יובל דלוס 305211880

מבוא למערכות לומדות – תרגיל 4

PAC Learnability

1. Let A be a learning algorithm, $\mathcal{D}$ be any distribution, and our loss function is in the range $[0,1]$ (e.g., the 0-1 loss). Prove that the following two statements are equivalent:

   (a) For every $\epsilon, \delta > 0$, there exists $m(\epsilon, \delta)$ such that $\forall m \geqslant m(\epsilon, \delta)$:

$$\mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) \leqslant \epsilon] \geqslant 1-\delta$$

   (b)

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

1. ראשית נראה כי $(b) \to (a)$:
נניח כי b נכון, נסדר מחדש את הביטוי ב-a ונפעיל עליו א״ש מרקוב:

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\big(A(S)\big) \le \varepsilon\right] \ge 1-\delta \Longrightarrow$$
$$1-\mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\big(A(S)\big) > \varepsilon\right] \ge 1-(1-\delta) \Longrightarrow$$
$$1-\mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\big(A(S)\big) > \varepsilon\right] \ge \delta \Longrightarrow$$
$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\big(A(S)\big) > \varepsilon\right] \le 1-\delta$$

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\big(A(S)\big) > \varepsilon\right] \overset{\text{Markov's inequality}}{\lneqq} \frac{\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\big(A(S)\big)\right]}{\varepsilon}$$

ומכיוון שהנחנו את (b) מתקיים:

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\big(A(S)\big) > \varepsilon\right] = 0$$

מכיוון שההסתברות יכולה להיות בין 0 ל-1 בלבד.
נהפוך את אי השוויון בתוך ההסתברות ונקבל:

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\big(A(S)\big) \ge \varepsilon\right] = 1$$

כעת אנו יודעים כי $0 < \delta < 0$ ולכן $1-\delta \le 1$ ונקבל:

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\big(A(S)\big) \ge \varepsilon\right] \ge 1-\delta$$

כנדרש.

נראה כעת את הגרירה $(a) \to (b)$:

מהנחת (a)אנו יודעים כי לכל $\varepsilon, \delta \in (0,1)$ קיים $m(\varepsilon, \delta)$ כך שלכל $m \ge m(\varepsilon, \delta)$ מתקיים $\underset{S\sim\mathcal{D}^m}{\mathbb{P}}\left[L_{\mathcal{D}}\big(A(S)\big) \ge \varepsilon\right] \ge 1-\delta$

ומכך שיכלנו לבחור כל $\varepsilon, \delta \in (0,1)$ נוכל לבחור את $\delta = \varepsilon = \frac{1}{m}$ ואז מתקיים:

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\big(A(S)\big)\right] = \int_{-\infty}^{\infty} x \cdot \mathbb{P}\big(L_{\mathcal{D}}\big(A(S)\big) = x\big)\, dx$$

כעת נתחשב בעובדה כי $L_{\mathcal{D}}\big(A(S)\big)$ הוא מ״מ חסום בין 0 ל-1 ונתחם את האינטגרל בהתאם:

$$= \int_0^1 x \cdot \mathbb{P}\big(L_{\mathcal{D}}\big(A(S)\big) = x\big)\, dx = \int_0^{\varepsilon} x \cdot \mathbb{P}\big(L_{\mathcal{D}}\big(A(S)\big) = x\big)\, dx + \int_{\varepsilon}^1 x \cdot \mathbb{P}\big(L_{\mathcal{D}}\big(A(S)\big) = x\big)\, dx \le$$

$$\le \int_0^{\varepsilon} x \cdot \mathbb{P}\big(L_{\mathcal{D}}\big(A(S)\big) \le \varepsilon\big)\, dx + \int_{\varepsilon}^1 x \cdot \mathbb{P}\big(L_{\mathcal{D}}\big(A(S)\big) > \varepsilon\big)\, dx =$$

$$= \int_0^{\varepsilon} x \cdot \mathbb{P}\big(L_{\mathcal{D}}\big(A(S)\big) \le \varepsilon\big)\, dx + \int_{\varepsilon}^1 x \cdot \Big(1 - \mathbb{P}\big(L_{\mathcal{D}}\big(A(S)\big) \le \varepsilon\big)\Big)\, dx$$

נציב את ההנחה ב- (a) ונקבל:

$$\le \int_0^{\varepsilon} x \cdot (1-\delta)\, dx + \int_{\varepsilon}^1 x \cdot \big(1-(1-\delta)\big)\, dx \le \varepsilon + \delta - \varepsilon\delta \le \varepsilon + \delta$$

כעת נסתכל על גבול התוחלת כאשר $m \to \infty$:

$$\lim_{m\to\infty} \mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\big(A(S)\big)\right] \le \lim_{m\to\infty} (\varepsilon + \delta) = \lim_{m\to\infty} \left(\frac{1}{m} + \frac{1}{m}\right) = 0$$

ניזכר בעובדה כי מדור על תוחלת של מ״מ אי-שלילי ולכן תוחלת זו חסומה ע״י 0, ומכלל הסנדוויץ׳ נקבל:

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

כנדרש.

305211880 יובל דלוס

2. **Sample Complexity of Concentric Circles in the Plane** 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., $\mathcal{H} = \{h_r : r \in \mathbb{R}_+\}$, where $h_r(x) = \mathbf{1}[\|x\|_2 \leqslant r]$. Prove that $\mathcal{H}$ is PAC learnable and its sample complexity is bounded by

$$m_{\mathcal{H}}(\epsilon, \delta) \leqslant \frac{\log(1/\delta)}{\epsilon} \ .$$

2. ראשית נשים לב בשאלה כי מחלקת ההיפותזות היא כל המעגלים הקנוניים סביב הראשית, וככזאת יש מעגל סביב הראשית שתופס בתוכו את כל הנקודות החיוביות, ונראה שנוכל תמיד למצוא מעגל כזה עבור m דגימות כך שהוא יטעה לכל היותר ב-ε בסבירות של $1-\delta$ לכל הפחות.

נניח שרדיוס המעגל האמיתי (ההיפותזה הנכונה) הוא r_0, כעת, בקבלינו m דגימות נשרטט רדיוס חדש r_h שיכיל בתוכו את כל הדגימות החיוביות מתוך m הדגימות שלנו, כלומר הנקודה "הרחוקה" ביותר מהראשית תתווה את היקף המעגל הפנימי (בעל רדיוס r_h) שהוא ההיפותזה אותה אנו מציעים. נראה איזה מספר $m_{\mathcal{H}}(\varepsilon, \delta)$ של דגימות נצטרך בשביל להבטיח שהמעגל שלנו יטעה לכל היותר ב-ε.

