מבוא למערכות לומדות - הרצאה 3 - הוכחת המשפט היסודי

22 ביוני 2015

היום נוכיח את המשפט היסודי של למידה חישובית. המשפט, שהוכח ע"י ופניק וצ'רבוננקינס בשנת 71, מראה שעבור בעיות קלסיפיקציה בינארית, סיבוכיות המדגם פרופורציונאלית למימד VC.

1 תזכורת - מודל PAC, אלגוריתמי ERM, מימד VC והמשפט היסודי

נקבע מרחב מדגם X מחלקת היפותזות $\mathcal{H} \subset Y^X$. כאמור, אנו נתרכז בקלסיפיקציה בינארית ולכן נניח ש־$Y = \{0,1\}$ ונקבע את פונקצייה ההפסד להיות

$$l(\hat{y}, y) = l_{0-1}(\hat{y}, y) = \begin{cases} 0 & \hat{y} = y \\ 1 & \hat{y} \neq y \end{cases}$$

.

הגדרה 1.1 אלגוריתם למידה מקבל בתור קלט מדגם אימון

$$S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \in (X \times Y)^m$$

ומחזיר פונקציה $h : X \to Y$.

הגדרה 1.2 השגיאה של $h : X \to Y$ ביחס להתפלגות $\mathcal{D}$ על $X \times Y$ היא

$$L_{\mathcal{D}}(h) = E_{(x,y)\sim\mathcal{D}} l(h(x), y) = \Pr_{(x,y)\sim\mathcal{D}}(h(x) \neq y)$$

השגיאה של $\mathcal{H}$ ביחס ל־$\mathcal{D}$ היא

$$L_{\mathcal{D}}(\mathcal{H}) = \inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h)$$

הגדרה 1.3 (סיבוכיות המדגם) יהא $\mathcal{A}$ אלגוריתם למידה. עבור $\delta, \epsilon > 0$ נסמן ב־$m_{\mathcal{A}}(\epsilon, \delta)$ את המספר המינימאלי כך שלכל התפלגות $\mathcal{D}$ ו־$m \geq m_{\mathcal{A}}(\epsilon, \delta)$ מתקיים

$$\Pr_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}(\mathcal{A}(S)) > L_{\mathcal{D}}(\mathcal{H}) + \epsilon\right) \leq \delta$$

הגדרה 1.4 אלגוריתם למידה יקרא **ERM** אם לכל מדגם S הוא מחזיר $h \in \mathcal{H}$ המקיימת $L_S(h) = \inf_{h'\in\mathcal{H}} L_S(h')$. כאן,

$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m} l(h(x_i), y_i)$$

היא **השגיאה האמפירית** של h ביחס למדגם S.

הגדרה 1.5 תת קבוצה $A \subset X$ **מנותצת ע"י** $\mathcal{H}$ אם מתקיים $\mathcal{H}|_A = \{0,1\}^A$. נסמן ב־$\mathrm{VC}(\mathcal{H})$ את הגודל המקסימאלי של תת קבוצה מנותצת.

משפט 1.6 (המשפט היסודי) קיימים קבועים $C_1, C_2 > 0$ כך שלכל $\mathcal{H} \subset \{0,1\}^X$ ולכל $\mathcal{A}$ ERM מתקיים

$$m_{\mathcal{A}}(\epsilon,\delta) \leq C_1 \cdot \frac{\mathrm{VC}(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)}{\epsilon^2},\ \ m^r_{\mathcal{A}}(\epsilon,\delta) \leq C_1 \cdot \frac{\mathrm{VC}(\mathcal{H})\log\left(\frac{1}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)}{\epsilon}$$

יתר על כן, לכל אלגוריתם (לאו דווקא ERM) מתקיים

$$m_{\mathcal{A}}(\epsilon,\delta) \geq C_2 \cdot \frac{\mathrm{VC}(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)}{\epsilon^2},\ \ m^r_{\mathcal{A}}(\epsilon,\delta) \geq C_2 \cdot \frac{\mathrm{VC}(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)}{\epsilon}$$

