Theoretical Part

PAC learnability

1. כדי להראות ש- $\mathcal{H}$ היא למידת PAC, נראה שעבור אלגוריתם לומד A ופונקציית $m_{\mathcal{H}}$ התנאי עומד.

נגדיר אלגוריתם לומד באופן הבא: בהינתן דגימה $\{(x_i, y_i)\}_{i=1}^{m}$ האלגוריתם יחזיר את העיגול $\hat{r}$ שנותן את המדד הכי קרוב לדגימות החיוביות. כלומר A יחזיר $h_{\hat{r}}$ כך ש- $\hat{r} = \max_{i: y_i = 1} \|x_i\|_2$ כך שכל דגימה X בתוך המעגל, הלייבל שלה יהיה 1. ואחרת, 0.
כעת, נראה שעבור כל מעגל קונצנטרי r ולכל התפלגות D על X ולכל $\varepsilon, \delta \in (0,1)$ התנאי PAC מתקיים.
נשים לב שעבור המעגל הקטן ביותר מוכל בתוך המעגל האמיתי.
כלומר, $\hat{r} \le r$. כלומר שהטעות של האלגוריתם יכולה לנבוע רק מנקודות שנופלות בקטע $r \setminus \hat{r}$ לכן

$$L_D(h_s) = P[x \in r \setminus \hat{r}] = P[\hat{r} \le \|x\|_2 \le r]$$

נסמן ב- L את הקבוצה הבאה: $L = \{x \in \mathbb{R}^2 : \hat{r} \le \|x\|_2 \le r\}$.
נסמן ב- r^* מה״מ עבור כל המקרים של $\hat{r}$ בהם $P_{x \sim D}[x \in L] > \varepsilon$
אז נקבל:

$$P_{S \sim D^m}[L_D(h_S) > \varepsilon] = P_{S \sim D^m}\left[\bigwedge_{i=1}^{m} x_i \notin L^*\right] = \prod_{i=1}^{m} P_{x_i \sim D}[x_i \notin L^*]$$
$$\le (1-\varepsilon)^m < e^{-\varepsilon m}$$

כאשר $L^* = \{x \in \mathbb{R}^2 : r^* \leq \|x\|_2 \leq r\}$ לכן

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

$\Leftarrow$

$$1-e^{-\varepsilon m} < 1-\delta \iff -e^{-\varepsilon m} > -\delta \iff e^{-\varepsilon m} \leq \delta$$
$$\iff -\varepsilon m \leq \log(\delta) \iff m \geq \frac{\log(\frac{1}{\delta})}{\varepsilon}$$

2.2 VC - dimension

2. נניח בשלילה שמתקיים: $VC\text{-}dim(H_1) > VC\text{-}dim(H_2)$

נסמן: $VC\text{-}dim(H_2) = d_2$, $VC\text{-}dim(H_1) = d_1$. אז מהנחה $d_1 > d_2$. כעת, מהנתון $H_1 \subseteq H_2$. לכן, כל $h \in H_1$ מקיימת $h \in H_2$. ולכן, אם d_1 היא קבוצה של נקודות שמנותצת ע"י H_1 אזי היא מנותצת ע"י H_2. כלומר מתקיים $d_1 = VC\text{-}dim(H_2)$ בסתירה להנחה ש- $VC\text{-}dim(H_2) = d_2$. כלומר, H_2 לא יכולה לנתץ קבוצה שגדולה יותר מ- d_2.

3. נוכיח ש- $d = VC\text{-}dim(H_{con})$ עבור $d \in \mathbb{N}$.

• נראה ש- $d \leq VC\text{-}dim(H_{con})$:

נגדיר את הקבוצה d כך שעבור הליטרל ה- x_i המשתנה ה-i שווה 1-ל והשאר שווים ל-0 (לדוגמה עבור $d=4$ נקבל אפשרויות $(1,0,0,0)$, $(0,1,0,0)$, $(0,0,1,0)$, $(0,0,0,1)$).

אז לכל השמה של לייבל נוכל ליצור ביטוי בוליאני לכל אחד מהמשתנים.
אם $x_i = 1$ ניקח את הליטרל x_i. אחרת, ניקח את הליטרל $\bar{x}_i$.
באופן כזה, ניצור ביטוי בוליאני שמתייג f בדיוק את כל הנקודות
שהלייבל שלהם 1 בלבד. כלומר $vc\text{-}dim(H_{con}) \geq d$.

