Shelly Gamlielly

IML Ex4

1) יהי A אלגוריתם למידה, ותהי D התפלגות.
פונקציית ההפסד חסומה $[0,1]$
צ"ל: a ו- b שקולים

a) $\mathbb{P}_{S\sim D^m}[L_D(A(S)) \le \varepsilon] \ge 1-\delta$

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

$a \Rightarrow b$

נתון כי $\mathbb{P}_{S\sim D^m}[L_D(A(S)) \le \varepsilon] \ge 1-\delta$

נשים ♡ כי

$$\mathbb{P}_{S\sim D^m}[L_D(A(S)) \le \varepsilon] \ge 1-\delta \iff \mathbb{P}_{S\sim D^m}[L_D(A(S)) > \varepsilon] \le \delta$$

מאי שוויון מרקוב, ונמשיך בהמשך

$$\mathbb{P}_{S\sim D^m}[L_D(A(S)) \ge \varepsilon] \le \frac{\mathbb{E}_{S\sim D^m}[L_D(A(S))]}{\varepsilon}$$

נרצה להראות שלכל $\varepsilon'>0$ קיים $m'>0$ כך שלכל $m>m'$ מתקיים

$$\mathbb{E}L_D(A(S)) = |\mathbb{E}[L_D(A(S))]| < \varepsilon'$$

נבחר $m' = m(\varepsilon,\delta)$, $\varepsilon = \delta = \min\{\frac{1}{2}, \frac{\varepsilon'}{2}\}$

I) אם $\frac{1}{2} < \frac{\varepsilon'}{2} \Leftarrow 1 < \varepsilon'$ אזי $\varepsilon=\delta=\frac{1}{2}$

II) אם $\frac{1}{2} \ge \frac{\varepsilon'}{2} \Leftarrow 1 \ge \varepsilon'$ אזי $\varepsilon=\delta=\frac{\varepsilon'}{2}$

I)
$$\mathbb{E}_{S\sim D^m}[L_D(A(S))] \le \mathbb{P}[L_D(A(S)) > \tfrac{1}{2}]\cdot 1 + \mathbb{P}[L_D(A(S)) \le \tfrac{1}{2}]\cdot\tfrac{1}{2}$$
$$\le \mathbb{P}[L_D(A(S)) > \tfrac{1}{2}] + 1\cdot\tfrac{1}{2}$$ מצורף הסבר בהמשך
$$= \mathbb{P}[L_D(A(S)) > \tfrac{1}{2}] + \tfrac{1}{2} \le \delta + \tfrac{1}{2} = \tfrac{1}{2} + \tfrac{1}{2} \le 1$$
$$< \varepsilon' \Rightarrow \mathbb{E}_{S\sim D^m}[L_D(A(S))] < \varepsilon'$$

II)
$$\mathbb{E}_{S\sim D^m}[L_D(A(S))] \le \mathbb{P}[L_D(A(S)) > \tfrac{\varepsilon'}{2}]\cdot 1 + \mathbb{P}[L_D(A(S)) \le \tfrac{\varepsilon'}{2}]\cdot\tfrac{\varepsilon'}{2}$$
$$\le \mathbb{P}[L_D(A(S)) > \tfrac{\varepsilon'}{2}] + 1\cdot\tfrac{\varepsilon'}{2} \le \delta + \tfrac{\varepsilon'}{2} = \tfrac{\varepsilon'}{2} + \tfrac{\varepsilon'}{2} = \varepsilon'$$
$$\Rightarrow \mathbb{E}_{S\sim D^m}[L_D(A(S))] < \varepsilon'$$

ולכן עפ"י הגדרת הגבול

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

הוכחה: (I

$$\mathbb{E}\, L_D(A(S)) = \int_{-\infty}^{\infty} L_D(A(S))\, f(L_D(A(S)))\, dS$$

$$\overset{(*)}{=} \int_0^1 L_D(A(S))\, f(L_D(A(S)))\, dS = \int_0^{\varepsilon} L_D(A(S))\, f(L_D(A(S)))\, dS$$

$$+ \int_{\varepsilon}^{1} L_D(A(S))\, f(L_D(A(S)))\, dS \le \varepsilon \int_0^{\varepsilon} f(L_D(A(S)))\, dS$$

$$+ 1 \cdot \int_{\varepsilon}^{1} f(L_D(A(S)))\, dS \overset{(**)}{=} \varepsilon\Big(\underbrace{\mathbb{P}(L_D(A(S)) \le \varepsilon) - \mathbb{P}[L_D(A(S)) \le 0]}_{F_{L_D(A(S))}(\varepsilon) - F_{L_D(A(S))}(0)}\Big)$$

$$+ 1\Big(\mathbb{P}(L_D(A(S)) \le 1 - \mathbb{P}(L_D(A(S)) \le \varepsilon)\Big)$$

$$= \varepsilon \cdot \mathbb{P}[L_D(A(S)) \le \varepsilon] - 0 + 1 - \mathbb{P}[L_D(A(S)) \le \varepsilon]$$

$$= \varepsilon\, \mathbb{P}[L_D(A(S)) \le \varepsilon] + \mathbb{P}[L_D(A(S)) > \varepsilon]$$

(*) נתון כי $L_D(A(S)) \in [0,1]$

(**) נזכיר כי מאומר מהתפלגות מצטברת $\int_0^{\varepsilon} f_X(x)\, dx = F_X(\varepsilon) - F_X(0)$

$= \mathbb{P}(X \le \varepsilon) - \mathbb{P}(X \le 0)$

$a \Leftarrow b$

נתון כי $\lim_{m\to\infty} \mathbb{E}[L_D(A(U))] = 0$

כעת נשתמש באי שוויון מרקוב

מהגדרת הגבול, $\mathbb{E}_{S\sim D^m} L_D(A(S)) \to 0$ קיים גבול אם לכל $\varepsilon'>0$

קיים M' טבעי כך שלכל $m > M'$ מתקיים

$$\mathbb{E}_{S\sim D^m} L_D(A(S)) = |\mathbb{E}_{S\sim D^m} L_D(A(S))| < \varepsilon'$$

כעת

$$\mathbb{P}_{S\sim D^m}[L_D(A(S)) \geq \varepsilon] \leq \frac{\mathbb{E}_{S\sim D^m} L_D(A(S))}{\varepsilon}$$

$$< \frac{\varepsilon'}{\varepsilon} = \delta$$ נבחר $\varepsilon' = \varepsilon\delta$ ונקבל כנדרש

$$\Rightarrow \mathbb{P}_{S\sim D^m}[L_D(A(S)) \geq \varepsilon] < \delta$$

$$\Rightarrow \mathbb{P}_{S\sim D^m}[L_D(A(S)) \leq \varepsilon] \geq 1-\delta$$

Sample Complexity of concentric Circle in the Plane (2)

domain set $\mathcal{X} = \mathbb{R}^2$, label set $\mathcal{Y} = \{0,1\}$

תהי H מחלקת היפותזות כך ש $H = \{h_r : r\in\mathbb{R}_+\}$

כאשר $h_r(x) = \mathbb{1}\{\|x\|_2 \leq r\}$

צ"ל: H היא למידה PAC, וה-sample complexity חסום ע"י:

$$m_H(\varepsilon,\delta) \leq \frac{\log\frac{1}{\delta}}{\varepsilon}$$

הוכחה: נתאר את האלגוריתם הלמידה:
בהינתן m דוגמאות אקראיות, נבחר את המעגל C' שייתן לנו את החסימה ההדוקה ביותר לדוגמאות החיוביות
כלומר C' יהיו המעגל עם השטח הקטן ביותר שמכיל את כל הדוגמאות שמתויגות עם 1 + ואף דוגמה עם תיוג 0.
אם אין אף דוגמה חיובית אז $C' = \emptyset$

C' המעגל עם הרדיוס המינימלי האפשרי כך ש: $h_r(x) = \begin{cases} 1 & \|x\|_2 \leq r \\ 0 & \text{else} \end{cases}$

נרצה להראות שלכל $\delta,\varepsilon \in (0,1)$ קיים $m = m_H(\varepsilon,\delta)$ כך שלכל התפלגות D מעל $\mathbb{R}^2$ ולכל מעגל C, עם הסתברות לכל הפחות $1-\delta$ מעל הבחירה של ה-training set S מגודל לפחות m, השגיאה של האלגוריתם $L_{D,C}(C')$ היא לכל היותר ε

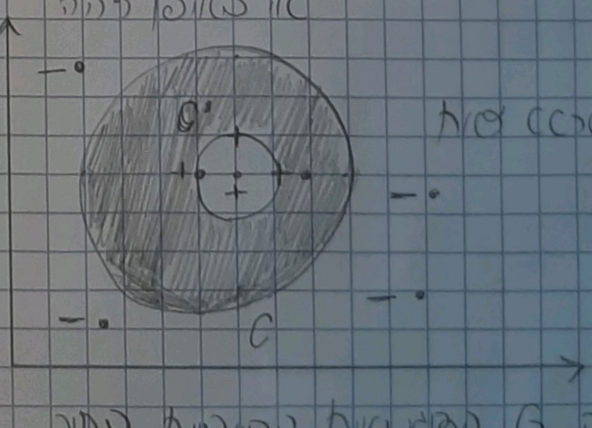
ובכתיב מתמטי:

$$\mathbb{P}[L_{D,C}(C') \leq \varepsilon] \geq 1-\delta$$

או באופן זהה

$$\mathbb{P}[L_{D,C}(C') > \varepsilon] \leq \delta$$

נשרטט זאת



* נסמן תיוג 0 ב- −

נשים לב כי השגיאה המקסימלית היינו השגיאה שבה אנו טועים כאשר נסווג לפי C'

כל הדוגמאות שמחוץ למעגל הפנימי יתוייגו כ-0, אך הדוגמאות שמחוץ למעגל הפנימי ובתוך המעגל החיצוני הן בעצם עם תיוג 1

