פרק 12: יסודות בלמידה חישובית

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. נציין כי מדובר בתחום רחב אשר התפתח באופן משמעותי בשני העשורים האחרונים. אנו נסתפק בהצגת מספר תוצאות ומושגים יסודיים, וזאת עבור **בעיית הסיווג הבינארי בלבד**.

מקור בסיסי:

Mitchell, *Machine Learning*, 1997, Chapter 7.

לקריאה נוספת בנושא:

Kearns and Vazirani, An introduction to computational learning theory,
MIT Press, 1994.

L. Devroye, L. Györfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer, 1996.

G. Lugosi, Pattern classification and learning theory, http://www.econ.upf.es/~lugosi/lecturenotes.ps

12.1 מודל הלמידה הבסיסי

נזכור כי בבעיית הלמידה המודרכת אנו נדרשים "ללמוד" פונקציה $f_0 : X \to Y$ בעזרת אוסף דוגמאות $\{x^{(k)}, y^{(k)}\}_{k=1}^{n}$. המודל הבסיסי בו נעסוק כולל את המרכיבים הבאים:

א. פונקצית המטרה: פונקציה $f_0 : X \to Y$ ממרחב הקלט X, למרחב הפלט Y, אותה ברצוננו ללמוד. נזכיר כי $Y = \mathbb{R}$ לבעיית הרגרסיה, $Y = \{-1,+1\}$ לבעיית הסיווג הבינארי.

- בפרק זה נתעלם מרעש ונניח שהתיוג דטרמיניסטי. כלומר: לכל קלט x מתאימה יציאה יחידה, דהיינו $y = f_0(x)$.

ב. מודל בחירת הדוגמאות: דוגמאות הקלט נבחרות באופן בלתי תלוי ולפי פילוג הסתברות קבוע (אך לא בהכרח ידוע), כלומר $x^{(k)} \sim P_X,\ k = 1,\ldots,n$.
הדוגמאות מתויגות באופן מושלם לפי f_0, כלומר $y^{(k)} = f_0\left(x^{(k)}\right)$.

ג. מודל פרמטרי: אוסף H של פונקציות $H : X \to Y$, שמתוכו נבחר את הפונקציה $\hat{h}$ [או $\hat{f}$] אשר משערכת את פונקצית המטרה f_0. H תכונה כאן **מחלקת ההשערות**.

מדד הביצועים עבור השערה $h \in H$ כלשהי יהיה מהצורה:

$$L(h) := E\ell(h(x), f_0(x))$$

כאשר:

- $\ell(\hat{y}, y)$ הינה פונקצית מחיר מתאימה. למשל:
  $\ell(\hat{y}, y) = (\hat{y} - y)^2$ לבעיית הרגרסיה, $\ell(\hat{y}, y) = 1\{\hat{y} \neq y\}$ לבעיית הסיווג.
- התוחלת היא על (המשתנה המקרי) x, לפי הפילוג $x \sim P_X$. פילוג זה זהה לפילוג לפיו נבחרו הדוגמאות.
- לבעיית הסיווג הבינארי נקבל: $L(\hat{h}) = P\{\hat{h}(x) \neq f_0(x)\} \equiv P_e(\hat{h})$

מטרת תהליך הלימוד היא, אם כן, לבחור פונקציה $\hat{h} \in H$ (כתלות במדגם) אשר מביאה את מדד הביצועים $L(h)$ למינימום. הבעיה כמובן ש $L(h)$ אינו ניתן לחישוב מתוך מדגם סופי! ניתן רק להעריכו.

הערות:

א. חשוב להדגיש כי הדוגמאות $\left\{x^{(k)}\right\}$ נבחרות לפי אותו פילוג P_X המשמש בהגדרת מדד הביצועים. דבר זה יאפשר קבלת חסמים על קצב ושגיאת הלימוד שאינם תלויים ב- P_X.

ב. המודל הנ"ל מניח כי הקשר בין x ו- y הינו דטרמיניסטי. ניתן להרחיב את התוצאות להלן למקרה של קשר אקראי ("רועש"), כלומר להחליף את הפונקציה $y=f_0(x)$ בפילוג מותנה $p(y\,|\,x)$.

המודל ההסתברותי שהגדרנו מאפשר התייחסות כמותית לשאלות הבאות:

א. **דיוק הלמידה:** באיזה דיוק ניתן ללמוד את פונקצית המטרה $f_0(x)$ מתוך n דוגמאות?

ב. **קצב הלמידה:** כמה דוגמאות נדרשות כדי להשיג דיוק נתון?