• כעת, נראה ש- $vc\text{-}dim(H_{con}) < d+1$:

נניח שקיימות $d+1$ נקודות מ- $\{0,1\}^d$ אזי משובך היונים, קיימות
2 נקודות שמקבלות את אותו ערך באותו משתנה בוליאני.
נסמן ב- $\hat{x}$ ו $\hat{y}$ כך ש- $\hat{x}_k = \hat{y}_k$ עבור $k \in [d]$.
נשים לב שעבור כל ביטוי בוליאני $x_i, \bar{x}_i$ לא יכולים להופיע ביחד
לכן, לא יהיה ניתן ליצור h כך שתבדיל בין $\hat{x}$ ל $\hat{y}$. כלומר, לא
נוכל לתייג את כל $d+1$ הנקודות ע"י H_{con}.
סה"כ, קיבלנו ש- $d = vc\text{-}dim(H_{con})$.

2.3 Sample complexity

4. • מהגדרה: עבור ε_1 מתקיים:

$$P[L_{D,f}(A(S_m)) < \varepsilon_1] \geq 1-\delta \iff m \geq m_H(\varepsilon_1, \delta)$$

ובאופן דומה, עבור ε_2 מתקיים

$$P[L_{D,f}(A(S_m)) < \varepsilon_2] \geq 1-\delta \iff m \geq m_H(\varepsilon_2, \delta)$$

עבור $m \in N$ אם $L_{D,f}(A(S_m)) < \varepsilon_1$ אזי $L_{D,f}(A(S_m)) < \varepsilon_2$
כיון ש- $m_H(\varepsilon_1, \delta) = m$ והוא מקיים את המשוואה הראשונה

נקבל שהוא גם מקיים את המשוואה השנייה ולכן נקבל:
$m = m_H(\varepsilon_1, \delta) \geq m_H(\varepsilon_2, \delta)$.

- מהגדרה, עבור δ_1 מתקיים:

$$P[L_{D,f}(A(S_m)) < \varepsilon] \geq 1-\delta_1$$

באופן דומה, עבור δ_2 מתקיים:

$$P[L_{D,f}(A(S_m)) < \varepsilon] \geq 1-\delta_2$$

עבור $m \in \mathbb{N}$ אם $P[L_{D,f}(A(S_m)) < \varepsilon] > 1-\delta_1 \Leftarrow$
$P[L_{D,f}(A(S_m)) < \varepsilon] > 1-\delta_2$ כיון ש- $m_H(\varepsilon,\delta_1) = m$ מקיים את המשוואה הראשונה, הוא מקיים גם את המשוואה השנייה
ולכן נקבל: $m = m_H(\varepsilon, \delta_1) \geq m_H(\varepsilon, \delta_2)$.

