

**פרויקט מסכם לתואר בוגר במדעים (B.Sc)
במתמטיקה שימושית**

התיאוריה המתמטית של למידה

רזי פח'ראלדין

Mathematical Learning Theory

Razi Fachareldeen

Advisor:

Assoc. Prof Haggai
Katriel

מנחה:

פרופ"ח חגי כתריאל

Karmiel

2019

כרמיאל

תוכן עניינים

3	1	מבוא
5	2	כלים מתמטיים
5	2.1	מרחבים מטריים
5	2.1.1	הגדרות ומושגים בסיסיים
7	2.1.2	סדרות במרחב מטרי
9	2.1.3	רציפות של פונקציה בין מרחבים מטריים
9	2.1.4	מרחבי פונקציות כמרחבים מטריים
10	2.1.5	שלמות של מרחבים מטריים
13	2.1.6	קומפקטיות במרחבים מטריים
20	2.2	כלים מתורת ההסתברות
20	2.2.1	הגדרות בסיסיות
21	2.2.2	אי־שוויונות הסתברותיים
27	3	מבוא ללמידה מונחית
28	3.1	דוגמאות הלמידה והבדיקה
28	3.2	למידה כבעיית חיפוש במרחב פונקציות $\mathcal{H}$
29	3.3	דוגמה - רגרסיה ליניארית
30	3.4	ריגולריזציה - קנס על הנורמה
32	4	למידה סטטיסטית
32	4.1	הגדרה פורמלית של למידה
33	4.1.1	פונקציית הרגרסיה

34		חסם לשגיאה עבור H סופי	4.2
37		המקרה הכללי	4.3
42		שימושים	5
42		רגרסיה ליניארית	5.1
44		סימולציות	5.1.1
48		רשת נוירונים	5.2
51		סיכום	6
53		ביבליוגרפיה	7
55		נספח: תכניות Matlab לסימולציות בפרק 5	8

פרק 1

מבוא

למידת מכונה היא משפחה של אלגוריתמים שבעזרתם אפשר לגרום למחשב ללמוד לבצע משימות מורכבות בעזרת למידה מנתונים, בלי לכתוב קוד שמתאר במדויק את השלבים השונים כדי לבצע את המשימה. הדוגמה הקלאסית שבעזרתה אפשר להבין את החשיבות והכוח של למידת מכונה היא המשימה של סיווג תמונות, נניח שנרצה לכתוב תוכנית מחשב שבה אנו מקבלים כקלט תמונה ואנו רוצים לדעת אם בתמונה נמצא חתול. כמובן תמונה מיוצגת על ידי מטריצה של פיקסלים ולכן כתיבת קוד שלוקח בחשבון את כל הקומבינציות האפשריות של הפיקסלים שמתארות חתול היא בלתי אפשרית. לעומת זאת נוכל לאסוף כמות מספקת של תמונות שתויגו (חתול או לא) על ידי בני אדם ולהעביר את הנתונים האלה לאלגוריתם למידה שילמד מנתונים אלה ויחזיר לנו מודל שיודע לסווג תמונות ברמת דיוק גבוהה. אלגוריתם למידה כזה, שמקבל נתונים מתויגים כקלט שייך למחלקה של אלגוריתמי למידה שנקראים למידה מונחית. ישנן מחלקות אחרות של אלגוריתמים כמו למידה בלתי מונחית/למידת חיזוק שאינם במסגרת עבודה זו. כמו בדוגמא שלמעלה, למידת מכונה מאוד שימושית למקרים שבהם קשה מאוד או בלתי אפשרי לתאר איך פותרים את המשימה בעזרת רצף של חוקים.

למידת מכונה אינה תחום חדש: תוכנית המחשב הראשונה שבה השתמשו באלגוריתם למידה [1] הייתה ב-1952 שבה כתבו אלגוריתם שגורם למחשב ללמוד לשחק דמקה, אבל בעשור האחרון התחום הזה נהיה הרבה יותר פופולרי משתי סיבות עיקריות:

- זמינות של כמויות אדירות של נתונים שהולכת וגדלה.

- כוח מחשוב שגם הוא גדל בצורה משמעותית.

השילוב של שני הגורמים הנ"ל מאפשר בניית מודלים הרבה יותר מורכבים ומדויקים ממה שהיה אפשרי לפי כמה שנים, ולכן היום אפשר לראות שימושים של למידת מכונה בתחומים רבים כמו:

- רפואה: זיהוי וסיווג מחלות שונות בעזרת נתונים שנאספו מדימות רפואי, למשל חיזוי אלצהיימר בשלבים מוקדמים [2], סגמנטציה של גידול במוח [3], סיווג של רקמות סרטניות שנלקחו מחולים עם סרטן עור [4] ועוד.

- פינטק [5]: זיהוי הונאות פיננסיות, חיזוי מחירים של מניות, ניהול סיכונים.

- רובוטיקה: רכבים/רחפנים אוטונומיים.

- אפליקציות שונות: מערכת התרגום של גוגל [6], אפליקציות של זיהוי קול, זיהוי פנים, השלמה אוטומטית של משפטים ועוד..

בעבודה הזו נתמקד בתיאוריה המתמטית של למידה סטטיסטית. בעזרת תיאוריה זו אפשר להראות שלמידה היא אכן אפשרית וגם להבין באופן כמותי איך פרמטרים שונים של האלגוריתמים משפיעים על תהליך הלמידה ועל רמת הדיק של המודל שהם מחזירים. התוצאה הראשונה והחשובה ביותר בתיאוריה הזו פותחה על ידי ואפניק וצרבונינקס (VC) [7] ומראה שלמידה אפשרית כאשר מדובר במשימות של קלסיפקציה (סיווג), כלומר כאשר המודל מתבקש להחליט לאיזה קטגוריה אובייקט מסוים שייך מתוך מספר סופי של קטגוריות אפשריות. למרות שבעזרת התיאוריה הזו אפשר להבין את ההנחות הבסיסיות של למידה ואת האלמנטים המרכזיים שמשפיעים על תהליך הלמידה, בעבודה זו בחרנו להציג את התיאוריה עבור המקרה שבו האלגוריתם מתבקש להחזיר פונקציה רציפה (בעיות רגרסיה). למשל, חיזוי מחיר של מניה/רווח עתידי ממוצר מסוים על סמך מדדים כמותיים לגבי המניה או המוצר. תיאוריה זו היא מעט יותר מורכבת ומצריכה יותר כלים מתמטיים מהתיאוריה הנוגעת לבעיות קלסיפקציה, כך שבד"כ קורסי מבוא לא מלמדים אותה.

בפרק 2 אנו מציגים את הרקע המתמטי הדרוש להבנת התיאוריה שמכיל נושאים מאנליזה פונקציונלית ותורת ההסתברות. בפרק 3 נציג מבוא כללי ללמידה מונחית שיעזור לנו להבין את הקשר של המשפטים המתמטיים לתהליך הלמידה. בפרק 4 נציג את התיאוריה של למידה סטטיסטית שברובה נלקחה ממאמר על היסודות המתמטיים של הלמידה, של Cucker ו-Smale [8], מאמר שדורש רקע מתמטי נרחב יחסית. בפרק 5 נממש את התיאוריה על מודל למידה חשוב של רגרסיה ליניארית ונציג את התובנות המעשיות שניתן להשיג מהתיאוריה המתמטית.

פרק 2

כלים מתמטיים

בפרק זה נציג מושגים ומשפטים מתמטיים שיהיו חשובים להבנת התיאוריה בהמשך.

2.1 מרחבים מטריים

2.1.1 הגדרות ומושגים בסיסיים

הגדרה 1 תהי X קבוצה כלשהי, פונקציה $d : X \times X \mapsto \mathbb{R}$ תיקרא מטריקה אם היא מקיימת:

$$1. d(x, y) \geq 0, d(x, y) = 0 \iff x = y$$

$$2. d(x, y) = d(y, x)$$

$$3. d(x, z) \leq d(x, y) + d(y, z), \forall x, y, z \in X$$

דוגמה 2.1.1 יהי $X = \mathbb{R}^n$ אזי

$$d_2(x, y) = \|x - y\|_2 = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

היא מטריקה.

הגדרה 2 תהי X קבוצה כלשהי ו d מטריקה. אזי נקרא לזוג (X, d) מרחב מטרי

הגדרה 3 יהי (X, d) מרחב מטרי, נגדיר כדור פתוח ברדיוס x סביב r :

$$S_r(x) = \{y \in X \mid d(y, x) < r\}$$

כאשר r מספר ממשי חיובי.

הגדרה 4 יהי (X, d) מרחב מטרי ו $A \subset X$. נגיד ש A קבוצה חסומה אם קיימים $x \in A$ ו $r > 0$ כך ש:

$$A \subset S_r(x)$$

הגדרה 5 תהי $A \subset X$. $x \in A$ תיקרא נקודה פנימית של A אם קיים $r > 0$ כך ש: $S_r(x) \subset A$

הגדרה 6 קבוצה X תיקרא קבוצה פתוחה אם כל הנקודות בה הן פנימיות.

טענה 2.1.1 איחוד כלשהו של קבוצות פתוחות הוא פתוח

הוכחה:

יהי $\{U_\alpha\}_{\alpha \in I}$ אוסף אינסופי כלשהו של קבוצות פתוחות U_α , נרצה להוכיח כי

$$\bigcup_{\alpha \in I} U_\alpha$$

מהווה קבוצה פתוחה.

יהי $x \in \bigcup_{\alpha \in I} U_\alpha$, אזי קיים $\alpha^* \in I$ כך ש $x \in U_{\alpha^*}$ ומכיוון ש U_{α^*} היא קבוצה פתוחה, קיים $r > 0$ כך ש:

$$S_r(x) \subset U_{\alpha^*} \subset \bigcup_{\alpha \in I} U_\alpha$$

■

ולכן $\bigcup_{\alpha \in I} U_\alpha$ קבוצה פתוחה

טענה 2.1.2 חיתוך סופי של קבוצות פתוחות הוא פתוח

הוכחה: יהי $\{U_i\}_{i=1}^n$ אוסף סופי של קבוצות פתוחות U_i , נרצה להוכיח כי

$$\bigcap_{i=1}^n U_i$$

מהווה קבוצה פתוחה.

יהי $x \in \bigcap_{i=1}^n U_i$, אזי לכל $i \in 1, 2, \dots, n$ קיים $r_i > 0$ כך ש $S_{r_i}(x) \subset U_i$ נסמן

$$r = \min_{i=1, 2, \dots, n} r_i$$

אזי מתקיים לכל $i \in \{1, 2, \dots, n\}$ ש

$$S_r(x) \subset S_{r_i}(x) \subset U_i$$

ולכן

$$S_r(x) \subset \bigcap_{i=1}^n U_i$$

■

הגדרה 7 יהי (X, d) מרחב מטרי ו $A \subset X$. $x \in A$ תיקרא נקודה גבולית של A אם לכל $r > 0$: $S_r(x) \cap A \neq \{x\}$ מכילה לפחות נקודה אחת $y \neq x$

למה 2.1.1 (מסקנה מיידית מההגדרה) יהי (X, d) מרחב מטרי ו $A \subset X$.
תהי $x_0 \in A$ נקודה גבולית של A , אזי לכל $r > 0$ הכדור הפתוח $S_r(x_0)$ מכיל אינסוף נקודות.

הגדרה 8 קבוצה X תיקרא קבוצה סגורה אם היא מכילה את כל הנקודות הגבוליות שלה.

טענה 2.1.3 יהי (X, d) מרחב מטרי ו $A \subseteq X$. אזי A קבוצה סגורה ב X אם ורק אם A^c פתוחה.

הוכחה: אם $A = \emptyset$ או $A = X$ (כלומר $A^c = \emptyset$ או $A^c = X$ בהתאמה) זה טריוויאלי מההגדרה של קבוצה פתוחה. נניח כי $A \neq \emptyset \neq A^c$.

$\Leftarrow$ נניח כי A סגורה ב X . נרצה להראות ש A^c פתוחה ב X .

תהי $x \in A^c$ מכיון ש A סגורה ו $x \notin A$, לא יכולה להיות נקודה גבולית ולכן קיים $r > 0$ כך ש $S_r(x) \subseteq A^c$. מכיון שזה נכון עבור כל $x \in A^c$ אזי A^c פתוחה.

$\Rightarrow$ נניח כי A^c פתוחה, נרצה להראות ש A סגורה. תהי $x \in X$ נקודה גבולית של A . נניח בשלילה כי $x \notin A$ ולכן $x \in A^c$ ואנחנו יודעים כי A^c פתוחה ולכן קיים $r > 0$ כך ש $S_r(x) \subseteq A^c$. כלומר $S_r(x) \cap A = \emptyset$ ולכן x לא יכולה להיות נקודה גבולית של A בסתירה להנחה. לכן A סגורה. ■

2.1.2 סדרות במרחב מטרי

הגדרה 9 יהי (X, d) מרחב מטרי, נגיד שהסדרה $\{x_n\}_{n=1}^\infty$, $x_n \in X$ מתכנסת ל $x \in X$ אם ורק אם $\lim_{n \rightarrow \infty} d(x_n, x) = 0$ או במונחים מתמטיים:

$$\forall \epsilon > 0, \exists N \in \mathbb{N} : \forall n > N \quad d(x_n, x) \leq \epsilon$$

טענה 2.1.4 יהי (X, d) מרחב מטרי ו $A \subset X$ היא קבוצה סגורה אזי הגבול של כל סדרה המתכנסת בתוכה שייך ל A

הוכחה: תהי $\{x_n\}_{n=1}^\infty$ סדרה כלשהי ב A המתכנסת ל x ונניח בשלילה כי $x \in X \setminus A$. מהסגירות של A נובע כי $X \setminus A$ פתוחה, לכן קיים $r > 0$ כך ש $S_r(x) \subset X \setminus A$ כלומר קיימת סביבה של x שאינה מכילה אף נקודה מ A ולכן x איננה גבול של סדרת הנקודות $\{x_n\}_{n=1}^\infty$. זו סתירה ולכן $x \in A$. ■

הגדרה 10 הסדרה $\{x_n\}_{n=1}^\infty$ נקראת סדרת קושי אם:

$$\forall \epsilon > 0, \exists N \in \mathbb{N} : \forall n, m > N \quad d(x_n, x_m) \leq \epsilon$$

למה 2.1.2 כל סדרה מתכנסת היא סדרת קושי

הוכחה: נניח כי $\{x_n\}_{n=1}^\infty$ מתכנסת ל x . אזי

$$\forall \epsilon > 0, \exists N \in \mathbb{N} : \forall n > N \quad d(x_n, x) \leq \frac{\epsilon}{2},$$

אזי $\forall n, m > N$ מתקיים

$$d(x_n, x_m) \leq d(x_n, x) + d(x, x_m) \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

■

הגדרה 11 תהי $\{x_n\}_{n=1}^\infty$ סדרה כלשהי ויהיו $n_1 < n_2 < \dots \in \mathbb{N}$ אזי $\{x_{n_i}\}_{i=1}^\infty$ היא תת-סדרה

למה 2.1.3 תהי $\{x_n\}_{n=1}^\infty$ סדרת קושי. אם אפשר לבחור תת-סדרה $\{x_{n_k}\}_{k=1}^\infty$ מתכנסת, אזי $\{x_n\}_{n=1}^\infty$ מתכנסת.

הוכחה: מההתכנסות של $\{x_{n_k}\}_{k=1}^\infty$ נובע ישירות מההגדרה (9)

$$\exists N_1 \in \mathbb{N} : \forall n_k > N_1 \quad d(x_{n_k}, x) \leq \frac{\epsilon}{2}$$

ומההנחה ש $\{x_n\}_{n=1}^\infty$ היא סדרת קושי אזי מ (10)

$$\exists N_2 \in \mathbb{N} : \forall n, m > N_2 \quad d(x_n, x_m) \leq \frac{\epsilon}{2}$$

נגדיר : $N = \max(N_1, N_2)$ ואז לכל $n > N$ ולכל $\epsilon > 0$

$$d(x_n, x) \leq d(x_n, x_{n_k}) + d(x_{n_k}, x) \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

ולפי הגדרה (9) הסדרה $\{x_n\}_{n=1}^\infty$ מתכנסת. ■

משפט 2.1.1 (בולצאנו-ויירשטראס)
לכל סדרה חסומה ב $\mathbb{R}$ יש תת-סדרה מתכנסת

2.1.3 רציפות של פונקציה בין מרחבים מטריים

הגדרה 12 (רציפות בנקודה) יהי (X, d_X) ו (Y, d_Y) מרחבים מטריים, נגיד שהפונקציה

$f : X \mapsto Y$ רציפה ב $x_0 \in X$ אם : לכל $\epsilon > 0$ קיימת $\delta > 0$ כך ש

$$d_X(x, x_0) < \delta \Rightarrow d_Y(f(x), f(x_0)) < \epsilon$$

טענה 2.1.5 (הגדרה שקולה) יהיו (X, d_X) ו (Y, d_Y) מרחבים מטריים, אזי הפונקציה

$f : X \mapsto Y$ רציפה ב $x_0 \in X$ אם"ם לכל סדרה $\{x_n\}_{n=1}^\infty \subset X$ מתקיים כי אם $\lim_{n \rightarrow \infty} x_n \rightarrow x_0$ אזי $\lim_{n \rightarrow \infty} f(x_n) \rightarrow f(x_0)$

הוכחה: ($\Leftarrow$) נניח כי f רציפה ב x_0 . יהי $\epsilon > 0$ כלשהו אזי קיים $\delta > 0$ כך ש $d_Y(f(x), f(x_0)) < \epsilon$ כאשר $d_X(x, x_0) < \delta$. תהי $\{x_n\}_{n=1}^\infty$ סדרה כלשהי המקיימת $x_n \rightarrow x_0$ אזי קיים $N \in \mathbb{N}$ כך ש $d_X(x_n, x_0) < \delta$ לכל $n > N$. לכן $d_Y(f(x_n), f(x_0)) < \epsilon$ לכל $n > N$. כלומר $f(x_n) \rightarrow f(x_0)$

($\Rightarrow$) נניח בשלילה כי f אינה רציפה ב x_0 . אזי קיים ϵ^* כך שלכל $n \in \mathbb{N}$ קיים x_n המקיים $d_X(x_n, x_0) < \frac{1}{n}$ וגם $d_Y(f(x_n), f(x_0)) \geq \epsilon^*$ כלומר קיבלנו ש $x_n \rightarrow x_0$ אבל $f(x_n) \not\rightarrow f(x_0)$ וזו סתירה. ■

הערה 2.1.1 אם f רציפה בכל הנקודות של X נגיד ש f רציפה ב X

2.1.4 מרחבי פונקציות כמרחבים מטריים

הגדרה 13 תהי X קבוצה שרירותית ויהי (Y, d) מרחב מטרי, אזי הפונקציה $f : X \rightarrow Y$ נקראת חסומה אם

$$d(f(X)) = \sup_{x, y \in X} d_Y(f(x), f(y)) < \infty$$

הגדרה 14 נסמן ב $B(X, Y)$ את מרחב הפונקציות החסומות $f : X \rightarrow Y$.

