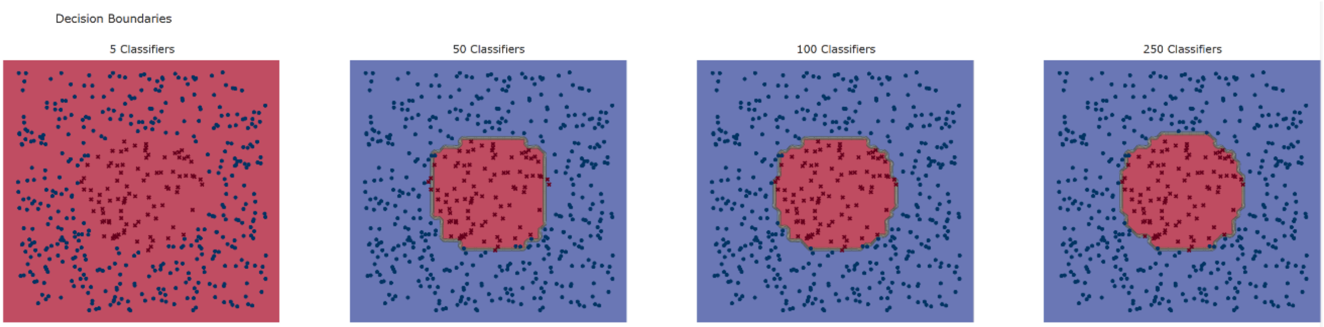
3. <u>**חלק פרקטי**</u>

3.1 Boosting

3.1.1. ניתן לראות כי ככל שמספר הלומדים עולה כך יורדת השגיאה על האימון והטסט. אמנם יש עליות קטנות, ברור כי הכיוון הכללי הוא ירידה. ניתן להסביר את העליות הקטנות מכיוון שיכול להיות שהאנסמבל מוסיף לומדנים שהורסים לו את הלוס במקרים מסויימים. באופן כללי, מכיוון שהשגיאה בטסט גדולה ממש בקצת מהשגיאה באימון, ניתן להסיק כי המודל עושה את העבודה ושאין חשש להתאמת יתר (*overfitting*) משמעותי, שכן אם היה אוברפיטינג, היינו רואים שגיאה נמוכה מאוד עבור סט האימון ושגיאה גבוהה בסט הבוחן.