נבחר מעגל כזה שהרצועה בינו לבין המעגל האמיתי r_0 היא בשטח ε, נשים לב שכל הנקודות שבתוך המעגל הזה הן בהכרח חיוביות כי הן בתוך מעגל ההיפותזה r_0, וכל מה שמחוץ למעגל בהכרח מתויג כשלילי, לכן הטעויות שיכולות לקרות לנו הן אך ורק ברצועה בין שני המעגלים.

ההסתברות שנקודה תיפול מחוץ לרצועה היא $(1-\varepsilon)$, ומכיוון שכל m הנקודות ב"ת ההסתברות שכולן יפלו בתוך המעגל הפנימי היא $(1-\varepsilon)^m$.

נשתמש בזהות $(1-x) < e^{-x}$ ונקבל:

$$(1-\varepsilon)^m \leq \mathrm{e}^{-\varepsilon\mathrm{m}}$$

ומכאן:

$$\mathrm{e}^{-\varepsilon\mathrm{m}} < 1-\delta \Longrightarrow -\varepsilon m < \log(1-\delta) \Longrightarrow m > \frac{\log\left(\frac{1}{\delta}\right)}{\varepsilon} \Longrightarrow$$

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

VC dimension

3. **Boolean Conjunctions** Let $\mathcal{X} = \{0,1\}^d$ and $\mathcal{Y} = \{0,1\}$, and assume $d \geqslant 2$. 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 $\overline{x}_k = 1 - x_k$. The class $\mathcal{H}_{\text{con}}$ is defined by boolean conjunctions over any subset of these $2d$ literals. For example: let $d = 5$ and consider the hypothesis that labels $\mathbf{x}$ according to the following conjunction

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

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. Compute the VC dimension of $\mathcal{H}_{\text{con}}$ and prove your answer.

3. ראשית נראה ש- $VC\ dimensiton\ \ngeq d+1$

עבור היפותזה עם $d+1$ ליטרים בהכרח יהיה ליטרל x_i כלשהוא כך שבהיפותזה הוא יופיע פעמיים $x_1, \dots x_i, \dots \overline{x}_i, \dots x_{d+1}$

כעת בצורה כזאת, בהכרח לא משנה איזו השמה של הליטרלים נבחר תמיד נקבל שהתגית שלנו 0, מכיוון ש- $x_i \wedge \overline{x}_i = 0$ ובעצם לא נוכל לקבל אף פעם תגית 1 ולא הצלחנו להגיע לכל האפשרויות ולנפץ את הדוגמא.

נראה שכל סט דוגמאות מגודל d ומטה נוכל לנתץ.

נניח שיש לנו d דגימות, ונניח שכל הדוגמאות הן וקטור היחידה המתאים למיקום שלהן (כלומר הווקטור e_i יהיה הדגימה ה-i)

כעת נרצה להראות שנוכל להגיע לכל התגיות האפשריות $\{0,1\}^d$ נסדר אותן לפי סדר, לדוגמא פה כאשר $d=3$:

X			y							
0	0	1	0	1	0	0	1	1	0	1
0	1	0	0	0	1	0	1	0	1	1
1	0	0	0	0	0	1	0	1	1	1

כעת, נוכל להראות שאין אף תגית y שלא נוכל להגיע אליה עם היפותזה באמצעות וקטורי היחידה שלנו.

- נניח כי כל התגיות שוות ל-0, נוכל להגיע לזאת באמצעות בחירת ההיפותזה "נבחר תמיד 0"
- אם יש לנו תגית אחת בלבד שהיא "1", נגיד תגית i, נוכל לבחור בהיפותזה x_i, כל שאר עמודות יסתדרו גם כן כי בהן $x_i = 0$ וגם התגית בהן שווה 0
- נניח יש לנו שתי תגיות "1", נוכל לבחור (בדוגמא שלנו) את ההיפותזה $\overline{x}_1$ עבור y המסומן באדום
- באופן כללי אם יש לנו $k < d$ תגיות מסומנות ב-1, נניח בה"כ כי אלה תגיות $[1,k]$, נוכל תמיד לבחור את ההיפותזה
$$\mathcal{H} \ni h_k = \bigwedge_{i \notin [k]} \overline{x}_i$$
- ואם כל התגיות הן 1 נבחר בהיפותזה "תמיד 1"

ובעצם הראינו שנוכל להגיע לכל תוצאה $y = \{1,0\}^d$ עם d וקטורים וניפצנו כל קבוצה מגודל d.

עבור כל קבוצה C קטנה מגודל d, נוכל לנתץ אותה באותו אופן, כאשר נתעלם כליל מ- $\{x_j, \dots, x_d\}$ ונפעל על השאר באותה דרך.

יובל דלוס 305211880

4. Prove that if $\mathcal{H}$ has the uniform convergence property with function $m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$ then $\mathcal{H}$ is Agnostic-PAC learnable with sample complexity $m_{\mathcal{H}}(\epsilon, \delta) \leqslant m^{UC}(\epsilon/2, \delta)$.

4. נניח כי ל- $\mathcal{H}$ יש את התכונה של התכנסות במידה שווה כמתואר, אזי, לכל $\varepsilon, \delta \in (0,1)$, ו-S קבוצת דגימות בגודל m, כך ש- $m \geq m^{UC}\left(\frac{\varepsilon}{2}, \delta\right)$, מתקיים שבהסתברות שלכל הפחות $1-\delta$ S היא $\frac{\varepsilon}{2}$-מייצגת, כלומר, לכל $h \in \mathcal{H}$ מתקיים:

$$L_{\mathcal{D}}(h) \leq L_S(h) + \frac{\varepsilon}{2}$$

כלומר

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \frac{\varepsilon}{2} \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon$$

ולכן $\mathcal{H}$ היא agnostic PAC-learnable עם $m_{\mathcal{H}}(\varepsilon, \delta) = m^{UC}\left(\frac{\varepsilon}{2}, \delta\right)$.

7. **of VC-Dimension** 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 $\text{VC}\,\mathcal{H}_1 \leqslant \text{VC}\,\mathcal{H}_2$.

