

הרצאה 2: הסכנות עם אדפטיביות

Source: Lecture notes by
Aaron Roth and Adam Smith

מרצה: אורי שטמר

ננתח את תרגיל הרשות משיעור שעבר (הוא מראה שגם אנליסטים "תמימים" עלולים לעשות overfit):

תזכורת: הדומיין הוא $X = \{0,1\}^d \times \{0,1\}$ כאשר עבור $(x, y) \in X$ אנחנו נתייחס ל- x כאל "איבר" או "דוגמה" ונתייחס ל- y כאל "התיוג" של x . בהינתן מדגם S (שנדגם מתוך התפלגות לא ידועה \mathcal{D} מעל הדומיין X), המטרה שלנו היא למצוא פונקציית תיוג $f: \{0,1\}^d \rightarrow \{0,1\}$ שתחזה, טוב ככל האפשר, תיוגים של דוגמאות חדשות מ- \mathcal{D} . כלומר, אנחנו מחפשים פונקציה f כך שהביטוי הבא נמוך ככל האפשר:

$$\text{error}_{\mathcal{D}}(f) = \Pr_{(x,y) \sim \mathcal{D}} [f(x) \neq y] = \mathbb{E}_{(x,y) \sim \mathcal{D}} [\mathbb{1}\{f(x) \neq y\}]$$

הערות:

- * לפונקציה f כזאת קוראים לפעמים "מסווג" או "השערה" או "כלל תחזית"
- * ל- $\text{error}_{\mathcal{D}}(f)$ קוראים "שגיאת ההכללה" של f
- * שימו לב ש- $\text{error}_{\mathcal{D}}(f)$ היא שאילתא סטטיסטית

איך נוכל למצוא פונקציה f כזאת? הפרוצדורה הבאה נראית סבירה לכאורה:

$$1. \text{ לכל קואורדינטה } 1 \leq i \leq d, \text{ חשב } c_i = \mathbb{E}_{(x,y) \sim S} [\mathbb{1}\{x_i = y\}] = \frac{1}{n} \sum_{(x,y) \in S} \mathbb{1}\{x_i = y\}$$

כלומר, לכל קואו' i אנחנו בודקים כמה "טוב" היא חוזה את התיוג

$$2. \text{ נאמר שקואו' } i \text{ "חוזה טוב" את התיוג אם } c_i \geq \frac{1}{2} + \frac{1}{\sqrt{n}}$$

תהי $P \subseteq \{1, 2, \dots, d\}$ קבוצת כל הקואו' שחוזות טוב את התיוג.

3. נגדיר את כלל התחזית הבא, אשר מבצע החלטת רוב מבין הקואו' שחוזות טוב:

$$f(x) = \begin{cases} 1 & , \sum_{i \in P} x_i \geq \frac{|P|}{2} \\ 0 & , \text{ otherwise} \end{cases}$$

4. נעריך את השגיאה האמפירית של f על המדגם שלנו, כלומר נחשב את

$$\text{error}_S(f) = \mathbb{E}_{(x,y) \sim S} [\mathbb{1}\{f(x) \neq y\}] = \frac{1}{n} \sum_{(x,y) \in S} \mathbb{1}\{f(x) \neq y\}$$

שימו לב שבפרוצדורה הזאת ניגשנו למדגם רק דרך הצגת שאילתות סטטיסטיות (ולמידת הערך האמפירי שלהם). בסה"כ שאלנו $d + 1$ שאילתות סטטיסטיות. לכן, אם המשפט שהוכחנו בשיעור שעבר היה נכון גם לאנליסטים שנראים "תמימים" ונגשים לנתונים רק דרך שאילתות סטטיסטיות, אז היינו יכולים לצפות שבהסתברות לפחות $1 - \beta$ נקבל

$$|\text{error}_S(f) - \text{error}_{\mathcal{D}}(f)| \leq O\left(\sqrt{\frac{\log(d/\beta)}{n}}\right)$$

האם זה נכון?

משפט 0: קיים קבוע $c > 0$ כך שהטענה הבאה מתקיימת. לכל $d \leq c \cdot n$, בהסתברות לפחות 0.9 מתקיים

$$|\text{error}_S(f) - \text{error}_D(f)| \leq 0.01$$

(למה משפט 0 נכון?)

משפט 1: קיים קבוע $c > 1$ כך שהטענה הבאה מתקיימת. לכל $d \geq c \cdot n$, אם \mathcal{D} היא ההתפלגות האחידה על פני $\{0,1\}^d \times \{0,1\}$, אזי בהסתברות לפחות 0.9 מתקיים

$$|\text{error}_S(f) - \text{error}_D(f)| \geq 0.49$$

שימו לב:

- (1) אנחנו יודעים ש $\text{error}_D(f) = \frac{1}{2}$ ולכן המשפט הנ"ל אומר שבהסתברות גבוהה נמצא כל תחזית עם שגיאה אמפירית קרובה מאוד לאפס
- (2) זה אומר שבאופן כללי אנחנו לא יכולים לצפות שהתשובות האמפיריות לשאלות סטטיסטיות יתנו דיוק לא טריוויאלי עבור יותר ממספר לינארי (ב- n) של שאלות, כאשר השאלות נבחרות בצורה אדפטיבית כתלות בתשובות לשאלות הקודמות.

