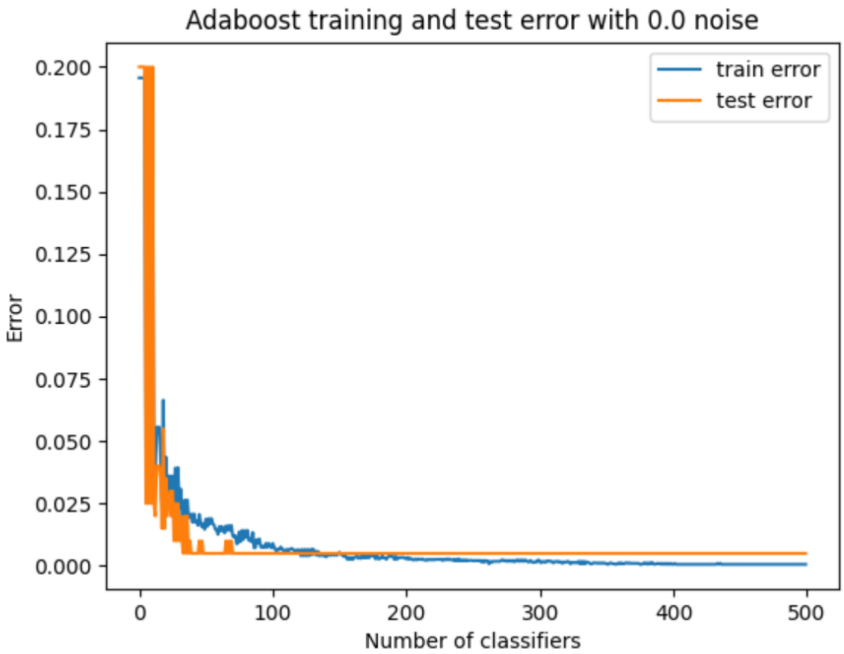
316163260

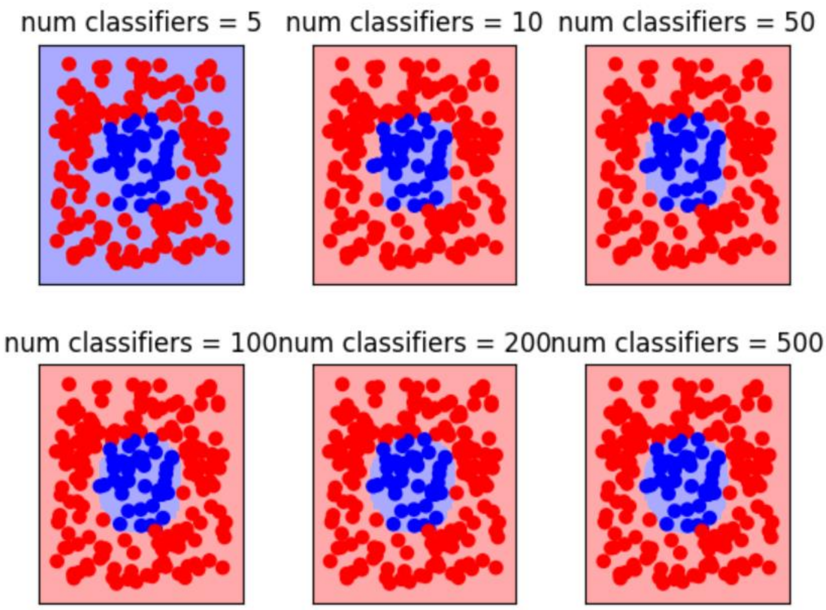
נועה בן דרור

IML – תרגיל 4 – חלק מעשי

(13

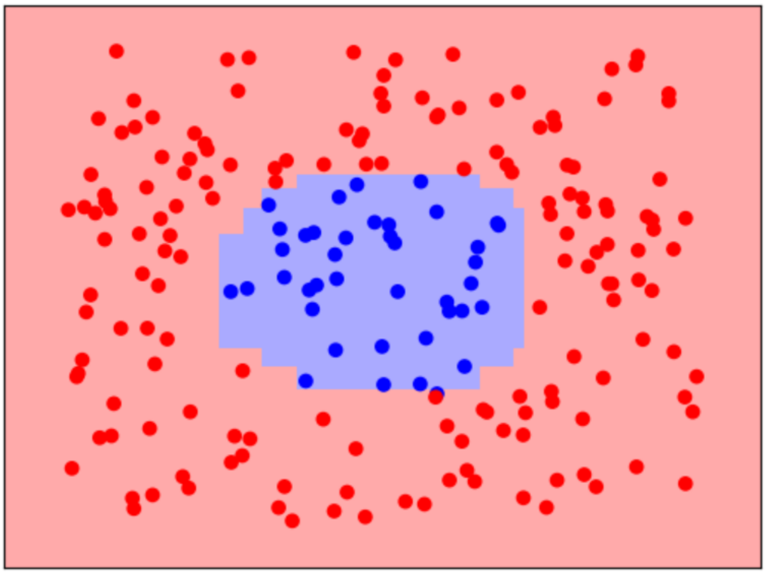