7. נסתכל על ההגדרה:

$$VCdim = \max\{\ |C| : \mathcal{H}\ shatters\ C\ \}$$

כלומר המימד VC נקבע ע״י גודל תת-הקבוצה המקסימלית $C = \{x_1, \dots, x_{|C|}\} \subseteq \mathcal{X}$ שמחלקת ההיפותזות מצליחה לנפץ.
כעת, בהנחה ש- $\mathcal{H}_1 \subseteq \mathcal{H}_2$ נניח בשלילה כי $VC\ \mathcal{H}_1 > VC\ \mathcal{H}_2$, אבל מכיוון ש- $\mathcal{H}_1 \subseteq \mathcal{H}_2$, ומחלקת ההיפותזות $\mathcal{H}_1$ מנתצת תת-קבוצה בגודל C שכזאת, על אותו מרחב דגימות גם $\mathcal{H}_2$ תצליח לנתץ קבוצה באותו הגודל, לפחות ע״י צמצום שלה למחלקה $\mathcal{H}_1$ בלבד, וזו בסתירה ל- $VC\ \mathcal{H}_1 > VC\ \mathcal{H}_2$.

יובל דלוס 305211880

Theoretical Claim

8. Let X be a sample space and $\mathcal{Y} = \{\pm 1\}$. Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class. For $C \subset \mathcal{X}$, recall the notation $\mathcal{H}_C$ for the restriction of $\mathcal{H}$ to the subset C. Define the function $\tau_m(\mathcal{H}) : \mathbb{N} \to \mathbb{N}$ corresponding to $\mathcal{H}$ to be

$$\tau_{\mathcal{H}}(m) := \max \left\{ |\mathcal{H}_C| \,\middle|\, C \subseteq \mathcal{X} \,,\, |C| = m \right\}.$$

(a) Explain, in your own words, the meaning of $\tau_{\mathcal{H}}$.

(a) טאו היא הגודל המקסימלי של מחלקת ההיפותזות המצומצמת לתת-קבוצה C בגודל m, שמצליחה לנפץ את אותה תת-קבוצה.
כלומר, נניח והצטמצמנו לקבוצה C כלשהי בגודל m, וישנן $\binom{m}{|X|}$ קבוצות כאלה, טאו מהווה את הגודל המקסימלי של מחלקת ההיפותזות שמצליחה לנפץ מי מהקבוצות האלה.

(b) Suppose that $VCdim(\mathcal{H}) = \infty$. Find an expression for the value of $\tau_{\mathcal{H}}(m)$ for $m \in \mathbb{N}$.

(b) כאשר יש לנו מחלקת היפותזות $\mathcal{H}$ בעלת $VC\ dim = \infty$ זה אומר שכל צמצום לקבוצה C בגודל m נקבל כי היא מכילה את כל האפשרויות לפונקציות $|C| \to y$, כלומר עבור כל m $|\mathcal{H}_C| = 2^m$, ומכיוון שטאו הוא הגודל המקסימלי של קבוצת היפותזות כזאת, נקבל כי $\forall m, \tau_{\mathcal{H}}(m) = 2^m$, ומכיוון ש- $VC\ dim = \infty$ נקבל כי $\tau_{\mathcal{H}}(m) \to \infty$.

(c) Now suppose that $VCdim(\mathcal{H}) = d$. Find an expression for the value of $\tau_{\mathcal{H}}(m)$ for $m \leqslant d$.

(c) באותו אופן טאו מסמן את הגודל המקסימלי של קבוצת ההיפותזות, ומכיוון שעבור כל $m \leq d$ מחלקת ההיפותזות מנפצת את קבוצה C נקבל כי $\tau_{\mathcal{H}}(m) = 2^m$.

(d) You will now prove the following important result: suppose that $VCdim(\mathcal{H}) = d$ and let $m > d$. Then

$$\tau_{\mathcal{H}}(m) \leqslant \left(\frac{em}{d}\right)^d ,$$

where e is the natural logarithm base. You'll do this in three steps:

i. Using induction, show that for any finite $C \subset \mathcal{X}$,

$$|\mathcal{H}_C| \leqslant \left| \left\{ B \subseteq C \,\middle|\, \mathcal{H} \text{ shatters } B \right\} \right| .$$

(d)

i. נרצה להשתמש באינדוקציה להוכיח טענה זו.

ראשית, עבור $m = 1$, $\mathcal{H}$ תמיד יוכל לנפץ כל קבוצה שנבחר, וגודל שני צדדי א״ש תמיד יהיה 1.

כעת, נניח עבור $k < m$ ונוכיח עבור סט דגימות בגודל m.

נקבע את מחלקת ההיפותזות שלנו $\mathcal{H}$ ואת תת הקבוצה של הדגימות $C = \{c_1, \dots, c_m\}$, ונגדיר תת קבוצה חדשה של דגימות ללא הדגימה הראשונה $C' = \{c_2, \dots, c_m\}$.

כעת נוכל לקבוע את הקבוצות הבאות של תוצאות אפשריות להיפותזות:

$$Y_0 = \{\, (y_2, \dots, y_m) : \ (0, y_2, \dots, y_m) \in \mathcal{H}_C \ \vee \ (1, y_2, \dots, y_m) \in \ \mathcal{H}_C \,\}$$

$$Y_1 = \{\, (y_2, \dots, y_m) : \ (0, y_2, \dots, y_m) \in \mathcal{H}_C \wedge \ (1, y_2, \dots, y_m) \in \ \mathcal{H}_C \,\}$$

במילים, קבוצה Y_0 אלה כל ההיפותזות האפשריות ל-$(y_2, \dots, y_m)$ שהיינו יכולים להתאים להן קורדינאטה ראשונה 0 או 1 ועדיין הייתה היפותזה כלשהי ב-$\mathcal{H}_C$ שהייתה מנבאת את התוצאה הזאת בכללותה.

קבוצה Y_1 אלה רק ההיפותזות האפשריות ל-$(y_2, \dots, y_m)$, כך שעבור $(y_2, \dots, y_m)$ מסוים גם הקורדינאטה הראשונה 1 וגם 0 מתקיימות והיו מנובאות ע״י היפותזה כלשהי ב-$\mathcal{H}_C$.

כלומר, נוכל להבין כי $|\mathcal{H}_C| = |Y_0| + |Y_1|$, וגם $Y_0 = \mathcal{H}_{C'}$. נפעיל את הנחת האינדוקציה שלנו על $\mathcal{H}$ ו- C' ונקבל

$$|Y_0| = |\mathcal{H}_{C'}| \leq |\{B \subseteq C' : \mathcal{H} \ shatters \ B\}| = |\{B \subseteq C : c_1 \notin B \ \wedge \ \mathcal{H} \ shatters \ B\}|$$

כעת נגדיר את מחלקת ההיפותזות $\mathcal{H}' \subseteq \mathcal{H}$:

$$\mathcal{H}' = \{h \in \mathcal{H} : \exists h' \in \mathcal{H} \ s.t \left(1 - h'(c_1), h'(c_2), \dots, h'(c_m)\right) = \left(h(c_1), h(c_2), \dots, h(c_m)\right)\}$$

כלומר, כל ההיפותזות שמסכימות על כל הקורדינאטות מלבד הראשונה, ניתן היה לכתוב גם כך: $\mathcal{H}' = \overline{Y_0 \cap Y_1}$.