טענה 2.1.6 תהי X קבוצה שרירותית ויהי (Y, d) מרחב מטרי אזי ההעתקה

$$d_\infty(f, g) = \sup_{x \in X} d_Y(f(x), g(x))$$

היא מטריקה מעל $B(X, Y)$.

הוכחה:

$$1. d_\infty(f, g) \geq 0 \text{ נובע ישירות מזה ש } d_Y(f(x), g(x)) \geq 0$$

2.

$$\begin{aligned} \sup_{x \in X} d_Y(f(x), g(x)) &= 0 \\ \implies \forall x \in X : d_Y(f(x), g(x)) &= 0 \quad (\text{מהגדרת הסופרמום}) \\ \implies \forall x \in X : f(x) &= g(x) \quad (d_Y \text{ מטריקה}) \\ \implies f &= g \end{aligned}$$

3. $d_\infty(f, g) = d_\infty(g, f)$ נובע ישירות מהסימטריות של d_Y
 4. יהיו $f, g, h \in \mathcal{B}(X, Y)$ לכל $x \in X$ מתקיים

$$\begin{aligned} \sup_{x \in X} d_Y(f(x), h(x)) &\leq \sup_{x \in X} [d_Y(f(x), g(x)) + d_Y(g(x), h(x))] \\ &\leq \sup_{x \in X} d_Y(f(x), g(x)) + \sup_{x \in X} d_Y(g(x), h(x)) \\ &= d_\infty(f, g) + d_\infty(g, h) \end{aligned}$$

לכן

$$\sup_{x \in X} d_Y(f(x), h(x)) \leq d_\infty(f, g) + d_\infty(g, h)$$

■

הערה 2.1.2 בעבודה זו בכל פעם שמדובר במרחב של פונקציות נשתמש במטריקה d_∞

הגדרה 15 מרחב הפונקציות הממשיות הרציפות בקטע הסגור $[a, b]$ יסומן על ידי $C[a, b]$

2.1.5 שלמות של מרחבים מטריים

הגדרה 16 מרחב מטרי (X, d) נקרא מרחב שלם אם כל סדרת קושי בו מתכנסת לנקודה ב X .

טענה 2.1.7 המרחב $\mathbb{R}$ עם המטריקה הסטנדרטית הוא שלם

הוכחה: נניח כי $\{a_n\}_{n=1}^\infty$ היא סדרת קושי, מההגדרה נובע כי לכל $\epsilon > 0$ קיים $N \in \mathbb{N}$ כך שלכל $m, n > N$ מתקיים $|a_n - a_m| < \epsilon$.

אזי לכל $n > N$ מתקיים $a_N - \epsilon < a_n < a_N + \epsilon$ כלומר הסדרה חסומה עבור $n \in [N, \infty]$ ולכן חסומה לכל $n > 0$.

לפי משפט בולצאנו-ויירשטראס (2.1.1) קיימת סדרה $\{a_{n_k}\}_{k=1}^\infty$ מתכנסת, לכן לפי למה (2.1.3) הסדרה $\{a_n\}_{n=1}^\infty$ מתכנסת.

■

טענה 2.1.8 המרחב $\mathbb{R}^d$ עם המטריקה הסטנדרטית הוא שלם

הוכחה: נניח כי $\{a_n\}_{n=1}^\infty \subset \mathbb{R}^d$ היא סדרת קושי.

נכתוב $a_n = (a_{n,1}, a_{n,2}, \dots, a_{n,d})$ וכמובן ש

$$\|a_n - a_m\| = \sqrt{\sum_{i=1}^d (a_{n,i} - a_{m,i})^2} \geq |a_{n,k} - a_{m,k}|$$

לכל $k = 1, 2, \dots, d$

ומכיון ש $\{a_n\}_{n=1}^\infty$ סדרת קושי אזי גם $\{a_{n,k}\}_{n=1}^\infty$ היא סדרת קושי ב $\mathbb{R}$ לכל k . לכן היא גם מתכנסת.

נסמן $a_k = \lim_{n \rightarrow \infty} a_{n,k}$ ונקבל כי $\lim_{n \rightarrow \infty} a_n = (a_1, a_2, \dots, a_d)$

■

כלומר הסדרה $\{a_n\}_{n=1}^\infty$ מתכנסת.

טענה 2.1.9 המרחב הוקטורי $C[a, b]$ עם המטריקה $d_\infty(f, g) = \sup_{x \in [a, b]} |f(x) - g(x)|$ הוא מרחב שלם

הוכחה: תהי $\{f_n\}_{n=1}^\infty$ סדרת קושי ב $C[a, b]$. כלומר

$$\forall \epsilon > 0, \exists N \in \mathbb{N} : \forall n, m > N \quad d(f_n, f_m) \leq \epsilon$$

צריך להראות שקיימת פונקציה $f \in C[a, b]$ כך ש:

$$\lim_{m \rightarrow \infty} d_\infty(f_m, f) = 0$$

ניתן לראות כי לכל $x \in [a, b]$ מתקיים

$$|f_n(x) - f_m(x)| \leq \max_{x \in [a, b]} |f_n(x) - f_m(x)| = d(f_n, f_m) < \epsilon$$

אזי מהגדרת סדרת קושי נובע ישירות כי:

$$(2.1.5.1) \quad \forall \epsilon > 0, \exists N \in \mathbb{N} : \forall n, m > N \quad |f_n(x) - f_m(x)| \leq \epsilon$$

כלומר לכל x קבוע סדרת המספרים $\{f_n(x)\}_{n=1}^\infty$ היא קושי (ב $\mathbb{R}$). ומכיון ש $\mathbb{R}$ הוא מרחב שלם, הסדרה הזו מתכנסת. נקרא לגבול שלה $f(x)$.

נקח את הגבול כאשר $m \rightarrow \infty$ ב (2.1.5.1) ונקבל

$$(2.1.5.2) \quad \forall x \in [a, b], \forall n > N : |f_n(x) - f(x)| < \epsilon$$

ומכיון שזה לכל x אז ברור ש

$$\forall n > N : \max_{x \in [a, b]} |f_n(x) - f(x)| = d(f_n, f) < \epsilon$$

ועכשיו נשאר להראות כי $f \in C[a, b]$, כלומר ש- f רציפה.
לפי (2.1.5.2)

$$\forall \epsilon' > 0, \exists N \in \mathbb{N} : \forall n > N, \forall x \in [a, b] \quad |f_n(x) - f(x)| < \epsilon'$$

נבחר $\epsilon' = \frac{\epsilon}{3}$

$$(2.1.5.3) \quad \exists N \in \mathbb{N} : \forall n > N, \forall x \in [a, b] \quad |f(x) - f_n(x)| < \frac{\epsilon}{3}$$

יהי $x_0 \in [a, b]$ מהרציפות של f_n נובע כי

$$(2.1.5.4) \quad \exists \delta > 0 : |x - x_0| < \delta \Rightarrow |f_n(x) - f_n(x_0)| < \frac{\epsilon}{3}$$

מכיוון ש

$$|f(x) - f(x_0)| \leq |f(x) - f_n(x)| + |f_n(x) - f_n(x_0)| + |f_n(x_0) - f(x_0)|$$

נקבל מ (2.1.5.3) ו (2.1.5.4) כי

$$|x - x_0| < \delta \implies |f(x) - f(x_0)| < \epsilon$$

כלומר $f \in C[a, b]$. וזה משלים את ההוכחה.

■

למה 2.1.4 יהי (X, d) מרחב מטרי שלם ו $A \subset X$ אזי A קבוצה סגורה אם ורק אם A היא מהווה מרחב שלם עם המטריקה המושרית d .

הוכחה: ($\Leftarrow$) נניח כי A סגורה ותהי $\{a_n\}_{n=1}^{\infty}$ סדרת קושי ב A , הסדרה הזו מתכנסת מכיוון שהיא נמצאת גם ב X שהוא מרחב שלם, ומהסגירות של A נובע כי הגבול נמצא ב A ולכן A שלם.

($\Rightarrow$) תהי A תת-קבוצה של X שמהווה מרחב מטרי שלם עם המטריקה המושרית d ותהי a נקודה גבולית של A , כלומר קיימת סדרה $\{a_n\}_{n=1}^{\infty}$ שמתכנסת ל a . מההתכנסות נובע שסדרה זו היא גם קושי ומהשלמות של A נובע כי $a \in A$. מכיוון שזה נכון לכל נקודה גבולית של A אזי A מהווה קבוצה סגורה.

■

הגדרה 17 יהי (X, d) מרחב מטרי ו $A \subset X$ הקוטר של קבוצה A הוא

$$d(A) = \sup_{a, b \in A} d(a, b)$$

בהנחה ש $d(A) < \infty$.

הגדרה 18 סדרת קבוצות $\{A_n\}_{n=1}^\infty$ תיקרא מקוננת אם מתקיים:

$$A_1 \supset A_2 \supset A_3 \supset \dots$$

משפט 2.1.2 יהי (X, d) מרחב מטרי שלם. תהי $\{A_n\}_{n=1}^\infty$ סדרה מקוננת של קבוצות סגורות ולא ריקות עם קוטר שואף לאפס. אזי

$$\bigcap_{n=1}^\infty A_n \neq \emptyset$$

הוכחה:

עבור $n = 1, 2, \dots$ נבחר $a_n \in A_n$. נראה שהסדרה $\{a_n\}_{n=1}^\infty$ היא קושי :
יהי $\epsilon > 0$ נבחר N_ϵ כך ש $d(A_n) < \epsilon$ עבור $n \geq N_\epsilon$.
אז עבור $m, n \geq N_\epsilon$, מתקיים :

$$(2.1.5.5) \quad A_n, A_m \subset A_{N_\epsilon} \Rightarrow x_n, x_m \in A_{N_\epsilon}$$

$$(2.1.5.6) \quad d(A_n) < \epsilon \Rightarrow^{(2.1.5.5)} d(a_n, a_m) < \epsilon$$

ולכן $\{a_n\}_{n=1}^\infty$ היא קושי ומכיוון ש (X, d) שלם לכן היא גם מתכנסת.

נסמן $a = \lim_{n \rightarrow \infty} a_n$ ונראה כי $a \in A_n$ לכל n .

נבחר $k \in \mathbb{N}$ כלשהו, עבור $m \geq k$ מתקיים $a_m \in A_m \subset A_k$ ולכן $\{a_m\}_{m=k}^\infty \subset A_k$.

היות ו $\{a_m\}_{m=k}^\infty$ מתכנסת והקבוצה A_k סגורה נובע כי $a \in A_k$.

לכן

$$a \in \bigcap_{n=1}^\infty A_n$$

■

2.1.6 קומפקטיות במרחבים מטריים

הגדרה 19 יהי (X, d) מרחב מטרי, כיסוי פתוח של $A \subset X$ הוא אוסף של קבוצות פתוחות $\{G_\alpha\}_{\alpha \in I}$ שהאיחוד שלהן מכיל את A . כלומר

$$A \subseteq \bigcup_{\alpha \in I} G_\alpha$$

כאשר I היא קבוצת אינדקסים (סופית או אינסופית).

הגדרה 20 תת־כיסוי של $\{G_\alpha\}$ זו תת־קבוצה $\{G_{\alpha_\gamma}\}$ שגם היא כיסוי של A .

הגדרה 21 יהי (X, d) מרחב מטרי ו $A \subset X$. נגיד ש A קבוצה קומפקטית אם לכל כיסוי פתוח קיים תת־כיסוי סופי.

משפט 2.1.3 קבוצה קומפקטית היא חסומה.

הוכחה:

תהי A קבוצה קומפקטית ויהי $x \in A$. נסתכל על אוסף הקבוצות $\{S_r(x)\}_{r=1}^\infty$ ברור שזה כיסוי פתוח ולכן מהקומפקטיות נובע כי קיים תת־כיסוי סופי, כלומר קיים $R \geq 1$ כך ש $A \subseteq \bigcup_{r=1}^R S_r(x)$ וברור ש $S_{r_1}(x) \subset S_{r_2}(x)$ לכל $r_2 > r_1$ לכן $A \subseteq S_R(x)$ ■

משפט 2.1.4 קבוצה קומפקטית היא סגורה

הוכחה:

יהי (X, d) מרחב מטרי ו $A \subset X$.

נראה כי $A^c = X \setminus A$ פתוחה. נבחר נקודה כלשהי $y \in A^c$ ונראה שהיא פנימית.

לכל $x \in A$, $x \neq y$ ולכן קיימות סביבות סביב x ו y זרות. כלומר קיים $\epsilon_x > 0$ כך ש:

$$(2.1.6.1) \quad S_{\epsilon_x}(x) \cap S_{\epsilon_x}(y) = \emptyset$$

ובנוסף, ברור כי

$$A \subset \bigcup_{x \in A} S_{\epsilon_x}(x)$$

אבל מכיוון ש A קומפקטית חייבים להיות קיימים $x_1, x_2, \dots, x_n$ כך ש:

$$A \subset \bigcup_{i=1}^n S_{\epsilon_{x_i}}(x_i)$$

נגדיר:

$$B = \bigcap_{i=1}^n S_{\epsilon_{x_i}}(y)$$

שהיא קבוצה פתוחה המכילה את y . נרצה להראות ש $B \subset A^c$ (או לחלופין $B \cap A = \emptyset$) כדי להראות ש y נקודה פנימית.

$$\begin{aligned} (B \cap A) &\subset (B \cap \bigcup_{i=1}^n S_{\epsilon_{x_i}}(x_i)) = \bigcup_{i=1}^n (B \cap S_{\epsilon_{x_i}}(x_i)) \subset \\ &\subset \bigcup_{i=1}^n (S_{\epsilon_{x_i}}(y) \cap S_{\epsilon_{x_i}}(x_i)) = \bigcup_{i=1}^n \emptyset = \emptyset \end{aligned}$$

■

משפט 2.1.5 תהי A קבוצה קומפקטית ו $B \subset A$ קבוצה סגורה. אזי B קומפקטית

הוכחה:

יהי $\{G_\alpha\}_{\alpha \in I}$ כיסוי פתוח של B . (נבחין כי B^c היא קבוצה פתוחה) אזי

$$\bigcup_{\alpha \in I} G_\alpha \cup B^c$$

הינו כיסוי פתוח של A . ומכיוון ש A קומפקטית אזי חייב להיות לה תת־כיסוי סופי, כלומר

$$A \subset \bigcup_{i=1}^n G_{\alpha_i} \cup B^c$$

ולכן

$$B \subset \bigcup_{i=1}^n G_{\alpha_i}$$

■

משפט 2.1.6 קטע סגור $[a, b]$ הוא קבוצה קומפקטית ב $\mathbb{R}$

הוכחה:

נניח בשלילה כי זה לא נכון, אזי קיים כיסוי פתוח $\{G_\alpha\}$ כך שאין לו תת־כיסוי סופי. נחלק את הקטע לשני חלקים שווים (נקרא לאמצע הקטע c_1) ברור ש $\{G_\alpha\}$ מכסה את שניהם. אזי לפחות לאחד מהקטעים אין תת־כיסוי סופי, בלי הגבלת הכלליות נניח שזה $[a, c_1]$ ונקרא לו I_1 . באותו אופן אפשר לחלק את הקטע I_1 לשני חלקים שווים (נקרא לאמצע הקטע c_2) ושוב בלי הגבלת הכלליות נבחר את $I_2 = [a, c_2]$ להיות הקטע שאין לו תת־כיסוי סופי. נמשיך כך ונקבל סדרת קטעים מקוננים $I_n = [a, c_n]$ עם קוטר שואף לאפס.

