# תרגול 10: למידה בצורה PAC
# (Probably Approximately Correct)

## תקציר התאוריה

**<u>סימונים</u>**:

$X$ - מרחב הכניסה.

$Y$ - מרחב היציאה ( $Y=\mathbb{R}$ לבעיית הרגרסיה, $Y=\{-1,+1\}$ לבעיית הסיווג).

$f_0 : X \to Y$ - פונקציית המטרה אותה רוצים ללמוד (עבור בעיית הסיווג זהו המסווג הנלמד).

$p_X(x)$ - פילוג הסתברות על מרחב הכניסה $X$. פילוג זה קובע את פילוג ה"כניסה הטיפוסית" $x \in X$.

$\{x_k, y_k\}_{k=1}^n$ - דוגמאות (סדרת הלימוד).

$F$ - מרחב ההיפותזות, מתוכו עלינו לבחור את הפונקציה $\hat{f}$ אשר משערכת את $f_0$.

**<u>קריטריון ביצועים</u>**:

מגדירים פונקציית מחיר $\ell(\hat{y},y)$. עבור בעיית הרגרסיה נהוג לבחור ב- $\ell(\hat{y},y)=(\hat{y}-y)^2$, ואילו עבור בעיית הסיווג ב- $\ell(\hat{y},y)=\mathbb{I}\{\hat{y}\neq y\}$.

קריטריון ביצועים עבור פונקציה (היפותזה) כלשהיא $\hat{f}\in F$ הוא המחיר הממוצע, הנתון ע"י:

$$L(\hat{f}) \triangleq \mathbb{E}\left\{\ell(\hat{f}(x),y)\right\}$$

כאשר התוחלת היא לפי $p_X(x)$, ו- $y=f_0(x)$.

מטרתנו למצוא פונקציה (היפותזה) $\hat{f}\in F$ אשר מביאה את קריטריון הביצועים למינימום:

$$\hat{f}^* \in \arg\min_{\hat{f}\in F} L(\hat{f}) \,.$$

נשים לב כי קריטריון זה לא ניתן לחישוב במקרה ולא יודעים את $p_X(x)$ ו- $f_0(x)$.

במקרה זה משתמשים בסדרת הלימוד ומגדירים פונקציית המחיר האמפירי:

$$\hat{L}_n(\hat{f}) \triangleq \frac{1}{n}\sum_{k=1}^{n} \ell(\hat{f}(x_k), y_k)$$

ומחפשים:

$$\hat{f}_n = \arg\min_{\hat{f}\in F} \hat{L}_n(\hat{f}) \,.$$

לסיכום, סימונים קריטריון ביצועים:

| | |
|---|---|
| $\ell(\hat{y}, y)$ | פונקציית מחיר |
| $L(\hat{f}) \triangleq \mathbb{E}\left\{\ell(\hat{f}(x), y)\right\}$ | המחיר ממוצע |
| $\hat{L}_n(\hat{f}) \triangleq \frac{1}{n}\sum_{k=1}^{n} \ell(\hat{f}(x_k), y_k)$ | המחיר הממוצע האמפירי |
| $\hat{f}^* \in \arg\min_{\hat{f}\in F} L(\hat{f})$ <br> $L^* \triangleq L(\hat{f}^*) = \min_{\hat{f}\in F} L(\hat{f})$ | הפונקציה האופטימאלית ב- $\hat{f}\in F$, <br> והמחיר הממוצע עבור פונקציה זו. |
| $\hat{f}_n = \arg\min_{\hat{f}\in F} \hat{L}_n(\hat{f})$ | הפונקציה האופטימאלית האמפירית ב- $\hat{f}\in F$ |

<u>**חסם PAC עבור קבוצת היפותזות סופית**</u> ($|F| < \infty$):

זהו חסם הסתברותי המציין עד כמה הקירוב שלנו, $\hat{f}_n$, קרוב לפונקציה האופטימלית, $\hat{f}^*$, במובן קריטריון הביצועים $L(\cdot)$.

נסמן $L^* \triangleq \min_{\hat{f}\in F} L(\hat{f})$. אזי:

$$\text{(1)} \quad \forall\varepsilon > 0: \ \mathbb{P}\left\{L(\hat{f}_n) - L^* > \varepsilon\right\} < 2|F|e^{-\varepsilon^2 n/2}$$

אם בנוסף יודעים כי $L^* = 0$, אזי יש **חסם הדוק יותר**:

$$\text{(2)} \quad \forall\varepsilon > 0: \ \mathbb{P}\left\{L(\hat{f}_n) > \varepsilon\right\} < |F|e^{-\varepsilon n}$$

# שאלות

## **שאלה** 1

נניח כי מעוניינים ללמוד מסווג בינארי $f_0 : \{-1,1\} \times \{-1,1\} \mapsto \{-1,1\}$.