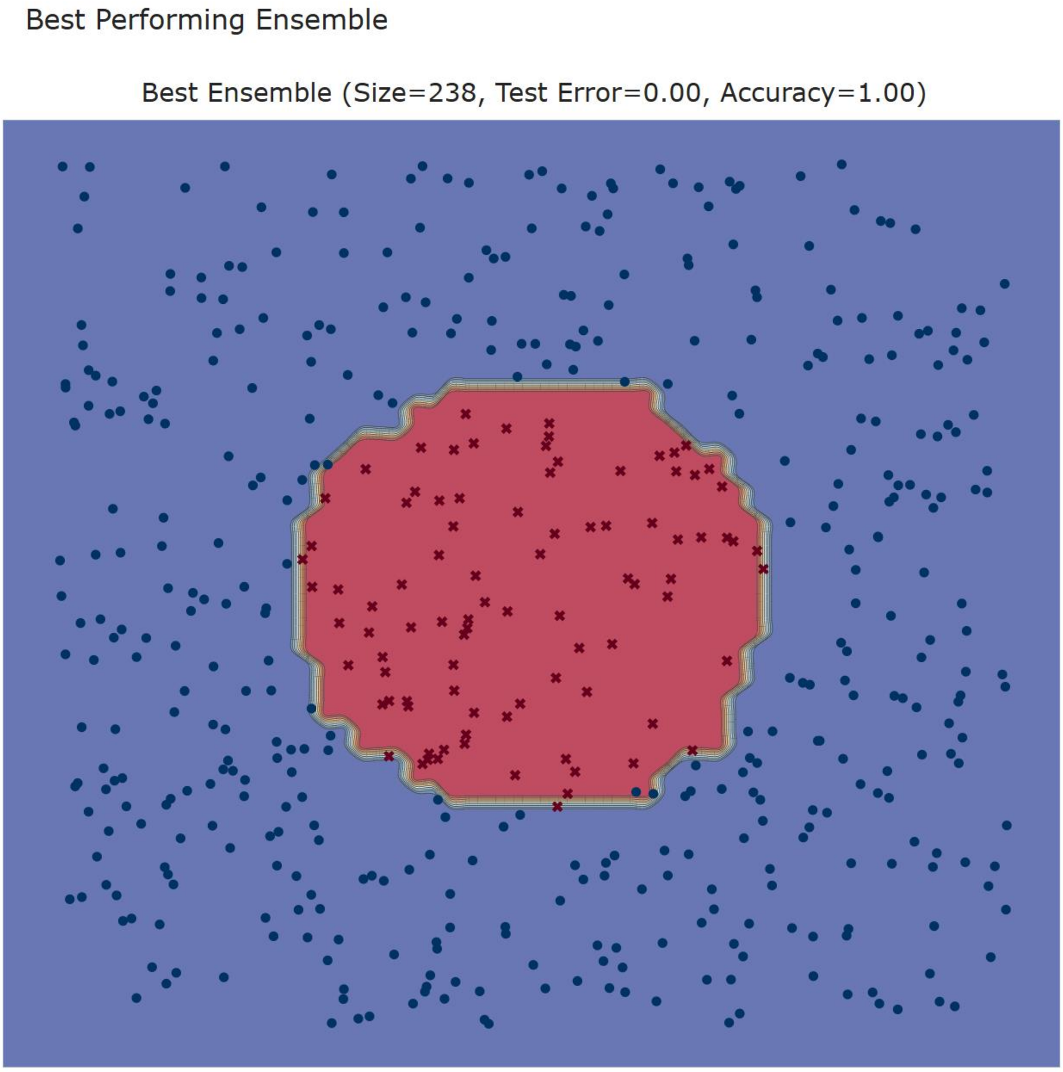


3.1.2. המשימה היא להפריד את האיקסים האדומים מהעיגולים הכחולים, וליצור לכל אחד מהם "קלאסטר". נשים לב כי מידע זה אינו פריד לינארית (לא נוכל להעביר שום קו שיחלק את המידע מעל/מתחת/משמאל/מימין לצבע אחד ומה שמכיוון הנגדי לצבע השני). אמנם, ככל שאנו מגדילים



את גודל האנסמבל כך המפריד נהיה חלק יותר ומדויק יותר ומצליח לעמוד במשימה בצורה יחסית טובה. לדוגמא, נשים לב כי כשיש בועידה 50 לומדים אז עדיין יש כמה נקודות אדומות שלא נתחמות לקלאסטר האדום. לבסוף רואים שעם 250 המשימה בוצעה בצורה טובה משמעותית מאשר עם 5 אשר סיווג את הכל כאדום.