כעת קל לראות כי כל היפותזה ב-$\mathcal{H}'$ שהייתה מנפצת תת קבוצה $B \subseteq C'$ הייתה מנפצת גם את תת הקבוצה $B \cup \{c_1\}$, וזאת מכיוון שההיפותזה לא משתמשת בקורדינאטה של c_1 לצורך הניתוץ, ולכן מה שיהיה רשום בא לא משנה.

כעת נפעיל את הנחת האינדוקציה שלנו על $\mathcal{H}'$ ו- C', ונתחשב בעובדה כי $Y_1 = \mathcal{H}'_{C'}$ ונקבל:

$$|Y_1| = |\mathcal{H}'_{C'}| \leq |\{B \subseteq C' : \mathcal{H}' shatters \ B\}| = \left|\{B \subseteq C' : \mathcal{H}' shatters \ B \cup \{c_1\}\}\right| =$$
$$= |\{B \subseteq C : c_1 \in B \ \wedge \ \mathcal{H}' \ shatters \ B\}| \leq |\{B \subseteq C : c_1 \in B \ \wedge \mathcal{H} \ shatters \ B\}|$$

סה״כ קיבלנו:

$$|\mathcal{H}_C| = |Y_0| + |Y_1| \leq |\{B \subseteq C : c_1 \notin B \ \wedge \ \mathcal{H} \ shatters \ B\}| + |\{B \subseteq C : c_1 \in B \ \wedge \mathcal{H} \ shatters \ B\}| =$$
$$= |\{B \subseteq C : \ \mathcal{H} \ shatters \ B\}|$$

כנדרש.

יובל דלוס 305211880

ii. Explain in your own words the meaning of this inequality.

ii. אי השוויון גורס כי מחלקת ההיפותזות תמיד תהיה קטנה ממספר תתי-הקבוצות B אותה היא מסוגלת לנפץ, כלומר, תמיד תהיה היפותזה שיכולה לנתץ מספר תתי קבוצות שונות, שזו תוצאה מעניינת כי היא מרמזת שאימון על תת קבוצה B מסוימת היה יכול לתת לנו לחזות במדויקת את התוצאות על תת קבוצה B' אחרת. ומנגד, ישנן תתי קבוצות ב-C שאינן מוסיפות לנו מידע חדש על גבי מידע שכבר קיבלנו מתת קבוצה אחרת.

iii. Show that, for any finite $C \subseteq \mathcal{X}$, we have

$$\left|\{B \subseteq C \mid \mathcal{H} \text{ shutters } B\}\right| \leqslant \sum_{k=0}^{d}\binom{m}{k}$$

iii. אי השווין הנ"ל מציג מצידו הימיני את מספר כל תתי הקבוצות האפשריות של C, בכל גודל אפשרי מ-0 ועד כל הקבוצה כולה, מספר זה מבוטא בבינום כפי שהוא רשום. לעומת זאת, בצד השמאלי של הביטוי יש לנו את מספר כל תתי הקבוצות ש-$\mathcal{H}$ מצליח לנתץ, בהנחה והוא מצליח לנתץ את כל תתי הקבוצות לא משנה מאיזה גודל נקבל שוויון, אחרת, אם יש תתי קבוצות שהוא לא מצליח לנפץ נקבל את אי השוויון.

iv. Use the following inequality (which you are not required to prove)

$$\sum_{k=0}^{d}\binom{m}{k} \leqslant \left(\frac{em}{d}\right)^{d}$$

to finish the proof that $\tau_{\mathcal{H}}(m) \leqslant \left(\frac{em}{d}\right)^{d}$.

iv. לפי ההגדרה של טאו:

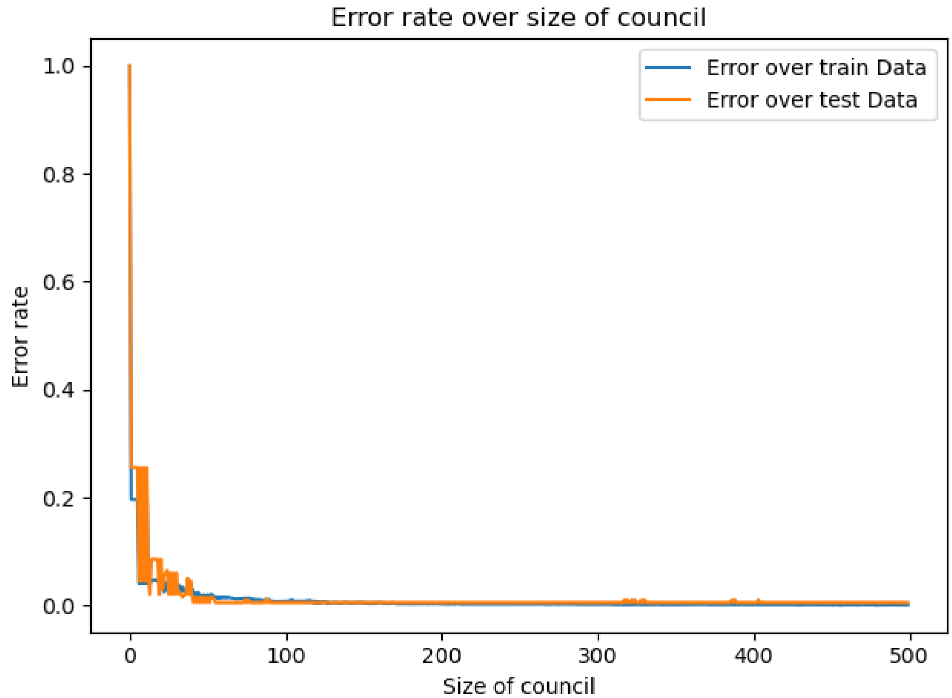
$$\tau_{\mathcal{H}}(m) = \max\{\,|\mathcal{H}_C| : C \subseteq \mathcal{X}, |C| = m\,\} \leq |\{B \subseteq C :\ \mathcal{H}\ shatters\ B\}| \leq \sum_{k=0}^{d}\binom{m}{k} \leq \left(\frac{em}{d}\right)^{d}$$

(e) If $m = d$, does the inequality $\tau_{\mathcal{H}}(m) \leqslant \left(\frac{em}{d}\right)^{d}$ hold? If it does hold, is it tight?