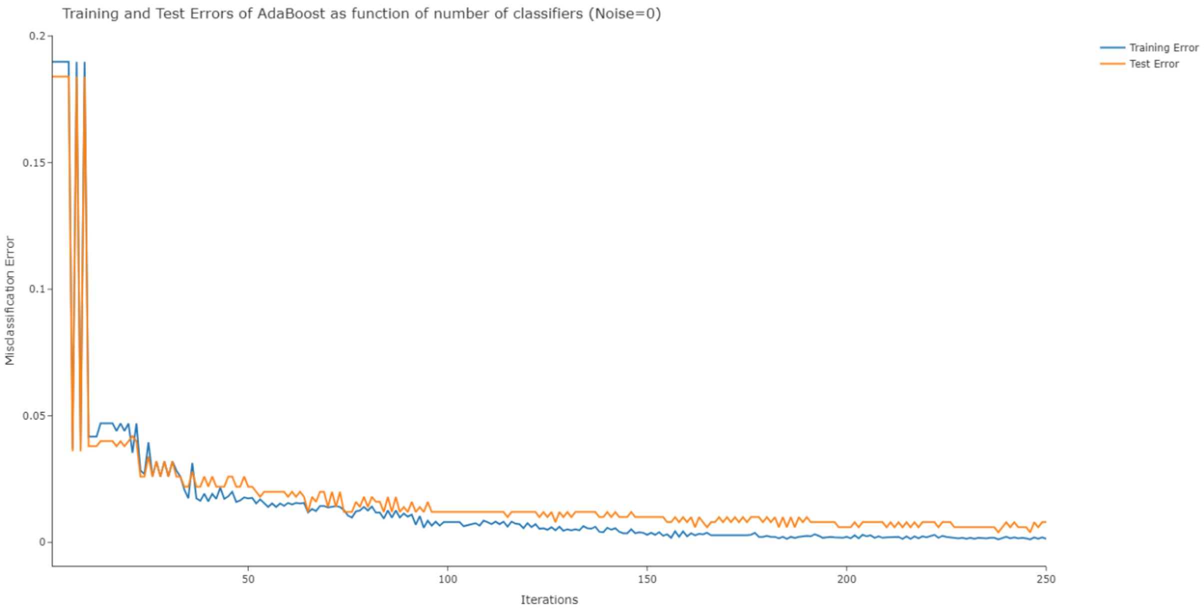
2.4 Agnostic PAC

5. H אינה למידת agnostic PAC.
נסתכל על קבוצת H הבאה: $H = \{0,1\}^n$ כך ש- $n \in \mathbb{N}$.
ראינו בהרצאה שמתקיים $VC\text{-}dim(H) = \infty$ ולכן מטענה שראינו בהרצאה H אינה Agnostic PAC. אבל נשים לב שהתנאי בשאלה מתקיים כי קיימת פונקציה m_H כך שלכל התפלגות D על Z קיים אלג' A כך שעבור $S \sim D^m$ תמיד יחזיר $h_S = \underset{h \in H}{\operatorname{argmin}}(L_D(h))$
ולכן נקבל: $L_D(h_S) = \min(L_D(h)) \leq \min(L_D(h)) + \varepsilon$
בהסתברות 1, כנדרש.

Practical part – IML

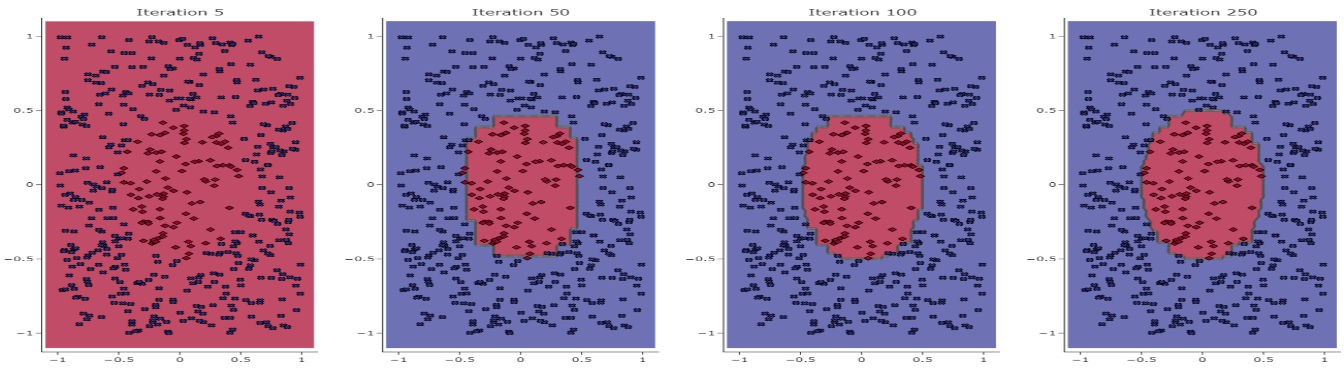
3.1 boosting

1. Explain your results:



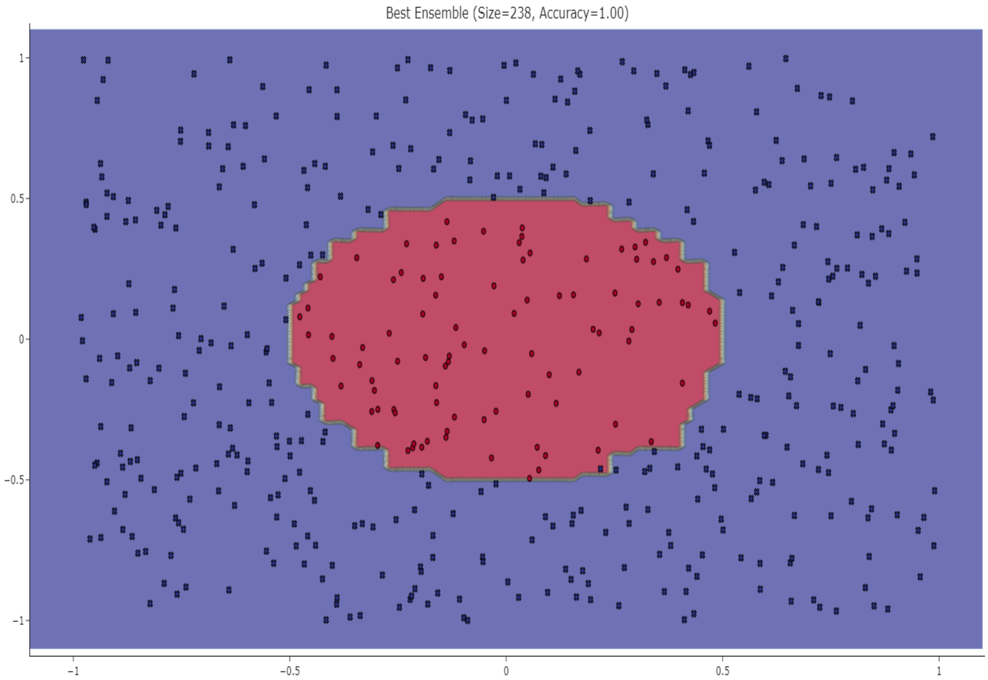
מהגרף נסיק עבור קבוצת האימון וקבוצת הטסט, שככל שקבוצת הלומדים גדלה, הטעות הולכת ושואפת לאפס. הטעות של קבוצת הטסט מעט גבוהה מהטעות של קבוצת האימון אבל עדיין גם היא הולכת ונעשית קטנה.