.3.1.3

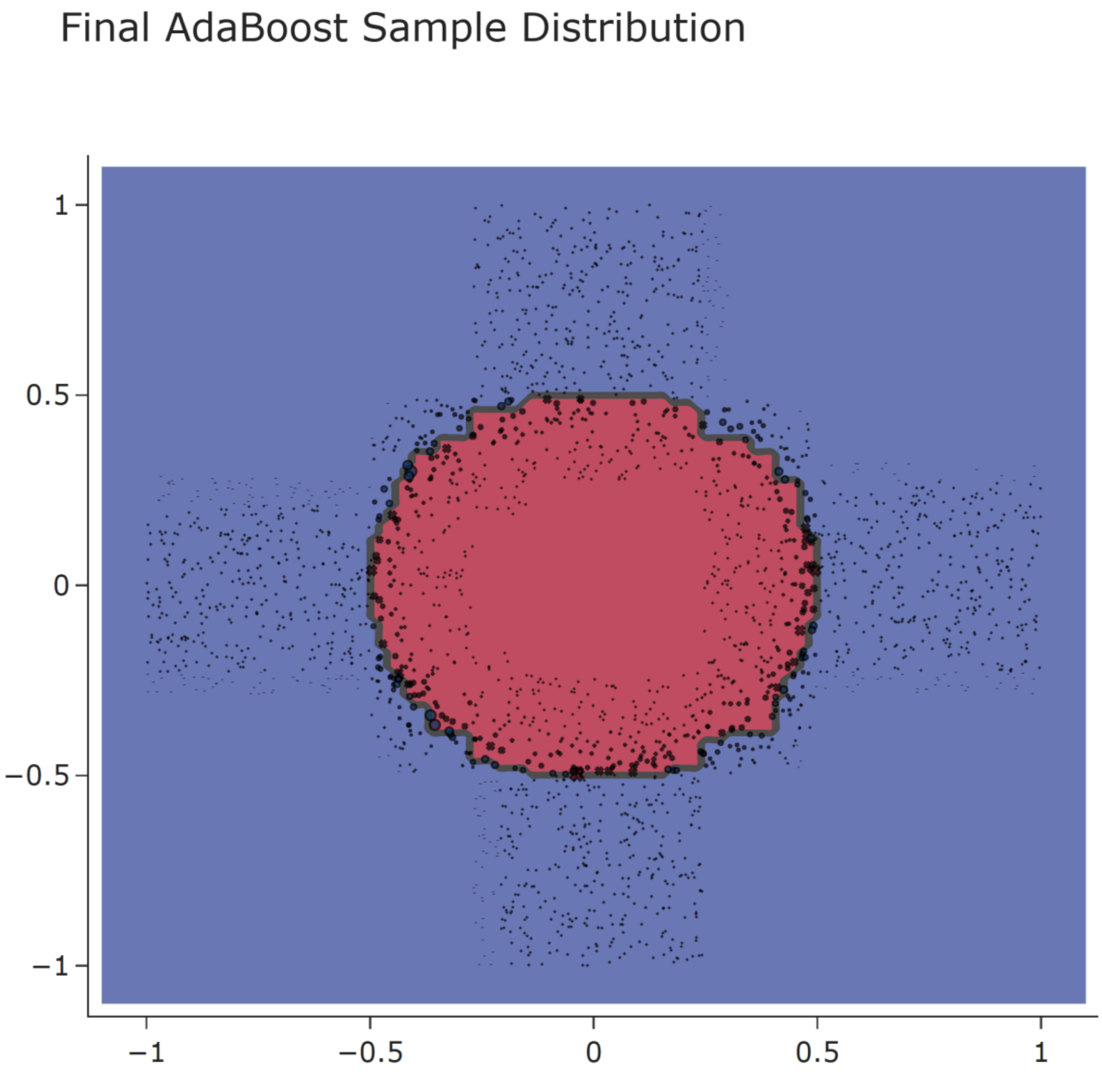


3.1.4. באופן כללי, הפלוט ממחיש את התפלגות המדגם של ה*adaboost* הסופי, עם גדלי נקודות פרופורציונליים למשקלם מהאיטרציה האחרונה. הנקודות הקטנות יותר מייצגות דגימות עם משקלים קטנים יותר, דהיינו, כנראה שהיה קל לסווג אותן נכון, כפי שלמדנו בכיתה. מן הצד השני אפשר לראות נקודות גדולות יותר שהיו מאתגרות יותר עבור הלומדים.

