מבוא למערכות לומדות - תרגיל 4

גיא קורנבליט, ת.ז 308224948

שאלה 1

יהי $\mathcal{A}$ אלגוריתם לומד, $\mathcal{D}$ התפלגות, ופונקציית ההפסד ℓ^{0-1}. נוכיח כי הטענות הבאות שקולות:

(1) לכל $\varepsilon,\delta>0$ קיימת פונקציה $m\left(\varepsilon,\delta\right)\in\mathbb{R}$ כך שלכל $m\geq m\left(\varepsilon,\delta\right)$ מתקיים

$$\mathbb{P}_{S\sim\mathcal{D}^{m}}\left(L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right)\leq\varepsilon\right)\geq1-\delta$$

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

הוכחה:

ראשית, נסיק כי הטענות מוגדרות על בעיית סיווג, לומד $\mathcal{A}:2^{\mathcal{X}}\to\mathcal{Y}^{\mathcal{X}}$ (ללא הנחת קיום של מחלקת היפותזות), ועם הנחת הרלייזביליות. כלומר, קיימת $f:\mathcal{X}\to\mathcal{Y}\in\{0,1\}$ עבורה נוכל לכתוב $S=\left\{ \left(x_{i},f\left(x_{i}\right)\right)\right\} _{i=1}^{m}$. כאשר כל $X_{1},...,X_{m}\sim\mathcal{D}$ לכל התפלגות $\mathcal{D}$ מעל $\mathcal{X}$, וגם $L_{\mathcal{D}}\left(f\right)=0$.

בכיוון הראשון, יהיו $\varepsilon,\delta>0$ כך שקיימת $m\left(\varepsilon,\delta\right)$, עבורה לכל $m\geq m\left(\varepsilon,\delta\right)$ מתקיים

$$(*)\ \mathbb{P}_{S\sim\mathcal{D}^{m}}\left(L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right)\leq\varepsilon\right)\geq1-\delta$$

צ"ל כי קיים $\varepsilon_{0}>0$ כך שקיים $N\in\mathbb{N}$ המקיים כי לכל $m>N$ מתקיים

$$\left|\mathbb{E}_{S\sim\mathcal{D}^{m}}\left(L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right)\right)\right|<\varepsilon_{0}$$

. ראשית, נבחין כי מאחר ו-$L_{\mathcal{D}}$ מוגדרת עם ℓ^{0-1} אז

$$L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right)=\mathbb{P}_{X\sim\mathcal{D}}\left(h_{S}(x)\neq f(x)\right)$$

כלומר $L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right)\in[0,1]$ (#).

מאחר ו-S הוא משתנה מקרי עם התפלגות $\mathcal{D}^m$, אז גם $L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right)$ משתנה מקרי ולכן מהגדרת התוחלת מתקיים -

$$\begin{aligned}
\mathbb{E}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right)\right) &= \int\limits_{S\subseteq\mathcal{X}^m} L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right)\cdot f_{\mathcal{D}^m}\left(S\right) \\
&\overset{\text{additivity}}{=} \int\limits_{\substack{S\subseteq\mathcal{X}^m \\ L_{\mathcal{D}}\left(\mathcal{A}(S)\right)\leq\varepsilon}} L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right)\cdot f_{\mathcal{D}^m}\left(S\right) + \int\limits_{\substack{S\subseteq\mathcal{X}^m \\ L_{\mathcal{D}}\left(\mathcal{A}(S)\right)>\varepsilon}} L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right)\cdot f_{\mathcal{D}^m}\left(S\right) \\
&\overset{(\#)L_{\mathcal{D}}\leq 1}{\leq} \int_{L_{\mathcal{D}}\left(\mathcal{A}(S)\right)\leq\varepsilon} \varepsilon\cdot f_{\mathcal{D}^m}\left(S\right) + \int_{L_{\mathcal{D}}\left(\mathcal{A}(S)\right)>\varepsilon} 1\cdot f_{\mathcal{D}^m}\left(S\right) \\
&\leq \varepsilon\cdot\int_{L_{\mathcal{D}}\left(\mathcal{A}(S)\right)\leq\varepsilon} \cdot f_{\mathcal{D}^m}\left(S\right) + \int_{L_{\mathcal{D}}\left(\mathcal{A}(S)\right)>\varepsilon} f_{\mathcal{D}^m}\left(S\right) \\
&= \varepsilon\cdot\overbrace{\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right)\leq\varepsilon\right)}^{\leq 1} + \overbrace{\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right)>\varepsilon\right)}^{(*)\ \leq\delta} \\
&\leq \varepsilon+\delta
\end{aligned}$$

נגדיר $\varepsilon_0=\varepsilon+\delta>0$, $N=m\left(\varepsilon,\delta\right)$ וקיבלנו שלכל $m\geq N$ מתקיים

$$\left|\mathbb{E}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right)\right)\right|\leq\varepsilon_0$$

כאשר השקילות לביטוי בערך מוחלט מאחר ו-$L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right)\in[0,1]$, אז התוחלת תמיד אי-שלילית.
■

בכיוון השני, יהיו $\varepsilon,\delta>0$. נגדיר $\varepsilon_0=\varepsilon\delta>0$, אז מההנחה, קיים $N\in\mathbb{N}$ כך שלכל $m\geq N$ מתקיים $\left|\mathbb{E}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right)\right)\right|\leq\varepsilon_0$. כעת, מאחר ו- $L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right)$ הינו פונקציה של המשתנה המקרי S, אזי שגיאת ההכללה הינה משתנה מקרי אי-שלילי. לפיכך, תנאי אי-שוויון מרקוב מתקיימים ולכן

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right)>\varepsilon\right)\leq\frac{\mathbb{E}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right)\right)}{\varepsilon}\leq\frac{\varepsilon_0}{\varepsilon}=\delta$$

והמאורע המשלים הינו $\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}\left(\mathcal{A}\left(S\right)\right)\leq\varepsilon\right)\geq 1-\delta$ כנדרש. ■

שאלה 2

יהי $\mathcal{X}=\mathbb{R}^2$, $\mathcal{Y}=\{0,1\}$ ותהי $\mathcal{H}_r=\{h_r : r\in\mathbb{R}^2\}$, כך ש-$h_r(x)=\mathbf{1}_{\left\{\|x\|_2\le r\right\}}$. נראה כי $\mathcal{H}$ היא למידה-PAC וסיבוכיות המדגם חסומה ע"י $m_{\mathcal{H}}\left(\varepsilon,\delta\right)\le\frac{\log(1/\delta)}{\varepsilon}$.

