I M 2
6 7 5 7 7
4 סמסטר
יולי 2021
3 ב 2021

① A learning algorithm, D a distribution
loss function in $[0,1]$

(a) $\forall \varepsilon,\delta>0 \ \exists m(\varepsilon,\delta)$ s.t $\forall m \geq m(\varepsilon,\delta)$

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

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

$\Leftarrow$
נניח: (b)

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

לפי הגדרת הגבול קיים $\varepsilon_1>0$ כך ש $\forall m \geq m(\varepsilon_1)$:

$$\mathbb{E}_{S\sim D^m}[L_D(A(S))] < \varepsilon_1$$

לכל $\varepsilon,\delta>0$ $\varepsilon_1=\delta\varepsilon$ אזי לפי מרקוב:

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

$$\Downarrow$$

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

$(a\Leftarrow b)$ $\quad \Leftarrow$ מ.ש.ל

$\Rightarrow$ נניח (a)

$\forall \varepsilon,\delta>0 \ \exists m(\varepsilon,\delta)$ s.t $\forall m \geq m(\varepsilon,\delta)$

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

$\lim\limits_{m\to\infty} \mathbb{E}_{S\sim D^m}[L_D(A(S))] \neq 0$ $\leftarrow$ נניח בשלילה:

אזי:

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

כי נתון אי-שלילי ש$L_D(A(S))$.

לפי הגדרת הגבול קיים $\varepsilon^*>0$ כך שלכל $m \geq m(\varepsilon,\varepsilon)$:

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

$$\Downarrow$$

$$\int_0^1 x\,\mathbb{P}_{S\sim D^m}[L_D(A(S))=x]\,dx > \varepsilon^* \qquad (*)$$

עבור $\varepsilon=(\varepsilon^*)^2, \ \delta=\frac{\varepsilon^*}{2}$ אזי:

$$\mathbb{P}_{S\sim D^m}[L_D(A(S))\leq(\varepsilon^*)^2] \geq 1-\frac{\varepsilon^*}{2} \qquad (**)$$

$$\Downarrow$$

מ-(*) $\Downarrow$

$$\epsilon^* < \int_0^1 x\,\mathbb{P}_{S\sim D^m}[L_D(A(S))]\,dx = \int_0^{\sqrt{\epsilon^*}} x\,\mathbb{P}_{S\sim D^m}[L_D(A(S))=x]\,dx +$$

$$+\int_{\sqrt{\epsilon^*}}^1 x\,\mathbb{P}_{S\sim D^m}[L_D(A(S))=x]\,dx \le$$

$$\le \int_0^{\sqrt{\epsilon^*}} \sqrt{\epsilon^*}\,\mathbb{P}_{S\sim D^m}[L_D(A(S))=\sqrt{\epsilon^*}]\,dx + \int_{\sqrt{\epsilon^*}}^1 1\cdot\mathbb{P}_{S\sim D^m}[L_D(A(S))=1]\,dx \le$$

$$< \int_0^{\sqrt{\epsilon^*}} \sqrt{\epsilon^*}\left(1-\frac{\epsilon^*}{2}\right)dx + \int_{\sqrt{\epsilon^*}}^1 \frac{\epsilon^*}{2}\,dx$$

מ-(**)

$$\epsilon^* < (\sqrt{\epsilon^*})^2\left(1-\frac{\epsilon^*}{2}\right) + \frac{\epsilon^*}{2}\left(1-\frac{\epsilon^*}{2}\right)(1-\sqrt{\epsilon^*}) <$$

$$\epsilon^* < \epsilon^*\left(1-\frac{\epsilon^*}{2}\right) + \frac{\epsilon^*}{2}(1-\sqrt{\epsilon^*}) < \epsilon^*$$

$\Downarrow$

סתירה

מכאן שלא תתכן הנחה (b) הם נכונה

לכן מש.ל $\Rightarrow$ $(a \Rightarrow b)$

הוכחנו את שני הכיוונים ולכן מתקיימת שקילות.

ב

② $X=\mathbb{R}^2$, $Y=\{0,1\}$ $\mathcal{H}$ - class of concentric circles:

$$\mathcal{H}=\{h_r : r\in\mathbb{R}^+\},\ h_r(x)=\mathbb{1}[\|x\|_2\le r]$$

נגדיר אלגוריתם ERM A שעבור קבוצת דגימות S מחזירה h_r, כך ש h_r הוא המעגל הקטן ביותר שמכיל את הדגימות עם תיוג חיובי. עבור מעגל ברדיוס r עד מעגלים בגודל r' המוגדר "רווח", נגדיר $0<\epsilon\le 1$

נסמן את השטח:

$$G=\{x\in\mathbb{R}^2 \mid r'<\|x\|\le r\}$$

$\Downarrow$

הסתברות שלא דגמנו נקודה ב G:

$$D(x\in\mathbb{R}^2 \mid 0<\|x\|\le r \Rightarrow D(x\in G)\le\epsilon$$

$\Downarrow$

$$\mathbb{P}_{S\sim D}(L_D(A(S))\le\epsilon) = \mathbb{P}(D(x\in G)\le\epsilon)=1\ge 1-\delta$$

מכאן H למידה ע"י PAC.

אם נתון ε $D(x\in\mathbb{R}^2 \mid 0\le\|x\|\le r)\ge\varepsilon$ אז $\exists\, 0\le r^*_\varepsilon\le r$

כך ש- $D(x\in\mathbb{R}^2 \mid r^*\le\|x\|\le r)\ge\varepsilon$. אם $r^*<r$ מינימלי

$D(x\in\mathbb{R}^2 \mid r^*_\varepsilon\le\|x\|\le r)\ge\varepsilon$.

לפי הגדרת r^* נמצא מחוץ להסתברות ε עבור והסתברות כ 1-ε

עבור $\forall x\in S$ $(\|x\|<r^*)\vee(r\le\|x\|)$ מתקיים שאנו מחזירים לכן

כבסיס לא נמצא ב-G בהסתברות ε-1.

