Contents

מבוא למערכות לומדות תרגיל 4

עמית בסקין 312259013

1 למידות־PAC

1. יהיו A אלגוריתם למידה ו־ $\mathcal{D}$ התפלגות כלשהי.

   נניח שפונקציית ה־ Loss נמצאת בטווח $[0,1]$.

   צ.להוכיח ששתי הטענות הבאות שקולות:

   - לכל $\epsilon,\delta>0$ יש $m\left(\epsilon,\delta\right)$ כך שלכל $m\geq m\left(\epsilon,\delta\right)$ מתקיים:

$$\underset{S\sim\mathcal{D}^{m}}{\mathbb{P}}\left[L_{\mathcal{D}}\left(A\left(S\right)\right)\leq\epsilon\right]\geq1-\delta$$

   - מתקיים:

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

הוכחה:

כיוון ראשון: נניח שלכל $\epsilon,\delta>0$ יש $m\left(\epsilon,\delta\right)$ כך שלכל $m\geq m\left(\epsilon,\delta\right)$ מתקיים:

$$\underset{S\sim\mathcal{D}^{m}}{\mathbb{P}}\left[L_{\mathcal{D}}\left(A\left(S\right)\right)\leq\epsilon\right]\geq1-\delta$$

ונראה ש־

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

אכן:

יהא $m\geq m\left(\epsilon,\delta\right)$. ניקח $\epsilon=\delta=\frac{1}{m}$. נסמן $x=L_{\mathcal{D}}\left(A\left(S\right)\right)$. נתבונן בפונקציית ההצטברות $f\left(x\right)$ שמקיימת:

$$\mathbb{P}\left(a\leq x\leq b\right)=\int_{a}^{b}f\left(x\right)dx$$

נתון שפונקציית ה־ Loss נמצאת בטווח $[0,1]$ ולכן זהו תחום האינטגרציה. אז נקבל:

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(A\left(S\right)\right)\right]=\int\limits_0^1 xf\left(x\right)dx$$

נפצל לתחומים:

$$=\int\limits_0^{1/m} xf\left(x\right)dx+\int\limits_{1/m}^1 xf\left(x\right)dx\leq$$

$$\leq\frac{1}{m}\int\limits_0^{1/m} f\left(x\right)dx+\int\limits_{1/m}^1 f\left(x\right)dx$$

אבל לכל a,b מתקיים:

$$\mathbb{P}\left(a\leq x\leq b\right)=\int_a^b f\left(x\right)dx\leq 1$$

ולכן:

$$\frac{1}{m}\int\limits_0^{1/m} f\left(x\right)dx+\int\limits_{1/m}^1 f\left(x\right)dx\leq\frac{1}{m}+\int\limits_{1/m}^1 f\left(x\right)dx$$

ועתה מההנחה:

$$\underset{S\sim\mathcal{D}^m}{\mathbb{P}}\left[L_{\mathcal{D}}\left(A\left(S\right)\right)\leq\frac{1}{m}\right]\geq 1-\frac{1}{m}$$

נובע ש־

$$\underset{S\sim\mathcal{D}^m}{\mathbb{P}}\left[L_{\mathcal{D}}\left(A\left(S\right)\right)>\frac{1}{m}\right]<\frac{1}{m}$$

קרי:

$$\frac{1}{m}+\int\limits_{1/m}^1 f\left(x\right)dx<\frac{1}{m}+\frac{1}{m}=\frac{2}{m}$$

ניקח את m לאינסוף ונקבל:

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

כנדרש.

כיוון שני: נניח ש־

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

ונראה שלכל $\epsilon,\delta>0$ יש $m\left(\epsilon,\delta\right)$ כך שלכל $m\geq m\left(\epsilon,\delta\right)$ מתקיים:

$$\underset{S\sim\mathcal{D}^m}{\mathbb{P}}\left[L_{\mathcal{D}}\left(A\left(S\right)\right)\leq\epsilon\right]\geq 1-\delta$$

אכן: יהיו $\epsilon, \delta > 0$. לפי א"ש מרקוב מתקיים:

$$\underset{S\sim\mathcal{D}^m}{\mathbb{P}}\left[L_{\mathcal{D}}\left(A\left(S\right)\right)\leq\epsilon\right]\leq\frac{\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(A\left(S\right)\right)\right]}{\epsilon}$$

עתה מההנחה ש־

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

אזי שקיים M כך שלכל $m > M$ מתקיים ש־

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(A\left(S\right)\right)\right]\leq\epsilon\left(1-\delta\right)$$

אז יהא $m > M$ ונקבל:

$$\frac{\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(A\left(S\right)\right)\right]}{\epsilon}\leq\frac{\epsilon\left(1-\delta\right)}{\epsilon}=1-\delta$$

קרי:

$$\underset{S\sim\mathcal{D}^m}{\mathbb{P}}\left[L_{\mathcal{D}}\left(A\left(S\right)\right)\leq\epsilon\right]\leq1-\delta$$

כלומר:

$$m\left(\epsilon,\delta\right)=M$$

מש"ל.

2. יהיו $\mathcal{X}=\mathbb{R}^2, \mathcal{Y}=\{0,1\}$ ו־ $\mathcal{H}=\{h_r \mid r\in\mathbb{R}_+\}$ מחלקה של היפותיזות כאשר $h_r\left(x\right)=\mathbf{1}\left[\|x\|_2\leq r\right]$.

צ.להוכיח ש־ $\mathcal{H}$ היא למידה־PAC וש־

$$m_{\mathcal{H}}\left(\epsilon,\delta\right)\leq\frac{\log\left(1/\delta\right)}{\epsilon}$$

הוכחה:

מהנחת הריאליזביליות יש $h_r\in\mathcal{H}$ עם $L_{\mathcal{D}}\left(h_r\right)=0$.

תהא $S=\{(x_i,y_i)\}_{i=1}^m$ קבוצה של m דגימות שנדגמו באופן זהה ובלתי תלוי מהתפלגות $\mathcal{D}$ מעל $\mathcal{X}\times\mathcal{Y}$.

נשים לב שמאחר ש־ $L_{\mathcal{D}}\left(h_r\right)=0$ אזי ש־

$$\underset{(x,y)\sim\mathcal{D}}{\mathbb{P}}\left[\|x\|_2>r\text{ and }y=1\right]=0$$

ניקח:

$$r_{alg}=\begin{cases}\max_{i\in[m]}\left\{\|x_i\|_2\in\mathbb{R}_+\mid y_i=1\right\} & \exists i\in[m]\left(y_i=1\right)\\ 0 & \text{else}\end{cases}$$

יהא r' ממשי אי־שלילי עם $r' < r$ כך ש־

$$\underset{x\sim\mathcal{D}_{\mathcal{X}}}{\mathbb{P}}\left[r' \leq \|x\|_2 \leq r\right] = \epsilon$$

ונשים לב שכדי שיתקיים $L_{\mathcal{D}}\left(h_{r_{alg}}\right) \leq \epsilon$ מספיק שיתקיים ש־ $r_{alg} \geq r'$, כלומר:

$$\mathbb{P}\left[L_{\mathcal{D}}\left(h_{r_{alg}}\right) > \epsilon\right] \leq \mathbb{P}\left[r_{alg} < r'\right]$$

אבל $r_{alg} < r'$ אסם לכל $(x_i, y_i) \in S$, מתקיים ש־ $\|x_i\|_2 \notin [r', r]$, קרי:

$$= \underset{S\sim\mathcal{D}^m}{\mathbb{P}}\left[\forall i \in [m]\left(\|x_i\|_2 \notin [r', r]\right)\right] = (1-\epsilon)^m \leq e^{-\epsilon m}$$

ומאחר שאנחנו מניחים ש־ $m > \frac{\log(1/\delta)}{\epsilon}$, אזי ש־

$$\leq e^{-\epsilon\frac{\log(1/\delta)}{\epsilon}} = e^{-\log(1/\delta)} = \frac{1}{e^{\log(1/\delta)}} = \frac{1}{1/\delta} = \delta$$

ומכאן ש־ $\mathcal{H}$ הינה למידה־PAC וכן:

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \frac{\log(1/\delta)}{\epsilon}$$

כנדרש. ■

2 מימד VC

3. יהיו $\mathcal{Y} = \{0,1\}$, $\mathcal{X} = \{0,1\}^d$ עם $d \geq 2$.

כל דגימה $(\mathrm{x}, y) \in \mathcal{X} \times \mathcal{Y}$ מורכבת מהשמה של d משתנים בוליאניים ותווית.

לכל משתנה בוליאני x_k עם $k \in [d]$ ישנם שני ליטרלים: x_k ו־ $\overline{x_k} = 1 - x_k$.

נגדיר מחלקה $\mathcal{H}_{\text{con}}$: כל היפותזה ניתנת על ידי גימומים של תת־קבוצה כלשהי מתוך $2d$ הליטרלים הנ"ל.

צ.לחשב את מימד ה־ VC של $\mathcal{H}_{\text{con}}$.