1.1 מבנה ההוכחה

הוכחת החסם התחתון תועבר בתירגול. אנו נתרכז בחסם העליון, ונתמקד במקרה הלא פריד (הוכחת המקרה הפריד מתבססת על רעיונות דומים). מבנה ההוכחה יהיה דומה למבנה ההוכחה של החסם על מחלקות סופיות מלפני החג. אנו נראה שאם מספר הדוגמאות גדול מהחסם, אז, בהסתברות $1-\delta <$, השגיאה האמפירית של כל ההיפותזות ב־$\mathcal{H}$ קרובה לשגיאה האמיתית עד כדי $\frac{\epsilon}{2}$. כלומר, מתקיים

$$\Pr_{S\sim\mathcal{D}^m}\left(\exists h\in\mathcal{H},\ \ |L_S(h) - L_{\mathcal{D}}(h)| \geq \frac{\epsilon}{2}\right) \leq \delta \tag{1}$$

מכאן נובע החסם העליון, שכן, כפי שראינו בהוכחת החסם עבור מחלקות סופיות, לכל $\mathcal{A}$ ERM מתקיים

$$\Pr_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}(\mathcal{A}(S)) \geq L_{\mathcal{D}}(\mathcal{H}) + \epsilon\right) \leq \Pr_{S\sim\mathcal{D}^m}\left(\exists h\in\mathcal{H},\ \ |L_S(h) - L_{\mathcal{D}}(h)| \geq \frac{\epsilon}{2}\right) \tag{2}$$

על מנת להראות שמתקיים אי־שוויון (1) לא נוכל להשתמש, כמו בהוכחת החסם על מחלקות סופיות, בחסם האיחוד. למשל, כי $\mathcal{H}$ עשוייה להיות אינסופית, ואז החסם חסר משמעות. אנחנו נתגבר על המכשול הנ"ל בשני שלבים:

1. **(למת סאור, שלח, ופניק וצ'רבוננקינס, פרלס**[1]**)** אנו נראה שאם $\mathrm{VC}(\mathcal{H}) = d$, אז $\mathcal{H}$ היא "קטנה" במובן הבא - לכל תת־קבוצה $A \subset X$ מתקיים

$$|\mathcal{H}|_A| \leq (|A|+1)^d$$

נשים לב שכאשר כאשר A גדולה $(|A|+1)^d \ll 2^{|A|}$. כלומר, מספר הפונקציות ב־$\mathcal{H}|_A$ קטן מאד ממספר הפונקציות מ־A ל־$\{0,1\}$.

2. **(למת המדגם הכפול)** אנו נראה שמהעובדה ש־$\mathcal{H}$ קטנה במובן הנ"ל נובע שכאשר מספר הדוגמאות גדול מ־$C_1 \cdot \frac{\mathrm{VC}(\mathcal{H})+\log\left(\frac{1}{\delta}\right)}{\epsilon^2}$ מתקיים אי־שוויון (1) בהסתברות $1-\delta$.

2 למת המדגם הכפול

הגדרה 2.1 פונקציית הגידול של $\mathcal{H}$ היא הפונקציה $\pi_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ המוגדרת ע"י

$$\pi_{\mathcal{H}}(m) = \max_{A \subset X,\ |A| \leq m} |\mathcal{H}|_A|$$

באופן כללי, קשה לחשב את $\pi_{\mathcal{H}}$. נראה מספר מקרים בהם זה אפשרי

- אם $\mathcal{H} = \{0,1\}^X$ אז

$$\pi_{\mathcal{H}}(m) = \begin{cases} 2^m & m \leq |X| \\ 2^{|X|} & m > |X| \end{cases}$$

- נניח ש־X אינסופית ו־

$$\mathcal{H} = \{h : X \to \{0,1\} \mid |h^{-1}(1)| = 1\}$$

אז

$$\pi_{\mathcal{H}}(m) = m+1$$

(מדוע?)