(14



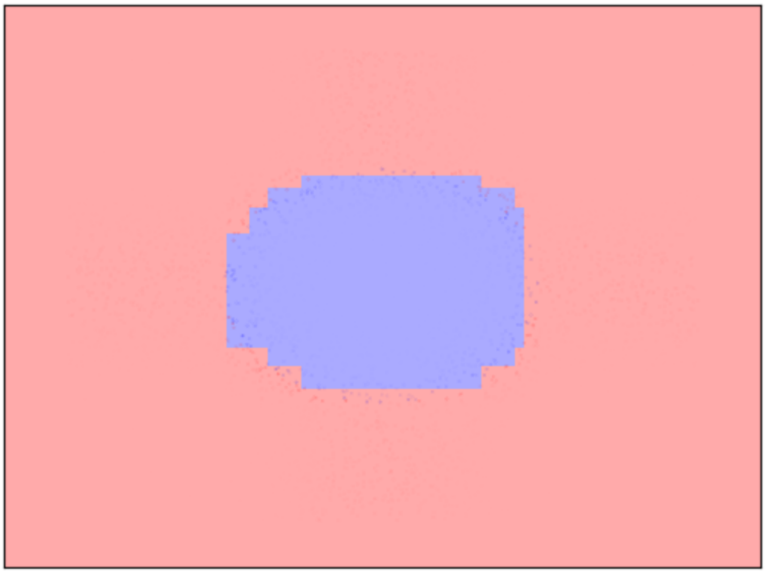
(15

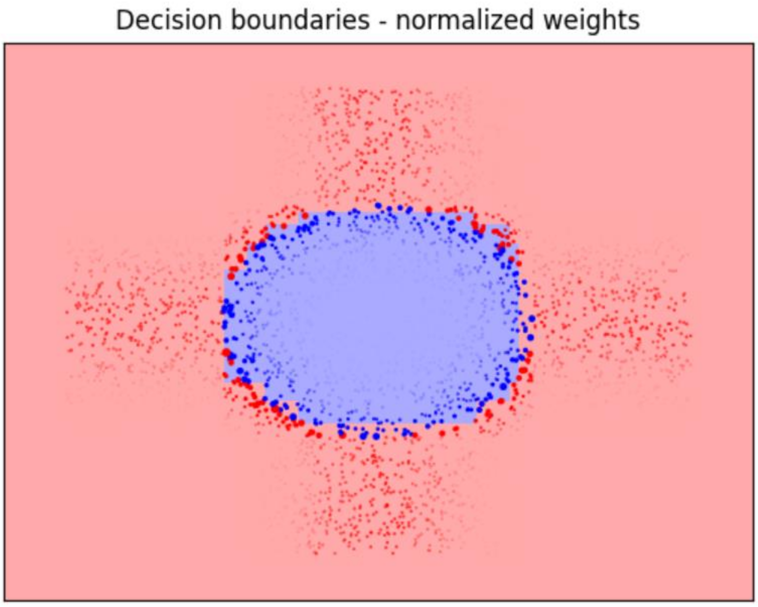
Decision boundaries for 34 classifiers
error = 0.02



