בס"ד

IML
Ex4

1) נוכיח שהביטויים שקולים: ולכן נתחיל בהנחה ש(a) נכון ונוכיח את (b).

$\Leftarrow$ כלומר, לכל $\varepsilon,\delta>0$, קיים $m(\varepsilon,\delta)$ כך שלכל $m \geq m(\varepsilon,\delta)$:

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

ולכן נבחר λ כך שנרצה להוכיח שקיים $m_0 \in \mathbb{N}$ כך שעבור $m>m_0$ מתקיים $E_{S\sim D^m}\left[L_D(A(S))\right] \leq \lambda$.

לפי (a) נוכל לבחור $\delta = \varepsilon = \min\left(1/2, \lambda/2\right)$ ואז $m_0 = m(\varepsilon,\delta)$

ונקבל:

$$E_{S\sim D^m}\left[L_D(A(S))\right] \leq \mathbb{P}\left(L_D(A(S)) > \frac{\lambda}{2}\right)\cdot 1 + \underbrace{\mathbb{P}\left(L_D(A(S)) \leq \frac{\lambda}{2}\right)}_{=1}\cdot\frac{\lambda}{2}$$

$$\leq \mathbb{P}\left(L_D(A(S)) > \frac{\lambda}{2}\right) + \frac{\lambda}{2}$$

$$\leq \mathbb{P}\left(L_D(A(S)) > \varepsilon\right) + \frac{\lambda}{2}$$

$$\text{(נובע מ-(a))} \leq \delta + \frac{\lambda}{2} \leq \frac{\lambda}{2}+\frac{\lambda}{2} = \lambda$$

וקיבלנו מהגדרת הגבול הרגיל: $\lim_{m\to\infty} E_{S\sim D^m}\left[L_D(A(S))\right] = 0$.

$\Rightarrow$ ועכשיו נניח ש(b) מתקיים ונוכיח שגם (a) מתקיים:

כלומר מתקיים $\lim_{m\to\infty} E_{S\sim D^m}\left[L_D(A(S))\right] = 0$

נשים לב שלכל $\geq 1-\delta$: $\mathbb{P}_{S\sim D^m}\left[L_D(A(S)) > \varepsilon\right] < \delta = \mathbb{P}_{S\sim D^m}\left[L_D(A(S)) \leq \varepsilon\right] \geq 1-\delta$

ולכן נשתמש באי' שוויון של Markov (כמו הנתון):

$$\mathbb{P}_{S\sim D^m}\left[L_D(A(S)) > \varepsilon\right] \leq \frac{E_{S\sim D^m}\left[L_D(A(S))\right]}{\varepsilon} \leq \frac{\varepsilon\delta}{\varepsilon} = \delta$$

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

בס״ד

$h_r(x) = 1[\|x\|_2 \le r]$, $Y = \{0,1\}$, $X = \mathbb{R}^2$ (2

$H = \{h_r : r \in \mathbb{R}_+\}$

כאשר נתון דגימה $S \in \mathbb{R}^2 x y$ ונחפש $h_{\hat{r}} \in ERM_H$

כך ש- $\hat{r}$ מינימלי, כאשר יש לה:

$$\begin{cases} ERM_H : S \Rightarrow \operatorname{argmin}_{h \in H} L_S(h) \\ L_S(h) = \frac{1}{m} |\{i : h(x_i) \neq y_i\}| \end{cases}$$

מההנחת הריאליזביליות, נובע כי קיים מעגל עם בפנים רק נקודות עם ערך 1 ומחוץ למעגל נקודות עם ערך 0.

ולכן נגדיר היפותזה h^* שמייצגת את המעגל הדיק ביותר שמכיל את כל המקרים החיוביים כלומר $L_{D,f}(h^*) = 0$

ו- r^* רדיוס של מעגל הזה.

נגדיר גם את E, השטח בין המעגל עם רדיוס r לבין המעגל עם רדיוס r^*: $E = \{x \in \mathbb{R}^2 \mid r^* \le \|x\|_2 \le \hat{r}\}$

ולכן, כאשר נמצאים באזור של E, נקבל:

$$D(x \in \mathbb{R} \mid r^* \le \|x\|_2 \le \hat{r}) \ge \varepsilon$$

הרצאה $\Rightarrow P(L_{D,f}(h_{r^*}) \le \varepsilon) \ge 1-\delta$

$\Rightarrow P(L_D(A(S)) \le \varepsilon) \ge 1-\delta$

מהשיעור, ידוע שלפי תנאים של למידה PAC, מתקיים:

$0 \le \hat{r} \le r$

אם $r^* \le \hat{r}$, נקבל $D(x \in \mathbb{R} \mid r^* \le \|x\|_2 \le \hat{r}) \le \varepsilon$ ולכן עומדים בתנאים.

אחרת, אזי $\hat{r} \ge \|x\|$ וגם $r < \|x\|$ ולכן הסתברות שהנקודה לא ב-E היא $1-\varepsilon$.

רמז

$\Rightarrow (1-\varepsilon)^m \le e^{-\varepsilon m} \le \delta$

$\Rightarrow \log e^{-\varepsilon m} \le \log \delta$

$\Rightarrow -\varepsilon m \le \log \delta \quad \Rightarrow m \ge \frac{-\log \delta}{\varepsilon} = \frac{\log(1/\delta)}{\varepsilon}$

$\Rightarrow m_H(\varepsilon, \delta) \le \frac{\log(1/\delta)}{\varepsilon}$

בס"ד

(3 נוכיח ש: $d = VC_{dim}(H_{con})$ ונגדיר $E=\{e_1,\dots,e_d\}$

כאשר $e_i \in \mathbb{R}^d$ וקטורי יחידה.

כאשר נבחר $h(x)=1$ אזי לקבוצה יש ערך 1

וכאשר נבחר $h(x) = x_1 \wedge \bar{x_1}$, יש ערכים להקבוצה.

נבחר ב- $h(x)=\bigwedge_{j\in J} x_j$ כאשר $J\subseteq[d]$ ונקבל $\{e_j\}_{j\in J}$

נשים לב $h(e_j)=1$.

$$\forall j \quad h_J(x_j) = \begin{cases}1 & j\in J\\ 0 & j\notin J.\end{cases}$$

נניח בשלילה שקיים $C=\{c_1,\dots,c_{d+1}\}$ כך ש: $C\subseteq X$

ו- $|C|=d+1$ שמתנפץ על ידי H.