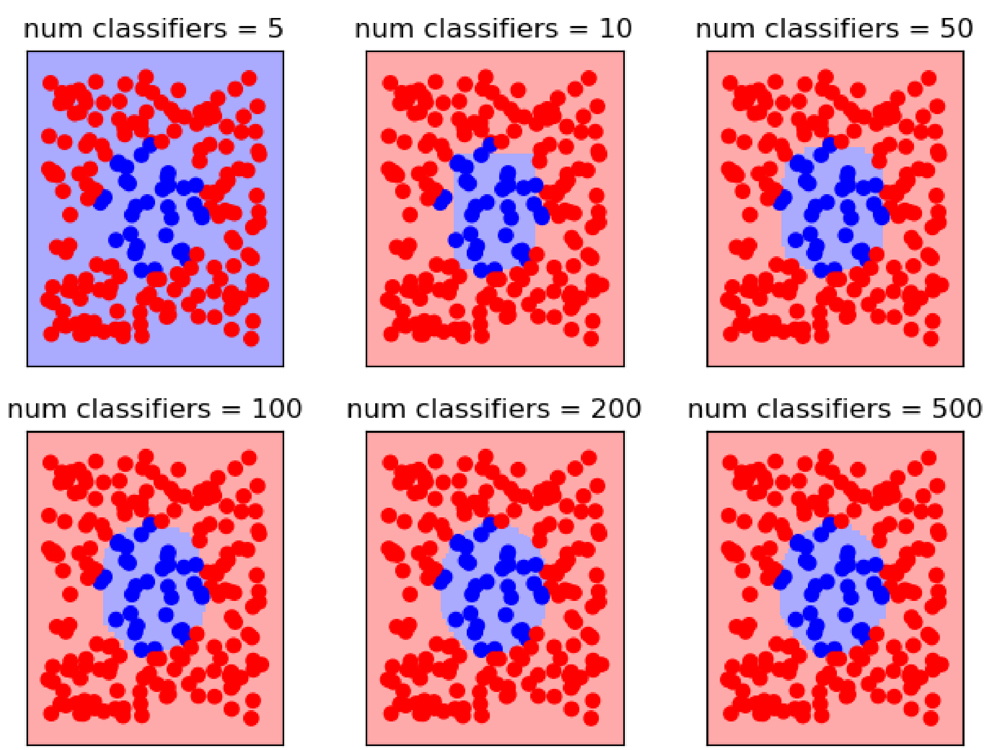
(e) נניח כי $a \coloneqq \frac{m}{d}$, אזי הביטוי באי שוויון ניתן לכתיבה כ- $\tau_{\mathcal{H}}(m) \leq e^d a^d$, כעת, ניתן להבין שכאשר $a < 1$ ככל ש-d יהיה יותר גדול ככה החסם יהיה יותר הדוק, אך כאשר $a = 1$, כלומר כמו בשאלה $m = d$ נקבל שא"ש הוא $\tau_{\mathcal{H}}(m) \leq e^d$, זה אמנם עדיין חסם לטאו, אך הוא פחות הדוק.

יובל דלוס 305211880

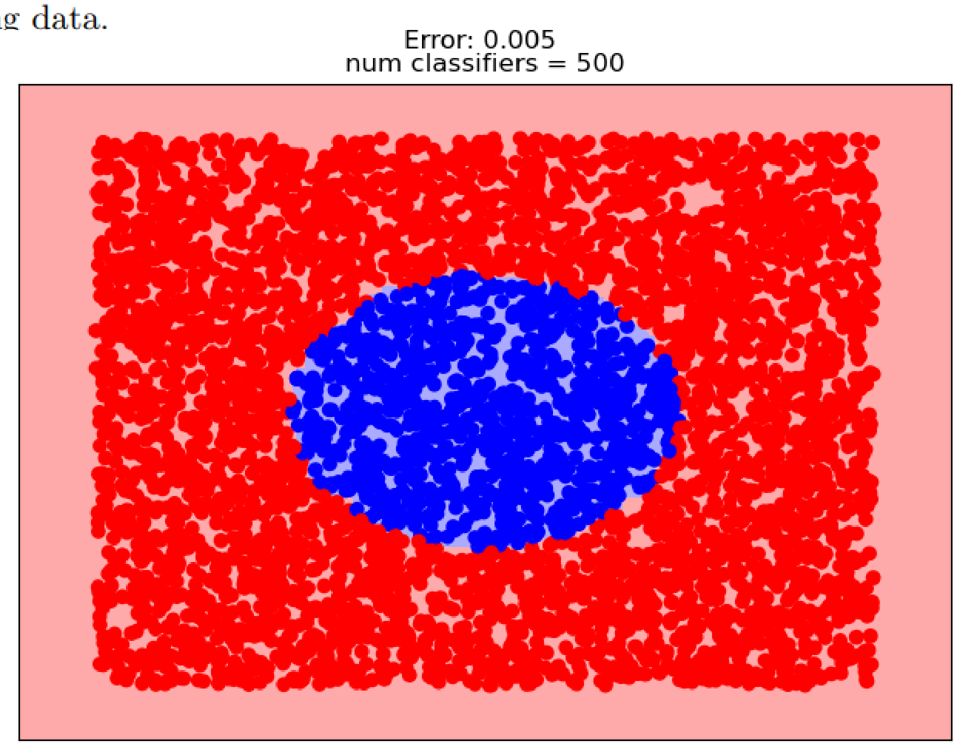
10. In ex4_tools you are provided with the function generate_data. Use it to generate 5000 samples without noise (i.e. noise_ratio=0). Train an Adaboost classifier over this data. Use the DecisionStump weak learner mentioned above, and $T = 500$. Generate another 200 samples without noise ("test set") and plot the training error and test error, as a function of T. Plot the two curves on the same figure.