הוכחה:

- נגדיר לכל $h\in\mathcal{H}$ את שגיאת ההכללה להיות

$$L_{\mathcal{D}}\left(h_S\right)=\mathbb{E}_{X\sim\mathcal{D}}\left(h_S(X)\neq h_\theta\left(X\right)\right)=\mathbb{P}_{X\sim\mathcal{D}}\left(\mathbf{1}_{h_S(X)\neq h^*(X)}\right)$$

עבור $h_\theta:\mathcal{X}\to\mathcal{Y}$, פונקציית Ground-truth (כדור היחידה הסגור) תחת הנחת הרלייזביליות, המקיימת $L_{\mathcal{D}}\left(h_\theta\right)=0$.

- **נגדיר אלגוריתם למידה $\mathcal{A}$ ($\mathcal{A}_m$)** עם עיקרון tightest fit for positive examples. בהינתן $S=\left(\boldsymbol{x}_i,y_i\right)_{i=1}^m$ מדגם בגודל $m\in\mathbb{N}$, $\mathcal{A}$ מחזיר היפותזה $h_{r_{alg}}\in\mathcal{H}$ כאשר

$$r_{alg}=\max_{y_i=1}\|x_i\|_2$$

אם לכל $i\in[m]$, מתקיים $y_i=1$ אז $r_{alg}=\infty$ (כלומר נסווג הכל ב-1), ואם $y_i=0$ לכל i אז $r_{alg}=-\infty$ (נסווג הכל כ-0).

- **הוכחת $\mathcal{H}$ למידה-PAC.** יהיו $\varepsilon,\delta>0$, ותהי התפלגות $\mathcal{D}$ מעל $\mathcal{X}$. מתקיים $r_{alg}\le\theta$, נניח כי $r_{alg}<\theta$. כמו כן, נבחין כי הסיווג יהיה נכון עבור דגימות שמקיימות $\|x\|_2\le r_{alg}$ ו-$\|x\|_2>\theta$, ולכן טעות הכללה של המודל תהיה בתחום $\{\boldsymbol{x}\in\mathbb{R}^2 : r_{alg}<\|x\|_2<\theta\}$. במילים אחרות, נוכל לעדכן את האינדיקטור שמגדיר את שגיאת ההכללה כך ש-

$$\mathbf{1}_{\left\{X\in\mathbb{R}^2 : r_{alg}<\|X\|_2<\theta\right\}}=\mathbf{1}_{h_S(X)\neq h^*(X)}$$

.

- **נחפש את ההסתברות למדגם שגורר טעות הכללה של לכל היותר ε.** נבחין כי אם $\mathbb{P}_{X\sim\mathcal{D}}\left(-\infty<\|X\|_2<\theta\right)<\varepsilon$ אז $L_{\mathcal{D}}\left(h_S\right)=\mathbb{P}_{X\sim\mathcal{D}}\left(r_{alg}<\|X\|_2<\theta\right)<\varepsilon$ לכל $S\subseteq\mathcal{X}$. כלומר ההיפותזה למידה PAC כמעט-תמיד (בהסתברות 1 עבור $\delta=0$). לכן, נניח כי $\mathbb{P}_{X\sim\mathcal{D}}\left(-\infty<\|X\|_2<\theta\right)\ge\varepsilon$, ונגדיר $\theta'\in\mathbb{R}$ כך ש-$\mathbb{P}_{X\sim\mathcal{D}}\left(\theta'<\|X\|_2<\theta\right)=\varepsilon$. נחלק למקרים.

– אם $\theta' < r_{alg}$, אז מתקיים $(\theta' \le r_{alg} < \theta)$ וממונוטוניות ההסתברות מתקיים

$$L_{\mathcal{D}}\left(h_S\right) = \mathbb{P}_{X\sim\mathcal{D}}\left(r_{alg} < \left\|X\right\|_2 < \theta\right) < \varepsilon$$

, כלומר ההסתברות לשגיאת ההכללה חסומה ע"י ε כמעט-תמיד.

– אם $r_{alg} \le \theta'$ (המעגל שמגדיר $\mathcal{A}$ מוכל במעגל שרדיוסו θ'), אז נקבל $L_{\mathcal{D}}\left(h_S\right) \ge \varepsilon$. ההסתברות למאורע זה הינה כאשר המדגם לא כולל נקודות בשטח בין המעגל ברדיוס θ' לבין המעגל ברדיוס θ, אחרת מאופן פעולת האלגוריתם היינו חוזרים למקרה בו $\theta' < r_{alg}$. מאחר וההסתברות לנקודה בודדת להיות בשטח זה הינה בדיוק ε, ומפני שכל הדגימות ב-S ב"ת נקבל כי

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}\left(h_S\right) \ge \varepsilon\right) = \left(1-\varepsilon\right)^m$$

כעת, נובע מפיתוח טור טיילור של אקספוננט כי $1 - x \le e^{-x}$ ולכן

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}\left(h_S\right) \ge \varepsilon\right) \le e^{-\varepsilon m}$$

נחסום את ההסתברות $e^{\varepsilon m} \le \delta \iff m \ge \frac{-\log(\delta)}{\varepsilon} = \frac{\log(1)-\log(\delta)}{\varepsilon} = \frac{\log\left(\frac{1}{\delta}\right)}{\varepsilon}$. נגדיר $m_{\mathcal{H}}\left(\varepsilon,\delta\right) = \frac{\log\left(\frac{1}{\delta}\right)}{\varepsilon}$ ומכאן הטענה. ■

שאלה 3

יהיו $d \ge 2$, $\mathcal{X} = \{0,1\}^d$ ו-$\mathcal{Y} = \{0,1\}$. כל דגימה $(\boldsymbol{x}, y)$ מורכבת מהשמה שונה של d משתנים ותיוג. לכל משתנה $x_k, k \in [d]$, נסמן את שני הליטרלים - x_k ו-$\overline{x_k} = 1 - x_k$. המחלקה $\mathcal{H}_{con}$ מוגדרת על תתי-קבוצות של $2d$ המשתנים הללו וקוניוקציה ביניהם. **נמצא את מימד VC של $\mathcal{H}_{con}$.**

הוכחה: נטען כי $\text{VC-dimension}\left(\mathcal{H}_{con}\right) = d$.