וזה בגלל שהשגיאה האפשרית שם אנו נסווג לא נכון

מכיוון כל הדוגמאות החיוביות מהוות מתוך C' וכל הדוגמאות השליליות מחוץ ל-C אנו נתייחס רק לאיזור $C \setminus C'$ אלו נסווג על פי C'

רק אנו נוכיח כי הוכחנו חיובית שסופית מעומר $C\setminus C'$

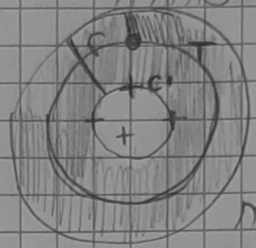
כעת נגדיר את T להיות טבעת ($r \ge \|x\| \ge \hat{r}$) מהמרכז (המעגל) C שרדיוסו r כך שהמשקל של T- עפ״י התפלגות סופית מתוכו שמוכן כל ההסתברות היא ε

נסמן את $T' = C\setminus C'$ ונטען כי T' כמה משקל גדול מ-ε תחת ההתפלגות D אזי הוא כולל את T.
כמו כן T' מכיל את T אזי אף נקודה מ-T לא מופיעה במדגם S - שכן אחרת נקודה $p \in T$ אף היא נקודה פנימית לכן חוג המינימלי שכן הוא מכיל, C', ואף על פי ההגדרה של C' שמעגל והרחק מאחר מחובו מהנקודות החיוביות C' אומרת להתרחב לתוך T הזי שגם כן הנקודה מחוברה

$T=\{x\in\mathbb{R}^2: \hat{r}\le\|x\|\le r\}$ ושרכב גבוה, והו $\hat{r} \ge r^*$ כך ש
$D_x(\{x ; \hat{r}\le\|x\|\le r\}) = \epsilon$
אם אף נקודה מ-S לא שייכת $\hat{r} \le \|x\| \le r$ אז T' במשוא ⇐ T

אם $p \in T$ מופיעה מ-S אז T' לא כולל את T

T לא מוכל ב-T'



T מוכל ב-T'

עפ״י ההגדרה של T - ההסתברות שהדגמו פונקציה לא תפסיק משקל T הוא ε
לכן ההסתברות ש-m דגימות כאלה שמגיעות מההתפלגות D - בלתי תלוי לא יפלו מ-T הוא $(1-\epsilon)^m$

נשתמש בעובדה שההסתברות של אירוע מאורעות ב״ת היא מכפלת ההסתברויות של כל מאורע בנפרד.

לכן ההסתברות שהשגיאה גדולה מ-ε חסומה ע״י ההסתברות שאף דגימה מ-S לא תיפול ב-T
$P[L_{D,c}(C') > \epsilon] \le (1-\epsilon)^m$
$\le e^{-\epsilon m}$ ⇐ $1-\epsilon \le e^{-\epsilon}$ ונזכור כי

$T' = C\setminus C'$ כמה משקל גדול מ-ε אם״ם $T \subseteq T'$ ⟺ אף נקודה מ-T לא מופיעה במדגם S

מכאן שנבחר m שמספק את $\delta \ge (1-\epsilon)^m$,
אזי בהסתברות $1-\delta$ מעל m דגימות אקראיות המשקל של איזור השגיאה $C\setminus C'$ יהיה חסום על ידי אפסילון-ε

אנו רוצים כי δ עבור m המקיים $e^{-\epsilon m} \le \delta$
$-\epsilon m \le \ln\delta$, $m \ge \frac{\ln\delta}{-\epsilon}$, $m \ge \frac{-\ln\delta}{\epsilon}$

$m \ge \frac{\ln\left(\frac{1}{\delta}\right)}{\epsilon}$

והוכחנו כי המחלקה הינה למידה PAC

* $P[L_D(h_S) \ge \epsilon]$ חסומה מלמעלה ע״י ההסתברות שאף נקודה מ-S לא שייכת למאורע $\{x\in\mathbb{R}^2 ; r' \le \|x\| \le r\}$

הערות:
קיימת פונקציה f עבורה השגיאה היא 0:
שמחלקת ההגדלות קיים מעגל המכיל בתוכו רק נקודות חיוביות ומחוץ למעגל נקודות עם תיוג 0
הרי פונקציה שהגדרנו מחזירה $h_S = C'(S)$ עבור $S \sim D^m$
האיזור שבו נמצא מוגדר ע״י $\{x\in\mathbb{R}^2 ; r' \le \|x\|_2 \le r\}$
כאשר r הרדיוס של מעגל C ו-r' הרדיוס של C'

אותה עם שגיאה ε

$r' = \inf\{a : \mathbb{P}[a \le \|x\|_2 \le r] < \epsilon\} \Rightarrow \mathbb{P}[\{x \in \mathbb{R}^2 : r' \le \|x\| \le r\}] \ge \epsilon$
$T=\{x\in\mathbb{R}^2 : \hat{r}\le\|x\|\le r\}$ $\Rightarrow \mathbb{P}[L_D(h_S) \ge \epsilon] \le \mathbb{P}[\{$אף נקודה מ-S לא נמצאת ב-T$\}]$

VC dimention

Boolean conjunction

יהיו $\mathcal{Y}=\{0,1\}$, $\mathcal{X}=\{0,1\}^d$ ונניח כי $d\geq 2$

לכל דוגמה $(x,y)\in \mathcal{X}\times\mathcal{Y}$ מכילה רשומה ל-d משתנים בוליאניים x ותיוג y

לכל משתנה בוליאני x_k, $k\in[d]$ קיימים שני ליטרלים: x_k , $\overline{x_k}=1-x_k$.

המחלקה H_{con} מוגדרת ע״י boolean conjunctions של 0 תת הקבוצה של $2d$ הליטרלים.

צ״ל: VC dimention של H_{con}

לפי הגדרה
$$VCdim(H_{con}) = \sup\{m\in\mathbb{N} : \exists C \text{ with } |C|=m \text{ and } H \text{ shatters } C\}$$

כאשר מחלקת היפותזות H_{con} מנתצת קבוצה סופית $C\subset\mathcal{X}$ אם הצמצום של H_{con} ל-C הוא מגודל $2^{|C|}$, כלומר אם הפונקציות מ-C יכולות לקבל את כל התיוגים האפשריים.

כדי להוכיח ש $VC\,dim(H_{con})=d$ נוכיח כי

I) קיימת קבוצה $C\subseteq\mathcal{X}$ מגודל d ש-H_{con} מנתצת אותה ולכן $VC\,dim(H_{con})\geq d$

II) לכל קבוצה $C\subseteq\mathcal{X}$ מגודל $d+1$, H_{con} לא מנתצת את C. כלומר כל הקבוצות מגודל $m\geq d+1$ לא מנותצות ע״י H_{con} ולכן $VC\,dim(H_{con})\leq d$

I) תהיי $C=\{e_1,\dots,e_d\}$ הקבוצה שמכילה את וקטורי היחידה ונראה כי H_{con} מנתצת את C מגודל d

לכל $(y_1,\dots,y_d)\in\{0,1\}^d$ נגדיר $h(e_i)=\bigwedge_{i\in I}\overline{x_i}$ כאשר $I=\{i : y_i=0\}$

לדוגמא עבור $d=2$ ו- $e_1=(1,0)$, $(y_1,y_2)=(0,0)$: $h(e_1)=0\wedge 1=0=y_1$

$d=2$, $e_2=(0,1)$, $(y_1,y_2)=(1,0)$: $h(e_2)=\overline{x_2}=0=y_2$

$d=2$, $e_2=(0,1)$, $(y_1,y_2)=(0,1)$: $h(e_2)=\overline{x_1}=1=y_2$

כלומר מהיות וקטור היחידה h מבצעת הפונקציות boolean conjunction על 0 ה- $\overline{x_i}$ לכן 0 - x_i עבור תיוג 0

מאופן ההגדרה של וקטור היחידה לכל e_i מתקיים $x_i=1$ ועבור $j\neq i$ $x_j=0$.

אם $x_i=1$ עבור תיוג $y_i=0$ אז היא יהיו מ boolean conjunction h אז יהיה תוך $\overline{x_i}=0$ ולכן $h(e_i)=0 \Leftarrow h(e_i)=y_i$
מספיק שליטרל 0 אחד $h(e_i)=\bigwedge_{k\in I}x_k=0$

אם $x_i=1$ עבור תיוג $y_i=1$ אז היא לא יכיל boolean conjunction h אף לא משנה איפה x_j עם $j\neq i$ יכיל מבנה של $x_j=0$
אז $\overline{x_j}=1$ ואז $h(e_i)=1 \Leftarrow h(e_i)=y_i$
וכל כולם עם 1 אז 1 הוא שלהם יהיה 1

משלב זה נקבל כי לכל $i\in[d]$ מתקיים $h(e_i)=y_i$ ולכן H_{con} מנתצת את C כנדרש

נבחין כי ה-boolean conjunction h עם קבוצה ריקה מקיים

$h(e_i) = 1$

כלומר אם $I=\emptyset$ נובע לכל $i\in[d]$ $y_i = 1$ אזי

$h(e_i) = 1 = y_i$

כמו כן $x_k \wedge \bar{x}_k = 0$ לכן אם $h(e_i) = \bigwedge_{i\in I} \bar{x}_i$

אם קיים $x_i \in \{0,1\}$ לכן ש $y_i = 0$ ו $x_j \in \{0,1\}$ לכן ש $y_j = 0$

מכיוון ש- e_i וקטור יחידה אזי $x_i \neq x_j$ ולכן $x_i \wedge x_j = 0$

ובאותו אופן $\bar{x}_i \neq \bar{x}_j$ ולכן $\bar{x}_i \wedge \bar{x}_j = 0$

$x_i = 1 \Rightarrow \bar{x}_i = 0$ $\Downarrow$ $x_j = 0 \Rightarrow \bar{x}_j = 1$ $\Downarrow$ $\bar{x}_i \wedge \bar{x}_j = 0$

$x_i = 0 \Rightarrow \bar{x}_i = 1$ $\Downarrow$ $x_j = 1 \Rightarrow \bar{x}_j = 0$ $\Downarrow$ $\bar{x}_i \wedge \bar{x}_j = 0$

II) נניח בשלילה שקיימת קבוצה $C \subset X$ מגודל $d+1$ שמנותצת ע"י H_{con}: $C = (x_1, \dots, x_d, x_{d+1})$

שמוגדרת כך נגדיר $h_1, \dots, h_{d+1} \in H$ כך ש $h_i(x_j) = \begin{cases} 1 & \text{if } i \neq j \\ 0 & \text{else} \end{cases}$

לכן

$h_1(x_1, \dots, x_{d+1}) = (0, 1, 1, \dots, 1)$

$h_2(x_1, \dots, x_{d+1}) = (1, 0, 1, \dots, 1)$

$h_{d+1}(x_1, \dots, x_{d+1}) = (1, 1, 1, \dots, 1, 0)$

נבחין כי בדוגמה $(x_i, y_i) \in (X \times Y)$ מכילה השמה ל-d משתנים בוליאנים x, כמו כן h_i מכילה $d+1$ ליטרלים

לכן, מעקרון שובך היונים קיימים שני ליטרלים שמתאימים לאותו משתנה בוליאני.

