

הרצאה 5 - יסודות בלמידה

חישובית

PDF

מה נלמד היום



בפרק זה נציג מעט מהתיאוריה הכמותית הקיימת בנושא למידה והכללה. המטרה הבסיסית של תיאוריה זו היא תיאור כמותי של בעיית הלמידה, אפיון הביצועים האפשריים עבור בעיית למידה נתונה, וחקר כמותי של השפעת המרכיבים השונים של הבעיה (כגון: סיבוכיות המודל, אופן בחירת הדגימות, מספר הדגימות, וכו') על הביצועים המתקבלים.

תיאוריה זו היא בעיקרה **בעלת אופי סטטיסטי**, כלומר מסתמכת על **כלים הסתברותיים**.

אנו נסתפק בהצגת מספר תוצאות ומושגים יסודיים, וזאת עבור **בעיית הסיווג הבינארי בלבד**.

מודל הלמידה הבסיסי

נזכור כי בבעיית הלמידה המודרכת אנו נדרשים "ללמוד" פונקציה $\hat{y} = h(\mathbf{x})$ על סמך מדגם $\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^N$.

המודל הבסיסי בו נעסוק כולל את המרכיבים הבאים:

- **פונקציית החיזוי** - פונקציה $\hat{y} = h(\mathbf{x})$ ממרחב הקלט \mathcal{X} למרחב היציאה \mathcal{Y} אותה אנו רוצים ללמוד. נזכיר כי עבור בעיית רגרסיה מתקיים $\mathcal{Y} = \mathbb{R}$ ועבור בעיית הסיווג הבינארי מתקיים $\mathcal{Y} = \{-1, 1\}$. נניח כי התיג דטרמיניסטי.
- **מודל בחירת הדוגמאות** - דוגמאות הקלט נבחרות באופן בלתי תלוי ולפי פילוג הסתברות קבוע (אך לא בהכרח ידוע), כלומר באופן i.i.d. כלומר מתקיים, $\mathbf{x}^{(i)} \sim P_X, i = 1, \dots, N$. הדוגמאות מתויגות באופן מושלם לפי הפונקציה h_0 , כלומר $y^{(i)} = h_0(\mathbf{x}^{(i)})$.
- **מודל פרמטרי** - אוסף H של פונקציות $H : \mathcal{X} \rightarrow \mathcal{Y}$, שמתוכו נבחר את הפונקציה \hat{h} אשר משערכת את פונקציית המטרה h . כאשר H תכונה כאן **מחלקת ההשערות**.

פונקציית הסיכון עבור השערה $h \in H$ כלשהי תהיה מהצורה

$$R(h) = E[l(h(\mathbf{x}), h_0(\mathbf{x}))]$$

כאשר:

- $l(\hat{y}, y)$ הינה פונקציית מחיר מתאימה. למשל פונקציית הפסד l_2 לבעיית רגרסיה או zero-one loss לבעיית סיווג.
- התוחלת היא על המשטנה המקרי \mathbf{x} לפי הפילוג $\mathbf{x} \sim P_X$. פילוג זה זהה לפילוג לפיו נבחרו הדוגמאות.
- עבור בעיית הסיווג הבינארי נקבל $R(\hat{h}) = P\{\hat{h}(\mathbf{x}) \neq h_0(\mathbf{x})\} = P_e(\hat{h})$. כאשר המעבר השני נכון בגלל תוחלת של אינדיקטור.

מטרת תהליך הלימוד היא, אם כן, לבחור את הפונקציה האופטימלית כתלות במדגם, h_D^* , מתוך מחלקת ההשערות H , שמביאה את פונקציית הסיכון למינימום.

הבעיה היא כמובן ש- $R(h)$ אינו ניתן לחישוב מתוך מדגם סופי!