12.2 מזעור מחיר האמפירי (Empirical Risk Minimization)

בהיעדר מידע לגבי הפילוג, ניתן להחליף את המזעור של קריטריון הביצועים $L(\hat{h})$ במזעור של פונקצית המחיר האמפירית (אותה אנו יכולים לחשב).

בהינתן המדגם $\{x^{(k)},y^{(k)}\}_{k=1}^n$, נבחר אם כן את ההשערה $\hat{h}_n$ באופן הבא:

$$\hat{h}_n \in \arg\min_{h\in H} \hat{L}_n(h)\,, \qquad \hat{L}_n(h) \doteq \frac{1}{n}\sum_{k=1}^{n} \ell(y^{(k)}, f(x^{(k)}))$$

לדוגמא:

א. עבור מחיר ריבועי, נקבל:
$$\hat{L}_n(h) \doteq \frac{1}{n}\sum_{k=1}^{n} (y^{(k)} - h(x^{(k)}))^2$$
זוהי פונקצית המחיר ששימשה אותנו בהקשר של רשת עצביות רב-שכבתית.

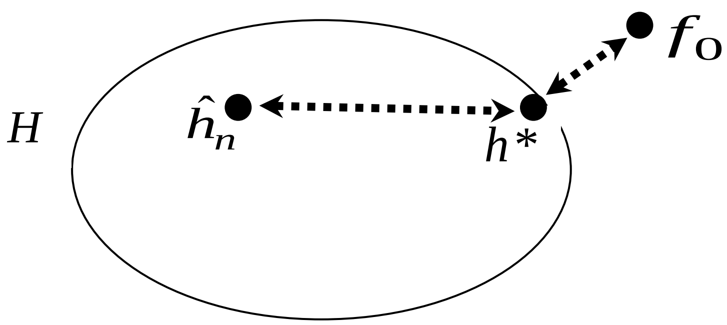
ב. עבור בעיית הסיווג, נקבל:
$$\hat{L}_n(h) \doteq \frac{1}{n}\sum_{k=1}^{n} 1\{y^{(k)} \neq h(x^{(k)})\}$$
זהו מספר השגיאות הממוצע של המסווג h על סט הלימוד.

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

<u>הערה:</u> למרות שאנו מניחים מזעור של השגיאה האמפירית לצורך הפיתוח התיאורטי, אין <u>לראות בכך המלצה לעשות זאת בפועל</u>! גישה זו יכולה להוביל להתאמת-יתר חמורה כאשר מרחב ההשערות גדול (ביחס למדגם הנתון).

<u>סוגי שגיאות: שגיאת ההכללה לעומת שגיאת הקירוב:</u>

נסמן $h^* \in \arg\min_{h \in H} L(h)$ – ההשערה האופטימאלית (שלצערנו אינה ניתנת לחישוב בפועל).



הערה: עבור פונ' שגיאות המקיימות את אי-שוויות המשולש (לדוג' שגיאת תיוג) אורך החיצים שקול למרחק ממש.

קריטריון הביצועים המתקבל עבור ניתן לרישום באופן הבא:

$$L(\hat{h}_n) = L(h^*) + [L(\hat{h}_n) - L(h^*)]$$

- האיבר הראשון הוא <u>שגיאת הקירוב</u> (בדומה למשתנה ההטיה, bias), אשר נובע מכך שאנו מגבילים את הפונקציה הנלמדת לקבוצת ההשערות H. הוא אינו תלוי במספר הדגימות n.
- האיבר השני הוא <u>שגיאת השערוך</u> (בדומה למשתנה השונות), ומבטא את השגיאה הנובעת <u>מסופיות המדגם</u> עקב כך שהפונקציה הנבחרת $\hat{h}_n$ אינה האופטימלית (מתוך H). זאת מכיוון שאנו מבצעים מינימיזציה של המחיר האמפירי $\hat{L}_n$ במקום של קריטריון הביצועים L.
- ככל שמחלקת ההשערות H עשירה (גדולה) יותר, אנו מצפים כי האיבר הראשון יקטן, והאיבר השני יגדל.

- עושר המודל (H) צריך להיות כזה המוצא איזון אופטימאלי בין שני איברים אלה (bias-variance tradeoff).

12.3 חסמים עבור מחלקת השערות סופית

נתמקד מעתה בבעיית הסיווג הבינארי: $Y=\{-1,+1\}$, $\ell(\hat{y},y)=1\{\hat{y}\neq y\}$. מטרתנו למצוא חסמים על קריטריון הביצועים $L(\hat{h}_n)$, כאשר $\hat{h}_n$ היא הפונקציה (ההשערה) המביאה למינימום את המחיר האמפירי $\hat{L}_n(h)$.