תזכורות:

- תהא $C \subseteq \mathcal{X}$. הצמצום של $\mathcal{H}$ ל־ C הוא:

$$\mathcal{H}_C = \left\{h_C \in \{0,1\}^C \mid \exists h \in \mathcal{H} \text{ s.t. } \forall x \in C\left(h_C(x) = h(x)\right)\right\}$$

- נאמר ש־ $\mathcal{H}$ מנתצת קבוצה סופית $C \subseteq \mathcal{X}$ אם $|\mathcal{H}_C| = 2^{|C|}$.
- הגדרנו:

$$\mathrm{VC}\dim(\mathcal{H}) = \sup\left\{m \in \mathbb{N} \mid \exists C \subseteq \mathcal{X}\left(|C| = m \text{ and } \mathcal{H} \text{ shatters } C\right)\right\}$$

- כדי להראות ש־ $\mathrm{VC}\dim(\mathcal{H})=d$ צריך להראות ש־
  - יש $C\subseteq\mathcal{X}$ מגודל d שמנותצת על ידי $\mathcal{H}$, קרי $\mathrm{VC}\dim(\mathcal{H})\geq d$
  - כל $C\subseteq\mathcal{X}$ מגודל $d+1$ אינה מנותצת על ידי $\mathcal{H}$, קרי $\mathrm{VC}\dim(\mathcal{H})\leq d$

פיתרון: נטען כי $\mathrm{VC}\dim(\mathcal{H})=d$. הוכחה:

- $\mathrm{VC}\dim(\mathcal{H})\geq d$: ניקח את קבוצת הוקטורים הסטנדרטיים ב־ $\mathbb{R}^d$, קרי: $C=\left\{e_i\in\mathbb{R}^d\mid i\in[d]\right\}$ ונראה שניתן לנתץ אותה על ידי $\mathcal{H}$:

אכן, יהא $\mathrm{y}\in\{0,1\}^d$ וקטור תוויות עבור הוקטורים ב־ C. אז ניקח:

$$h=\underset{y_i=0}{\wedge}\overline{x_i}$$

ראשית לכל $e_i\in C$ ולכל $j\neq i$ מתקיים שהכניסה ה־ j ב־ e_i היא אפס, ומאחר שכל המשתנים בגימום של h הם ליטרליים שליליים אזי שההשמה של כניסה j ב־ e_i בתוך משתנה $\overline{x_j}$ בגימום של h (אם הוא מופיע), נותן 1.

עתה, אם $y_i=1$ אז $\overline{x_i}$ לא משתתף בגימום ובסה"כ מתקבל גימום של אחדות ולכן $h(e_i)=1$. מצד שני אם $y_i=0$ אז $\overline{x_i}$ משתתף בגימום ומקבל השמה של 0 ובסה"כ מתקבל גימום של אחדות ואפס ולכן $h(e_i)=0$.

- $\mathrm{VC}\dim(\mathcal{H})\leq d$: תהא קבוצה C עם $|C|=d+1$. נניח בשלילה שניתן להתאים עבור C כל וקטור תוויות y. נמספר את הוקטורים ב־ $C=\{c_i\}_1^{d+1}$. אז בפרט לכל c_i יש h_C כך ש־ $h_C^i(c_i)=0$ ולכל $j\neq i$ מתקיים $h_C^i(c_j)=1$.

לכל היפותיזה h_C^i כנ"ל, נבחר ליטרל z_i עליו c_i מקבל 0, קיים כזה כי $h_C^i(c_i)$.

נתבונן בקבוצת הליטרלים שקיבלנו: $\{z_i\}_1^{d+1}$. מעקרון שובך היונים יש שני ליטרלים z_i,z_j שמתייחסים לאותו משתנה. נחלק למקרים:

  - אם $z_i=z_j$ אז c_i,c_j שניהם מאפסים את z_i בסתירה לכך שאחד מהם אמור לקבל תיוג 0 והשני 1.
  - אם $z_i=\overline{z_j}$ בה"כ $z_i=x_k$ ו־ $z_j=\overline{x_k}$. יהא $l\notin\{i,j\}$. אז l מקבל 1 גם x_k וגם על $\overline{x_k}$ בסתירה.

מש"ל

3 PAC־אגנוסטיות

4. צ.להוכיח שאם ל־ $\mathcal{H}$ יש את תכונת ההתפלגות האחידה עם פונקציה

$$m_{\mathcal{H}}^{UC}:\ (0,1)^2\to\mathbb{N}$$

אז $\mathcal{H}$ היא למידה־PAC־אגנוסטית עם סיבוכיות של

$$m_{\mathcal{H}}(\epsilon,\delta)\leq m^{UC}\left(\frac{\epsilon}{2},\delta\right)$$

תזכורת:

- קבוצת אימון S_m נקראת ϵ־מייצגת עבור $\mathcal{D},\mathcal{H},L$ אם לכל $h\in\mathcal{H}$ מתקיים: $|L_S(h)-L_{\mathcal{D}}(h)|<\epsilon$.
- נאמר שלמחלקת היפותיזות $\mathcal{H}$ יש את תכונת ההתכנסות האחידה אם יש $m_{\mathcal{H}}^{UC}:\ (0,1)^2\to\mathbb{N}$ כך שלכל $\epsilon>0$ ו־ $\delta<1$ ולכל התפלגות $\mathcal{D}$ על $\mathcal{X}\times\mathcal{Y}$ מתקיים שאם $m\geq m_{\mathcal{H}}^{UC}(\epsilon,\delta)$ אז $m\geq1-\delta$ $\mathcal{D}^m\left(\{S_m\in(\mathcal{X}\times\mathcal{Y})^m\mid S\text{ is }\epsilon\text{-representative }\}\right)\geq1-\delta$.

- מחלקת היפותיזות $\mathcal{H}$ נקראת למידה־PAC־אגנוסטית ביחס לפונקציית Loss $L\colon \mathcal{H}\times(\mathcal{X}\times\mathcal{Y})\to[0,\infty)$ אם קיימת פונקציה $\widetilde{m}_{\mathcal{H}}\colon (0,1)^2\to\mathbb{N}$ ואלגוריתם למידה $\mathcal{A}_m\colon (\mathcal{X}\times\mathcal{Y})^m\to\mathcal{H}$ עם התכונה הבאה: לכל $\epsilon,\delta\in(0,1)$ ולכל התפלגות $\mathcal{D}$ מעל $\mathcal{X}\times\mathcal{Y}$ ולכל $m\geq\widetilde{m}_{\mathcal{H}}(\epsilon,\delta)$ מתקיים ש־ $\mathcal{D}^m\left(\left\{S_m\mid L_{\mathcal{D}}\left(h_{S_m}\right)\leq\min_{h\in\mathcal{H}}L_{\mathcal{D}}(h)+\epsilon\right\}\right)\geq 1-\delta$ כאשר $S_m=\{(x_i,y_i)\}_1^m$ דגימות שנדגמו בצורה זהה ובלתי־תלויה מ־ $\mathcal{D}$, וכן $h_{S_m}=\mathcal{A}_m(S_m)$.

למה: תהא S_m קבוצת אימון $\frac{\epsilon}{2}$־מייצגת עבור $\mathcal{D},\mathcal{H},L$. תהא h_{S_m} פלט כלשהו של $ERM_{\mathcal{H}}(S_m)$, קרי $h_{S_m}\in\arg\min_{h\in\mathcal{H}}L_{S_m}(h)$.
אז $L_{\mathcal{D}}\left(h_{S_m}\right)<\min_{h\in\mathcal{H}}L_{\mathcal{D}}(h)+\epsilon$.

הוכחה: מכך ש־ S_m היא $\frac{\epsilon}{2}$־מייצגת עבור $\mathcal{D},\mathcal{H},L$, אזי שלכל $h\in\mathcal{H}$ מתקיים:

$$\left|L_{S_m}(h)-L_{\mathcal{D}}(h)\right|<\frac{\epsilon}{2}\implies\begin{cases}L_{S_m}(h)-L_{\mathcal{D}}(h)<\frac{\epsilon}{2} & \star_1\\ L_{\mathcal{D}}(h)-L_{S_m}(h)<\frac{\epsilon}{2} & \star_2\end{cases}$$

ובפרט זה מתקיים עבור $h=h_{S_m}$, אז מ־ $\star_2$ נקבל:

$$L_{\mathcal{D}}\left(h_{S_m}\right)-L_{S_m}\left(h_{S_m}\right)<\frac{\epsilon}{2}$$

קרי:

$$L_{\mathcal{D}}\left(h_{S_m}\right)<L_{S_m}\left(h_{S_m}\right)+\frac{\epsilon}{2}$$

עתה ניקח $h'\in\arg\min_{h\in\mathcal{H}}L_{\mathcal{D}}(h)$, ומאחר שב־ h' אילוץ המינמום חזק יותר, אזי ש־

$$\leq L_{S_m}(h')+\frac{\epsilon}{2}$$

ומ־ $\star_1$ ב־ $\epsilon/2$־רפרסנטטיביות:

$$\leq L_{\mathcal{D}}(h')+\epsilon$$

כנדרש.

הוכחת השאלה המרכזית:

ל־ $\mathcal{H}$ יש את תכונת ההתכנסות האחידה, קרי יש $m_{\mathcal{H}}^{UC}\colon (0,1)^2\to\mathbb{N}$ כך שלכל $0<\epsilon$ ו־ $1>\delta$ ולכל התפלגות $\mathcal{D}$ על $\mathcal{X}\times\mathcal{Y}$ מתקיים שאם $m\geq m_{\mathcal{H}}^{UC}(\epsilon,\delta)$ אז $\mathcal{D}^m\left(\{S_m\in(\mathcal{X}\times\mathcal{Y})^m\mid S\text{ is }\epsilon\text{-representative }\}\right)\geq 1-\delta$.