ניקח $m_H(\varepsilon,\delta)=\frac{\log(1/\delta)}{\varepsilon}$ אם כך נקבל

$|S|\ge m_H(\varepsilon,\delta)$ מ-ד"ו

$$(1-\varepsilon)^{|S|}\le(1-\varepsilon)^{m_H(\varepsilon,\delta)}\le\exp\left(-\varepsilon\cdot\frac{\log(1/\delta)}{\varepsilon}\right)=\exp(\log\delta)=\delta$$

$$\mathbb{P}(D(x\notin G))\le\delta \qquad \forall x$$

לכן נאמר שהסתברות שהדגימות מחוץ ל-G קטנה מ-δ

כלומר ההסתברות לקבל שגיאה קטנה מ-ε גדולה או שווה ל-δ-1

ולכן מרכיבות הדגימות הסופית $\frac{\log(1/\delta)}{\varepsilon}$,

(ב) $H=\{h_1,\dots,h_N\}$

מהגדרה: $\mathrm{VCdim}(H)=\max\{|C| \mid H \text{ shatters } C\}\overset{\text{סימון}}{=}n$

כלומר קיימות לפחות 2^n היפותזות שונות ב-H.

$$|H|\ge 2^n \Leftrightarrow n=\mathrm{VCdim}(H)\le\log_2(|H|)$$

(ג) $X=\{0,1\}^n$, $Y=\overset{\text{מ.ל.ה}}{\{0,1\}}$, עם $I\subseteq[n]$: $h_I(x)=\left(\sum_{i\in I}x_i\right)\bmod 2$

$H_{parity}=\{h_I \mid I\subseteq[n]\}$.

$|H_{parity}|=2^n$ כיוון שכל אחת מוגדרת לכל תת קבוצה של $[n]$

מש"ל ב $\mathrm{VCdim}(H_{parity})\le\log(|H_{parity}|)=\log_2(2^n)=n$

5

$k \in \mathbb{Z}$, $([a_i,b_i])_{i=1}^{k}$ קבוצה של אינטרוולים על $\mathbb{R}$

$A=\bigcup_{i=1}^{k}[a_i,b_i]$. $\mathcal{H}_{k\text{-intervals}} \subseteq \left(h_A(x)=\begin{cases} 0 & x\notin A \\ 1 & x\in A \end{cases} \right)$

נראה קודם [illegible] ונמצא כי [illegible] את הקבוצה הבאה:

$X=\{x_1,\dots,x_{2k}\}$, $x_1<x_2<\dots<x_{2k}$.

[illegible] עם 2^{2k} תיוגים אפשריים

נניח בשלילה כי קיים תיוג שלא ניתן להגיע אליו עבור כל [illegible]

(x_i,y_i) עבור $y_i=0$ [illegible] לא תיפסל בהגדרה. לכן [illegible] רק

$y_i=1$ עבור y_{i-1} [illegible] y_i . אם כמו שיש לנו

x_{i-1}, x_i בהכרח יש לנו קטע שעובר את הקטע $[x_{i-1},x_i]$.

אם הם שונים [illegible] x_{i-1}, x_i [illegible] קטע לכן נבחר את

הקטע ש x_{i-1} [illegible] עד ל- x_i .

[illegible] $\{1,0,1,\dots,0,1\}$ [illegible] עבור

[illegible] k , [illegible] ו[illegible]

$2k$ של $\mathcal{H}_{k\text{-intervals}}$ [illegible]

[illegible] עבור $2k+1$ [illegible]

[illegible] [illegible] מקודם:

$X=\{x_1,\dots,x_{2k+1}\}$

אם נבחר תיוג [illegible] $\{1,0,1,\dots\}$ [illegible]

[illegible] לכן לא ניתן [illegible] קטעים עבור התיוג

$VCdim(\mathcal{H}_{k\text{-intervals}}) = 2k$

אינו חסום: k

$\lim_{k\to\infty} VCdim = \lim_{k\to\infty} 2k = \infty$

(6)

Let $X=\{0,1\}^d$, $Y=\{0,1\}$, $d\geq 2$

נוכיח כי $VCdim(H_{con})=d$. עבור הבסיס הסטנדרטי:

$E=\{e_1 \dots e_d\}$

נראה שניתן לנפץ, לכל x_i קיים לסמן גם אפשר הבסיס:

$$h(x_i)=\begin{cases} x_i & y_i=1 \\ \overline{x_i} & o.w. \end{cases}$$

מצאנו קבוצה בגודל d ; $VCdim(H_{con})\geq d$

נותר להראות כי לא קיימת קבוצה מנופצת בגודל $d+1$ נניח בשלילה שקיימת מנופצת כזו $C=\{x_1,\dots,x_{d+1}\}$ (נקרא:)

$h_i(x_j)=\delta_{ij}$

משפט היונים אז יהיו שיהיו 2

אינדקסים לפחות המופיעים באותה קורדינטה x, ונמצאו של ...

באותו ... ונמחק ... כלומר x מפה את ...

... לכן d ... $d+1 > VCdim$

$\Downarrow$

$VCdim(H_{con})=d$

מ.ש.ל.

(8) H מתנהג באופן אחיד עם הסוף $m_H^{uc}:(0,1)^2\to\mathbb{N}$

צ"ל H למידה ב Agnostic-PAC עם מורכבות מדגם $m_H(\epsilon,\delta)\leq m^{uc}(\epsilon/2,\delta)$

אנחנו יודעים:

$\forall \epsilon,\delta>0, \forall D, \forall m\geq m_H^{uc}(\frac{\epsilon}{2},\delta)$

$\mathbb{P}_{S\sim D^m}[L_D(A(S))\leq \min_{h\in H}\{L_D(A(h))+\epsilon\}]$ מתקיים:

כיוון עבור S ... כן ... מינימום ...

$\mathbb{P}_{S\sim D^m}(L_D(A(S)))\leq \min_{h\in H}\{L_D(A(h))+\epsilon)$

נסמן $\epsilon/2$ ... δ ... לקבל:

$\mathbb{P}_{S\sim D^m}[L_D(A(S))\leq \min_{h\in H}(L_D(A(h))+\epsilon/2)]\geq D^m(\{\xi\in X\times Y \mid \xi$ is $\epsilon/2$ represents $\xi\})\geq 1-\delta$

מכאן מורכבות המדגם מוכח.