נניח בלי הגבלת הכלליות שליטרלים l_1, l_2 מתאימים לאותו משתנה בוליאני, אז קיימות 2 אפשרויות:

1) אם $l_1 = l_2$:

מאופן הגדרת $h_1(\cdot)$ מתקיים $h_1(x_1)=0 \wedge h_1(x_2)=1$

מכיוון ש $l_1 = l_2$ מתקיים $h_2(x_1)=0 \wedge h_2(x_2)=1$

בסתירה לאופן הגדרת $h_2(\cdot)$:

$h_2(x_1)=1 \wedge h_2(x_2)=0$

2) אם $l_1 \neq l_2$:

מאופן הגדרת $h_1(\cdot)$ מתקיים $h_1(x_3)=1$

מכיוון ש $l_1 \neq l_2$ מתקיים $h_2(x_3) \neq 1$

$\Rightarrow h_2(x_3)=0$

בסתירה לאופן הגדרת $h_2(\cdot)$

$h_2(x_k)=0 \iff k=2$

אך ממקרה זה $k=3$

ובצורה יותר כללית:

עבור $i \neq j$

$h_i(x_i)=0 \wedge h_i(x_j)=\underline{1}$

$\Rightarrow h_j(x_i)=0 \wedge h_j(x_j)=\underline{1}$

אם קיימים שני ליטרלים המתאימים לאותו משתנה מונוטוני $l_i = l_j$

בסתירה לכך ש $h_j(x_i)=0 \iff i=j$

אם $l_i \neq l_j$ אזי עבור $k \neq i,\ i \neq j,\ k \neq j$ $h_i(x_k)=\underline{1}$

$\Rightarrow h_j(x_k) \neq \underline{1}$

$\Rightarrow h_j(x_k)=0$

בסתירה לכך ש $h_j(x_k)=0 \iff j=k$

לכן ההנחה שבשלילה לא נכונה קיימת C מגודל d כך ש-H_{con} מנפצת אותה

מכאן ש- $VCdim(H_{con}) \geq d$

נתון: H מקיימת תכונת uniform convergence עם פונקציה (H
$m_H^{UC}:(0,1)^2\to\mathbb{N}$

צ"ל: H היא למידה Agnostic PAC עם sample complexity
$m_H(\epsilon,\delta)\le m_H^{UC}(\epsilon/2,\delta)$

עפ"י הנתון קיימת פונקציה $m_H^{UC}:(0,1)^2\to\mathbb{N}$ כך שלכל $\epsilon,\delta\in(0,1)$
ולכל התפלגות D מעל $X\times Y$, אם S היא
מדגם של $m\ge m_H^{UC}(\frac{\epsilon}{2},\delta)$ דוגמאות iid מ-D, אז
בהסתברות לפחות $1-\delta$, S היא $\frac{\epsilon}{2}$-representative

כלומר מתקיים: $\mathbb{P}_{S\sim D^m}[S \mid S \text{ is } \frac{\epsilon}{2} \text{ representative}]\ge 1-\delta$

כאשר training set S הוא $\frac{\epsilon}{2}$ representative אם

$$\forall h\in H\quad |L_S(h)-L_D(h)|\le\frac{\epsilon}{2}\iff -\frac{\epsilon}{2}\le L_S(h)-L_D(h)\le\frac{\epsilon}{2}$$

נראה כי אם $h_S\in\arg\min_{h\in H}L_D(h)+\epsilon$ מתקיים $L_D(h_S)\le\min_{h\in H}L_D(h)+\epsilon$

הוכחה: מכיוון ש-S היא $\frac{\epsilon}{2}$ representative מתקיים

$$L_D(h_S)\le L_S(h_S)+\frac{\epsilon}{2},\quad L_S(h)\le L_D(h)+\frac{\epsilon}{2}$$

ומכיוון כי h_S היא משערך ERM לכל $h\in H$ מתקיים

$$L_S(h_S)\le L_S(h)$$

$$L_D(h_S)\le L_S(h_S)+\frac{\epsilon}{2}\le L_S(h)+\frac{\epsilon}{2}\le L_D(h)+\frac{\epsilon}{2}+\frac{\epsilon}{2}$$

S היא $\frac{\epsilon}{2}$ representative ולכן $L_D(h)\le L_S(h)+\frac{\epsilon}{2}$

$$=L_D(h)+\epsilon\Rightarrow\boxed{L_D(h_S)\le\min_{h\in H}L_D(h)+\epsilon}$$

לכן לכל $S\in(X\times Y)^m$

$$\mathbb{P}_{S\sim D^m}[L_D(h)\le\min_{h\in H}L_D(h)+\epsilon]$$
$$\ge\mathbb{P}_{S\sim D^m}[S\mid S\text{ is }\frac{\epsilon}{2}\text{ representative}]\ge 1-\delta$$

מתכונת ה-UC

ולכן קיבלנו כי לכל $\epsilon,\delta\in(0,1)$ ולכל התפלגות D מעל $X\times Y$
ו-$m\ge m_H^{UC}(\frac{\epsilon}{2},\delta)$ מתקיים

$$\mathbb{P}_{S\sim D^m}[L_D(h_S)\le\min_{h\in H}L_D(h)+\epsilon]\ge 1-\delta$$

ומכאן ש-H היא למידה Agnostic PAC

Monotonicity

7) יהיו H_1, H_2 מחלקות של קלסיפיקציה בינארית כך ש $H_1 \subseteq H_2$.

הוכח: $VCdim(H_2) \geq VCdim(H_1)$

נזכיר כי מחלקת היפותזות H_1 מנתצת קבוצה סופית $C \subseteq X$ אם $|H_{1C}| = 2^{|C|}$, כלומר אם ההגבלות על C יכולות להביא את כל ה- labeling האפשריים

$$VCdim(H_1) = \sup\{m \in \mathbb{N} : \exists C \text{ with } |C|=m \text{ and } H_1 \text{ shatters } C\}$$

תהי $C = \{c_1, \dots, c_m\}$ ונניח ש $VCdim(H_1) = m$

כך ש- H_1 מנתצת את C $\Rightarrow$ $|H_{1C}| = 2^m$

נראה ש $H_1 \subseteq H_2$ אזי $h_1 \in H_1 \Rightarrow h_1 \in H_2$

ההגבלה של H_1 על C: אוסף כל ההיפותזות h_C כך ש-

$$H_{1C} = \{h_{1C} = h_1(c_1), \dots, h_1(c_m) \mid h_1 \in H_1\}$$

$$H_{2C} = \{h_{2C} = h_2(c_1), \dots, h_2(c_m) \mid h_2 \in H_2\} \supseteq \{h_{1C} = h_1(c_1), \dots, h_1(c_m) \mid h_1 \in H_2\}$$

וכעת, ההגבלה של H_2 על C

אזי $H_1 \subseteq H_2 \Rightarrow |H_1| \le |H_2| \Rightarrow |H_{1C}| \le |H_{2C}| \Rightarrow 2^m = |H_{1C}| \le |H_{2C}| \le 2^m$

לכן $|H_{2C}| = |H_{1C}| = 2^m \Rightarrow H_2$ מנתצת את C

כל היפותזה H_i עבור $i \in \{1,2\}$ ניתן להגדיר שמתאר אותן היא מחזיקת את התיוגים הספציפיים ב- C, ובכך ההיפותזות שונות שמחזיקות את C אותו הדבר

מסקנה: תהי $h \in H_1$ אזי $h \in H_2$ ולכן הגודל של כל הקבוצה ש- H_2 מנתצת אותה היא לפחות בגודל $|C| = m$ שאותה H_1 מנתצת

H_1 מנתצת את C $\xRightarrow{H_1 \subseteq H_2}$ H_2 מנתצת את C

$$VCdim(H_1) \le VCdim(H_2)$$

$VCdim(H_1)$ - הגודל המקסימלי של קבוצה $C \subseteq X$ ש- H_1 יכולה לנתץ במקרה שלנו $|C| = m$

$VCdim(H_2)$ - הגודל המקסימלי של קבוצה $C \subseteq X$ ש- H_2 יכולה לנתץ

מכיוון ש- H_2 מנתצת קבוצה בגודל m לכן $m \le VCdim(H_2)$

8) יהי $\mathcal{X}$ מרחב המופעים ו-$\mathcal{Y}=\{\pm1\}$.
תהי $H\subseteq \mathcal{Y}^{\mathcal{X}}$ מחלקת היפותזות.
לכל $C\subseteq \mathcal{X}$ יהיו H_C הצמצום של H ל-C,
נגדיר פונקציה $\tau_H(m) : \mathbb{N}\to\mathbb{N}$
ע"י $\tau_H(m)=\max\{|H_C| ; C\subseteq \mathcal{X}, |C|=m\}$

a) נסביר במילים את המשמעות של τ_H
זוהי פונקציה מהטבעיים לטבעיים עם מחלקת היפותזות H
כך שבהינתן מספר m מחזירה את גודל הצמצום המקסימלי
עבור דגימה כלשהי מגודל המופעים (m דוגמאות)