(1) נראה כי קיימת קבוצה $C \subseteq \mathcal{X}$ כך ש-$|C| = d$, $\mathcal{H}_{con}$ מנתצת C. תהי $C = \left(e_1, ..., e_d\right)$.

לכל וקטור תיוגים $\vec{y} = \left(y_1, ..., y_d\right)^\top \in \{0,1\}^d$. נגדיר את הקבוצה $I = \{i \in [d] : y_i = 0\}$.

נגדיר את ההיפותזה באופן הבא -

$$h = \underset{i\in I}{\wedge} \overline{x_i}$$

כלומר, ההיפותזה מתייגת איבר ב-1 אם"ם הליטרלים שלו מספקים את נוסחת הגימום.
יהי $j \in [d]$, אם $j \in I$ אז ל-e_j יש 1 בכניסה ה-j ולכן הליטרל $\overline{x_j}$ מופיע בגימום ולכן $h(e_j) = 0 = y_i$. אם $j \notin I$, אז ל-e_j יש אפס בכל קוארדינטה $i \in I$ ו-$\overline{x_j}$ לא נמצא בגימום ולכן e_j מספקת את הנוסחה ולכן $h(e_j) = 1 = y_i$. כנדרש. הראינו כי לכל תיוג אפשרי של C קיימת היפותזה שתתייג את הקבוצה לפיו, ולכן $\mathcal{H}_{con}$ מנתצת את C.

(2) **נראה כי לכל קבוצה** $C \subseteq \mathcal{X}$ **כך ש-** $|C| = d+1$, $\mathcal{H}_{con}$ **לא מנתצת את** C**.** תהי $C \subseteq \mathcal{X}$ כך ש- $|C| = d+1$. נסמן $C = (x_1, ...x_{d+1})$ כאשר לכל דגימה x_i $i \in [d+1]$ נסמן ב-x_i^k את הקוארדינטה ה-$k \in [d]$, כלומר את המשתנה הבוליאני ה-k. נניח בשלילה כי $\mathcal{H}_{con}$ מנתצת את C, כלומר לכל וקטור תיוג $y \in \{0,1\}^{d+1}$ קיימת היפותזה $h \in \mathcal{H}_C$ שמשיגה את אותו התיוג ב-C.

נגדיר קבוצת תיוגים

$$L = \left\{(y_1, y_2, ..., y_{d+1}) \in \{0,1\}^{d+1} : \exists i \in [d+1]\; s.t\; y_i = 0, \forall j \neq i\; y_j = 1\right\}$$

L מכילה כל האפשרויות לוקטור שמכיל קוארדינטה 0 יחידה, לכן $|L| = d+1$. בנוסף, מאחר ו-$\mathcal{H}_{con}$ מנתצת את C והקבוצה L מוכלת בקבוצת התיוגים, אז בפרט קיימות $h_1, ..., h_{d+1} \in \mathcal{H}_C$ שמשיגות כל תיוג בקבוצה L. נניח כי לכל $i \in [d+1]$, $h_i(x_i) = 0$ וגם לכל $j \neq i$ $h_i(x_j) = 1$.

אבחנה: נשים לב כי x_i נשלח אל האפס אם"ם אינו מספק את נוסחת הגימום של h_i, לכן נבחין כי קיים ליטרל בנוסחת הגימום שמגדירה h_i. אחרת, נוסחת הגימום אינה מכילה ליטרלים, והגימום הריק מגדיר תיוג מלא של כל האיברים ב-C ב-1 (כי כל הליטרלים המוגדרים באיברי C מסופקים על ידי המשוואה הריקה). כלומר, לכל h_i קיים $k \in [d]$ כך ש- x_k או $\overline{x_k}$ מופיע בגימום.

בסה"כ, הפונקציות $h_1, ...h_{d+1}$ מגדירות תיוגים של איברי C שכוללים תיוג 0 יחיד, לכן בנוסחאות הגימום של הפונקציות מופיעים לפחות $d+1$ ליטרלים. אולם, קיימים d משתנים ולכן מעיקרון שובך היונים בהכרח קיים משתנה שמופיע בשתי נוסחאות גימום שונות, שמתאימות לשתי היפותזות שונות. נסמן את הקוארדינטה של המשתנה המשותף ב-k, ונניח כי הלה משותף לנוסחאות הגימום של h_1 ו-h_2, נסמן את הנוסחאות ב- l_1, l_2.
נחלק למקרים:

(א) אם $x^k \in l_1$ וגם $x^k \in l_2$, אזי $h_1(x_1) = 0$ אם"ם $x_1^k = 0$, ו-$h_2(x_2) = 0$ אם"ם $x_2^k = 0$ ולכן $h_1(x_2) = 0$ בסתירה להגדרת ההיפותזות (אם בשתי הנוסחאות מופיע $\overline{x^k}$ ההוכחה סימטרית).

(ב) אם $x_k \in l_1$ וגם $\overline{x_k} \in l_2$, אזי $h_1(x_1) = 0$ אם"ם $x_1^k = 0$ ו-$h_2(x_2) = 0$ אם"ם $x_2^k = 1$. נשים לב כי $d \geq 2$ ו-$|C| = d+1$ לכן קיים $x_3 \in C$. אם $x_3^k = 0$ אז $h_1(x_3) = 0$ ואם $x_3^k = 1$ אז $h_2(x_3) = 0$. בסתירה להגדרת ההיפותזות. (המקרה הסימטרי שקול).

