מבוא למערכות לומדות - הרצאה 2 - תורת ההכללה

29 ביוני 2015

בחמשת השבועות הראשונים של הקורס נלמד על **תורת ההכללה**. תורה זו תאפשר לנו להעריך את סיבוכיות המדגם של הרבה מחלקות. אנחנו נצטמצם **לקלסיפיקציה בינארית**. כלומר, לבעיות למידה בהן אנו רוצים ללמוד מיפוי הממפה כל קלט לאחת מבין שתי מחלקות, והשגיאה של המיפוי מוגדרת להיות פשוט הסיכוי שהתחזית של המיפוי שגויה. פורמאלית, אנו נביט בבעיות למידה $(X, Y, \mathcal{H}, l)$ שבהן $Y = \{0, 1\}$ ו־

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

נשים לב שבמקרה הזה השגיאה של $h : X \to Y$ ביחס להתפלגות $\mathcal{D}$ על $X \times Y$ היא פשוט, הסיכוי ש־$h(x) \neq y$ כאשר $(x, y) \sim \mathcal{D}$.

1 מודל PAC - תזכורת

בעיית למידה נקבעת ע"י רביעיה $(X, Y, \mathcal{H}, l)$, כאשר X היא קבוצה הנקראית **מרחב דוגמאות**, Y היא קבוצה הנקראית **מרחב פלטים**, $l : Y \times Y \to \mathbb{R}_+$ היא פונקצייה המקיימת $\forall y,\ l(y, y) = 0$ ונקראית **פונקצייה הפסד** ו־$\mathcal{H}$ היא אוסף של פונקציות מ־X ל־Y הנקרא **מחלקת היפותזות**.

הגדרה 1.1 אלגוריתם למידה הוא אלגוריתם המקבל בתור קלט מדגם אימון

$$(x_1, y_1), \ldots, (x_m, y_m)$$

ומחזיר פונקציה $h : X \to Y$.

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

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

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

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

נזכיר מספר מינוחים נוספים:

- נאמר ש־$\mathcal{D}$ **פרידה (או ממומשת או realizable) ע"י** $h^* : X \to Y$ אם, כאשר $(x,y) \sim \mathcal{D}$, מתקיים $y = h^*(x)$ בהסתברות 1.
- נאמר ש־$\mathcal{D}$ **פרידה ע"י** $\mathcal{H}$ אם $\mathcal{D}$ פרידה ע"י איזשהי $h^* \in \mathcal{H}$.
- אנו נסמן ב־$\mathcal{D}^m$ את ההתפלגות של מדגם בן m דוגמאות הנדגם לפי $\mathcal{D}$. כלומר, $\mathcal{D}^m$ היא ההתפלגות של מדגם

$$S = \{(x_{1,}, y_1), \ldots, (x_m, y_m)\} \in (X \times Y)^m$$

כאשר $(x_1, y_1), \ldots, (x_m, y_m)$ הם מ"מ ב"ת המתפלגים לפי $\mathcal{D}$.

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

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

סיבוכיות המדגם של $\mathcal{H}$ מוגדרת בתור סיבוכיות המדגם של האלגוריתם הטוב ביותר $m_{\mathcal{H}}(\epsilon, \delta) = \inf_{\mathcal{A}} m_{\mathcal{A}}(\epsilon, \delta)$. כמו כן, $\mathcal{H}$ **למידה** אם $m_{\mathcal{H}}(\epsilon, \delta) < \infty$ לכל $\epsilon, \delta > 0$.

הרבה פעמים נתייחס באופן פרטני למקרה בו ההתפלגות פרידה ע"י $\mathcal{H}$. לכן נגדיר:

הגדרה 1.4 (סיבוכיות המדגם במקרה הפריד) יהא $\mathcal{A}$ אלגוריתם למידה. עבור $\delta, \epsilon > 0$ נסמן ב־$m^r_{\mathcal{A}}(\epsilon, \delta)$ את המספר המינימאלי כך שלכל התפלגות $\mathcal{D}$ הפרידה ע"י $\mathcal{H}$ ו־$m \geq m^r_{\mathcal{A}}(\epsilon, \delta)$ מתקיים

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

באופן דומה מגדירים את המושגים **למידה במקרה הפריד, וסיבוכיות המדגם של** $\mathcal{H}$ **במקרה הפריד**.