ניגש ללמת המדגם הכפול. ניזכר כי בהוכחת החסם העליון עבור מחלקות סופיות ראינו שמתקיים

$$\Pr_{S \sim \mathcal{D}^m} \left(\exists h \in \mathcal{H} \text{ s.t. } |L_S(h) - L_{\mathcal{D}}(h)| \geq \epsilon\right) \leq |\mathcal{H}| \cdot 2\exp\left(-2\epsilon^2 m\right)$$

כאן, ה"מקור" של הנכפל $2\exp\left(-2\epsilon^2 m\right)$ הוא חסם הופדינג לפיו, עבור $h \in \mathcal{H}$ קבועה, מתקיים $\Pr_{S\sim\mathcal{D}^m}\left(|L_S(h) - L_{\mathcal{D}}(h)| \geq \epsilon\right) \leq 2\exp\left(-2\epsilon^2 m\right)$. לעומת זאת, המקור של הנכפל $|\mathcal{H}|$ הוא חסם האיחוד. למת המדגם הכפול תאפשר לנו להחליף את $|\mathcal{H}|$ בפונקציית הגידול.

[1] הלמה אכן הוכחה באופן בלתי תלוי ובהקשרים שונים, ע"י 4(!) קבוצות שונות של כותבים.

למה 2.2 (המדגם הכפול) לכל מחלקה מתקיים

$$\Pr_{S\sim\mathcal{D}^m}\left(\exists h\in\mathcal{H} \text{ s.t. } |L_S(h)-L_{\mathcal{D}}(h)|\geq\epsilon\right)\leq\pi_{\mathcal{H}}(2m)\cdot 4\exp\left(-\frac{\epsilon^2 m}{8}\right)$$

רעיון ההוכחה: אנו רוצים לחסום את

$$\Pr\left(\exists h\in\mathcal{H} \text{ s.t. } |L_S(h)-L_{\mathcal{D}}(h)|\geq\epsilon\right)$$

הרעיון המרכזי הוא כדלכמן. אנו נניח ש־S נדגם בשני שלבים: בשלב הראשון נדגם מעין "מדגם רפאים" גדול

$$S'=\{(x'_1,y'_1),\ldots,(x'_{2m},y'_{2m})\}$$

בשלב השני הדוגמאות ב־S' חולקו אקראית לשתי קבוצות זהות בגודלן, והקבוצה הראשונה נבחרה להיות המדגם S. הנחה זו לא מגבילה את הכלליות, שכן עדיין $S\sim\mathcal{D}^m$. כיצד ההנחה, הלכאורה מוזרה, הזו תשרת אותנו? ובכן, אנו נראה שהמדגם S' יכול להוות מעין "תחליף" להתפלגות $\mathcal{D}$. קונקרטית, נראה ש־

$$\Pr\left(\exists h\in\mathcal{H} \text{ s.t. } |L_S(h)-L_{\mathcal{D}}(h)|\geq\epsilon\right)\approx\Pr\left(\exists h\in\mathcal{H} \text{ s.t. } |L_S(h)-L_{S'}(h)|\geq\epsilon\right)$$

אבחנה זו מאפשרת לנו "להצטמצם ל־S'", ובפרט "להחליף" את $\mathcal{H}$ בצמצום שלה ל־S', ובכך להקטין את מספר ההיפתזות ה"אפקטיבי" ללכל היותר $\pi_{\mathcal{H}}(|S'|)=\pi_{\mathcal{H}}(2m)$.

ניגש להוכחה עצמה. אנו נסמן ב־S,S' זוג מדגמים שנדגמו באופן שתואר. קונקרטית אנו נניח שהדרך בה אנו מחלצים את S מתוך S' היא שמכל זוג דוגמאות עוקבות אנו בוחרים בדיוק אחת שתהיה ב־S. כמו כן, נעשה abuse of notation ונסמן ב־$\mathcal{H}|_{S'},h|_{S'}$ את הצמצומים של h ו־$\mathcal{H}$ ל־$\{x_1,\ldots,x_{2m}\}$. נוכיח ראשית ש־"ניתן להחליף את $\mathcal{D}$ ב־S'".

טענה 2.3 מתקיים

$$\Pr\left(\exists h\in\mathcal{H} \text{ s.t. } |L_S(h)-L_{\mathcal{D}}(h)|\geq\epsilon\right)\leq\Pr\left(\exists h\in\mathcal{H} \text{ s.t. } |L_S(h)-L_{S'}(h)|\geq\frac{\epsilon}{4}\right)+2\exp\left(-\frac{\epsilon^2 m}{2}\right)$$