(16

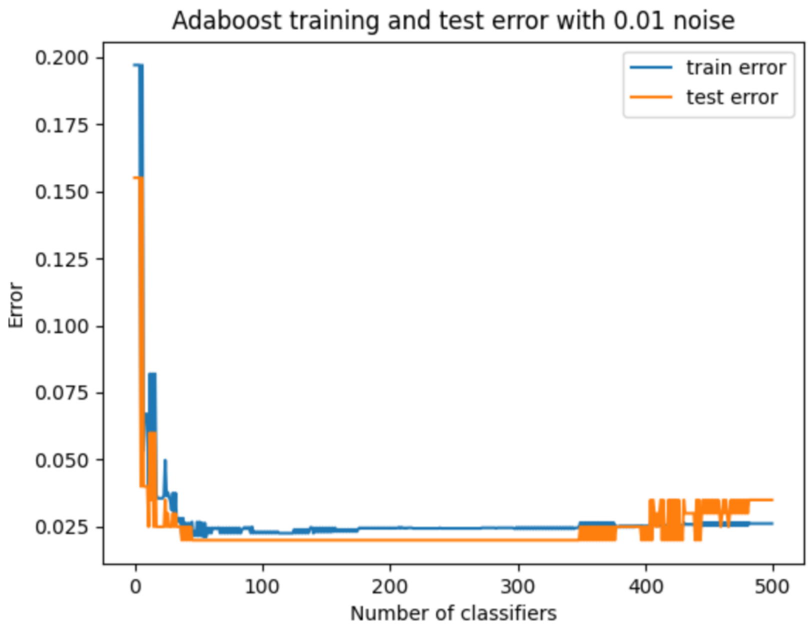
Decision boundaries - original weights



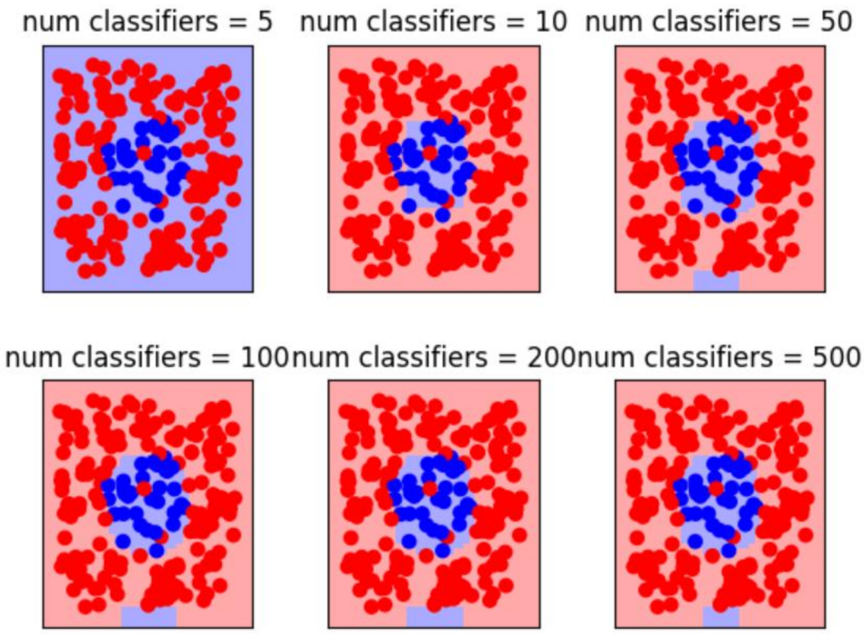


אלגוריתם ה-adaboost ממשקל מחדש את הנקודות בכל איטרציה, ונותן משקל גדול יותר לדגימות שבהן טעינו באיטרציה האחרונה. בגרף הראשון קשה לראות את הנקודות, ורואים בערך רק את ההפרדה. לאחר נרמול של הנקודות, בגרף השני ניתן לראות את הנקודות. ניתן לראות שהנקודות עם המשקל הגדול ביותר הן אלה שקרובות להפרדה, וזה אומר שאלה הנקודות שטעינו בהן באיטרציות האחרונות, ולכן זה הגיוני שההפרדה נמצאת קרוב אליהן.