2 אלגוריתמי ERM

בשעה טובה נלמד את אלגוריתם הלמידה הראשון שלנו. האלגוריתם הנ"ל, הנקרא Empirical Risk Minimizer (ERM), הוא מאד כללי (לכל מחלקה $\mathcal{H}$ קיים אלגוריתם כנ"ל), מאד אינטואיטיבי וכמעט תמיד (כמעט) אופטימאלי. הסיבה שהוא לא פותר את כל הבעיות בלמידה היא שבד"כ הוא לא יעיל. על מנת לתת מוטיבציה לאלגוריתם נניח שקיבלנו מדגם

$$S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$$

נניח, כמו כן, שבידנו כוח חישוב בלתי מוגבל. נזכור שאנחנו רוצים להחזיר היפותזה h עם שגיאה הקרובה ככל האפשר לשגיאה של ההיפותזה הכי טובה ב־$\mathcal{H}$. דרך אחת לעשות זאת

היא לעבור על כל ההיפותזות $h \in \mathcal{H}$ ולבחור את זו עם השגיאה, $L_{\mathcal{D}}(h)$, הכי קטנה. לצערנו לא ניתן לעשות זאת, אפילו אם יש בידנו כח חישוב בלתי מוגבל, מכיוון שאנחנו לא יודעים מהי ההתפלגות $\mathcal{D}$. עם זאת, המדגם שלנו מהווה "ייצוג" של ההתפלגות $\mathcal{D}$. לכן, אנלוג טבעי להצעה הנ"ל היא לנסות להעריך, באמצעות המדגם, את $L_{\mathcal{D}}(h)$, ולהחזיר ההיפותזה $h \in \mathcal{H}$ שהערכת השגיאה שלה נמוכה ככל האפשר.

2.1 השגיאה האמפירית

נגדיר את **השגיאה האמפירית** של h ביחס למדגם S להיות

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

השגיאה האמפירית מהווה הערכה של השגיאה האמיתית, $L_{\mathcal{D}}(h)$. הלמה הבאה מראה שעבור מדגם מספיק גדול $L_{\mathcal{D}}(h) \approx L_S(h)$.

למה 2.1 נסמן $B = \sup_{y,y' \in Y} l(\hat{y}, y)$. אזי

$$\Pr_{S \sim \mathcal{D}^m}\left(|L_S(h) - L_{\mathcal{D}}(h)| \geq \epsilon\right) < 2\exp\left(-\frac{2\epsilon^2 m}{B^2}\right)$$

הוכחת הלמה הינה פשוטה ונובעת מחסם הופדינג, יחד עם האבחנה שכאשר $S \sim \mathcal{D}^m$, $L_S(h)$ הוא סכום של m מ"מ ב"ת ש"ה שהתוחלת של כל אחד מהם היא $L_{\mathcal{D}}(h)$. נזכיר את חסם הופדינג ואח"כ נוכיח את הלמה.

משפט 2.2 (הופדינג) יהיו $Z_1, \ldots, Z_m \in [0, B]$ מ"מ ב"ת עם תוחלת μ. נסמן

$$\bar{Z} = \frac{1}{m}\sum_{i=1}^{m} Z_i$$

אזי

$$\Pr\left(\left|\bar{Z} - \mu\right| > \epsilon\right) \leq 2\exp\left(-\frac{2\epsilon^2 m}{B^2}\right)$$