הוכחה: רעיון ההוכחה פשוט - נניח שמתקיים המאורע בצד שמאל, כלומר, יש $h\in\mathcal{H}$ המקיימת $|L_S(h)-L_{\mathcal{D}}(h)|\geq\epsilon$. לשם פשטות, נניח שמתקיים $L_S(h)\geq L_{\mathcal{D}}(h)+\epsilon$. האבחנה העיקרית היא שהמדגם $S'\setminus S$ מתפלג לפי $\mathcal{D}^m$ **באופן בלתי תלוי ב־**S. לכן, בהסתברות גבוהה, עבור אותה h, $L_{\mathcal{D}}(h)\approx L_{S'\setminus S}(h)$. ובמקרה הזה מתקיים

$$L_{S'}(h)=\frac{L_S(h)+L_{S\setminus S'}(h)}{2}\gtrapprox\frac{L_{\mathcal{D}}(h)+\epsilon+L_{\mathcal{D}}(h)}{2}>L_{\mathcal{D}}(h)+\frac{\epsilon}{4}$$

כלומר, מתקיים המאורע המופיע בצד ימין.

ניגש להוכחה. נסמן ב־A את המאורע

$$\exists h \in \mathcal{H} \text{ s.t. } |L_S(h) - L_{\mathcal{D}}(h)| \geq \epsilon$$

וכמו כן ב־B את המאורע

$$\exists h \in \mathcal{H} \text{ s.t. } |L_S(h) - L_{S'}(h)| \geq \frac{\epsilon}{4}$$

נטען ש־

$$\Pr(B|A) \geq 1 - 2\exp\left(-\frac{\epsilon^2 m}{2}\right) \tag{3}$$

זה יספיק, כי אז נקבל

$$\begin{aligned}
\Pr(B) &= \Pr(B|A)\Pr(A) + \Pr(B|A^c)\Pr(A^c) \\
&\geq \Pr(B|A)\Pr(A) \\
&\geq \left(1 - 2\exp\left(-\frac{\epsilon^2 m}{2}\right)\right)\Pr(A) \\
&\geq \Pr(A) - 2\exp\left(-\frac{\epsilon^2 m}{2}\right)
\end{aligned}$$

נוכיח, אם כן, ש־(3) מתקיים. אכן, בהינתן A, יש $h \in \mathcal{H}$ המקיימת $|L_S(h) - L_{\mathcal{D}}(h)| \geq \epsilon$. נשים לב שהמדגם $S' \setminus S$ הינו בלתי תלוי ב־S. לכן, מחסם הופדינג נקבל שעבור אותה h, $|L_{S'\setminus S}(h) - L_{\mathcal{D}}(h)| < \frac{\epsilon}{2}$ בהסתברות לפחות $1 - 2\exp\left(-\frac{\epsilon^2 m}{2}\right)$. במקרה הזה B מתקיים, שכן,

$$\begin{aligned}
|L_S(h) - L_{S'}(h)| &= \left|L_S(h) - \frac{1}{2}\left(L_S(h) + L_{S'\setminus S}(h)\right)\right| \\
&= \frac{1}{2}\left|L_S(h) - L_{S'\setminus S}(h)\right| \\
&\geq \frac{1}{2}\left[|L_S(h) - L_{\mathcal{D}}(h)| - \left|L_{\mathcal{D}}(h) - L_{S'\setminus S}(h)\right|\right] \\
&\geq \frac{1}{2}\left[\epsilon - \frac{\epsilon}{2}\right] = \frac{\epsilon}{4}
\end{aligned}$$

■

הטענה הבאה תחסום את $\Pr\left(\exists h \in \mathcal{H} \text{ s.t. } |L_S(h) - L_{S'}(h)| \geq \epsilon\right)$

טענה 2.4 מתקיים

$$\Pr\left(\exists h \in \mathcal{H} \text{ s.t. } |L_S(h) - L_{S'}(h)| \geq \epsilon\right) \leq \pi_{\mathcal{H}}(2m) \cdot 2\exp\left(-2\epsilon^2 m\right) \tag{4}$$

הוכחה: אנו נראה טענה מעט יותר חזקה - **שלכל** מדגם קבוע S' בן $2m$ נקודות מתקיים אי־השוויון הדרוש, כאשר ההסתברות היא על הבחירה של S מתוך S'. בפרט, לאורך ההוכחה נניח ש־S' איננו אקראי, וההסתברויות יהיו רק על פני הבחירה של S מתוך S'. הטענה הזו חזקה יותר שכן היא תהיה נכונה בפרט עבור המדגם S' שיוגרל בשלב הראשון (לפני הבחירה של S).