אז יהיו $\epsilon,\delta\in(0,1)$ וניקח $m\geq m_{\mathcal{H}}^{UC}\left(\frac{\epsilon}{2},\delta\right)$.

מהלמה אנחנו יודעים שאם S_m היא $\frac{\epsilon}{2}$־מייצגת אז $L_{\mathcal{D}}\left(h_{S_m}\right)<\min_{h\in\mathcal{H}}L_{\mathcal{D}}(h)+\epsilon$.

אז ממונוטוניות פונקציית ההסתברות נקבל ש־

$$\underset{S\sim\mathcal{D}^m}{\mathbb{P}}\left[L_{\mathcal{D}}\left(h_{S_m}\leq\min_{h\in\mathcal{H}}L_{\mathcal{D}}(h)+\epsilon\right)\right]\geq\mathcal{D}^m\left(\left\{S_m\in(\mathcal{X}\times\mathcal{Y})^m\mid S\text{ is }\frac{\epsilon}{2}\text{-representative }\right\}\right)$$

ומקיום תכונת ההתכנסות האחידה:

$$\geq 1-\delta$$

כלומר קיבלנו את $m_{\mathcal{H}}^{UC}\colon\ (0,1)^2 \to \mathbb{N}$ מקיום תכונת ההתכנסות האחידה, ואלגוריתם הלמידה שלנו

$$\mathcal{A}_m\colon (\mathcal{X}\times\mathcal{Y})^m \to \mathcal{H}$$

הוא h_{S_m} והוא מקיים שלכל $\epsilon,\delta\in(0,1)$ ולכל התפלגות $\mathcal{D}$ מעל $\mathcal{X}\times\mathcal{Y}$ ולכל $m\geq \widetilde{m}_{\mathcal{H}}(\epsilon,\delta)$ מתקיים ש־

$$\mathcal{D}^m\left(\left\{S_m \mid L_{\mathcal{D}}(h_{S_m}) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h)+\epsilon\right\}\right)\geq 1-\delta$$

כאשר $S_m=\{(x_i,y_i)\}_1^m$ דגימות שנדגמו בצורה זהה ובלתי־תלויה מ־ $\mathcal{D}$, וזוהי בדיוק ההגדרה של למידות־PAC־אגנוסטית. מש"ל

■

5. תהא $\mathcal{H}$ מחלקת היפותיזות מעל $Z=\mathcal{X}\times\{\pm1\}$.

נתבונן בפונקציית ההפסד 0-1.

נניח שיש פונקציה $m_{\mathcal{H}}(\epsilon,\delta)$ כך שלכל התפלגות $\mathcal{D}$ מעל Z יש אלגוריתם A עם התכונה הבאה:

כאשר מריצים את A על $m\geq m_{\mathcal{H}}(\epsilon,\delta)$ דגימות שנדגמות בהתפלגות זהה ובאופן בלתי־תלוי שנלקחות על ידי $\mathcal{D}$, אז מובטח ש־ A יחזיר עם הסתברות של לפחות $1-\delta$, היפותיזה $h_S\colon\mathcal{X}\to\{\pm1\}$ עם $L_{\mathcal{D}}(h_S)\leq\min_{h\in\mathcal{H}}L_{\mathcal{D}}(h)+\epsilon$.

צ.לבדוק האם $\mathcal{H}$ היא למידה־PAC־אגנוסטית.

<u>בחרתי לענות על שאלה 4.</u>

4 מונוטוניות

6. תהא $\mathcal{H}$ מחלקת היפותיזות עבור משימת קלסיפקציה בינארית.

נניח ש־ $\mathcal{H}$ היא למידה־PAC ושהסיבוכיות ניתנת על ידי $m_{\mathcal{H}}(\cdot,\cdot)$.

צ.להראות ש־ $m_{\mathcal{H}}$ היא מונוטונית לא־עולה בכל אחד מהפרמטרים שלה, קרי בהינתן $\delta\in(0,1)$ ו־ $0<\epsilon_1\leq\epsilon_2<1$ מתקיים ש־

$$m_{\mathcal{H}}(\epsilon_1,\delta)\geq m_{\mathcal{H}}(\epsilon_2,\delta)$$

ובהינתן $\epsilon\in(0,1)$ ו־ $0<\delta_1\leq\delta_2<1$ מתקיים ש־

$$m_{\mathcal{H}}(\epsilon,\delta_1)\geq m_{\mathcal{H}}(\epsilon,\delta_2)$$

<u>בחרתי לענות על שאלה 7.</u>

7. יהיו $\mathcal{H}_1,\mathcal{H}_2$ מחלקות של קלסיפקציה בינארית כך ש־ $\mathcal{H}_1\subseteq\mathcal{H}_2$. צ.להראות ש־

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

תזכורות:

- תהא $C\subseteq\mathcal{X}$. הצמצום של $\mathcal{H}$ ל־ C הוא:

$$\mathcal{H}_C=\left\{h_C\in\{0,1\}^C \mid \exists h\in\mathcal{H}\text{ s.t. }\forall x\in C\,(h_C(x)=h(x))\right\}$$

- נאמר ש־ $\mathcal{H}$ מנתצת קבוצה סופית $C \subseteq \mathcal{X}$ אם $|\mathcal{H}_C| = 2^{|C|}$.
- הגדרנו:

$$\mathrm{VC}\dim\left(\mathcal{H}\right) = \sup\left\{m \in \mathbb{N} \mid \exists C \subseteq \mathcal{X}\left(|C| = m \text{ and } \mathcal{H} \text{ shatters } C\right)\right\}$$

- כדי להראות ש־ $\mathrm{VC}\dim\left(\mathcal{H}\right) = d$ צריך להראות ש־
  - יש $C \subseteq \mathcal{X}$ מגודל d שמנותצת על ידי $\mathcal{H}$, קרי $\mathrm{VC}\dim\left(\mathcal{H}\right) \geq d$
  - כל $C \subseteq \mathcal{X}$ מגודל $d+1$ אינה מנותצת על ידי $\mathcal{H}$, קרי $\mathrm{VC}\dim\left(\mathcal{H}\right) \leq d$

הוכחת הטענה:

$$\mathrm{VC}\dim\left(\mathcal{H}_1\right) = \sup\left\{m \in \mathbb{N} \mid \exists C \subseteq \mathcal{X}\left(|C| = m \text{ and } \mathcal{H}_1 \text{ shatters } C\right)\right\} =$$

$$= \sup\left\{m \in \mathbb{N} \mid \exists C \subseteq \mathcal{X}\left(|C| = m \text{ and } \left|\left(\mathcal{H}_1\right)_C\right| = 2^m\right)\right\} =$$

$$= \sup\left\{m \in \mathbb{N} \mid \exists C \subseteq \mathcal{X}\left(|C| = m \text{ and } \left|\left\{h_C \in \{0,1\}^C \mid \exists h \in \mathcal{H}_1 \text{ s.t. } \forall x \in C\left(h_C\left(x\right) = h\left(x\right)\right)\right\}\right| = 2^m\right)\right\}$$

ומאחר ש־ $\mathcal{H}_1 \subseteq \mathcal{H}_2$:

$$\leq \sup\left\{m \in \mathbb{N} \mid \exists C \subseteq \mathcal{X}\left(|C| = m \text{ and } \left|\left\{h_C \in \{0,1\}^C \mid \exists h \in \mathcal{H}_2 \text{ s.t. } \forall x \in C\left(h_C\left(x\right) = h\left(x\right)\right)\right\}\right| = 2^m\right)\right\} =$$

$$= \sup\left\{m \in \mathbb{N} \mid \exists C \subseteq \mathcal{X}\left(|C| = m \text{ and } \left|\left(\mathcal{H}_2\right)_C\right| = 2^m\right)\right\} =$$

$$= \sup\left\{m \in \mathbb{N} \mid \exists C \subseteq \mathcal{X}\left(|C| = m \text{ and } \mathcal{H}_2 \text{ shatters } C\right)\right\} =$$

$$= \mathrm{VC}\dim\left(\mathcal{H}_2\right)$$

כנדרש.

5 טענה תיאורתית

8. יהא $\mathcal{X}$ מרחב דגימות ו־ $\mathcal{Y} = \{\pm 1\}$.

תהא $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ מחלקת היפותיזות.

עבור $C \subseteq \mathcal{X}$, ניזכר בסימון $\mathcal{H}_C$ עבור הצמצום של $\mathcal{H}$ לקבוצה C.

נגדיר את הפונקציה $\tau_m\left(\mathcal{H}\right) : \mathbb{N} \to \mathbb{N}$ שמתאימה ל־ $\mathcal{H}$ על ידי:

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

א. צ.להסביר מהי המשמעות של הפונקציה $\tau_{\mathcal{H}}$. הסבר:

ראשית, עבור $|C| = m$ תמיד מתקיים ש־ $|\mathcal{H}_C| \leq 2^m$, כי מספר התיוגים המקסימלי הוא 2^m, קרי 0 או 1 לכל איבר ב־ C. ולכן $\tau_{\mathcal{H}}(m) \leq 2^m$לכל m.