הוכחה: (של למה 2.1) נשים לב שאם $(x, y) \sim \mathcal{D}$ אז $l(h(x), y)$ הינו מ"מ אי־שלילי החסום ע"י B ותוחלתו היא $L_{\mathcal{D}}(h)$. מכאן, אם

$$S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \sim \mathcal{D}^m$$

אז

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

הוא ממוצע של מ"מ ב"ת ש"ה, שתוחלתם היא $L_{\mathcal{D}}(h)$ והם מקבלים ערכים ב־$[0,B]$. מכאן, לפי חסם הופדינג (כאשר $L_S(h)$ מחליף את $\bar{Z}$ ו־$L_{\mathcal{D}}(h)$ מחליף את μ) מתקיים

$$\Pr\left(|L_S(h)-L_{\mathcal{D}}(h)|>\epsilon\right)\leq 2\exp\left(-\frac{2\epsilon^2 m}{B^2}\right)$$

■

2.2 אלגוריתם ה־ERM

הגדרה 2.3 אלגוריתם למידה יקרא **ERM** אם לכל מדגם S הוא מחזיר $h\in\mathcal{H}$ המקיימת $L_S(h)=\inf_{h'\in\mathcal{H}}L_S(h')$.

הערה 2.4 1. לכל מחלקה קיים אלגוריתם ERM. עם זאת, הוא לא בהכרח יחיד - יכולה להיות יותר מהיפותזה אחת ב־$\mathcal{H}$ הממזערת את השגיאה האמפירית.

2. אלגוריתמי ERM משחקים תפקיד חושב בתורת ההכללה. התורה שנפתח תראה שלכל בעיית קלסיפיקציה בינארית, סיבוכיות המדגם של אלגוריתם ה־ERM קרובה לאופטימלית.

3 סיבוכיות המדגם של מחלקות סופיות

המשפט הראשון שנוכיח ייתן חסם עליון על סיבוכיות המדגם של אלגוריתמי ERM עבור מחלקות סופיות. כפועל יוצא אותו חסם חוסם את סיבוכיות המדגם של $\mathcal{H}$

משפט 3.1 נסמן $B=\sup_{y,y'\in Y}l(\hat{y},y)$. לכל אלגוריתם ERM $\mathcal{A}$ מתקיים

$$\begin{aligned} m_{\mathcal{A}}(\epsilon,\delta) &\leq \left(\frac{B}{\epsilon}\right)^2\cdot 2\log\left(\frac{2|\mathcal{H}|}{\delta}\right) \\ m^r_{\mathcal{A}}(\epsilon,\delta) &\leq \frac{B}{\epsilon}\cdot\log\left(\frac{|\mathcal{H}|}{\delta}\right) \end{aligned}$$

הוכחה: נוכיח רק את החסם עבור סיבוכיות המדגם במקרה הכללי (המקרה הפריד מושאר כתרגיל). יהא $\mathcal{A}$ אלגוריתם ERM ויהא

$$S=\{(x_1,y_1),\ldots,(x_m,y_m)\}\sim\mathcal{D}^m$$

כאשר $m\geq\left(\frac{B}{\epsilon}\right)^2\cdot 2\log\left(\frac{2|\mathcal{H}|}{\delta}\right)$. צריך להראות שבהסתברות $\geq 1-\delta$ על פני בחירת המדגם מתקיים

$$L_{\mathcal{D}}(\mathcal{A}(S))\leq L_{\mathcal{D}}(\mathcal{H})+\epsilon$$

רעיון ההוכחה וההוכחה עצמה פשוטים מאד - אנו נראה שבהסתברות לפחות $1-\delta$, לכל ההיפותזות ב־$\mathcal{H}$, השגיאה האמפירית מקרבת שאת השגיאה האמיתית עד כדי $\frac{\epsilon}{2}$. כלומר

$$\forall h \in \mathcal{H}, \;\; |L_S(h) - L_{\mathcal{D}}(h)| \le \frac{\epsilon}{2} \tag{1}$$