- חשוב להדגיש כי הדוגמאות $\{\mathbf{x}^{(i)}\}$ נבחרות לפי פילוג P_X המשמש בהגדרת מדד הביצועים. דבר זה יאפשר קבלת חסמים על קצב ושגיאת הלימוד שאינם תלויים ב- P_X .
- המודל הנ"ל מניח קשר דטרמיניסטי בין \mathbf{x} ל- y . ניתן להרחיב את התוצאות הללו למקרה של קשר אקראי, כלומר להחליף את הפונקציה $y = h_0(\mathbf{x})$ בפילוג המותנה $p(y|\mathbf{x})$.

המודל ההסתברותי שהגדרנו מאפשר התייחסות כמותית לשאלות הבאות:

- **דיוק הלמידה** - באיזה דיוק ניתן ללמוד את פונקציית המטרה $h_0(\mathbf{x})$ מתוך N דוגמאות?
- **קצב הלמידה** - כמה דוגמאות נדרשות כדי להשיג דיוק נתון?

מזעור המחיר האמפירי (Empirical Risk) (Minimization)

בהיעדר מידע לגבי הפילוג, ניתן להחליף את המזעור של פונקציית הסיכון האמיתית, R , במזעור של פונקציית הסיכון האמפירית, \hat{R} , אותה אנחנו יכולים לחשב על סמך המדגם.

כלומר, בהינתן המדגם $\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^N$, נבחר את ההשערה $h_{\mathcal{D}}^*$ באופן הבא:

$$h_{\mathcal{D}}^* \in \arg \min_{h \in H} \hat{R}_{\mathcal{D}}(h), \quad \hat{R}_{\mathcal{D}}(h) = \frac{1}{N} \sum_{i=1}^N l(h(\mathbf{x}^{(i)}), h_0(\mathbf{x}^{(i)}))$$

לדוגמה:

- עבור בעיית רגרסיה עם פונקציית הפסד מסוג l_2 נקבל את פונקציית הסיכון הבאה:

$$\hat{R}_{\mathcal{D}}(h) = \frac{1}{N} \sum_{i=1}^N (h(\mathbf{x}^{(i)}) - h_0(\mathbf{x}^{(i)}))^2$$

- עבור בעיות סיווג נקבל:

$$\hat{R}_{\mathcal{D}}(h) = \frac{1}{N} \sum_{i=1}^N I(h(\mathbf{x}^{(i)}) \neq h_0(\mathbf{x}^{(i)}))$$

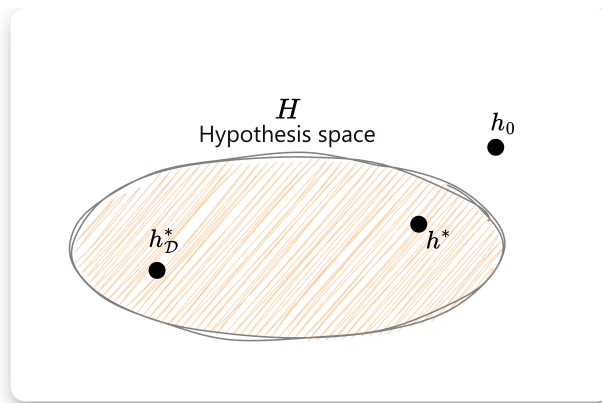
כאשר מדובר במספר השגיאות הממוצע של המסווג על סט הלימוד.

נניח מעתה כי $h_{\mathcal{D}}^*$ היא אכן הפונקציה הנבחרת על ידי אלגוריתם הלמידה שלנו. בפרט, אנו מניחים כי ניתן למצוא את המינימום הגלובאלי של $\hat{R}_{\mathcal{D}}(h)$, מבלי להתייחס לקושי החישובי הכרוך בכך.

הערה: למרות שאנו מניחים מזעור של השגיאה האמפירית אין לראות בכך המלצה לעשות זאת! גישה זו יכולה להוביל להתאמת-יתר חמורה עבור מרחב השערות גדול.