לצורך הלמידה, הוגרלו $n = 100$ דוגמאות בלתי תלויות, ונתונה משפחת ההיפותזות ממנה עלינו ללמוד את $f_0$:

$$F = \left\{ f : \{-1,1\} \times \{-1,1\} \mapsto \{-1,1\} \middle| \sum_{a,b \in \{-1,1\}} f(a,b) < 2 \right\}$$

א. אם ידוע כי המסווג הנלמד $f_0$ נתון ע"י:

| $a$ | $b$ | $f_0$ |
|---|---|---|
| $-1$ | $-1$ | $1$ |
| $-1$ | $1$ | $1$ |
| $1$ | $-1$ | $1$ |
| $1$ | $1$ | $-1$ |

מצאו $\varepsilon$ **קטן ככל האפשר**, כך שיתקיים:

$$\mathbb{P}\left\{ L(\hat{f}_n) - L^* \le \varepsilon \right\} > 0.98$$

ב. חשבו את $L^*$ במקרה של סעיף א', אם ידוע כי $p_X \sim U\left(\{-1,1\} \times \{-1,1\}\right)$.

ג. חזרו על א', אם ידוע כי המסווג הנלמד $f_0$ נתון ע"י:

| $a$ | $b$ | $f_0$ |
|---|---|---|
| $-1$ | $-1$ | $-1$ |
| $-1$ | $1$ | $-1$ |
| $1$ | $-1$ | $-1$ |
| $1$ | $1$ | $1$ |

ד. אם נתון כי $p_X \sim U\left(\{-1,1\} \times \{-1,1\}\right)$, מצאו את הסתברות לשגיאה 0.

**<u>פתרון 1</u>**

א. נחשב תחילה את $|F|$:

סה"כ יש $2^{2^2}=16$ פונקציות בינאריות בשני משתנים. $F$ מכילה פונקציות המקיימות $\sum_{a,b\in\{-1,1\}} f(a,b)<2$, ולכן $F$ לא מכילה פונקציות בינאריות עם תוצאה אחת או פחות של $-1$. מכאן, $|F|=16-4-1=11$.

מכיוון שנדרש למצוא $\varepsilon$ **קטן ככל האפשר**, נבדוק האם נוכל להשתמש בחסם ההדוק יותר ב- (2). כלומר נבדוק האם $f_0\in F$. אבל, לפי הנתון, $\sum_{a,b\in\{-1,1\}} f_0(a,b)=2$, ולכן $f_0\notin F$. כלומר $L^*>0$. מכאן שעלינו להשתמש בחסם (1) לעיל:

$$\mathbb{P}\left\{L(\hat{f}_n)-L^*\le\varepsilon\right\}>1-\ 2|F|e^{-\varepsilon^2 n/2}$$

כלומר נדרוש:

$$1-\ 2|F|e^{-\varepsilon^2 n/2}>0.98$$
$$2|F|e^{-\varepsilon^2 n/2}<0.02$$
$$\varepsilon>0.374$$

(אפשר להשתמש ישר בנוסחה של "מרווח הבטחון" של רשימות ההרצאה).

ב. לכל $f\in F$,

$$L(f)=\mathbb{P}\left\{f(x)\ne f_0(x)\right\}=\sum_{x:f(x)\ne f_0(x)} p_X(x)$$
$$=\frac{1}{4}\left|x:f(x)\ne f_0(x)\right|.$$

אבל: $\min_{f\in F}\left|x:f(x)\ne f_0(x)\right|=1$,

והוא מתקבל עבור פונקציות $f\in F$ הבאות:

| $a$ | $b$ | $f$ | $a$ | $b$ | $f$ | $a$ | $b$ | $f$ |
|---|---|---|---|---|---|---|---|---|
| $-1$ | $-1$ | -1 | $-1$ | $-1$ | 1 | $-1$ | $-1$ | 1 |
| $-1$ | 1 | 1 | $-1$ | 1 | -1 | $-1$ | 1 | 1 |
| 1 | $-1$ | 1 | 1 | $-1$ | 1 | 1 | $-1$ | -1 |
| 1 | 1 | $-1$ | 1 | 1 | $-1$ | 1 | 1 | $-1$ |