צעד א': ראשית, נקבע $h \in \mathcal{H}$ בודדת. מכיוון ש־S נבחר מתוך S' ע"י כך שאנו לוקחים בדיוק אחת מתוך כל זוג דוגמאות עוקבות, מתקיים

$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m} l(h(x'_{2i-1}), y'_{2i-1}) \cdot Z_i + l(h(x'_{2i}), y'_{2i}) \cdot (1 - Z_i)$$

כאשר $Z_1, \ldots, Z_m \in \{0,1\}$ הם משתנים מקריים בלתי תלויים המתפלגים אחיד ("מטבעות הוגנים"). נסמן

$$U_i = l(h(x'_{2i-1}), y'_{2i-1}) \cdot Z_i + l(h(x'_{2i}), y'_{2i}) \cdot (1 - Z_i)$$

נשים לב ש־$U_1, \ldots, U_m \in [0,1]$ הם מ"מ ב"ת וכמו כן, מתקיים

$$\begin{aligned}
E[L_S(h)] &= \frac{1}{m}\sum_{i=1}^{m} E[U_i] \\
&= \frac{1}{m}\sum_{i=1}^{m}\left[\frac{l(h(x'_{2i-1}), y'_{2i-1}) + l(h(x'_{2i}), y'_{2i})}{2}\right] \\
&= \frac{1}{2m}\sum_{i=1}^{2m} l(h(x'_i), y'_i) = L_{S'}(h)
\end{aligned}$$

לכן, לפני חסם הופדינג[2]

$$\Pr\left(|L_S(h) - L_{S'}(h)| \geq \epsilon\right) \leq 2\exp\left(-2\epsilon^2 m\right) \tag{5}$$

צעד ב': כעת עלינו לחסום את ההסתברות ש־$|L_S(h) - L_{S'}(h)| \geq \epsilon$ עבור איזשהי $h \in \mathcal{H}$. נשים לב ש־$|L_S(h|_{S'}) - L_{S'}(h|_{S'})| = |L_S(h) - L_{S'}(h)|$. לכן,

$$\Pr\left(\exists h \in \mathcal{H} \text{ s.t. } |L_S(h) - L_{S'}(h)| \geq \epsilon\right) = \Pr\left(\exists h \in \mathcal{H}|_{S'} \text{ s.t. } |L_S(h) - L_{S'}(h)| \geq \epsilon\right)$$

כעת, אי־שיוויון (4) נובע מחסם האיחוד, אי־שוויון (5) ומהעובדה ש־$|\mathcal{H}|_{S'}| \leq \pi_{\mathcal{H}}(2m)$. ■

כעת אנו מוכנים להסיק את למת המדגם הכפול.

הוכחה: (של למת המדגם הכפול) מתקיים

$$\begin{aligned}
\Pr\left(\exists h \in \mathcal{H} \text{ s.t. } |L_S(h) - L_{\mathcal{D}}(h)| \geq \epsilon\right) &\leq \Pr\left(\exists h \in \mathcal{H} \text{ s.t. } |L_S(h) - L_{S'}(h)| \geq \frac{\epsilon}{4}\right) + 2\exp\left(-\frac{\epsilon^2 m}{2}\right) \\
&\leq \pi_{\mathcal{H}}(2m) \cdot 2\exp\left(-\frac{\epsilon^2 m}{8}\right) + 2\exp\left(-\frac{\epsilon^2 m}{2}\right) \\
&\leq \pi_{\mathcal{H}}(2m) \cdot 4\exp\left(-\frac{\epsilon^2 m}{8}\right)
\end{aligned}$$

---

[2] כאן אנו מסתמכים על הגרסה הבאה של חסם הופדינג: אם $U_1, \ldots, U_m \in [0,1]$ הם מ"מ ב"ת אזי $\Pr\left(\left|\frac{1}{m}\sum_{i=1}^{m} U_i - \frac{1}{m}\sum_{i=1}^{m} E[U_i]\right| \geq \epsilon\right) \leq 2\exp\left(-2\epsilon^2 m\right)$.