זה יספיק, שכן, במקרה הזה, מכיוון שאלגוריתם ERM מחזיר היפותזה עם שגיאה אמפירית מינימאלית, השגיאה האמיתית תהיה אף היא קרובה למינימאלית. קונקרטית, מתקיים

$$L_{\mathcal{D}}(\mathcal{A}(S)) \le L_S(\mathcal{A}(S)) + \frac{\epsilon}{2} = \inf_{h'\in\mathcal{H}} L_S(h') + \frac{\epsilon}{2} \le \inf_{h'\in\mathcal{H}} L_{\mathcal{D}}(h') + \frac{\epsilon}{2} + \frac{\epsilon}{2} = L_{\mathcal{D}}(\mathcal{H}) + \epsilon$$

נותר, אם כן, להראות שבהסתברות $\le 1-\delta$ מתקיים שוויון (1). נסמן ב־U_h את המאורע ש־$|L_S(h) - L_{\mathcal{D}}(h)| > \frac{\epsilon}{2}$. נסמן, כמו כן, ב־$U = \cup_{h\in\mathcal{H}} U_h$ את המאורע ששוויון (1) לא מתקיים . די להראות ש־$\Pr_S(U) < \delta$.

מלמה 2.1 מתקיים

$$\Pr_S(U_h) = \Pr_S\left(|L_S(h) - L_{\mathcal{D}}(h)| > \frac{\epsilon}{2}\right) \le 2\exp\left(-\frac{\epsilon^2 m}{2B^2}\right)$$

מכאן, לפי חסם האיחוד, ומפני ש־$m \ge \left(\frac{B}{\epsilon}\right)^2 \cdot 2\log\left(\frac{2|\mathcal{H}|}{\delta}\right)$, נובע

$$\begin{aligned}
\Pr_S(U) &= \Pr_S(\cup_{h\in\mathcal{H}} U_h) \\
&\le \sum_{h\in\mathcal{H}} \Pr_S(U_h) \\
&\le 2\sum_{h\in\mathcal{H}} \exp\left(-\frac{\epsilon^2 m}{2B^2}\right) \\
&\le 2|\mathcal{H}| \exp\left(-\frac{\epsilon^2 m}{2B^2}\right) \\
&\le 2|\mathcal{H}| \exp\left(-\frac{\epsilon^2 \left(\frac{B}{\epsilon}\right)^2 \cdot 2\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}{2B^2}\right) \\
&= 2|\mathcal{H}| \exp\left(-\log\left(\frac{2|\mathcal{H}|}{\delta}\right)\right) = \delta
\end{aligned}$$

■

אפליקציה: סיבוכיות המדגם של מחלקות בנות d פרמטרים

כבר ממשפט 3.1 הפשוט אנחנו יכולים לקבל מסקנות מעניינות. תהא $(X, Y, \mathcal{H}, l_{0-1})$ בעיית למידה. נניח שניתן לתאר את ההיפותזות ב־$\mathcal{H}$ ע"י d פרמטרים, שכל אחד מהם מתואר ע"י q ביטים (פורמאלית קיימת פונקציה $R : (\{0,1\}^q)^d \to \mathcal{H}$ שהיא על). במקרה הזה, $|\mathcal{H}| \le 2^{qd}$. לכן, ממשפט 3.1, סיבוכיות המדגם של אלגוריתמי ERM חסומה ע"י

$$\frac{2qd + 2\log\left(\frac{2}{\delta}\right)}{\epsilon^2}$$

נשים לב שאם מתעלמים מהגורם $2\log\left(\frac{2}{\delta}\right)$ (שבד"כ יהיה קטן הרבה יותר מ-$2qd$) וקובעים את ϵ, נקבל שהחסם על סיבוכיות המדגם פרופורציונאלי למספר הפרמטרים כפול גודל הייצוג של כל פרמטר. נעיר שיש לא מעט מחלקות מעניינות מהצורה הנ"ל. למשל:

- **מחלקת ההיפותזות שניתן לממש ע"י קובץ exe. בן d ביטים.** בסופו של יום, נרצה לממש את ההיפותזה $h : X \to Y$ שנלמד. כמובן, לא נרצה שהקובץ שמפעיל את ההיפותזה יהיה גדול מידי. לכן, טבעי להסתכל על המחלקה הנ"ל.