עתה, אם $\tau_{\mathcal{H}}(m) = 2^m$ אז $\mathrm{VC}\dim(\mathcal{H}) \geq m$ כי $\tau_{\mathcal{H}}(m) = 2^m$ גורר שקיימת קבוצה מגודל m שניתנת לניתוץ. כמו כן, אם $\tau_{\mathcal{H}}(m) < 2^m$ אז $\mathrm{VC}\dim(\mathcal{H}) < m$ כי זה שלא קיימת קבוצה מגודל m שניתנת לניתוץ.

אם כן, במידה ש־ $\mathrm{VC}\dim(\mathcal{H}) < m$, הפונקציה $\tau_{\mathcal{H}}(m)$ יכולה לתת מושג עד כמה m רחוק מלהיות מימד ה־ VC של $\mathcal{H}$. אם ההפרש בין $\tau_{\mathcal{H}}(m)$ הוא "גדול" אז מימד ה־ VC של $\mathcal{H}$ "הרבה" יותר קטן מ־ m, ואם ההפרש קטן אז מימד ה־ VC של $\mathcal{H}$ הוא "כמעט" m. ובמילים אחרות המספר $\tau_{\mathcal{H}}(m)$ נותן את מספר ההיפותיזות המקסימלי האפשרי עבור תת־קבוצה מגודל m של קבוצת המדגם.

ב. נניח ש־ $\mathrm{VC}\dim(\mathcal{H}) = \infty$. צ.למצוא ביטוי עבור הערך של $\tau_{\mathcal{H}}(m)$ עבור $m \in \mathbb{N}$. פיתרון:

המשמעות של $\mathrm{VC}\dim(\mathcal{H}) = \infty$ היא ש־ $\mathcal{H}$ מנתצת כל קבוצה מגודל m לכל m טבעי ולכן, בהסתמך על הסעיף הקודם, לכל m מתקיים ש־ $\tau_{\mathcal{H}}(m) = 2^m$.

ג. נניח ש־ $\mathrm{VC}\dim(\mathcal{H}) = d$. צ.למצוא ביטוי עבור הערך של $\tau_{\mathcal{H}}(m)$ עבור $m \leq d$. פיתרון:

d הוא הגודל המקסימלי של קבוצה ש־ $\mathcal{H}$ מנתצת, אז מכך ש־ $\mathrm{VC}\dim(\mathcal{H}) = d$ נובע שקיימת קבוצה $C \subseteq \mathcal{X}$ עם $|C| = d$ כך ש־ $|\mathcal{H}_C| = 2^d$. ואז אם $m \leq d$ אזי שכל קבוצה C' עם $|C'| = m$ ו־ $C' \subseteq C$, מתקיים ש־ $|\mathcal{H}_{C'}| = 2^m$ שהרי:

$$\mathcal{H}_C = \left\{ h_C \in \{0,1\}^C \mid \exists h \in \mathcal{H} \text{ s.t. } \forall x \in C \left(h_C(x) = h(x) \right) \right\}$$

קרי אם יש $h \in \mathcal{H}$ כך ש־ לכל $x \in C$ מתקיים ש־ $h_C(x) = h(x)$, אז בפרט לכל $x \in C' \subseteq C$ $h_{C'}(x) = h(x)$, ולכן אם מתקבלת כל קלסיפקציה אפשרית עבור C אז בהכרח גם עבור C' ומכאן ש־

$$\tau_{\mathcal{H}}(m) = 2^m$$

מש"ל

ד. נניח ש־ $\mathrm{VC}\dim(H) = d$. יהא $m > d$. נראה ש־

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

עבור e בסיס הלוגריתם הטבעי. נעשה זאת בשלבים הבאים:

i. צ.להראות באינדוקציה שלכל $C \subseteq \mathcal{X}$ סופית, מתקיים כי:

$$|\mathcal{H}_C| \leq |\{ B \subseteq C \mid \mathcal{H} \text{ shutters } B \}|$$

הוכחה:

באינדוקציה על $m = |C|$:

בסיס האינדוקציה: עבור $m = 1$: יש איבר אחד ב־ C.

אם ניתן לתייג אותו רק באפס או רק באחד אז $|\mathcal{H}_C| = 1$ ואז הקבוצה היחידה שמוכלת ב־ C ושמנותצת היא הקבוצה הריקה ולכן גם צד ימין שווה לאחד. אם ניתן לתייג את האיבר ב־ C גם באפס וגם באחד אז צד שמאל בא"ש שווה לשתיים ואז ניתן לנתץ גם את C עצמה וגם את הקבוצה הריקה ולכן גם צד ימין הוא 2. מש"ל בסיס האינדוקציה.

צעד: נניח שהטענה נכונה לכל $k < m$ עבור $m > 1$ ונוכיח עבור m: נמספר את האיברים ב־ C: $C = \{c_i\}_1^m$ וכן נגדיר $C' = \{c_i\}_2^m$.

כמו כן נגדיר:

$$Y_0 = \{(y_2, \dots, y_m) \mid (0, y_2, \dots, y_m) \in \mathcal{H}_C \vee (1, y_2, \dots, y_m) \in \mathcal{H}_C\}$$

$$Y_1 = \{(y_2, \dots, y_m) \mid (0, y_2, \dots, y_m) \in \mathcal{H}_C \wedge (1, y_2, \dots, y_m) \in \mathcal{H}_C\}$$

נשים לב שמתקיים כי $Y_1 \subseteq Y_0$ כי לכל $(y_2, \dots, y_m)$ עם $(0, y_2, \dots, y_m) \in \mathcal{H}_C \wedge (1, y_2, \dots, y_m) \in \mathcal{H}_C$ בפרט מתקיים ש־ $(0, y_2, \dots, y_m) \in \mathcal{H}_C \vee (1, y_2, \dots, y_m) \in \mathcal{H}_C$. כמו כן, לכל $(y_2, \dots, y_m) \in Y_1$ מתקיים ש־ $(0, y_2, \dots, y_m) \in \mathcal{H}_C$ וגם $(1, y_2, \dots, y_m) \in \mathcal{H}_C$ ולכל $(y_2, \dots, y_m) \in Y_0 \backslash Y_1$ מתקיים ש־ $(0, y_2, \dots, y_m) \in \mathcal{H}_C$ XOR $(1, y_2, \dots, y_m) \in \mathcal{H}_C$ ולכן $|\mathcal{H}_C| = 2|Y_1| + |Y_0 \backslash Y_1| = |Y_0| + |Y_1|$.

עבור היפותיזה $h_C \in \mathcal{H}_C$ שנותנת את התווית y_i עבור c_i, נסמן: $h_C = (y_1, y_2, \dots, y_m)$ ובאופן דומה נסמן גם עבור $h_{C'} \in \mathcal{H}_{C'}$.

אם כן, על פי ההגדרה מתקיים:

$$\mathcal{H}_{C'} = \left\{h_{C'} \in \{0,1\}^{C'} \mid \exists h_C \in \mathcal{H}_C \text{ s.t. } \forall x \in C' \left(h_{C'}(x) = h(x)\right)\right\} =$$

$$= \left\{(y_2, \dots, y_m) \in \{0,1\}^{C'} \mid (0, y_2, \dots, y_m) \in \mathcal{H}_C \vee (1, y_2, \dots, y_m) \in \mathcal{H}_C\right\} = Y_0$$

אז מהנחת האינדוקציה:

$$|Y_0| = |\mathcal{H}_{C'}| \leq |\{B \subseteq C' \mid \mathcal{H} \text{ shutters } B\}|$$

אבל $C' = C \backslash \{c_1\}$ ולכן:

$$= |\{B \subseteq C \mid \mathcal{H} \text{ shutters } B \text{ and } c_1 \notin B\}|$$

נגדיר $\mathcal{H}' \subseteq \mathcal{H}$ על ידי:

$$\mathcal{H}' = \{h \in \mathcal{H} \mid \exists h' \in \mathcal{H} \text{ s.t. } \left(1 - h'(c_1), h'(c_2), \dots, h'(c_m)\right) = \left(h(c_1), h(c_2), \dots, h(c_m)\right)\}$$

ונשים לב ש־

$$\mathcal{H}'_{C'} = \left\{h_{C'} \in \{0,1\}^{C'} \mid \exists h' \in \mathcal{H}' \text{ s.t. } \forall x \in C' \left(h'_{C'}(x) = h'(x)\right)\right\} =$$

$$= \left\{h_{C'} \in \{0,1\}^{C'} \mid \exists h \in \mathcal{H}' \text{ s.t. } \exists h' \in \mathcal{H} \text{ s.t. } \left(1 - h'(c_1), h'(c_2), \dots, h'(c_m)\right) = \left(h(c_1), h(c_2), \dots, h(c_m)\right) \text{ and } \forall x \in C' \left(h'_{C'}(x) = h(x)\right)\right\} =$$

$$= \left\{(y_2, \dots, y_m) \in \{0,1\}^{C'} \mid (\overline{y_1}, y_2, \dots, y_m) \in \mathcal{H}_C \wedge (y_1, y_2, \dots, y_m) \in \mathcal{H}_C\right\} = Y_1$$