נשים לב כי במקרה זה המחיר האמפירי איננו אלא השגיאה האמפירית.

א. המקרה שבו $f_0\in H$: התוצאה הבאה עוסקת במקרה שבו פונקציית המטרה f_0 כלולה בקבוצת ההשערות H, כלומר $L^*\doteq\min_{h\in H}L(h)=0$.

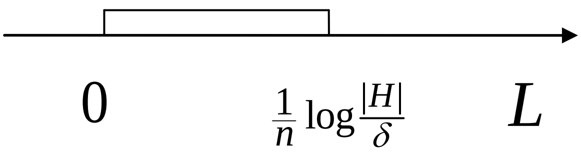
משפט 1: נניח כי: $|H|<\infty$, וכן $f_0\in H$ (כלומר $L^*=0$). אזי, השערה $\hat{h}_n$ הממזערת את השגיאה האמפירית מקיימת לכל $\varepsilon>0$

$$P\{L(\hat{h}_n)>\varepsilon\}\;<\;|H|e^{-\varepsilon n}$$

חשוב להדגיש: ההסתברות היא על פני כל המדגמים בגודל n. המשפט אומר שהחלק היחסי של מדגמים "רעים" הוא קטן. מדגם "רע" הוא כזה שבו למרות שהשגיאה האמיתית היא 0, השגיאה על המדגם גדולה מ- $\varepsilon>0$.

משפט 1 – ניסוח "מרווח סמך" (Confidence Interval): ע"י 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 "מרווח הסמך"):

- לכל $\delta>0$ מתקיים, בהסתברות של $(1-\delta)$ לפחות $L(\hat{f}_n)\;<\frac{1}{n}\log\frac{|H|}{\delta}$



<u>משפט 1 – ניסוח סיבוכיות המדגם</u> (Sample Complexity): 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 $L(\hat{h}_n) < \varepsilon$ בהסתברות $(1-\delta)$ לפחות.

<u>מספר מונחים בסיסיים בלמידה חישובית</u>: אלגוריתם כלשהו לבחירת $\hat{h}_n \in H$ שעבורו $P\{L(\hat{h}_n) > \varepsilon\} \to 0$ כאשר $n \to \infty$ (לכל $f_0 \in H$) נקרא אלגוריתם **PAC** – Probably Approximately Correct. 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 <u>כל קבוצת השערות סופית היא ברת למידה</u>). יתר על כן, בהמשך נראה (בעזרת אותו אלגוריתם) כי <u>כל קבוצת השערות בעלת מימד VC (גודל שיוגדר בהמשך) סופי היא ברת למידה</u>.

ב. המקרה שבו $f_0 \notin H$: נעבור כעת למקרה הכללי יותר שבו פונקצית המטרה f_0 אינה כלולה בהכרח בקבוצת ההשערות H, ולמעשה איננו מניחים הנחה כלשהי לגביה ("למידה אגנוסטית")[1]. במקרה זה $L^* \neq 0$.

<u>משפט 2 (Agnostic Learning):</u>

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

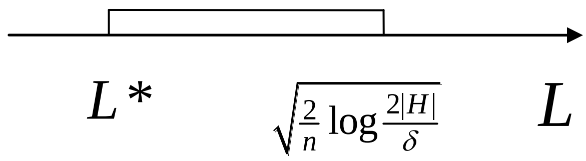
$$P\{L(\hat{h}_n) > L^* + \varepsilon\} < 2|H|e^{-\varepsilon^2 n/2}.$$

<u>הערות למשפט 2</u>:

- ניתן לראות כי חסם זה <u>חלש מהקודם</u>, כיוון שקצב הדעיכה המעריכי של הסתברות הטעות הינו ε^2.
- ניסוח "מרווח סמך": $L(\hat{h}_n) < L^* + \sqrt{\frac{2}{n}\log\frac{2|H|}{\delta}}$, בהסתברות $(1-\delta)$ לפחות.
  האיבר הראשון (L^*) מבטא את שגיאת הקירוב, והשני את שגיאת השערוך.

[1] Agnostic - a person who denies or doubts the possibility of ultimate knowledge in some area of study. From Greek, *ágnōtos* not known.

- ניסוח "סיבוכיות המדגם" : תרגיל.