2. Explain your results:



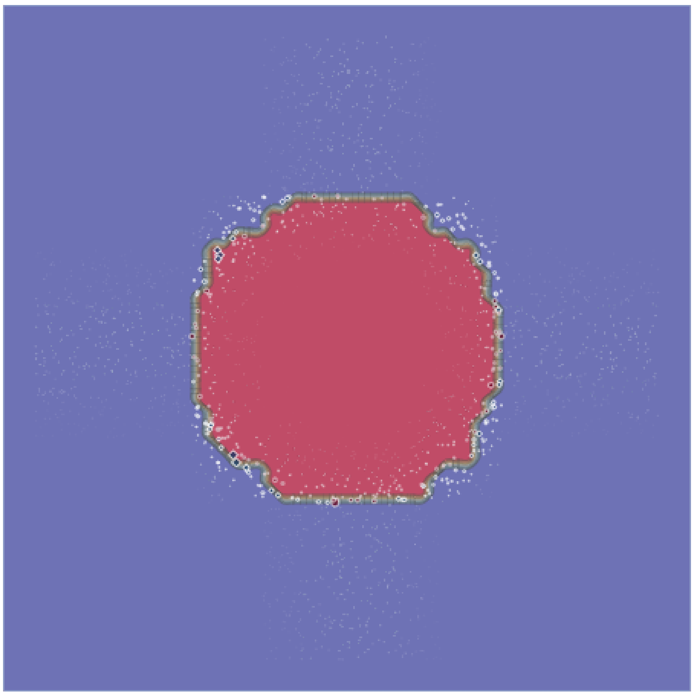
מהגרף ניתן לראות שככל שמספר הלומדים גדל, כך ההצלחה שלהם לסווג נכון את הנקודות עולה והטעות קטנה. בגרף הראשון של 5 לומדים, הם מצליחים לסווג יחסית טוב אבל ככל שמספר הלומדים גדל, כבר בגודל 50 ניתן לראות שיפור משמעותי וגודל 100 ו250 הטעות ממש קטנה ושואפת לאפס.

3. size=238, Accuracy=1.00



4. <u>Explain your results, and specifically explain which samples are "easier" and which are more "challenging" for the classifier:</u>