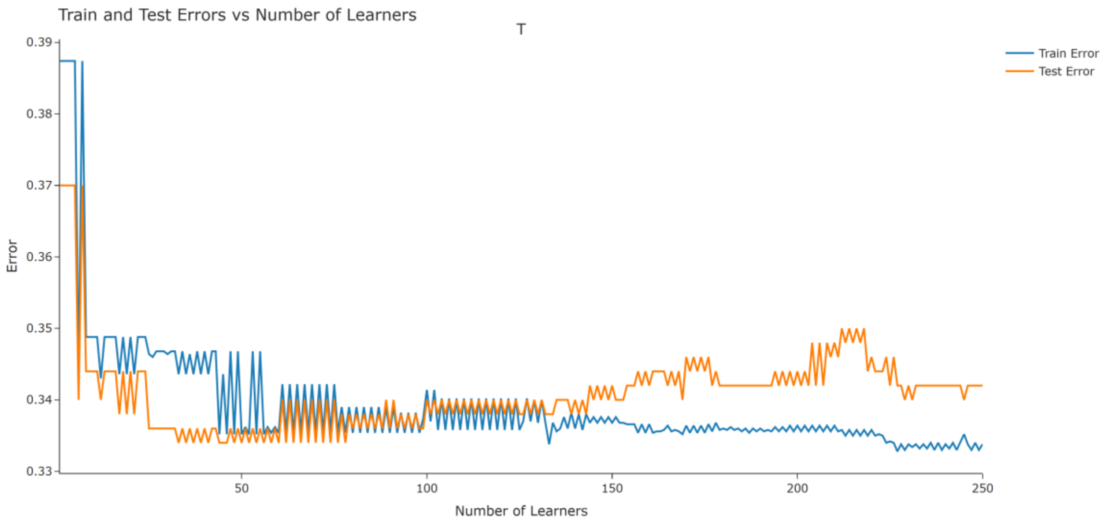
הנקודות האדומות שעמוק בתוך המעגל כאמור קטנות יותר ובפרט בעלות משקל נמוך יותר. ממה שלמדנו בכיתה ולפי איך שמוגדר אלגוריתם אדאבוסט, זה סימן שהצלחנו לסווג אותן יחסית בתחילת האלגוריתם. דהיינו הם כנראה סווגו בהצלחה כבר על ידי ה*weak-learners* הראשונים שהאדאבוסט הריץ (*decision stumps*). גם הנקודות הכחולות שנמצאות מחוץ לעיגול אבל מוצגות באופן קטן סווגו בהצלחה כנראה יחסית בתחילת האלגוריתם ומכאן קיבלו משקל נמוך

בדיוק כמו הנקודות האדומות הקטנות במעגל. באופן כללי את הנקודות הקטנות היה ניתן לסווג טוב בקלות יחסית בגלל שהן שוכנות באזורים שבהם ה-*decision stumps* מתפקדים היטב.

בנוסף, ניתן לראות כי הנקודות שקרובות לגבול המעגל קיבלו את המשקל הכי גדול כי הן היו המאתגרות ביותר עבור המסווגים. מכאן שכנראה הלומדנים טעו על סיווג נקודות אלו מה שגרם לעלייה במשקל של הנקודות לפי הגדרת אלגוריתם אדאבוסט. מה שעוד ניתן להבין זה שלא במקרה הנקודות הגדולות יותר נמצאות על גבול המעגל. יותר קל ללומדנים לסווג את נקודות שנמצאות על צורת הצלב (ניתן לראות שנוצר צלב מנקודות קטנות), וזאת בשל שהלומדנים באנסמבל אדאבוסט הם גדמי החלטה (*decision stump*) שמיושרים לצירים. זה אומר שהם יכולים לעשות רק פיצולים בניצב לצירים. כתוצאה מכך, דגימות שאינן ניתנות להפרדה בקלות על ידי פיצולים כאלה (כמו אלה ליד הגבול או לאורך הצירים) מהוות אתגר וקשה יותר לסווג אותן בצורה נכונה.