לסיכום, הראינו כי לא קיימות היפותזות ב-$\mathcal{H}_{con}$ שמסוגלות לתייג את C לפי L, כלומר $|\mathcal{H}_C| \neq C \to \mathcal{Y}$, ולכן $\mathcal{H}_{con}$ לא מנתצת את C. מכאן, נובע מיידית כי $\mathcal{H}_{con}$ אינה מנתצת כל $C \subseteq \mathcal{X}$ כך ש- $|C| \geq d+1$. לפיכך, $\text{VC-dimension}(\mathcal{H}_{con}) = d$. ■

שאלה 4

טענה: אם ל-$\mathcal{H}$ מקיימת את תכונת התכנסות במ"ש עם פונקציה $m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$, אז $\mathcal{H}$ היא למידה Agnostic-PAC עם סיבוכיות מדגם $m_{\mathcal{H}}(\varepsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\varepsilon/2, \delta)$.

הוכחה: יהיו $0 < \varepsilon, \delta < 1$, התפלגות $\mathcal{D}$ על $\mathcal{X} \times \mathcal{Y}$ ופונקציית הפסד ℓ המגדירה את שגיאת ההכללה ואת השגיאה האמפירית. מהגדרה, מדגם $S \subseteq (\mathcal{X} \times \mathcal{Y})^m$ הינו ε-מייצג עבור $\mathcal{D}, \ell, \mathcal{H}$ אם מתקיים $|L_S(h) - L_{\mathcal{D}}(h)| < \varepsilon$ לכל $h \in \mathcal{H}$. באופן שקול, מתקיים

$$|L_S(h) - L_{\mathcal{D}}(h)| < \varepsilon \iff L_{\mathcal{D}}(h) - \frac{\varepsilon}{2} < L_S(h) < L_{\mathcal{D}}(h) + \frac{\varepsilon}{2}$$
$$\iff L_{\mathcal{D}}(h) < L_S(h) + \frac{\varepsilon}{2} < L_{\mathcal{D}}(h) + \varepsilon$$

נסמן ב-h_S את הפלט של אלגוריתם הלמידה $ERM_{\mathcal{H}}(S)$, כלומר $h_S = \underset{h \in \mathcal{H}}{\operatorname{argmin}}(L_S(h))$, וב-$h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}}(L_{\mathcal{D}}(h))$ את המסווג שמשיג את שגיאת ההכללה המינימלית. נבחין כי מתקיים עבור מדגם S $\frac{\varepsilon}{2}$-מייצג מתקיים $L_{\mathcal{D}}(h_S) < L_S(h_S) + \frac{\varepsilon}{2}$ וגם $L_S(h^*) + \frac{\varepsilon}{2} < L_{\mathcal{D}}(h^*) + \varepsilon$. כמו כן, מאופן פעולת $ERM_{\mathcal{H}}$, מתקיים כי $L_S(h_S) \leq L_S(h^*)$ $(*)$ ולכן בסה"כ מתקיים

$$L_{\mathcal{D}}(h_S) < L_S(h_S) + \frac{\varepsilon}{2} \overset{(*)}{\leq} L_S(h^*) + \frac{\varepsilon}{2} < L_{\mathcal{D}}(h^*) + \varepsilon$$

כלומר, מתקיים $L_{\mathcal{D}}\left(h_S\right) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}\left(h\right)+\varepsilon$ (נוכל לעדן לא"ש חלש). הראינו כי

$$S_0 \in\left\{S \subseteq \mathcal{X} \times \mathcal{Y}: \text{S is } \frac{\varepsilon}{2}\text{-representative }\right\} \Rightarrow S_0 \in\left\{S \subseteq \mathcal{X} \times \mathcal{Y}: L_{\mathcal{D}}\left(h_S\right) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)+\varepsilon\right\}$$

ולכן $\left\{\text{S is } \frac{\varepsilon}{2}\text{-representative }\right\} \supseteq \left\{L_{\mathcal{D}}\left(h_S\right) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)+\varepsilon\right\}$ (**).

כעת, מאחר ו-$\mathcal{H}$ מקיימת את תכונת התכנסות במידה שווה, אז קיימת פונקציה $m_{\mathcal{H}}^{UC}$ כך שלכל $m \in \mathbb{N}$ המקיים $m \geq m_{\mathcal{H}}^{UC}\left(\frac{\varepsilon}{2}, \delta\right)$ מתקיים

$$\mathbb{P}_{S \sim \mathcal{D}^m}\left(L_{\mathcal{D}}\left(h_S\right) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)+\varepsilon\right) \stackrel{(**)}{\geq} \mathbb{P}_{S \sim \mathcal{D}^m}\left(S \text{ is } \frac{\varepsilon}{2}\text{-representative}\right) \geq 1-\delta$$

כלומר, $\mathcal{H}$ למידה Agnostic-PAC, עם סיבוכיות מדגם $m_{\mathcal{H}}(\varepsilon, \delta) \leq m_{\mathcal{H}}^{UC}\left(\frac{\varepsilon}{2}, \delta\right)$. כלומר, $\mathcal{H}$ המקיימת את תכונת התכנסות במ"ש, מקיימת את חוק המספרים הגדולים (L_S מתכנס בהסתברות ל-$L_{\mathcal{D}}$) לכל $\mathcal{D}, \varepsilon, \delta$ ולכן בפרט קיים חסם לשגיאת ההכללה. ■

שאלה 7 - מונוטוניות ב-VC-Dimension

טענה: יהיו $\mathcal{H}_1, \mathcal{H}_2$ מחלקות היפותזות של סיווג בינארי כך ש- $\mathcal{H}_1 \subseteq \mathcal{H}_2$, אזי

$$\mathrm{VC}\left(\mathcal{H}_1\right) \leq \mathrm{VC}\left(\mathcal{H}_2\right)$$

.