שגיאת ההכללה לעומת שגיאת הקירוב

נסמן - $h^* \in \arg \min_{h \in H} \hat{R}(h)$ - בתור ההשערה האופטימלית שאינה ניתנת לחישוב.



ניתן לרשום את פונקציית הסיכון המתקבלת בצורה הכאה:

$$R(h_D^*) = R(h^*) + [R(h_D^*) - R(h^*)]$$

- האיבר הראשון הוא שגיאת הקירוב (בדומה למשתנה ההטיה, bias), אשר נובע מכך שאנו מגבילים את הפונקציה הנלמדת לקבוצת ההשערות H . הוא אינו תלוי במספר הדגימות.
- האיבר השני הוא שגיאת השערוך (בדומה למשתנה השונות), ומבטא את השגיאה הנובעת מסופיות המדגם עקב כך שהפונקציה הנבחרת h_D^* אינה האופטימלית (מתוך H). זאת מכיוון שאנו מבצעים מינימיזציה של הסיכון האמפירי ולא של הסיכון האמיתי.
- ככל שמחלקת ההשערות H עשירה (גדולה) יותר, אנו מצפים כי האיבר הראשון (איבר ההטיה) יקטן, והאיבר השני (איבר השונות) יגדל.
- עושר המודל (H) צריך להיות כזה המוצא איזון אופטימאלי בין שני איברים אלה.

חסמים עבור מחלקת השערות סופית

נתמקד מעתה בבעיית הסיווג הבינארי, כלומר בעיות סיווג עם פונקציית הפסד מסוג zero-one loss:

$$l(\hat{y}, y) = I\{\hat{y} \neq y\}, \mathcal{Y} = \{-1, +1\}$$

מטרתנו למצוא חסמים על פונקציית הסיכון $R(h_D^*)$, כאשר h_D^* היא הפונקציה (ההשערה) המביאה למינימום את המחיר האמפירי $\hat{R}_D(h)$.

נשים לב כי במקרה הבינארי המחיר האמפירי איננו אלא השגיאה האמפירית (למה?).

ראשית נעסוק במקרה בו $h_0 \in H$, כלומר במקרה בו פונקציה המטרה h_0 כלולה בתוך קבוצת ההשערות H .

כלומר:

$$R^* = \min_{h \in H} R(h) = 0$$

משפט 1

נניח כי $|H| < \infty$ וכן $h_0 \in H$, כלומר $R^* = 0$. אזי, השערה h_D^* הממזערת את הסיכון האמפירי מקיימת לכל $\varepsilon > 0$

$$P(R(h_D^*) > \varepsilon) < |H|e^{-\varepsilon n}$$

ניתן להגדיר את המשפט גם בצורה שקולה באמצעות "רווח סמך" (confidence interval).

רווח סמך הוא מושג מסטטיסטיקה. מושג זה מתאר, עבור פרמטר לא ידוע כלשהו, קטע שמחושב מתוך תוצאות המדגם, כך שהסיכוי שהקטע שנקבל יכלול את הפרמטר הוא קבוע, הקרוי רמת הסמך של הקטע. המשלים לרמת הסמך קרוי רמת המובהקות.

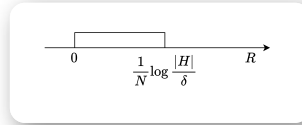
ניסוח מתמטי של רווח סמך הוא:

בהינתן מדגם $\mathcal{D} = \{\mathbf{x}^{(i)}\}$ מהתפלגות F_θ הידועה למעט ערכו של הפרמטר θ , רווח סמך בעל רמת מובהקות α הוא קטע שקצוותיו מחושבים על פי המדגם כך שהסתברות של הפרמטר θ להיות בתוך קטע זה היא $1 - \alpha$.