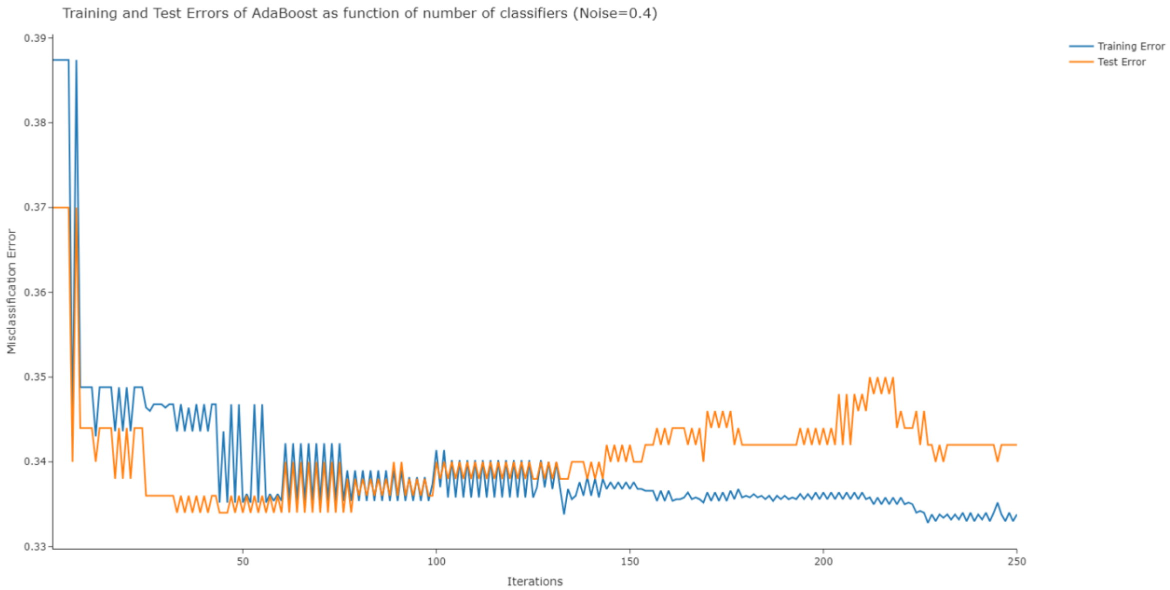
Final AdaBoost Sample Distribution

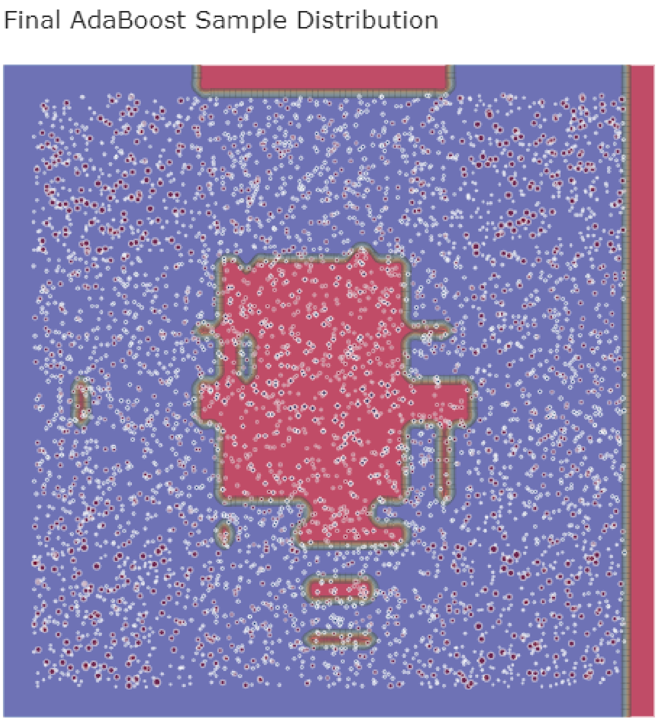


מהגרף ניתן להבחין איזה נקודות היו קלות ללומדים לסווג ואיזה נקודות היו קשות ללומדים לסווג.

נקודות קטנות בתוך העיגול האדום ומחוצה לו על הרקע הכחול היו קלות לסווג, זאת ניתן לראות ע"י גודל הנקודות. משקלן הוא קטן ולכן הן קטנות. הדגימות שהיו קשות יותר לסווגן היו הנקודות שנמצאות על גבול העיגול ושצבען כחול והן בצורת יהלום.

5. <u>Show graphs as in question (1) and question (4). Explain the results. In your answer explain what is seen in the plot of the loss in terms of the bias-variance tradeoff:</u>





מהגרפים ניתן להבחין בהשפעתו של הרעש של היכולת של הלומדים לסווג את הדגימות.

בגרף (1) הטעות של שני הסטים הולכת ונעשית קטנה ככל שמספר הלומדים גדל אך מדי פעם יש "קפיצות" בגרפים נוצרים מרעש בדגימות. בעוד סט האימון מאיטרציה 50 הולך ושואף לאפס, סט הטסט קופץ מעט, בהתאם לרעש בדגימות. אבל מאיטרציה 250 מתייצב גם הוא אך לא ליד האפס אלא גבוה יותר, בהשפעת הרעש.

באותו אופן, בגרף (4) ניתן להבחין שהרעש מקשה על הלומדים לסווג נכון את הנקודות וניתן לראות זאת על זה שיש חפיפה בין נקודות ממחלקה אחת על מחלקת אחרת ולהיפך.