■

3 למת סאור־שלח

על מנת להשתמש בלמת המדגם הכפול, אנו נזדקק לחסם על $\pi_{\mathcal{H}}(m)$. החסם הנ"ל מתקבל מלמת סאור־שלח(־פרלס־ופניק וצ'רבוננקיס).

למה 3.1 (סאור־שלח) נסמן $|X| = m$. לכל מחלקה $\mathcal{H} \subset \{0,1\}^X$ עם $\mathrm{VC}(\mathcal{H}) = d$ מתקיים[3]

$$|\mathcal{H}| \leq \binom{m}{0} + \binom{m}{1} + \ldots + \binom{m}{d}$$

מסקנה 3.2 לכל מחלקה $\mathcal{H} \subset \{0,1\}^X$ עם $\mathrm{VC}(\mathcal{H}) = d$ מתקיים

$$\pi_{\mathcal{H}}(m) \leq (m+1)^d$$

הוכחה: (של המסקנה) על מנת להראות ש־$\pi_{\mathcal{H}}(m) \leq (m+1)^d$ צ"ל שלכל $A \subset X$ בגודל m מתקיים $|\mathcal{H}|_A| \leq (m+1)^d$. אבל $\mathcal{H}|_A$ היא מחלקה ממימד VC לכל היותר d ולכן מלמת שאור־שלח נובע ש־

$$|\mathcal{H}|_A| \leq \binom{m}{0} + \binom{m}{1} + \ldots + \binom{m}{d} \leq (m+1)^d$$

כאשר אי השוויון הימני נובע מכך ש־$\binom{m}{0}+\binom{m}{1}+\ldots+\binom{m}{d}$ הוא בדיוק מספר תתי הקבוצות של A בגודל $\leq d$, והמספר הנ"ל הוא לכל היותר $(m+1)^d$ (למה?) ■

לפני שנוכיח את למת סאור־שלח, נעיר כי אי השוויון הוא הדוק. עבור קבוצה X בגודל m, נביט ב־

$$\mathcal{H} = \{h : X \to \{0,1\} \mid |h^{-1}(1)| \leq d\}$$

לא קשה לראות ש־$\mathrm{VC}(\mathcal{H}) = d$ וכמו כן ש־$|\mathcal{H}| = \binom{m}{0} + \binom{m}{1} + \ldots + \binom{m}{d}$.

3.1 הוכחת הלמה

נניח בה"כ ש־$X = [m] := \{1, \ldots, m\}$. הוכחת המשפט תהיה באינדוקציה על $d+m$.

[3] שימו לב שתמיד $d \leq m$.

בסיס האידוקציה

ראשית, נשים לב ש־d יכול להיות כל מספר טבעי, כולל אפס. כמו כן m יכול להיות כל מספר טבעי הגדול או שווה ל־d. בבסיס האינדוקציה נוכיח את הלמה למקרה בו $d = 0$ (ו־m כלשהו), וכמו כן למקרה בו $m = d$.

מקרה 1. נניח ש־$d = 0$ ו־m הוא מספר כלשהו. מכיוון ש־$d = 0$, נובע ש־$\mathcal{H}$ מכילה לכל היותר פונקציה בודדת (למה?). לכן מתקיים

$$|\mathcal{H}| \leq 1 = \binom{m}{0}$$

מקרה 2. נניח ש־$m = d$. במקרה הזה,

$$|\mathcal{H}| \leq 2^m = \binom{m}{0} + \binom{m}{1} + \ldots + \binom{m}{m}$$

שלב האינדוקציה

המקרה $m = d$ כוסה בבסיס האינדוקציה. נניח, אם כן, ש־$m > d$. נגדיר

$$\begin{aligned} \mathcal{H}_1 &= \mathcal{H}|_{[m-1]} \\ \mathcal{H}_2 &= \{h \in \mathcal{H}|_{[m-1]} \mid \exists h_1, h_2 \in \mathcal{H} \text{ s.t. } h_1|_{[m-1]} = h_2|_{[m-1]} = h \text{ and } h_1 \neq h_2\} \end{aligned}$$

במילים, $\mathcal{H}_1$ היא אוסף כל הפונקציות מ־$[m-1]$ ל־$\{0,1\}$ שניתן להרחיב לפונקציה מ־$[m]$ ל־$\{0,1\}$ השייכת ל־$\mathcal{H}$. כמו כן, $\mathcal{H}_2$ היא אוסף כל הפונקציות מ־$[m-1]$ ל־$\{0,1\}$ שניתן להרחיב **בשתי דרכים שונות** לפונקציה מ־$[m]$ ל־$\{0,1\}$ השייכת ל־$\mathcal{H}$.