כלומר אם נעבור על כל המופעים האפשריים בגודל m מתוך מרחב
המופעים, ולכל אחד נבדוק את הצמצום למחלקת ההיפותזות
שלה ומספר כמה איברים שונים יש.
τ_H יהיה המספר המקסימלי מתוך כל הדוגמאות.
($\tau_H(m)$ הוא מס' הפונקציות השונות מקבוצה C בגודל m ל-$\{\pm1\}$ שניתנות מהגבלה של H)

b) נניח כי $VCdim\, H=\infty$, ונמצא מה יהיה הערך של $\tau_H(m)$
עבור $m\in\mathbb{N}$. (כמובן $m<VCdim\, H$)

ראשית נבחין אילו ערכים $\tau_H(m)$ יכולה לקבל עבור קבוצה בגודל
$m\in\mathbb{N}$:

$$H_C\subseteq\{\pm1\}^m \Rightarrow |H_C|\le 2^m$$

כי יש לכל היותר 2^m פונקציות

$VCdim(H)=\infty \Leftarrow$ לכל $m\in\mathbb{N}$ מתקיים $\tau_H(m)=2^m$

עבור $VCdim<\infty$
הערכים ש-$\tau_H(\cdot)$ יכולה לקבל עבור קבוצה בגודל $m\in\mathbb{N}$ הם לכל
היותר 2^m

מההגדרה של VCdim - הוא מוגדר להיות המס' הגדול
ביותר שעבורו $\tau_H()$ הוא בגודל 2^m

כלומר אם נעבור על כל המספרים לפי הסדר עם שומרים את המספר
הגדול שבו מתקיים $\tau_H(m)=2^m$, אז המספר הזה יהיה
שווה ל-VCdim
ולכן
$$VCdim(H)=\max_m\{m|\tau_H(m)=2^m\}$$

במקרה ש-$VCdim(H)=\infty$ אז כשנעבור על כל המספרים כולם
יקיימו $\tau_H(m)=2^m$ (לעולם לא נמצא מספר שלא מקיים אותו).

c) נניח כי $VCdim(H)=d$, ונמצא מה יהיה הערך של $\tau_H(m)$ עבור
$m\le d$.

כלומר
$$d=\max_m\{m|\tau_H(m)=2^m\}$$

$\Leftarrow$ לכל $m\le d$ מתקיים $|H_C|\le 2^m$
$\Leftarrow$ גודל הצמצום המקסימלי עבור דגימה כלשהי מגודל
המופעים
$$\tau_H(m)=\max\{|H_C| ; C\subseteq\mathcal{X}, |C|=m\}=2^m$$

שכן אין צורך לעבור רק m דגימות שמנפצות את H.
מכיוון ש-d הוא ה-m הגדול ביותר כך ש-$\tau_H(m)=2^m$

כלומר אם $VCdim=d$ אז לכל $m\le d$ מתקיים $\tau_H(m)=2^m$.

6

נניח כי $VCdim = d$ ויהיי $m > d$

נניח כי $\tau_H(m) = \left(\frac{em}{d}\right)^d$

i) נוכיח באינדוקציה כי לכל $C \subseteq X$ סופית

$$|H_C| \le |\{B \subseteq C \mid H \text{ shatters } B\}|$$

זו המשמעות של אי השוויון היא שאם $VCdim = d$ אז אין קבוצה C שגודלה יהיה גדול מ-d רק ש-H תנפץ אותה

<u>בסיס האינדוקציה</u>:

עבור $m=1$ מתקיים $C = \{c_1\}$

נבחין כי הצמצום של H על C הוא $H_C = \{h_C = h(x_1), \dots, h(x_m) : h \in H\}$

ולכן $H_C \subseteq \{\pm 1\}$ $\Leftarrow$ $|H_C| \le 2^1$

נשים ♡ כי הקבוצה הריקה מנותצת ע"י H

כעת עבור $B \subseteq C$ מתקיים $B = \emptyset$ או $B = C$

וקיימות שתי אפשרויות

• H מנתצת את C $\Leftarrow$ H מנתצת את B

$\Leftarrow$ $|\{B \subseteq C \mid H \text{ shatters } B\}| = 2$ שכן H מנתצת את $B = \emptyset$

כמו כן H מנתצת את C $\Leftarrow$ $|H_C| = 2^1$

ולכן $|H_C| = 2 = |\{B \subseteq C \mid H \text{ shatters } B\}|$

• H לא מנתצת את C $\Leftarrow$ H לא מנתצת את $B = C$

$\Leftarrow$ $|\{B \subseteq C \mid H \text{ shatters } B\}| = 1$ שכן H מנתצת את $B = \emptyset$

כמו כן H לא מנתצת את C $\Leftarrow$ $|H_C| = 1$

ולכן

$|H_C| = 1 = |\{B \subseteq C \mid H \text{ shatters } B\}|$

<u>צעד האינדוקציה</u>

נניח נכונות הטענה עבור $k < m$ ונוכיח כי הטענה נכונה עבור m

תהיי H וקבוצה $C = \{c_1, \dots, c_m\}$, $C' = \{c_2, \dots, c_m\}$

נסמן כי $|H_{C'}| = |\{(y_2, \dots, y_m) : (0, y_2, \dots, y_m) \in H_C \vee (1, y_2, \dots, y_m) \in H_C\}|$

כמו כן נסמן $Y = \{(y_2, \dots, y_m) : (0, y_2, \dots, y_m) \in H_C \wedge (1, y_2, \dots, y_m) \in H_m\}$

כמו כן $|H_C| = |H_{C'}| + |Y|$

כעת מהנחת האינדוקציה מתקיים

$|H_{C'}| \le |\{B \subseteq C' \mid H \text{ shatters } B\}| = |\{B \subseteq C : c_1 \notin B \wedge H \text{ shatters } B\}|$

כעת נגדיר $H' \subseteq H$ ע"י $H' = \{h \in H : \exists h' \in H \text{ s.t. } (1-h'(c_1), h'(c_2), \dots, h'(c_m)) = h(c_1), h(c_2), \dots, h(c_m)\}$

כלומר H' היא ההשערות שמגיבות אותו דבר על הנקודות שמסומנות עם $C' = \{c_2, \dots, c_m\}$ ונבדלות ב-$\{c_1\}$ כי $1 - h'(c_1) = h(c_1)$

כעת מהנחת האינדוקציה $|Y| = |H'_{C'}| = |\{B \subseteq C' : H' \text{ shatters } B\}| =$

$|\{B \subseteq C' : H' \text{ shatters } B \cup \{c_1\}\}| = |\{B \subseteq C' : c_1 \in B \wedge H' \text{ shatters } B\}| \le$

$|\{B \subseteq C' : c_1 \in B \wedge H \text{ shatters } B\}|$

המשך d i)

מהנחת ה $|H(C')| \leq |\{B \subseteq C' \mid c_1 \notin B \land H \text{ shatters } B\}|$

$|H'(C')| \leq |\{B \subseteq C' \mid c_1 \in B \land H \text{ shatters } B\}|$

נשים לב ש- $|H(C)| = |H(C')| + |H'(C')| \leq |\{B \subseteq C' \mid c_1 \notin B \land H \text{ shatters } B\}|$

$+ |\{B \subseteq C' \mid c_1 \in B \land H' \text{ shatters } B\}| = |\{B \subseteq C' \cup \{c_1\} \mid$

$H \text{ shatters } B\}| = |\{B \subseteq C \mid H \text{ shatters } C\}|$

נשים לב כי $C = \{c_1, \ldots c_m\} = \{c_1\} \cup C' = \{c_1\} \cup \{c_2, \ldots c_m\}$

לפיכך הוכחנו את נכונות הטענה (כשהפירוק C ל- C' ו {c_1})

עבור כך

d) iii) ii)+ נוכיח כי לכל $C \subseteq X$ סופית מתקיים $|\{B \subseteq C;\ H \text{ shatters } B\}| \le \sum_{k=0}^{d} \binom{m}{k}$

יהי $C = \{c_1, \dots, c_m\}$

אוסף כל תתי הקבוצות של C מסומן כ- 2^C

לכל ייחוד B תתי הקבוצות של C שמנותצות ע"י H

ולכן |אוסף B תתי הקבוצות של C| $\ge |\{B \subseteq C;\ H \text{ shatters } B\}|$

בהינתן קבוצה סופית C בעלת m איברים כמות תתי הקבוצות של C שיש בהן בדיוק k איברים היא $\binom{m}{k}$ ולכן

$$|2^C| = \sum_{k=0}^{m} \binom{m}{k}$$

$$|\{B \subseteq C : H \text{ shatters } B\}| \le |2^C| = \sum_{k=0}^{d} \binom{m}{k}$$

נתון כי $VCdim(H) = d = \sup\{m \in \mathbb{N} : \exists C \text{ with } |C| = m \text{ and } H \text{ shatters } C\}$

ולכן הגודל המקסימלי של תת קבוצה של C ש-H מנתצת יהיה d

iv) נתון כי $\sum_{k=0}^{d} \binom{m}{k} \le \left(\frac{em}{d}\right)^d$

$$\begin{aligned}
T_H(m) &= \max\{|H_C| ;\ C \subseteq X, |C| = m\} \\
&\le \max\{|\{B \subseteq C \mid H \text{ shatters } B\}| ;\ C \subseteq X, |C| = m\} \\
&\le \max\left\{\sum_{k=0}^{d} \binom{m}{k}\right\} \\
&\le \left(\frac{em}{d}\right)^d
\end{aligned}$$

e) אם $m = d$ נקבל מהסעיף כי $T_H(m) = 2^m = 2^d$

ולכן כי השוויון $T_H(m) \le \left(\frac{em}{d}\right)^d = e^d$ <u>אינו</u> הדוק

$(2^d < e^d)$

מ-d הסעיף כי $|\{B \subseteq C ;\ H \text{ shatters } B\}| \le \sum_{k=0}^{d} \binom{m}{k} \underset{m=d \text{ כאשר}}{=} \sum_{k=0}^{m} \binom{m}{k}$

f) $VCdim = \max\{d \mid \exists C \subseteq X \text{ s.t. } |C| = d \text{ and } T_H(C) = 2^d\}$