לפי משפט (2.1.2) קיים x ששייך לכל הקטעים בסדרה. ברור ש x היא נקודה פנימית של אחת מהקבוצות הפתוחות $\{G_\alpha\}$, נקרא לקבוצה הזו G_{α^*} . כלומר

$$\exists r > 0 : S_r(x) \subset G_{\alpha^*}$$

ומצד שני

$$\exists n : \forall n > N, S_r(x) \supset I_n$$

כלומר $I_N \subseteq G_{\alpha^*}$. בסתירה לעובדה שבנינו את הקטעים I_n כך שאין להם תת־כיסוי סופי. ■

משפט 2.1.7 תהי A קבוצה קומפקטית, אזי לכל תת קבוצה אינסופית שלה $E \subset A$ יש נקודה גבולית ב A

הוכחה: נניח בשלילה כי לא קיימת נקודה כזאת, אזי לכל $x \in E$ אפשר לבחור סביבה מספיק קטנה $S_{r_x}(x)$ כך שתכיל רק את הנקודה x . נבנה כיסוי פתוח

$$\{S_{r_x}(x) | x \in E\} \cup \{S_1(x) | x \in A \setminus E\}$$

■ שאין לו תת-כיסוי סופי. בסתירה להנחה ש A קומפקטית.

הגדרה 22 יהי (X, d) מרחב מטרי ו $A \subset X$

נגיד שקבוצת נקודות סופית $\{a_1, a_2, \dots, a_\ell\}$ מהווה ϵ -כיסוי של A אם $A \subset \bigcup_{k=1}^\ell S_\epsilon(a_k)$ או במילים אחרות לכל $a \in A$ קיים $1 \leq i \leq \ell$ כך ש $d(a, a_i) < \epsilon$.

הגדרה 23 יהי (X, d) מרחב מטרי ו $A \subset X$. נגיד ש A חסומה בהחלט אם לכל $\epsilon > 0$ קיים ϵ -כיסוי של A .

טענה 2.1.10 יהי (X, d) מרחב מטרי ו $A \subset X$. אזי A חסומה בהחלט אם ורק אם לכל $\epsilon > 0$ קיים כיסוי סופי $\{G_i\}_{i=1}^n$ כך ש $d(G_i) < \epsilon$.

הוכחה: ($\Leftarrow$) יהי $\epsilon > 0$ אזי קיימת קבוצת נקודות סופית $\{a_1, a_2, \dots, a_\ell\}$ כך ש $A \subseteq \bigcup_{k=1}^\ell S_{\frac{\epsilon}{2}}(a_k)$ נגדיר $G_i = S_{\frac{\epsilon}{2}}(a_i)$, $i = 1, 2, \dots, \ell$ ויהי $x, y \in G_i$ אזי מתקיים

$$d(x, y) \leq d(x, a_i) + d(a_i, y) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

($\Rightarrow$) יהי $\epsilon > 0$ אז לפי ההנחה קיימות $G_1, G_2, \dots, G_n$ המקיימות $d(G_i) < \epsilon$ וכי $A \subset \bigcup_{i=1}^n G_i$ נניח כי G_i אינן קבוצות ריקות ונבחר $a_i \in G_i$. יהי $a \in A$ כלשהו אז ברור שקיים $1 \leq k \leq n$ כך שמתקיים $a \in G_k$ ולכן $d(a, a_k) \leq d(G_k) < \epsilon$ כלומר $a \in S_\epsilon(a_k)$ מכיוון שזה נכון לכל $a \in A$ אז $A \subset \bigcup_{i=1}^n S_\epsilon(a_i)$ ■

טענה 2.1.11 תהי A חסומה בהחלט אזי $B \subseteq A$ גם חסומה בהחלט.

הוכחה: A חסומה בהחלט אזי לפי טענה קודמת קיימות $\{G_i\}_{i=1}^\ell$ כך ש $A \subseteq \bigcup_{k=1}^\ell G_k$ ושלכל $i = 1, 2, \dots, \ell$ $d(G_i) < \epsilon$. נשיב לב כי לכל $i = 1, 2, \dots, \ell$ מתקיים:

$$1. d(G_i \cap B) \leq d(G_i) < \epsilon$$

$$2. \bigcup_{i=1}^\ell (G_i \cap B) = (\bigcup_{i=1}^\ell G_i) \cap B \supseteq A \cap B = B$$

■ כלומר לכל ϵ מצאנו כיסוי $\{G_i\}_{i=1}^\ell$ לכן B חסומה בהחלט.

הגדרה 24 יהי (X, d) מרחב מטרי ו $A \subset X$. מספר הכיסוי שנסמן ע"י $\mathcal{N}(A, \epsilon)$ הוא מספר הכדורים המינימלי שמחווים ϵ -כיסוי כלומר:

$$\mathcal{N}(A, \epsilon) := \min\{\ell : \exists(a_1, \dots, a_\ell) A \subseteq \bigcup_{k=1}^\ell S_\epsilon(a_k)\}$$

הערה 2.1.3 $\mathcal{N}(A, \epsilon)$ יכול להיות סופי או אינסופי (כאשר אין ϵ כיסוי).

דוגמה 2.1.2 תהי $B_{R, \|\cdot\|_\infty} = \{\theta \in \mathbb{R}^d : \|\theta\|_\infty \leq R\}$ (קוביה במרחב $\mathbb{R}^d$) נרצה למצוא חסם ל $\mathcal{N}(B_{R, \|\cdot\|_\infty}, \delta)$ עם המטריקה הסטנדרטית $(d(\theta, \tilde{\theta}) = \|\theta - \tilde{\theta}\|_2)$. כלומר נרצה למצוא ℓ נקודות $\{\theta_i\}_{i=1}^\ell$ כך שלכל $\theta \in B_{R, \|\cdot\|_\infty}$ קיים $1 \leq i \leq \ell$ כך ש: $d(\theta, \theta_i) = \|\theta - \theta_i\|_2 \leq \delta$
נתבונן קודם במקרה ש $d = 2$:

המרחב שלנו הוא הריבוע $[-R, R] \times [-R, R]$ שאותו נרצה לכסות ע"י מעגלים ברדיוס δ . קודם נכסה את המרחב בעזרת ריבועים עם אורך צלע δ , חישובי שטחים נותנים חסם תחתון למספר הריבועים הדרוש כדי לכסות את הריבוע הגדול: $2^2 \left\lceil \frac{R}{\delta} \right\rceil^2$ וכדי לכסות את המרחב ע"י מעגלים נשאר לקחת מעגלים ברדיוס δ במרכז כל אחד מהריבועים הקטנים. ומכיוון שזה נכון לכל d נקבל כי:

$$(2.1.6.2) \quad \mathcal{N}(B_{R, \|\cdot\|_\infty}, \epsilon) \leq 2^d \left\lceil \frac{R}{\delta} \right\rceil^d$$

הערה 2.1.4 מכיוון ש $\forall \theta \in \mathbb{R}^d : \|\theta\|_\infty \leq \|\theta\|_2$ אזי ברור ש $B_{R, \|\cdot\|_2} \subseteq B_{R, \|\cdot\|_\infty}$ ולכן:

$$(2.1.6.3) \quad \mathcal{N}(B_{R, \|\cdot\|_2}, \delta) \leq \mathcal{N}(B_{R, \|\cdot\|_\infty}, \delta) \leq 2^d \left\lceil \frac{R}{\delta} \right\rceil^d$$

הגדרה 2.5 יהי (X, d) מרחב מטרי ו $A \subset X$. נגיד ש A קומפקטית סדרתית אם לכל סדרה $\{x_n\}_{n=1}^\infty$ ב A קיימת תת-סדרה מתכנסת שהגבול שלה נמצא ב A .

משפט 2.1.8 יהי (X, d) מרחב מטרי ו $A \subset X$. אזי הטענות האלו שקולות:

1. A קבוצה קומפקטית

2. A קומפקטית סדרתית

3. A שלמה וחסומה בהחלט

הוכחה:

1 $\implies$ 2

נקח סדרה $\{x_n\}_{n=1}^\infty$ כלשהי ב A , ונסתכל על אוסף האיברים $x_1, x_2, x_3, \dots$. אם האוסף הזה סופי, אזי קיים איבר שמופיע בסדרה אינסוף פעמים ולכן אפשר לבחור את האיברים האלה כתת-סדרה וברור שהיא מתכנסת. אם האוסף אינו סופי, לפי טענה (2.1.7) באוסף הזה חייבת להיות נקודה גבולית, ולכן אפשר לבחור תת-סדרה מתכנסת והגבול נמצא ב A כי A סגורה לפי משפט 2.1.4.

$$2 \Rightarrow 3$$

(שלמות)

תהי $\{x_n\}_{n=1}^\infty$ סדרת קושי ב A . מהקומפקטיות הסדרתית (הגדרה 2.5) נובע כי קיימת תת-סדרה $\{x_{n_k}\}_{k=1}^\infty$ מתכנסת. ומלמה 2.1.3 נובע שגם $\{x_n\}_{n=1}^\infty$ מתכנסת.

(חסימות בהחלט)

נניח בשלילה ש A אינה חסומה בהחלט, אזי קיים $r > 0$ כך שאי אפשר לכסות את A על ידי מספר סופי של כדורים ברדיוס r . נבנה סדרה $\{x_n\}_{n=1}^\infty$ ב A כך שאין לה תת-סדרה מתכנסת ונגיע לסתירה.

נקח $x_1 \in A$ כלשהו, מכיוון ש $S_r(x_1)$ אינה מכסה את A אז קיימת לפחות נקודה אחת $x_2 \in A \setminus S_r(x_1)$.

נמשיך באותו אופן ע"י בחירת הנקודות כך:

$$x_{n+1} \in A \setminus \bigcup_{i=1}^n S_r(x_i)$$

ויקבלנו סדרה המקיימת

$$d(x_n, x_m) \geq r, \forall m, n > 0$$

לסדרה כזאת אי אפשר למצוא תת-סדרה מתכנסת. כי אם הסדרה x_{n_k} מתכנסת אזי היא קושי (2.1.2) ולכן מתקיים

$$d(x_{n_k}, x_{n_m}) < r$$

עבור k, m גדולים מספיק. זו סתירה להנחה ש A קומפקטית סדרתית

$$3 \Rightarrow 1$$

נניח בשלילה כי זה לא נכון, כלומר קיים כיסוי פתוח $G = \{G_\alpha\}_{\alpha \in I}$ שאין לו תת-כיסוי סופי. הקבוצה A חסומה בהחלט ולכן קיימת קבוצת נקודות סופית $\{x_1, x_2, \dots, x_\ell\}$ כך ש

$$A \subset \bigcup_{i=1}^\ell S_{\frac{1}{2}}(x_i)$$

נסתכל על האוסף

$$\{S_{\frac{1}{2}}(x_i) \cap A\}, i = 1, 2, \dots, \ell$$

לפחות לאחת מהקבוצות אין תת-כיסוי סופי (אחרת A קומפקטית), נקרא לה A_1 . $A_1 \subseteq A$ גם היא חסומה בהחלט ולכן קיימת קבוצת נקודות סופית $\{x'_1, x'_2, \dots, x'_{\ell'}\}$ כך ש $A_1 \subset \bigcup_{i=1}^{\ell'} S_{\frac{1}{2}}(x'_i)$, נבחר את הקבוצה שאין לה תת-כיסוי סופי ונקרא לה A_2 . וכן הלאה. בצורה הזאת נקבל סדרת קבוצות מקוננות A_n שהקוטר שלהן שואף ל-0. על ידי בחירת נקודות $x_n \in A_n$ נקבל סדרת קושי $\{x_n\}_{n=1}^\infty$. מהשלמות של A נובע כי הסדרה גם מתכנסת ל $y \in A$. ולכן $\exists \alpha_0 : y \in G_{\alpha_0}$ ומכיוון ש G_{α_0} קבוצה פתוחה אזי קיים $r > 0$ כך ש $S_r(y) \subset G_{\alpha_0}$.

נקח n מספיק גדול כך ש $d(x_n, y) < \frac{r}{2}$ וגם $1/2^n < \frac{r}{2}$

אזי

$$\forall x \in A_n : d(x, y) \leq d(x, x_n) + d(x_n, y) < \frac{r}{2} + \frac{r}{2} = r$$

ולכן ל A_n קיים תת־כיסוי סופי (G_{α_0}) בסתירה לאופן שבו בנינו את הקבוצות A_n לכן A קומפקטית. ■

מסקנה 2.1.1 קבוצה קומפקטית היא חסומה לחלוטין ולכן לפי הגדרה מספר הכיסוי שלה הוא סופי

משפט 2.1.9 (מסקנה) $A \subset \mathbb{R}^n$ קומפקטית אם ורק אם A חסומה וסגורה

הוכחה: הכיוון הראשון הוא מסקנה ישירה של המשפטים 2.1.3 ו 2.1.4. עכשיו נניח כי נתונה לנו קבוצה חסומה וסגורה $A \subset \mathbb{R}^n$. נראה ש A חסומה בהחלט ושלמה ולכן לפי משפט 2.1.8 היא קומפקטית.

(חסימות בהחלט) : מהחסימות של A נובע כי קיים $R > 0$ כלשהו כך ש: $A \subset B_{R, \|\cdot\|_\infty}$ שלפי דוגמה 2.1.2 היא חסומה בהחלט. לכן לפי טענה 2.1.11 גם A חסומה בהחלט.

(שלמות) : לפי טענה 2.1.8 $\mathbb{R}^n$ הוא שלם, ולפי למה 2.1.4 והסגירות של A נקבל כי A הוא תת־מרחב שלם. ■

משפט 2.1.9 לא נכון עבור מרחב מטרי כלשהו:

דוגמה 2.1.3 יהי $X = C[a, b]$ והמטריקה

$$d_\infty(f, g) = \sup_{x \in [a, b]} |f(x) - g(x)|$$

אזי כדור סגור במרחב המטרי (X, d) לא מהווה קבוצה קומפקטית.

הוכחה: תהי $f_n(x)$ פונקציה שהגרף שלה הוא משולש שמיוצר על ידי שלושת הנקודות:

$$\left(\frac{1}{2}\left(\frac{1}{n+1} + \frac{1}{n}\right), 0\right), \left(\frac{1}{n}, 1\right), \left(\frac{1}{2}\left(\frac{1}{n} + \frac{1}{n-1}\right), 0\right)$$

מכיוון שהחיתוך בין הגרפים של $f_n(x)$ ו $f_{n+1}(x)$ הוא רק בנקודה $\left(\frac{1}{2}\left(\frac{1}{n+1} + \frac{1}{n}\right), 0\right)$ וכי

$$\forall n \geq 2 : \sup f_n(x) = 1$$

$$(2.1.6.4) \quad \forall n, m > 1 : |f_n(x) - f_m(x)| = 1$$

נבחר את האוסף $\{S_{\frac{1}{2}}(f)\}_{f \in X}$ ברור שזה כיסוי פתוח של X . אבל כל אחת מהקבוצות באוסף לא יכולה להכיל יותר מפונקציה אחת מסדרת הפונקציות f_n , אחרת נסתור את (2.1.6.4) ולכן אי אפשר לוותר על אף קבוצה מהאוסף, ולכן לא קיים תת־כיסוי סופי.

כלומר הכדור הסגור במרחב מטרי שלם אינו קומפקטי.

משפט 2.1.10 תהי $f : X \rightarrow Y$ פונקציה רציפה בין מרחבים מטרים. אם X קומפקטית אזי $f(X)$ קבוצה קומפקטית.

הוכחה: תהי סדרה $\{y_n\}_{n=1}^\infty$ ב $f(X)$, אנחנו צריכים להראות כי קיימת תת-סדרה שמתכנסת לנקודה ב $f(X)$.

נבנה סדרה $\{x_n\}_{n=1}^\infty$ המוכלת ב X כך ש $f(x_n) = y_n$. מהקומפקטיות של X נובע כי קיימת תת-סדרה $\{x_{n_k}\}_{k=1}^\infty$ המתכנסת לנקודה ב X ומכיון ש f רציפה אזי לפי טענה 2.1.5 $\{f(x_{n_k})\}_{k=1}^\infty$ מתכנסת ל $f(x^*)$. וכמובן ש $f(x^*) \in Y$ כי $x^* \in X$. ■

מסקנה 2.1.2 (וירשטראס) יהי (X, d) מרחב קומפקטי ותהי $f : X \rightarrow \mathbb{R}$ פונקציה רציפה, אזי f מקבלת מקסימום ומינימום.