ולכן:

$$L^* = \min_{f\in F} L(f) = \frac{1}{4}.$$

ג. במקרה זה כמובן $f_0 \in F$, כלומר $L^* = 0$, ולכן נוכל להשתמש בחסם הדוק יותר (2). כלומר:

$$\mathbb{P}\left\{L(\hat{f}_n) \le \varepsilon\right\} > 1 - |F|e^{-\varepsilon n}.$$

לכן נדרש:

$$1 - |F|e^{-\varepsilon n} > 0.98$$

$$|F|e^{-\varepsilon n} < 0.02$$

$$\varepsilon > 0.0631$$

כצפוי ה-$\varepsilon$ המינימלי קטן יותר במקרה זה מאשר בסעיף א'.

ד. נתון כי כל כניסה מופיעה בהסתברות $1/4$:

| $a$ | $b$ | הסתברות |
|---|---|---|
| $-1$ | $-1$ | $1/4$ |
| $-1$ | $1$ | $1/4$ |
| $1$ | $-1$ | $1/4$ |
| $1$ | $1$ | $1/4$ |

נסמן את $n$ הדגימות האקראיות (i.i.d.) ב $\{X_k\}_{k=1}^n$. אזי:

$$\mathbb{P}\begin{Bmatrix}\text{error=0 after}\\ \text{seeing } n \text{ samples}\end{Bmatrix} = \mathbb{P}\begin{Bmatrix}n \text{ samples contains}\\ \text{all 4 possible inputs}\end{Bmatrix}$$

$$= 1 - \mathbb{P}\begin{Bmatrix}n \text{ samples doesn't contain}\\ \text{all 4 possible inputs}\end{Bmatrix}$$

$$= 1 - \mathbb{P}\begin{pmatrix}\{X_1 \ne (-1,-1),\dots,X_n \ne (-1,-1)\} \cup \{X_1 \ne (-1,1),\dots,X_n \ne (-1,1)\} \cup \\ \cup \{X_1 \ne (1,-1),\dots,X_n \ne (1,-1)\} \cup \{X_1 \ne (1,1),\dots,X_n \ne (1,1)\}\end{pmatrix}$$

$$\ge 1 - \mathbb{P}\{X_1 \ne (-1,-1),\dots,X_n \ne (-1,-1)\} - \mathbb{P}\{X_1 \ne (-1,1),\dots,X_n \ne (-1,1)\}$$

$$- \mathbb{P}\{X_1 \ne (1,-1),\dots,X_n \ne (1,-1)\} - \mathbb{P}\{X_1 \ne (1,1),\dots,X_n \ne (1,1)\}$$

$$= 1 - \left(\frac{3}{4}\right)^n - \left(\frac{3}{4}\right)^n - \left(\frac{3}{4}\right)^n - \left(\frac{3}{4}\right)^n = 1 - 4\left(\frac{3}{4}\right)^n$$

כאשר האי-שוויון נובע מחסם איחוד, והשיוין הלפני אחרון נובע מהעובדה כי הדגימות הן i.i.d. עם פילוג אחיד. עבור $n = 100$ נקבל אם כן:

$$\mathbb{P}\left\{\begin{matrix}\text{error=0 after}\\ \text{seeing 100 samples}\end{matrix}\right\} \geq 1-4\left(\frac{3}{4}\right)^{100} = 0.9999... \approx 1$$
$$\Rightarrow \boxed{\mathbb{P}\left\{\begin{matrix}\text{error=0 after}\\ \text{seeing 100 samples}\end{matrix}\right\} \approx 1}$$

תוצאה זו מדגישה כי חסמי PAC הם כלליים (נכונים לכל פילוג של כניסה ופונקציית מטרה), וכי אם יש לנו מידע נוסף לגבי המודל, ניתן לקבל חסמים הרבה יותר טובים.