עבור $m \le VCdim(H)$

$VCdim = \max\{d \mid \exists C \subseteq X \text{ s.t. } |C| \text{ and } T_H(C) \le \left(\frac{em}{d}\right)^d\}$

עבור $m > VCdim(H)$

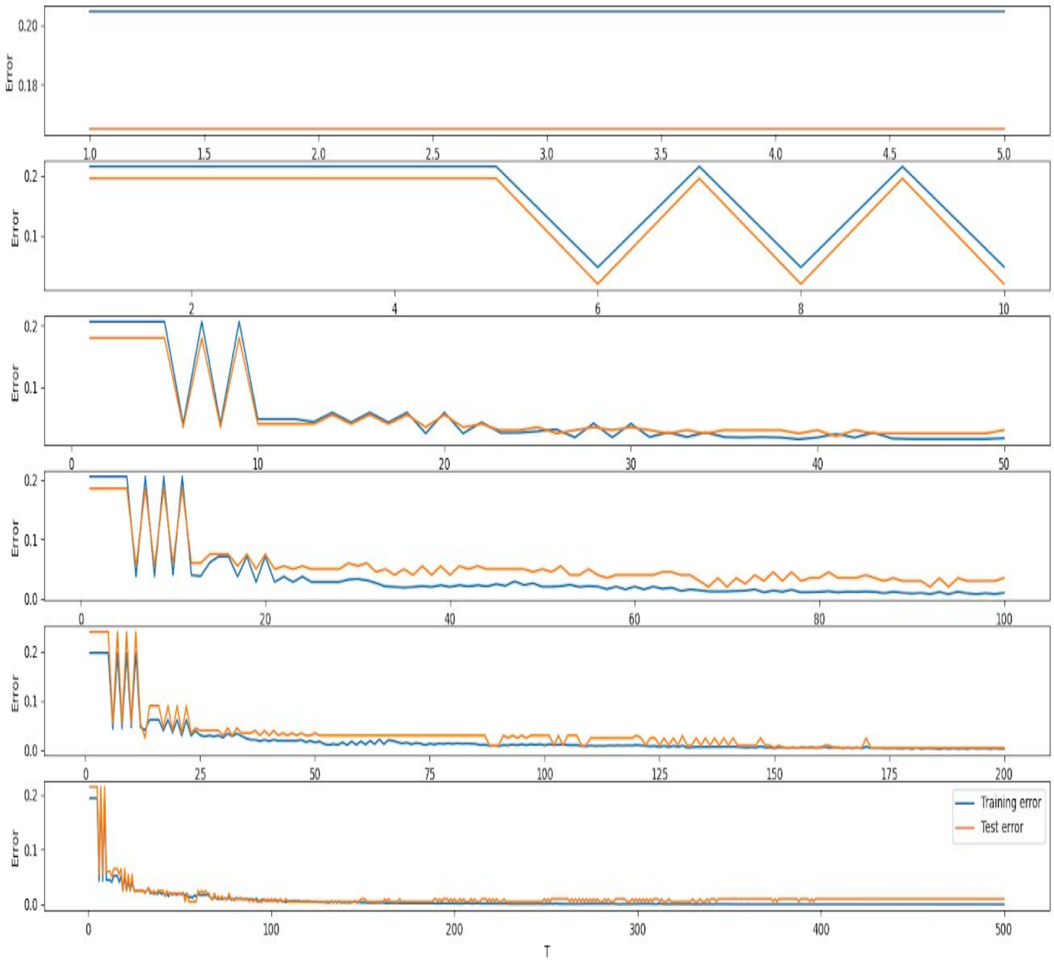
IML EX4

10.

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.

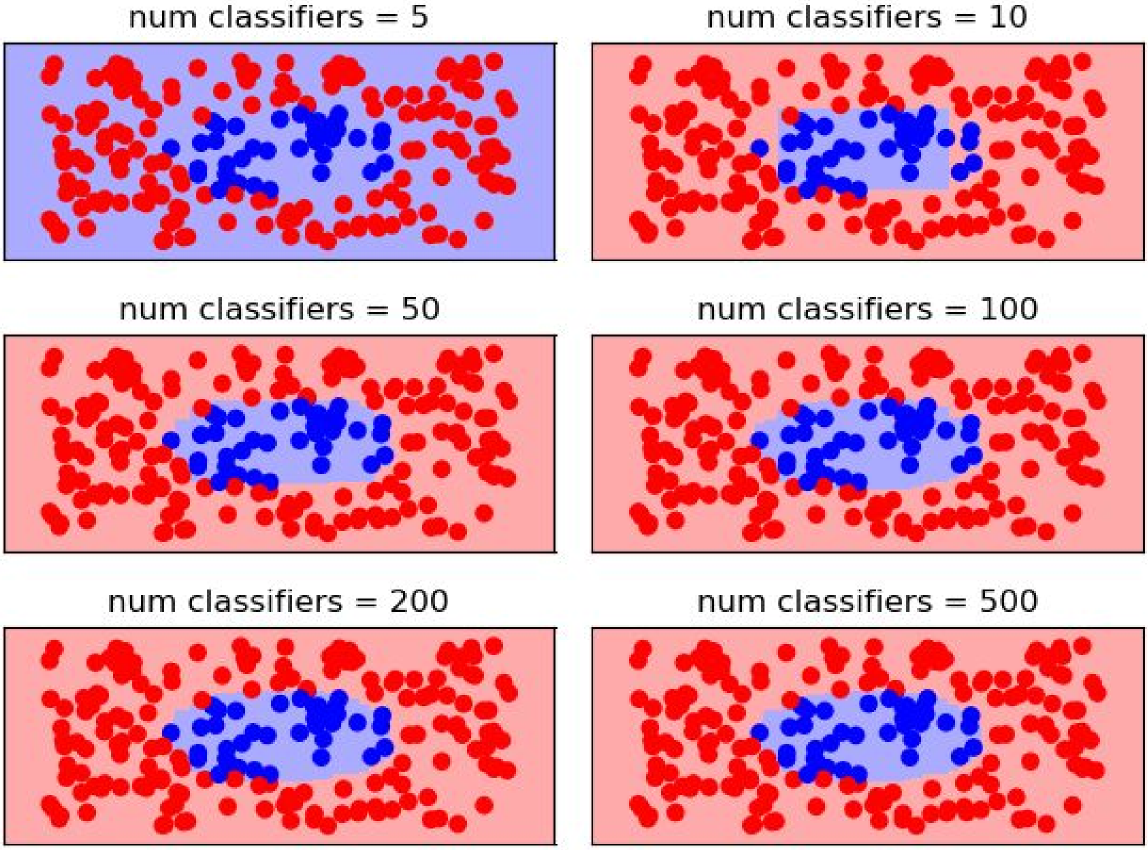
בנוסף צרפתי גרפים שמציגים את השגיאה של האימון והשגיאה של הטסט על T בסדר הבא :

T = 5
T=10
T=50
T=100
T=200
T=500



11.

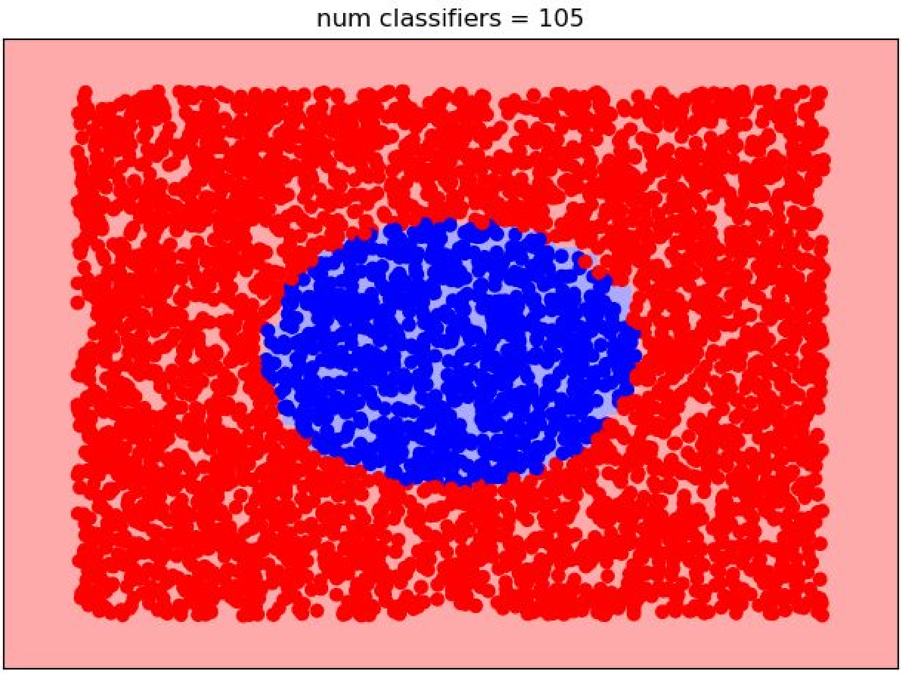
11. Plot the decisions of the learned classifiers with $T \in \{5, 10, 50, 100, 200, 500\}$ together with the test data. You can use the function **decision_boundaries** together with **plt.subplot** for this purpose.



12.