הוכחה: נסמן $d_1 := \mathrm{VC}\left(\mathcal{H}_1\right)$ ו-$d_2 := \mathrm{VC}\left(\mathcal{H}_2\right)$. מההכלה, נוכל להסיק כי $\mathcal{H}_1, \mathcal{H}_2$ פועלות על מדגמים מעל אותו מרחב מדגם, כלומר לכל $h \in \mathcal{H}_1, \mathcal{H}_2$ מתקיים $h: \mathcal{X} \rightarrow\{0,1\}$.

מהגדרת מימד VC, מאחר ו-$\mathrm{VC}\left(\mathcal{H}_1\right)=d_1$ קיימת קבוצה $C \subseteq \mathcal{X}$ בת d_1 דגימות כך ש- $\mathcal{H}_1$ מנתצת את C. מאחר וכל ההיפותזות ב-$\mathcal{H}_1$ מוכלות ב-$\mathcal{H}_2$ אז בפרט $\mathcal{H}_2$ מנתצת את C, כלומר קיימות 2^{d_1} היפותזות ב-$\mathcal{H}_C$. לכן נוכל להסיק כי מימד VC של $\mathcal{H}_2$ הינו לכל הפחות מימד VC של $\mathcal{H}_2$, כלומר $d_1 \leq d_2$. כנדרש. ■

שאלה 8

יהי $\mathcal{X}$ מרחב מדגם, ו-$\mathcal{Y}=\{\pm 1\}$, $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ מחלקת היפותזות. נגדיר את הפונקציה : $\tau_{\mathcal{H}}(m)$ $\mathbb{N} \rightarrow \mathbb{N}$ באופן הבא:

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

(1) הפונקציה מודדת את קצב הגידול המירבי של מחלקת ההיפותזות על מדגם בגודל m. בפרט, מחזירה את המספר המקסימלי של היפותזות ב-$\mathcal{H}$ עבור תת-קבוצה של מרחב המדגם $\mathcal{X}$ (דגימות) בגודל m.

(2) נניח כי $VCdim\left(\mathcal{H}\right)=\infty$, אזי $\tau_{\mathcal{H}}\left(m\right)\equiv2^{m}$ לכל $m\in\mathbb{N}$.

(3) אם מימד $VCdim\left(\mathcal{H}\right)=d$, אז לכל $m\leq d$ מתקיים $\tau_{\mathcal{H}}\left(m\right)=2^{m}$.

(4) **טענה:** אם $VCdim\left(\mathcal{H}\right)=d$ ו-$m>d$ אזי

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

הוכחה:

(א) **טענת עזר 1:** לכל $C\subseteq\mathcal{X}$ בגודל סופי ולכל היפותזה $\mathcal{H}$, מתקיים

$$\left|\mathcal{H}_{C}\right|\leq\left|\left\{ B\subseteq C\right\} :\mathcal{H}\text{ shatters }B\right|$$

הוכחה: תהי $\mathcal{H}$, נוכיח באינדוקציה על $\left|C\right|=m$. נסמן $C=\left\{ c_{1},...,c_{m}\right\}$.

בסיס: עבור $m=1$, נבחין כי $B\in\left\{ \emptyset,\left\{ c_{1}\right\} \right\}$. ראשית, נבחין כי הקבוצה הריקה מנותצת באופן ריק על ידי $\mathcal{H}$. נתבונן ב- $\left\{ c_{1}\right\}$. מאחר ו-$\mathcal{Y}=\left\{ \pm1\right\}$, אז קיימת לפחות היפותזה אחת ב-$\mathcal{H}_{C}$, ולכל היותר 2 (אם $\mathcal{H}$ מנתצת את $\left\{ c_{1}\right\}$). לכן, אי-השוויון מתקיים עם 1 בשני הצדדים או 2 בשני צידיו.

צעד: נניח כי הטענה נכונה עבור קבוצות מגודל $k<m$, נוכיח עבור m. נסמן $C=\left\{ c_{2},...,c_{m}\right\}$. הראינו כי נוכל לייצג מחלקת היפותזות מצומצמת על קבוצה C בתור וקטור התיוג אליהם כל היפותזה ממפה את המדגם. נגדיר

$$Y_{0}=\left\{ \left(y_{2},...,y_{m}\right):\left(0,y_{2},...,y_{m}\right)\in\mathcal{H}_{C}\vee\left(1,y_{2},...,y_{m}\right)\in\mathcal{H}_{C}\right\}$$
$$Y_{1}=\left\{ \left(y_{2},...,y_{m}\right):\left(0,y_{2},...,y_{m}\right)\in\mathcal{H}_{C}\wedge\left(1,y_{2},...,y_{m}\right)\in\mathcal{H}_{C}\right\}$$

ראשית, נבחין כי Y_{0} ו-Y_{1} מוגדרות להיות קבוצות התיוגים, כלומר ההיפותזות, על הקבוצה C'. בפרט, ההיפותזות מיוצגות כוקטורי הערכים אליהם הן ממפות את איברי

C, נוכל לכתוב $Y_0 = \mathcal{H}_{C'}$. כמו כן, נבחין כי התיוגים ב-Y_1 על C' זהים למעט התיוג של c_1. לכן, האיחוד של Y_1 ו-Y_0 הינו כל התיוגים האפשריים על הקבוצה C. נתבונן ב-$|Y_0|$, $v \in Y_0$ כאשר $v \in \{0,1\}^{m-1}$ מייצג היפותזה שיתכן ומיפתה את c_1 לאפס או לאחת. כלומר, אם קיימת היפותזה כזו, היא תיספר ב-Y_0 פעם אחת. מנגד, Y_1 מחזיקה $u \in \{0,1\}^{m-1}$ כאשר ההיפותזה ממפה את c_1 לשתי התגיות. לפיכך, אם קיימות שתי היפותזות זהות למעט בקואורדינטה הראשונה, אז נספור אחת ב-Y_0 ואת השניה ב-Y_1. בסה"כ, נסיק כי $|\mathcal{H}_C| = |Y_1| + |Y_0|$.