## <u>שאלה 2</u>

כצעד ראשון לקראת הרחבה של חסמי PAC עבור קבוצות היפותזות אינסופיות נרצה להעריך את ה"עושר"/"גיוון" של אוסף היפותזות מסוים. נתחיל בלהראות את התכונה הבאה:

**<u>משפט (Steele Dudley)</u>** יהי $G$ מרחב וקטורי של פונקציות ממשיות ב-$R^d$, נסמן את מימדו ב-$r=\dim(G)$. נגדיר את משפחת המסווגים (היפותזות) הבאה:

$$G'=\left\{g'(x)=sign(g(x)) \mid g\in G\right\}$$

לכל קבוצה $\{x_i\}_{i=1}^m$ של $m=r+1$ תבניות קיימת סדרת תיוגים בינאריים $\{y_i\}_{i=1}^m\in\{-1,1\}^m$ כך שלא ניתן לסווג את הנקודות $\{x_i,y_i\}_{i=1}^m$ בצורה מושלמת ע"י מסווגים ב-$G'$.

<u>תזכורת</u>: $sign(x)=\begin{cases}1 & x\geq 0\\ -1 & x<0\end{cases}$.

<u>סקיצת ההוכחה</u>:

1. תהי קבוצת נקודות כלשהיא ב-$R^d$: $D=\{x_i\}_{i=1}^m$. בהמשך ההוכחה נבנה סדרת תיוגים $\{y_i\}_{i=1}^m$ כך שלא יהיה ניתן לסווג את הנקודות $\{x_i,y_i\}_{i=1}^m$ בצורה מושלמת ע"י מסווגים ב-$G'$.
2. נגדיר אופרטור לינארי $L:G\to R^m$ באופן הבא: $L(g)=\begin{pmatrix}g(x_1) & \ldots & g(x_m)\end{pmatrix}^T$. נשים לב כי האופרטור $L(\cdot)$ הוא לינארי (הומוגני ואדיטיבי) למרות שהפונקציות $g\in G$ אינן בהכרח לינאריות.

3. נגדיר $L(G)$ להיות הטווח של האופרטור $L$ מ- $G$, כלומר,

$$L(G) = \left\{ x \in R^m \mid \exists g \in G : L(g) = x \right\}$$

$L(G)$ הינו תת מרחב של $R^m$, כלומר, סגור לחיבור וכפל בסקלר.

4. נשים לב כי מימד המרחב $L(G)$ אינו יכול להיות יותר גבוה ממימד של $G$. לדוגמא, ניתן להראות בקלות כי אם $\{g_1, g_2, ..., g_r\}$ הוא בסיס של $G$ אזי אוסף הוקטורים $\{L(g_1), ..., L(g_r)\}$ פורש את המרחב $L(G)$. כלומר, $\dim L(G) \le \dim(G) = r = m - 1$ או במילים אחרות $\dim L(G) < m$.

5. מהעובדות $L(G) \subseteq R^m$ ו- $\dim L(G) < \dim(R^m)$ נובע כי קיים $P = (P_1, ..., P_m) \neq \underline{0}$ כך ש- $P \perp L(g)$ לכל $g \in G$. באופן שקול ניתן לרשום כי

$$\sum_{i=1}^{m} P_i g(x_i) = 0 \ (*)$$

לכל $g \in G$. נניח כי קיים לפחות רכיב אחד של $P$ שלילי ממש, אחרת ניתן להסתכל על וקטור $(-P)$ שהוא גם מאונך ל- $L(G)$.

6. נבחר את התיוגים הבאים: $y_i = sign(P_i)$. ע"מ לסיים את ההוכחה נותר להראות כי לא קיים מסווג $g' \in G'$ שמצליח לסווג את הסדרה $\{x_i, y_i\}_{i=1}^m$ בצורה מושלמת.

נניח בשלילה כי קיים $\hat{g} \in G$ כך ש- $g' = sign(\hat{g})$ מסווג בצורה המושלמת את הסדרה $\{x_i, y_i\}_{i=1}^m$. מזה נובע כי $sign(P_i) = sign(\hat{g}(x_i))$ לכל $i$ או, באופן שקול, ש- $P_i \hat{g}(x_i) \ge 0$ לכל $i$. לכן, עבור $j$ כך ש- $P_j < 0$ (לפי ההנחה סעיף 5 קיים אחד כזה) מתקיים כי $g(x_j) < 0$ ולכן $P_j \hat{g}(x_j) > 0$.

נתבונן בסכום $\sum_{i=1}^{m} P_i \hat{g}(x_i)$ - כאן כל אחד מהגורמים הוא אי-שלילי ואחד מהגורמים הוא חיובי ממש, לכן $\sum_{i=1}^{m} P_i \hat{g}(x_i) > 0$. אך זה עומד בסתירה לתנאי הניצבות (*) בסעיף 5.

<u>מש"ל</u>