336224076 דוד בוסביב

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.,

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

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

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

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 $1 - \varepsilon \le e^{-\varepsilon}$
Hint 2: The main idea is similar to section 1.2 in recitation 6

1. כדי להראות ש H_r היא למידה Pac נראה אלגוריתם למידה $A:\{(x,y)\}^m \to H_r$ ופונ' סיבוכיות דגימות $m_H:(0,1)^2 \to N$ כך שמתקיימות התכונות של למידה pac.

$$h_{r'}(r) = \begin{cases} 1 & |r|_2 \le |r'|_2 \\ -1 & |r|_2 > |r'|_2 \end{cases}$$

נגדיר את $\hat{r} = \max_{\forall j,i\,:\, y_i = 1} |r|_2$

המספר המקסימלי עבורו אנחנו טועים

תהי D התפלגות כלשהי על דגימות, נראה שיש מספר m עבורו

$$P_{S\sim D^m}[L_D(h_S) \le \varepsilon) \ge 1-\delta$$

נסתכל על האפשרויות שלנו לטעות. הטעות הם אם בחרנו $r\in R^2$ עבורו $\hat{r} < |r|$

$$L_D(h_S) = P(|r| \in [|\hat{r}|, r'])$$

ולכן -

$$\Rightarrow P_{S\sim D^m}(L_D(h_S) \le \varepsilon) \underset{\text{מאורע משלים.}}{=} 1 - P_{S\sim D^m}(L_D(h_S) > \varepsilon)$$

לפי התרגול נוכל להסיק ישירות ש- $P_{S\sim D^m}(L_D(h_S) > \varepsilon) < \exp(-\varepsilon m)$

כלמ צריך ש $1-\exp(-\varepsilon m)\geq 1-\delta$ ולכן צריך ש-

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

מכאן ש $\mathcal{H}_r$ למידה פאק עם סיבוכיות דגימה

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

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

$$x_1 \wedge x_2 \wedge \bar{x}_3$$

$$VCdim(\mathcal{H}) = \max\{|C| \mid \mathcal{H} \text{ shatters } C\}$$

1. מהגדרה של VCD נסיק כי

נסמן $VCD(H_1) = C_1$

$VCD(H_2) = C_2$

נניח בשלילה כי הטענה אינה נכונה לכן מתקיים $C_1 > C_2$

כלומר H_1 מנתצת ע"י קבוצה S_{C_1} שהיא גדולה יותר מהקבוצה שמנתצת את H_2 אותה נסמן S_{C_2}.

אך לפי הגדרה $H_1 \subseteq H_2$ כלומר לכל $\forall h \in H_1 \Leftarrow h \in H_2$ בפרט מתקיים שאם H_1 מנתצת ע"י C_1 גם H_2 יכולה להיות מנותצת עם קבוצה בגודל C_1 בסתירה!

ולכן $VCD(H_1) \leq VCD(H_2)$

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

3. כאשר

נראה ש $VC\text{-}dim(H)=d$.

B ליטרל ניתן לרשום אותו כ- x_i וגם בתור $\bar{x}_i = x_i - 1$

נראה שניתן לנפץ את הקבוצה $C=\{e_1,\dots,e_d\}$ כאשר e_i מוגדר להיות הבסיס הסטנדרטי של d, כלומר עבור e_i הליטרל x_i דלוק.

נראה שעבור לבנות conj שיספיק עבור ה - $2d$ ליטרלים

עבור $y\in\{0,1\}^d$ נגדיר $h_y(x)=\bigwedge_{i\in S} x_i \wedge \bigwedge_{i\notin S^*} \bar{x}_i$

כלומר אם $i\in S$ אזי $x_i=1$ ואם $i\notin S$ אזי $x_i=0$

אפשר לראות שעם הפונק' h מנפצים את הקבוצה C בגודל d.

לכן $VC\text{-}dim(H)\geq d$

נראה שלא קיים C בגודל $d+1$ שניתן לנפץ.

נניח בשלילה שכן אזי קיימת קבוצה $C_{d+1}=\{x_1,\dots,x_{d+1}\}$ כאשר $x_i\in\{0,1\}^d$, מכיוון שיש לנו $d+1$ וקטורים ב R^d נסיק שיש תלות לינארית

לפי המשלבת הלינארית נסמן את x_{d+1} באופן הבא:

$x_{d+1} = -\frac{a_1}{a_{d+1}}x_1 \dots -\frac{a_d}{a_{d+1}}x_d$ ← נובע מהגדרה של תלות לינארית.

