תרגול 11 : מימד VC

תקציר התאוריה

<u>סימונים</u> :

X - מרחב הכניסה.

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

F - משפחת ההיפותזות, מתוכו עלינו ללמוד את המסווג.

<u>המטרה</u> :

אנו צריכים מדד ל"עושר" של משפחת ההיפותזות. למשל, אם יש לנו מספר סופי של היפותזות, אזי המדד הוא $|F|$. אבל מה נעשה כאשר $|F|$ גדול מאוד (או אפילו $|F|=\infty$ - למשל משפחה פרמטרית) ? במקרה זה המדד ההגיוני הוא מספר פרמטרים המתארים את F. בעצם, מימד VC קשור (בד"כ) למספר הפרמטרים <u>האפקטיבי</u> המתארים את משפחת ההיפותזות.

<u>דיכוטומיות וניתוץ</u> :

תהי $\{x_k\}_{k=1}^n$ קבוצת נקודות נתונה, $x_k \in X$.

<u>דיכוטומיה</u> הינה חלוקה של קבוצה זו לשתי קבוצות זרות. סה"כ יש 2^n דיכוטומיות.

כל היפותזה $f\in F$ משרה דיכוטומיה על $\{x_1,\dots,x_n\}$, דהיינו החלוקה לשתי הקבוצות : $\{x_k : f(x_k)=+1\},\ \{x_k : f(x_k)=-1\}$.

נסמן : $S_F\{x_1,\dots,x_n\} \triangleq |\{(f(x_1),\dots,f(x_n)) : f\in F\}|$, כלומר $S_F\{x_1,\dots,x_n\}$ זהו מספר הדיכוטומיות השונות שמשרים איברי F על $\{x_k\}_{k=1}^n$. נשים לב כי זהו גודל שתלוי ב-$\{x_k\}_{k=1}^n$ ו-F. כמו כן, ברור כי : $S_F\{x_1,\dots,x_n\}\le 2^n$.

נאמר כי הקבוצה $\{x_k\}_{k=1}^n$ <u>מנותצת לחלוטין</u> ע"י F אם F משרה עליה את כל הדיכוטומיות האפשריות, כלומר $S_F\{x_1,\dots,x_n\}=2^n$.

<u>מקדם הניתוץ</u> מוגדר ע"י המקסימום של $S_F\{x_1,\dots,x_n\}$ על כל קבוצות הנקודות האפשריות, כלומר :

$$S_F(n) \triangleq \max_{\{x_1 \dots x_n\} \subset X} S_F\{x_1, \dots, x_n\}.$$

נשים לב כי $S_F(n)$ תלוי במספר הנקודות n ובקבוצת ההיפותזות F בלבד.

<u>מימד VC</u>:

מימד VC של F הינו המספר הגדול ביותר של נקודות שניתן לנתץ לחלוטין ע"י F:

$$\begin{aligned} VC(F) &\triangleq \max\{n \geq 1 \ : \ S_F(n) = 2^n\} \\ &\equiv \max\{n \geq 1 \ : \ \max_{\{x_1 \dots x_n\} \subset X} S_F\{x_1 \dots x_n\} = 2^n\} \end{aligned}$$

<u>הערה:</u> מימד VC הוא תכונה גאומטרית של מרחב ההיפותזות F, וכאמור, ניתן להתייחס אליו (בד"כ) כאל מספר הפרמטרים האפקטיבי במשפחת ההיפותזות.

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

נסמן $V = \mathrm{VCdim}(F)$, ונניח כי $V < \infty$. אזי:

$$\forall \varepsilon > 0: \ \mathbb{P}\left\{L(\hat{f}_n) - L^* > \varepsilon\right\} < \ 4(2n+1)^V e^{-\varepsilon^2 n/32}$$

<u>שאלה 1</u>

נתונה משפחת מסווגים לינאריים ב-$\mathbb{R}^d$, כלומר:

$$F = \left\{f(x) = sign(w^T x + b) : w \in \mathbb{R}^d, b \in \mathbb{R}\right\}.$$

הראו כי $VC(F) = d + 1$.

<u>תזכורת (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:

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

אזי 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>פתרון</u>

נוכיח את הטענה בשני שלבים: **(1)** נראה כי $VC(F) \geq d+1$; **(2)** נראה כי $VC(F) < d+2$.