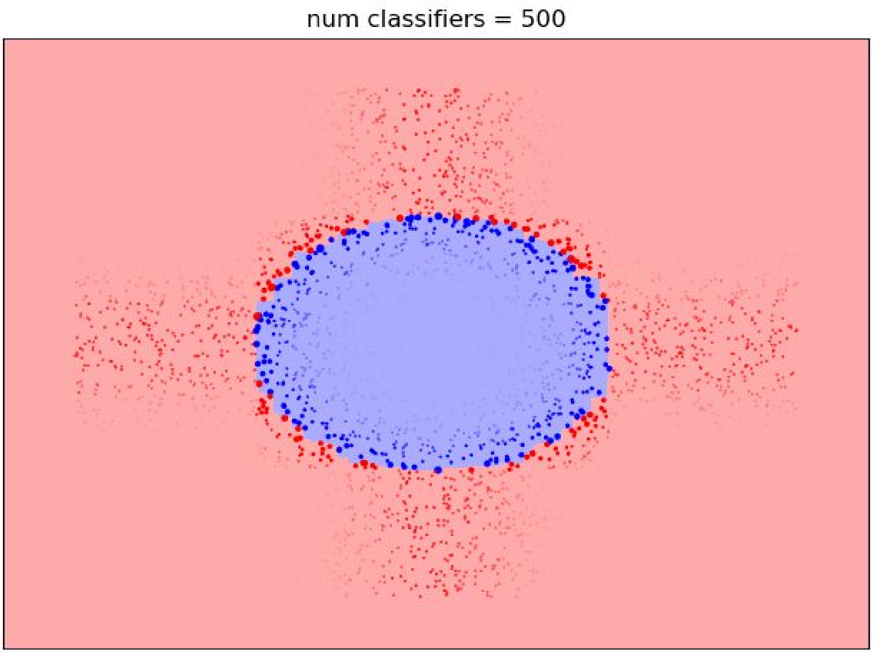
12. Out of the different values you used for T, find $\hat{T}$, the one that minimizes the test error. What is $\hat{T}$ and what is its test error? Plot the decision boundaries of this classifier together with the training data.

minimal test error is : 0.005 with T : 105



13.

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?



השימוש במשקלים המתקבלים באיטרציה האחרונה במהלך האימון בא לידי ביטוי בהצגה של הדאטא ביחס לגודל הנקודות.

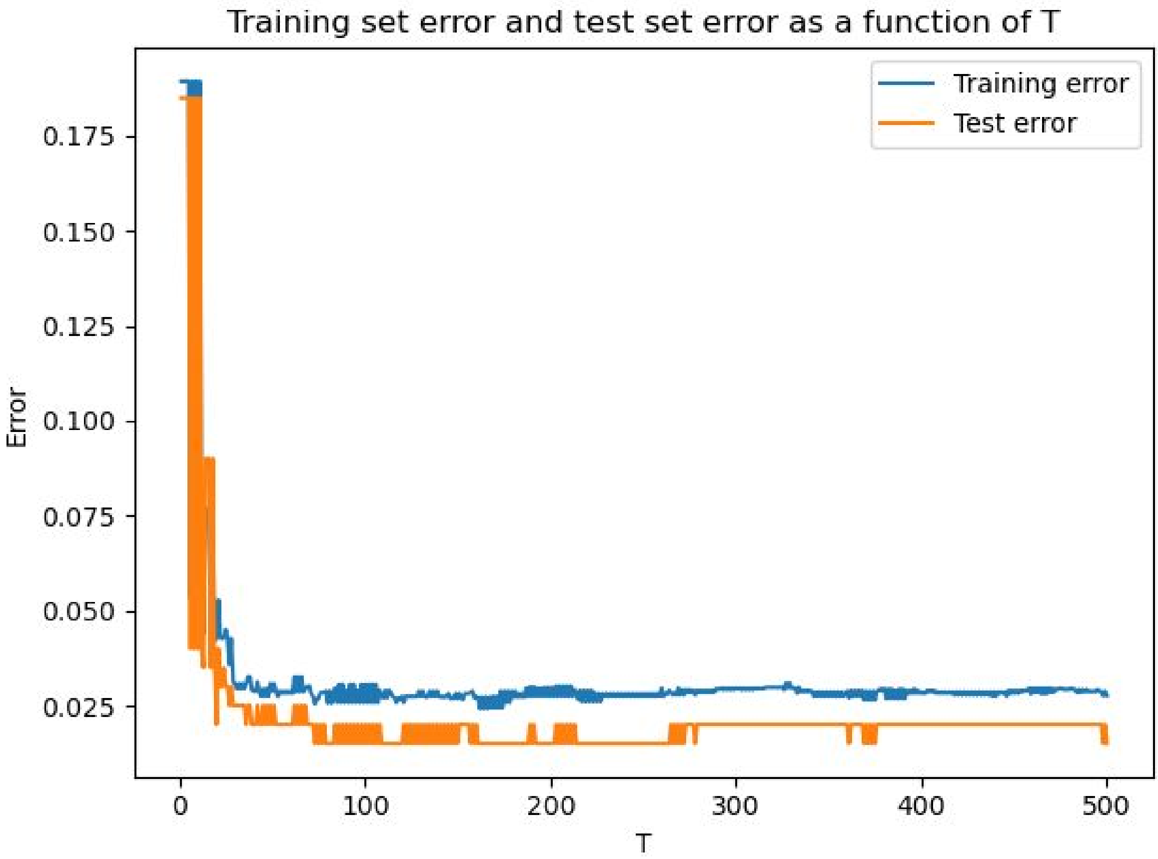
מאופן תיאור האלגוריתם של adaboost אנו נותנים יותר משקל לדוגמאות שטעינו עליהן לעומת דוגמאות שצדקנו עליהן. כך שנצפה שבאיטרציה האחרונה חלוקת המשקלים תהיה בהתאם, כאשר עיקר הטעויות יהיו באזורים בהם יש קרבה בין נקודות עם תיוגים שונים. במילים אחרות יהיה קשה יותר לסווג את הנקודות הקרובות לגבול במובן שהאלגוריתם יטעה עליהן פעמים רבות.
כפי שניתן לראות הנקודות הבולטות ביותר נמצאות על הגבול שמפריד בין דוגמאות עם תיוגים שונים. וזה מתיישב עם אופן פעולת האלגוריתם של adaboost שיטעה פעמים רבות בסיווג של נקודות אלו (כלומר ייתן להן תיוג שגוי) ולכן בכל איטרציה המשקל שלהן יגדל. לעומת זאת עבור הנקודות הרחוקות מאוד מהגבול ונמצאות בקצוות יהיו קלות לסיווג כלומר האלגוריתם לאחר מספר קטן של איטרציות יסווג אותן עם התיוג הנכון. מכאן שהן יהיו בעלות משקלים קטנים יותר משמעותית.

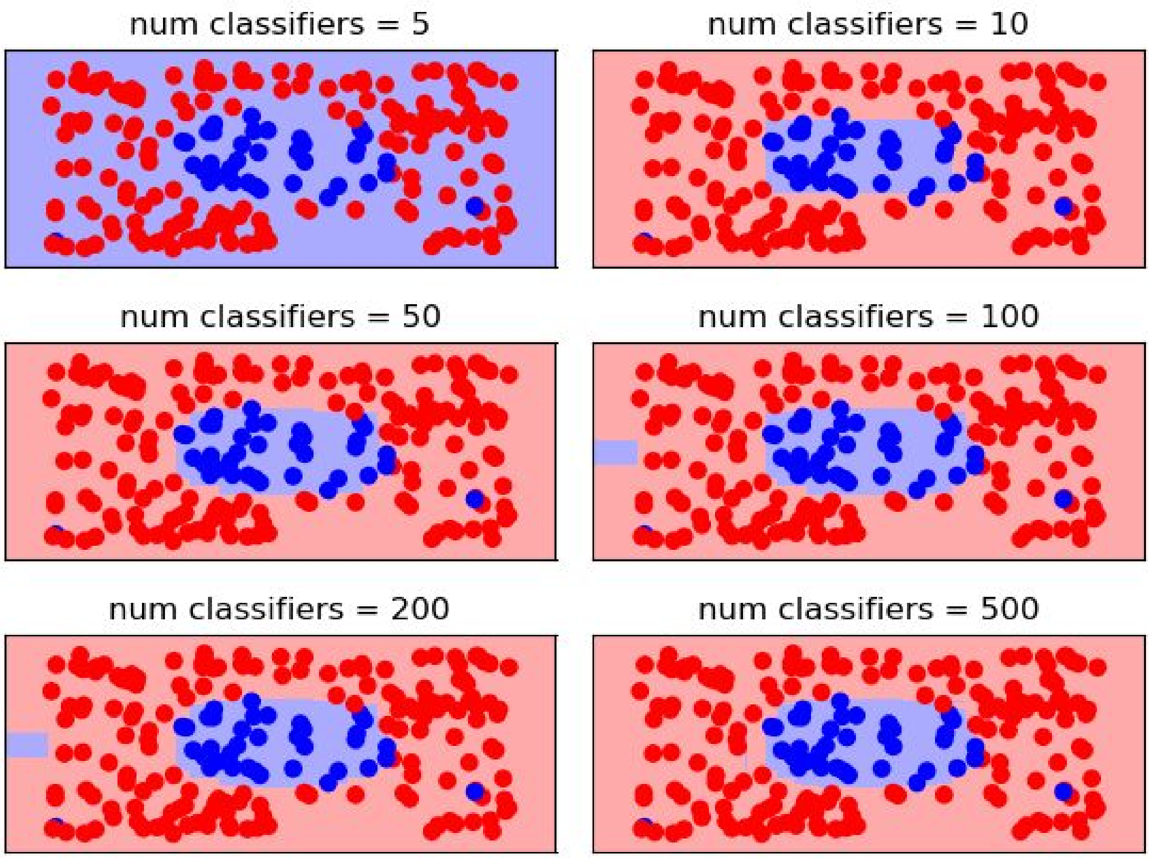
14.

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.