כעת נראה שלא ניתן לנפץ את הקבוצה כאשר מתקיים

$$y_i = \begin{cases} 0 & i = d+1 \\ 1 & \text{else } i\in\{0,d\} \end{cases}$$

ידוע כי $y_{d+1}=0$, אזי יש לפחות מקום אחד שנעדה

יהיה כלומר נניח 1 שצריך לשים אפס או

נשים 0 שצריך לשים 1. כפי שאמרנו יש עוד וקטור אחד לבחון שבאותה קורדינטה שמנו אותה השמה (אחרת סתירה למינימליות). נשים לב שעבור אותו הווקטור (נסמן x_j) $h(x_j) = 1$ שכן כמובן סתירה!, הרי אם $h(x_j)=0$ בגלל אותו אינדקס משותף אנחנו גם צריכים להחזיר 0 עבור $h(x_j)$ שכן סתירה! (הרי באותו מקום אנחנו מחזירים הפוך ממה שצריך להחזיר.)

ולכן אין קבוצה בגודל $d+1$ שמנותצת סה״כ

$$\boxed{VC\text{-}dim(H) = d}$$

* S הקבוצה המנ״נת איך ה j מוכל שומר 1 או אפס אם נרצה אבל x_j או את $\bar{x}_i$ בהתאמה.

כמה אני רחוק מהכי טוב.

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 $m_{\mathcal{H}}(\varepsilon,\delta_1) \ge m_{\mathcal{H}}(\varepsilon,\delta_2)$.

יהי $\delta, \epsilon \in (0,1)$

אני יודעים לפי הנתונים ש- $\epsilon_1 \le \epsilon_2$ לכן גם מתקיים

$$\forall m > m_H(\epsilon_1,\delta) \qquad P_{S\sim D^m}[L_D(h_S) \le \epsilon_1] \ge 1-\delta$$

$$\forall m > m_H(\epsilon_2,\delta) \qquad P_{S\sim D^m}[L_D(h_S) \le \epsilon_2] \ge 1-\delta$$

• $\epsilon_1 \le \epsilon_2$, לכן נקבל

$$P_{S\sim D^m}[L_D(h_S) \le \epsilon_2] \ge P_{S\sim D^m}[L_D(h_S) \le \epsilon_1]$$

⇐ לכן לכל $m \ge m_H(\epsilon_1,\delta)$ מתקיים

$$P_{S\sim D^m}[L_D(h_S) \le \epsilon_2] \ge P_{S\sim D^m}[L_D(h_S) \le \epsilon_1] \ge 1-\delta$$

ולכן נוכל להסיק ש $m_H(\epsilon_1,\delta) \overset{*}{\ge} m_H(\epsilon_2,\delta)$ כנדרש.

(שכן $\epsilon_2 \ge \epsilon_1$ אז נצטרך פחות דגימות כדי להגיע לתנאי הנדרש לעומת כאן כדי לקיים את התנאי עבור ϵ_1 נצטרך להגדיל את מספר הדגימות כדי שהתנאי יתקבל).

2.

$$\forall m > m_H(\epsilon,\delta_1) \qquad P_{S\sim D^m}[L_D(h_S) \le \epsilon] \ge 1-\delta_1$$

$$\forall m > m_H(\epsilon,\delta_2) \qquad P_{S\sim D^m}[L_D(h_S) \le \epsilon] \ge 1-\delta_2$$

בנוסף ידוע ש $\delta_1 \le \delta_2$ ⇐ $1-\delta_1 \ge 1-\delta_2$

$$1-\delta_2 \le 1-\delta_1 \le P_{S\sim D^m}[L_D(h_S) \le \epsilon] \qquad \Leftarrow \forall m \ge m_H(\epsilon,\delta_1)$$

$$\Rightarrow m_H(\epsilon,\delta_1) \ge m_H(\epsilon,\delta_2)$$

ככל שיש הסתברות יותר גדולה נצטרך יותר מדגמים.

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.

H agnostic-pac $\Leftrightarrow$ VC-dim(H) סופי.