(1) כדי להראות ש-$VC(F) \geq d+1$ לכל $d \in \{1,2,3,...\}$ יש למצוא סידור של $d+1$ נקודות שניתן לנתץ ע"י F. נסתכל על הקבוצה הבאה של וקטורים ממימד d: $D = \{\mathbf{0}, \mathbf{e_1}, \mathbf{e_2}, ..., \mathbf{e_d}\}$, כאשר $\mathbf{e_i}$ הוא וקטור יחידה בכיוון i, כלומר, הרכיב ה-i-י של וקטור $\mathbf{e_i}$ שווה לאחד וכל השאר לאפס. וקטור $\mathbf{0}$ הוא וקטור של אפסים.
נשים לב כי קבוצה D מכילה $d+1$ איברים.
תהי $g: R^d \to \{-1,1\}$ היפותזה כלשהי (לאו דווקא $g \in F$), כמובן ש-g משרה דיכוטומיה מסוימת על D. נראה כי קיימת $f \in F$ שמנתצת את הדיכוטומיה:
<u>טענה</u> יהיו

$$w_i = \begin{cases} 1, & g(e_i) = 1 \\ -1, & g(e_i) = -1 \end{cases} \text{ ו- } b = \begin{cases} 1/2, & g(0) = 1 \\ -1/2, & g(0) = -1 \end{cases}$$

אזי $f(x) = sign(w^T x + b)$ השייכת ל-F מנתצת את הדיכוטומיה.
<u>הוכחת הטענה:</u> פשוט נציב את הנקודות ב-D לפונקציה f.
עבור $\mathbf{e_i}$:

$$\begin{aligned} f(\mathbf{e_i}) &= sign(w^T \mathbf{e_i} + b) = sign(w_i + b) \\ &= sign\left(\begin{cases} \geq 1/2, & g(\mathbf{e_i}) = 1 \\ \leq -1/2, & g(\mathbf{e_i}) = -1 \end{cases} \right) \\ &= \begin{cases} 1, & g(\mathbf{e_i}) = 1 \\ -1, & g(\mathbf{e_i}) = -1 \end{cases} = g(\mathbf{e_i}) \end{aligned}$$

עבור $\mathbf{0}$:

$$\begin{aligned} f(\mathbf{0}) &= sign(w^T \mathbf{0} + b) = sign(b) \\ &= sign\left(\begin{cases} 1/2, & g(\mathbf{0}) = 1 \\ -1/2, & g(\mathbf{0}) = -1 \end{cases} \right) \\ &= g(\mathbf{0}) \end{aligned}$$

<u>מש"ל טענה</u>

הראינו כי לכל g קיימת $f \in F$ כך $f(x) = g(x)$ לכל $x \in D$, אזי F מנתצת לחלוטין את הקבוצה D, לכן $VC(F) \geq |D| = d+1$.

(2) נראה ש- $VC(F) < d+2$. נגדיר $\tilde{F} = \left\{ \tilde{f}(x) = w^T x + b : w \in R^d, b \in R \right\}$ אזי $F = \left\{ f(x) = sign\left(\tilde{f}(x)\right) : \tilde{f} \in \tilde{F} \right\}$. נשים לב כי $\dim(\tilde{F}) = d+1$, לכן לפי משפט Steele-Dudley מתקבל כי $VC(F) < d+2$.

<u>שאלה</u> <u>2</u>

נתונה משפחת ההיפותזות:

$$F = \left\{ f(x) = sign(ax+b) : a,b \in \mathbb{R} \right\}$$

$$. f : \mathbb{R} \mapsto \{-1,1\}$$

א. מצאו n_0, כך שלכל $n \geq n_0$ יהיה $L(\hat{f}_n) - L^* \leq 0.2$ בהסתברות 0.97 לפחות.

ב. נתון כי מסווג המטרה הינו $f_0(x) = sign(x^2 - 4)$, וכן נתון כי הדגימות מוגרלות לפי $p_X \sim U(-4,4)$. חשבו את L^* במקרה זה, וכן את המסווג $\hat{f}^*$ המגשים את L^*.

<u>פתרון</u> <u>2</u>

א. במקרה זה $|F| = \infty$, ולכן נשתמש בחסם (3). ראינו (בתרגול קודם) כי מימד VC למשפחה הנתונה הוא $V = 2$. כמו כן, $\varepsilon = 0.2$.