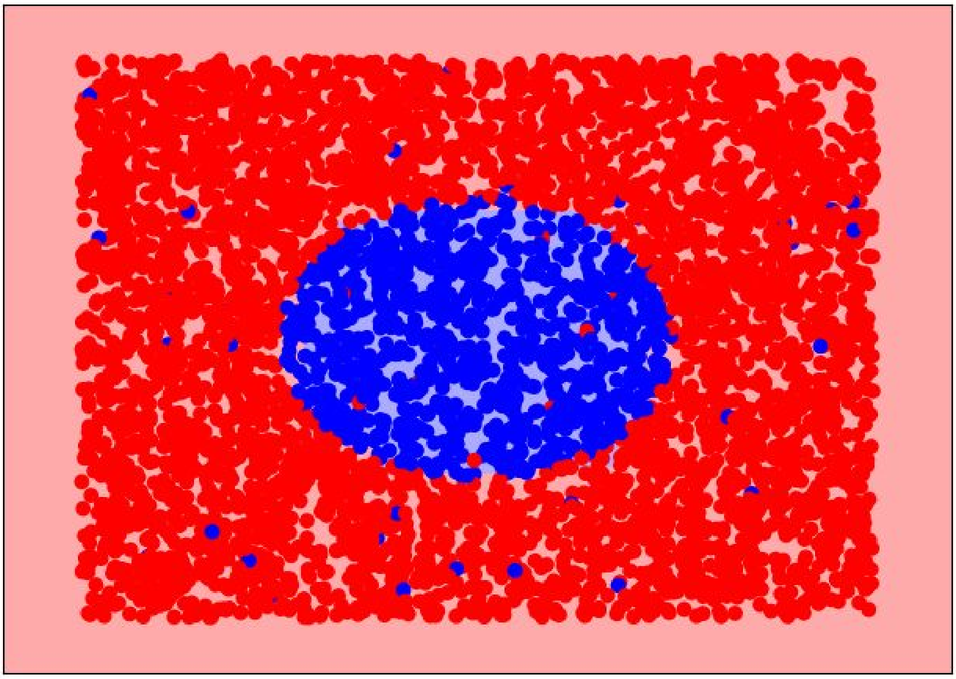
noise ratio = 0.01


Training set error and test set error as a function of T
Training error
Test error
0.175
0.150
0.125
0.100
0.075
0.050
0.025
Error
0
100
200
300
400
500
T

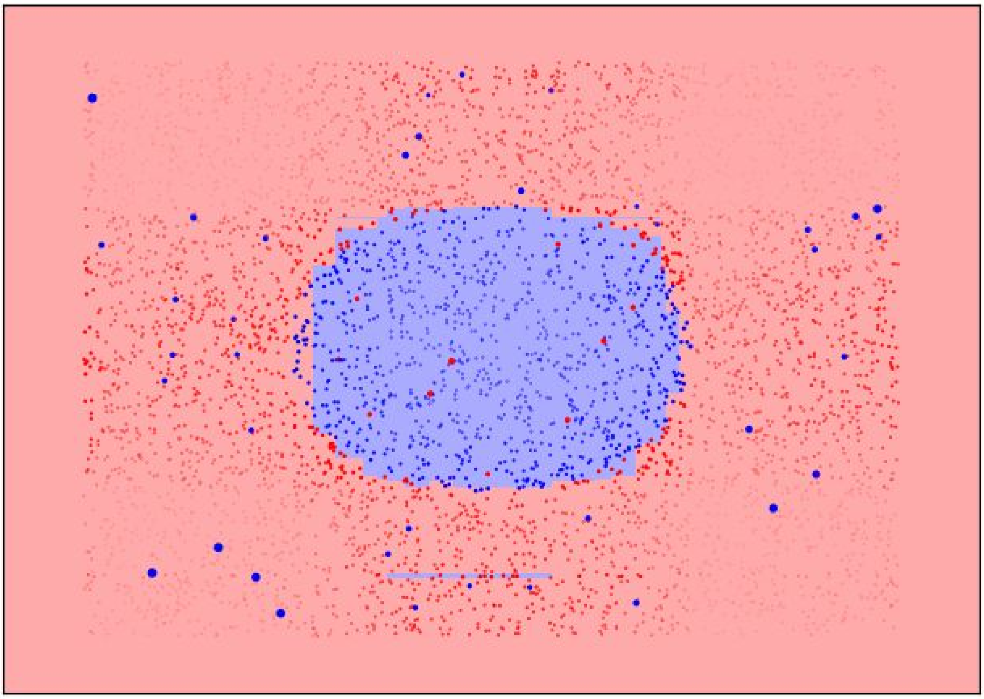

num classifiers = 5
num classifiers = 10
num classifiers = 50
num classifiers = 100
num classifiers = 200
num classifiers = 500

minimal test error is : 0.015 with T : 72

num classifiers = 72

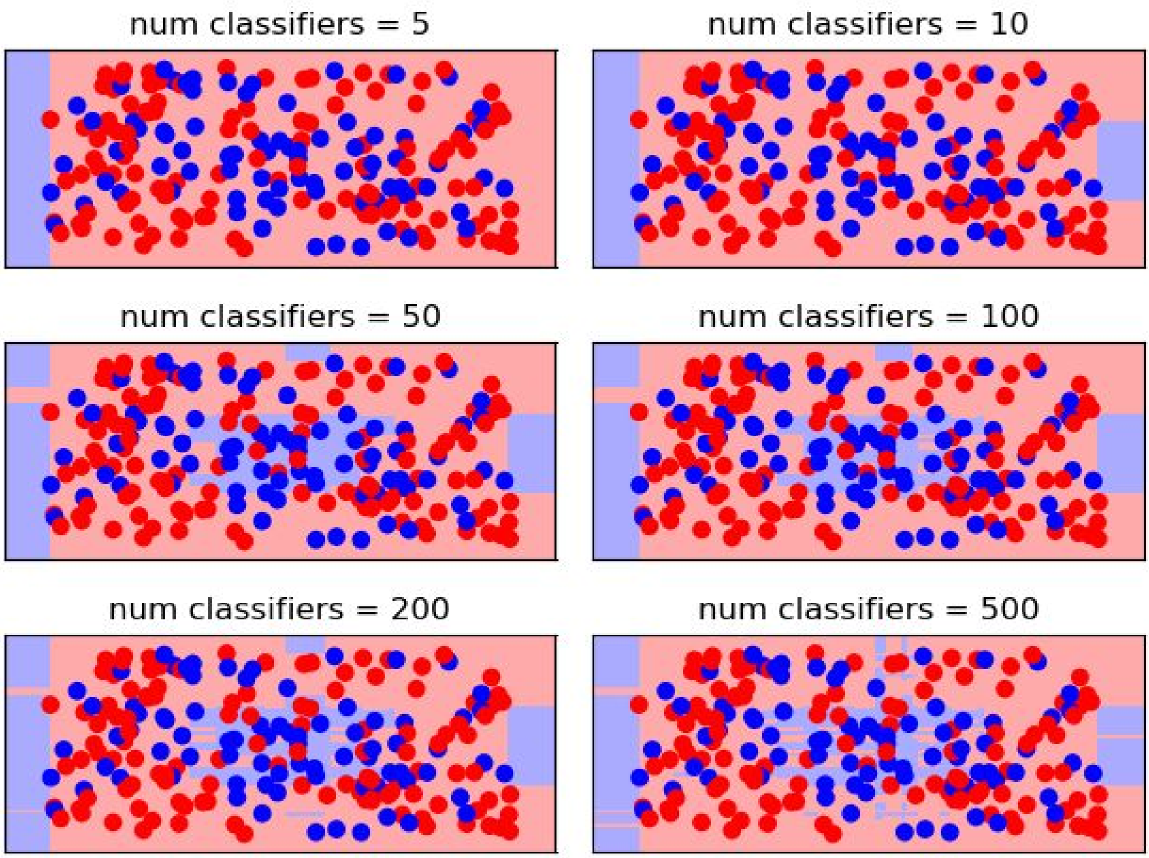


num classifiers = 500



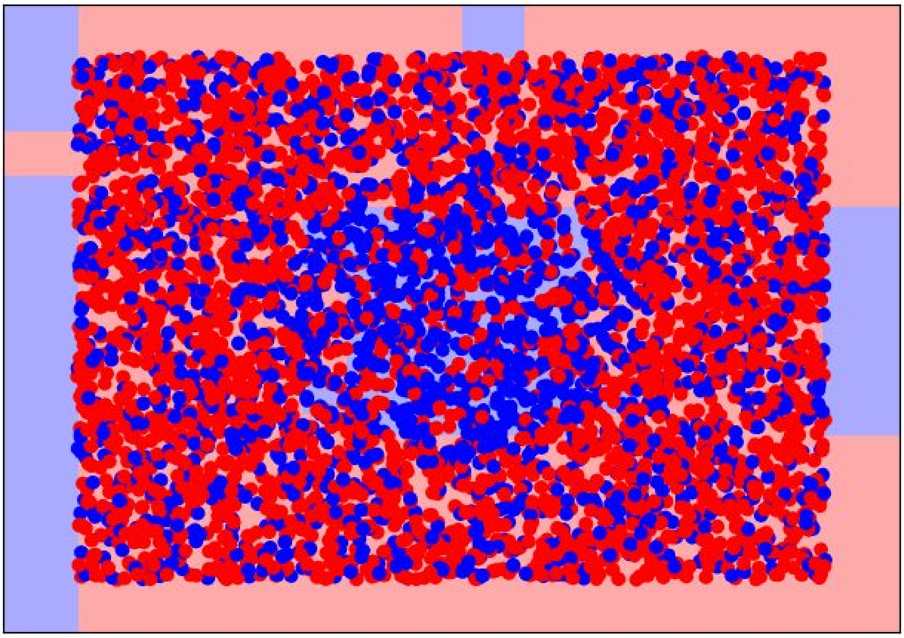
<u>**noise ratio = 0.4**</u>


Training set error and test set error as a function of T
Training error
Test error
Error
T
0.41
0.40
0.39
0.38
0.37
0.36
0.35
0.34
0
100
200
300
400
500