תהי $h_1,\dots,h_{d+1}\in H^d_{con}$ כך ש:

$$\forall i,j\in[d+1],\quad h_i(c_j)=\begin{cases}0 & i=j\\ 1 & i\neq j\end{cases}$$

ולכן לכל $h_i, i\in[d+1]$ קיים ליטרל l_i שמחזיר שקר על c_i ונאמת על c_j (לכל $i\neq j$)

מעקרון שובך היונים, קיים זוג l_i, l_j שמשתמשים באותו x_k כאשר $l_i=l_j$. כ"ש ($l_j = x_k$ וגם $l_i=x_k$)

או ($l_j=\bar{x_k}$ וגם $l_i=\bar{x_k}$)

$\Leftarrow$ l_i אמת על c_i (l_j אמת על c_j)

$\leftarrow$ וזו סתירה להגדרה של l_i.

$\Leftarrow$ $l_i\neq l_j$ $\Leftarrow$ $k\neq i\neq j$ כך ש: $h_i(c_k)=0$

וזו סתירה להגדרה של h_i ■

4) לכל $\varepsilon,\delta\in(0,1)$, כאשר S דגימה מגודל m כך ש: $m \geq m^{UC}(\frac{\varepsilon}{2},\delta)$ אז דגימה S בהסתברות לפחות $1-\delta$ וגם S $\varepsilon/2$-Representative. כלומר לכל $h\in H$:

$$\begin{cases} L_D(h) \leq L_S(h) + \varepsilon/2 \\ L_D(h) \leq \min_{h'\in H} L_D(h') + \frac{\varepsilon}{2} \leq \min_{h'\in H} L_D(h') + \varepsilon \end{cases}$$

כאשר H agnostic PAC עם $m_H(\varepsilon,\delta) = m^{UC}(\frac{\varepsilon}{2},\delta)$

$$\Rightarrow \mathbb{P}_{S\sim D}\left[L_D(h) \leq \min_{h'\in H} L_D(h') + \varepsilon\right] \geq 1-\delta$$

$$\Rightarrow \underbrace{\mathbb{P}_{S\sim D}\left[L_D(h) \leq \min_{h'\in H} L_D(h') + \varepsilon\right]}$$

$$\geq \mathbb{P}_{S\sim D^m}\left[\{S\in(X,Y)^m \mid S \text{ is } \tfrac{\varepsilon}{2} \text{ representative}\}\right]$$

$$\geq 1-\delta$$

כאשר $m \geq m_H^{UC}(\frac{\varepsilon}{2},\delta)$.

ד) ידוע $H_1 \subseteq H_2$ ולכן לכל $C=\{c_1,\dots,c_n\}\subseteq X$ מתקיים $H_{1C}\subseteq H_{2C}$.
אז C מתנפץ (shattered) ע"י H_1 אז גם C מתנפץ ע"י H_2.
כלומר $VCdim(H_1) \leq VCdim(H_2)$

Scanné avec CamScanner

בס"ד

8) נגדיר $\tau_H(m) = \max\{|H_C| \mid C \subseteq X, |C| = m\}$

(a) ב' לנו $\tau_H(m)$ שמייצג את גודל הגדול ביותר (max) של קבוצת ההיפותזות H המתנפצת ע"י C.

(b) אם $VCdim(H) = \infty$, אין הגבלה ל-$VCdim(H)$. 2^m כלומר $\tau_H(m)$ מקבל את גודל מקסימלי כלומר גודל

(c) אם $VCdim(H) = d$ כאשר $m \geq d$ אזי $\tau_H(m) = 2^m$. (H היא קבוצה של כל פונק' $C \to \{0,1\}$)

(d) (i) $|H_C| \leq |\{B \subseteq C \mid H \text{ is shatters } B\}|$
נוכיח את הטענה על $m \in \mathbb{N}$ באינדוקציה:

בסיס: $m = 0$, ולכן קבוצה ריקה כלומר הטענה מתקיימת באופן ריק.

הנחה: נניח שהטענה נכונה עבור $m-1$ ונוכיח זאת עבור m.

צעד: נגדיר:

$$y_1 = \{(y_2, \dots, y_m) \mid (0, y_2, \dots, y_m) \in H_C \vee (1, y_2, \dots, y_m) \in H_C\}$$
$$C_1 = (c_1, \dots, c_m)$$
$$y_2 = \{(y_2, \dots, y_m) \mid (0, y_2, \dots, y_m) \in H_C \wedge (1, y_2, \dots, y_m) \in H_C\}$$
$$C_2 = (c_2, \dots, c_m)$$

כלומר יש לנו: $|H_C| = |A_1| + |A_2|$

ונדע ש $H_{C'}$ ול-y_0 יש אותם וקטורים, ומהנחת האינדוקציה:

$$|y_1| = |H_{C'}| = |\{B \subseteq C' \mid H \text{ shatters } B\}|$$

ונגדיר H' שמתנפצת ל-C':

$$H' = \{h \in H \mid \exists h' \in H, (1 - h'(c_1), h_2'(c_2), \dots h_m'(c_m))\} = \{h_1(c_1), \dots, h_m(c_m)\}$$

מכיוון ש $|c'| < m$, מהנחת האינדוקציה מתקיים!

$$|Y_1| = |H_{C'}| \leq |\{B \subseteq C' \mid H' \text{ shatters } B\}|$$

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

$$\leq |\{B \subseteq C \mid c_1 \in B \wedge H \text{ shatters } B\}|$$

$$\Rightarrow |H_C| = |Y_0| + |Y_1|$$

$$\leq |\{B \subseteq C \mid c_1 \notin B \wedge H \text{ shatters } B\}|$$
$$+ |\{B \subseteq C \mid c_1 \in B \wedge H \text{ shatters } B\}|$$

$$\Rightarrow |H_C| \leq |\{B \subseteq C \mid H \text{ shatters } B\}| \quad \blacksquare$$

(ii) $VCdim(H) = d$ כלומר d גודל מקסימלי של תת קבוצה של C כך ש H מנפצת אותה. (H הוא הסט מכיון ל H_C)

(iii) מתקיים לכל $C \subseteq \mathcal{X}$:

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

$$\Rightarrow |\{B \subseteq C \mid H \text{ shatters } B\}| \subseteq \{B \mid B \subseteq C \wedge |B| \leq d\}$$

כלומר קיבלנו שאיבר האחרון קטן שווה ל- $\sum_{k=0}^{d} \binom{m}{k}$

ידוע לנו ש- $VCdim(H) = d$ כלומר אין קבוצה מגודל גדול מ-d שמנותפצת ע"י H

$$\Rightarrow |\{B \subseteq C \mid H \text{ shatters } B\}| \leq \sum_{k=0}^{d} \binom{m}{k}$$