טענה 3.3 $\mathrm{VC}(\mathcal{H}_1) \leq d$

הוכחה: לא קשה לראות שאם $A \subset [m-1]$ מנותצת ע"י $\mathcal{H}_1$ אז היא מנותצת גם ע"י $\mathcal{H}$ ומכיוון ש־$\mathrm{VC}(\mathcal{H}) \leq d$ נקבל ש־$|A| \leq d$. ■

טענה 3.4 $\mathrm{VC}(\mathcal{H}_2) \leq d - 1$

הוכחה: לא קשה לראות שאם $A \subset [m-1]$ מנותצת ע"י $\mathcal{H}_2$ אז $A \cup \{m\}$ מנותצת גם ע"י $\mathcal{H}$ ומכיוון ש־$\mathrm{VC}(\mathcal{H}) \leq d$ נקבל ש־$|A| = |A \cup \{m\}| - 1 \leq d - 1$. ■

טענה 3.5 $|\mathcal{H}| = |\mathcal{H}_1| + |\mathcal{H}_2|$

הוכחה: עבור $h \in \mathcal{H}_1$ נסמן

$$G_h = \{h' \in \mathcal{H} : h'|_{[m-1]} = h\}$$

לא קשה לראות ש־ $\mathcal{H} = \dot\cup_{h\in\mathcal{H}_1} G_h$ ולכן

$$|\mathcal{H}| = \sum_{h\in\mathcal{H}_1} |G_h|$$

כמו כן, לכל $h \in \mathcal{H}_1$, $|G_h| = 1$ או $|G_h| = 2$. והמקרה השני מתקיים אמ"מ $h \in \mathcal{H}_2$ לכן

$$\sum_{h\in\mathcal{H}_1} |G_h| = |\mathcal{H}_1| + |\mathcal{H}_2|$$

■

טענה 3.6 עבור $m \le d$ נסמן $N(d,m) = \binom{m}{0} + \binom{m}{1} + \ldots + \binom{m}{d}$. אזי, לכל $d < m$ מתקיים

$$N(d,m) = N(d,m-1) + N(d-1,m-1)$$

הוכחה: (הוכחה 1 - סיפור קומבינטורי) $N(d,m)$ סופר את מספר הדרכים לבחור תת קבוצה בת $\le d$ איברים מתוך $\{1,\ldots,m\}$. אבל, מספר הדרכים לעשות זאת הוא גם $N(d,m-1) + N(d-1,m-1)$ - על מנת לבחור קבוצה בת d איברים, אפשר לבחור קודם אם מכניסים את m לקבוצה או לא. אח"כ, אם החלטנו להכניס את m, עלינו לבחור $\le (d-1)$ איברים מתוך $\{1,\ldots,m-1\}$ ויש $N(d-1,m-1)$ דרכים לעשות זאת. אם החלטנו לא להכניס את m, עלינו לבחור $\le d$ איברים מתוך מתוך $\{1,\ldots,m-1\}$ ויש $N(d,m-1)$ דרכים לעשות זאת. ■

הוכחה: (הוכחה 2 - הוכחה אלגברית) מהשוויון $\binom{m}{j} = \binom{m-1}{j} + \binom{m-1}{j-1}$ הנכון לכל $j < m$ נקבל

$$\begin{aligned}
N(d,m-1) + N(d-1,m-1) &= \binom{m-1}{0} + \binom{m-1}{1} + \ldots + \binom{m-1}{d-1} \\
&\quad + \binom{m-1}{0} + \binom{m-1}{1} + \ldots + \binom{m-1}{d} \\
&= \binom{m-1}{0} + \left[\binom{m-1}{0} + \binom{m-1}{1}\right] + \ldots + \left[\binom{m-1}{d-1} + \binom{m-1}{d}\right] \\
&= \binom{m}{0} + \binom{m}{1} + \ldots + \binom{m}{d} = N(d,m)
\end{aligned}$$