8) נתון $Z = X \times \{\pm 1\}$ ופונקציית loss function 0-1, נניח שקיימת פונ' m_H

שמקיימת גם התנאים מעל Z קיים אלגוריתם A שמקיים:

כאשר מפעילים את A על m_H דוגמאות i.i.d. מ-D

מובטחת הצלחה בהסתברות של לפחות $1-\delta$ (ההכללה):

$$L_D(h(S)) \leq \min_{h \in H} L_D(h) + \epsilon \qquad h(S): X \to \{\pm 1\}$$

נכון שזה על מחבר לפי agnostic PAC. נניח בשלילה שהוא כן,

אז קיים אלגוריתם A ופונ' $m_H(\epsilon,\delta)$ כך שלכל

$0 < \epsilon, \delta < 1$ ולכל התפלגות D מעל Z ולכל

$f \in H$ מגדירים באופן דטרמיניסטי:

$$D(y|x) = \begin{cases} 1 & y = f(x) \\ 0 & \text{אחרת} \end{cases}$$

נשים לב כי עם התפלגות מוגדרת כזו, ניתן לייצר ע"י הנחת D_x

לכל x ולכל $f \in H$, המשמעות היא ש-H היא

למידה לפי PAC הריאליזבילי. לכן הטענה שלנו נכונה. מ.ש.ל

9) לא מתקיימת הטענה עבור קטנים מספיק:

H היא learnable PAC ו-$\epsilon \in (0,1)$ $\quad 0 < \epsilon_1 \leq \epsilon_2 < 1$

ניח בשלילה שמתקיים כאשר $m_H(\epsilon_2,\delta) \leq m_H(\epsilon_1,\delta)$, ו-$m_H$ מוגדר.

אז המספר המינימלי של דוגמאות כך ש-H הוא למיד PAC.

$\epsilon_1 \leq \epsilon_2$, וגם $m_H(\epsilon_1,\delta) \geq m_H(\epsilon_2,\delta)$ לכן ניתן לקבל דיוק יותר

בסבירות עם יותר של דוגמאות, כי דורש הסתברות לכך ש-m_H

מוגדר מס' מינימלי של דוגמאות לקבלת דיוק זה.

סב ϵ $\epsilon \in (0,1)$ נניח בשלילה שקיימים δ כך ש- $0 < \epsilon_1 \leq \epsilon_2 < 1$

ומתקיים $m_H(\epsilon_1,\delta) < m_H(\epsilon_2,\delta)$ בסתירה להגדרה לעיל

למינימליות ההגדרה.