הוכחה: ממשפט קודם $f(X)$ היא תת-קבוצה קומפקטית של $\mathbb{R}$ ולכן חסומה וסגורה. מהחסימות נובע כי קיים סופרימום, נסמן $M = \sup_{x \in X} |f(X)|$ ניקח סדרה $\{y_n\}_{n=1}^\infty$ המקיימת: $M - \frac{1}{n} < y_n \leq M$ (זה אפשרי מהגדרת הסופרימום). נשים לב כי $y_n \xrightarrow{n \rightarrow \infty} M$ ומהסגירות נובע כי הגבול שייך ל $f(X)$. כלומר קיים $x_0 \in X$ כך ש $f(x_0) = M$. בעזרת טיעון דומה אפשר להוכיח שקיים מינימום. ■

2.2 כלים מתורת ההסתברות

2.2.1 הגדרות בסיסיות

הגדרה 26 יהיו x, y משתנים מקריים ממשיים, אזי הפונקציה

$$F(x, y) = P(x \leq x, y \leq y)$$

נקראת פונקציית ההתפלגות המשותפת של הוקטור המקרי $z = (x, y)$

הגדרה 27 יהיו $x_1, x_2 \dots x_n$ מ"מ, אם קיימת פונקציה

$$\rho_{x_1, x_1 \dots x_n} : \mathbb{R}^n \rightarrow [0, \infty)$$
 כך שלכל קבוצה פתוחה $A \subset \mathbb{R}^n$ מתקיים:

$$P((x_1, x_1 \dots x_n) \in A) = \int \dots \int_A \rho_{x_1, x_1 \dots x_n}(x_1, x_2 \dots x_n) dx_1 dx_2 \dots dx_n$$

נקרא לה פונקציית הצפיפות המשותפת של $(x_1, x_2 \dots x_n)$

הגדרה 28 יהיו y, x משתנים מקריים בעלי צפיפות משותפת $\rho(x, y)$, נגדיר את פונקציית ההתפלגות השולית של x ע"י:

$$\rho_x(x) = \int_{-\infty}^{\infty} \rho(x, y) dy$$

הגדרה 29 יהיו y, x משתנים מקריים, נגיד שהם בלתי תלויים אם מתקיים לכל $A, B \in \mathbb{R}$

$$P(x \in A, y \in B) = P(x \in A)P(y \in B).$$

אם קיימת פונקציית הצפיפות המשותפת תנאי זה שקול ל

$$\rho(x, y) = \rho_x(x)\rho_y(y).$$

הגדרה 30 יהיו y, x משתנים מקריים בעלי צפיפות משותפת $\rho(x, y)$, נגדיר את פונקציית הצפיפות המותנת של x בהינתן y ע"י:

$$\rho_{y|x}(y|x) = \frac{\rho(x, y)}{\rho_x(x)}$$

עבור x כך ש $\rho_x(x) \neq 0$

הגדרה 31 יהיו y, x משתנים מקריים בעלי צפיפות משותפת $\rho(x, y)$, התוחלת המותנת של y בהינתן $x = x$ מוגדרת ע"י

$$E_{y|x}[y|x] = \int_{-\infty}^{\infty} y \rho_{y|x}(y|x) dy$$

למה 2.2.1 (חוק התוחלת השלמה) יהיו y, x משתנים מקריים בעלי צפיפות משותפת $\rho(x, y)$, אזי מתקיים:

$$E_y[y] = E_x[E_{y|x}[y|x]]$$

2.2.2 אי-שוויונות הסתברותיים

משפט 2.2.1 (אי-שוויון מרקוב)

יהי x מ"מ המקיים: $E|x| < \infty$

אזי לכל $\lambda > 0$ מתקיים:

$$P(|x| \geq \lambda) \leq \frac{E|x|}{\lambda}$$

בפרט, אם $x \geq 0$ אזי

$$P(x \geq \lambda) \leq \frac{E(x)}{\lambda}$$

הוכחה: עבור x בדיד :

$$\begin{aligned} P(|x| \geq \lambda) &= \sum_{|x_j| \geq \lambda} P(x = x_j) \leq \sum_{|x_j| \geq \lambda} \frac{|x_j|}{\lambda} P(x = x_j) \\ &\leq \sum_{j=1}^{\infty} \frac{|x_j|}{\lambda} P(x = x_j) = \frac{1}{\lambda} E|x| \end{aligned}$$

■ ההוכחה עבור x רציף דומה כאשר מחליפים סכום באינטגרל.

משפט 2.2.2 (מסקנה מאי־שוויון מרקוב)

תהי $f : [0, \infty) \rightarrow [0, \infty)$ פונקציה מונוטונית עולה ממש ויהי $y \geq 0$ מ"מ. אזי לכל $\lambda > 0$ מתקיים:

$$P(y \geq \lambda) \leq \frac{1}{f(\lambda)} E(f(y))$$

הוכחה: מכיוון ש $f(y)$ עולה ממש מתקיים : $y \geq \lambda \iff f(y) \geq f(\lambda)$ לכן

$$P(y \geq \lambda) = P(f(y) \geq f(\lambda))$$

נפעיל את אי־שוויון מרקוב על $x = f(y) \geq 0$

$$P[y \geq \lambda] = P[f(y) \geq f(\lambda)] \leq \frac{1}{f(\lambda)} E(f(y))$$

■

משפט 2.2.3 (אי־שוויון צ'בישב) יהי x משתנה מקרי בעל שני מומנטים עם תוחלת μ ושונות σ^2 אזי לכל $\lambda > 0$ מתקיים:

$$P(|x - \mu| \geq \lambda) \leq \frac{\sigma^2}{\lambda^2}$$

הוכחה: ניישם את (2.2.2) עם

$$y = |x - \mu|, f(y) = y^2$$

ונקבל :

$$P(|x - \mu| \geq \lambda) \leq \frac{1}{\lambda^2} E[|x - \mu|^2] = \frac{1}{\lambda^2} \sigma^2$$

■

משפט 2.2.4 (בנט) יהיו $\{\xi_i\}_{i=1}^m$ **משתנים מקריים ב"ת כך ש:** $\mathbf{E}(\xi_i) = \mu_i$ $\text{var}(\xi_i) = \sigma_i^2$ **וגם** $|\xi_i - \mu_i| \leq M$ **אזי לכל** $\epsilon > 0$:

$$P\left(\frac{1}{m} \sum_{i=1}^m [\xi_i - \mu_i] \geq \epsilon\right) \leq \exp\left\{\frac{-\epsilon m}{M} \left[\left(1 + \frac{\Sigma^2}{mM\epsilon}\right) \log\left(1 + \frac{mM\epsilon}{\Sigma^2}\right) - 1\right]\right\}$$

$$\Sigma^2 = \sum_{i=1}^m \sigma_i^2 \quad \text{כאשר}$$

הוכחה: בלי הגבלת הכלליות נניח כי $\mu_i = 0 \quad \forall i$

נישם את משפט (2.2.2) עם

$$\mathbf{y} = \sum_{i=1}^m \xi_i \quad f(\mathbf{y}) = e^{c\mathbf{y}} (c > 0) \quad \lambda = m\epsilon$$

ונקבל:

$$P\left(\frac{1}{m} \sum_{i=1}^m \xi_i \geq \epsilon\right) = P\left(\sum_{i=1}^m \xi_i \geq m\epsilon\right) \leq e^{-cm\epsilon} \mathbf{E}\left[e^{\sum_{i=1}^m c\xi_i}\right] = \quad (2.2.2.1)$$

$$\stackrel{\text{(אי תלות)}}{=} e^{-cm\epsilon} \prod_{i=1}^m \mathbf{E}(e^{c\xi_i})$$

מפיתוח טיילור סביב אפס של e^x **נקבל :**

$$\begin{aligned} \mathbf{E}(e^{c\xi_i}) &= \mathbf{E}\left(\sum_{k=0}^{\infty} \frac{c^k \xi_i^k}{k!}\right) = 1 + c\mathbf{E}(\xi_i) + \mathbf{E}\left(\sum_{k=2}^{\infty} \frac{c^k \xi_i^k}{k!}\right) \\ &= 1 + \mathbf{E}\left(\sum_{k=2}^{\infty} \frac{c^k \xi_i^{k-2} \mathbf{x}_i^2}{k!}\right) \leq 1 + \sum_{k=2}^{\infty} \frac{c^k M^{k-2} \sigma_i^2}{k!} \end{aligned}$$

נשתמש באי-שוויון $1 + t \leq e^t$ **ונקבל :**

$$\mathbf{E}(e^{c\xi_i}) \leq \exp\left\{\sum_{k=2}^{\infty} \frac{c^k M^{k-2} \sigma_i^2}{k!}\right\} = \exp\left\{\frac{\sigma_i^2}{M^2} \sum_{k=2}^{\infty} \frac{c^k M^k}{k!}\right\} = \exp\left\{\frac{\sigma_i^2}{M^2} (e^{cM} - 1 - cM)\right\}$$

ולכן :

$$\begin{aligned} P\left(\frac{1}{m} \sum_{i=1}^m \xi_i \geq \epsilon\right) &\leq e^{-cm\epsilon} \prod_{i=1}^m \mathbf{E}(e^{c\xi_i}) \quad (2.2.2.2) \\ &\leq \exp\left\{-mc\epsilon + \frac{e^{cM} - 1 - cM}{M^2} \Sigma^2\right\} \end{aligned}$$

נקח את $c = \frac{1}{M} \ln(1 + \frac{Mm\epsilon}{\Sigma^2})$ שממוזער את צד ימין ב (2.2.2.2),

■

נציב ב (2.2.2.2) ונקבל את תוצאת המשפט.

הערה 2.2.1 אם נגדיר : $g(\lambda) = (1 + \lambda) \ln(1 + \lambda) - \lambda$ נוכל לרשום את אישויון בנט בצורה :

$$(2.2.2.3) \quad P\left(\frac{1}{m} \sum_{i=1}^m [\xi_i - \mu_i] \geq \epsilon\right) \leq \exp\left\{-\frac{\Sigma^2}{M^2} g\left(\frac{mM\epsilon}{\Sigma^2}\right)\right\}$$

נבחין כי :

$$1. \quad g(0) = 0$$

$$2. \quad g'(\lambda) = \ln(1 + \lambda) \Rightarrow g'(0) = 0$$

$$3. \quad g''(\lambda) = \frac{1}{1+\lambda}$$

משפט 2.2.5 (ברנשטיין)

יהיו $\{\xi_i\}_{i=1}^m$ משתנים מקריים ב"ת כך ש: $\mathbf{E}(\xi_i) = \mu_i \quad \text{var}(\xi_i) = \sigma_i^2$

וגם $|\xi_i - \mu_i| \leq M$ אזי לכל $\epsilon > 0$ מתקיים:

$$P\left(\frac{1}{m} \sum_{i=1}^m [\xi_i - \mu_i] \geq \epsilon\right) \leq \exp\left\{-\frac{m^2 \epsilon^2}{2(\Sigma^2 + \frac{1}{3}mM\epsilon)}\right\}$$

$$\Sigma^2 = \sum_{i=1}^m \sigma_i^2 \quad \text{כאשר}$$

הוכחה: נראה כי

$$g(\lambda) \geq \frac{3\lambda^2}{6 + 2\lambda}$$

לכל $\lambda \geq 0$. התוצאה נובעת ישירות אחרי הצבה ב (2.2.2.3) כי

$$-\frac{\Sigma^2}{M^2} g\left(\frac{mM\epsilon}{\Sigma^2}\right) \leq -\frac{\Sigma^2}{M^2} \frac{3\left(\frac{mM\epsilon}{\Sigma^2}\right)^2}{6 + 2\frac{mM\epsilon}{\Sigma^2}} = -\frac{m^2 \epsilon^2}{2(\Sigma^2 + \frac{1}{3}mM\epsilon)}$$

נגדיר :

$$h(\lambda) = (6 + 2\lambda)g(\lambda) - 3\lambda^2$$

ונרצה להראות ש $h(\lambda) \geq 0$ לכל $\lambda \geq 0$.

נראה כי $\lambda_0 = 0$ היא נקודת מינימום של h :

$$h'(\lambda) = 2g(\lambda) + (6 + 2\lambda)g'(\lambda) - 6\lambda \xrightarrow{g(0)=g'(0)=0} h'(0) = 0$$

$$\begin{aligned}
h''(\lambda) &= 2g'(\lambda) + (6 + 2\lambda)g''(\lambda) + 2g'(\lambda) - 6 \\
&= 4\ln(1 + \lambda) + \frac{6 + 2\lambda}{1 + \lambda} - 6 \\
&= \frac{4}{1 + \lambda} [\ln(1 + \lambda)(1 + \lambda) + 1.5 + 0.5\lambda - 1.5(1 + \lambda)] \\
&= \frac{4}{1 + \lambda} g(\lambda) \geq 0
\end{aligned}$$

מכיוון שהנגזרת השנייה חיובית לכל $\lambda \geq 0$, הפונקציה h קמורה ולכן הנקודה $\lambda_0 = 0$ היא נקודת מינימום גלובלית ו $h(0) = 0$, לכן $h(\lambda)$ חיובית לכל $\lambda \geq 0$ ■

טענה 2.2.1 (הפדינג)

יהיו $\{\xi_i\}_{i=1}^m$ משתנים מקריים ב"ת כך ש $\mathbf{E}(\xi_i) = \mu_i$ וגם $|\xi_i - \mu_i| \leq M$. אזי לכל $\epsilon > 0$ מתקיים:

$$P\left(\frac{1}{m} \sum_{i=1}^m [\xi_i - \mu_i] \geq \epsilon\right) \leq \exp\left\{-\frac{m\epsilon^2}{2M^2}\right\}$$

הוכחה: כמו בהוכחת אי־שוויון בנט נשתמש ב (2.2.2.1). מכיוון ש $f(y) = e^y$ היא פונקציה קמורה וגם $-cM \leq y \leq cM$, האינטרפולציה הליניארית

$$\ell(y) = f(-cM) \cdot \frac{cM - y}{2cM} + f(cM) \cdot \frac{y - (-cM)}{2cM}$$

המחברת בין שתי הנקודות

$$(-cM, f(-cM)), (cM, f(cM))$$

תקיים $f(y) \leq \ell(y)$ לכל $y \in [-cM, cM]$

עבור $y = c\xi_i$ נקבל:

$$f(c\xi_i) = e^{c\xi_i} \leq \frac{c\xi_i - (-cM)}{2cM} e^{cM} + \frac{cM - c\xi_i}{2cM} e^{-cM}$$

ולכן

$$(\mathbf{E}(\xi_i) = 0) \implies \mathbf{E}(e^{c\xi_i}) \leq \frac{1}{2}e^{-cM} + \frac{1}{2}e^{cM}$$

ע"י פיתוח טיילור נקבל:

$$\begin{aligned}
\frac{1}{2}e^{-cM} + \frac{1}{2}e^{cM} &= \frac{1}{2} \sum_{k=0}^{\infty} \frac{(-cM)^k}{k!} + \frac{1}{2} \sum_{k=0}^{\infty} \frac{(cM)^k}{k!} = \sum_{j=0}^{\infty} \frac{(cM)^{2j}}{(2j)!} \\
&= \sum_{j=0}^{\infty} \frac{((cM)^2/2)^j}{j!} \prod_{\ell=1}^j \frac{1}{2\ell - 1} \leq \sum_{j=0}^{\infty} \frac{((cM)^2/2)^j}{j!} = \exp\left\{\frac{(cM)^2}{2}\right\}
\end{aligned}$$

ולכן נקבל כי

$$P\left(\frac{1}{m} \sum_{i=1}^m \xi_i \geq \varepsilon\right) \leq e^{-mc\varepsilon} \prod_{i=1}^m \mathbf{E}(e^{c\xi_i}) \leq \exp\left\{-cm\varepsilon + \frac{m(cM)^2}{2}\right\}$$

נבחר את c שממזער את צד ימין $c^* = \frac{\varepsilon}{M^2}$.

נציב ונקבל את מה שרצינו.

הערה 2.2.2 יכולנו להראות את הטענות הנ"ל עבור " $\frac{1}{m} \sum_{i=1}^m [\xi_i - \mu_i] \leq -\varepsilon$ " בדיוק באותה צורה ולקבל את אותו חסם. ולכן נוכל לחסום את

$$P\left(\left|\frac{1}{m} \sum_{i=1}^m [\xi_i - \mu_i]\right| \geq \varepsilon\right)$$

ע"י פעמיים החסם שמופיע בטענות הקודמות.

פרק 3

מבוא ללמידה מונחית

למידה מונחית הוא סוג של למידת מכונה שבו האלגוריתם מקבל קבוצה של דוגמאות $\{x^{(i)}, y^{(i)}\}_{i=1}^m$ ובעזרת דוגמאות אלה הוא מתבקש לפתח את היכולת לחזות את הערך y בהינתן x .

כל דוגמה i מכילה ייצוג של אובייקט כלשהו ע"י וקטור של ערכים $x = [x_1, x_2, \dots, x_n]$ כאשר x_i מייצג תכונה מסוימת של האובייקט ועבור כל דוגמה כזאת הוא מקבל גם כן את "התשובה הנכונה" או המטרה של האלגוריתם y . אם, למשל, המשימה היא לחזות מחירים של דירות, x יכול להכיל נתונים על גודל, מיקום, מספר חדרים וכו'. y הוא מחיר הדירה.

בעזרת נתונים אלה האלגוריתם מתבקש להחזיר פונקציה $f(x)$ שבעזרתה אפשר יהיה לחזות את הערך של y . כאשר $y \in \mathbb{R}$ כמו בדוגמא זו הבעיה נקראת בעיית רגרסיה וכאשר $y \in \{1, 2, \dots, k\}$ אז זו בעיית קלסיפיקציה שבה האלגוריתם מתבקש להחליט לאיזה מ k הקטגוריות האובייקט x שייך.

כדי שנוכל לאמוד את רמת הדיוק של f אנחנו צריכים ערך מספרי שייצג לנו את גודל השגיאה המתקבלת מהשימוש ב f כדי לחזות את y . למשל בבעיות רגרסיה בד"כ אומדים את השגיאה עבור דוגמה (x, y) ע"י

$$Err(x, y) = (f(x) - y)^2$$

ובבעיות קלסיפיקציה

$$Err(x, y) = \begin{cases} 1, & f(x) \neq y \\ 0, & \text{אחרת} \end{cases}$$

מכיוון שהאלגוריתם רואה רק את אוסף הדוגמאות שקיבל, נרצה בדרך כלל שהאלגוריתם יבחר פונקציה f שממזערת את השגיאות על דוגמאות הלמידה, כלומר נרצה למזער את הגודל

$$\frac{1}{m} \sum_{i=1}^m Err(x^{(i)}, y^{(i)})$$

כאשר $(x^{(i)}, y^{(i)})$ היא הדוגמה ה- i .

עד כה הבעיה נראית כמו בעיית אופטימיזציה של השגיאה אבל ההבדל הוא שבלמידת מכונה יש חשיבות קריטית ליכולת של הפונקציה שהאלגוריתם מחזיר לבצע תחזיות מוצלחות על דוגמאות חדשות שהוא לא ראה.