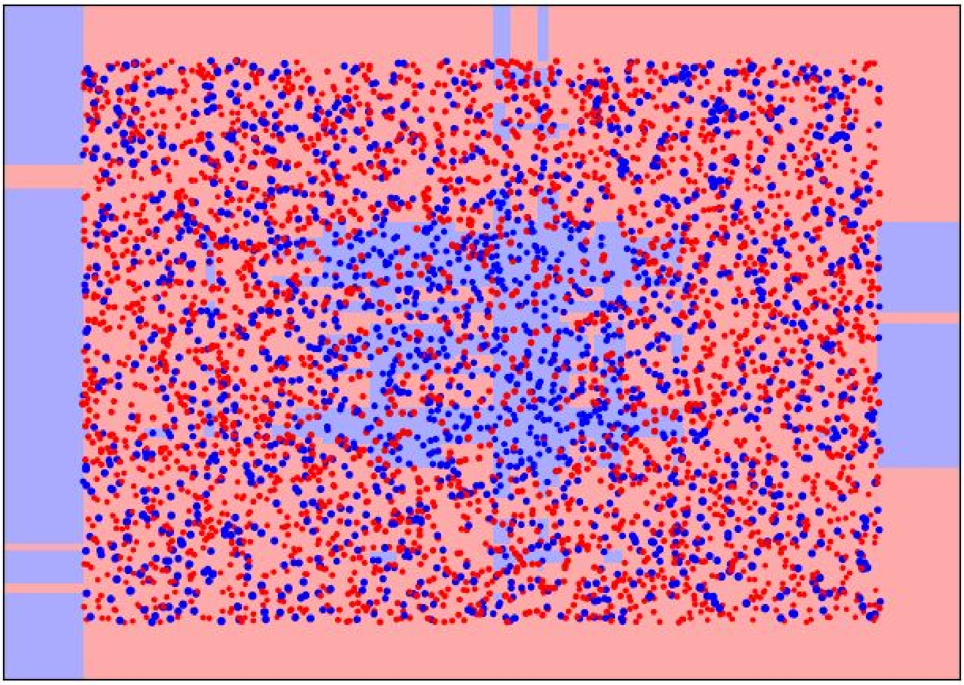

num classifiers = 5
num classifiers = 10
num classifiers = 50
num classifiers = 100
num classifiers = 200
num classifiers = 500

minimal test error is : 0.34 with T : 59

num classifiers = 59



num classifiers = 500



<u>ניתוח הגרפים :</u>

	noise ratio = 0	noise ratio = 0.01	noise ratio = 0.4
minimaL test error	0.005	0.015	0.34
T that minimize the error	105	72	59

נבחין כי ע"פ המדד של שגיאה מינימלית על קבוצת הטסט ככל שהערך של ה noise ratio יותר גדול כך השגיאה המינימלית גדלה.
המשמעות של noise ratio היא יחס הדגימות עבורו נהפוך את התיוגים בסט של האימון.
כמו כן אנו רואים כי עבור דאטא ללא רעש השגיאה המינימלית מתקבלת עבור הT הגדול ביותר ביחס לדאטא מורעש. כאשר יש הבדל בין דאטא מאוד מורעש עם noise ratio של 0.4 בו השגיאה המינימלית מתקבלת עם 59 מסווגים ולעומת זאת עבור דאטא מעט מורעש עם noise ratio של 0.01 השגיאה המינימלית מתקבלת עם 72 מסווגים.
בסך הכל ככל שהדאטא יותר מורעש נקבל שגיאה מינימלית עם פחות מסווגים שכן הרבה מסווגים יגרמו ל Overfitting, וכן השגיאה המינימלית גדלה ככל שיש יותר רעש.

<u>השינויים בגרפים :</u>

- גרף השגיאה של הtest set ושל ה training set

אנו רואים שינוי משמעותי כאשר 40% מהדאטא מתוייג לא נכון - כלומר רעש חזק. השגיאה מתחילה ב ~ 0.41 ועבור סט האימון יורדת עד ~ 0.33 ועבור הטסט סט יורדת ל 0.34 אך עולה בהוספת עוד מסווגים ל ~ 0.37.
זאת לעומת המקרה בו אין רעש כלל או רעש נמוך יחסית של 0.01 בהם השגיאה מתחילה ב~ 0.2 ויורדת ל 0.015 עבור דאטא עם רעש 0.01 ו 0.005 עבור דאטא לא מורעש ביחס לטסט סט. מבחינת סט האימון השגיאה קטנה יותר במקרה של דאטא לא מורעש.

- גרף של ההחלטות שקיבלו המסווגים המאומן על הtest data עבור כל T שמייצג את מספר המסווגים

בהקשר של טסט סט עבור דאטא לא רועש ככל שמספר מסווגים גדל כך השגיאה קטנה. בשלב כלשהו ניתן לראות שאין שינוי משמעותי בשגיאה ביחס למספר המסווגים T.
עבור דאטא רועש מאוד (0.4) יותר מ - 100 מסווגים נתנו תוצאות פחות טובות כלומר שגיאה גדולה יותר לעומת פחות מ 100 מסווגים.

- גרף של ה decision boundaries של המסווג שעבורו קבלנו את השגיאה המינימלית על סט האימון

כאשר הדאטא רועש מאוד 0.4 - הנקודות מבולגנות וקשה לראות הפרדה לעומת המקרה שבו רואים הפרדה מוחלטת בין נקודות כחולות ואדומות עם דאטא לא רועש.
כאשר הדאטא רועש במידה מועטה 0.01 - יש טעויות כלומר ניתן לראות נקודות אדומות בשטח של הכחולות ולהיפך אבל מעט יחסית.

- גרף של ה decision boundaries עם המשקלים של האיטרציה האחרונה

מבחינת הנקודות הבולטות שיש להן משקלים גדולים ניתן לראות במקרה של דאטא מורעש במידה מועטה 0.01 - יש נקודות כחולות בולטות שרחוקות מאיזור הגבול המפריד, בניגוד למה שהיינו רוצים. שכן בדאטא לא מורעש ראינו כי הנקודות מסביב לגבול הן אלה שקבלו משקלים חזקים ונקודות רחוקות יותר קבלו קבלו משקלים נמוכים מאוד.

כאשר הדאטא רועש מאוד 0.4 - הרבה נקודות קבלו משקלים גדולים, בנוסף לא ניתן לראות כלל הימצאות של נקודות עם משקלים גבוהים ספציפית על הגבול המפריד בין נקודות עם תיוגים שונים.

Bias-complexity tradeoff

נוכל לפרק את שגיאה הERM לסכום של שגיאת הקירוב ושגיאת השערוך.
משום שאנו מעוניינים למזער את השגיאה הכוללת אנחנו מתמודדים עם הBias-complexity tradeoff.
מצד אחד בחירת מחלקת היפותזות שהיא עשירה תפחית את שגיאת הקירוב, אך עלולה להגדיל את שגיאת השערוך שכן עלול להיות overfitting.
מצד שני אם נבחר מחלקת היפותזות שהיא קטנה נקטין את שגיאת השיערוך אך ייתכן כי נגדיל את שגיאת הקירוב שכן זה עלול להוביל ל- underfitting.

ה learner מתמודד עם בעיה של בחירת tradeoff טוב בין שני השיקולים, הboosting מסייע לו בבחירת מחלקת היפותזות כך ששגיאת הקירוב ושגיאת השערוך לא יהיו גדולות. AdaBoost מאפשר לנו לשלוט ב tradeoff בין שגיאות הקירוב והשערוך על ידי שינוי פרמטר יחיד - T מספר המסווגים.

השימוש בboosting מאפשר לlearner להתחיל עם מחלקה בסיסית שעבורה ייתכן כי שגיאת הקירוב תהיה גדולה, וככל שאנו מתקדמים באיטרציות ומספר המסווגים גדל, מחלקת ההיפותזות תהיה עשירה יותר. כפי שאנו רואים בגרפים ניתן לראות שהחל מT מסויים אין שיפור משמעותי.

מהגרף נסיק כי **עבור דאטא ללא רעש** הגדלה של מספר המסווגים T תגרום להפחתה ב bias ותשמר variance נמוך. מכאן ששגיאת mse תהיה קטנה יותר ככל שT גדול יותר.
ייתכן שעבור T גדול מאוד (T=500) השיפור מבחינת הbias יהיה מינורי (הפחתה לא משמעותית) בהשוואה ל variance. בהתאם החל מT=200 ועד T=500 אנו רואים בגרף של השגיאה כי יש התייצבות ואין שינוי משמעותי בירידה של השגיאה.

במקרה של **דאטא עם רעש 0.01** עבור T קטן יותר תתקבל שגיאה קטנה יותר לעומת דאטא ללא רעש בו T גדול יותר ייתן שגיאה קטנה יותר. במקרה של דאטא עם רעש 0.01 השגיאה המינימלית תתקבל עם 72 מסווגים לעומת 105 בדאטא ללא רעש.
במקרה של **דאטא מורעש מאוד עם 0.4** אנו רואים שיפור מבחינת סט האימון עד 150 ~T והטסט סט עד T~59, מבחינת הטסט סט יש עלייה בשגיאה החל מ T>59 .
כאשר הדאטא רועש שימוש בT גדול יגרום ל **overfitting** ושגיאת ההערכה תגדל כך שהשגיאה המיוחסת לטסט סט תהיה גדולה ואנו רואים אף עלייה בשגיאה.

שגיאת הקירוב - The Approximation Error

This term measures **how much risk we have because we restrict ourselves to a specific class**, namely, how much inductive bias we have. The approximation error does not depend on the sample size and is **determined by the hypothesis class chosen**. Enlarging the hypothesis class can decrease the approximation error.

שגיאת השערוך - The Estimation Error

The difference between the approximation error and the error achieved by the ERM predictor. The estimation error results because the empirical risk (i.e., training error) is only an estimate of the true risk, and so the predictor minimizing the empirical risk is only an estimate of the predictor minimizing the true risk.