10) $H_1 \subseteq H_2$ אז $VCdim(H_1) \leq VCdim(H_2)$

מינימליות והגדרה הנתונה $VCdim(H_1) \leq VCdim(H_2)$. עבור כל קבוצה

של נקודות $C \subseteq X$, $|C| \leq VCdim(H_1)$ מתקיים ש-H_1 מנתצת

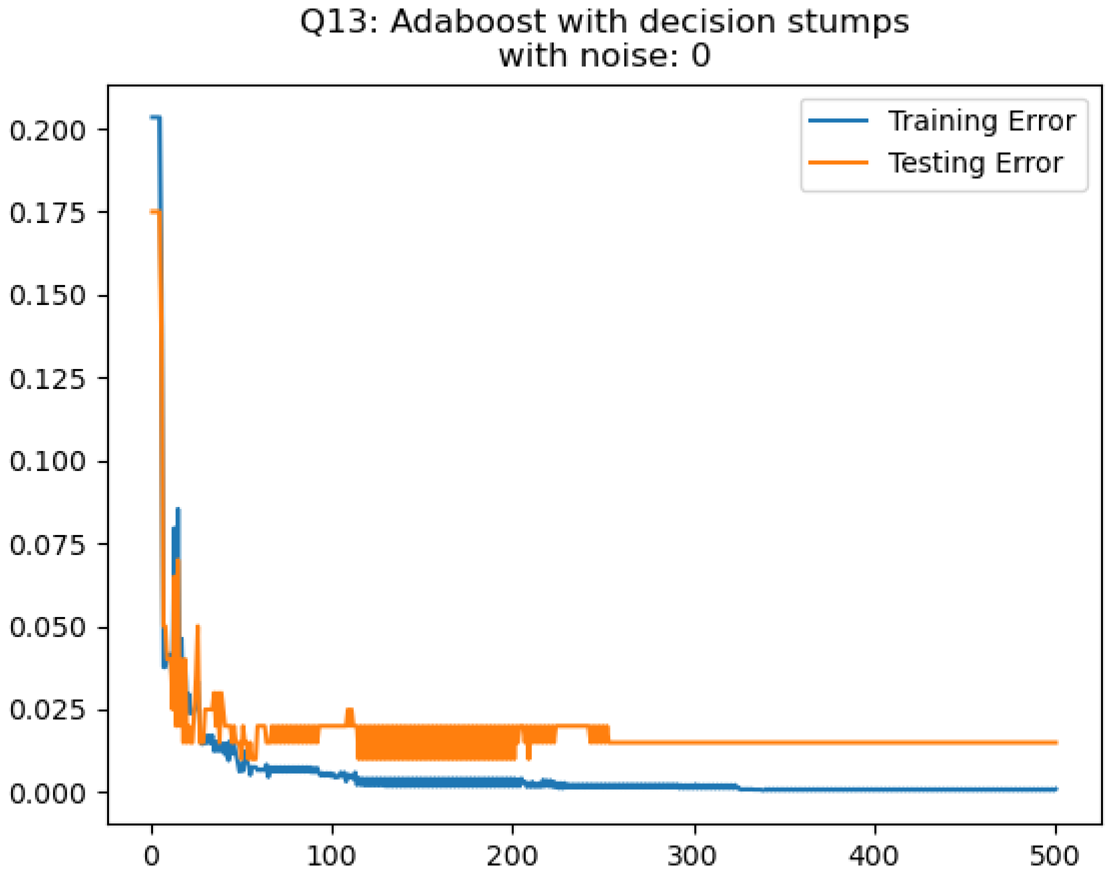
אותה אז C מנותצת גם ב-H_2, ומכאן מתקיים

$VCdim(H_1) \leq VCdim(H_2)$ מ.ש.ל

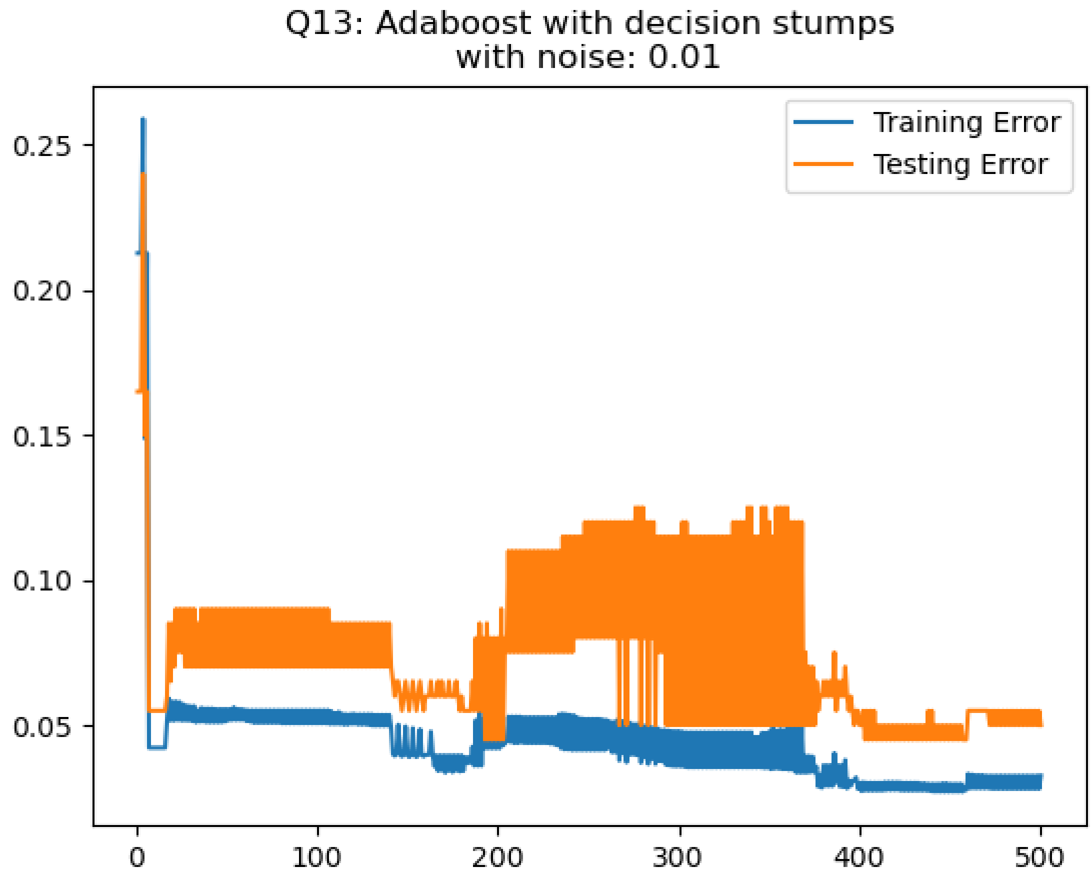
Code questions:

Q13

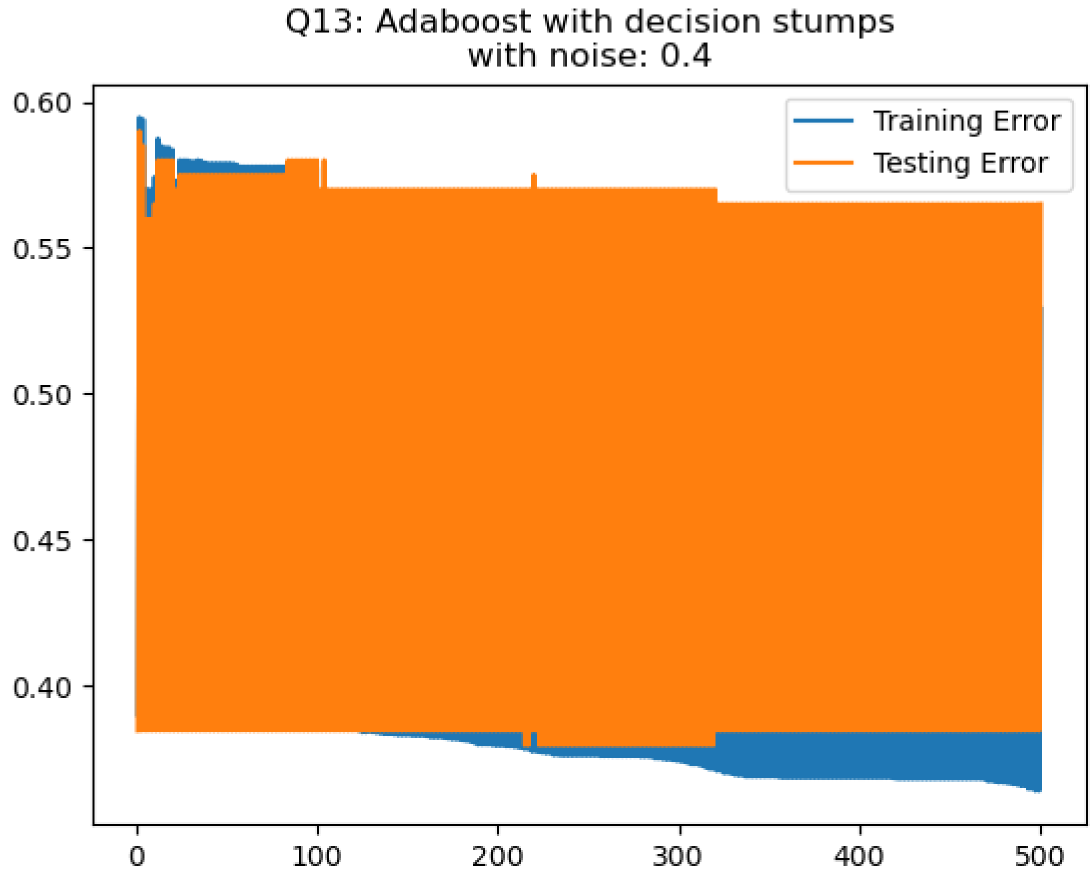
noise=0



noise= 0.01

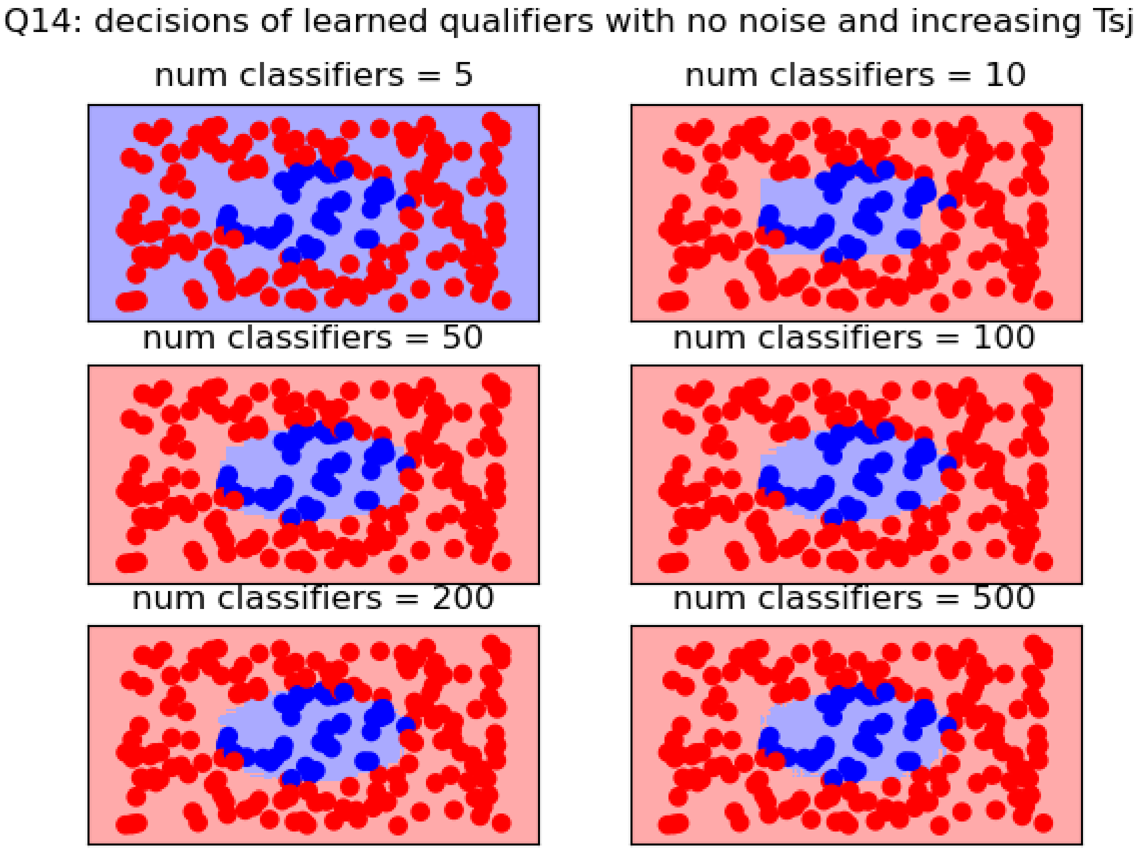

Q13: Adaboost with decision stumps
with noise: 0.01
Training Error
Testing Error
0.25
0.20
0.15
0.10
0.05
0
100
200
300
400
500

noise=0.4


Q13: Adaboost with decision stumps
with noise: 0.4
Training Error
Testing Error
0.60
0.55
0.50
0.45
0.40
0
100
200
300
400
500