3.1 דוגמאות הלמידה והבדיקה

קלט אנחנו מקבלים אוסף דוגמאות $\{x^{(i)}, y^{(i)}\}_{i=1}^m$, שאותו נחלק לשתי קבוצות :

1. דוגמאות שישמשו את האלגוריתם ללמידה (training set).
 2. דוגמאות בדיקה שהאלגוריתם לא יראה בתהליך הלמידה (test set), שישמשו אותנו לבדיקת איכות התחזיות של האלגוריתם.
- נזכור כי האלגוריתם יבחר את הפונקציה f כך שהשגיאה על דוגמאות הלמידה היא מינימלית אבל מה שחשוב לנו הוא איכות התחזיות על דוגמאות הבדיקה (כאומדן לשגיאה האמיתית). השאלה המרכזית בלמידת מכונה היא איך בעזרת תכנון של תהליך הלמידה נוכל להשפיע על השגיאה שהאלגוריתם יבצע על דוגמאות הבדיקה, כפי שנראה בהמשך יש מתח מרכזי בתהליך הלמידה בין :

1. הקטנת השגיאה על דוגמאות הלמידה.
 2. שמירה על פער קטן בין השגיאה על דוגמאות הלמידה לבין השגיאה על דוגמאות הבדיקה.
- בספרות המתח הזה נקרא מתח בין תת-התאמה (underfitting) להתאמת-יתר (overfitting) כאשר תת-התאמה מתייחס למקרה שבו האלגוריתם לא מצליח להקטין את השגיאה על דוגמאות הלמידה, התאמת-יתר מתייחס למקרה שבו הפער בין השגיאה על דוגמאות הלמידה לשגיאה על דוגמאות הבדיקה הוא גדול.

3.2 למידה כבעיית חיפוש במרחב פונקציות $\mathcal{H}$

אלגוריתם הלמידה מתבקש להחזיר פונקציה $f : [0, 1]^d \rightarrow \mathbb{R}$, ולשם כך יש להגדיר מרחב פונקציות $\mathcal{H}$ שממנו הוא רשאי לבחור את f . אחת הדרכים לאזן בין תת-התאמה להתאמת-יתר היא בחירת גודל המרחב או במילים אחרות עד כמה מורכבות הפונקציות שהמרחב מכיל.

ככל שהמרחב גדול יותר, לאלגוריתם יש יכולת בחירה בין יותר פונקציות ולכן הסבירות שנגיע למצב שבו יש התאמת-יתר עולה, לעומת זאת אם מרחב הפונקציות מכיל רק פונקציות פשוטות אז הסבירות לתת-התאמה עולה.

בד"כ כל פונקציה במרחב $\mathcal{H}$ תהיה מוגדרת על ידי מערך סופי של פרמטרים ולכן חיפוש הפונקציה הטובה ביותר מתבצע ע"י מציאת מערך הפרמטרים שימזער את השגיאה על דוגמאות הלמידה כלל האפשר.

3.3 דוגמה - רגרסיה ליניארית

רגרסיה ליניארית היא אולי האלגוריתם הפשוט ביותר עבור פתרון של בעיית רגרסיה, שבה המודל מניח כי $y^{(i)} = w^T x^{(i)} + \varepsilon_i$ כאשר $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ משתנים מקריים בלתי תלויים.

כלומר כתחזית ל y המודל יחזיר פונקציה ליניארית של הקלט $\hat{y} = w^T x$ כאשר $w \in \mathbb{R}^d$ הוא וקטור של פרמטרים.

בהינתן קבוצת דוגמאות למידה $\{x^{(i)}, y^{(i)}\}_{i=1}^m$ נוכל למדוד את השגיאה של המודל ע"י השגיאה הריבועית הממוצעת (MSE):

$$MSE = \frac{1}{m} \sum_{i=1}^m (y^{(i)} - \hat{y}^{(i)})^2$$

$$\text{ואם נסמן } y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix}, \hat{y} = \begin{bmatrix} \hat{y}^{(1)} \\ \vdots \\ \hat{y}^{(m)} \end{bmatrix} \text{ נוכל לרשום}$$

$$MSE = \frac{1}{m} \|y - \hat{y}\|_2^2$$

$$\text{אם נגדיר מטריצה } X \text{ שבה כל שורה היא דוגמה כלומר } X = \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(m)} \end{bmatrix} \text{ נוכל לרשום}$$

$$MSE = \frac{1}{m} \|y - Xw\|_2^2$$

כך שכדי למצוא את הפונקציה בעלת השגיאה הנמוכה ביותר במרחב הפונקציות הליניאריות $\mathcal{H} = \{w^T x \mid w \in \mathbb{R}^d\}$ נמצא את סט הפרמטרים w כך שהשגיאה הריבועית הממוצעת על דוגמאות הלמידה היא מינימלית.

הערה 3.3.1 בד"כ מודל רגרסיה ליניארית מתייחס למודל מעט יותר מורכב $\hat{y} = w^T x + b$

כאשר $b \in \mathbb{R}$, אבל קל לראות שאם מגדירים $x' = \begin{bmatrix} 1 \\ x \end{bmatrix}$, $w' = \begin{bmatrix} b \\ w \end{bmatrix}$ נגיע לאותה נוסחה

$$\hat{y} = w'^T x'$$

הערה 3.3.2 נבחין כי רגרסיה ליניארית היא פונקציה ליניארית של w כך שאנו יכולים בעזרת מודל זה לייצר פונקציות לא ליניאריות ביחס ל x . למשל במקרה הדו-מימדי נוכל

לקבל קשר פולינומי ע"י כך שבמקום להשתמש ב $x = \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix}$ אפשר לייצר משתנה חדש

$x' = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1 x_2 \\ x_1^2 \\ x_2^2 \end{bmatrix}$ ואז הנוסחה $w^T x'$ מייצגת קשר פולינומי מסדר 2. נבחין כי d גדל כאשר סדר הפולינום גדל.

3.4 ריגורליזציה - קנס על הנורמה

עד כה הצגנו דרך אחת שבה אנחנו יכולים להשפיע על התנהגות האלגוריתם והיא בחירת מרחב הפונקציות $\mathcal{H}$ אבל בבעיות אמיתיות לפעמים בלתי אפשרי לדעת מראש איזה מרחב פונקציות לבחור כך שנקבל מודל בעל דיוק מספק.

ריגורליזציה היא דרך שבה אנו אומרים לאלגוריתם שאנחנו מעדיפים פונקציות מסוימות מהמרחב על פני אחרות, בד"כ הפונקציות המועדפות הן פונקציות פשוטות יותר כך שריגורליזציה מורידה את הסבירות למצב שיש בו התאמת יתר.

לדוגמה, עבור רגרסיה ליניארית נוכל לשנות את הבעיה כך שהאלגוריתם יעדיף w עם נורמה קטנה יותר ע"י כך שבמקום למצוא w שימזער את

$$(3.4.0.1) \quad MSE(w) = \frac{1}{m} \|y - Xw\|_2^2$$

נבחר את w שימזער את

$$(3.4.0.2) \quad MSE_{reg}(w) = MSE(w) + \lambda \|w\|_2^2$$

כאשר $\lambda \geq 0$ זה פרמטר שאנחנו בוחרים לפני תחילת תהליך הלמידה. כאשר $\lambda = 0$ נחזור לבעיה המקורית כך שאין לנו עדיפות לפתרון עם נורמה קטנה וככל שנגדיל את λ האלגוריתם יבחר פתרון עם נורמה קטנה יותר. הריגורליזציה שתיארנו נקראת בספרות רגורליזציה $\mathcal{L}^2$ בגלל השימוש בנורמה $\|\cdot\|_2$.

לרגרסיה ליניארית עם ריגורליזציה $\mathcal{L}^2$ קוראים "Ridge regression".

לסיכום, ריגורליזציה היא הוספת פונקצית קנס $\Omega(w)$ לפונקצית השגיאה על דוגמאות הלמידה.

הערה 3.4.1 עבור רגרסיה ליניארית אפשר להראות [9] כי בעיית האופטימיזציה (3.4.0.2) שקולה לבעיית אופטימיזציה עם אילון על הנורמה

$$\min_{w: \|w\|_2 \leq R} MSE(w)$$

עבור $R > 0$ כלשהו. וכי הגדלת הערך λ תגרום לכך שהנורמה קטנה יותר כלומר R קטן יותר.

הערה 3.4.2 ישנן שיטות נוספות לביצוע ריגולרזציה שאנו לא מציגים פה, רגולרזציה עם קנס על הנורמה תהיה שימושית ביישום התיאוריה שנציג בפרק הבא ולכן בחרנו להציג אותה.

פרק 4

למידה סטטיסטית

בפרק זה נציג חלק מהתיאוריה של למידה סטטיסטית עבור בעיות רגרסיה, שמראה כי תחת הנחות מסוימות למידה היא אפשרית. כלומר שבעזרת דוגמאות הלמידה אפשר להסיק מסקנות על איכות התחזיות של המודל עבור דוגמאות חדשות. בעזרת התיאוריה והמשפטים שנציג נראה איך בחירה של המודל והפרמטרים שלו תשפיע על הלמידה ובפרט על המתח בין תהיכתאמה להתאמתית.

4.1 הגדרה פורמלית של למידה

למידה סטטיסטית היא מסגרת שבה אפשר לבחון את היעילות של תהליך הלמידה שתיארנו קודם, על ידי הנחה לגבי המודל שיוצר את הנתונים (דוגמאות למידה/בדיקה):

1. קיימת פונקציה צפיפות משותפת $\rho_{\mathbf{x}}$ שאינה ידועה ללומד וממנה נדגמו הוקטורים x באופן בלתי תלוי.

2. עבור כל וקטור x המציאות מחזירה ערך y לפי $\rho_{y|x}(y|x)$ שגם היא אינה ידועה.

בעית הלמידה היא בחירה של פונקציה $h(x)$ מתוך מרחב פונקציות $\mathcal{H}$ כך שהתחזיות קרובות ככל האפשר לערכים שהמציאות מחזירה. בחירת הפונקציה מתבצעת בעזרת דוגמאות הלמידה שמהוות מדגם של m וקטורים מקריים בלתי תלויים $\{(x^{(i)}, y^{(i)})\}_{i=1}^m$ שנדגמו מ $\rho_{\mathbf{x}, y}(x, y) = \rho_{\mathbf{x}}(x)\rho_{y|x}(y|x)$.

כדי לבחור את הפונקציה שהתחזיות שלה הכי קרובות למציאות אנו אומדים את השגיאה המתקבלת מהשימוש ב f עבור קלט x כדי לחזות את y על ידי $Err(y, f(x))$ ולכן המטרה היא למצוא את הפונקציה מתוך $\mathcal{H}$ שממזערת את תוחלת השגיאות

$$\mathcal{E}(f) = \mathbf{E}[Err(y, f(x))] = \int Err(y, f(x))\rho_{\mathbf{x}, y}(x, y)dx dy$$

הערה 4.1.1 בעבודה זו נניח בלי הגבלת הכלליות כי $x \in [0, 1]^d$ מכיוון שמבחינה מעשית

גם כאשר המשתנים מקבלים ערכים אחרים בד"כ מנרמלים את המשתנים שיהיו בטווח $[0, 1]$ כדי לאפשר התכנסות יותר מהירה של אלגוריתמי אופטימיזציה נומריים.

עבור בעיות רגרסיה נהוג להשתמש ב $Err(y, f(x)) = (f(x) - y)^2$ ולכן כדי לאמוד את השגיאה של המודל נשתמש ב

$$(4.1.0.1) \quad \mathcal{E}(f) = \mathbf{E}(f(\mathbf{x}) - \mathbf{y})^2 = \int_{[0,1]^d \times \mathbb{R}} (f(x) - y)^2 \rho_{\mathbf{x}, \mathbf{y}}(x, y) dx dy$$

4.1.1 פונקצית הרגרסיה

נגדיר :

$$f_\rho(x) = E[\mathbf{y} | \mathbf{x} = x] = \int_{-\infty}^{\infty} y \rho_{\mathbf{y} | \mathbf{x}}(y | x) dy$$

ונראה כי בהינתן $\mathcal{E}(f)$ שבחרנו, פונקציה זו היא הטובה ביותר שנוכל למצוא אם היינו יודעים את ההתפלגות $\rho_{\mathbf{x}, \mathbf{y}}$.

טענה 4.1.1 לכל $f : [0, 1]^d \rightarrow \mathbb{R}$ מתקיים :

$$\mathcal{E}(f) = \mathbf{E}_{\mathbf{x}} [(f(\mathbf{x}) - f_\rho(\mathbf{x}))^2] + \mathcal{E}(f_\rho)$$

הוכחה:

מנוסחת התוחלת השלמה (2.2.1) נקבל כי

$$(4.1.1.1) \quad \mathcal{E}(f) = \mathbf{E}(f(\mathbf{x}) - \mathbf{y})^2 = \mathbf{E}_{\mathbf{x}}[\mathbf{E}_{\mathbf{y} | \mathbf{x}}[(f(\mathbf{x}) - \mathbf{y})^2 | \mathbf{x}]]$$

נקבע $x \in [0, 1]^d$ ונקבל

$$\begin{aligned} \mathbf{E}_{\mathbf{y} | \mathbf{x}}[(f(x) - \mathbf{y})^2 | \mathbf{x} = x] &= \mathbf{E}[(f(x) - f_\rho(x) + f_\rho(x) - \mathbf{y})^2 | \mathbf{x} = x] \\ &= \mathbf{E}[(f(x) - f_\rho(x))^2 | \mathbf{x} = x] + \mathbf{E}[(f_\rho(x) - \mathbf{y})^2 | \mathbf{x} = x] \\ &\quad + 2 \mathbf{E}[(f(x) - f_\rho(x))(f_\rho(x) - \mathbf{y}) | \mathbf{x} = x] \\ &\stackrel{(*)}{=} \mathbf{E}[(f(x) - f_\rho(x))^2 | \mathbf{x} = x] + \mathbf{E}[(f_\rho(x) - \mathbf{y})^2 | \mathbf{x} = x] \end{aligned}$$

$$\begin{aligned} (*) \mathbf{E}[(f(x) - f_\rho(x))(f_\rho(x) - \mathbf{y}) | \mathbf{x} = x] &= (f(x) - f_\rho(x)) E[f_\rho(x) - \mathbf{y} | \mathbf{x} = x] \\ &= (f(x) - f_\rho(x))(f_\rho(x) - E[\mathbf{y} | \mathbf{x} = x]) \\ &= 0 \end{aligned}$$

עכשיו נציב בחזרה ב equation 4.1.1.1

ונקבל :

$$\begin{aligned}\mathcal{E}(f) &= \mathbf{E}_{\mathbf{x}}[E_{\mathbf{y}|\mathbf{x}}[(f(\mathbf{x}) - \mathbf{y})^2|\mathbf{x}]] \\ &= \mathbf{E}_{\mathbf{x}}[\mathbf{E}[(f(x) - f_{\rho}(x))^2|\mathbf{x} = x]] + \mathbf{E}_{\mathbf{x}}[\mathbf{E}[(f_{\rho}(x) - \mathbf{y})^2|\mathbf{x} = x]] \\ &= \mathbf{E}_{\mathbf{x}}[(f(\mathbf{x}) - f_{\rho}(\mathbf{x}))^2] + \mathcal{E}(f_{\rho})\end{aligned}$$

■

מסקנה ישירה מהטענה: הפונקציה $f_{\rho}(x)$ היא הפונקציה שנותנת לנו את השגיאה המינימלית, כלומר:

$$f_{\rho}(x) = \arg \min_f \mathcal{E}(f)$$

ולכן היא מבחינתנו "פונקציית המטרה".

הערה 4.1.2 ישנם אלגוריתמים שמנסים לאמוד ישירות את $f_{\rho}(x)$ בעזרת דוגמאות הלמידה, למשל אלגוריתם KNN (k -השכנים הקרובים ביותר) שבהינתן x במקום להחזיר ממוצע של ה $y^{(i)}$ עם $x^{(i)} = x$ (מכיוון שבד"כ יש מעט מאוד נקודות אם בכלל כך ש $x^{(i)} = x$), משתמשים ב:

$$\hat{f}(x) = \frac{1}{k} \sum_{i=1}^k \left(y^{(i)} | x^{(i)} \in N_k(x) \right)$$

כאשר $N_k(x)$ זו סביבה של x המכילה את k הנקודות הכי קרובות ל x מהדוגמאות למידה $x^{(i)}$.

4.2 חסם לשגיאה עבור $\mathcal{H}$ סופי

עכשיו נסתכל על דוגמאות הלמידה $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^m$ ונגדיר :

$$\mathcal{E}_{\mathcal{D}}(f) = \frac{1}{m} \sum_{i=1}^m (f(x^{(i)}) - y^{(i)})^2$$

שמייצג את השגיאה הממוצעת שהמודל מבצע על דוגמאות הלמידה (לכן התלות ב $\mathcal{D}$).

נביט בביטוי $|\mathcal{E}(f) - \mathcal{E}_{\mathcal{D}}(f)|$ שמודד את ההבדל בשגיאה שהמודל מבצע על דוגמאות הלמידה לעומת השגיאה הכללית, מכיוון שאין לנו גישה לחשב את השגיאה הכללית של המודל נרצה למצוא חסם לביטוי הזה.