משפט 1 - בניסוח רווח סמך

על ידי השוואת אגף ימין ל- δ , כלומר בחירת $\varepsilon = \frac{1}{N} \log \frac{|H|}{\delta}$, ניתן לקבל את הצורה הבאה של המשפט, כאשר הפרמטר δ נקרא רווח הסמך:

$$\bullet \text{ לכל } \delta > 0 \text{ מתקיים בהסתברות של } (1 - \delta) \text{ לפחות } R(h_{\mathcal{D}}^*) < \frac{1}{N} \log \frac{|H|}{\delta}$$



משפט 1 - ניסוח סיבוכיות המדגם

החסם שקיבלנו מאפשר לנו לבחור את גודל המדגם N המבטיח שגיאה קטנה כרצוננו, ובהסתברות גבוהה כרצוננו, אם $N > \frac{1}{\varepsilon} \log \frac{|H|}{\delta}$, נקבל כי $R(h_{\mathcal{D}}^*)$ בהסתברות $1 - \delta$ לפחות.

משפט 1 - ניסוח חסם על התוחלת

ננסה בנוסף חסם עבור התוחלת.

עבור השערה $h_{\mathcal{D}}^*$ אי שלילית, התוחלת שלה, $E[R(h_{\mathcal{D}}^*)]$, חסומה על ידי

$$E[R(h_{\mathcal{D}}^*)] < \frac{1 + \log(|H|)}{N} = \mathcal{O}\left(\frac{1}{N}\right)$$

מספר מונחים בסיסיים בלמידה חישובית:

אלגוריתם כלשהו לבחירת $h_{\mathcal{D}}^* \in H$ שעבורו $P(R(h_{\mathcal{D}}^*) > \varepsilon) \rightarrow 0$ כאשר $N \rightarrow \infty$ (לכל $h_0 \in H$) נקרא אלגוריתם Probably Approximately Correct או בקיצור PAC. קבוצת השערות H שעבורה קיים אלגוריתם PAC נקראת ברת-למידה (Learnable).

משפט 1 מראה כי האלגוריתם הממזער את השגיאה האמפירית הוא אלגוריתם PAC עבור כל קבוצת השערות סופית (ולפיכך כל קבוצת השערות סופית היא ברת למידה).

נעבור כעת למקרה הכללי יותר שבו פונקציית המטרה h_0 אינה כלולה בהכרח בקבוצת השערות H , ולמעשה איננו מניחים הנחה כלשהי לגביה. במקרה זה $R^* \neq 0$.

משפט 2

נניח כי $|H| < \infty$ ונסמן שוב $R^* = \min_{h \in H} R(h)$. אזי, לכל $\varepsilon > 0$

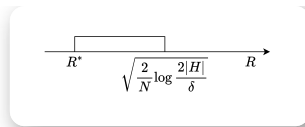
$$P(R(h_{\mathcal{D}}^*) > R^* + \varepsilon) < 2|H|e^{-\frac{1}{2}\varepsilon^2 n}$$

הערות:

- ניתן לראות כי חסם זה חלש מהקודם, כיוון שקצה הדעיכה המעריכי של הסתברות הטעות הינו ε^2
- מהי סיבוכיות המדגם?

משפט 2 - בניסוח רווח סמך

ניסוח רווח סמך עבור משפט זה הוא $R(h_{\mathcal{D}}^*) < R^* + \sqrt{\frac{2}{N} \log \frac{2|H|}{\delta}}$ - בהסתברות $1 - \delta$ לפחות. האיבר הראשון (R^*) מבטא את שגיאת הקירוב, אותה אי אפשר למזער, והשני את שגיאת השערוך.



משפט 2 - ניסוח חסם על התוחלת

ננסח בנוסף חסם עבור התוחלת.