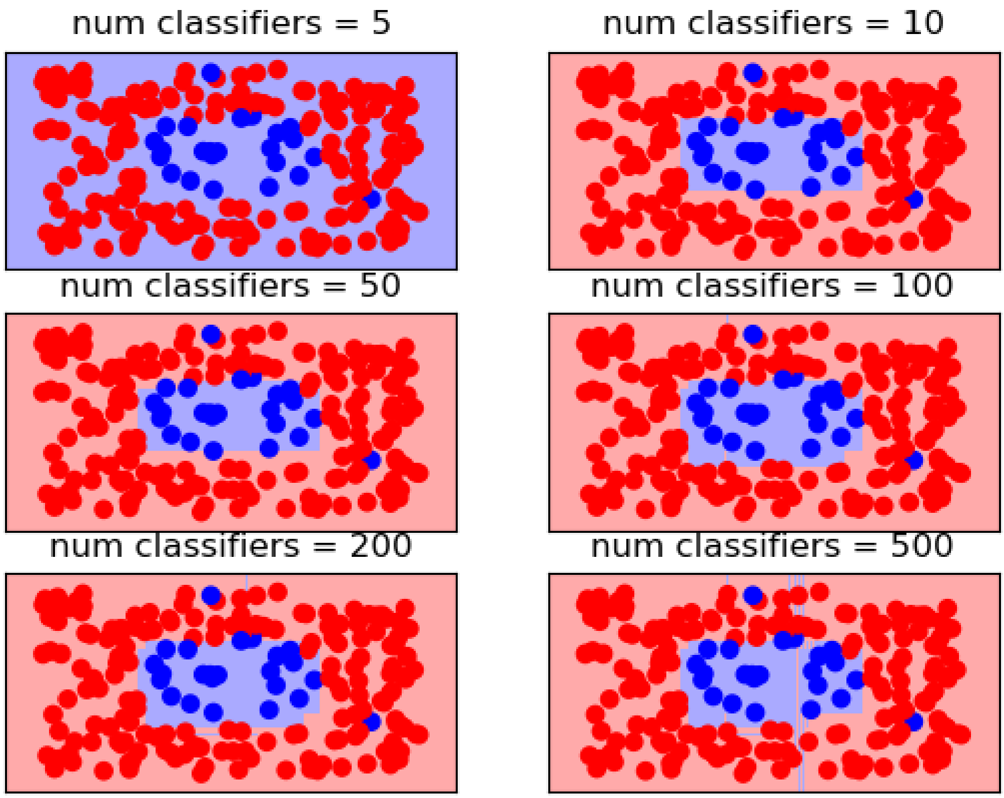
Q14

noise=0



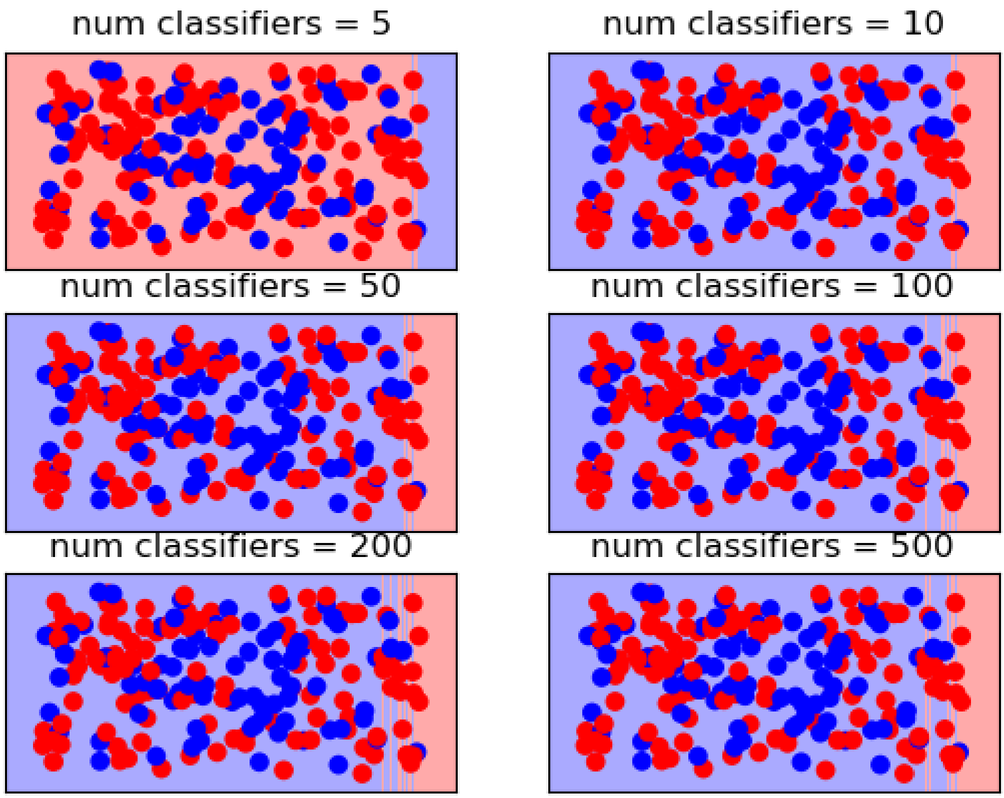
noise= 0.01

Q14: decisions of learned qualifiers with noise: 0.01, and increasing Ts



noise=0.4

Q14: decisions of learned qualifiers with noise: 0.4, and increasing Ts

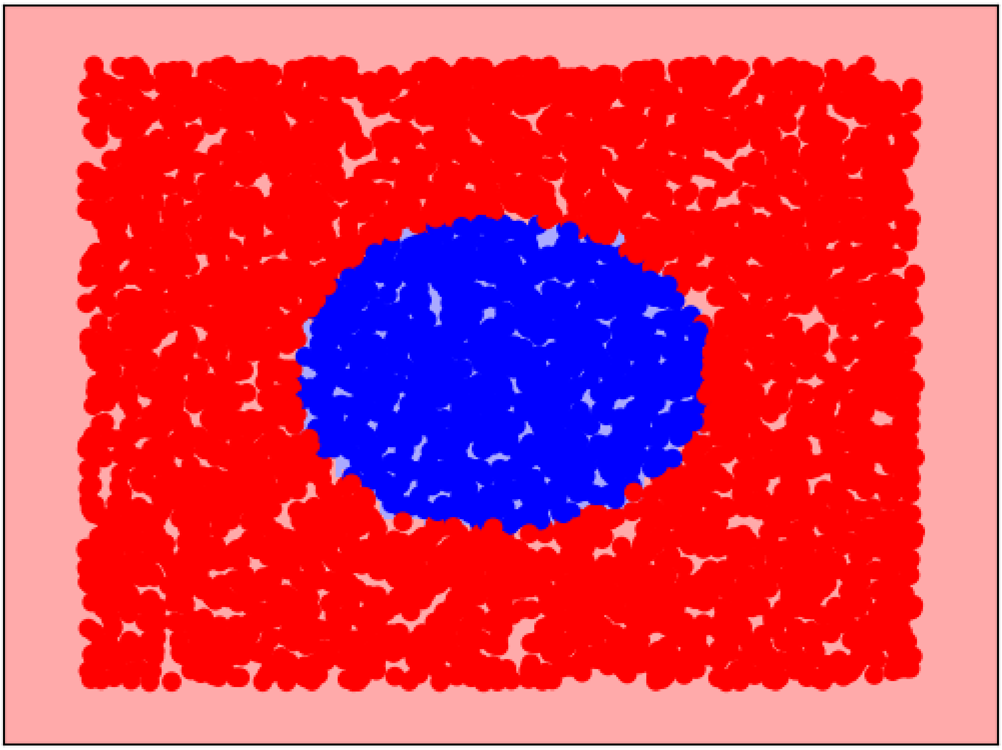


Q15

noise=0