מאחר ו-$Y_0 = \mathcal{H}_{C'}$, אז מהנחת האינדוקציה על $\mathcal{H}$ ו-C' מתקיים

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

כעת, נגדיר

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

כלומר $\mathcal{H}'$ הינה מחלקת היפותזות כך שלכל $h \in \mathcal{H}'$ קיימת היפותזה שלא מסכימה איתה על c_1 ב-$\mathcal{H}$. נבחין כי מתקיים $Y_1 = \mathcal{H}'_{C'}$, וכי מההגדרה אם $\mathcal{H}'$ מנתצת קבוצה $B \subseteq C'$, כלומר אם $\mathcal{H}'_B$ מכילה את כל התיוגים האפשריים עבור קבוצה B, אז $\mathcal{H}'_{B\cup\{c_1\}}$ מכילה את כל התיוגים האפשריים עבור $B \cup \{c_1\}$, כלומר $\mathcal{H}'$ מנתצת גם את $B \cup \{c_1\} \subseteq C$. לסיכום, נקבל מהנחת האינדוקציה על $\mathcal{H}'$ ו-C' כי

$$|Y_1| = |\mathcal{H}'_{C'}| \overset{I.H}{\leq} |\{B \subseteq C'\} : \mathcal{H}' \text{ shatters } B| = |\{B \subseteq C'\} : \mathcal{H}' \text{ shatters } B \cup \{c_1\}|$$
$$= |\{B \subseteq C\} : c_1 \in B,\ \mathcal{H}' \text{ shatters } B| \leq |\{B \subseteq C\} : c_1 \in B,\ \mathcal{H} \text{ shatters } B|$$

לפיכך

$$|\mathcal{H}_C| \leq |\{B \subseteq C\} : c_1 \notin B,\ \mathcal{H} \text{ shatters } B| + |\{B \subseteq C\} : c_1 \in B,\ \mathcal{H} \text{ shatters } B|$$
$$= |\{B \subseteq C\} : \mathcal{H} \text{ shatters } B|$$

■

(ב) החסם שהוכחנו מבטא שכמות ההיפותזות בצמצום על C (כלומר הפונקציות שמשיגות תיוגים שונים עבור איברי C) חסומה מלמעלה על ידי כמות תתי-הקבוצות של C ש-$\mathcal{H}$ מנתצת. כמות תתי הקבוצות של C היא 2^m, ולכן מההגדרה אם C עצמה מנותצת על ידי $\mathcal{H}$ אז $|\mathcal{H}_C| = 2^m$ ולכן החסם הדוק.

(ג) **טענת עזר 2:** נראה כי לכל קבוצה $C \subseteq \mathcal{X}$ מתקיים

$$|\{B \subseteq C | \mathcal{H} \text{ shatters } B\}| \leqslant \sum_{k=0}^{d} \left(\begin{array}{c} m \\ k \end{array} \right)$$

הוכחה: המספר $2^m = \sum_{k=0}^{m} \binom{m}{k}$ מייצג את כמות תתי הקבוצות ב-C, והראינו כי מספר זה חוסם את מספר תתי-הקבוצות ש-$\mathcal{H}$ מנתצת. עם זאת, מתקיים $d = \mathrm{VC}dim\left(\mathcal{H}\right)$ ו- $m > d$, אז מהגדרה לכל $B \subseteq \mathcal{X}$ ובפרט לכל $B \subseteq C$ כך ש- $|B| > d$ $\mathcal{H}$ לא מנתצת את B. לכן, נוכל להדק את החסם באמצעות חיסור תתי-הקבוצות שבהכרח אינן מנותצות, כלומר כאשר $k > d$. כלומר, כמות תתי הקבוצות המנותצות ע"י $\mathcal{H}$ הינה לכל היותר $\sum_{k=0}^{d} \left(\begin{array}{c} m \\ k \end{array} \right)$.

(ד) **מסקנה:** עבור $m > d$, מתקיים $\tau_{\mathcal{H}}(m) \leqslant \left(\frac{em}{d}\right)^d$.

הוכחת המסקנה:

$$\tau_{\mathcal{H}}(m) = \max_{C \subseteq \mathcal{X}, |C| = m} |\mathcal{H}_C| \underset{\textbf{Lemma 1}}{\leq} |\{B \subseteq C | \mathcal{H} \textbf{ shatters } B\}| \underset{\textbf{Lemma 2}}{\leq} \sum_{k=0}^{d} \left(\begin{array}{c} m \\ k \end{array} \right) \underset{\textbf{Hint}}{\leq} \left(\frac{em}{d}\right)^d$$

■

(ה) אם $m = d$, אז נוכל לבחור מדגם C שינותץ על ידי $\mathcal{H}$, כלומר יתקיים $|\mathcal{H}_C| = 2^m$. במקרה זה, נבחין כי טענת עזר 1 וטענת עזר 2 מתקיימות ($\sum_{k=0}^{d} \begin{pmatrix} m \\ k \end{pmatrix} = 2^m$) ולכן מתקיים

$$2^m = \tau_{\mathcal{H}}(m) \leq \left(\frac{em}{m}\right)^m = e^m$$

כלומר, החסם לא הדוק.

(ו) נאפיין במילים את התנהגות $\tau_{\mathcal{H}}(m)$. כאשר $m \leq VCdim\left(\mathcal{H}\right)$ נקבל אזי $\mathcal{H}_C$ גדלה מעריכית ב-m, וכאשר $m > VCdim\left(\mathcal{H}\right)$ $\mathcal{H}_C$ גדלה פולינומיאלית ב-m. מכאן, נוכל להגדיר אפיון שקול למימד VC. אם קיים $m_0 \in \mathbb{N}$ עבורו לכל $m > m_0$ הפונקציה $\tau_{\mathcal{H}}\left(m\right)$ פולינומית ב-m, אז מימד VC של $\mathcal{H}$ הינו m_0.