אז מהנחת האינדוקציה נקבל ש־

$$|Y_1| = |\mathcal{H}'_{C'}| \leq |\{B \subseteq C' \mid \mathcal{H}' \text{ shutters } B\}|$$

אבל לכל $(y_2, \ldots, y_m) \in \mathcal{H}'_{C'}$ מתקיים ש־ $(y_1, y_2, \ldots, y_m) \in \mathcal{H}_C \wedge (\overline{y_1}, y_2, \ldots, y_m) \in \mathcal{H}_C$, ולכן אם $\mathcal{H}'$ מנתצת את B אז היא מנתצת גם את $B \cup \{c_1\}$, קרי:

$$= |\{B \subseteq C' \mid \mathcal{H}' \text{ shutters } B \cup \{c_1\}\}| =$$

$$= |\{B \subseteq C' \mid \mathcal{H}' \text{ shutters } B \text{ and } c_1 \in B\}|$$

ומאחר ש־ $\mathcal{H}' \subseteq \mathcal{H}$:

$$\leq |\{B \subseteq C' \mid \mathcal{H} \text{ shutters } B \text{ and } c_1 \in B\}|$$

ולבסוף:

$$|\mathcal{H}_C| = |Y_0| + |Y_1| \leq$$

$$\leq |\{B \subseteq C \mid \mathcal{H} \text{ shutters } B \text{ and } c_1 \notin B\}| + |\{B \subseteq C' \mid \mathcal{H} \text{ shutters } B \text{ and } c_1 \in B\}| =$$

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

כנדרש. ■

ii. צ.להסביר את המשמעות של אי־השוויון הנ"ל. הסבר:

אם יש ב־ $\mathcal{H}_C$ m היפותיזות אז יש ב־ C לפחות m תת־קבוצות ש־ $\mathcal{H}$ מנתצת אותן.

iii. צ.להראות שלכל קבוצה סופית $C \subseteq \mathcal{X}$, מתקיים:

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

הוכחה:

נתון ש־ $d =$VCdim$(\mathcal{H})$ ולכן $\mathcal{H}$ יכולה לנתץ רק קבוצות שהן מגודל d ומטה. אז כל הקבוצות ב־ C ש־ $\mathcal{H}$ יכולה לנתץ הן תתי־הקבוצות מגודל שהוא קטן או שווה ל־ d. אז לכל $k \leq d$, יש $\binom{m}{k}$ תתי־קבוצות ב־ C שהן מגודל k, ולכן מתקיים הא"ש הנ"ל.

iv. צ.להשתמש באי־השוויון הבא:

$$\sum_{k=0}^{d} \binom{m}{k} \leq \left(\frac{em}{d}\right)^d$$

כדי לסיים את ההוכחה ש־

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

הוכחה:

$$\tau_{\mathcal{H}}(m) := \max\{|\mathcal{H}_C| \mid C \subseteq \mathcal{X} \text{ and } |C| = m\}$$

ומאחר שהטענות שהוכחנו לעיל נכונות לכל C מגודל m:

$$\le \max\{|\{B \subseteq C \mid \mathcal{H} \text{ shutters } B\}| \mid C \subseteq \mathcal{X} \text{ and } |C| = m\} \le$$

$$\le \sum_{k=0}^{d} \begin{pmatrix} m \\ k \end{pmatrix}$$

ומאחר ש־ $m > d+1$:

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

כנדרש. ■

ה. נניח ש־ $m = d$. צ.לבדוק האם מתקיים הא"ש $\tau_{\mathcal{H}}(m) \le \left(\frac{em}{d}\right)^{d}$. אם כן, האם הוא הדוק? פיתרון:

הא"ש נשמר כי:

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

וככל ש־ d גדול יותר, כך החסם יותר הדוק כי ככל ש־ d גדל כך קטן ההפרש $e^{d} - 2^{d} = (e-2)^{d} \approx (0.718)^{d}$.

ו. צ.לאפיין במילים את ההתנהגות של $\tau_{\mathcal{H}}(m)$ עבור $m \le \mathrm{VC}\dim(\mathcal{H})$ ועבור $m > \mathrm{VC}\dim(\mathcal{H})$, ולהשתמש בו כדי להציע הגדרה אלטרנטיבית עבור $\mathrm{VC}\dim(\mathcal{H})$. פיתרון:

עבור $m > \mathrm{VC}\dim(\mathcal{H})$ מתקיים ש־ $\tau_{\mathcal{H}}(m) < 2^{m}$, קרי מתקבלת התנהגות פולינומיאלית ועבור $m \ge \mathrm{VC}\dim(\mathcal{H})$ מתקיים $\tau_{\mathcal{H}}(m) = 2^{m}$ קרי מתקבלת התנהגות אקספוננציאלית.

מכאן נובע שאם גרף הפונקציה של $\tau_{\mathcal{H}}(m)$ הוא אקספוננציאלי ב־ m תמיד, אז $\mathcal{H}$ אינה למידה־PAC כי אז $\mathrm{VC}\dim(\mathcal{H}) = \infty$, ואילו אם הגרף של $\tau_{\mathcal{H}}(m)$ הינו פולינומי החל ממקום מסויים אזי ש־ $\mathcal{H}$ הינה למידה־PAC כי מתקיים ש־ $\mathrm{VC}\dim(\mathcal{H}) < \infty$.

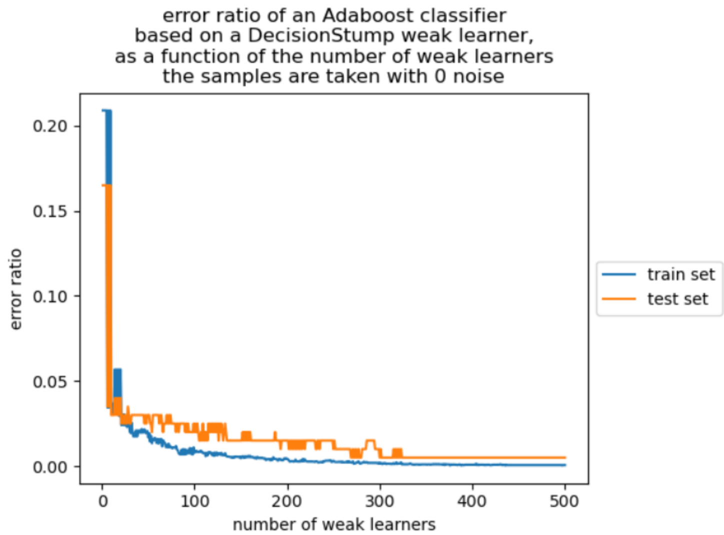
אז נוכל להגדיר:

$$\mathrm{VC}\dim(\mathcal{H}) = \min\{m \in \mathbb{N} \mid \tau_{\mathcal{H}}(m) < 2^{m}\}$$

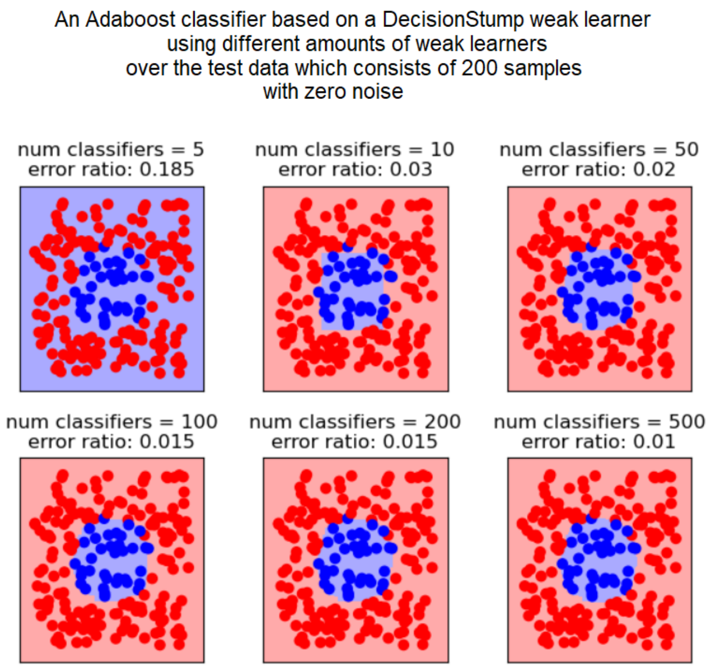
ואז אם נסמן ב־ $\mathrm{VC}\widehat{\dim}(\mathcal{H})$ את מימד ה־ VC לפי ההגדרה החדשה, אז נקבל ש־ $\mathrm{VC}\dim(\mathcal{H}) = \mathrm{VC}\widehat{\dim}(\mathcal{H}) - 1$ כאשר $\mathrm{VC}\dim(\mathcal{H})$ הוא מימד ה־ VC לפי ההגדרה המקורית: $\mathrm{VC}\widehat{\dim}(\mathcal{H}) = d$ אזי שכל קבוצה מגודל d ומעלה לא ניתנת לניתוץ ואילו $d-1$ ניתנת לניתוץ ולכן $\mathrm{VC}\dim(\mathcal{H}) = d$. מש"ל

6 להפריד את הבלתי ניתן להפרדה - Adaboost

.10



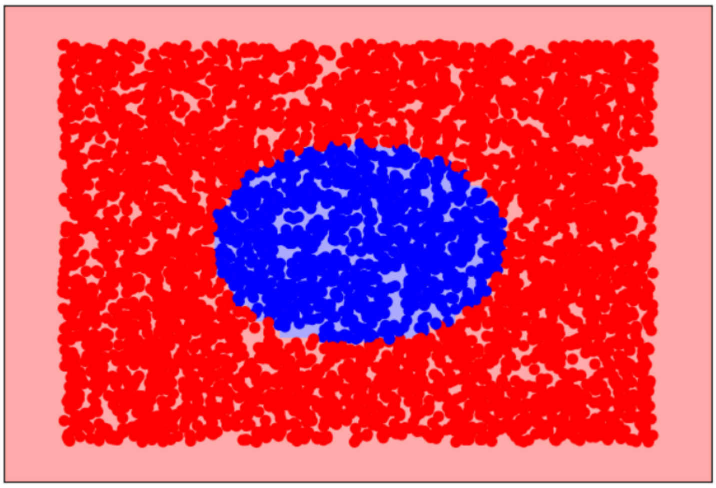
.11



.12 בשאלה הקודמת ניתן לראות ש־ $T = 500$ ממזער את השגיאה עבור ה־ test set.