משפט 4.2.1 יהי $M > 0$ ותהי $f : [0, 1]^d \rightarrow \mathbb{R}$ פונקציה כלשהי כך ש :

$\max_{x,y} |f(x) - y| \leq M$, ויהי $\mathcal{D}$ אוסף דוגמאות למידה בגודל m שהתקבל ע"י דגימה מפונקצית הצפיפות המשותפת $\rho_{\mathbf{x},y}$, אזי לכל $\varepsilon > 0$ מתקיים :

$$(4.2.0.1) \quad P(|\mathcal{E}(f) - \mathcal{E}_{\mathcal{D}}(f)| \geq \varepsilon) \leq 2 \exp \left\{ -\frac{m\varepsilon^2}{2M^4} \right\}$$

$$(4.2.0.2) \quad P(|\mathcal{E}(f) - \mathcal{E}_{\mathcal{D}}(f)| \geq \varepsilon) \leq 2 \exp \left\{ -\frac{m\varepsilon^2}{2(\sigma^2 + \frac{1}{3}M^2\varepsilon)} \right\}$$

כאשר $\sigma^2 = \text{Var} [(f(\mathbf{x}) - y)^2]$

הערה 4.2.1 זה המשפט הבסיסי שבעזרתו נראה כי למידה היא אפשרית, נשיב לב לאופי ההסתברותי של המשפט, שאומר שבהסתברות גבוהה ההבדל בין השגיאה על דוגמאות הלמידה לשגיאה האמיתית יהיה קטן כרצוננו (בהנחה שניקח מספר דוגמאות למידה גדול מספיק). בספרות הוכחת הלמידה בצורה הזו נקראת

PAC - Probably, Approximately Correct

הוכחה: מכיוון שדוגמאות הלמידה $\mathcal{D}$ נדגמו מפונקצית הצפיפות המשותפת $\rho_{\mathbf{x},y}$ נקבל m וקטורים מקריים ב"ת $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$, $i = 1, 2, \dots, m$. נגדיר $\xi_i = (f(\mathbf{x}^{(i)}) - \mathbf{y}^{(i)})^2$ ונבחין כי :

$$1. \quad \forall i = 1, 2, \dots, m, \quad \mathbf{E}(\xi_i) = \mathcal{E}(f)$$

$$2. \quad \frac{1}{m} \sum_{i=1}^m \xi_i = \mathcal{E}_{\mathcal{D}}(f)$$

$$3. \quad \text{Var}(\xi_i) = \text{Var} [(f(\mathbf{x}) - \mathbf{y})^2] = \sigma^2$$

$$4. \quad \sum_{i=1}^m \text{Var}(\xi_i) = m\sigma^2$$

מהנתון $|f(x) - y| \leq M$ נובע כי $\xi_i \in [0, M^2]$, $\mathbf{E}(\xi_i) \in [0, M^2]$ ולכן

$$|\xi_i - \mathbf{E}(\xi_i)| \leq M^2$$

כדי להוכיח את (4.2.0.2) נציב $\xi_i = (f(\mathbf{x}^{(i)}) - \mathbf{y}^{(i)})^2$ באי-שוויון ברנשטיין (2.2.5) ונקבל

$$\begin{aligned}
P(|\mathcal{E}(f) - \mathcal{E}_{\mathcal{D}}(f)| \geq \varepsilon) &= P\left(\left|\frac{1}{m} \sum_{i=1}^m [\xi_i - \mathbf{E}(\xi_i)]\right| \geq \varepsilon\right) \\
&= P\left(\left|\frac{1}{m} \sum_{i=1}^m \xi_i - \mathbf{E}(\xi)\right| \geq \varepsilon\right) \\
&\leq 2 \exp\left\{-\frac{m^2 \varepsilon^2}{2(\Sigma^2 + \frac{1}{3}mM^2\varepsilon)}\right\} \\
&= 2 \exp\left\{-\frac{m^2 \varepsilon^2}{2(m\sigma^2 + \frac{1}{3}mM^2\varepsilon)}\right\} \\
&= 2 \exp\left\{-\frac{m\varepsilon^2}{2(\sigma^2 + \frac{1}{3}M^2\varepsilon)}\right\}
\end{aligned}$$

כדי להוכיח את (4.2.0.1) נשתמש באופן דומה באי־שוויון הפדינג (2.2.1) ונקבל כי

$$\begin{aligned}
P\left(\left|\frac{1}{m} \sum_{i=1}^m \xi_i - \mathbf{E}(\xi)\right| \geq \varepsilon\right) &\leq 2 \exp\left\{-\frac{m\varepsilon^2}{2(M^2)^2}\right\} \\
&= 2 \exp\left\{-\frac{m\varepsilon^2}{2M^4}\right\}
\end{aligned}$$

■

המשפט הנ"ל מוצא חסם עבור פונקציה מסויימת f אבל נזכור כי בבעיה שלנו אנו מחפשים את הפונקציה $\hat{f}$ שתמזער את השגיאה על דוגמאות הלמידה מתוך מרחב פונקציות $\mathcal{H}$.

נניח לרגע שמרחב הפונקציות הוא סופי, כלומר $\mathcal{H} = \{f_1, f_2 \dots f_N\}$. נרצה למצוא חסם לשגיאה הכללית של המודל הנלמד $\hat{f}$, כלומר חסם לביטוי $P\left(\left|\mathcal{E}(\hat{f}) - \mathcal{E}_{\mathcal{D}}(\hat{f})\right| \geq \varepsilon\right)$. אבל מכיוון שאנו לא יכולים לדעת איזה $f \in \mathcal{H}$ יתן את השגיאה המינימלית על דוגמאות הלמידה ננסה לחסום את $P(|\mathcal{E}(f) - \mathcal{E}_{\mathcal{D}}(f)| \geq \varepsilon)$ לכל $f \in \mathcal{H}$.

נבחין כי

$$\begin{aligned}
\text{" } \sup_{f \in \mathcal{H}} |\mathcal{E}(f) - \mathcal{E}_{\mathcal{D}}(f)| \geq \varepsilon \text{ " } &\iff \text{" } |\mathcal{E}(f_1) - \mathcal{E}_{\mathcal{D}}(f_1)| \geq \varepsilon \\
&\text{or } |\mathcal{E}(f_2) - \mathcal{E}_{\mathcal{D}}(f_2)| \geq \varepsilon \\
&\dots \\
&\text{or } |\mathcal{E}(f_N) - \mathcal{E}_{\mathcal{D}}(f_N)| \geq \varepsilon \text{ "}.
\end{aligned}$$

לכן מתקיים :

$$\begin{aligned}
P\left(\sup_{f \in \mathcal{H}} |\mathcal{E}(f) - \mathcal{E}_{\mathcal{D}}(f)| \geq \varepsilon\right) &\leq \sum_{n=1}^N P(|\mathcal{E}(f_n) - \mathcal{E}_{\mathcal{D}}(f_n)| \geq \varepsilon) \\
&\leq 2N \exp\left\{-\frac{m\varepsilon^2}{2M^4}\right\} \quad (4.2.0.3)
\end{aligned}$$

בחסם הנ"ל יש שלושה ערכים שחשוב שנשים לב אליהם, ε, m , וההסתברות של הטעות, שבעזרת החסם הנ"ל נוכל לחסום כל אחד מהם בעזרת השניים האחרים:

מסקנה 4.2.1 בהינתן m דוגמאות למידה ו $\delta > 0$ מתקיים שבהסתברות לפחות $1 - \delta$:

$$\mathcal{E}_{\mathcal{D}}(\hat{f}) - \sqrt{\frac{2M^4}{m} \ln\left(\frac{2N}{\delta}\right)} < \mathcal{E}(\hat{f}) < \mathcal{E}_{\mathcal{D}}(\hat{f}) + \sqrt{\frac{2M^4}{m} \ln\left(\frac{2N}{\delta}\right)}$$

וזהו חסם שמראה באופן ברור את הקשר בין השגיאה הכוללת לשגיאה על דוגמאות הלמידה.

הוכחה: נוכל לרשום את (3.4.2.0) כך: בהסתברות של לפחות

$$1 - 2N \exp\left\{-\frac{m\varepsilon^2}{2M^4}\right\}$$

מתקיים ש

$$\sup_{f \in \mathcal{H}} |\mathcal{E}(f) - \mathcal{E}_{\mathcal{D}}(f)| < \varepsilon \xrightarrow{(\hat{f} \in \mathcal{H})} \mathcal{E}_{\mathcal{D}}(\hat{f}) - \varepsilon < \mathcal{E}(\hat{f}) < \mathcal{E}_{\mathcal{D}}(\hat{f}) + \varepsilon$$

$$\varepsilon = \sqrt{\frac{2M^4}{m} \ln\left(\frac{2N}{\delta}\right)} \text{ אזי } \delta = 2N \exp\left\{-\frac{m\varepsilon^2}{2M^4}\right\}$$

■

מסקנה 4.2.2 בהינתן ε ו δ גדולים מאפס, אזי אם

$$m > \frac{2M^4}{\varepsilon^2} \ln\left(\frac{2N}{\delta}\right)$$

מתקיים שבהסתברות לפחות $1 - \delta$ השגיאה על הדוגמאות הלמידה קרובה עד כדי ε לשגיאה הכוללת של המודל.

החסמים הנ"ל עוזרים לנו להבין את המתח שקיים בין הגדלים m, δ, ε אבל רואים כי החסמים נהיים יותר גרועים ככל ש N גדל ובבעיות אמיתיות החיפוש הוא במרחב $\mathcal{H}$ אינסופי ולא מתוך קבוצה סופית כמו שהנחנו פה, לכן נרצה לקבל חסמים שאינם תלויים ב N .

4.3 המקרה הכללי

מטרתנו להחליף את הגודל N בחסם (3.4.2.0) בגודל סופי גם עבור $\mathcal{H}$ אינסופי. אחת ההנחות היסודיות בפרק זה היא הקומפקטיות של מרחב הפונקציות $\mathcal{H}$, שמאפשרת לנו לעשות זאת יחסית בקלות.

בתיאוריה והמשפטים שנציג נניח כי $\mathcal{H}$ היא תת־קבוצה קומפקטית של מרחב הפונקציות הרציפות $\mathcal{C}([0, 1]^d, \mathbb{R})$ עם המטריקה $d_\infty(f, g)$.

משפט 4.3.1 יהי $\mathcal{H}$ כמתואר למעלה ו־ $M > 0$ כך שלכל $f \in \mathcal{H}$:

$$\max_{x,y} |f(x) - y| \leq M$$

יהי $\mathcal{D}$ אוסף דוגמאות למידה בגודל m שהתקבל ע"י דגימה מפונקצית הצפיפות המשותפת $\rho_{\mathbf{x}, \mathbf{y}}$, אזי לכל $\varepsilon > 0$ מתקיים:

$$P \left(\sup_{f \in \mathcal{H}} |\mathcal{E}(f) - \mathcal{E}_{\mathcal{D}}(f)| \geq \varepsilon \right) \leq 2\mathcal{N} \left(\mathcal{H}, \frac{\varepsilon}{8M} \right) \exp \left\{ -\frac{m\varepsilon^2}{8M^4} \right\}$$

הערה 4.3.1 באגף ימין של אי־השוויון ניתן לראות את הדמיון ל (3.4.2.0) כאשר ההבדל העיקרי הוא שמחליפים את הגודל N במספר הכיסוי $\mathcal{N}(\mathcal{H}, \frac{\varepsilon}{8M})$ שהוא סופי עבור $\mathcal{H}$ קומפקטי (מסקנה 2.1.1)

נתחיל בהוכחת טענה שתהיה שימושית להוכחת המשפט:

טענה 4.3.1 יהיו $f_1, f_2 : [0, 1]^d \rightarrow \mathbb{R}$ כך שעבור $i = 1, 2$ $|f_i(x) - y| \leq M$ $\forall x, y$ דוגמאות למידה כמו קודם, אזי:

$$(4.3.0.1) \quad |(\mathcal{E}(f_1) - \mathcal{E}_{\mathcal{D}}(f_1)) - (\mathcal{E}(f_2) - \mathcal{E}_{\mathcal{D}}(f_2))| \leq 4M \|f_1 - f_2\|_\infty.$$

הוכחה (טענה 4.3.1) ראשית נבחין כי

$$(*) \quad (f_1(\mathbf{x}) - \mathbf{y})^2 - (f_2(\mathbf{x}) - \mathbf{y})^2 = (f_1(\mathbf{x}) - f_2(\mathbf{x}))(f_1(\mathbf{x}) + f_2(\mathbf{x}) - 2\mathbf{y})$$

ולכן מתקיים ש:

$$\begin{aligned} |\mathcal{E}(f_1) - \mathcal{E}(f_2)| &= |\mathbf{E}[(f_1(\mathbf{x}) - \mathbf{y})^2] - \mathbf{E}[(f_2(\mathbf{x}) - \mathbf{y})^2]| \\ &\stackrel{(*)}{=} |\mathbf{E}[(f_1(\mathbf{x}) - f_2(\mathbf{x}))(f_1(\mathbf{x}) + f_2(\mathbf{x}) - 2\mathbf{y})]| \\ &\leq \mathbf{E}|(f_1(\mathbf{x}) - f_2(\mathbf{x}))(f_1(\mathbf{x}) + f_2(\mathbf{x}) - 2\mathbf{y})| \\ &\leq \|f_1 - f_2\|_\infty \mathbf{E}|(f_1(\mathbf{x}) - \mathbf{y}) + (f_2(\mathbf{x}) - \mathbf{y})| \\ &\leq 2M \|f_1 - f_2\|_\infty \end{aligned} \quad (4.3.0.2)$$

באופן דומה מתקיים ש :

$$\begin{aligned}
|\mathcal{E}_{\mathcal{D}}(f_1) - \mathcal{E}_{\mathcal{D}}(f_2)| &= \frac{1}{m} \left| \sum_{i=1}^m (f_1(\mathbf{x}^{(i)}) - \mathbf{y}^{(i)})^2 - \sum_{i=1}^m (f_2(\mathbf{x}^{(i)}) - \mathbf{y}^{(i)})^2 \right| \\
&\stackrel{(*)}{=} \frac{1}{m} \left| \sum_{i=1}^m (f_1(\mathbf{x}^{(i)}) - f_2(\mathbf{x}^{(i)}))(f_1(\mathbf{x}^{(i)}) + f_2(\mathbf{x}^{(i)}) - 2\mathbf{y}^{(i)}) \right| \\
&\leq \|f_1 - f_2\|_{\infty} \frac{1}{m} \sum_{i=1}^m \left| (f_1(\mathbf{x}^{(i)}) - \mathbf{y}^{(i)}) + (f_2(\mathbf{x}^{(i)}) - \mathbf{y}^{(i)}) \right| \\
&\leq \|f_1 - f_2\|_{\infty} \frac{1}{m} \sum_{i=1}^m 2M \leq 2M \|f_1 - f_2\|_{\infty} \quad (4.3.0.3)
\end{aligned}$$

לכן

$$\begin{aligned}
|(\mathcal{E}(f_1) - \mathcal{E}_{\mathcal{D}}(f_1)) - (\mathcal{E}(f_2) - \mathcal{E}_{\mathcal{D}}(f_2))| &= |(\mathcal{E}(f_1) - \mathcal{E}(f_2)) - (\mathcal{E}_{\mathcal{D}}(f_1) - \mathcal{E}_{\mathcal{D}}(f_2))| \\
&\leq |\mathcal{E}(f_1) - \mathcal{E}(f_2)| + |\mathcal{E}_{\mathcal{D}}(f_1) - \mathcal{E}_{\mathcal{D}}(f_2)| \\
&\leq 4M \|f_1 - f_2\|_{\infty}
\end{aligned}$$

■

הוכחה (משפט 4.3.1) נסמן $\ell = \mathcal{N}(\mathcal{H}, \frac{\varepsilon}{8M})$, ויהיו $f_1, f_2 \dots f_{\ell}$ מרכזי הכדורים הפתוחים ברדיוס $\frac{\varepsilon}{8M}$ שמכסים את $\mathcal{H}$. נסמן

$$S_j = S_{\frac{\varepsilon}{8M}}(f_j), \quad j = 1, 2 \dots \ell.$$

אזי לכל $f \in S_j$ מתקיים :

$$|(\mathcal{E}(f) - \mathcal{E}_{\mathcal{D}}(f)) - (\mathcal{E}(f_j) - \mathcal{E}_{\mathcal{D}}(f_j))| \stackrel{(4.3.0.1)}{\leq} 4M \|f - f_j\|_{\infty} \leq 4M \frac{\varepsilon}{8M} = \frac{\varepsilon}{2}.$$

מכיוון שזה נכון לכל $f \in S_j$ אז נקבל ש :

$$\sup_{f \in S_j} |\mathcal{E}(f) - \mathcal{E}_{\mathcal{D}}(f)| \geq \varepsilon \implies |\mathcal{E}(f_j) - \mathcal{E}_{\mathcal{D}}(f_j)| \geq \frac{\varepsilon}{2}$$

ולכן לכל $j = 1, 2, \dots, \ell$ מתקיים לפי משפט 4.2.1 :

$$\begin{aligned} P\left(\sup_{f \in S_j} |\mathcal{E}(f) - \mathcal{E}_{\mathcal{D}}(f)| \geq \varepsilon\right) &\leq P\left(|\mathcal{E}(f_j) - \mathcal{E}_{\mathcal{D}}(f_j)| \geq \frac{\varepsilon}{2}\right) \\ &\leq 2 \exp\left\{-\frac{m(\frac{\varepsilon}{2})^2}{2M^4}\right\} \\ &= 2 \exp\left\{-\frac{m\varepsilon^2}{8M^4}\right\} \end{aligned} \quad (4.3.0.4)$$

