318442241 - יאיר שטק          IML - תרגיל 4 - חלק תאורטי

## 2.1 PAC Learnability

1. For $\mathcal{A}$ some learning algorithm, $\mathcal{D}$ a probability distribution over $\mathcal{X}$ and the 0-1 loss function (i.e misclassification), prove the following are equivalent:

$(a)\quad \forall\varepsilon,\delta>0 \quad \exists m(\varepsilon,\delta) \quad \text{s.t} \quad \forall m\geq m(\varepsilon,\delta) \quad \mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(\mathcal{A}(S))\leq\varepsilon]\geq 1-\delta$

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

Hints:

- For $(a)\Rightarrow(b)$ show that $\forall\varepsilon,\delta>0$ and $\forall m\geq m(\varepsilon,\delta)$ it holds that $\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(\mathcal{A}(S))]\leq\varepsilon+\delta$.
- For $(b)\Rightarrow(a)$ use Markov's inequality

כיון ראשון: $(\Uparrow)$

יהיו $\varepsilon,\delta>0$ ונניח ש- $\lim_{m\to\infty}\mathbb{E}_{S\sim D^m}[L_D(A(S))]=0$

מא"ש מרקוב

$$P_{S\sim D^m}\left[L_D(A(S))\le\varepsilon\right]=1-P_{S\sim D^m}\left[L_D(A(S))>\varepsilon\right]\ge 1-\frac{\mathbb{E}[L_D(A(S))]}{\varepsilon}$$

כעת נראה ש- $1-\frac{\mathbb{E}[L_D(A(S))]}{\varepsilon}\ge 1-(1-\delta)=\delta$

נבחין שמספיק שנראה שלכל $m\ge m(\varepsilon,\delta)$ מתקיים $\mathbb{E}[L_D(A(S))]\le\delta\varepsilon$

ואכן בהינתן $\lim_{m\to\infty}\mathbb{E}_{S\sim D^m}[L_D(A(S))]=0$ - מהגדרת הגבול לכל $m\ge m_0$

מתקיים $\lim_{m\to\infty}\mathbb{E}_{S\sim D^m}[L_D(A(S))]\le\varepsilon\delta$ לכן אם נגדיר $m(\varepsilon,\delta)=m_0$ נקבל

מה שרצינו.

כיון שני: $(\Downarrow)$

נרצה להוכיח ש- $\lim_{m\to\infty}\mathbb{E}_{S\sim D^m}[L_D(A(S))]=0$ ז"א לפי הגדרת הגבול

$$\forall\varepsilon>0\quad\exists m_0\in\mathbb{N}\quad\forall m>m_0\quad\mathbb{E}_{S\sim D^m}[L_D(A(S))]<\varepsilon$$

כאשר נתון $P_{S\sim D^m}[L_D(A(S))\le\varepsilon]\ge 1-\delta$ לכל $m\ge m(\varepsilon,\delta)$

$$\mathbb{E}_{S\sim D^m}[L_D(A(S))]=\int_0^1 x\cdot P[L_D(A(S))=x]\,dx\le$$

$$\le\int_0^{\frac{\varepsilon}{4}}\frac{\varepsilon}{4}\cdot P[L_D(A(S))=x]\,dx+\int_{\frac{\varepsilon}{4}}^{1}xP[L_D(A(S))=x]\,dx=$$

$$=\frac{\varepsilon}{4}P\left[L_D(A(S))\le\frac{\varepsilon}{4}\right]+P\left[L_D(A(S))>\frac{\varepsilon}{4}\right]=\frac{\varepsilon}{4}(1-\delta)+(1-(1-\delta))$$

$$=\frac{\varepsilon}{4}+\frac{\delta\varepsilon}{4}$$

ואם נבחר $\delta=\frac{\varepsilon}{2}$ נקבל את הנדרש

2. Let $\mathcal{X}:=\mathbb{R}^2$, $\mathcal{Y}:=\{0,1\}$ and let $\mathcal{H}$ be the class of concentric circles in the plane, i.e.,

$$\mathcal{H}:=\{h_r : r\in\mathbb{R}_+\}\quad\text{where}\quad h_r(\mathbf{x})=\mathbb{1}_{[||\mathbf{x}||_2\le r]}$$