עבור השערה $h_{\mathcal{D}}^*$ אי-שלילית, מתקיים כי $E[R(h_{\mathcal{D}}^*)] - R^*$ חסומה על ידי

$$E[R(h_{\mathcal{D}}^*)] - R^* = \mathcal{O}\left(\frac{\log |H|}{N}\right)$$

הוכחת המשפטים

על מנת להוכיח את המשפטים נגדיר את ההגדרות הבאות:

אוסף ההשערות ב- H העקביות עם הנתונים מוגדר להיות ה-**version space**. אוסף השערות זה מוגדר בצורה הבאה

$$VS_H = \{h_j \in H : \hat{R}_{\mathcal{D}}(h^{(j)}) = 0, j = 1, 2, \dots, |H|\}$$

עבור אלגוריתם המזער את השגיאה האמפירית ידוע כי מתקיים $h_{\mathcal{D}}^* \in VS_H$.

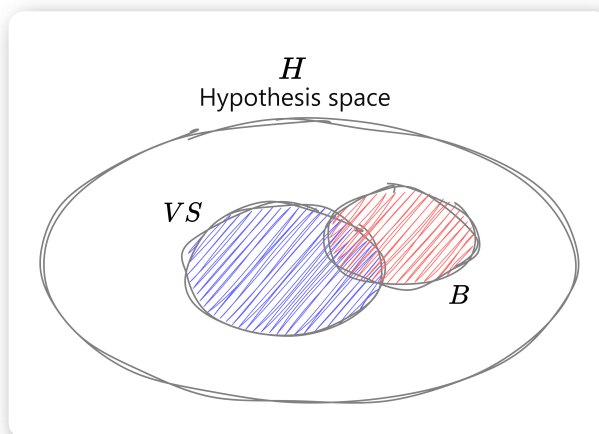
אוסף השערות הרעות ב- H מוגדר בצורה הבאה:

$$B = \{h_j \in H : R(h^{(j)}) > \epsilon, j = 1, 2, \dots, |H|\}$$

הערות:

- שימו לב שההשערות העקביות מוגדרות ע"י השגיאה האמפירית, בעוד שההשערות הרעות מוגדרות ע"י השגיאה ה"אמיתית".
- הקבוצה B אינה אקראית, כלומר אינה תלויה במדגם.
- ככל שגודל המדגם גדל, הקבוצה VS_H (התלויה במדגם) קטנה.

אנו מעוניינים להעריך את ההסתברות שקיימת השערה רעה שהיא עקבית, כלומר, $h \in (VS_H \cap B)$.



בשביל ההוכחה נצטרך את חסם האיחוד (union bound) שהוא

$$P(\cup_{i=1}^N A_i) \leq \sum_{i=1}^N P(A_i) \leq N \max_{1 \leq i \leq N} P(A_i)$$

כאשר שוויון מתקיים אם המאורעות זרים.

הוכחת משפט 1

נתבונן בהשערה מסוימת h_j כך שמתקיים

$$P(h_j(x^{(i)}) = y^{(i)} \text{ and } h_j \in B) < 1 - \varepsilon$$

נשים לב שההסתברות היא רק ביחס למשתנה האקראי $x^{(i)}$ כאשר אנו מגבילים את עצמנו ל- B שכן $h_j \in B$ היא קבוצה לא אקראית.

בגלל שהדגימות i.i.d מתקיים

$$P(h_j \in (VS_H \cap B)) < (1 - \varepsilon)^N$$

נגדיר את $h_j \in (VS_H \cap B)$ להיות המאורע A_j ונשתמש בחסם האיחוד כך שנקבל

$$P(\exists h_j \in (VS_H \cap B)) \leq |B|(1 - \varepsilon)^N$$

הגודל של הקבוצה B אינו ידוע ולכן נרשום

$$P(\exists h_j \in (VS_H \cap B)) \leq |H|(1 - \varepsilon)^N \leq |H|e^{-\varepsilon N}$$

כאשר האי שוויון האחרון נובע מתוך $1 - \varepsilon \leq e^{-\varepsilon}$.

מ.ש.ל.