H הינה למידת agnostic-pac אם

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

הטענה אינה נכונה, דוג': H מחלקת ההיפותזות שמכילה את כל הפונקציות מ X ל $\{\pm 1\}$ לכל $B \subseteq X$ קיים $h \in H$ כך ש $\forall\, h(x_i) = y_i$ ולכן לכל קבוצה בגודל $m \in \mathbb{N}$ קיימת קבוצה בגודל m שמנותצת ע"י H ולכן $VC\text{-}dim(H) = \infty$ מכאן ש H היא לא agnostic pac

נראה שהתנאי בשאלה מתקיים:

בהינתן התפלגות D ו $m \geq m_H$ נגדיר את האלגוריתם הבא, A תחזיר h_D כאשר $h(x) = y$ $\forall x \subset H$ כאשר y מוגדר להיות

$y = \arg\max P_D((x,y) \mid x)$ $\Leftarrow$ ואז, לכל $\varepsilon, \delta \in (0,1)$ מתקיים

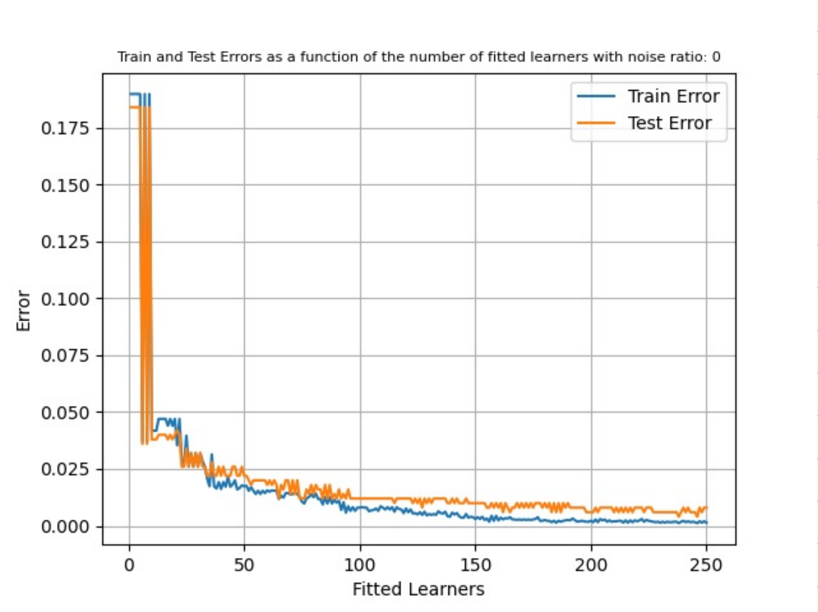
$$P_{S \sim D^m}\left(L_D(h) \leq \min_{h \in H}(L_D(h)) + \varepsilon\right) \geq 1-\delta$$

(שכן $L_D(H) \equiv \min(L_D(h))$ לכן התנאי יהיה כל הזמן נכון כמו שרצינו, שכן אנחנו ב agnostic אז נוכל ללמוד את הפונקציה בצורה מושלמת ((i.i.d)).

הראנו שהתנאי מתקיים ובנוסף $VC\text{-}dim(H) = \infty$ לכן H אינה agnostic pac

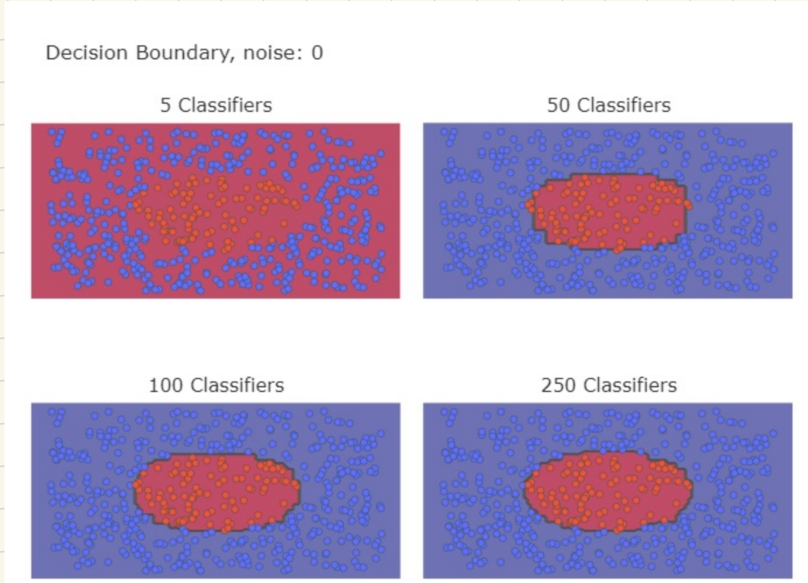
Practical Part

ג.