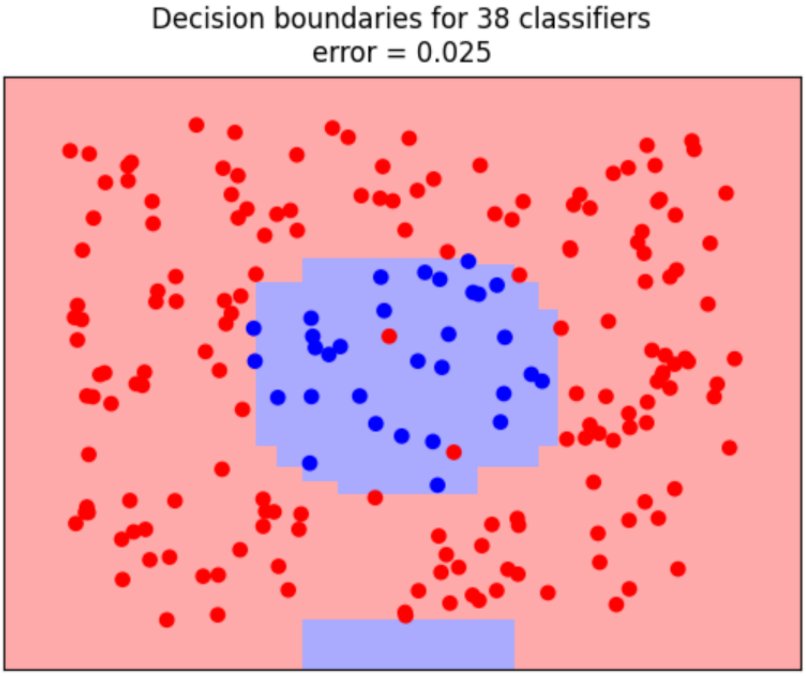
(17

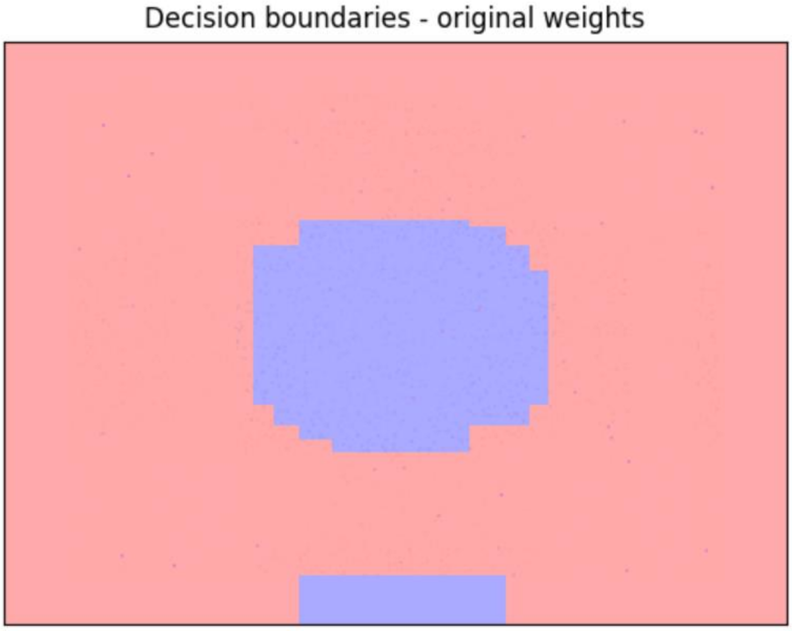


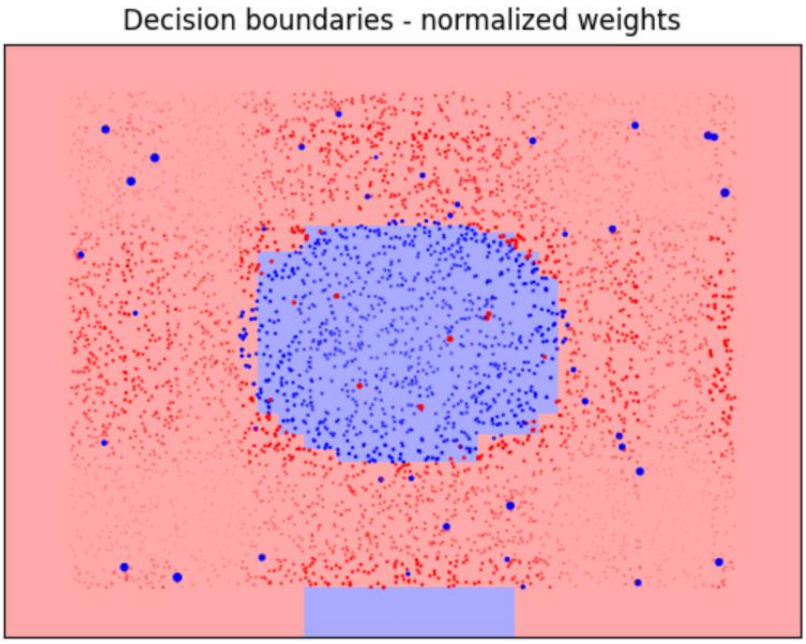
ניתן לראות שהגרף די דומה לגרף שקיבלנו עבור רעש=0. ההבדל העיקרי הוא שכאן ה- error מתישהו מתקבע יחסית על 0.025, ואילו כאשר רעש=0 ה-error התקבע מתישהו על ערך נמוך יותר מ-0.025. אז אפשר לראות שגם ה- test error וגם ה- training error פוחתים ככל שמוסיפים classifiers, ולכן נראה שאין over-fitting.



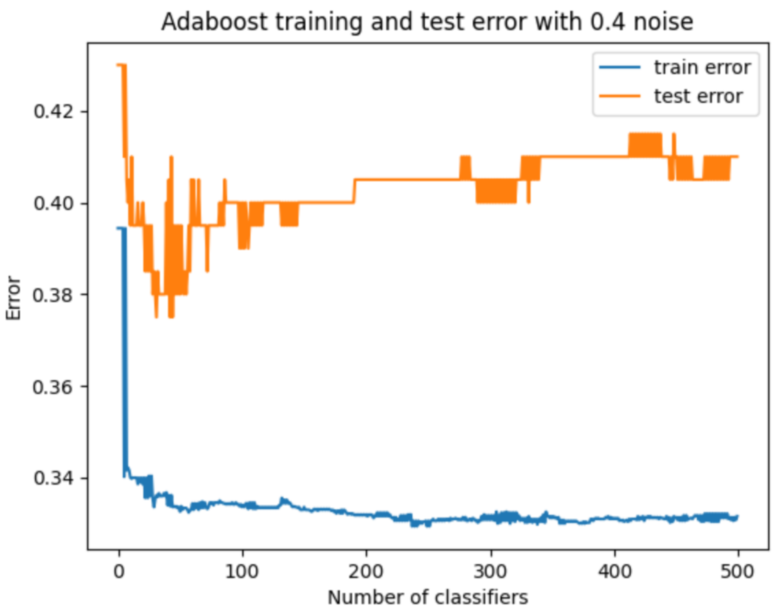
ניתן לראות שהרעש קצת מבלבל את המודל, וגורם לו לעשות קצת טעויות. בכל זאת, המודל עובד די טוב.