An Adaboost classifier based on a DecisionStump weak learner
over the training data which consists of 5,000 samples
with zero noise

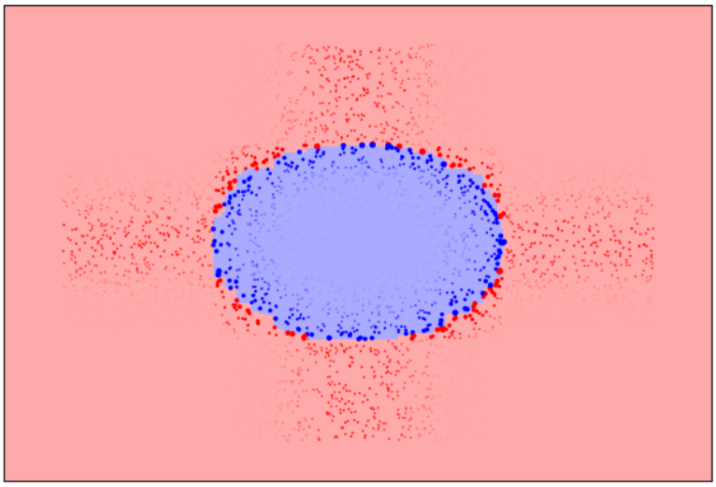
num classifiers = 500
error ratio = 0.0



.13

The training set of an Adaboost classifier
with size proportional to the transpose of the weights
of the samples in the last iteration of the training
with 5,000 training samples
with zero noise

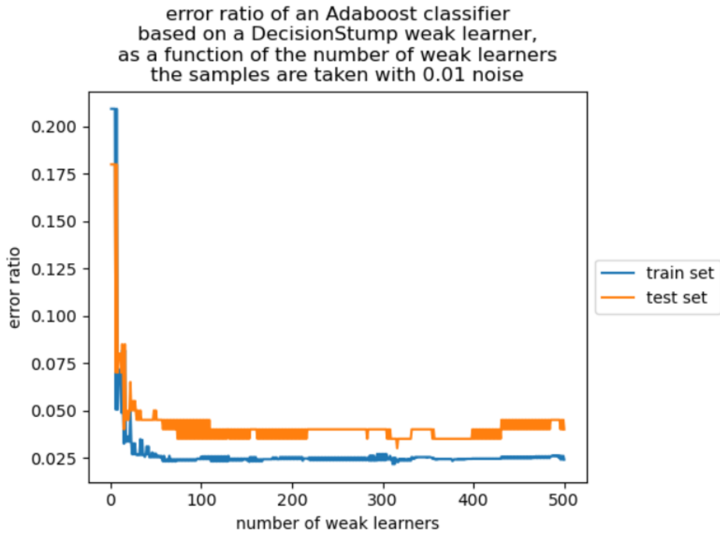
num classifiers = 500
error ration = 0.0002

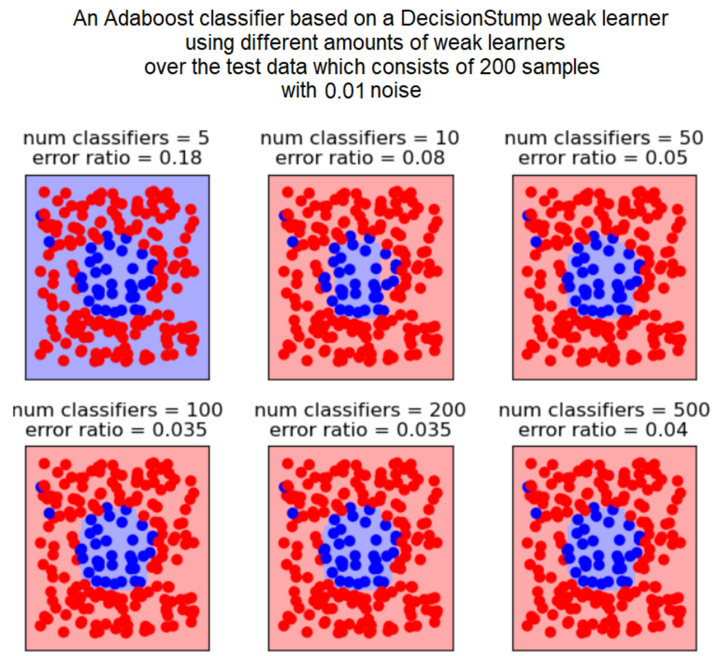


בשוליים של העיגול הנקודות גדולות יותר כי האלגוריתם טעה עליהן הרבה פעמים במהלך תהליך הלמידה ולכן הוא נתן להן משקל גדול יותר מלנקודות באזורים אחרים. וזאת משום שקשה לסווג את השוליים באמצעות חלוקות אופקיות ואנכיות בלבד.

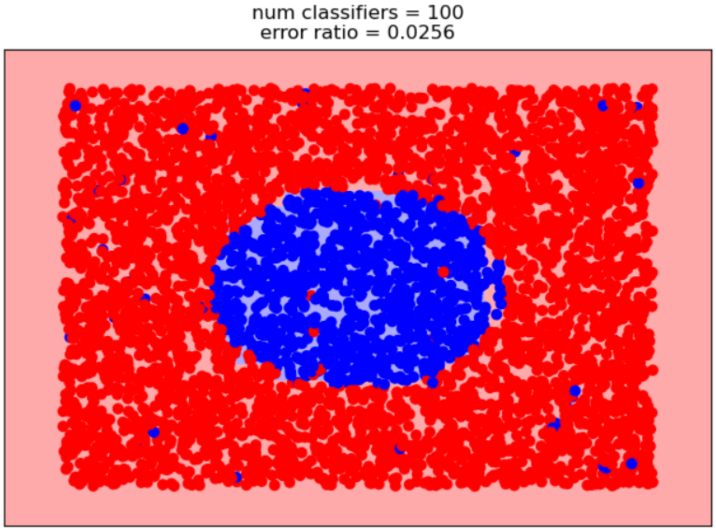
כמו כן, הנקודות מחוץ לעיגול קטנות יותר כי האלגוריתם טעה עליהן פחות באופן יחסי, ולכן נתן להן משקל קטן יותר. וזאת משום שקל יותר לסווג אותן על ידי חלוקות אופקיות ואנכיות.

.14

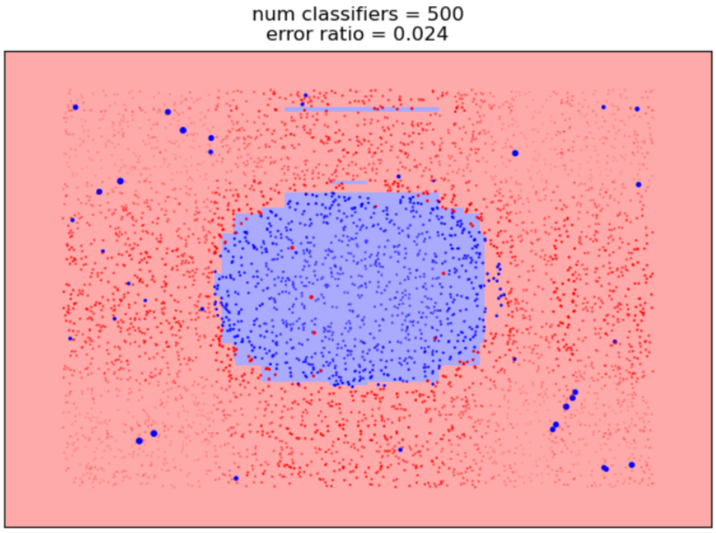

error ratio of an Adaboost classifier
based on a DecisionStump weak learner,
as a function of the number of weak learners
the samples are taken with 0.01 noise
error ratio
number of weak learners
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0
100
200
300
400
500
train set
test set