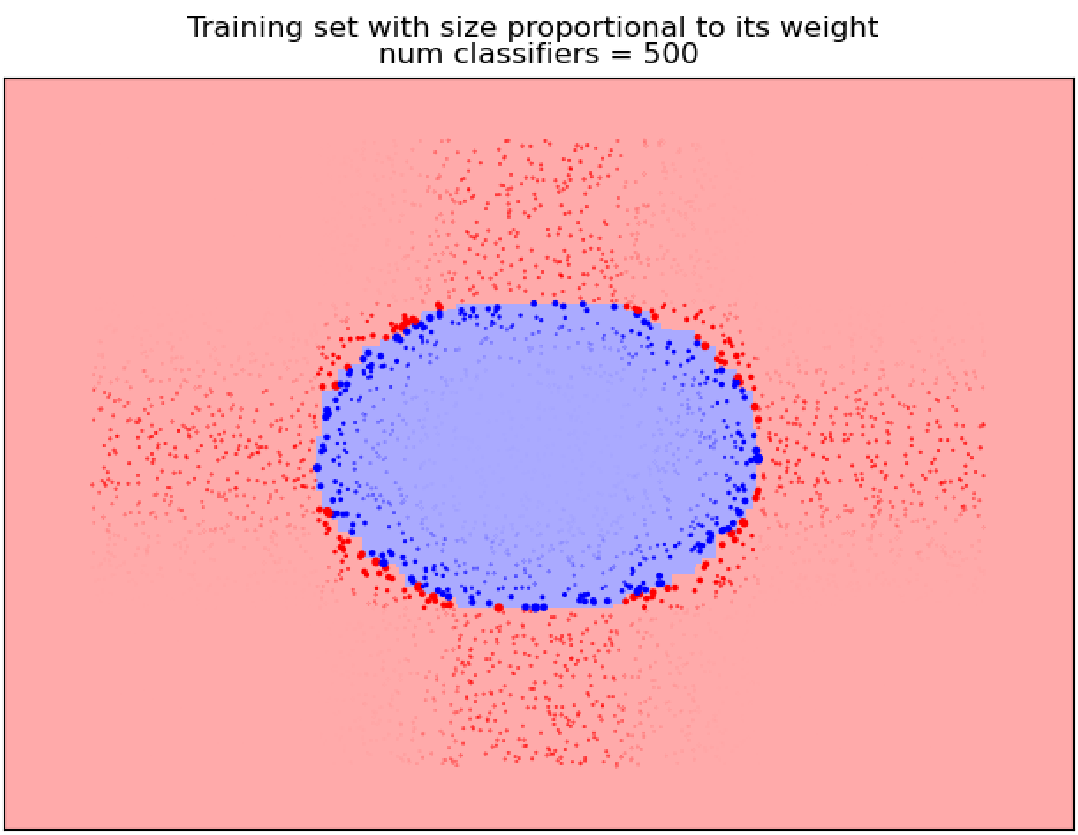
11. Plot the decisions of the learned classifiers with $T \in \{5, 10, 50, 100, 200, 500\}$ together with the test data. You can use the function decision_boundaries together with plt.subplot for this purpose.



12. Out of the different values you used for T, find $\hat{T}$, the one that minimizes the test error. What is $\hat{T}$ and what is its test error? Plot the decision boundaries of this classifier together with the training data.



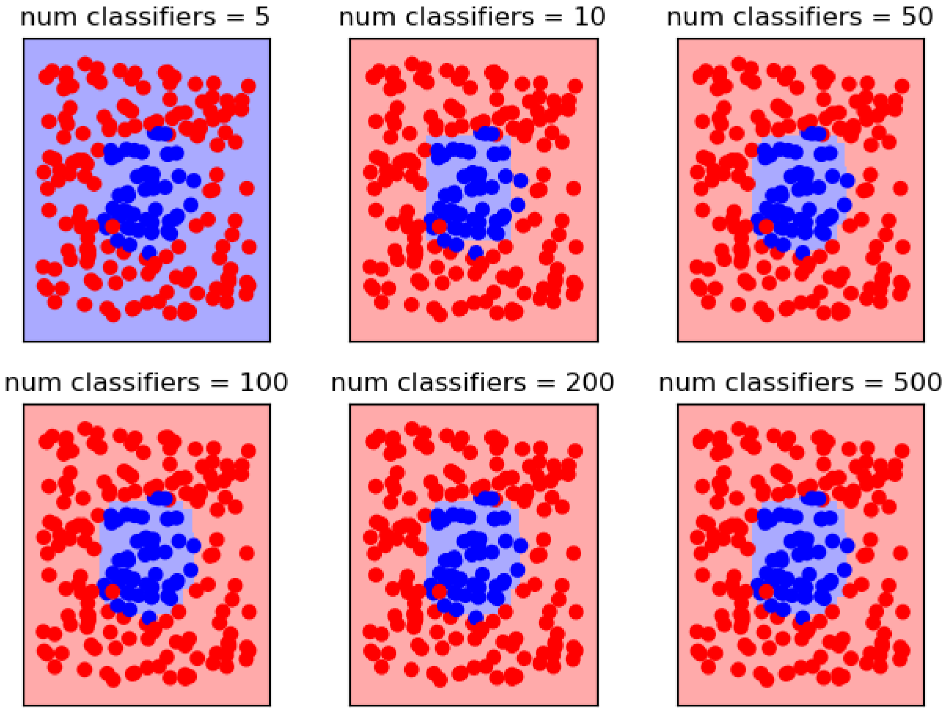
13. Look into the AdaBoost: Take the weights of the samples in the last iteration of the training (D^T). Plot the training set with size proportional to its weight in D^T, and color that indicates its label (again, you can use decision_boundaries). Oh! we cannot see any point! the weights are to small... so we will normalize them: D = D / np.max(D) * 10. What do we see now? can you explain it?