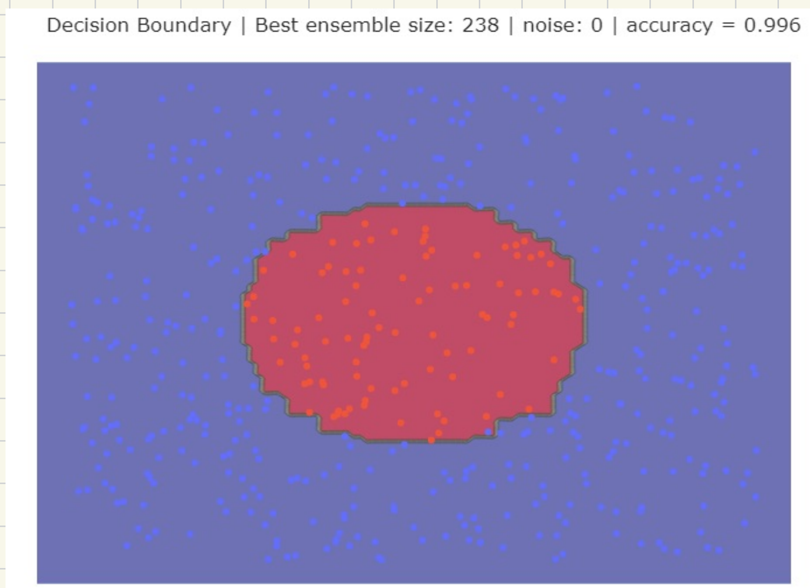


אפשר לראות בגרף את השגיאה ביחס למספר הלומדים כאשר אין
רעש (עבור desicion tree)
אפשר לראות שבסוף השגיאה הולכת לאפס, שכן אין
במודל רעש, וכמו שלמדנו בשיעור ככל שנעשה יותר קלאספריים
נגיע לשגיאה אפס.