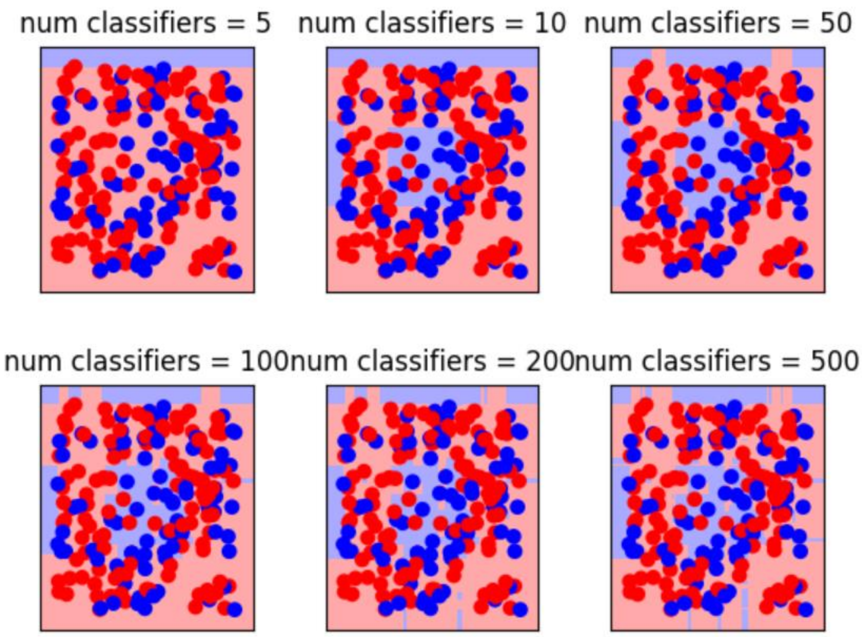




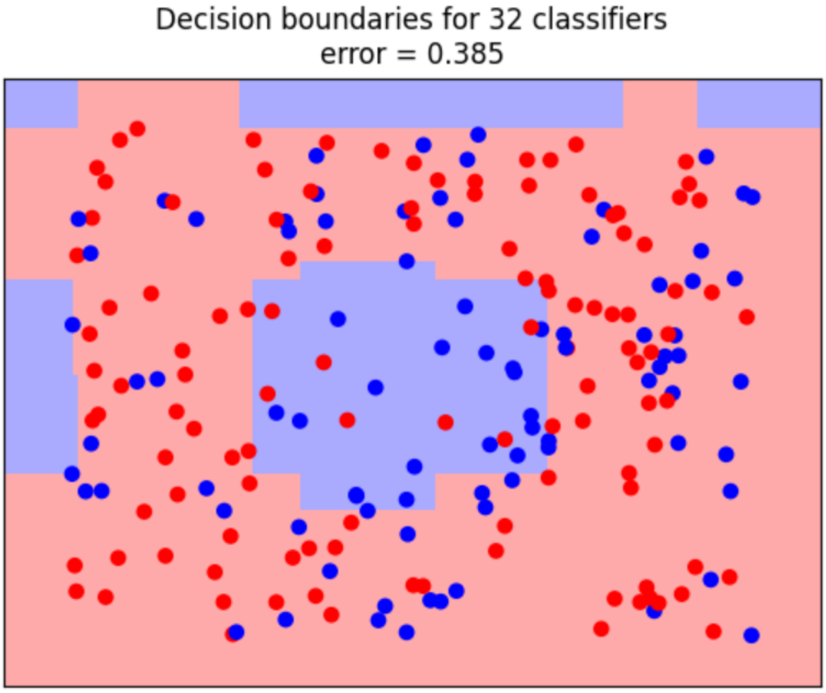
ניתן לראות שהנקודות עם המשקל הגדול ביותר כבר לא בהכרח נמצאות קרוב להפרדה, כי אלה נקודות שמהוות רעש, ולכן סביר שהמודל טעה לגביהן (ולכן נתן להן משקל גבוה).



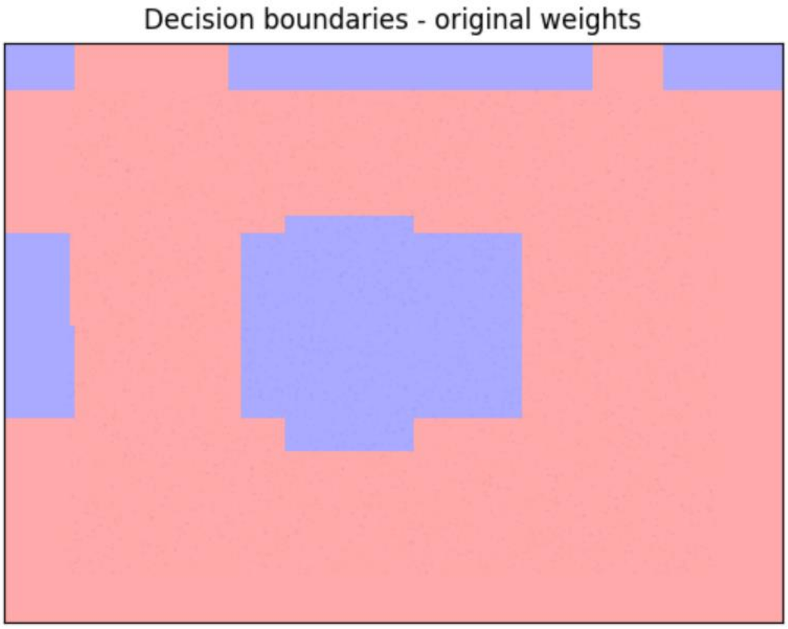
כאן ניתן לראות את ה- variance-bias trade-off בא לידי ביטוי. הדאטה היה מלא ברעשים, ולכן ככל שהוספנו יותר classifiers רואים שה-training error פוחת. למרות זאת, ה-test set יותר מעניין, רואים שבהתחלה יש ירידה ב-error, אבל ככל שאימנו יותר את המודל, הוא ניסה להתאים כל נקודה בדאטה שהיא רעש, ולכן יש כאן overfit, ולכן ה-test error עולה בשלב מסוים, ודי מתקבע.

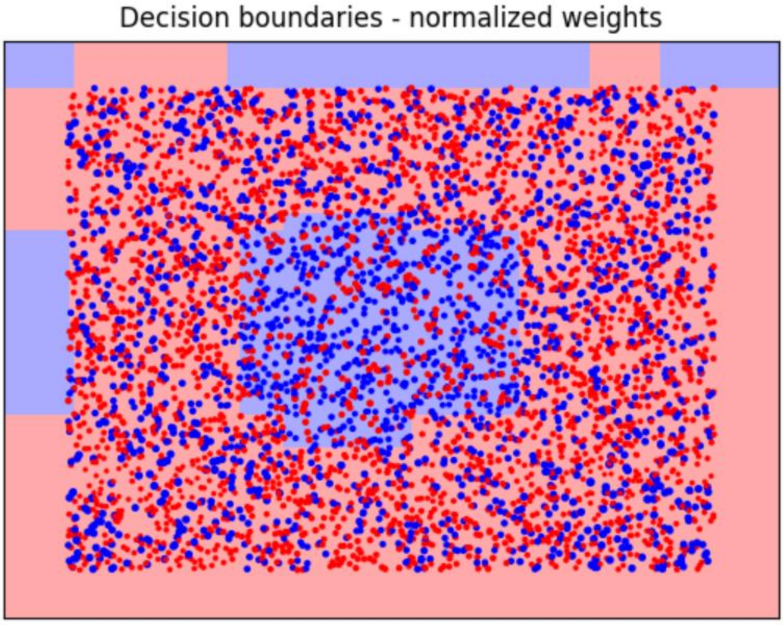


יש לנו הרבה רעש, אז ההפרדה הרבה פחות מסודרת לעומת לפני כן, כשהיה לנו פחות רעש או בלי רעש בכלל. כמו כן, רואים שיש הרבה יותר טעויות בסיווג.