לצורך הוכחת משפט 1, נגדיר :

<u>Version Space</u> : אוסף ההשערות ב- H <u>העקביות עם הנתונים</u>, קרי

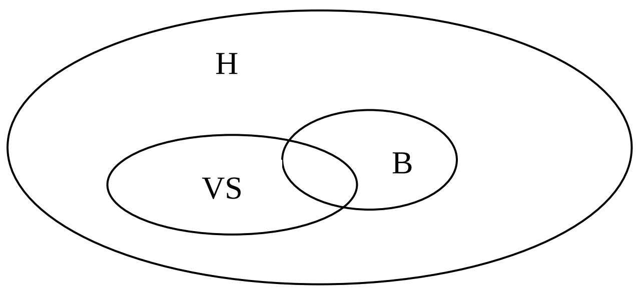
$$VS_H = \left\{ h_i \in H :\ \hat{L}_n\left(h_i\right) = 0,\ i = 1, 2, \cdots, |H| \right\}$$

עבור אלגוריתם הממזער את השגיאה האמפירית, ידוע כי $\hat{h}_n \in VS_H$.

אוסף ההשערות <u>הרעות</u> ב H - מוגדרות ע"י $B = \left\{ h_i \in H : L(h_i) > \varepsilon,\ i = 1, 2, \cdots, |H| \right\}$

<u>הערות</u> :

1. הקבוצה B <u>אינה אקראית</u> , כלומר אינה תלויה במדגם.
2. ככל שגודל המדגם גדל, הקבוצה VS_H (התלויה במדגם) מצטמקת והולכת.



אנו מעוניינים להעריך את ההסתברות שקימת <u>השערה רעה שהיא עקבית</u> $h \in \left(VS_H \cap B\right)$.

לפני ההוכחה ניזכר בחסם האיחוד (union bound)

$$P\left\{ \bigcup_{i=1}^{k} E_i \right\} \le \sum_{i=1}^{k} P\left(E_i\right) \le k \max_{1 \le i \le k} P\left(E_i\right).$$

שוויון קיים אם המאורעות זרים.

<u>הוכחת משפט 1:</u>

נתבונן בהשערה h_i <u>מסוימת</u>, מתקיים

$$P\left\{h_i\left(x^{(1)}\right)=y^{(1)}\wedge h_i\in B\right\}<1-\varepsilon$$

שימו לב, ההסתברות רק ביחס למשתנה האקראי $x^{(1)}$, כאשר אנו מגבילים את עצמנו ל-$h_i\in B$.

בגלל שהדגימות בת"ס

$$P\left\{h_i\in\left(VS_H\bigcap B\right)\right\}<\left(1-\varepsilon\right)^n$$

בעזרת חסם האיחוד (נגדיר מאורע $E_i\Leftrightarrow h_i\in\left(VS_H\bigcap B\right)$) נסיק ש

$$P\left\{\exists h_i\in\left(VS_H\bigcap B\right)\right\}<\left|B\right|\left(1-\varepsilon\right)^n$$

הגודל של הקבוצה B אינו ידוע, לכן נרשום

$$P\left\{\exists h_i\in\left(VS_H\bigcap B\right)\right\}\le\left|H\right|\left(1-\varepsilon\right)^n\le\left|H\right|e^{-\varepsilon n}.$$

האי-שוויון האחרון נובע מתוך $1-\varepsilon\le e^{-\varepsilon}$. מ.ש.ל.

<u>* הוכחת משפט 2:</u>

<u>תזכורת:</u> בקורס הסתברות לומדים את אי-שוויון צ'בישף,

$$P\{\left|X-E\left[X\right]\right|>\varepsilon\}\le\frac{\text{Var}\left[X\right]}{\varepsilon^2}$$

אנו נזדקק לחסם <u>הדוק יותר</u> במקרה שבו $X=\frac{1}{n}\sum_{k=1}^{n}Z^{(k)}$ ו-$\left\{Z^{(1)},\cdots,Z^{(k)}\right\}$ שווי-פילוג בת"ס.

חסם צ'בישף נותן במקרה זה

$$P\left\{\left|\tfrac{1}{n}\sum_{k=1}^{n}(Z^{(k)}-E(Z^{(k)}))\right|>\varepsilon\right\}\le\frac{\text{Var}\left[\sum_{k=1}^{n}Z^{(k)}\right]}{n^2\varepsilon^2}=\frac{\text{Var}\left[Z^{(1)}\right]}{n\varepsilon^2}$$

<u>אי-שוויון Hoeffding</u>: יהיו $\{Z^{(k)}\}_{k=1}^{n}$ משתנים אקראיים שווי-פילוג ובלתי תלויים סטטיסטית, המוגבלים בקטע סופי $a\le Z^{(k)}\le b$. אזי

$$P\left\{\left|\tfrac{1}{n}\sum_{k=1}^{n}(Z^{(k)}-E(Z^{(k)}))\right|>\varepsilon\right\}\le 2\exp\left(-\tfrac{2n\varepsilon^2}{(b-a)^2}\right)$$

היתרון המשמעותי של חסם זה ע"פ חסם צ'בישף הוא <u>קצבו המעריכי</u>.