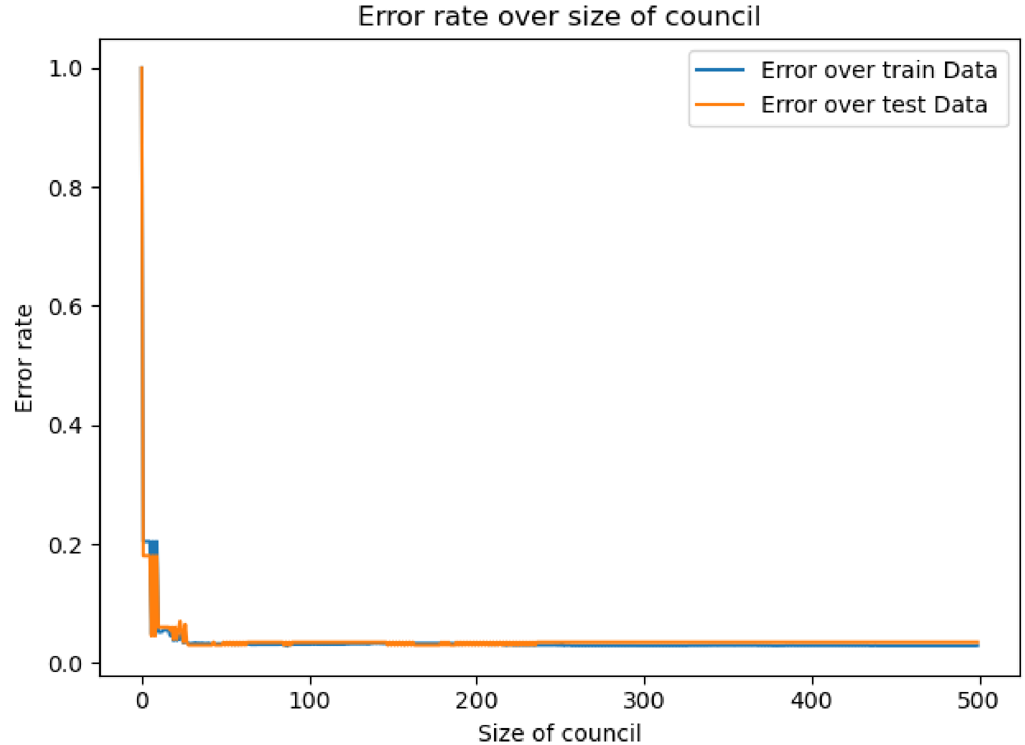


13. ניתן לראות שבשלב הזה כמעט ואין נקודות בכל ההיקף, הן כן שם אבל עם כל פעם שצדקנו בסיווג שלהן הן הלכו וקטנו, וכעת בסיבוב האחרון צדקנו בהן כל כך הרבה עד שהן כבר זניחות.
בנוסף ניתן לראות כי הנקודות הגדולות ביותר, גם הכחולות וגם האדומות, נמצאות בהיקף של המעגל משני צדדיו, זו תוצאה שהיינו יכולים לצפות כי זה האזורים בהם הסיווג הכי עדין, ובהם טעינו הכי הרבה פעמים, ניתן לראות שככל שמתקרבים למרכז הכחול הנקודות הולכות וקטנות, זה אזור שכמעט תמיד צדקנו בו לאורך זמן.

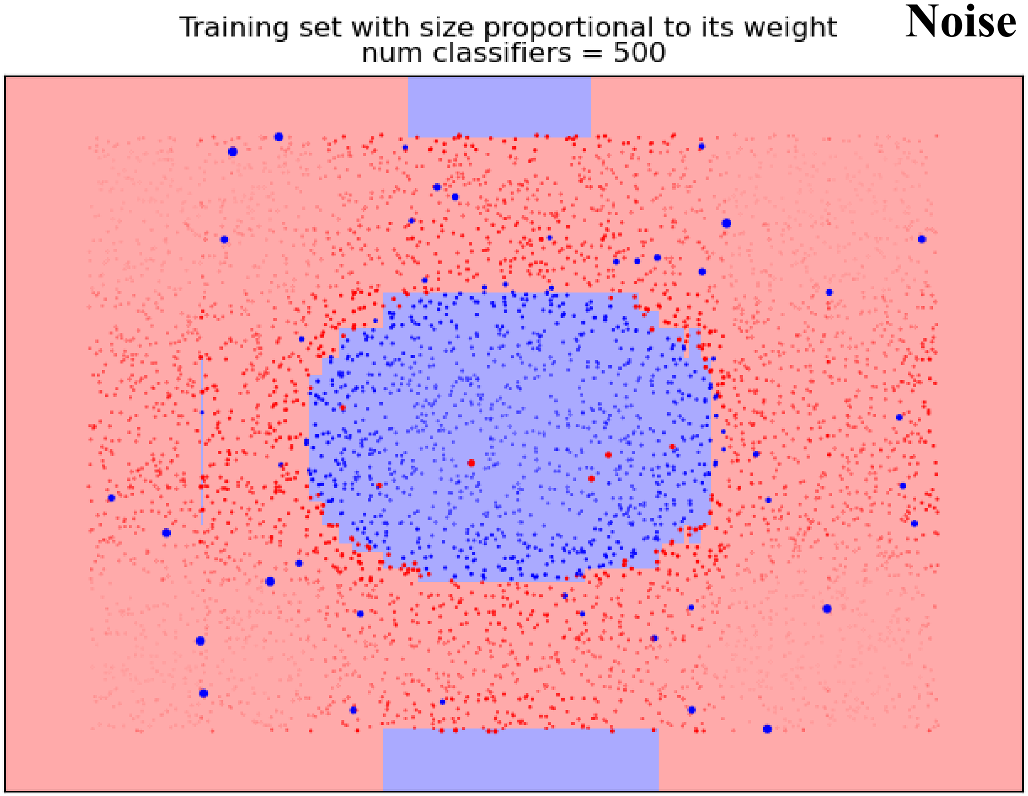
14. Repeat 10,11,12,13 with noised data. Try noise_ratio=0.01 and noise_ratio=0.4.

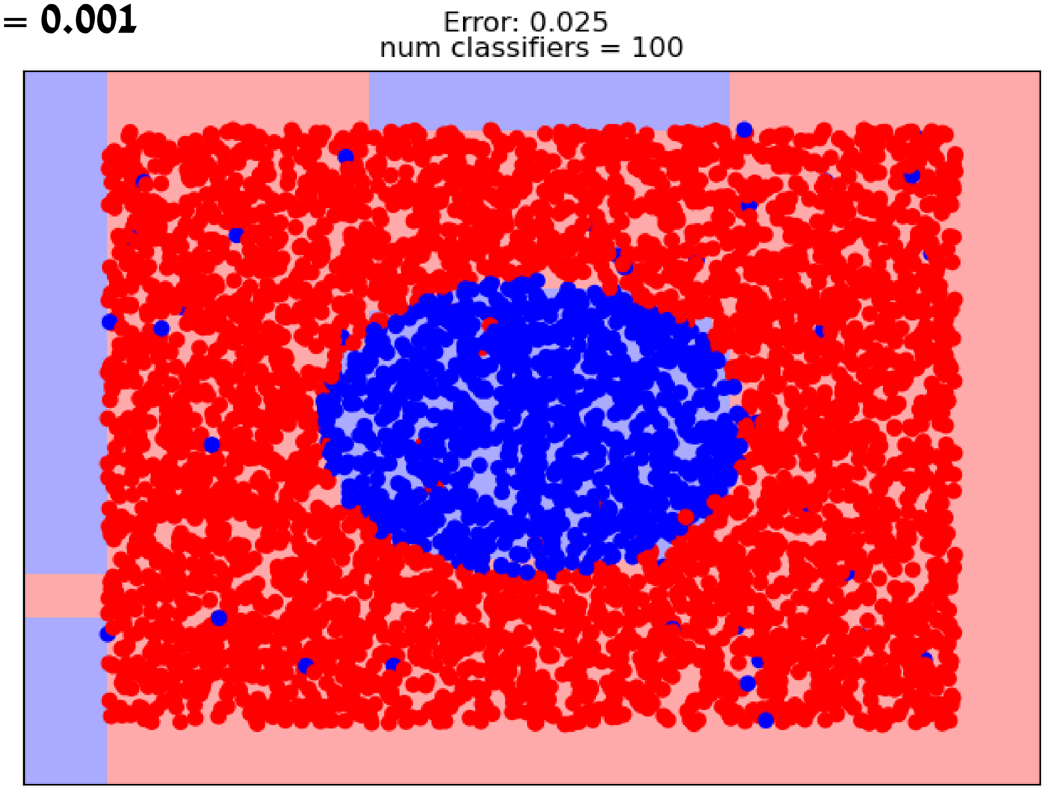
- Add all the graphs to the pdf.
- Describe the changes.
- Explain 10 in terms of the bias complexity tradeoff.
- Explain the differences in 12.