מצד שני מכיוון ש $\mathcal{H} \subseteq S_1 \cup \dots \cup S_j$ אזי

$$\begin{aligned} \text{" } \sup_{f \in \mathcal{H}} |\mathcal{E}(f) - \mathcal{E}_{\mathcal{D}}(f)| \geq \varepsilon \text{" } &\iff \text{" } \sup_{f \in S_1} |\mathcal{E}(f) - \mathcal{E}_{\mathcal{D}}(f)| \geq \varepsilon \\ &\text{or } \sup_{f \in S_2} |\mathcal{E}(f) - \mathcal{E}_{\mathcal{D}}(f)| \geq \varepsilon \\ &\dots \\ &\text{or } \sup_{f \in S_\ell} |\mathcal{E}(f) - \mathcal{E}_{\mathcal{D}}(f)| \geq \varepsilon \text{"}. \end{aligned}$$

ומזה נובע כי :

$$P\left(\sup_{f \in \mathcal{H}} |\mathcal{E}(f) - \mathcal{E}_{\mathcal{D}}(f)| \geq \varepsilon\right) \leq \sum_{j=1}^{\ell} P\left(\sup_{f \in S_j} |\mathcal{E}(f) - \mathcal{E}_{\mathcal{D}}(f)| \geq \varepsilon\right)$$

ובעזרת (4.3.0.4) נקבל :

$$\begin{aligned} &\leq \sum_{j=1}^{\ell} 2 \exp\left\{-\frac{m\varepsilon^2}{8M^4}\right\} \\ \left(\ell = \mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{8M}\right)\right) &= 2\mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{8M}\right) \exp\left\{-\frac{m\varepsilon^2}{8M^4}\right\} \end{aligned}$$

■

הערה 4.3.2 נוכל לרשום את התוצאה האחרונה כך :

$$P(\forall f \in \mathcal{H}. |\mathcal{E}(f) - \mathcal{E}_{\mathcal{D}}(f)| < \varepsilon) \geq 1 - 2\mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{8M}\right) \exp\left\{-\frac{m\varepsilon^2}{8M^4}\right\}$$

או במילים: לכל $\varepsilon > 0$ מתקיים ש $|\mathcal{E}(f) - \mathcal{E}_{\mathcal{D}}(f)| < \varepsilon$ לכל $f \in \mathcal{H}$

בהסתברות לפחות $1 - 2\mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{8M}\right) \exp\left\{-\frac{m\varepsilon^2}{8M^4}\right\}$

המשפט הנ"ל עונה על השאלה הראשונה שחשובה לנו בלמידה, שהיא האם השגיאה הכוללת קרובה לשגיאה על דוגמאות הלמידה. השאלה החשובה השניה היא האם המודל הנלמד (בעל השגיאה המינימלית על דוגמאות הלמידה) $\hat{f} = \arg \min_{f \in \mathcal{H}} \mathcal{E}_{\mathcal{D}}(f)$ הוא בעל שגיאה קרובה לזו של המודל הכי טוב במרחב החיפוש שלנו $f_{\mathcal{H}} = \arg \min_{f \in \mathcal{H}} \mathcal{E}(f)$? לפני שנענה על שאלה זו נראה כי $\hat{f}$ ו $f_{\mathcal{H}}$ קיימים, כלומר נרצה להראות כי לפונקציונלים $\mathcal{E}(f)$, $\mathcal{E}_{\mathcal{D}}(f)$ קיים מינימום כאשר $f \in \mathcal{H}$. לפי (4.3.0.2) ו (4.3.0.3) קל להראות כי $\mathcal{E}(f), \mathcal{E}_{\mathcal{D}}(f) : \mathcal{H} \rightarrow \mathbb{R}$ רציפות ומכיון ש $\mathcal{H}$ קומפקטי, אזי לפי משפט ווירשטראס (2.1.2) הן מקבלות מקסימום ומינימום.

משפט 4.3.2 נניח שתנאי המשפט 4.3.1 מתקיימים, ותהי

$$\hat{f} = \arg \min_{f \in \mathcal{H}} \mathcal{E}_{\mathcal{D}}(f)$$

אזי לכל $\varepsilon > 0$

$$P \left(\left| \mathcal{E}(\hat{f}) - \mathcal{E}(f_{\mathcal{H}}) \right| \leq \varepsilon \right) \geq 1 - 2\mathcal{N} \left(\mathcal{H}, \frac{\varepsilon}{16M} \right) \exp \left\{ -\frac{m\varepsilon^2}{32M^4} \right\}$$

הוכחה:

נסמן

$$f_{\mathcal{H}} = \arg \min_{f \in \mathcal{H}} \mathcal{E}(f)$$

ונבחין כי בהסתברות לפחות $1 - 2\mathcal{N} \left(\mathcal{H}, \frac{\varepsilon}{8M} \right) \exp \left\{ -\frac{m\varepsilon^2}{8M^4} \right\}$ מתקיים :

$$\begin{aligned} \mathcal{E}(\hat{f}) &\leq \mathcal{E}_{\mathcal{D}}(\hat{f}) + \epsilon \\ &\leq \mathcal{E}_{\mathcal{D}}(f_{\mathcal{H}}) + \epsilon \\ &\leq \mathcal{E}(f_{\mathcal{H}}) + 2\epsilon \end{aligned}$$

כאשר השורה הראשונה נובעת ממשפט 4.3.1, השורה השניה נובעת מההגדרה של $\hat{f}$ והשורה השלישית נובעת ממשפט 4.3.1 שוב. לסיום הוכחת המשפט נותר להחליף את ε ב $\varepsilon/2$.

■

פרק 5

שימושים

בפרק זה נראה איך אפשר ליישם את התיאוריה המתמטית על דוגמאות מעשיות כך שנוכל להסיק מסקנות על הדרך הנכונה לבנות מודל יעיל.

גם פה נניח כי הקלט $x \in [0, 1]^d$ ולכן כל מרחבי הפונקציות שיוצגו יהיו תת-מרחב של מרחב הפונקציות הרציפות על $[0, 1]^d$ עם נורמת המקסימום.

5.1 רגרסיה ליניארית

נממש את התיאוריה על מודל רגרסיה ליניארית עם ריגורליזציה $\mathcal{L}^2$ (Ridge Regression) שכמו שראינו קודם היא שקולה לבעיית חיפוש ב $\mathcal{H}_\theta = \{\theta^T x : \|\theta\|_2 \leq R\}$ עבור $R > 0$ כלשהו. נזכור כי מודל רגרסיה ליניארית מניח כי אפשר לקרב את y ע"י פונקציה ליניארית של הקלט $y = \theta^{*T} x + \varepsilon$ כאשר ε הוא מ"מ המייצג את הרעש או הטעות של השימוש ב $w^T x$ כדי לחזות את y . בפרק זה נניח כי $\varepsilon \sim U(-a, a)$.

כדי ליישם את התיאוריה מהפרק הקודם אנחנו צריכים למצוא חסם למספר הכיסוי עבור $\mathcal{H}_\theta$ עם מטריקת הסופרימום.

טענה 5.1.1 יהי $\mathcal{H}_\theta = \{f_\theta(x) = \theta^T x : \theta \in B_{R, \|\cdot\|_2}\}$

אזי

$$(5.1.0.1) \quad \mathcal{N}(\mathcal{H}_\theta, \epsilon) \leq 2^d \left\lceil \frac{R\sqrt{d}}{\epsilon} \right\rceil^d$$

הוכחה:

נבחין כי :

$$\begin{aligned} \|f_\theta - f_{\tilde{\theta}}\|_\infty &= \sup_{x \in [0,1]^d} |\theta^T x - \tilde{\theta}^T x| = \sup_{x \in [0,1]^d} |(\theta - \tilde{\theta})^T x| \\ (5.1.0.2) \quad & \leq \sup_{x \in [0,1]^d} \|\theta - \tilde{\theta}\|_2 \|x\|_2 \\ & \leq \|\theta - \tilde{\theta}\|_2 \sqrt{d} \end{aligned}$$

ומזה נובע

$$\|\theta - \tilde{\theta}\|_2 \leq \frac{\epsilon}{\sqrt{d}} \implies \|f_\theta - f_{\tilde{\theta}}\|_\infty \leq \epsilon$$

ומכיון שכל θ מגדירה פונקציה $f \in \mathcal{H}_\theta$ אחת ויחידה ברור ש:

$$\mathcal{N}(\mathcal{H}_\theta, \epsilon) \leq \mathcal{N}\left(B_{R, \|\cdot\|_\infty}, \frac{\epsilon}{\sqrt{d}}\right)$$

נשתמש בדוגמה 2.1.2 ונציב $\delta = \frac{\epsilon}{\sqrt{d}}$ ב (2.1.6.3):

$$\mathcal{N}\left(B_{R, \|\cdot\|_2}, \frac{\epsilon}{\sqrt{d}}\right) \leq 2^d \left\lceil \frac{R}{\epsilon/\sqrt{d}} \right\rceil^d = 2^d \left\lceil \frac{R\sqrt{d}}{\epsilon} \right\rceil^d$$

לכן

$$\mathcal{N}(\mathcal{H}_\theta, \epsilon) \leq \mathcal{N}\left(B_{R, \|\cdot\|_2}, \frac{\epsilon}{\sqrt{d}}\right) \leq 2^d \left\lceil \frac{R\sqrt{d}}{\epsilon} \right\rceil^d$$

■

כעת כל מה שנותר כדי להשלים את ההוכחה זה למצוא M כך שלכל $f \in \mathcal{H}_\theta$:
 $\max_{x,y} |f(x) - y| \leq M$ נבחין כי

$$\begin{aligned} \max_{x,y} |f(x) - y| &= \max_x |\theta^T x - \theta^{*T} x - \epsilon| \leq \max_x |\theta^T x - \theta^{*T} x| + |\epsilon| \\ &\leq \|\theta - \theta^*\|_2 \sqrt{d} + a \\ &\leq R\sqrt{d} + a \end{aligned}$$

להציב במשפט 4.3.1 את (5.1.0.1) ואת $M = R\sqrt{d} + a$ כך שנקבל :

$$\begin{aligned}
P\left(\left|\mathcal{E}(\hat{f}) - \mathcal{E}_{\mathcal{D}}(\hat{f})\right| \geq \epsilon\right) &\leq 2\mathcal{N}\left(\mathcal{H}_{\theta}, \frac{\epsilon}{8M}\right) \exp\left\{-\frac{m\epsilon^2}{8M^4}\right\} \\
&\leq 2^d \left\lceil \frac{8M}{\epsilon} \right\rceil^d \exp\left\{-\frac{m\epsilon^2}{8M^4}\right\} \\
&\leq 2^d \left\lceil \frac{8(R\sqrt{d} + a)}{\epsilon} \right\rceil^d \exp\left\{-\frac{m\epsilon^2}{8(R\sqrt{d} + a)^4}\right\}
\end{aligned}$$

מכיוון שאי־שוויון זה הוא חסם לפער בין הטעות על דוגמאות הלמידה לטעות האמיתית (בפרט דוגמאות הבדיקה) עבור המודל הנלמד $\hat{f}$ תוצאה זו עוזרת לנו להבין איך פרמטרים שונים של המודל משפיעים על הסבירות שלו להתאמת־יתר :

1. המימד d -

ככל ש d גדל ההסתברות לפער גדול בין השגיאה על דוגמאות הלמידה לשגיאה האמיתית גדל, כלומר הסיכוי למצב של התאמת־יתר גדל.

2. פרמטר הריגורויזציה λ -

ככל שנגדיל את פרמטר הריגורויזציה λ אנחנו מקטינים את R כך שהסיכוי לפער בין הטעויות קטן ולכן הסבירות ל התאמת־יתר קטנה. כמובן אם נקטין את R יותר מדי אז הפונקציה האמיתית לא תהיה במרחב $\mathcal{H}$ ולכן נעבור למצב של תת־התאמה.

3. גודל דוגמאות הלמידה -

ניתן לראות שהתלות ב־ m היא אקספוננציאלית, כך שאגף ימין קטן בצורה מאוד מהירה עם ההגדלה של m . לכן הסבירות להתאמת־יתר קטנה.

4. גודל הרעש - כאשר הרעש גדול יותר, הסבירות להתאמת־יתר גדלה.

5.1.1 סימולציות

בעזרת סימולציות נדגים את המסקנות שהגענו אליהן בפרק זה. עבור כל סימולציה :

1. הדוגמאות יוצרו כך: $x_j \sim U(0, 1), j = 1, 2, \dots, d$ ו $y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$ כאשר $\epsilon \sim U(-a, a)$ ו $\theta_j \sim U(-5, 5)$ נבחרו באופן רנדומלי בכל סימולציה.

יוצרו 10000 דוגמאות, שחולקו לדוגמאות למידה ובדיקה.

2. תהליך הלמידה: מצאנו את מערך הפרמטרים θ שמזער את

$$\min_{\theta: \|\theta\|_2 \leq R} \frac{1}{m} \|y - X\theta\|_2^2$$

על דוגמאות הלמידה בעזרת האלגוריתם הבא:

- נמצא את θ עם הנורמה הקטנה ביותר שפותרת את הבעיה בלי האילוץ ע"י: $\theta = X^+ y$ כאשר X^+ היא המטריצה ההפוכה המוכללת של X .

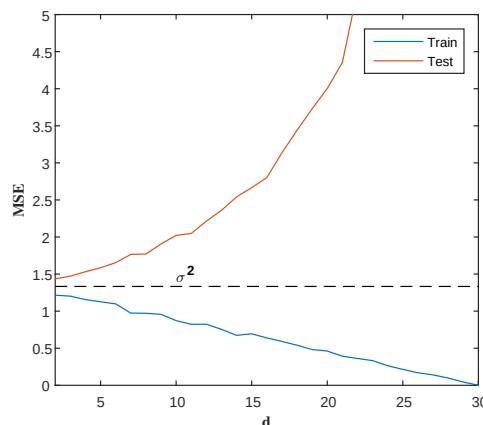
- אם θ שמצאנו מקיימת את האילוץ על הנורמה, כלומר $\|\theta\|_2 \leq R$ סיימנו.
 - אחרת, לפי תנאי KKT: הפתרון יהיה מהצורה $\theta = (X^T X + 2\lambda I)^{-1} X^T y$ והוא צריך לקיים את האילוץ $\|\theta\|_2 = R$. ולכן כדי למצוא את θ שנותנת פתרון אופטימלי עבור הבעיה, נמצא את λ הפותרת את המשוואה:

$$\|(X^T X + 2\lambda I)^{-1} X^T y\|_2 = R.$$

3. חישוב גודל השגיאה: עבור כל סימולציה חישבנו את השגיאה הריבועית הממוצעת על דוגמאות הלמידה ודוגמאות הבדיקה (כקירוב לטעות האמיתית).
 * כל נקודה בגרפים שנציג בהמשך היא ממוצע השגיאות על גבי 200 סימולציות.

1. ההשפעה של המימד d :

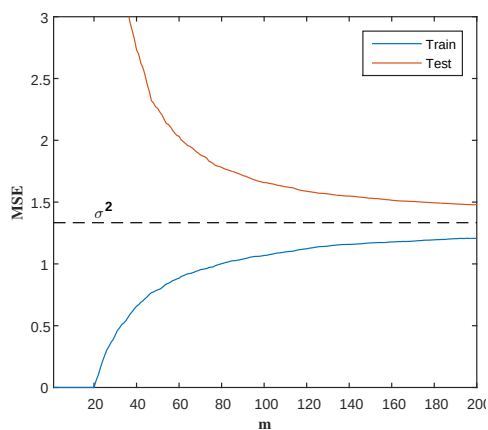
בחרנו $m = 30, a = 2, R = 100$. בחרנו $R = 100$ מכיוון שאנו רוצים לבדוק את ההשפעה של גודל המימד d על הפתרון כאשר כל הפרמטרים האחרים שווים. כמובן שכאשר d גדול יותר אז $\|\theta\|_2$ גדול יותר ולכן כדי להימנע מהשפעות נוספות לא הגבלנו את גודל הנורמה ($R = 100$) אינו מגביל את החיפוש בהינתן הדרך שאנו מייצרים את θ).



- ניתן לראות שהפער בין השגיאות הולך וגדל בקצב אקספוננציאלי ככל שמגדילים את d , כלומר הסיכוי למצב של התאמת יתר הולך וגדל.
 - הטעות על דוגמאות הלמידה הולכת וקטנה עד שמגיעה לאפס שהמודל מצליח ללמוד את הרעש (כאשר $d = m$ יש פתרון שעובר דרך כל הנקודות).

2. ההשפעה של מספר דוגמאות הלמידה m :

עבור $R = 100, a = 2, d = 20$.



- עבור $m \leq d = 20$ המודל יכול ללמוד את הרעש ולכן השגיאה על דוגמאות הלמידה שווה לאפס אבל כמובן שמודל כזה לא יצליח לבצע תחזיות על דוגמאות אחרות.

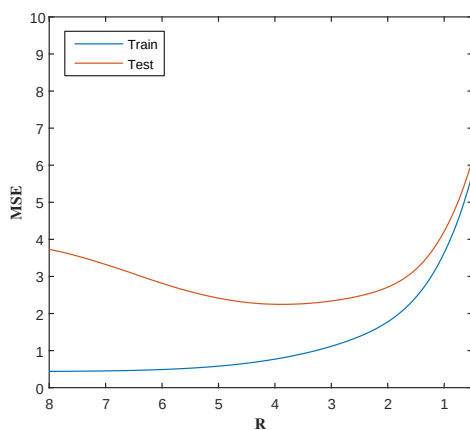
- ככל שנגדיל את גודל דוגמאות הלמידה, הפער בין השגיאה על דוגמאות הלמידה לבין השגיאה האמיתית קטן.