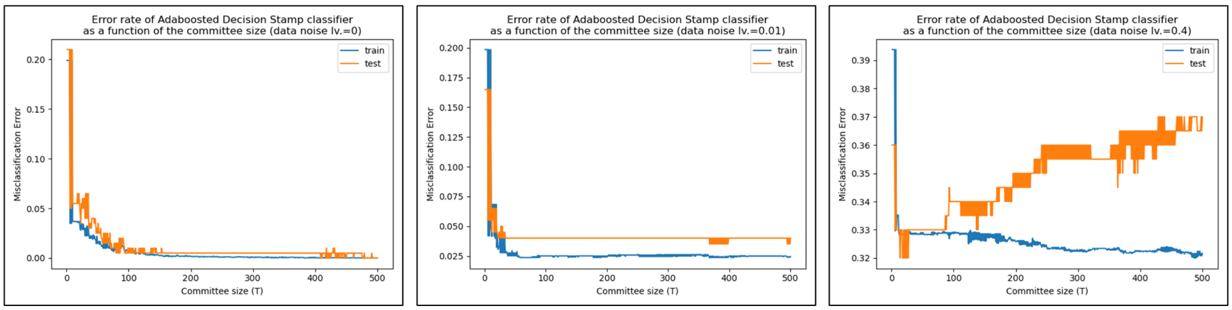
פרק מעשי

שאלה 10

נתבונן בהתפלגות הטעות של האלגוריתם כפונקציה של מספר המסווגים שהשתתפו בחיזוי, כאשר כל גרף מתאר את ההתפלגות מעל נתונים שנוצרו עם רמות רעש שונות. ניתן לראות עקומת שיפור משמעותית במעבר עבור מסווג שנסמך על מספר יחסית נמוך של מסווגים בודדים, כפי שניתן לראות בדעיכה של קצב הטעות ב-train וב-test. זאת, בהתאם לקצב הדעיכה האקספוננציאלי בטעות האמפירית כתוצאה מהשימוש ב-boosting. כמו כן, נבחין כי בהתאמה למידת הרעש בדאטה, עקומת הטעות ב-test משיגה תוצאה נמוכה יותר ככל שכמות הרעש נמוכה יותר, וכן כי הטעות במדגם ה-test וה-train יחסית מתייצבת עבור מספר גדול של מסווגים. מנגד, עבור רמת רעש של 0.4, נראה כי הטעות ב-test עולה החל משלב מסוים, הגם שבניסיונות נוספים (כתוצאה מהאקראיות בשליפת המדגמים) מדגם ה-test כן התייצב על ערך מסוים, גבוה מהערך עליו התייצב המסווג עבור דאטה מורעש ברמה 0.01.

בפועל, ככל שרמת הרעש עולה אנו מתקרבים למדגם יותר מציאותי של נתונים, ולכן ההבדלים בין רמות הטעות בגרפים השונים טמונים בהשפעות של תהליך ה-boosting על המסווג. הגדלת מספר המסווגים (גודל הועדה) בתהליך הסיווג הינם אינדיקציה לסיבוכיות מחלקת ההיפותזות, ולכן כפי שראינו בכיתה - ככל שכמות המסווגים גדלה, כך גדל הדיוק (קטן ה-bias), תוך גידול יחסית נשלט בשונות. עם זאת, ככל שהבעיה הופכת למציאותית יותר (קרי, נכנס רעש לדאטה), עבור

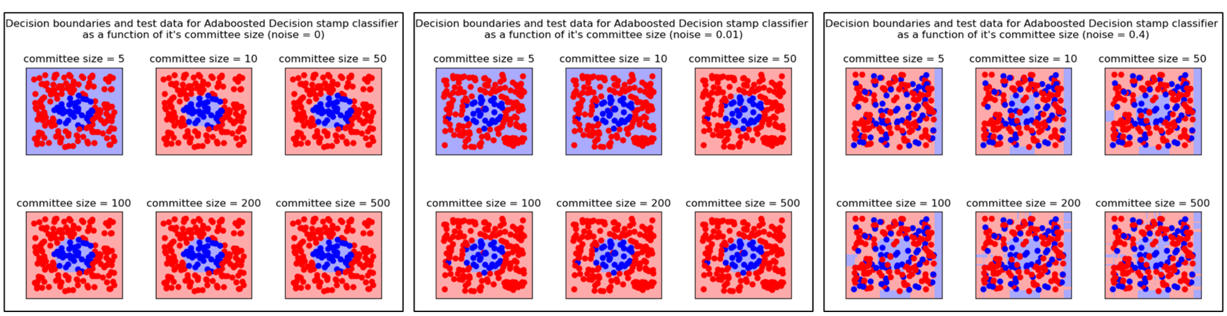
ערכי T קטנים נשיג שיפור בדיוק תוך גידול יחסית מינורי בשונות, אבל עבור T גדולים השונות תגדל בצורה משמעותית ולכן נקבל שגיאת הכללה גבוהה. בפרט, אלגוריתם adaboost רגיש לדאטה מורעש, מפני שלכל טעות בסיווג יש השפעה אקספוננציאלית על הניסיון לגשר על הטעות באמצעות עדכון ההתפלגות D^t, ולכן המסווגים h^t $t \in [T]$ "ירדפו" קודם כל אחרי השגיאות ומאחר שכל הדאטה מלא בהן אז נקבל שונות גדולה.



איור 0.1: test and train error of the adaboos decision stamp classifier on generated data as a function of the number of classifiers used in the prediction

שאלה 11

בגרף זה נוכל לראות את ה-decision boundary של המסווג לפי כמות חברי הועדה, יחד עם מדגם ה-test. נבחין כי ככל שהדאטה מורעש יותר, ה-decision boundary הופך מורכב יותר, ומשתנה יחסית לפי גודל הועדה (בהתאם לשונות שהראינו בשאלה הקודמת). בהתאמה, נראה שה-boundary נותר יחסית זהה חרף הגדלת חברי הועדה כאשר הרעש נמוך.