(iv) וידוע ש- $\tau_H(m) = \max\{|H_C| \mid C \subseteq \mathcal{X}, |C| = m\}$

$$\leq \max_B |\{B \subseteq C \mid H \text{ is shattering } C\}|$$

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

ולכן מ· (iii) נקבל:

$$\tau_H(m) \le \sum_{k=0}^{d}\binom{m}{k} \le \frac{(em)^d}{d}$$

מקסימאלית

$$\Rightarrow \tau_H(m) \le \frac{(em)^d}{d}$$

(e) עבור $m=d$, מתקיים: $\tau_H(m) = 2^m = 2^d$

$$\Rightarrow \tau_H(m) = 2^d \le e^d \overset{m=d}{=} \frac{(em)^d}{d}$$

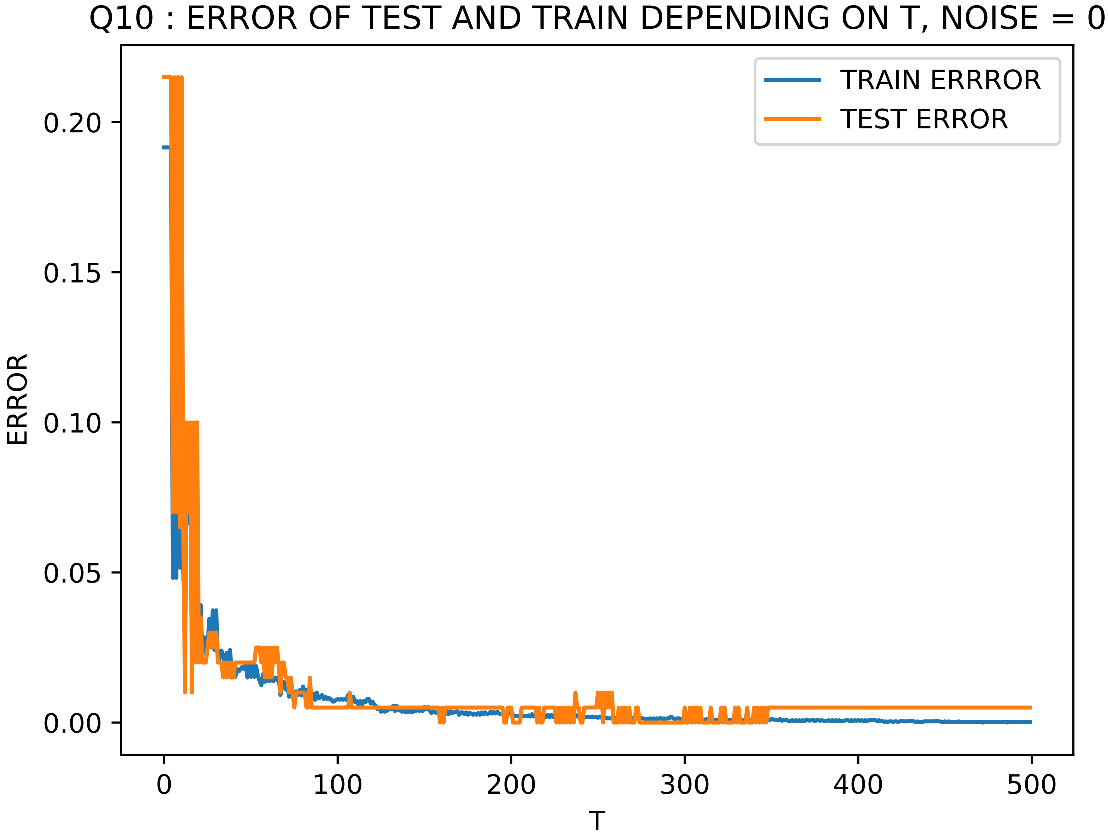
(f) (*) $m > VCdim(H)$:

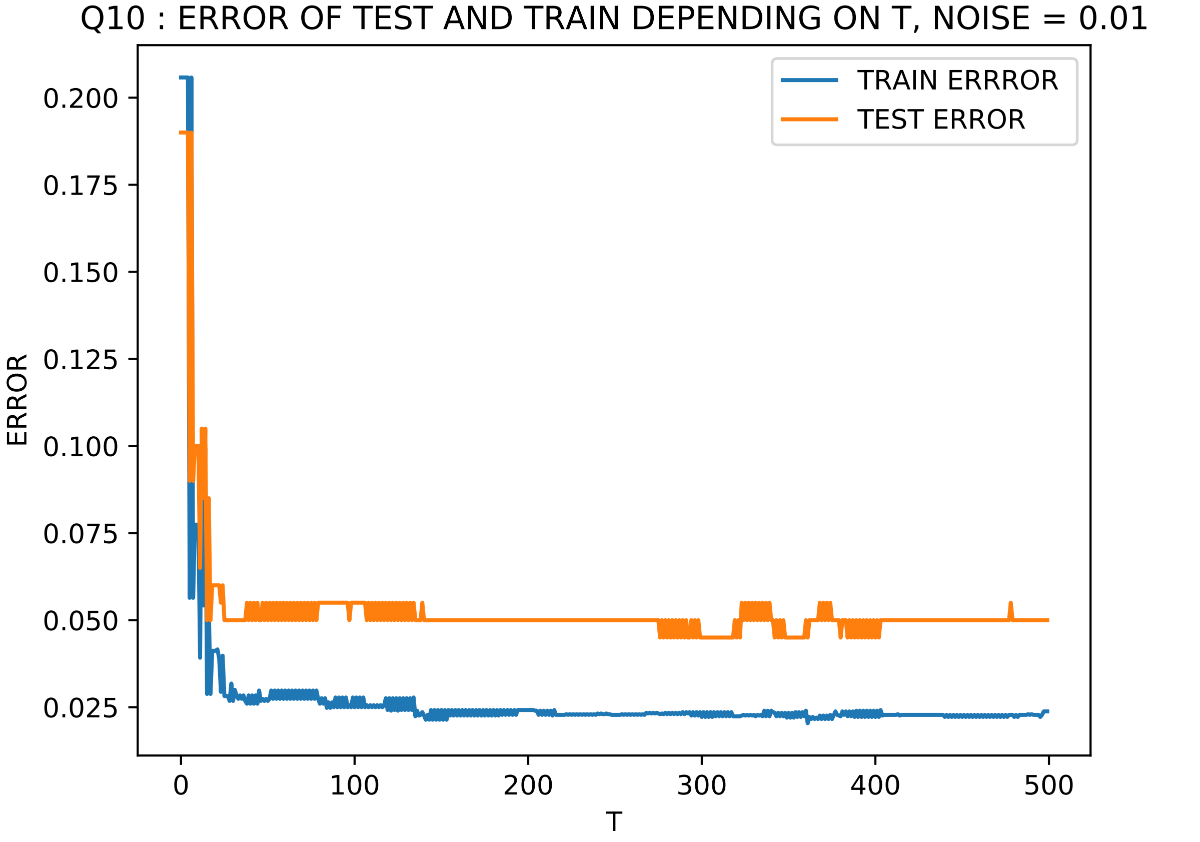
$$VCdim(H) = \max\{d \mid \exists C \subseteq X \mid |C| = d \lor \tau_H(m) \le \frac{(em)^d}{d}\}$$