Prove that $\mathcal{H}$ is PAC-learnable and its sample complexity is bounded by

$$m_{\mathcal{H}}(\varepsilon,\delta)\le\frac{\log(1/\delta)}{\varepsilon}$$

When proving, *do not* use a VC-Dimension argument. Instead prove the claim directly from the PAC learnability definition by showing a specific algorithm and analyzing its sample complexity.
Hint: Remember that for every $\varepsilon>0$ it holds that $1-\varepsilon\le e^{-\varepsilon}$

אינטואיציה: בהינתן $S$ קב' דגימות על מעגלים קונצנטריים נרצה לבחור

את המעגל עם הרדיוס המקסימלי שמקבל 1 - זהו ממש ERM

פורמליזציה: אם מניחים ריאליזביליות כאשר $f=h_r\in\mathcal{H}$

ותהי $S$ קב' דגימות הנדגמות באופן בת"ל ומתפלגות כלשהו $D$

נבחר את הרדיוס המקסימלי:

$$r_{max}=\max_i\|x_i\|\quad(y_i=1)\quad,\quad h_r^{max}=\mathbb{1}\{\|x\|\le max\}$$

כעת יהי $\varepsilon,\delta>0$ (נחשב את הסיכוי שהשגיאה שלנו גדולה מ-$\varepsilon$ וזה

הסיכוי שכל $m$ הדוגמאות בדגימה $X$ כך שהרדיוס של ה $x_i$ לא במרחק $\varepsilon$

מ-$r$ האמיתי ולכן:

$$P\left[L_{D,f}\left(h_r^{max}\right)\geq\varepsilon\right]=P\left[\left\{\forall i\in[m]:\left|\|x_i\|-r\right|>\varepsilon\right\}\right)=$$

$$\overset{\text{iid}}{=}P\left(\left|\|x_1\|-r\right|>\varepsilon\right)^m=(1-\varepsilon)^m\leq e^{-\varepsilon m}$$

ממונוטוניות נקבל $\ln(e^{-\varepsilon m})\leq\ln(\delta)\iff-\varepsilon m\leq\ln(\delta)$

$$\iff m\geq\frac{\ln(\delta)}{\varepsilon}$$

וסה״כ $m_H(\varepsilon,\delta)\leq\frac{\log(\delta)}{\varepsilon}$ כנדרש.

## 2.2 VC-Dimension

3. Let $\mathcal{X}=\{0,1\}^n$ and $\mathcal{Y}=\{0,1\}$, for each $I\subseteq[n]$ define the parity function:

$$h_I(\mathbf{x})=\left(\sum_{i\in I}x_i\right)\bmod 2.$$

לכל וקטור $(x_1,...,x_n)$ - $h$ תחזיר את השארית חלוקה

הסכום שלהם ב-2.

ישנן $2^n$ אפשרויות לבחור את $I$ עבור $2^n$ תתי קבוצות של $[n]$

הגודל סופי ובפרט $|h_I(x)|=2^n$ ולכן יש חסם עליון על VCdim

$$VCdim(H_I)\leq\log(|H|)\leq\log(2^n)=n$$

נוכיח שזה שוויון לכך שנמצא תת קבוצה $C$ רק ש $|C|=n$

וגם $H_I$ מנתצת את C.

לדוגמא: $e_1,\dots,e_n$ וקטורי היחידה, ותהי השמה $y_1,\dots,y_n$

עבור הוקטורים הללו, תהי תת קבוצה $I\subseteq[n]$

מספיק שנבחר מסווג $h$ כך ש- $h(e_i)=1$ לכל $j\in I$ כי לכל

$j\in I$ סכום הקואורדינטות הוא 1 ולכל $j\notin I$ הסכום 0

פורמלית:
$$h(e_j)=\begin{cases}1 & j\in I\\ 0 & j\notin I\end{cases}$$

4. Given an integer $k$, let $([a_i,b_i])_{i=1}^{k}$ be any set of $k$ intervals on $\mathbb{R}$ and define their union $A=\cup_{i=1}^{k}[a_i,b_i]$. The hypothesis class $\mathcal{H}_{k-intervals}$ includes the functions: $h_A(x):=\mathbb{1}_{[x\in A]}$, for all choices of $k$ intervals. Find the VC-dimension of $\mathcal{H}_{k-intervals}$ and prove your answer. Show that if we let $A$ be any finite union of intervals (i.e. $k$ is unlimited), then the resulting class $\mathcal{H}_{intervals}$ has VC-dimension $\infty$.