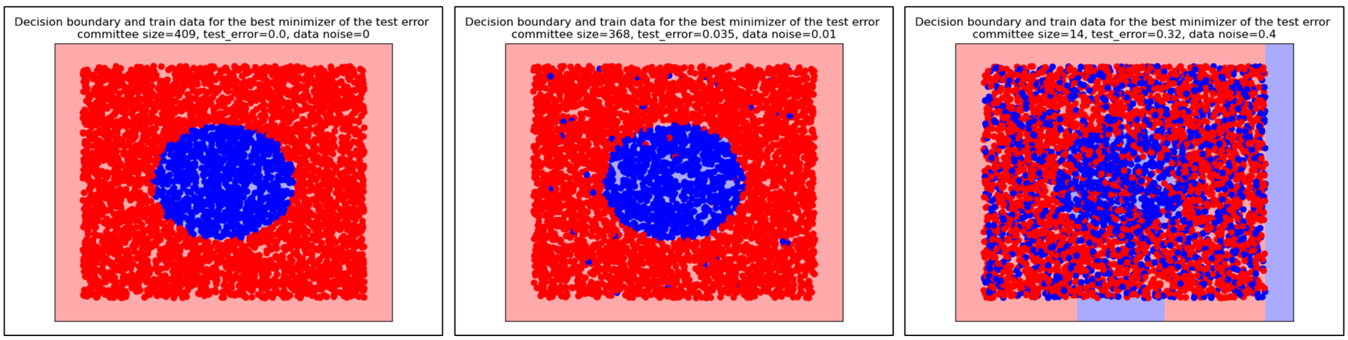


איור 0.2: Decision boundaries for different committees

שאלה 12

בגרף זה נוכל לראות את ה-decision boundary של גודל הועדה המיטבית במונחי מזעור שגיאת ההכללה על מדגם ה-test, לפי כמות הרעש בדאטה. ראשית, נבחין כי מאחר והדאטה מסונתז, אז

הנחת הרלייזביליות מתקיימת (עבור פונקציה שמתארת את העיגול), בפרט עבור הדאטה הלא מורעש - ולכן שגיאת ההכללה על ה-test מתאפסת. כמו כן, נבחין כי ככל שהדאטה מורעש יותר, גודל הועדה המיטבית הולך וקטן. הראינו בגרף בשאלה 10 כי השונות הולכת ועולה ככל שהדאטה מורעש יותר (עבור אלגוריתם adaboost) ובהתאם שגיאת ההכללה שהינה שקלול של השונות וההטייה תגדל, ומכאן האלגוריתם ישיג שיפור משמעותי כתוצאה מ-boosting, אך רק עבור גודל ועדה חסום (נמוך יחסית). למרות זאת, חרף גדלי הועדה השונים עבור הדאטה המורעש - 368 עבור רעש 0.1 ו-14 עבור רעש 0.4 - נבחין כי הטעות על מדגם ה-test דומה. עובדה זו ממחישה את היכולת המוגבלת של המסווג ללמוד מדוגמאות מורעשות.

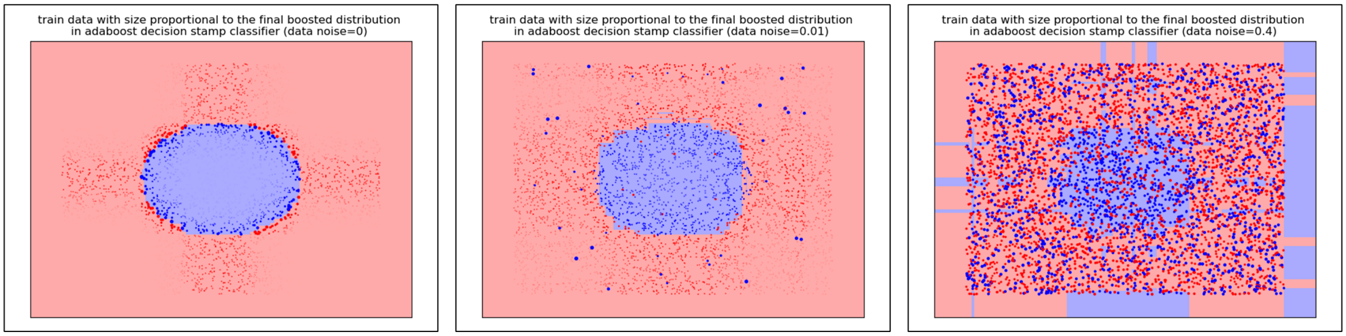


איור 0.3: Decision boundary over the train data for best minimizer committee

שאלה 13

הגרף מתאר את המשקולות של ההתפלגות המדומה D^T בסוף ריצת Adaboost. בהתאמה לאופן פעולת האלגוריתם, נקודות קטנות מסמנות דגימות שהמסווגים הצליחו להתגבר עליהן (בהכללה גסה), בעוד הדגימות הגדולות מעידות על קושי מתמשך בסיווג לאורך התאמת המסווגים h_i $i \in [T]$, בין אם המסווג ה-h_T הצליח או לא לסווג אותן. במילים אחרות, גודל הדגימות מסמן את מידת הקושי של המסווגים השונים לסווג אותן נכונה. ה-Decision boundary המשורטט בכל גרף מתאר את המסווג הראשי שמסתמך על 500 המסווגים לפי המשקולות שלהם. ראשית, נבחין כי כאשר הדאטה אינו מורעש, הדגימות הגדולות הינן בדיוק סביב השפה של המעגל, בגבול בין התיוגים. מאחר והראינו כי טעות ההכללה על ה-test של האלגוריתם הינה 0, וכפי שניתן לראות ב-decision boundary עצמו, המסווג $h_{boosted}$ הצליח לסווג את כל הדגימות נכונה, והדגימות הגדולות לאורך תהליך בניית T המסווגים סייעו בהגדרה של גבולות המעגל. עבור הדאטה המורעש, ניכר כי הטעות הנגררת הולכת וגדלה, ובעיקר כי גם בסוף התהליך האלגוריתם לא הצליח לסווג נכונה את הדגימות. נשים לב

לצורת הצלב של הדגימות של הדגימות שנוצרה בשני הגרפים משמאל - בהתאם לאופן הסיווג של Decision stamp שמחלק את המרחב לפי הצירים (עבור דגימה ב-$\mathbb{R}^2$).



איור 0.4: Train samples proportional to their weights in the end of the iterative proccess