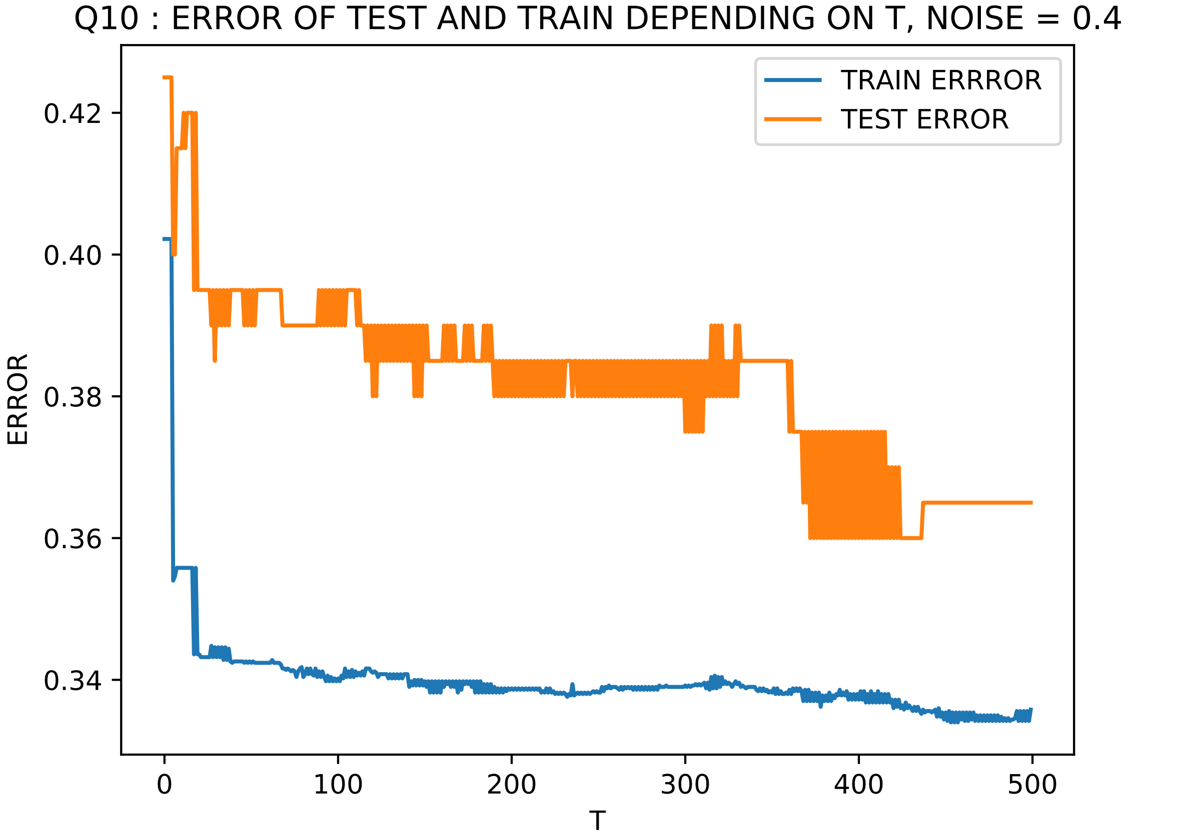
$$\boxed{\tau_H(m) \le \left(\frac{em}{d}\right)^d}$$

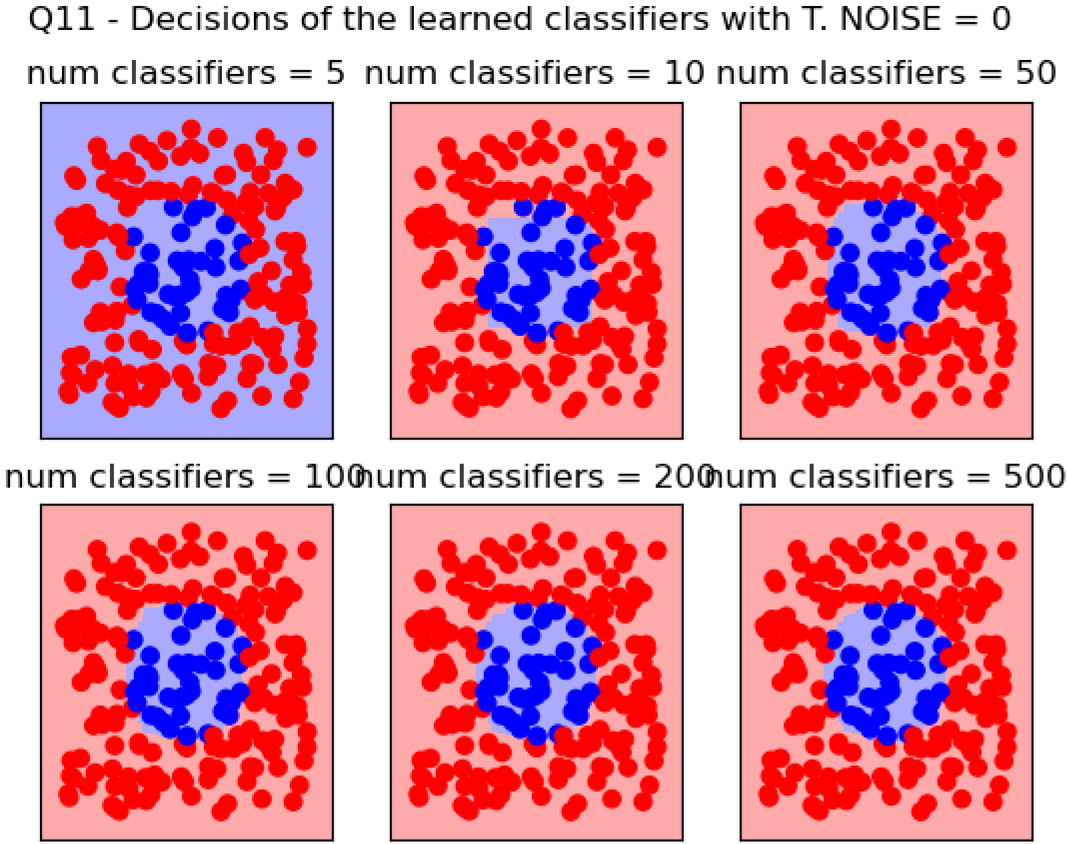
(+) $m \le VCdim(H)$:

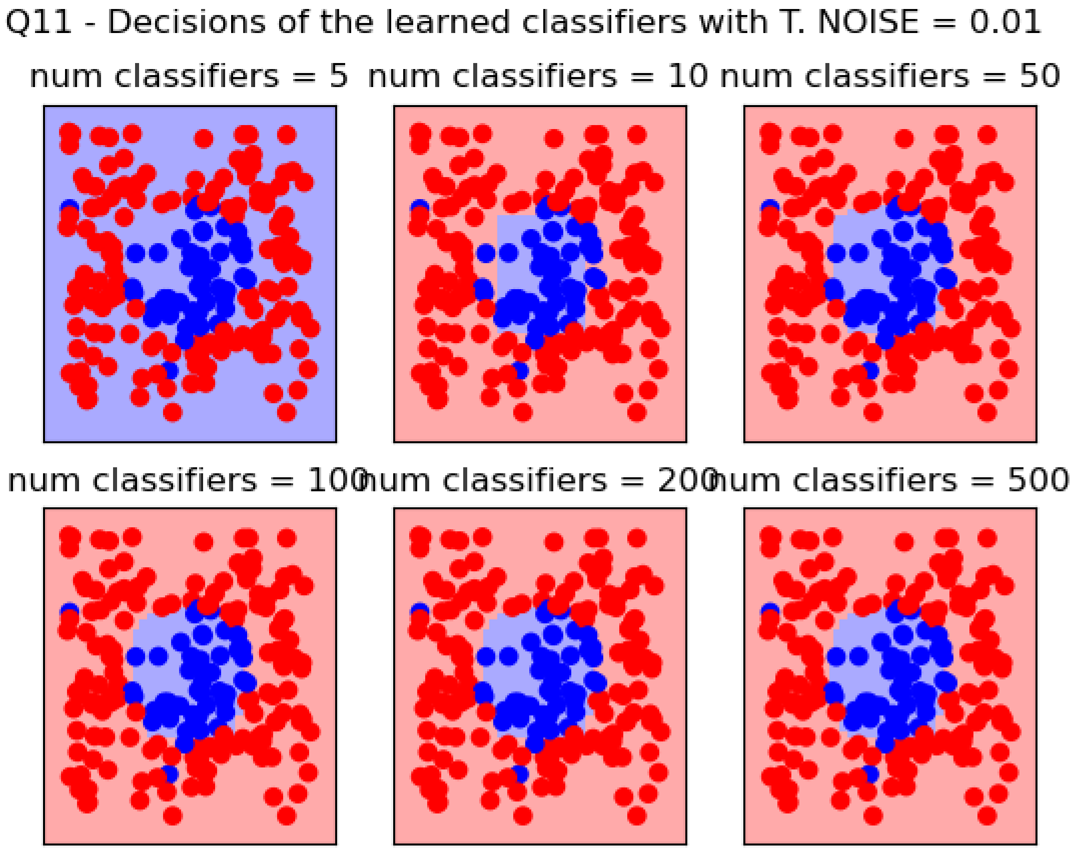
$$\tau_H(m) = 2^m$$

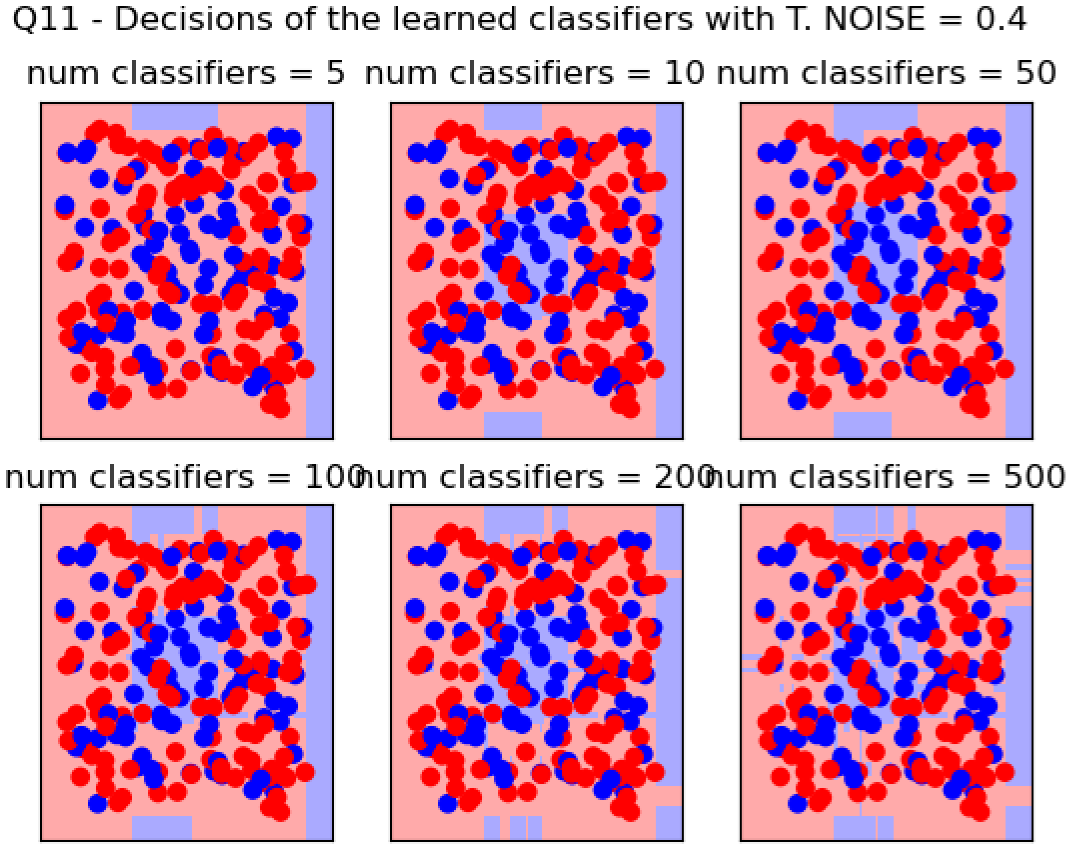

Q10 : ERROR OF TEST AND TRAIN DEPENDING ON T, NOISE = 0
TRAIN ERRROR
TEST ERROR
ERROR
0.00
0.05
0.10
0.15
0.20
0
100
200
300
400
500
T