נראה ש- $vcdim\geq 2k$. צ"ל קיימת קבוצה בגודל $2k$ ש- $H_{k\text{-int}}$ מנתצת

תהי - $|C|=2k$ $C\subseteq\mathbb{R}$ ונסדר את איברי c בסדר עולה $C=\{c_1<c_2<\dots<c_{2k}\}$

$H_{k\text{-int}}$ מנתצת את C כולו - כי יהי סיווג כלשהו של איברי c

אזי בכל פעם שאיבר $c_i$ מקבל סיווג 1 נצטרך לכסות אותו ע"י קטע שימשך עד

עוד האיברים אחריו גם מקבלים 1 - צ"ל הסיווג "הגרוע" ביותר הוא כשמסווג

באופן זוגי $1,0,1,0\dots$ ונקבל $k$ מקטעים ונוכל לכסות עם $k$ קטעים

כעת נראה ש $vcdim<2k+1$: תהי $|C|=2k+1$ ונסדרם שוב בסדר עולה

כאשר נסווג את האי זוגיים ל-1. נקבל סיווג ש-$H_{k\text{-}int}$ לא מנפץ

כי יש $k+1$ איזורים עם 1 המופרדים ב-$k$ איזורים עם 0,

לעומת זאת אם ניתן ל-$H$ להיות איחוד סופי כלשהו של קטעים

ללא הגבלה על $k$ - המימד יהיה אינסוף.

### 2.3 Monotonicity

5. Let $\mathcal{H}$ be a hypothesis class for a binary classification task. Suppose that $\mathcal{H}$ is PAC learnable and its sample complexity is given by $m_{\mathcal{H}}(\cdot,\cdot)$. Show that $m_{\mathcal{H}}$ is monotonically non-increasing in each of its parameters. That is:
   - Show that given $\delta \in (0,1)$, and given $0 < \varepsilon_1 \le \varepsilon_2 < 1$, we have that $m_{\mathcal{H}}(\varepsilon_1, \delta) \ge m_{\mathcal{H}}(\varepsilon_2, \delta)$.
   - Similarly, show that given $\varepsilon \in (0,1)$, and given $0 < \delta_1 \le \delta_2 < 1$, we have that $m_{\mathcal{H}}(\varepsilon, \delta_1) \ge m_{\mathcal{H}}(\varepsilon, \delta_2)$.

• תהי $\delta \in (0,1)$ ויהי $0 < \varepsilon_1 \le \varepsilon_2 < 1$

לכל מדגם בגודל $m \ge m_H(\varepsilon_1, \delta)$ או גדול מובטח לנו כי בהסתב׳

של $1-\delta$ לפחות נקבל כלל החלטה $\hat{h}$ עם שגיאה $\varepsilon_1$ לכל היותר.

$$m \ge m_H(\varepsilon_1, \delta) \Rightarrow P[L_{D,h}(\hat{h}) \le \varepsilon_1 \le \varepsilon_2] \ge 1-\delta$$

אם נניח בשלילה ש- $m_H(\varepsilon_1, \delta) < m_H(\varepsilon_2, \delta)$ נקבל סתירה

למינימליות של סיבוכיות המדגם.

באופן דומה ויהי $\varepsilon \in (0,1)$ ו- $0 < \delta_1 \le \delta_2 \le 1$

$$m \ge m_H(\varepsilon, \delta_1) \Rightarrow P[L_{D,h}(\hat{h}) \le \varepsilon] \ge 1-\delta_1 \ge 1-\delta_2$$

6. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two classes for binary classification, such that $\mathcal{H}_1 \subseteq \mathcal{H}_2$. Show that $VC-dim(\mathcal{H}_1) \leq VC-dim(\mathcal{H}_2)$.

יהיו $H_1, H_2$ 2 מחלקות היפותזות ותהי C קבוצה כלשהי

דהיינו $H_1 \subseteq H_2$ נקבל ש- אם $H_1$ מנתצת את C אז:

אז $H_2$ מנתצת את C. ולכן $vcdim(H_1) \leq vcdim(H_2)$

שהרי: אם $H_1$ מנתצת את C אז: $vcdim(H_1) \geq |C|$ ואם $H_2$

מנתצת את C $vcdim(H_2) \geq |C|$

## 2.4 Agnostic-PAC

7. Prove that if $\mathcal{H}$ has the uniform convergence property with function $m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$, then $\mathcal{H}$ is Agnostic-PAC learnable with sample complexity $m_{\mathcal{H}}(\varepsilon,\delta) \leq m^{UC}(\varepsilon/2,\delta)$.