- **חצאי המרחבים המוגדרים ע"י floating points.** אחת המחלקות הבסיסיות בלמידה היא המחלקה של חצאי מרחבים. כזכור, מחלקה זו מכילה את כל הפונקציות $h : \mathbb{R}^d \to \{\pm 1\}$ מהצורה

$$h(x) = \text{sign}\left(a_1 x_1 + \ldots + a_d x_d + b\right)$$

בפרט, כל היפותזה במחלקה מתוארת ע"י $(d+1)$ המספרים $a_1, \ldots, a_d, b$. בפועל, הרבה פעמים נייצג כל מספר ע"י 32 ביטים. במקרה הנ"ל, כל היפותזה במחלקה ניתנת לתיאור ע"י $(d+1)$ פרמטרים שכל אחד מתואר ע"י $q = 32$ ביטים.

4 מימד VC והמשפט היסודי

משפט 3.1 נותן לנו חסם עליון על סיבוכיות המדגם של מחלקות סופיות. החסם הנ"ל לא תמיד הדוק. בפרט, במקרה ש-$\mathcal{H}$ אינסופית, החסם הנ"ל חסר משמעות. התורה שנלמד השיעור ובשיעור הבא תאפשר לנו להבין טוב יותר מהי סיבוכיות המדגם של הרבה בעיות. כאמור, **אנו נצטמצם לבעיות קלסיפיקציה בינארית**. בפרט, עד להודעה חדשה, פונקצייה ההפסד שלנו תהיה l_{0-1} ומרחב הפלטים יהיה $Y = \{0,1\}$.

4.1 מימד VC

מימד ה-VC (על שם Vapnik and Chervonenkis) מתאים מספר טבעי, $\text{VC}(\mathcal{H})$, לכל מחלקה $\mathcal{H}$. כפי שיראה המשפט היסודי, המספר הנ"ל מאפיין את סיבוכיות המדגם של $\mathcal{H}$.

הגדרת המימד נסובה סביב המושג של **תת קבוצה מנותצת**. נקבע מרחב מדגם X. עבור $h : X \to \{0,1\}$ ותת קבוצה $A \subset X$ נסמן ב-$h|_A$ את **הצמצום של h ל-A**. כלומר את הפונקציה שתחומה A ומקיימת $\forall x \in A,\ h|_A(x) = h(x)$. נסמן, כמו כן,

$$\mathcal{H}|_A = \{h|_A : h \in \mathcal{H}\}$$

הגדרה 4.1 תת קבוצה $A \subset X$ **מנותצת ע"י** $\mathcal{H}$ אם מתקיים $\mathcal{H}|_A = \{0,1\}^A$.

דוגמא: אם נגדיר $\mathcal{H} = \{h_1, h_2\}$ כאשר $h_1, h_2 : \{a,b\} \to \{0,1\}$ הן הפונקציות המוגדרות ע"י

$$h_1(a) = 0,\ h_1(b) = 1,\ h_2(a) = 0,\ h_2(b) = 0,$$

אז הקבוצות המנותצות הן $\{\},\{a\}$ (הוכיחו!)

נניח ש־A מנותצת. אינטואיטיבית, אם אנחנו רוצים ללמוד מיפוי $h\in\mathcal{H}$, ידיעת הערכים של h על תת קבוצה $A'\subset A$ לא תאפשר לנו להסיק את הערכים של h על $A\setminus A'$. לכן, על מנת ללמוד עלינו לראות את "רוב" הערכים ב־A. בפרט, סיבוכיות המדגם חסומה מלמטה ע"י $|A|$. מהטיעון הנ"ל, הגודל המקסימאלי של תת קבוצה מנותצת $A\subset X$ מהווה אף הוא חסם תחתון על סיבוכיות המדגם. הגודל הנ"ל הוא המימד VC של $\mathcal{H}$. באופן אולי קצת מפתיע, המשפט היסודי יראה שהחסם התחתון הנ"ל הוא אף חסם עליון.