Q10 : ERROR OF TEST AND TRAIN DEPENDING ON T, NOISE = 0.01
TRAIN ERRROR
TEST ERROR
ERROR
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0
100
200
300
400
500
T

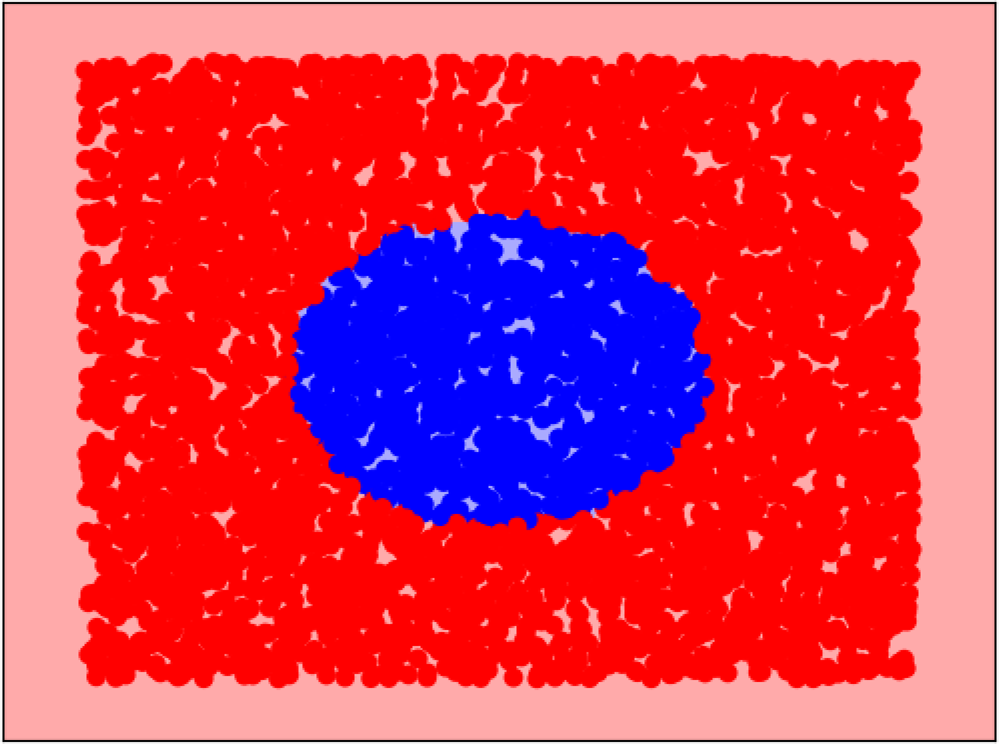

Q10 : ERROR OF TEST AND TRAIN DEPENDING ON T, NOISE = 0.4
TRAIN ERRROR
TEST ERROR
ERROR
0.42
0.40
0.38
0.36
0.34
0
100
200
300
400
500
T


Q11 - Decisions of the learned classifiers with T. NOISE = 0
num classifiers = 5
num classifiers = 10
num classifiers = 50
num classifiers = 100
num classifiers = 200
num classifiers = 500

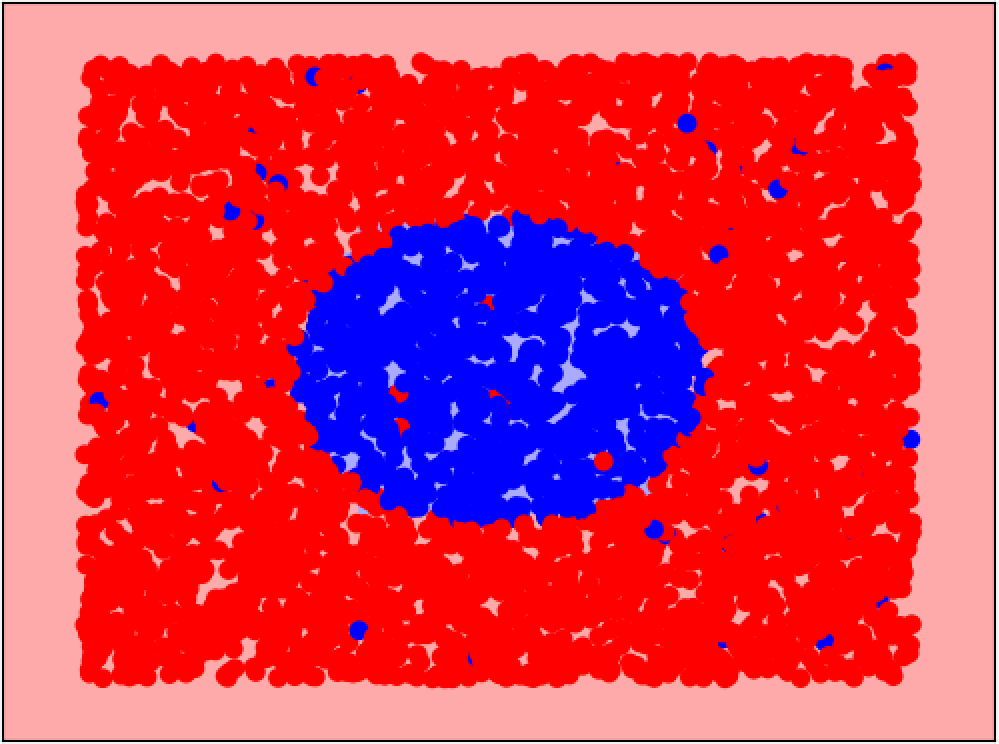

Q11 - Decisions of the learned classifiers with T. NOISE = 0.01
num classifiers = 5
num classifiers = 10
num classifiers = 50
num classifiers = 100
num classifiers = 200
num classifiers = 500


Q11 - Decisions of the learned classifiers with T. NOISE = 0.4
num classifiers = 5
num classifiers = 10
num classifiers = 50
num classifiers = 100
num classifiers = 200
num classifiers = 500