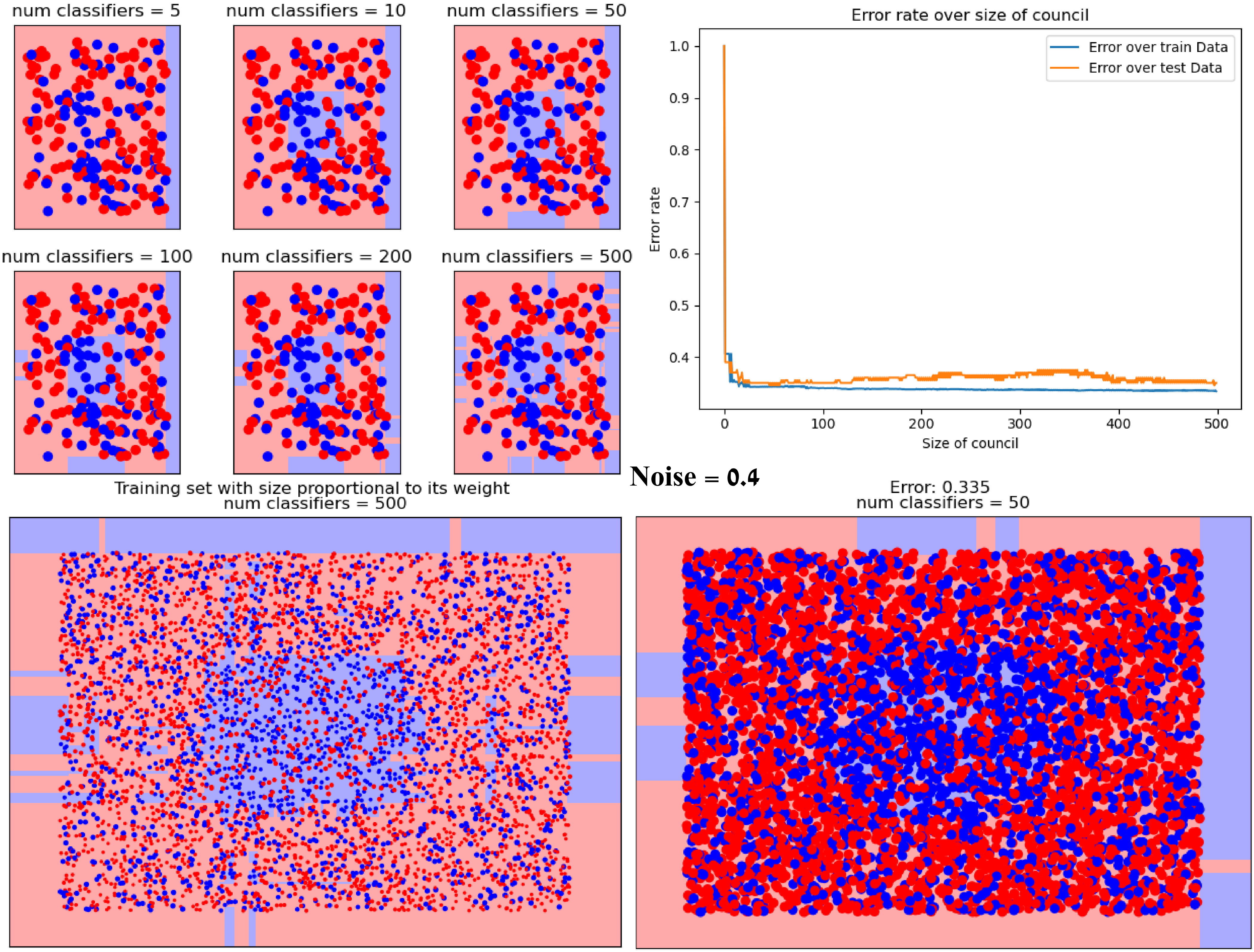

num classifiers = 5
num classifiers = 10
num classifiers = 50
num classifiers = 100
num classifiers = 200
num classifiers = 500


Error rate over size of council
Error over train Data
Error over test Data
Error rate
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500
Size of council

Noise = 0.001


Training set with size proportional to its weight
num classifiers = 500


Error: 0.025
num classifiers = 100


num classifiers = 5
num classifiers = 10
num classifiers = 50
num classifiers = 100
num classifiers = 200
num classifiers = 500
Error rate over size of council
Error over train Data
Error over test Data
Error rate
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0
100
200
300
400
500
Size of council
Training set with size proportional to its weight
num classifiers = 500
Error: 0.335
num classifiers = 50

Noise = 0.4

14. ניתן לראות שינויים רבים כתלות ברמת הרעש בדאטה, ניתן להבחין ככלל אצבע שככל שיש לנו יותר רעש יש יותר אזורי חלוקה רבים יותר אותם ה- AdaBoost מנסה להכליל, כמו כן כצפוי הטעות המינימלית במדגם הרועש יותר גבוהה מהטעות במדגם הרועש פחות, ואולי התוצאה המעניינת ביותר בעיניי היא שניתן לראות בגרף השמאלי התחתון, המייצג את גודל הנקודה כמספר הפעמים שסיווגנו אותה נכונה, כי הנקודות על המדגם הרועש יותר הן באופן כללי גדולות יותר מאלה שבמדגם השקט, הדגמה וויזואלית לאיך שהרעש מפריע לנו לסווג נכון נקודות לאורך זמן.

ניתן לראות כי ההבדלים ב-12 (גרף ימני תחתון) הם שכשהרעש גדול יותר יש יותר נקודות מפוזרות שהמודל מנסה לתפוס, זה גורם לשברור של אזורי הסיווג וליצירה של הרבה יותר אזורים ספורדיים כאשר הרעש גבוהה יותר. ניתן בנוסף לראות שהדגם הרועש יותר הוא בעל יחס טעות גבוהה יותר, כצפוי.