הגדרה 4.2 מימד VC, $\mathrm{VC}(\mathcal{H})$, של $\mathcal{H}$ הינו הגודל המקסימאלי של קבוצה מנותצת. כלומר,

$$\mathrm{VC}(\mathcal{H})=\max\{|A|\ :\ A\text{ is shattered by }\mathcal{H}\}$$

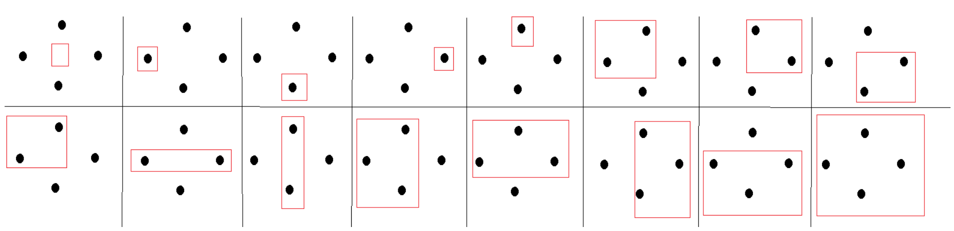
דוגמאות

- **עבור המחלקה** $\mathcal{H}$ מהדוגמא הקודמת מתקיים $\mathrm{VC}(\mathcal{H})=1$
- **מחלקת כל הפונקציות** $\mathcal{H}=\{0,1\}^X$. במקרה הזה, $\mathrm{VC}(\mathcal{H})=|X|$
- **מלבנים מקבילים לצירים.** כאן, $\mathcal{H}$ מכילה את כל הפונקציות $h:\mathbb{R}^2\to\{0,1\}$ כך שקיימים $a_1<a_2,b_1<b_2$ עבורם

$$h(x,y)=\begin{cases}1 & x\in[a_1,a_2]\text{ and }y\in[b_1,b_2]\\ 0 & \text{otherwise}\end{cases}$$

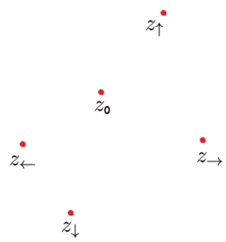
מתברר ש־$\mathrm{VC}(\mathcal{H})=4$. על מנת להראות את השיוויון הנ"ל, צריך להצביע על תת קבוצה $A\subset\mathbb{R}^2$ בת 4 איברים המנותצת ע"י $\mathcal{H}$, ובנוסף, להראות שאין תת קבוצה בת 5 איברים המנותצת ע"י $\mathcal{H}$ (מדוע אין צורך להראות שאין תת קבוצה מנותצת בגודל 6?). אכן:

- כפי שמעיד הציור הבא, הקבוצה $A=\{(-1,0),(1,0),(0,-1),(0,1)\}$ מנותצת.



- נניח בשלילה ש־$A\subset\mathbb{R}^2$ היא תת קבוצה מנותצת בת 5 איברים. יהא $z_{\leftarrow}\in A$ האיבר השמאלי ביותר (אם יש יותר מאחד, נבחר אחד מהם שרירותית). באופן דומה, יהיו $z_{\rightarrow},z_{\uparrow},z_{\downarrow}$ האיבר התחתון, העליון והימני. יהא $z_0\in A\setminus$

$\{z_{\rightarrow}, z_{\uparrow}, z_{\downarrow}, z_{\leftarrow}\}$. כפי שמעיד הציור הבא, לא קשה להראות שהפונקציה $f : A \rightarrow \{0,1\}$ המוגדרת ע"י $f(z) = \begin{cases} 0 & z = z_0 \\ 1 & \text{otherwise} \end{cases}$ לא נמצאת ב־$\mathcal{H}|_A$.