הערה: חסם זה מתעלם משונות המשתנה האקראי. ניתן לקחתו בחשבון לצורך שיפור החסם.

מטרתנו לחסום את $P\{L(\hat{h}_n)-L^*>\varepsilon\}$. לשם כך נשתמש באי-השוויונות הבאים:

א. $L(\hat{h}_n)-L^*\le 2\max_{h\in H}|L(h)-\hat{L}_n(h)|$.

הוכחה: נניח (לשם פשטות) כי קיים $h^*\in H$ כך ש $L^*=L(h^*)$. אזי

$$
\begin{aligned}
L(\hat{h}_n)-L^* &= L(\hat{h}_n)-\hat{L}_n(\hat{h}_n)+\hat{L}_n(\hat{h}_n)-L(h^*)\\
&\le [L(\hat{h}_n)-\hat{L}_n(\hat{h}_n)]+[\hat{L}_n(h^*)-L(h^*)]\\
&\le 2\max_{h\in H}|L(h)-\hat{L}_n(h)|
\end{aligned}
$$

ב. $\hat{L}_n(h)=\frac{1}{n}\sum_{k=1}^{n}Z^{(k)}$, כאשר $Z^{(k)}=1\{h(x^{(k)})\ne y^{(k)}\}$, וכן $E(Z^{(k)})=L(h)$. מכאן,

ע"י הצבה בחסם Hoeffding עם $a=0,b=1,\varepsilon/2$ נקבל, עבור כל h:

$$P\{|L(h)-\hat{L}_n(h)|>\varepsilon/2\}\le 2|H|\exp(-n\varepsilon^2/2)$$

מכל האמור לעיל ומחסם האיחוד נובע:

$$
\begin{aligned}
P\{L(\hat{h}_n)-L^*>\varepsilon\} &\le P\{\max_{h\in H}|L(h)-\hat{L}_n(h)|>\varepsilon/2\}\\
&\le |H|\max_{h\in H}P\{|L(h)-\hat{L}_n(h)|>\varepsilon/2\}\}\\
&\le 2|H|\exp(-n\varepsilon^2/2)
\end{aligned}
$$

מ.ש.ל.

<u>מגבלות החסמים שפותחו:</u>

ראינו חסם מהצורה הבאה: בהסתברות $(1-\delta)$ לפחות, $L(\hat{h}_n)<L^*+\sqrt{\frac{2}{n}\log\frac{2|H|}{\delta}}$. אנו יכולים לפרש את האיבר השני כאיבר המודד את <u>מורכבות מחלקת ההשערות</u> – במקרה זה מורכבות נמדדת ע"ס גודל הקבוצה.

אבל חסם זה אינו תלוי ב: פילוג הדוגמאות, המדגם, ספציפי לאלגוריתם מזעור השגיאה האמפירית. <u>מקור עוצמתו הוא גם מקור חולשתו</u>, 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 (לאו בהכרח מזעור שגיאה אמפירית) הבוחר השערה $\hat{h}_n$ מקיים

$$L(\hat{h}_n)<L^*+\Omega\left(\hat{h}_n,D_n,H\right)$$

כאשר $\Omega\left(\hat{h}_n, D_n, H\right)$ איבר מורכבות הדועך לאפס עבור $n \to \infty$.

12.4 מימד VC

החסמים שתארנו עד כה הינם חסרי תועלת כאשר הלמידה מתבצעת עם קבוצת השערות H שהיא גדולה מאוד, או אף אין-סופית ($|H| = \infty$). במקרה זה עלינו להשתמש במדדים אחרים ל"גודל האפקטיבי" (סיבוכיות) של קבוצת ההשערות המגדירה את המודל.

מדוע לא מספר פרמטרים? מדד אופייני בסטטיסטיקה למורכבות המודל עבור מודל פרמטרי אינסופי הינו מספר הפרמטרים המגדירים את המודל. למשל, המחלקה (האינסופית) של מסווגים ליניאריים מוגדרת על ידי $m = d + 1$ פרמטרים. אם מדד זה מתאים לבעיית הסיווג?

כפי שנראה, סיבוכיות המודל אכן קשורה במספר הפרמטרים (m). אולם ניתן לראות בקלות כי מספר זה אינו מהווה מדד מהימן לעושר המודל. למשל:

(1) רשת עצבית ליניארית רב שכבתית בעלת מספר כלשהו של נוירונים זהה מבחינת יכולת התיאור שלה לפרספטרון ליניארי בודד.