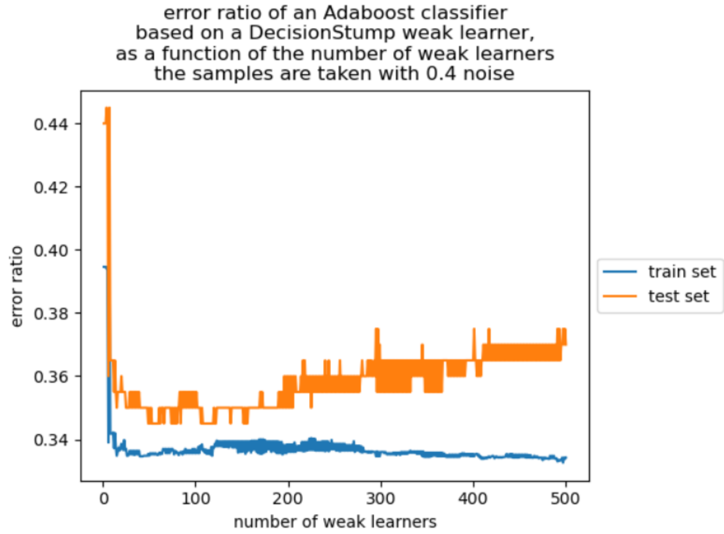

An Adaboost classifier based on a DecisionStump weak learner
using different amounts of weak learners
over the test data which consists of 200 samples
with 0.01 noise
num classifiers = 5
error ratio = 0.18
num classifiers = 10
error ratio = 0.08
num classifiers = 50
error ratio = 0.05
num classifiers = 100
error ratio = 0.035
num classifiers = 200
error ratio = 0.035
num classifiers = 500
error ratio = 0.04

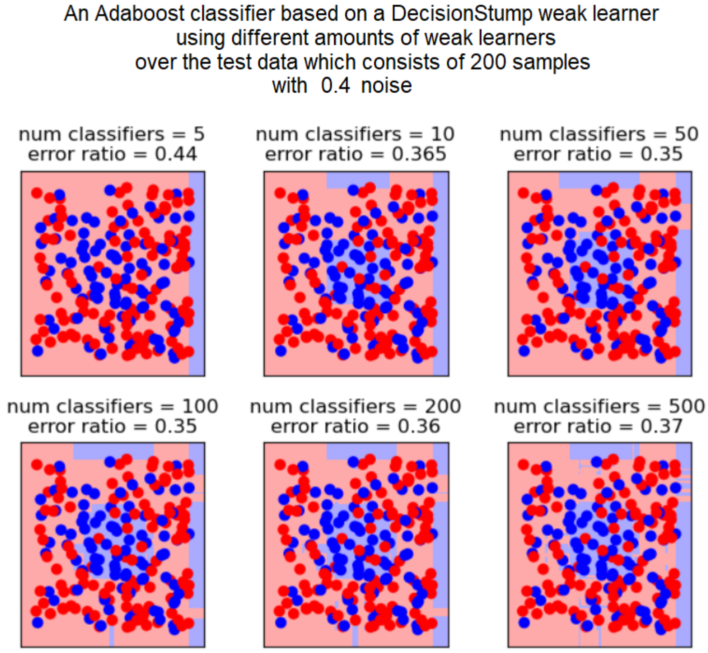
An Adaboost classifier based on a DecisionStump weak learner
over the training data which consists of 5,000 samples
with 0.01 noise



The training set of an Adaboost classifier
with size proportional to the transpose of the weights
of the samples in the last iteration of the training
with 5,000 training samples
with 0.01 noise

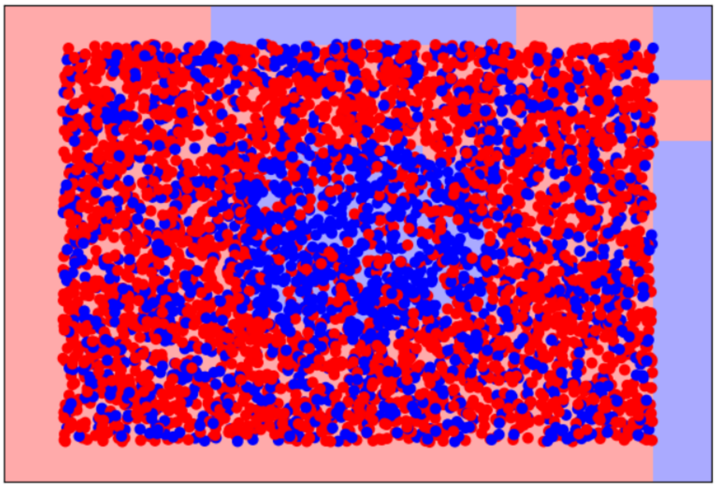



error ratio of an Adaboost classifier
based on a DecisionStump weak learner,
as a function of the number of weak learners
the samples are taken with 0.4 noise
error ratio
0.44
0.42
0.40
0.38
0.36
0.34
0
100
200
300
400
500
number of weak learners
train set
test set


An Adaboost classifier based on a DecisionStump weak learner
using different amounts of weak learners
over the test data which consists of 200 samples
with 0.4 noise
num classifiers = 5
error ratio = 0.44
num classifiers = 10
error ratio = 0.365
num classifiers = 50
error ratio = 0.35
num classifiers = 100
error ratio = 0.35
num classifiers = 200
error ratio = 0.36
num classifiers = 500
error ratio = 0.37

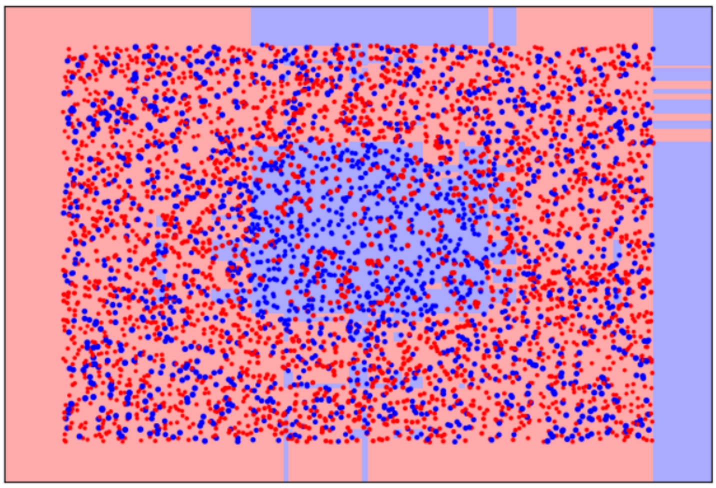
An Adaboost classifier based on a DecisionStump weak learner
over the training data which consists of 5,000 samples
with 0.4 noise

num classifiers = 50
error ratio = 0.3348



The training set of an Adaboost classifier
with size proportional to the transpose of the weights
of the samples in the last iteration of the training
with 5,000 training samples
with 0.4 noise

num classifiers = 500
error ratio = 0.3342



הסבר כללי לשינויים:

כשהרעש יחסית קטן, קרי 0.01, ניתן לראות שהשגיאה בסיווג היא יחסית קטנה ומספר המסווגים האופטימלי הוא עדיין המקסימלי מבין האפשרויות, וכשהרעש יחסית גדול, קרי 0.4, ניתן לראות שהשגיאה גדלה משמעותית, וכך גם מספר המסווגים האופטימלי מבין האפשרויות, נעשה קטן יותר באופן משמעותי.

הסבר של השינויים ביחס לשאלה 10:

ככל שהדאטא נתון עם יותר רעש, כך bias קטן יותר, קרי שימוש במספר גדול של לומדים חלשים, גורם להתאמת־יתר חמורה יותר, קרי variance גדול יותר. לכן ככל שהרעש גדל כך ניתן לראות שהשגיאה קבוצת המבחן גדלה, והפער בין מידת השגיאה בקבוצת האימון ובין השגיאה בקבוצת המבחן, גדל.

הסבר של השינויים ביחס לשאלה 12:

בהסתמך על ההסבר שנתתי לעיל עבור השינויים ביחס לשאלה 10, נובע שנרצה להשתמש בפחות לומדים חלשים על מנת להקטין את ה־ variance, ואכן כשהרעש יחסית גדול, קרי 0.4, נקבל שגיאה מינימלית עבור 50 מסווגים ולא עבור 500 כמו במקרה ללא הרעש.