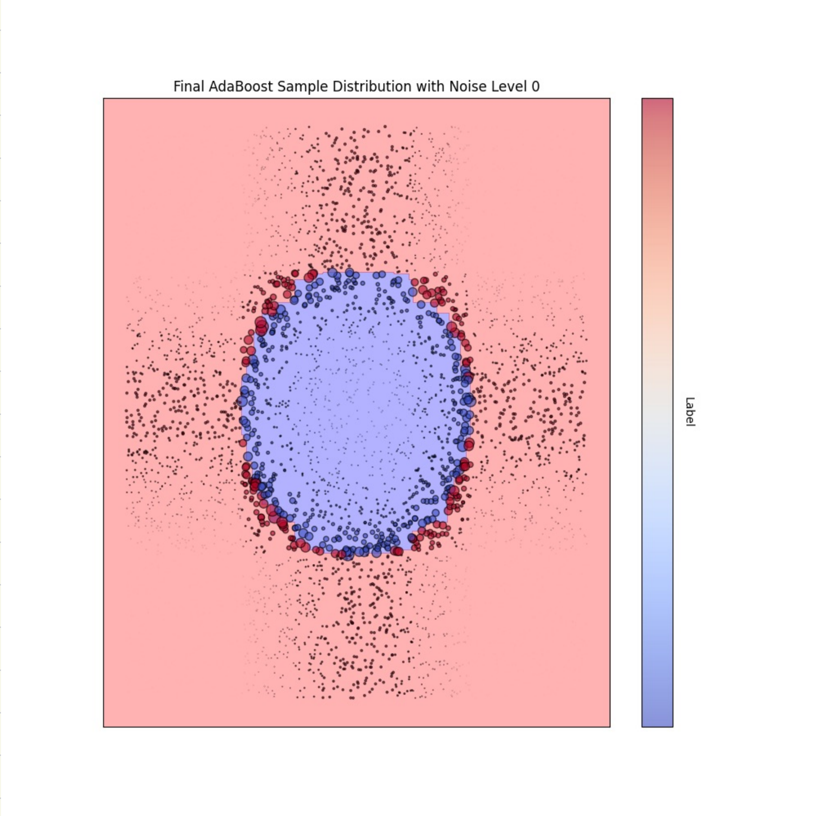
(2

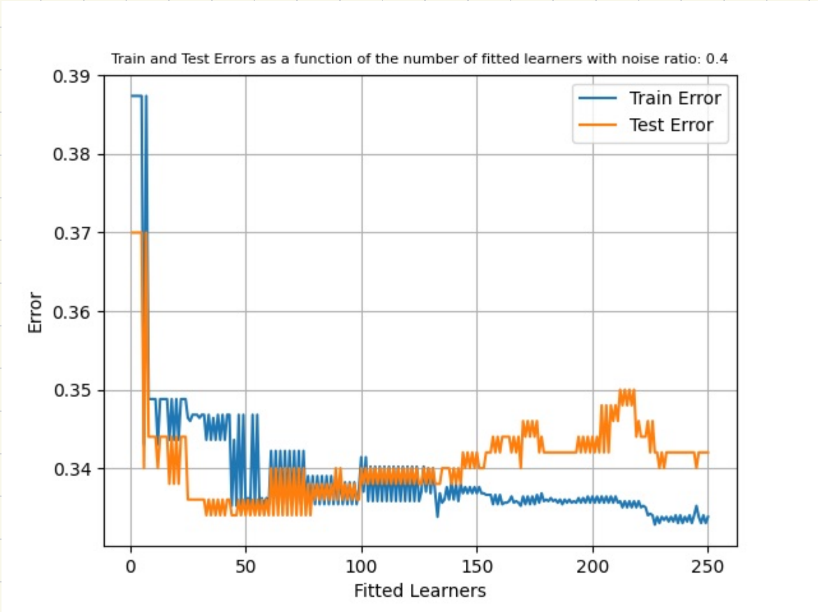


• אפשר לראות גם כאן ככל שיש יותר classifiers המדידות יוצאות יותר טוב.



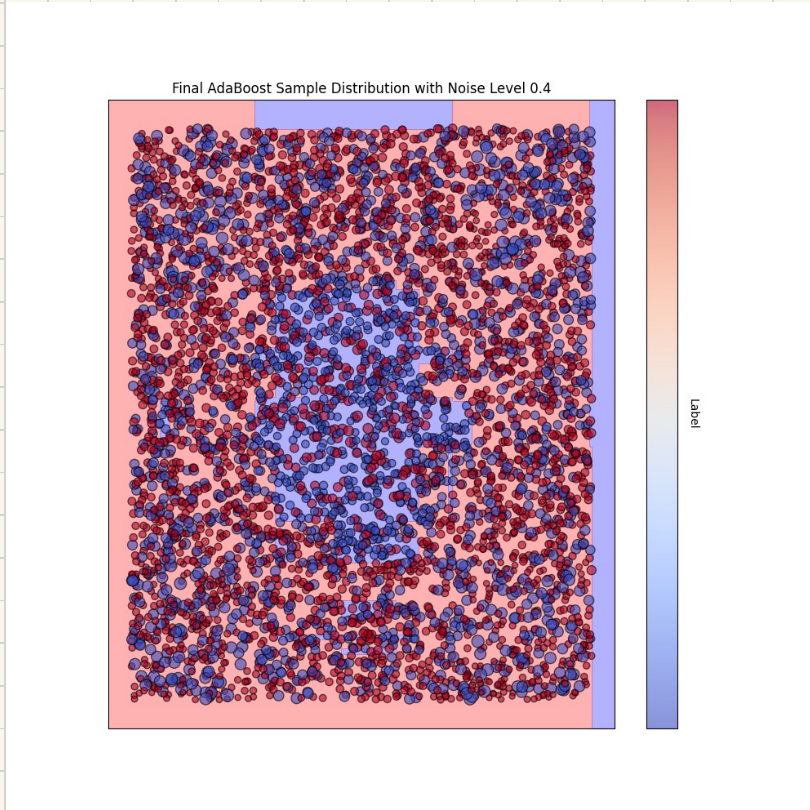
- עבור הגרף הזה אפשר לראות שהנק' עם המשקל הגבוה נמצאים בגבולות בין החלטה של האדום - לכחול זה מפני שקשה לקלסיפייר לסווג את הנקודות האלה ולכן האלגוריתם טועה בסיווג שלהם מספר פעמים, ולכן נשים להם משקל יותר גבוה.





- אפשר לראות שכאשר יש רעש השגיאה לא בהכרח קטנה ביחס לכמות הלומדים, שכן ככל שנשים יותר לומדים יהיה יותר סיכוי שתהיה שגיאה. לאחר בערך 120 איטרציות נקבל את השגיאה

הנמוכה ביותר, אך אחר כך השגיאה עולה, בגלל overfitting שאנחנו
עושים.



• כאשר יש רעש לרוב הנקודות
אותו משקל ולכן קשה לקלסיפייר
לסווג את הנקודות ולכן
הדיוק שנקבל גרוע כזה.
שלכולם בערך אותו משקל.