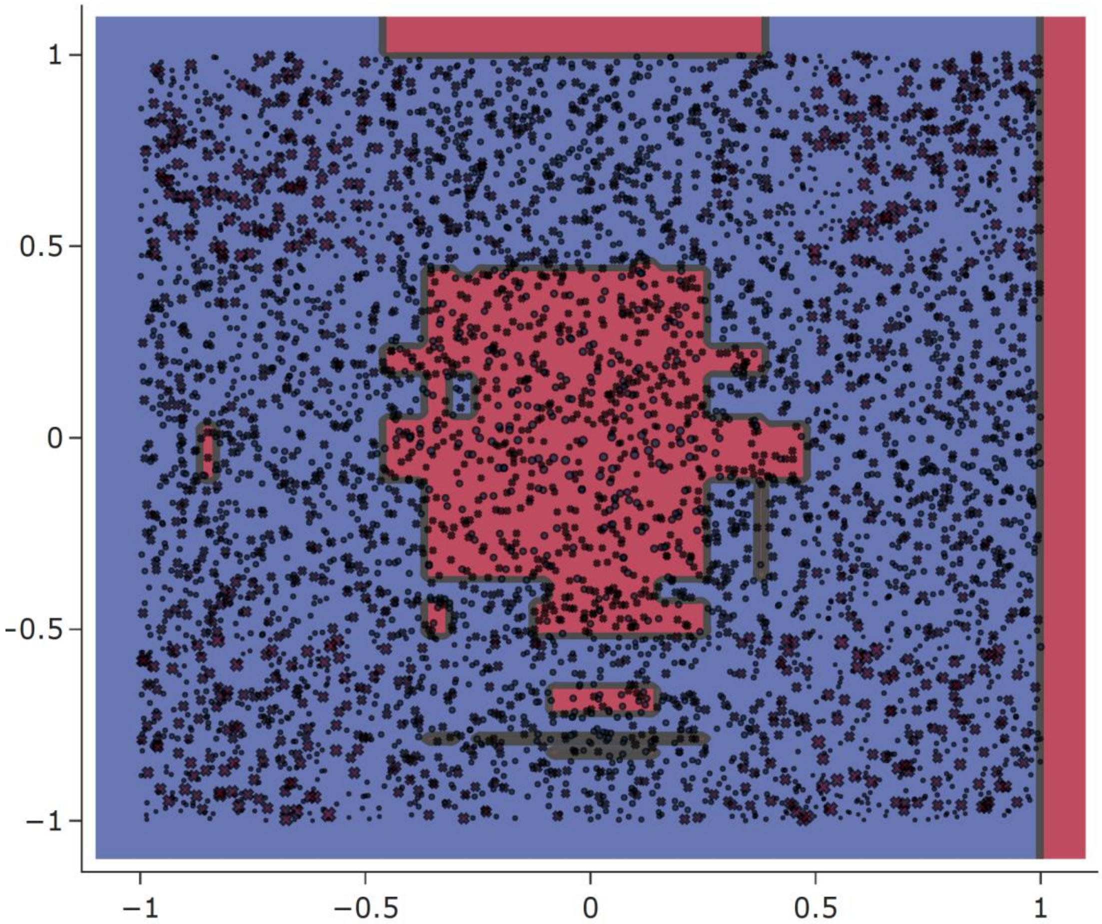


3.1.5. בפלוט הבא ניתן לראות כי כאשר מאמנים את המודל על סט אימון עם רעש ועם מספר נמוך של לומדנים, האדאבוסט לא מצליח ללמוד מספיק טוב ומתקבלת שונות נמוכה ועם *bias* גבוה. ניתן לומר כי הנקודה האופטימלית עבור הטריידאוף בין ה-*bias* לבין השונות היא באיזור ה-125 לומדנים, שכן ככל שמגדילים את מספר הלומדים ניתן להבחין שהשגיאה באימון יותר קטנה באופן משמעותי מהשגיאה בטסט. זה ככל הנראה נובע מהתאמת יתר לאימון, שכן השגיאה נמוכה מאוד ספציפית לסט אימון, כלומר השונות גבוהה וה-*bias* נמוך. דהיינו, אין הכללה לתופעה אלא יש תשובה ספציפית מאוד לכל דגימה מסט האימון.



בגרף המראה את חלוקת הדוגמאות של ה-*AdaBoost* הסופי עם רעש ברמה של 0.4, ניתן לראות שנקודות פנימיות בתוך המעגל האדום והכחולות הרחק ממנו קיבלו משקלים נמוכים, מכיוון שהן היו קלות לסיווג. נקודות קרובות לגבול המעגל ונקודות באזורי הצלב היו קשות יותר לסיווג וקיבלו משקלים גבוהים יותר, מה שמראה שהן גרמו לאלגוריתם יותר בעיות בסיווג נכון. הרעש בנתונים הוביל לקשיים נוספים ולמשקלים גבוהים יותר עבור דוגמאות מסוימות. באופן כללי ניתן לראות כי המודל טועה על סט אימון המכיל רעש משמעותית יותר מאשר בסט ללא רעש.

Final AdaBoost Sample Distribution



שימוש ב-*Chat GPT*

השתמשתי בצ׳אט ג׳יפיטי ברוב החלק התכנותי. כמו כן נעזרתי במורה פרטי שביצע את הקורס בעבר. לצערי אין לי ניסיון ב-*numpy* ולכן נעזרתי בשניהם לא מעט, בעיקר בהדפסת הגרפים. אני כן מבין את החומר ומתעניין בו, וכן לאחר מכן מבין את מה שהוא ביצע, ואני משתדל לפרט המון בדו״ח.