גם כאן אפשר לראות שהיה יותר מדי רעש, ושהמודל עושה הרבה טעויות סיווג.





גם כאן אפשר לראות שהיה יותר מדי רעש, ושהמודל עושה הרבה טעויות סיווג. כמו כן, נשים לב שהמשקלים של הנקודות יחסית זהים (או לפחות של הרבה מהן), וזה אולי מעיד שהיו הרבה נקודות שסווגו לא נכון ולכן הן קיבלו אותו משקל.

נועה בן דרור
316163260

תרגיל 4 IML - חלק תיאורטי

① צריך להוכיח כי:

$\forall \epsilon,\delta>0 \ \exists m(\epsilon,\delta) \ \forall m\geq m(\epsilon,\delta):$

$$\mathbb{P}_{S\sim D^m}\left[L_D(A(S))\le\epsilon\right]\geq 1-\delta \iff \lim_{m\to\infty} \mathbb{E}_{S\sim D^m}\left[L_D(A(S))\right]=0$$

ראשית נוכיח a=>b:

אנחנו יודעים כי $\lim_{m\to\infty}\mathbb{E}_{S\sim D^m}[L_D(A(S))]=0$ וזה אומר שלכל $\epsilon'>0$ קיים N

כך שלכל $n>N$ מתקיים $\mathbb{E}_{S\sim D^m}[L_D(A(S))]<\epsilon'$. נוכל לבחור $\epsilon'=\epsilon\cdot\delta$, ואז קיים n

כך ש- $n>m(\epsilon,\delta)$ כך ש- $\mathbb{E}_{S\sim D^m}[L_D(A(S))]<\epsilon'=\epsilon\cdot\delta$.

כעת, לכל $\epsilon,\delta>0$ קיים $m(\epsilon,\delta)$ כך שלכל $m\geq m(\epsilon,\delta)$:

$$\mathbb{P}_{S\sim D^m}\left[L_D(A(S))\geq\epsilon\right] \underset{\text{אי שוויון מרקוב}}{\le} \frac{\mathbb{E}_{S\sim D^m}[L_D(A(S))]}{\epsilon} \le \delta$$

ומכאן נסיק כי

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

Scanned with CamScanner

כעת נוכיח $a => b$:

אנחנו יודעים כי לכל $\varepsilon,\delta>0$ קיים $m(\varepsilon,\delta)$ כך שלכל $m \geq m(\varepsilon,\delta)$

$$\underbrace{\underset{S\sim D^m}{P}[L_D(A(S))>\varepsilon]<\delta}_{(*)} \iff \underset{S\sim D^m}{P}[L_D(A(S))\leq\varepsilon]\geq 1-\delta$$

לכל $m>m(\varepsilon,\delta)$ מתקיים:

$$0<\underset{S\sim D^m}{E}[L_D(A(S))]=\sum_{S\in D^m}P(S)\cdot L_D(A(S))=$$

$$\sum_{\substack{S\in D^m\\ \text{s.t. } L_D(A(S))<\varepsilon}}P(S)\cdot L_D(A(S))+\sum_{\substack{S\in D^m\\ \text{s.t. } L_D(A(S))\geq\varepsilon}}P(S)\cdot L_D(A(S)) \leq$$

$$\sum_{\substack{S\in D^m\\ \text{s.t. } L_D(A(S))<\varepsilon}}P(S)\cdot\varepsilon+\sum_{\substack{S\in D^m\\ \text{s.t. } L_D(A(S))\geq\varepsilon}}P(S)\cdot 1=P\Big(\bigcup_{\substack{S\in D^m\\ \text{s.t. } L_D(A(S))<\varepsilon}}S\Big)\cdot\varepsilon+P\Big(\bigcup_{\substack{S\in D^m\\ \text{s.t. } L_D(A(S))\geq\varepsilon}}S\Big)=$$

$$P(L_D(A(S))<\varepsilon)\cdot\varepsilon+P(L_D(A(S))\geq\varepsilon)\underset{\text{לפי } (*)}{<}P(L_D(A(S))<\varepsilon)\cdot\varepsilon+\delta\underset{P(L_D(A(S))<\varepsilon)<1}{<}$$

$$\varepsilon+\delta$$

לכן לכל $\varepsilon'>0$ קיים $m>m(\varepsilon,\delta)$ כך ש: $\underset{S\sim D^m}{E}[L_D(A(S))]<\varepsilon+\delta=\varepsilon'$

ולפי הגדרת הגבול נקבל שאכן $\lim_{m\to\infty}\underset{S\sim D^m}{E}[L_D(A(S))]=0$.

(2) ראשית, יש לנו m דגימות ב- $S=(x_i,y_i)_{i=1}^{m} \in R^2$.