נניח ש- H בעלת תכונת ההתכנסות הסטנדרטית עם הפונ' $m_h^{uc}: (0,1)^2 \to \mathbb{N}$

כ"א לכל $\varepsilon, \delta \in (0,1)$ ולכל התפלגות D מעל $X \times Y$ מתקיים:

$$D^m(\{S \in (X,Y)^m \mid S \text{ נציג } \varepsilon\text{-}\}) \geq 1-\delta$$

S נציג-$\varepsilon$ עם $D, H, l$ אם לכל $h \in H$ מתקיים $|L_S(h) - L_D(h)| < \varepsilon$. ולכן לכל $\varepsilon, \delta \in (0,1)$ נקבל:

$$D^m(\{S \in (X,Y)^m \mid S \text{ נציג } \tfrac{\varepsilon}{2}\text{-}\}) \geq 1-\delta$$

כ"א לכל $h \in H$ מתקיים:

$$|L_S(h) - L_D(h)| > \frac{\varepsilon}{2} \Rightarrow -\frac{\varepsilon}{2} < L_S(h) - L_D(h) < \frac{\varepsilon}{2} \Rightarrow$$

$$L_D(h) < L_S(h) + \frac{\varepsilon}{2} \wedge L_S(h) < L_D(h) + \frac{\varepsilon}{2}$$

נגדיר $h_D = \arg\min_{h\in H} L_D(h)$, $h_S = \arg\min_{h\in H} L_S(h)$ ונקבל:

$$L_D(h_S) \le L_S(h_S) + \frac{\varepsilon}{2} \le L_S(h_D) + \frac{\varepsilon}{2} \le L_D(h_D) + \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = L_D(h_D) + \varepsilon$$

ולכן:

$$P_{S\sim D^m}\left(L_D(h_S) \le \min_{h\in H} L_D(h) + \varepsilon\right) \ge$$

$$\ge D^m\left(\{S \in (X\times Y)^m \mid \varepsilon\text{-מייצג-}S\}\right) \ge 1-\delta$$

ובפרט $H$ למידה PAC - אגנוסטי עם $m_H(\varepsilon,\delta) \le m^{K}(\frac{\varepsilon}{2},\delta)$

8. Let $\mathcal{H}$ be a hypothesis class over a domain $\mathcal{Z} = \mathcal{X} \times \{\pm 1\}$, and consider the 0-1 loss function. Assume that there exists a function $m_{\mathcal{H}}$, for which it holds that for every distribution $\mathcal{D}$ over $\mathcal{Z}$ there is an algorithm $\mathcal{A}$ with the following property: when running $\mathcal{A}$ on $m \ge m_{\mathcal{H}}$ i.i.d. examples drawn from $\mathcal{D}$, it is guaranteed to return, with probability at least $1-\delta$, a hypothesis $h_S : \mathcal{X} \to \{\pm 1\}$ with $L_{\mathcal{D}}(h_S) \le \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon$. Is $\mathcal{H}$ agnostic PAC learnable? Prove or show a counter example.

לא. (הבעיה שהאלגו' נבחר בהתאם ל- D)

דוגמא נגדית: $H$ תיהיה מח' היפותזות של איחוד כל האיחודים הסופיים של קטעים משאלה 5 שמימד VC שלה כמובן אינסופי

ולכן היא לא למידה PAC.

בהינתן התפלגות D ו- $m_H$ דגימות הנדגמות אחיד מהתל

ויהי כלל למידה - $h_S$ כך שבהינתן דגימה $x$ $h_S$ מחשב את ההסתברות $p$ ש- $x$ יקבל ערך 1 - מטילים מטבע עם הסת' $p$ לקבל 1 ומחזירים את התוצאה.

H לא למיד PAC כמו שראינו בשאלה 5

אבל כן מתקיים $L_D(h_S) \le \min_{h \in H} L_D(h) + \varepsilon$ כי אנחנו מסתמכים

על ההתפלגות האמיתית.

- התנאים מתקיימים אבל H לא למידה PAC - לכן הטענה שגויה.