■

מארבעת הטענות לעיל ומהנחת האינדוקציה נקבל ש־

$$|\mathcal{H}| = |\mathcal{H}_1| + |\mathcal{H}_2| \le N(d,m-1) + N(d-1,m-1) = N(d,m)$$

4 הוכחת המשפט היסודי

כעת אנו מוכנים להסיק את המשפט היסודי. כאמור, נוכיח רק את החסם העליון במקרה הלא פריד. כמו כן, אנו נוכיח חסם מעט יותר חלש - נוכיח שקיים $C > 0$ כך שלכל $\mathcal{H}$ ממימד VC d ולכל ERM $\mathcal{A}$ מתקיים

$$m_{\mathcal{A}}(\epsilon, \delta) \leq C\frac{d\log\left(\frac{d}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)}{\epsilon^2}$$

אכן, לפי אי־שיוויון (2) למת החסם הכפול יחד עם למת סאור־שלח מתקיים

$$\begin{aligned}
\Pr_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}(\mathcal{A}(S)) \geq L_{\mathcal{D}}(\mathcal{H}) + \epsilon\right) &\leq \Pr_{S\sim\mathcal{D}^m}\left(\exists h \in \mathcal{H},\ \ |L_S(h) - L_{\mathcal{D}}(h)| < \frac{\epsilon}{2}\right) \\
&\leq \pi_{\mathcal{H}}(2m)\cdot 4\exp\left(-\frac{\epsilon^2 m}{32}\right) \\
&\leq (2m+1)^d\cdot 4\exp\left(-\frac{\epsilon^2 m}{32}\right)
\end{aligned}$$

נשים לב שהנכפל $(2m+1)^d$ גדל פולינומיאלית ביחס ל־m, בעוד הנכפל השני, $\exp\left(-\frac{\epsilon^2 m}{32}\right)$ קטן אקספוננציאלית. לכן, עבור m מספיק גדול החסם יהיה קטן מ־δ. קונקרטית, הלמה (הטכנית למדי) הבאה, שמסיימת את הוכחת המשפט היסודי, מראה שקיים קבוע $C > 0$ כך שאם $m > C\cdot\frac{d\log\left(\frac{d}{\epsilon}\right)+\log\left(\frac{1}{\delta}\right)}{\epsilon^2}$ אז החסם קטן מ־δ.

למה 4.1 קיים קבוע $C > 0$ כך שאם

$$m > C\cdot\frac{d\log\left(\frac{d}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)}{\epsilon^2}$$

אז

$$(2m+1)^d\cdot 4\exp\left(-\frac{\epsilon^2 m}{32}\right) < \delta$$

אנו נשתמש בחסם הבא (למה $A.2$ בספר של שי ושי):

טענה 4.2 עבור $a \geq 1, b > 0$, אם $x \geq 4a\log(2a) + 2b$ אז $x \geq a\log(x) + b$.

הוכחה: (של למה 4.1) אכן, אם

$$m > 32\cdot\frac{4d\log\left(\frac{64d}{\epsilon^2}\right) + 2d\log(12) + 2\log\left(\frac{1}{\delta}\right)}{\epsilon^2}$$

אז מהטענה (עבור $a = \frac{32d}{\epsilon^2},\ b = \frac{32d\log(12)+32\log(1/\delta)}{\epsilon^2}$)

$$m > 32\cdot\frac{d(\log(m) + \log(12)) + \log\left(\frac{1}{\delta}\right)}{\epsilon^2} = 32\cdot\frac{d(\log(12m)) + \log\left(\frac{1}{\delta}\right)}{\epsilon^2}$$

ואז, לפי אי־שיוויון (2) למת החסם הכפול יחד עם למת סאור־שלח

$$\begin{aligned}(2m+1)^d \cdot 4\exp\left(-\frac{\epsilon^2 m}{32}\right) &\leq (12m)^d \cdot \exp\left(-\frac{\epsilon^2 m}{32}\right)\\ &= \exp\left(d\log(12m) - \frac{\epsilon^2 m}{32}\right)\\ &\leq \exp\left(d\log(12m) - d\log(12m) + \log\left(\frac{1}{\delta}\right)\right) = \delta\end{aligned}$$

■