האלגוריתם A ימצא את הנקודה הרחוקה ביותר מהראשית, ב-S. לנקודה זו תהיה הנורמה האוקלידית הגדולה ביותר. נסמן את הניחוש שלנו כמעגל $C_{r_{estimated}}$ שהמרכז שלו

נמצא בראשית הצירים, ועם רדיוס $r_{estimated} = \max\limits_{\substack{s\in S \\ s \text{ is labeld } y=1}} \|s\|$

נגדיר כי $\max \emptyset = 0$.

נסמן כ- $C_{r_{real}}$ את המעגל האמיתי שמרכזו בראשית, ובעל רדיוס r_{real}.

נשים לב כי $r_{estimated} \le r_{real}$. אז אנחנו יכולים לתייג לא נכון רק נקודות שהתיוג שלהן

הוא 1, $MISS = C_{real} / C_{estimated}$. צריך להוכיח כי לכל δ, ϵ, $D \sim R^2$,

$\underset{S\sim D^m}{\mathbb{P}}[L_D(A(S)) \le \epsilon] \ge 1-\delta$. כדי להוכיח, נניח של MISS הסתברות ϵ

ב- $D \sim R^2$. אז נקבל שהמקרה של m הדגימות ליפול ב- $R^2 / MISS$

הוא $(1-\epsilon)^m$. אז נרצה ש- $(1-\epsilon)^m < \delta$. אנחנו יודעים כי $(1-\epsilon)^m \le e^{-m\epsilon}$

ואז $m \ge \frac{\log(\frac{1}{\delta})}{\epsilon}$, וזה אומר ש- $m_H(\epsilon,\delta) \le \frac{\log(\frac{1}{\delta})}{\epsilon}$.

(3) נניח בשלילה שלא. כלומר קיים $C \subseteq X$ כך ש- $|C| > \log_2(H)$ וגם H מנפצת את C, כלומר $|H_C| = 2^{|C|}$. לכן: $|H| \ge |H_C| = 2^{|C|} \underset{\text{הנחת השלילה}}{>} 2^{\log_2(H)} = |H|$. וזו סתירה, כנדרש.

4) ראשית נשים לב כי $|H| = 2^n$, ולפי שאלה 3 מתקיים $VCdim(H) \le \log_2|H|$

לכן $VCdim(H) < n$.

בנוסף, הקבוצה $\{e_1,\dots,e_n\}$ - ניתן לנתץ אותה. (משום שאם $I=\{i \mid y_i=1\}$

אז $h_I(e_i)=y_i$). לכן $VCdim(H) \geq n$.

אז קיבלנו $VCdim(H)=n$.

5) נראה כי לכל $k\in\mathbb{N}$ $VC\text{-}dim(H_k)=2k$. נראה ש- $VCdim(H)\geq 2k$. תהי קבוצה $C\subseteq\mathbb{R}$

כך ש- $|C|=2k$ ותהי השמה $y=\{y_i\}_{i=1}^{2k=|C|}$. נראה שניתן לספק אותה. נגדיר סדרת קטעים באופן הבא:

נגדיר את B להיות C בסדר עולה.

(*) כל עוד B לא ריקה: עבור הנקודה הכי קטנה $x_i\in B$ כך ש- $y_i=1$, נגדיר $a_j=x_i$ ו- b_j

להיות הנקודה החיובית הכי גדולה כך שאין ביניהן נקודות x_i כך ש- $y_i=0$. נעדכן $B=B\setminus[a_j,b_j]$

ונחזור לשלב (*).

כעת, נסמן $A=\bigcup_{j=1}^{m}[a_j,b_j]$ הקטעים שבנינו. h_A אכן מספקת השמה זו - לפי

הגדרת הקטעים. וכן $m\leq k$. נניח שלא, כלומר יש לפחות $k+1$ קטעים, כך שלפני בחירת הקטעים:

יש לפחות $k+1$ נקודות חיוביות ולפחות k נקודות שליליות, כלומר $|C|\geq 2k+1$ אך $|C|=2k$ בסתירה.

אז לכל קבוצה $|C|=2k$ H מנתצת את C, ולכן $VC\text{-}dim(H)\geq 2k$.

כעת נראה כי $VC\text{-}dim(H_k)\leq 2k$. תהי $C\subseteq\mathbb{R}$ כך ש- $|C|=2k+1$. נגדיר

$f(x_i)=i \bmod 2$. ונתבונן בהשמה $y=(1,0,1,0,\dots,1)$. לא ייתכן שקיים

$A=\bigcup_{j=1}^{k}[a_j,b_j]$ כך ש- h_A מספקת השמה זו. כי אם כן, בהכרח כל נקודה עם

תווית חיובית תהיה בקטע נפרד אך יש $k+1$ נקודות חיוביות, בסתירה.

6

$H_{con}=\{h:2^{2d}\rightarrow\{0,1\} \mid h \text{ is a boolean conjuction}\}$.

נזכר כי: $VC\text{-}H_{con}=\max\{|C| \mid C\subseteq X,\ |H_C|=2^{|C|}\}$.

נראה כי מימד ה VC של H_{con} שווה ל-d. ראשית נראה $VC\text{-}H_{con}\geq d$:

יהי: $C=\{x_1,\dots,x_d \mid x_i=e_i\}$, אזי לכל $\{y_i\in\{0,1\}^d \mid i\in[d]\}$ נגדיר $h(x_j)=\bigwedge_{y_i=0}\overline{x_{ji}}$,

כאשר הקורדינטה הראשונה היא הוקטור, והקורדינטה השנייה היא המיקום בפנים.