(2) עבור $x \in \mathbb{R}$, למודל הליניארי $f(x) = \mathrm{sgn}(x + b)$ ולפונקצית הסינוס $f(x) = \mathrm{sgn}\{\sin(ax)\}$ פרמטר אחד. האם הם בעלי סיבוכיות שווה? מסתבר שהמסווג האחרון יכול לממש כל דיכוטומיה על קבוצת נקודות סופית!

קיימות תוצאות רבות העוסקות במדדי מורכבות משופרים למחלקות גדולות, ומתוכן נתאר תוצאה מייצגת אחת המשתמשת במדד למורכבות של מחלקת השערות המכונה ממד VC על שם החוקרים Vapnik & Chervonenkis. בסעיף זה נתאר גודל זה, ובסעיף הבא את החסם המבוטא בעזרתו.

הרעיון הבסיסי של מדד VC הוא לבחון את יכולת ההפרדה של המודל ביחס לקבוצת נקודות (מדגם) סופית.

בהגדרות הבאות נתייחס לנקודות הקלט של המדגם: $X_n = \left\{x^{(1)}, x^{(2)}, \cdots, x^{(n)}\right\} \subseteq X$.

<u>**דיכוטומיה**</u>: נתבונן בקבוצת נקודות נתונה $X_n = \left\{x^{(1)}, x^{(2)}, \cdots, x^{(n)}\right\}$. חלוקה של קבוצה זו לשתי קבוצות זרות נקראת דיכוטומיה (חלוקה בינארית). מספר הדיכוטומיות השונות הינו 2^n.

כל פונקצית סיווג $H \in h$ משרה דיכוטומיה על X_n, דהיינו החלוקה לשתי הקבוצות:

$$A_+ = \{x^{(k)} : f(x^{(k)}) = +1\},\ \ A_- = \{x^{(k)} : f(x^{(k)}) = -1\}.$$

העקרון: במקום לספור את מספר ההשערות במחלקה, קרי $|H|$, נספור את <u>מספר החלוקות</u> (דיכוטומיות) השונות המתקבלות ע"י כלל איברי H עבור המדגם הנתון. ברור שמספר הדיכוטומיות האפשרי הוא סופי וחסום ע"י 2^n, אך עבור מחלקה נתונה H הוא עשוי להיות קטן במידה ניכרת.

<u>מקדם הניתוץ (Shattering coefficient)</u>:

- נסמן ב- $S_H\{X_n\}$ את מספר החלוקות הבינאריות <u>השונות</u> שמשרות פונקציות הסיווג הכלולות במודל H על X_n, דהיינו $S_H\{X_n\} \doteq |\{(h(x^{(1)}),\ldots,h(x^{(n)})) : h \in H\}|$.

  נשים לב כי $(h(x^{(1)}),\ldots,h(x^{(n)})) \in \{-1,+1\}^n$.

- <u>מקדם הניתוץ</u> $S_H(n)$ של המודל H, עבור n נקודות, מתקבל עתה ע"י לקיחת המכסימום על פני כל הקבוצות האפשריות של n נקודות:

$$S_H(n) \doteq \max_{\{X_n\}\subset X} S_H\{X_n\}$$

שימו לב כי מקדם הניתוץ אינו תלוי במדגם הנתון X_n.

<u>הגדרת מימד VC</u> (Vapnik-Chervonenkis dimension)

- נאמר כי קבוצת הנקודות X_n <u>מנותצת לחלוטין</u> ע"י מחלקת השערות H אם H משרה עליה את כל הדיכוטומיות האפשריות, כלומר $S_H\{X_n\} = 2^n$.

- <u>מימד VC</u> של מחלקת השערות H הינו <u>המספר הגדול ביותר</u> של נקודות שניתן לנתץ לחלוטין ע"י H:

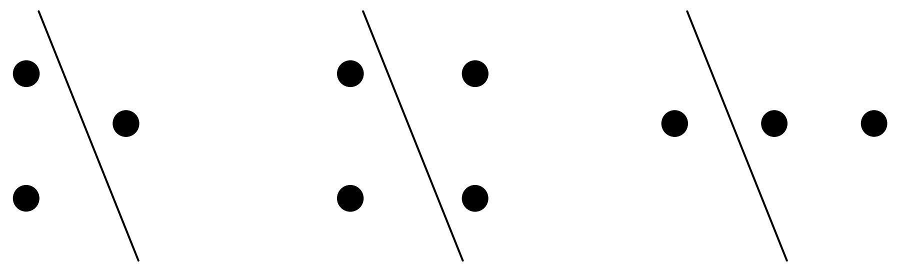
$$\mathrm{VCdim}(H) = \max\{n \ge 1 :\ S_H(n) = 2^n\}$$

יש להדגיש כי לצורך מימד VC אנו מאפשרים לבחור את הנקודות X_n שהן ה"<u>נוחות ביותר לניתוץ</u>". ביתר פירוט, $\mathrm{VCdim}(H) = V$ אם מתקיימים שני התנאים הבאים:

א. קיימת קבוצה של V נקודות שהיא מנותצת לחלוטין ע"י H.