הערה 4.3

- היכולת לחשב מימד VC חשובה על מנת להפעיל את התיאוריה שנלמד על בעיות למידה קונקרטיות (כלומר, על מחלקות $\mathcal{H}$ ספציפיות). בתרגול ובתרגיל שאחרי פסח תלמדו מספר שיטות לחישוב מימד VC.
- כלל אצבע לא פורמאלי שעובד הרבה פעמים אומר ש־$\mathrm{VC}(\mathcal{H})$ שווה (או לפחות פרופורציונאלי) למספר הפרמטרים הנדרשים על מנת להגדיר היפותזה ב־$\mathcal{H}$. למשל, על מנת להגדיר מלבן אנו נזקקים ל־4 מספרים, ואכן המימד VC של המחלקה המתאימה הוא 4.

4.2 המשפט היסודי (Vapnik and Chervonenkis, 1971)

המשפט היסודי של למידה חישובית מראה שסיבוכיות המדגם של $\mathcal{H}$ פרופורציונאלית ל־$\mathrm{VC}(\mathcal{H})$.

משפט 4.4 (המשפט היסודי) קיימים קבועים $C_1, C_2 > 0$ כך שלכל $\mathcal{H} \subset \{0,1\}^X$ ולכל ERM $\mathcal{A}$ מתקיים

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

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

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

- בתרגול תראו שמימד VC של חצאי מרחבים ב־$\mathbb{R}^d$ הוא $(d+1)$. מהמשפט היסודי נובע, אם כן, שסיבוכיות המדגם של המחלקה המתאימה הינה $\Theta\left(\frac{d+\log\left(\frac{1}{\delta}\right)}{\epsilon^2}\right)$. לשם השוואה, החסם על מחלקות סופיות (משפט 3.1) היה חסר משמעות עבור המחלקה הנ"ל, ואפילו כשהנחנו שכל אחד מהפרמטרים המגדיר את חצי המרחב מתואר ע"י q ביטים הסקנו שסיבוכיות המדגם חסומה רק ע"י $O\left(\frac{dq+\log\left(\frac{1}{\delta}\right)}{\epsilon^2}\right)$.
- עבור המחלקה של מלבנים מקבילים לצירים, שוב החסם על מחלקות סופיות חסר משמעות, בעוד המשפט היסודי מראה שסיבוכיות המדגם היא $\Theta\left(\frac{\log\left(\frac{1}{\delta}\right)}{\epsilon^2}\right)$.

4.3 מבנה ההוכחה

השיעור הבא יוקדש להוכחת המשפט. הוכחת החסם התחתון מתבססת על הטיעון שהופיע לפני ההגדרה של מימד VC. לגבי החסם העליון, מבנה ההוכחה יהיה דומה למבנה ההוכחה של החסם על מחלקות סופיות (משפט 3.1). כלומר, אנו נראה שאם מספר הדוגמאות גדול מהחסם, אז, בהסתברות $\geq 1-\delta$, השגיאה האמפירית של כל ההיפותזות ב־$\mathcal{H}$ קרובה לשגיאה האמיתית עד כדי $\frac{\epsilon}{2}$. כלומר, מתקיים

$$\forall h \in \mathcal{H},\ \ |L_S(h) - L_{\mathcal{D}}(h)| < \frac{\epsilon}{2} \tag{2}$$

מכאן, כמו בהוכחת משפט 3.1, ינבע שהפלט של כל אלגוריתם ERM יקיים $L_{\mathcal{D}}(h) \leq L_{\mathcal{D}}(\mathcal{H}) + \epsilon$.

על מנת להראות שמתקיים אי־שוויון (2) לא נוכל להשתמש, כמו בהוכחת משפט 3.1, בחסם האיחוד יחד עם למה 2.1. 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

$$|\mathcal{H}|_A| \leq \sum_{i=0}^{d} \binom{d}{i} \leq (|A|+1)^d$$

נשים לב שכאשר כאשר A גדולה $2^{|A|} \ll |A|^d$. 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 (2) בהסתברות $1-\delta$.

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