הוכחת משפט 2

ראשית נזכר באי שיוון צ'בישב

$$P(|X - E[X]| > \varepsilon) \leq \frac{\text{Var}[X]}{\varepsilon^2}$$

אנו מעוניינים במקרה שבו $X = \frac{1}{N} \sum_{i=1}^N Z^{(i)}$ ו- $\{Z^{(i)}\}_{i=1}^N$ משתנים i.i.d. במקרה זה חסם צ'בישב הוא

$$P\left(\left|\frac{1}{N} \sum_{i=1}^N (Z^{(i)} - E[Z^{(i)}])\right| > \varepsilon\right) \leq \frac{\text{Var}\left[\sum_{i=1}^N Z^{(i)}\right]}{N^2 \varepsilon^2} = \frac{\text{Var}[Z^{(1)}]}{N \varepsilon^2}$$

נשים לב לכך שחסם צ'בישב דועך בצורה איטית.

היינו רוצים להשיג חסם טוב יותר, כלומר חסם שדועך בצורה יותר מהירה. לשם כך נציג את אי שוויון Hoeffding.

אי שוויון Hoeffding: יהי $\{Z^{(i)}\}_{i=1}^N$ משתנים אקראיים i.i.d המוגבלים בקטע סופי $a \leq Z^{(i)} \leq b$ אזי

$$P\left(\left|\frac{1}{N} \sum_{i=1}^N (Z^{(i)} - E(Z^{(i)}))\right| > \varepsilon\right) \leq 2 \exp\left(-\frac{2N\varepsilon^2}{(b-a)^2}\right)$$

לחסם זה יש קצב מעריכי.

שימו לב שחסם זה מתעלם משונות המשתנה האקראי. ניתן לקחת אותה בחשבון לצורך שיפור החסם.

מטרתנו בהוכחה זאת היא לחסום את $P(R(h_{\mathcal{D}}^*) - R^* > \varepsilon)$

לשם כך נוכיח את אי השוויונות הבאים:

$$R(h_{\mathcal{D}}^*) - R^* < 2 \max_{h \in H} |R(h) - R(h_{\mathcal{D}}^*)|$$

לשם פשטות נניח כי קיים $h^* \in H$ כך שמתקיים $R^* = R(h^*)$ אזי

$$\begin{aligned}
R(h_{\mathcal{D}}^*) - R^* &= R(h_{\mathcal{D}}^*) - \hat{R}_{\mathcal{D}}(h_{\mathcal{D}}^*) + \hat{R}_{\mathcal{D}}(h_{\mathcal{D}}^*) - R^* \\
&\leq [R(h_{\mathcal{D}}^*) - \hat{R}_{\mathcal{D}}(h_{\mathcal{D}}^*)] + [\hat{R}_{\mathcal{D}}(h_{\mathcal{D}}^*) - R^*] \\
&\leq 2 \max_{h \in H} |R(h) - \hat{R}_{\mathcal{D}}(h)|
\end{aligned}$$

כעת, נרצה להשתמש בחסם Hoeffding. לשם כך נשים לב כי מתקיים:

$$\hat{R}_{\mathcal{D}}(h) = \frac{1}{N} \sum_{i=1}^N Z^{(i)}, \quad Z^{(i)} = I\{h(x^{(i)}) \neq y^{(i)}\}, \quad E(Z^{(i)}) = L(h)$$

כעת נציב בחסם Hoeffding עם $a = 0, b = 1, \frac{\varepsilon}{2}$ ונקבל

$$P(|R(h) - \hat{R}_{\mathcal{D}}(h)| > \frac{\varepsilon}{2}) \leq 2 \exp\left(-N \frac{\varepsilon^2}{2}\right)$$

סה"כ, נוכל להשתמש באי השוויונות שהוכחנו ובחסם האיחוד כך שנקבל

$$\begin{aligned}
P(|R(h_{\mathcal{D}}^*) - R^*| > \varepsilon) &\leq P\left(\max_{h \in H} |R(h) - \hat{R}_{\mathcal{D}}(h)| > \frac{\varepsilon}{2}\right) \\
&\leq |H| \max_{h \in H} P\left(|R(h) - \hat{R}_{\mathcal{D}}(h)| > \frac{\varepsilon}{2}\right) \\
&\leq 2|H| \exp\left(-N \frac{\varepsilon^2}{2}\right)
\end{aligned}$$

נשים לב שמאורע המקסימום שקול למאורע איחוד המאורעות ולכן נוכל להשתמש בחסם האיחוד במעבר מהשורה הראשונה לשנייה.