Q15: T that minimizes error: 111, Error: 0.015

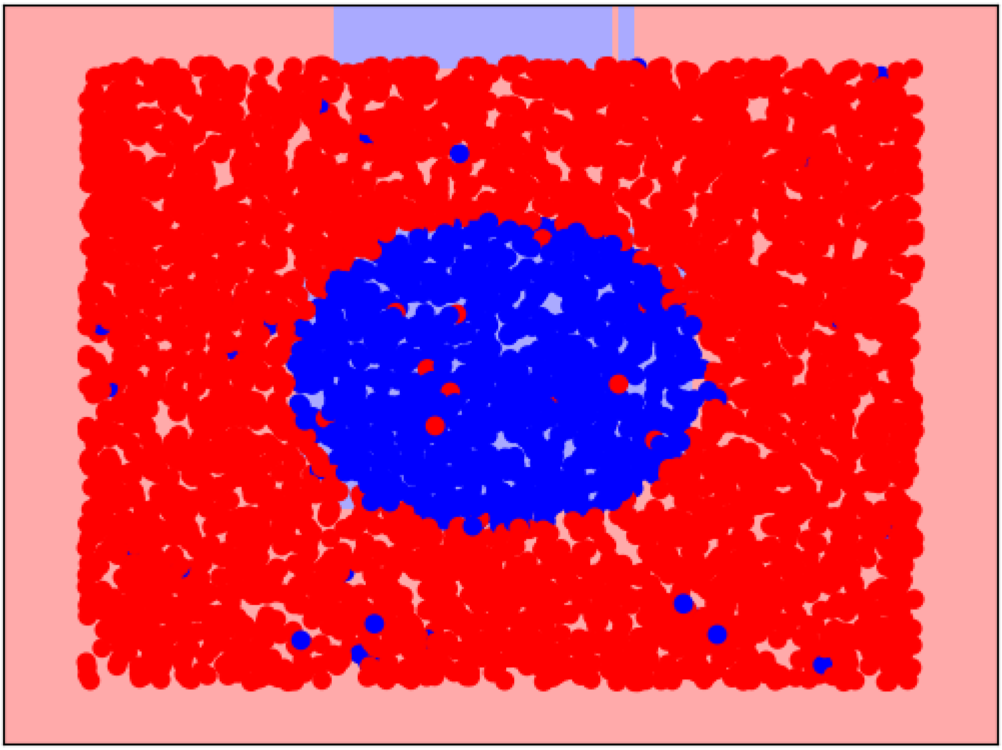
num classifiers = 111



noise= 0.01

Q15: T that minimizes error: 51, Error: 0.03, noise: 0.01

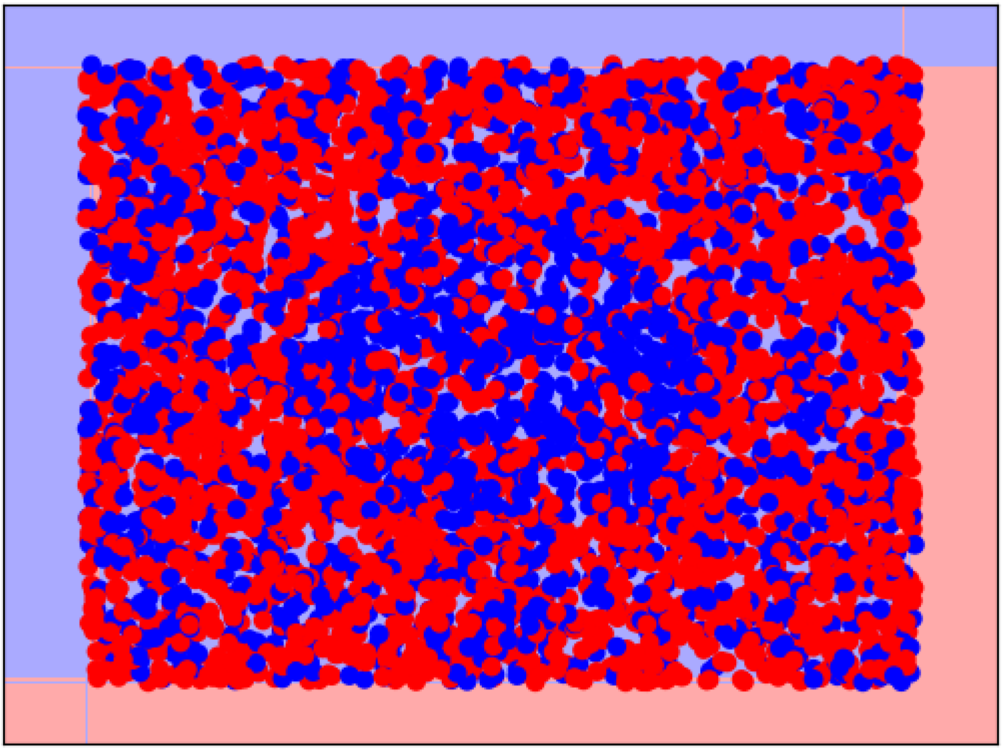
num classifiers = 51



noise=0.4

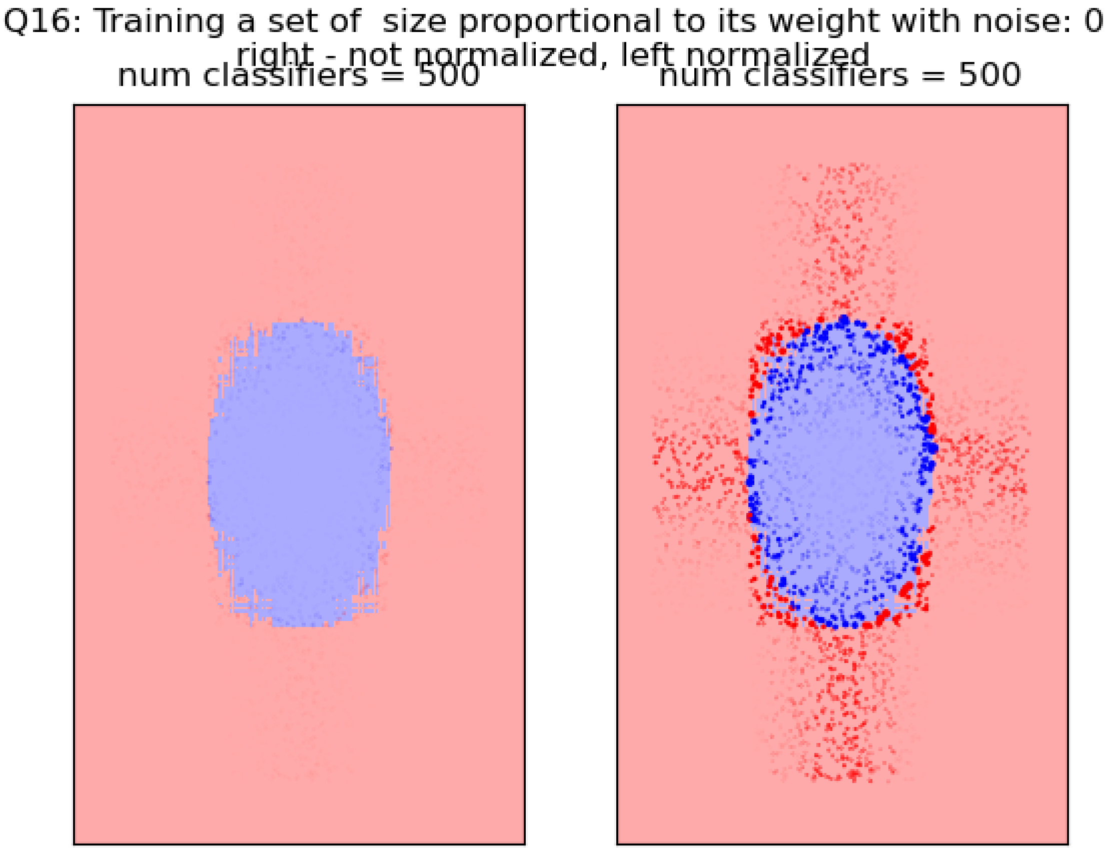
Q15: T that minimizes error: 62, Error: 0.4, noise: 0.4