ב. אין אף קבוצה של $V+1$ נקודות שהיא מנותצת לחלוטין ע"י F.

דוגמא: על מישורים ב $\mathbb{R}^2$.



עבור מערך הנקודות השמאלי כל הדיכוטומיות אפשריות, בעוד עבור המערך האמצעי 2 דיכוטומיות ה-XOR אינן אפשריות. לכן $\mathrm{VCdim}(H)=3$. נשים לב כי המערך הימני אינו מפריע לנו, שכן די למצוא קבוצה אחת בת 3 נקודות הניתנת לניתוץ.

מימד VC למודלים מסוימים:

נניח מרחב כניסה רציף: $x\in\mathbb{R}^d$. כזכור המודל H הינו אוסף של פונקציות סיווג בינאריות $h: X\rightarrow\{-1,1\}$.

- **מחלקה סופית:** $H=\{h_1,\ldots,h_K\}$, ברור כי $S_H(n)\le K$ (לכל n), ולכן הערך הגדול ביותר של n המקיים $2^n=K$ הוא $V_H\le\log_2 K$.
- **מסווגים ליניאריים:** ממד VC של מחלקת המסווגים הליניאריים הוא $d+1$ (ראו תרגיל הכיתה).
- **מלבנים:** נניח כי H הינה מחלקת המלבנים המקבילים לצירים. אזי $V_H=2d$.

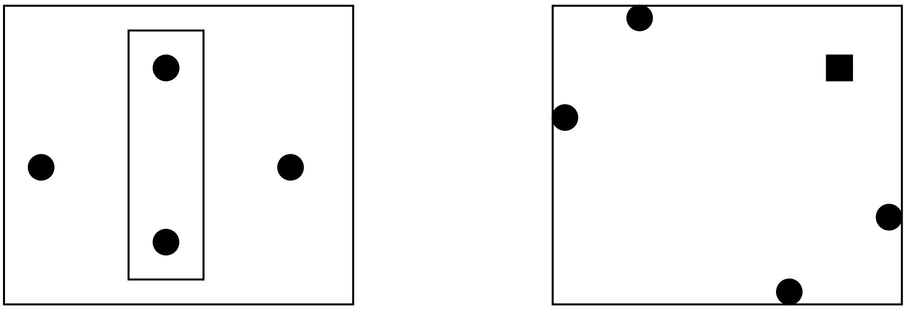
הוכחה (ראו איור): יש למצוא קבוצה (כלשהי) של $2d$ נקודות הניתנות לניתוץ. כמוכן יש להראות שאין אף קבוצה של $2d+1$ נקודות הניתנת לניתוץ.

ראשית, קל להראות כי הקבוצה הבאה של $2d$ נקודות אכן ניתנת לניתוץ:

$$(1,0,\cdots,0),\quad (0,1,\cdots,0),\cdots,\quad (0,0,\cdots,1)$$
$$(-1,0,\cdots,0),(0,-1,\cdots,0),\cdots,(0,0,\cdots,-1)$$

[הוכחה: נראה כי לכל תת-קבוצה של נקודות קיים מלבן המכיל תת-קבוצה זו בלבד. נבחר תת-קבוצה $B \subseteq A$. הנקודות ב- B מגדירות מלבן R מקביל לצירים ע"י לקיחת הערך המקסימלי והמינימלי של כל רכיב. נבנה מלבן R_ε מקביל לצירים הגדול במעט מ- R. מלבן זה מכיל אך ורק את הנקודות מ- B.]

נראה עכשיו שאין קבוצה בעלת $2d+1$ נקודות הניתנת לניתוץ. בכל קבוצה כזו יש נקודה אחת בעלת רכיב ראשון קטן ביותר ונקודה בעלת רכיב ראשון גדול ביותר, ובאופן דומה לגבי שאר הרכיבים. נשים לב שאם נקודה מסוימת היא בעלת רכיב ראשון ושני קטנים ביותר (למשל) הרי שתכונה זו רק תחזק את טענתנו. מקמירות המלבן, ברור שאין מלבן המכיל נקודות אלה ואף לא נקודה נוספת.