ובכך מסתכמת הוכחת המשפט השני.

הוכחת החסם על התוחלת

יהי Z משתנה אקראי אי-שלילי כך שמתקיים $P(Z > t) \leq ce^{-2Nt^2}$.

נחשב את התוחלת של Z^2 . לשם כך נצטרך להשתמש בנוסחת הזנב.

תזכורת, יהי X משתנה קראי אי שלילי עם פוקציית התפלגות P . התוחלת של X מקיימת

$$E[X] = \int_0^\infty P(X > x) dx$$

מתקיים

$$\begin{aligned}
E[Z^2] &= \int_0^\infty P(Z^2 > t) dt \\
&= \int_0^u P(Z^2 > t) dt + \int_u^\infty P(Z^2 > t) dt \\
&\leq u + \int_u^\infty P(Z^2 > t) dt \\
&\leq u + c \int_u^\infty e^{-2Nt} dt \\
&= u + \frac{c}{2N} e^{-2Nu}
\end{aligned}$$

כאשר המעבר בין השורה הראשונה לשנייה נכון עבור כל $u \geq 0$.

ניתן למצוא על ידי גזירה והשוואה ל-0 את הערך u שמביא למינימום את הביטוי, $u = \frac{\log c}{2N}$.

כלומר,

$$E[Z^2] \leq \frac{\log c}{2N}$$

לכן, עבור משתנה אקראי מהסוג $R(h_{\mathcal{D}}^*) - R$ שמקיים

$$P(R(h_{\mathcal{D}}^*) > R^* + \varepsilon) < 2|H|e^{-\frac{1}{2}\varepsilon^2 n}$$

מתקיים כי

$$E[R(h_{\mathcal{D}}^*)] - R^* = \mathcal{O}\left(\frac{\log |H|}{N}\right)$$

ובכך מסתכמת הוכחת החסם על התוחלת.

מגבלות החסמים שפותרו

ראינו חסם מהצורה הבאה, $R(h_{\mathcal{D}}^*) < R^* + \sqrt{\frac{2}{N} \log \frac{2|H|}{\delta}}$, בהסתברות $(1 - \delta)$ לפחות.

אנו יכולים לפרש את האיבר השני כאיבר המודד את מורכבות מחלקת ההשערות - במקרה זה מורכבות נמדדת ע"ס גודל הקבוצה.

אבל חסם זה אינו תלוי בפילוג הדוגמאות, במדגם והוא ספציפי לאלגוריתם מזעור השגיאה האמפירית.

מקור עוצמתו הוא גם מקור חולשתו, שכן הוא מטפל במקרה הגרוע ביותר ואינו מנצלים את המבנה של בעיה נתונה. חסמים משופרים קיימים היום, אך קשים להוכחה במידה ניכרת. חסמים אלה הם מהצורה:

בהסתברות גדולה מ $1 - \delta$, אלגוריתם נתון (לא בהכרח מזעור שגיאה אמפירית) הבוחר השערה $h_{\mathcal{D}}^*$ מקיים

$$R(h_{\mathcal{D}}^*) < R^* + \Omega(h_{\mathcal{D}}^*, \mathcal{D}, H)$$

כאשר $\Omega(h_{\mathcal{D}}^*, \mathcal{D}, H)$ איבר מורכבות הדועך לאפס עבור $n \rightarrow \infty$.