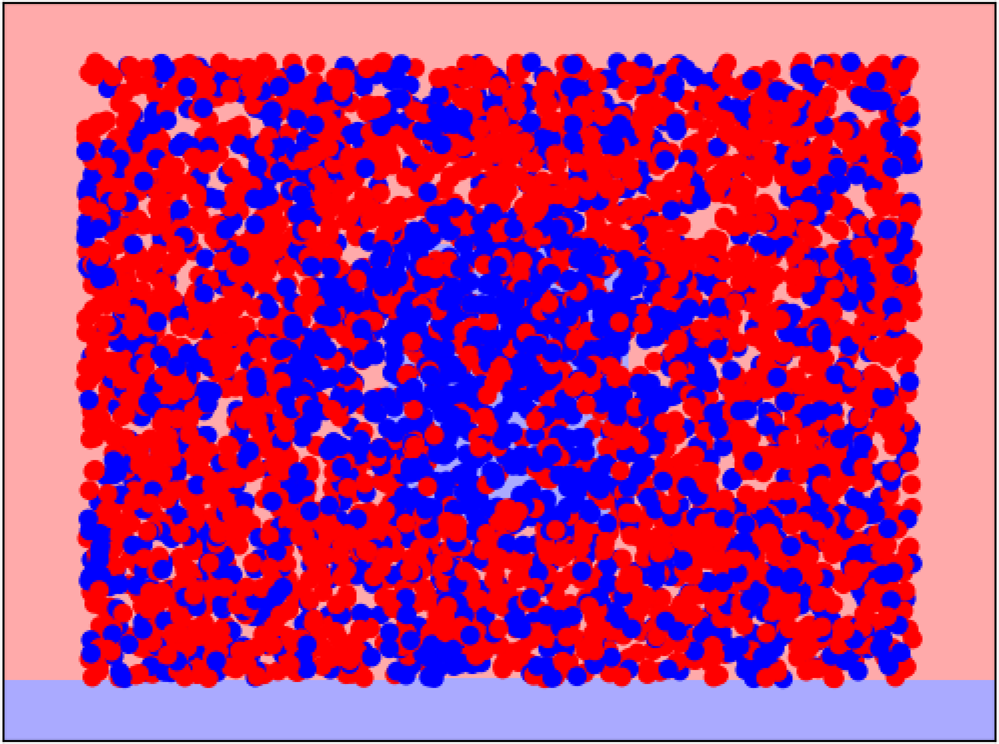
Q12 - Classifier with T = 200, NOISE = 0 and ERROR = 0.005



Q12 - Classifier with T = 50, NOISE = 0.01 and ERROR = 0.035

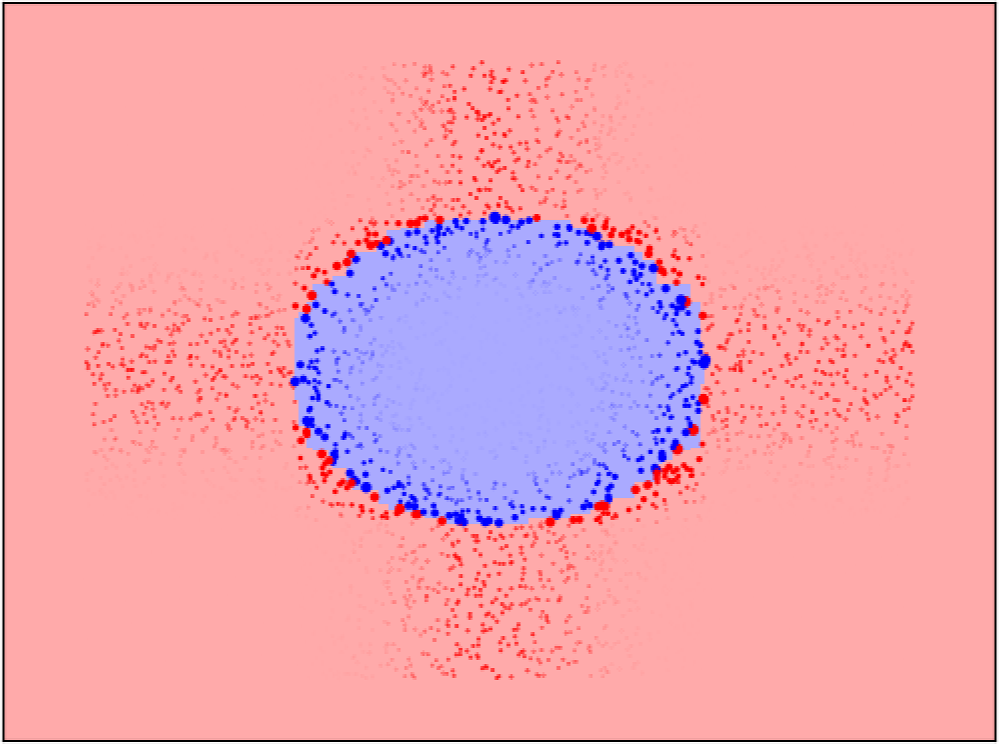


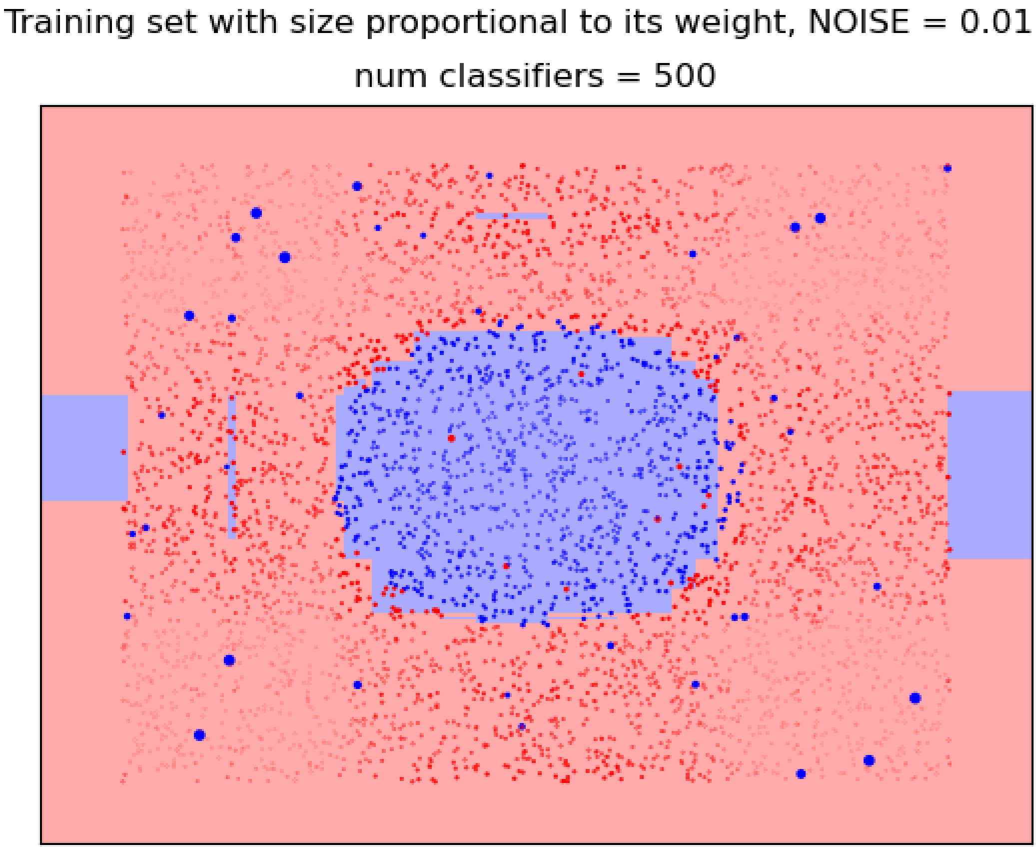
Q12 - Classifier with T = 10, NOISE = 0.4 and ERROR = 0.37