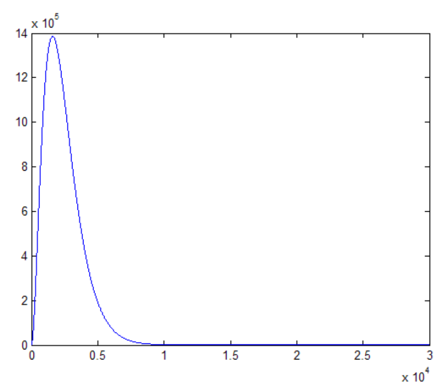
נדרוש:

$$1 - 4(2n+1)^V e^{-\varepsilon^2 n/32} > 0.97$$

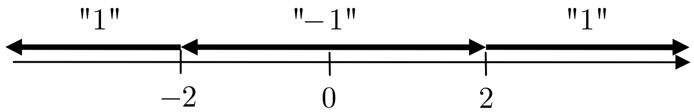
$$4(2n+1)^V e^{-\varepsilon^2 n/32} < 0.03$$

$$(2n+1)^V e^{-\varepsilon^2 n/32} < 0.0075$$

פותרים נומרית ומקבלים: $n > 20940 \triangleq n_0$.



ב. תחילה, ניראה איכותית את הפתרון. המסווג הנלמד נראה כך:



השאלה בעצם היא איפה הכי כדאי לשים נקודת החלטה בודדת (זאת קבוצת ההיפותזות שלנו), כך שהסתברות השגיאה תהיה מינימלית. כמובן שהדבר תלוי בפילוג ההסתברות של הדוגמאות, אך במקרה זה נתון כי הפילוג אחיד ולכן קל לענות שנקודת החלטה בודדת האופטימלית תהיה ב-2 (או -2). כך נקבל שגיאה רק על החלק של $x < -2$.

חישוב כמותי:

$$f(x) = sign(ax+b)$$

$$f_0(x) = sign(x^2-4)$$

$$L(f) = \mathbb{P}\left\{f(x) \neq f_0(x)\right\} = \int\limits_{x:f(x)\neq f_0(x)} p_X(x)dx$$

$$p_X(x) = \begin{cases} \frac{1}{8}, & -4 < x < 4 \\ 0, & \text{otherwise.} \end{cases}$$.

נניח $a > 0$. (הפתרון עבור $a < 0$ סימטרי)

את הקבוצה $\left\{x : f(x) \neq f_0(x)\right\}$ נחלק ל:

$$\begin{cases} x^2 - 4 < 0 \\ ax + b > 0 \end{cases} \Rightarrow \boxed{\begin{matrix} -2 < x < 2 \\ x > -\frac{b}{a} \end{matrix}}$$

$$\begin{cases} x^2 - 4 > 0 \\ ax + b < 0 \end{cases} \Rightarrow \boxed{\begin{matrix} x > 2, x < -2 \\ x < -\frac{b}{a} \end{matrix}}$$

נסמן $\alpha=-\frac{b}{a}$. נחלק לשלושה מקרים:

$$\alpha<-2: L(f)=\int_{-\infty}^{\alpha} p_X(x)dx+\int_{-2}^{2} p_X(x)dx$$
$$=\int_{-4}^{\alpha}\frac{1}{8}dx+\int_{-2}^{2}\frac{1}{8}dx=\frac{1}{8}(\alpha+4)+\frac{1}{8}4\xrightarrow{\min}\frac{1}{2}$$
$$-2<\alpha<2: L(f)=\int_{-\infty}^{-2} p_X(x)dx+\int_{\alpha}^{2} p_X(x)dx$$
$$=\int_{-4}^{-2}\frac{1}{8}dx+\int_{\alpha}^{2}\frac{1}{8}dx=\frac{2}{8}+\frac{1}{8}(2-\alpha)\xrightarrow{\min}\frac{1}{4}$$
$$\alpha>2: L(f)=\int_{-\infty}^{-2} p_X(x)dx+\int_{2}^{\alpha} p_X(x)dx$$
$$=\int_{-4}^{-2}\frac{1}{8}dx+\int_{2}^{\alpha}\frac{1}{8}dx=\frac{2}{8}+\frac{1}{8}(\alpha-2)\xrightarrow{\min}\frac{1}{4}$$

לכן, מקבלים כי: $L^*=\min_{f\in F}L(f)=\frac{1}{4}$.

והמינימום מתקבל עבור $\alpha=-\frac{b}{a}=2$:

$\hat{f}^*=sign(ax-2a)$.

כלומר זוהי נקודת החלטה ב-$x=2$, כצפוי.

(עבור $a<0$ מקבלים $\hat{f}^*=sign(ax+2a)$).