אם קיים j כך ש- $h(x_j)=1$ אז נקבל conjuction עם 1-ים בלבד, כי

$x_{ji}=0$ לכל j, $h(x_j)=y_j=1$.

אם קיים j כך ש- $h(x_j)=0$ אז נקבל conjuction עם לפחות 0 אחד, כי $x_{ji}=1$

ואז $h(x_j)=y_j=0$. נשים לב שלא קיים j כך ש- $y_j=0$, ונקבל conjuction ריק,

שהוא נכון (true) באופן ריק, ונקבל כי לכל j $h(x_j)=y_j=1$. אז נקבל שלכל קבוצה C

בגודל d נוכל למצוא conjuction שיגיע לכל התמונות $\{0,1\}$ ולכן $|H_C|=2^{|C|}$ ולכן

C מנותצת ע"י H_C

כעת נראה כי $VC\text{-}H_{con}\leq d$.

נניח בשלילה כי קיים $C=\{x_i \mid i\in[d+1]\}$ שמנותץ ע"י H_{con}.

כלומר קיים $\{y_i\in\{0,1\}^d \mid i\in[d+1]\}$ כך ש-

$\{h_i \mid i\in[d+1]\}$, $|h_i:\{0,1\}^{d+1}\to\{0,1\}|=2^{d+1}$.

כלומר גם שניתן לבחור כאלו כי $h_i(x_j)=\overline{\delta_{kronecker}(i,j)}$.

מהאופן שבה מחלקת ההיפותזות בנויה אנו יודעים שקיים לכל ששולל את x_j ושולל של

אף משתנה אחר ב-C, או קיימת קורדינטה x_{ij} ששונה מכל המשתנים האחרים ב-C.

כי h_i מקיים כאמור, אז לכל x_i קיים j כך ש- x_{ij} שונה מכל שאר המשתנים ב-C.

ולכן ניתן להסיק משובך היונים שיש כאן סתירה, כי יש $d+1$ בקבוצה, אבל רק d קורדינטות.

אין קבוצה בגודל $d+1$ שלא יכולה לנתץ את H_{con} ולכן $d\le VC\text{-}H_{con}<d+1$, ולכן סה"כ

$VC\text{-}H_{con}=d$ כנדרש.

(7) נניח של-H יש את תכונת ההתכנסות במ"ש עם הפונקציה m_H^{UC}. לפי הגדרה,
מחלקת ההיפותזות H היא למידה agnostic-PAC אם:

$\forall D,\epsilon,\delta \quad \exists m_H^{UC}:(0,1)^2\to\mathbb{N}$ s.t $\forall m>m_H^{UC}(\frac{\epsilon}{2},\delta)$ מתקיים:

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

יהי: $m>m_H^{UC}(\frac{\epsilon}{2},\delta)$. אנחנו יודעים שאם S היא $\frac{\epsilon}{2}$-representative יתקיים כי:

$$L_D(h_S) \leq L_S(h_S) + \frac{\varepsilon}{2} \leq L_S(\underset{h' \in H}{\operatorname{argmin}} L_D(h')) + \varepsilon.$$

בגלל שהנחנו ש-H יש את תכונת ההתכנסות במ״ש עם הפונקציה m_H^{UC} נקבל:

$$\underset{S \sim D^m}{\mathbb{P}}\left[L_D(h_S) \leq \min_{h' \in H} L_D(h') + \varepsilon\right] \geq \mathbb{P}_D\left[S \text{ is } \frac{\varepsilon}{2}\text{-representetive}\right] \geq 1-\delta$$

וזה אומר שלכל $m > m_H^{UC}(\frac{\varepsilon}{2}, \delta)$:

$$\underset{S \sim D^m}{\mathbb{P}}\left[L_D(h_S) \leq \min_{h' \in H} L_D(h') + \varepsilon\right] \geq 1-\delta$$, כנדרש.

(8) H היא לא למידה agnostic PAC. דוגמה נגדית - f מחלקת הפונקציות הבינאריות מעל X.
זו מחלקה שהיא לא למידה PAC ולכן היא גם לא למידה agnostic PAC.
נשים לב כי המחלקה שבחרנו מתאימה לנתונים בשאלה.

(9) יהי $0 < \delta < 1$ ויהיו $0 < \epsilon_1 \le \epsilon_2 < 1$. צריך להוכיח ש-

$$\underbrace{m_H(\epsilon_1,\delta)}_{\tilde{m}} \ge \underbrace{m_H(\epsilon_2,\delta)}_{\hat{m}}.$$

בהינתן S אימון שנדגם i.i.d בגודל m כך ש- $m \ge \tilde{m}$, בהסתברות של לפחות $1-\delta$ A מחזיר היפותזה h כך ש- $L_{D,f}(h) \le \epsilon_1$
(וכן $\epsilon_1 \le \epsilon_2$). לכן נסיק כי $\hat{m} \le \tilde{m}$.

יהי $0 < \epsilon < 1$ ויהיו $0 < \delta_1 \le \delta_2 < 1$. צריך להוכיח ש-

$$\underbrace{m_H(\epsilon,\delta_1)}_{\tilde{m}} \ge \underbrace{m_H(\epsilon,\delta_2)}_{\hat{m}}.$$

(10) יהי $H_1 \subseteq H_2$. נשים לב כי אם מנותצת S בגודל m מנתץ את H_1 נקבל ש-S גם מנותצת את H_2, בגלל ש-$H_1 \subseteq H_2$. לכן $VC\text{-}H_2 \geq VC\text{-}H_1$ כנדרש.