num classifiers = 62



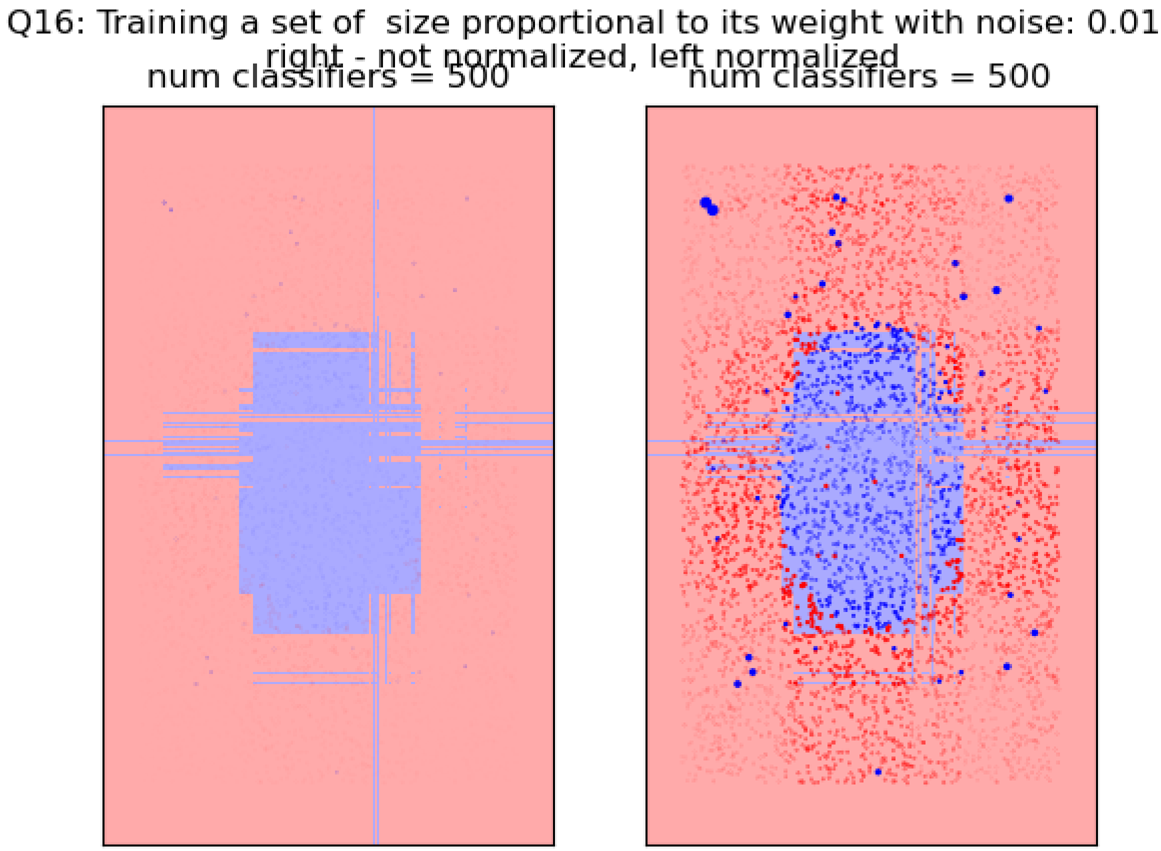
Q16

noise=0

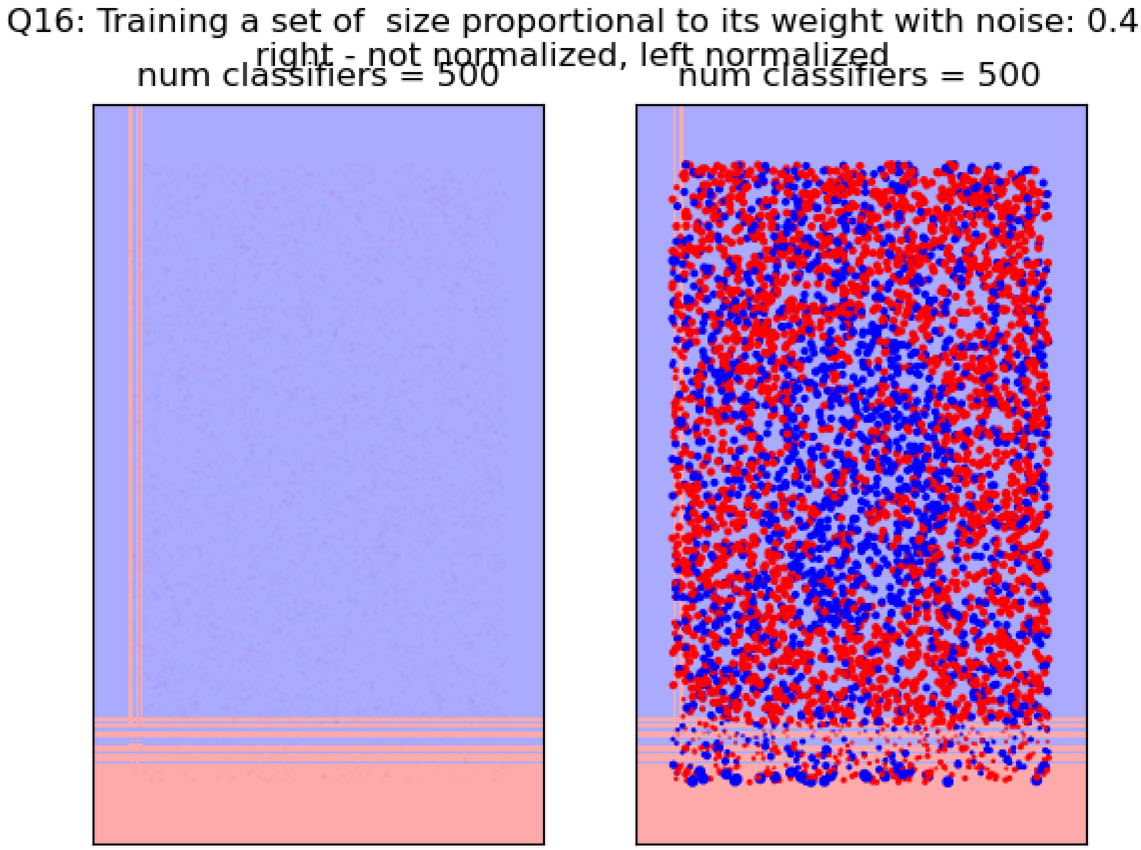


We can see that the points that have greater influence on the outcome in terms of weight, are those that are close to the boundary. This means that it is harder to classify them.

noise= 0.01


Q16: Training a set of size proportional to its weight with noise: 0.01
right - not normalized, left normalized
num classifiers = 500
num classifiers = 500

noise=0.4


Q16: Training a set of size proportional to its weight with noise: 0.4
right - not normalized, left normalized
num classifiers = 500
num classifiers = 500