2 adaboost.py

```python
"""
===================================================
    Introduction to Machine Learning (67577)
===================================================

Skeleton for the AdaBoost classifier.

Author: Gad Zalcberg
Date: February, 2019

"""
from ex4_tools import *


TRAIN_SAMPLES_AMOUNT = 5000
TEST_SAMPLES_AMOUNT = 200
DEFAULT_NOISE = 0
Q14_NOISES = [0.01, 0.4]
DEFAULT_CLASSIFIERS_AMOUNT = 500
CLASSIFIERS_AMOUNTS_LST = [5, 10, 50, 100, 200, 500]


class AdaBoost(object):

    def __init__(self, WL, T):
        """
        Parameters
        ----------
        WL : the class of the base weak learner
        T : the number of base learners to learn
        """
        self.WL = WL
        self.T = T
        self.h = [None]*T     # list of base learners
        self.w = np.zeros(T)  # weights

    def train(self, X, y):
        """
        Parameters
        ----------
        X : samples, shape=(num_samples, num_features)
        y : labels, shape=(num_samples)
        Train this classifier over the sample (X,y)
        After finish the training return the weights of the samples in the
        last iteration.
        """
        m = len(X)
        D = np.ones(m) / m
        for i in range(self.T):
            h = self.WL(D, X, y)
            h.train(D, X, y)
            y_pred = h.predict(X)
            self.h[i] = h
            zero_indices = np.where(y_pred == 0)[0]
            y_pred[zero_indices] = -1
            compare = (y != y_pred).astype(int)
            epsilon = D @ compare
            epsilon_inverse = 1 / epsilon
            in_log = epsilon_inverse - 1
```

```python
            curr_log = np.log(in_log)
            w = 0.5 * curr_log
            self.w[i] = w
            y_and_y_pred_product = np.multiply(y, y_pred)
            in_exp = -w * y_and_y_pred_product
            exp_arr = np.exp(in_exp)
            D = np.multiply(D, exp_arr)
            D /= np.sum(D)
        return D

    def predict(self, X, max_t):
        """
        Parameters
        ----------
        X : samples, shape=(num_samples, num_features)
        :param max_t: integer < self.T: the number of classifiers to use for
        the classification
        :return: y_hat : a prediction vector for X. shape=(num_samples)
        Predict only with max_t weak learners,
        """
        lst = [self.w[i] * self.h[i].predict(X) for i in range(max_t)]
        sign = np.sign(np.sum(lst, axis=0))
        return sign

    def error(self, X, y, max_t):
        """
        Parameters
        ----------
        X : samples, shape=(num_samples, num_features)
        y : labels, shape=(num_samples)
        :param max_t: integer < self.T: the number of classifiers to use for
        the classification
        :return: error : the ratio of the wrong predictions when predict only
        with max_t weak learners (float)
        """
        pred = self.predict(X, max_t)
        return np.sum(pred != y) / len(y)


class ExeSolver:
    def __init__(self, noise):
        self.noise = noise
        self.train_samples = generate_data(TRAIN_SAMPLES_AMOUNT, noise)
        self.X_train, self.y_train = self.train_samples[0], \
                                     self.train_samples[1]
        self.adaboost = AdaBoost(DecisionStump, DEFAULT_CLASSIFIERS_AMOUNT)
        self.D = self.adaboost.train(self.X_train, self.y_train)
        self.test_samples = generate_data(TEST_SAMPLES_AMOUNT, noise)
        self.X_test = self.test_samples[0]
        self.y_test = self.test_samples[1]


def q10(exe_solver):
    T = range(1, DEFAULT_CLASSIFIERS_AMOUNT+1)
    train_errors = \
        [exe_solver.adaboost.error(
            exe_solver.X_train, exe_solver.y_train, i) for i in T]
    test_errors = \
        [exe_solver.adaboost.error(
            exe_solver.X_test, exe_solver.y_test, i) for i in T]
    plt.plot(T, train_errors, label='train set')
    plt.plot(T, test_errors, label='test set')
    plt.legend(loc='center left', bbox_to_anchor=(1, 0.5))
    plt.title('error ratio of an Adaboost classifier\n'
              'based on a DecisionStump weak learner,\n'
              'as a function of the number of weak learners\n'
              'the samples are taken with {} noise'.format(exe_solver.noise))
    plt.xlabel('number of weak learners')
```

```python
    plt.ylabel('error ratio')
    plt.show()


def q11(exe_solver):
    counter = 1
    errors_and_classifiers_amounts = []
    for t in CLASSIFIERS_AMOUNTS_LST:
        plt.subplot(2, 3, counter)
        error = decision_boundaries(exe_solver.adaboost, exe_solver.X_test,
                                    exe_solver.y_test, t)
        errors_and_classifiers_amounts.append((error, t))
        counter += 1
    plt.show()
    return min(errors_and_classifiers_amounts, key=lambda x: x[0])[1]


def q12(exe_solver):
    amount = q11(exe_solver)
    decision_boundaries(
        exe_solver.adaboost, exe_solver.X_train, exe_solver.y_train, amount)
    plt.show()


def q13(exe_solver):
    new_D = (exe_solver.D / np.max(exe_solver.D) * 10).T
    decision_boundaries(exe_solver.adaboost, exe_solver.X_train,
                        exe_solver.y_train, DEFAULT_CLASSIFIERS_AMOUNT,
                        weights=new_D)
    plt.show()


def run_questions(noise):
    exe_solver = ExeSolver(noise)
    q10(exe_solver)
    q12(exe_solver)
    q13(exe_solver)


def run_zero_noise():
    run_questions(0)


def q14():
    for noise in Q14_NOISES:
        run_questions(noise)


# run questions 10-13 with zero noise
# run_zero_noise()

# run question 14
# q14()
```

3 ex4 tools.py

```
"""
===================================================
     Introduction to Machine Learning (67577)
===================================================

This module provides some useful tools for Ex4.

Author: Gad Zalcberg
Date: February, 2019

"""
import numpy as np
import matplotlib.pyplot as plt
from matplotlib.colors import ListedColormap
from itertools import product
from matplotlib.pyplot import imread
import os
from sklearn.model_selection import train_test_split


cm = ListedColormap(['#AAAAFF', '#FFAAAA'])
cm_bright = ListedColormap(['#0000FF', '#FF0000'])


def find_threshold(D, X, y, sign, j):
    """
    Finds the best threshold.
    D =  distribution
    S = (X, y) the data
    """
    # sort the data so that x1 <= x2 <= ... <= xm
    sort_idx = np.argsort(X[:, j])
    X, y, D = X[sort_idx], y[sort_idx], D[sort_idx]

    thetas = np.concatenate([[-np.inf], (X[1:, j] + X[:-1, j]) / 2, [np.inf]])
    minimal_theta_loss = np.sum(D[y == sign])  # loss of the smallest
    # possible theta
    losses = np.append(minimal_theta_loss, minimal_theta_loss -
                       np.cumsum(D * (y * sign)))
    min_loss_idx = np.argmin(losses)
    return losses[min_loss_idx], thetas[min_loss_idx]


class DecisionStump(object):
    """
    Decision stump classifier for 2D samples
    """

    def __init__(self, D, X, y):
        self.theta = 0
        self.j = 0
        self.sign = 0
        self.train(D, X, y)

    def train(self, D, X, y):
        """
        Train the classifier over the sample (X,y) w.r.t. the weights D over X
        Parameters
        ----------
```

```
        D : weights over the sample
        X : samples, shape=(num_samples, num_features)
        y : labels, shape=(num_samples)
        """
        loss_star, theta_star = np.inf, np.inf
        for sign, j in product([-1, 1], range(X.shape[1])):
            loss, theta = find_threshold(D, X, y, sign, j)
            if loss < loss_star:
                self.sign, self.theta, self.j = sign, theta, j
                loss_star = loss

    def predict(self, X):
        """
        Parameters
        ----------
        X : shape=(num_samples, num_features)
        Returns
        -------
        y_hat : a prediction vector for X shape=(num_samples)
        """
        y_hat = self.sign * ((X[:, self.j] <= self.theta) * 2 - 1)
        return y_hat


def decision_boundaries(classifier, X, y, num_classifiers=1, weights=None):
    """
    Plot the decision boundaries of a binary classfiers over X \subseteq R^2

    Parameters
    ----------
    classifier : a binary classifier, implements classifier.predict(X)
    X : samples, shape=(num_samples, 2)
    y : labels, shape=(num_samples)
    title_str : optional title
    weights : weights for plotting X
    """
    cm = ListedColormap(['#AAAAFF', '#FFAAAA'])
    cm_bright = ListedColormap(['#0000FF', '#FF0000'])
    h = .003  # step size in the mesh
    # Plot the decision boundary.
    x_min, x_max = X[:, 0].min() - .2, X[:, 0].max() + .2
    y_min, y_max = X[:, 1].min() - .2, X[:, 1].max() + .2
    xx, yy = np.meshgrid(np.arange(x_min, x_max, h), np.arange(y_min, y_max, h))
    Z = classifier.predict(np.c_[xx.ravel(), yy.ravel()], num_classifiers)
    # Put the result into a color plot
    Z = Z.reshape(xx.shape)
    plt.pcolormesh(xx, yy, Z, cmap=cm)
    # Plot also the training points
    if weights is not None:
        plt.scatter(X[:, 0], X[:, 1], c=y, s=weights, cmap=cm_bright)
    else:
        plt.scatter(X[:, 0], X[:, 1], c=y, cmap=cm_bright)
    plt.xlim(xx.min(), xx.max())
    plt.ylim(yy.min(), yy.max())
    plt.xticks([])
    plt.yticks([])
    error = classifier.error(X, y, num_classifiers)
    plt.title('num classifiers = {}\nerror ratio = {}'.format(
        num_classifiers, error))
    plt.draw()
    return error


def generate_data(num_samples, noise_ratio):
    '''
    generate samples X with shape: (num_samples, 2) and labels y with shape (num_samples).
    num_samples: the number of samples to generate
    noise_ratio: invert the label for this ratio of the samples
```

```
    '''
    X = np.random.rand(num_samples, 2) * 2 - 1
    radius = 0.5 ** 2
    in_circle = np.sum(X ** 2, axis=1) < radius
    y = np.ones(num_samples)
    y[in_circle] = -1
    y[np.random.choice(num_samples, int(noise_ratio * num_samples))] *= -1

    return X, y
```