הטבלה הבאה מתארת את מימד VC המתקבל עבור מספר קבוצות נוספות של מודלים בעלי עניין:

1. חצאי מרחבים: $f(x)=\text{sign}(a^T x+b)$ (לכל a,b)	$V=d+1$
2. צרוף לינארי של m פונקציות <u>נתונות</u>: $f(x)=\text{sign}(\sum_{i=1}^m \alpha_i g_i(x))$ (עם $\{\alpha_i\}$ כלשהם)	$V \le m$
3. כדורים ב- $\mathbb{R}^d$: $f(x)=\text{sign}(\|\|x-a\|\|^2-b^2)$	$V=d+1$
4. אליפסואידים ב- $\mathbb{R}^d$: $f(x)=\text{sign}(x^T\Sigma^{-1}x-1)$ כאשר $\Sigma>0$ (מטריצה סימטרית חיובית מוגדרת כלשהי)	$V \le \frac{1}{2}d(d+1)+1$

<u>הערה</u>: מימד VC קשור בד"כ למספר הפרמטרים המתארים את המודל, אולם קיימים יוצאים מכלל זה. למשל, קבוצת הפונקציות ההרמוניות $F=\{\mathrm{sgn}(\sin(ax):a>0\}$ (על $X=\mathbb{R}$) היא בעלת פרמטר אחד בלבד, ומימד VC אינסופי.

12.5 חסמי ביצועים עבור מחלקת השערות אינסופית

ההוכחה במקרה זה קשה באופן משמעותי מאשר במקרה הסופי. התוצאה הסופית מתבטאת למעשה בהחלפת גודל המחלקה $|H|$ במקדם הניתוץ $S_H(2n)$.

<u>משפט 3</u> (ללא הוכחה) נסמן שוב ב-$\hat{h}_n$ השערה הממזערת את השגיאה האמפירית, אז לכל $\varepsilon>0$, מתקיים:

$$P\{L(\hat{h}_n)-L^*>\varepsilon\} \;<\; 4S_H(2n)e^{-\varepsilon^2 n/32}.$$

<u>הערות</u>

- השוואה למשפט 2 מראה כי $S_H(2n)$ תופס את מקומו של $|H|$ בחסם זה.
- החסם האחרון אינו שלם כיוון שלא ברורה תלותו של $S_H(n)$ ב-n. (למשל, אם $S_H(n)=2^n$, החסם חסר תועלת). לקבלת חסם מפורש, נשתמש בקשר הבא בין מקדם הניתוץ למימד VC.
-

<u>משפט</u> (Sauer's Lemma):

אם $V \doteq \mathrm{VCdim}(F)<\infty$, אזי $S_H(n)\le\sum_{i=0}^{V}\binom{n}{i}\le(n+1)^V$.

הערה: שימו לב שמקדם הניתוץ גדל פולינומיאלית עם n, ולכן הוא משמעותי בחסם הביצועים.

הצבת החסם האחרון במשפט 3 נותנת את התוצאה הבאה:

<u>משפט 4</u>

א. לכל $\varepsilon>0$: $P\{L(\hat{f}_n)>L^*+\varepsilon\} \;<\; 4(2n+1)^V e^{-\varepsilon^2 n/32}$.

ב. מכאן כי בהסתברות $(1-\delta)$ לפחות: $L(\hat{f}_n)\le L^*+\sqrt{\dfrac{V\log(2n+1)+\log(4/\delta)}{n/32}}$.

<u>הערות:</u>

- חסם זה מדגים כי גודל המדגם נמדד ביחס לממד VC.
- חסם זה בעל משמעות עקרונית רבה, אך בעל חשיבות מעשית מועטה, שכן הוא אפקטיבי רק לגודלי מדגם גבוהים ביותר. למשל, אם נדרוש שגיאה מסדר גודל של 0.01 עבור בעיה בעלת ממד VC 50 יידרשו בערך $50,000$ דוגמאות.
- הסיבות לפסימיות של החסם:
    1. אינו תלוי התפלגות.
    2. ממד VC אינו תלוי מדגם.
    3. אלגוריתם מאוד נאיבי (מזעור השגיאה האמפירית).
- קיימים שיפורים משמעותיים של חסמי VC, הפותרים חלק ניכר מבעיותיהם. אך עדיין, השגת חסמי ביצועים משמעותיים לצורך בניית מסווגים מעשיים, היא אתגר משמעותי וחשוב.