Training set with size proportional to its weight, NOISE = 0

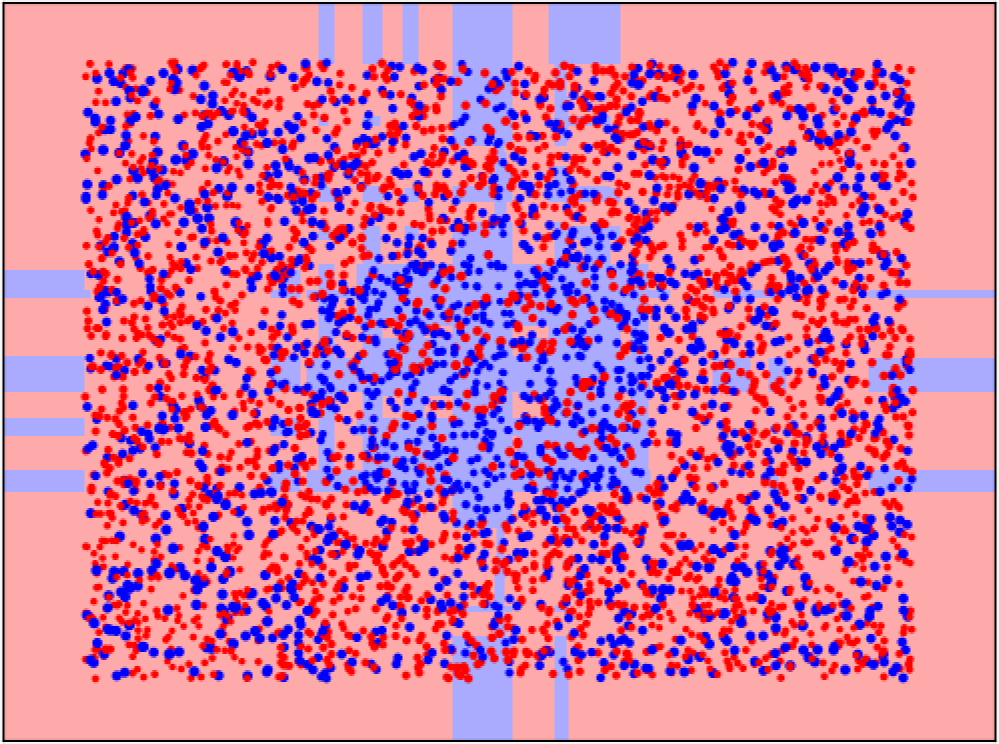
num classifiers = 500




Training set with size proportional to its weight, NOISE = 0.01
num classifiers = 500

Training set with size proportional to its weight, NOISE = 0.4

num classifiers = 500



13) מהגרפים שקיבלנו אפשר לשים לב שכאשר יש רעש (0 = Noise), יותר חזק את ההבדל במשקל של הנק'.
אפשר לראות שלנקודות שלא מקבלות ערך טוב שלהם (= falsely labeled) יש משקל יותר כבד (heavier weight).
כאמור שלנקודות שמקבלות ה label הנכון יש פחות משקל.

כמו כן, נשים לב שככל עוד שיש יותר רעש, המשקל של הנקודות בערך זהה משום שלכל נקודה יש [illegible] כלשהי) falsely labeled בגלל הרעש החשוב.

(14

QUESTION 10 (*)

מהגרפים, רואים שהרעש משפיע על ה - ERROR. כאשר הרעש יותר גדול אזי גם ה- ERROR יותר גדולה.

כאשר אין רעש: ה- train error - test error כמעט זהים, וה-T מקטין את ה- bias (אולי T = 200) ולכן שונות קטן (low variance).

כאשר רעש = 0.01: train error < test error, להוסיף רעש יגדיל את ה- variance, bias עדיין קטן אבל יש הגדלה של ה variance

כאשר רעש = 0.4: הוספנו הרבה רעש train error ≫ test error ושוב יש variance יותר גדולה מקודם

לסיכום, בגלל הרעש. ה variance יותר גדולה ולכן מגדילה את ההפרש בין ה- train error ל- test error

ספק

QUESTION 12

אפשר לראות מהגרפים שכאשר אין רעש, יש לנו ERROR-Rate
קטן ממש ולכן הקלסיפיקציה (classification) טובה. לכל T-
אבל כאשר מוסיפים רעש, ה- ERROR עולה ממש ויש לנו
נקודות שלא מקבלים label הנכון.
(בגלל שעם רעש ה decision boundaries מתחילות להיות יותר
קשות להפרדה).
ויותר יש רעש יותר יש נקודות עם label לא נכון.

QUESTION 12

מהגרפים רואים שכאשר מוסיפים רעש יש לנו נקודות מצבע
אחד שהיו בירוק על צבע אחר, כלומר יותר רעש יש.
יותר נקודות מקבלים label הלא נכון
רואים מגרף עם רעש= 0.4 שהטעויות הן בכמות גדולה ממש.

בנוסף, אפשר לשים לב ש- $\hat{T}$ ה T אופטימלי מקטין כאשר
הרעש יותר גדול.
(ידוע שאם T גדול יהיה הרבה overfitting וזה ישפיע
על ה noise ויגדיל אותה)
כי שכאשר יש רעש, נקבל פחות noise עם פחות overfitting.
ולכן נמנע מ overfitting.