- אפשר לראות שעבור הבעיה שלנו השגיאה הממוצעת מתכנסת לשונות של הרעש, זה נכון עבור הבעיה שלנו מכיוון שהפונקציה האמיתית נמצאת בתוך מרחב הפונקציות שבו אנו מחפשים $\mathcal{H}$.

3. ההשפעה של גודל האילוף R :

עבור $a = 2, m = 30, d = 20$.

θ נבחרה רנדומלית כמו קודם, אבל בחרנו לנרמל ל $\|\theta\|_2 = 5$.



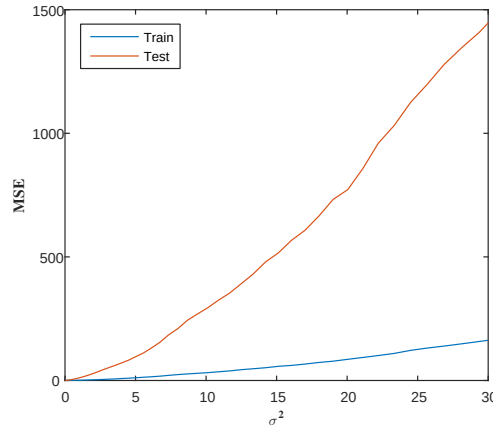
- השגיאה על דוגמאות הלמידה הולכת וגדלה ככל שאנו מקטינים את R מכיוון שאנו מקטינים את מרחב החיפוש.

- המרחק בין השגיאה על דוגמאות הלמידה והשגיאה האמיתית הולך ולקטן ככל שמקטינים את R .

- אפשר לראות שהשגיאה האמיתית קטנה עד לנקודה מסוימת (שהיא הנקודה הרצויה) אבל אם נקטין את R מעבר לנקודה זו אנחנו עוברים למצב של תת-התאמה.

4. ההשפעה של הרעש :

עבור $R = 100, m = 30, d = 20$.



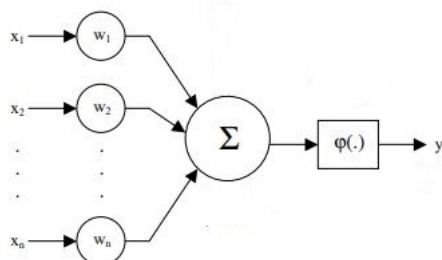
- כאשר אין רעש, כמובן שהשגיאות שוות לאפס כי $m > d$.
- ככל שנגדיל את הרעש השגיאה על דוגמאות הלמידה תלך ותגדל מכיוון שהדוגמאות יהיו יותר ויותר רחוקות מהקשר הליניארי שהמודל מניח.
- ככל שהרעש גדול יותר, ההבדל בין השגיאה על דוגמאות הלמידה והשגיאה האמיתית הולך וגדל.

5.2 רשת נוירונים

רשתות נוירונים זה אחד האלגוריתמים השימושיים ביותר שהראו יכולות דיוק גבוהות במיוחד בבעיות שונות, ישנן לא מעט סוגים של רשתות שבניות לפתירת בעיות מסוגים שונים, אבל הרעיון המרכזי הוא הרכבה של פונקציות פשוטות וכתוצאה מכך מקבלים מרחב פונקציות עשיר יותר שיכול לזהות תבניות לא ליניאריות בקלט. (אפשר להוכיח שבצורת רשת נוירונים אפשר לקרב כל פונקציה רציפה [10]).

רשת נוירונים מכילה מספר "נוירונים" שכל אחד מהם הוא פונקציה לא ליניארית פשוטה,

איור 5.1: נירון עם פונקציה φ



והפונקציות φ הכי פופולריות הן:

$$1. \varphi(y) = \frac{1}{1 + \exp^{-y}}$$

$$2. \varphi(y) = \max(0, y)$$

$$3. \varphi(y) = \tanh(y)$$

נמצא חסמים למספר הכיסוי עבור מרחב הפונקציות שמכיל את הנורונים :

$$\text{טענה 5.2.1 יהי } \mathcal{H} = \left\{ \varphi(\theta^T x) = \frac{1}{1 + \exp\{-\theta^T x\}} : \theta \in B_{R, \|\cdot\|_2} \right\}$$

אז

$$\mathcal{N}(\mathcal{H}, \epsilon) \leq 2^d \left\lceil \frac{R\sqrt{d}}{4\epsilon} \right\rceil^d$$

הוכחה: נסמן $g(y) = \frac{1}{1 + e^{-y}}$ ונשים לב כי $0 \leq g(y) \leq 1$ וגם

$$g'(y) = \frac{e^{-y}}{(1 + e^{-y})^2} = \frac{1}{1 + e^{-y}} \left(1 - \frac{1}{1 + e^{-y}} \right) = g(y)(1 - g(y))$$

ומזה נובע כי $|g'(y)| \leq \frac{1}{4}$.

לכן הפונקציה g היא ליפשיץ, כלומר : $|g(y) - g(\tilde{y})| \leq \frac{1}{4}|y - \tilde{y}|$: $\forall y, \tilde{y} \in \mathbb{R}$.

לכן :

$$\sup_{x \in [0,1]^d} |g(\theta^T x) - g(\tilde{\theta}^T x)| \leq \sup_{x \in [0,1]^d} \frac{1}{4} |(\theta - \tilde{\theta})^T x| \leq \frac{\sqrt{d}}{4} \|\theta - \tilde{\theta}\|_2$$

בעזרת אותו טיעון שנתנו בטענה הקודמת נקבל כי :

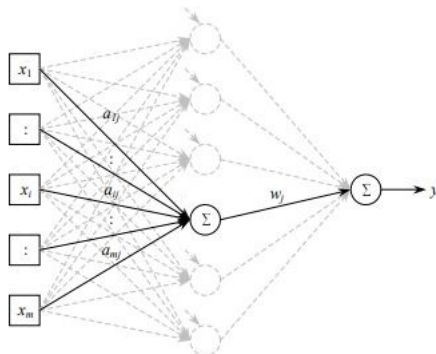
$$(5.2.0.1) \quad \mathcal{N}(\mathcal{H}, \epsilon) \leq \mathcal{N}\left(B_{R, \|\cdot\|_2}, \frac{4\epsilon}{\sqrt{d}}\right) \leq 2^d \left\lceil \frac{R\sqrt{d}}{4\epsilon} \right\rceil^d$$

■

הערה 5.2.1 ניתן להגיע לחסמים דומים גם עבור הנורוניהם מהסוגים האחרים בעזרת חסם לקבוע ליפשיץ שלהן.

מהטענה הזאת אפשר לראות שכל מה שאנו צריכים כדי למצוא חסם למספר הכיסוי הוא חסם לקבוע ליפשיץ של הפונקציה ביחס ל θ וכמובן הרכבה של פונקציות f_1, f_2 עם קבועי ליפשיץ k_1, k_2 בהתאמה גם היא ליפשיץ k כך ש: $k \leq k_1 \cdot k_2$. רשת נורונים עם שכבה אחת לביצוע רגרסיה לוקחת את הצירוף הליניארי של השכבה שיש בה מספר נורונים כמו בדוגמא למעלה.

איור 5.2: רשת נורונים עם שכבה אחת



כלומר אפשר לייצג רשת נורונים ע"י

$$f_{\theta, \mathbf{w}}(x) = \sum_{n=1}^n w_j \cdot \varphi(\theta_j^T x)$$

שבעזרת כתיב מטריצוני אפשר לכתוב :

$$f_{\theta, \mathbf{w}}(x) = \mathbf{w}^T \varphi(\Theta x)$$

וקל לראות שרשת נורונים מהווה הרכבה של פונקציות ליפשיציות ולכן גם היא ליפשיץ. לכן עבור מרחב הפונקציות שמתאר רשתות נורונים (עם אילוף על הפרמטרים $(\Theta, \mathbf{w})$ אפשר למצוא חסם למספר הכיסוי ולהשתמש בחסמים התיאורטיים שהצגנו בפרק 4.

פרק 6

סיכום

בעזרת התיאוריה שהצגנו ניתן לראות שלמידה היא אכן אפשרית כאשר קיימת תבנית כלשהי בנתונים שמקבלים כקלט ועל ידי בחירה נכונה של מרחב הפונקציות. ראינו עבור רגרסיה ליניארית איך פרמטרים שונים של האלגוריתם משפיעים על איכות התחזיות ועל האיוון בין תת-למידה ללמידת-יתר. מהחסמים שהוצגו בתיאוריה ניתן להבין עקרונות כלליים לתכנון אלגוריתמי למידה יעילים גם עבור בעיות אחרות, למשל:

1. ככל שמרחב הפונקציות שנבחר הוא יותר "עשיר" אנו מסתכנים בלמידת-יתר גם אם השגיאה על דוגמאות הלמידה קטנה.

2. כאשר השגיאה על דוגמאות הלמידה גדולה (תת-התאמה) אולי כדאי לנסות לבחור מרחב פונקציות יותר עשיר.

3. לחלופין אם אנו נתקלים בבעיה של התאמת-יתר אז ישנם כמה דברים שיכולים לעזור:

- בחירת מרחב פונקציות יותר קטן.

- הגדלת פרמטר הריגורויזציה.

- להוסיף עוד דוגמאות למידה.

המסקנות הנ"ל שימושיות עבור הרבה בעיות יישומיות אבל ישנן בעיות שבהן ההנחות שהצגנו אינן תקפות ולכן דרושות הרחבות של התיאוריה. למשל:

1. אנו מניחים כי הדוגמאות מיוצרות ע"י פונקציית צפיפות שלא משתנה עם הזמן, באופן מעשי ישנן הרבה בעיות שבהן הנחה זו אינה נכונה. למשל בעיה של זיהוי הונאות שבה סוגי ההונאות משתנים כל הזמן ולכן התבניות בנתונים שיעזרו לנו לזהות אותן גם הן משתנות.

2. הנחנו שהמודל מחזיר $y \in \mathbb{R}$, איך החסמים והתיאוריה משתנה עבור בעיות אחרות, למשל $y \in \mathbb{R}^n$?

3. את השגיאה חישבנו על ידי תוחלת ריבועי השגיאות, מה אם נחליף את זה בערך המוחלט של השגיאות?

עבור הבעיות שאכן עומדות בהנחות שהצגנו, האם אפשר למצוא חסמים יותר הדוקים למרחק בין השגיאה האמיתית לשגיאה על דוגמאות הלמידה? התיאוריה שפיתחנו בנויה

על כך שאנו מוצאים חסמים שתקפים עבור כל $f \in \mathcal{H}$, אבל בד"כ אנו מתעניינים רק בתתי־קבוצה של $\mathcal{H}$ שממזערת את השגיאה על דוגמאות הלמידה. האם ניתן להקל בהנחה זו? ואם כן איך זה ישפיע על החסמים?

פרק 7

ביביוֹלוגְרפיה

- [1] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM JOURNAL OF RESEARCH AND DEVELOPMENT*, pp. 71–105, 1959.
- [2] D. Yiming *et al.*, "A deep learning model to predict a diagnosis of alzheimer disease by using 18f-fdg pet of the brain," *Radiology*, vol. 290, no. 2, pp. 456–464, 2019.
- [3] M. Havaei *et al.*, "Brain tumor segmentation with deep neural networks," *Medical image analysis*, vol. 35, pp. 18–31, 2017.
- [4] E. Andre *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," 2017.
- [5] *Machine Learning in Finance*. [Online]. Available: "https://igniteoutsourcing.com/fintech/machine-learning-in-finance/"
- [6] M. Johnson *et al.*, "Google's multilingual neural machine translation system: Enabling zero-shot translation," *CoRR*, vol. abs/1611.04558, 2016.
- [7] V. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," vol. 16, no. 2, pp. 264–280.
- [8] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bulletin of the American Mathematical Society*, vol. 39, pp. 1–49, 2001.
- [9] K. Marius *et al.*, "Efficient and accurate lp-norm multiple kernel learning," in *Advances in Neural Information Processing Systems 22*, Y. Bengio *et al.*, Eds. Curran Associates, Inc., 2009, pp. 997–1005. [Online]. Available: <http://papers.nips.cc/paper/3675-efficient-and-accurate-lp-norm-multiple-kernel-learning.pdf>
- [10] B. C. Csáji, "Approximation with artificial neural networks," *MSc Thesis, Eötvös Loránd University (ELTE), Budapest, Hungary*, 2001.

פרק 8

נספח: תכניות Matlab לסימולציות בפרק 5

1. פתרון של הריבועים הפחותים עם אילוץ על הנורמה -

```
function [ theta ] = fitConsLeastSquares( X, y, R )
% find the solution with the smallest norm
theta = pinv(X) * y;
% if its feasible, end;
if (norm(theta, 2) <= R)
    return;
end
% otherwise : using KKT conditions-
% theta* = (X^T X + lambda I)^(-1) X^T y
% and ||theta*(lambda)||_2 = R:
I = eye(size(theta, 1));
XX = X.' * X;
Xy = X.' * y;
% find the lambda that solves the equation
KKT = @(Lambda) ...
    norm((XX + (2*Lambda * I)) \ Xy, 2) - R;
Lambda = fzero(KKT, 0);
% using lambda, return theta*
theta = (XX + (2* Lambda * I)) \ Xy;
end
```


2. קוד שייצר את הגרף עבור d משתנה:

```
sample = 100000;
m = 30; R = 100; a = 2; simulations = 200;
train_error = zeros(29,simulations);
test_error = zeros(29,simulations);
for s = 1:simulations
    for d = 2:30
        % pick d-dimension thetas using N(0,10)
        Theta_true = unifrnd(-5,5,d,1);
        %generate data;
        X_gen = unifrnd(0,1,sample,d);
        y_gen = X_gen*Theta_true + unifrnd(-a,a,sample,1);
        % split data to train and test sets.
        Xtrain = X_gen(1:m,:); ytrain = y_gen(1:m);
        Xtest = X_gen(m + 1:end,:);
        ytest = y_gen(m + 1:end);
        % fit the data & calculate errors
        [theta] = fitConsLeastSquares(Xtrain,ytrain,R);
        train_error(d-1,s) = MSE(Xtrain,ytrain,theta);
        test_error(d-1,s) = MSE(Xtest,ytest,theta);
    end
end
%averaging error across simulations
avg_train_error = mean(train_error,2);
avg_test_error = mean(test_error,2);
```

. קוד שייצר את הגרף עבור m משתנה:

```
sample = 10000;
d = 20; R = 100; simulations = 200;
max_m = 200;
train_error = zeros(max_m,simulations);
test_error = zeros(max_m,simulations);
for s = 1:simulations
    % generate data
    X_gen = unifrnd(0,1,sample,d);
    Theta_true = unifrnd(-5,5,d,1);
    y_gen = X_gen*Theta_true + unifrnd(-2,2,sample,1);
    for i = 1:max_m
        Xtrain = X_gen(1:i,:); ytrain = y_gen(1:i);
        Xtest = X_gen(i + 1:end,:);
        ytest = y_gen(i + 1:end);
        % fit the data using unconstrained LS
        [theta] = fitConsLeastSquares(Xtrain,ytrain,R);
        train_error(i,s) = MSE(Xtrain, ytrain, theta);
        test_error(i,s) = MSE(Xtest, ytest, theta);
    end
end
%averaging error across simulations
avg_train_error = mean(train_error,2);
avg_test_error = mean(test_error,2);
```

3. קוד שייצר את הגרף עבור R משתנה:

```
sample = 10000;
m = 30; d = 20; simulations = 200;
R_vec = linspace(0.01,8,100)';
train_error = zeros(length(R_vec),simulations);
test_error = zeros(length(R_vec),simulations);
for s = 1:simulations
    Theta_true = unifrnd(-5,5,d,1);
    % normalize theta to ||theta|| = 5 for visulaization
    Theta_true = 5 * Theta_true/norm(Theta_true,2);
    X_gen = unifrnd(0,1,sample,d);
    y_gen = X_gen*Theta_true + unifrnd(-2,2,sample,1);
    Xtrain = X_gen(1:m,:); ytrain = y_gen(1:m);
    Xtest = X_gen(m + 1:end,:);
    ytest = y_gen(m + 1:end);
    for i = 1:length(R_vec)
        %fit using Constrained LS
        [theta] = ...
            fitConsLeastSquares(Xtrain,ytrain, R_vec(i));
        train_error(i,s) = MSE(Xtrain, ytrain, theta);
        test_error(i,s) = MSE(Xtest, ytest, theta);
    end
end
%averaging error across simulations
avg_train_error = mean(train_error,2);
avg_test_error = mean(test_error,2);
```

4. קוד שייצר את הגרף עבור רעש משתנה:

```
sample = 100000;
m = 30; d = 20; R = 100; simulations = 200;
e = linspace(0,10,50); % errors will be U(-e(i),e(i))
% initialize error containers
train_error = zeros(length(e),simulations);
test_error = zeros(length(e),simulations);
for s = 1 : simulations
    %generate data
    Theta_true = unifrnd(-5,5,d,1);
    X_gen = unifrnd(0,1,sample,d);
    y_gen = X_gen*Theta_true;
    for i = 1:length(e)
        % add noise
        y_gen = y_gen + unifrnd(-e(i),e(i),sample,1);
        % split the data to training and testing
        Xtrain = X_gen(1:m,:); ytrain = y_gen(1:m);
        Xtest = X_gen(m + 1:end,:);
        ytest = y_gen(m + 1:end);
        % fit unconstrained LS & calculate errors
        [theta] = fitConsLeastSquares(Xtrain,ytrain,R);
        train_error(i,s) = MSE(Xtrain,ytrain,theta);
        test_error(i,s) = MSE(Xtest,ytest,theta);
    end
end
%averaging the simulation errors
avg_train_error = mean(train_error,2);
avg_test_error = mean(test_error,2);
```