טענת עזר: בהסתברות לפחות 0.95 מתקיים $|P| = \Omega(d)$

רעיון ההוכחה של טענת העזר:

א. המשתנים c_1, c_2, \dots, c_d הם בלתי תלויים, ולכל i מתקיים $\Pr[c_i \geq \frac{1}{2} + \frac{1}{\sqrt{n}}] = \Omega(1)$.

העובדה ש- c_1, c_2, \dots, c_d בלתי תלויים נובעת מכך שההתפלגות \mathcal{D} היא אחידה על פני $\{0,1\}^d \times \{0,1\}$ ולכן הקואורדינטות השונות הם בלתי תלויים. העובדה ש- $\Pr[c_i \geq \frac{1}{2} + \frac{1}{\sqrt{n}}] = \Omega(1)$ נובעת מכך ש- $c_i = \frac{1}{n} \sum_{(x,y) \in S} \mathbb{1}\{x_i = y\}$ הוא משתנה מקרי בינומי (מנורמל) עם תוחלת $\frac{1}{2}$ ועם סטיית תקן $\sigma = \frac{1}{2\sqrt{n}}$, ואפשר להראות שמשתנה מקרי כזה מתרחק מהתוחלת שלו בלפחות 2σ בהסתברות קבועה (לחסמים כאלה קוראים "אנטי-ריכוז"). ראו פרטים נוספים למטה.

ב. לכן, לפי צ'רנוף, עבור $d = \Omega(1)$ גדול מספיק, נקבל שבהסתברות לפחות 0.95 מתקיים $|P| = \mathbb{1}\{1 \in P\} + \dots + \mathbb{1}\{d \in P\} = \Omega(d)$

פרטים נוספים לגבי האנטי-ריכוז:

נסתכל על המשתנה המקרי $W = n \cdot c_i = \sum_{(x,y) \in S} \mathbb{1}\{x_i = y\}$. זה משתנה מקרי בינומי לא מנורמל עם תוחלת $\frac{n}{2}$. חסמי צ'רנוף והופדינג שראינו בשיעור שעבר נותנים לנו תוצאת "ריכוז" שאומרת שבהסתברות קבועה מתקיים ש- W קרוב לתוחלת שלו עד כדי $\pm\sqrt{n}$. ספציפית:

$$\Pr\left[\left|W - \frac{n}{2}\right| \geq \sqrt{n}\right] \leq 2e^{-2} \approx 0.27$$

עכשיו נראה תוצאת "אנטי-ריכוז" עבור המקרה הזה, שמראה שחסם הריכוז הנ"ל הוא בערך הדוק. כלומר תוצאת אנטי-ריכוז שמראה שבהסתברות קבועה המ"מ W $\ll \sqrt{n}$ מתרחק מהתוחלת שלו במשהו כמו \sqrt{n} .

משפט אנטי ריכוז:

$$\Pr \left[\left| W - \frac{n}{2} \right| \geq \frac{\sqrt{n}}{2} \right] \geq 0.2$$

סקיצת הוכחה:

ההסתברות $\Pr[W = i]$ מתמקסמת עבור $i = \frac{n}{2}$, ועבור המקרה הזה אנחנו מקבלים:

$$\Pr \left[W = \frac{n}{2} \right] = 2^{-n} \cdot \binom{n}{n/2} = 2^{-n} \cdot \frac{n!}{\left(\frac{n}{2}\right)! \cdot \left(\frac{n}{2}\right)!}$$

נשתמש בקירוב סטרלינג עבור עצרת, שאומר $n! \approx \sqrt{2\pi n} \cdot \left(\frac{n}{e}\right)^n$ ונקבל

$$\Pr \left[W = \frac{n}{2} \right] \approx 2^{-n} \cdot \frac{\sqrt{2\pi n} \cdot \left(\frac{n}{e}\right)^n}{\sqrt{\pi n} \cdot \left(\frac{n}{2e}\right)^{\frac{n}{2}} \cdot \sqrt{\pi n} \cdot \left(\frac{n}{2e}\right)^{\frac{n}{2}}} = \sqrt{\frac{2}{\pi n}}$$

כעת נוכל להוכיח את משפט האנטי-ריכוז שלנו:

$$\Pr \left[\left| W - \frac{n}{2} \right| < \frac{\sqrt{n}}{2} \right] \leq \sum_{i=\frac{n}{2}-\frac{\sqrt{n}}{2}}^{\frac{n}{2}+\frac{\sqrt{n}}{2}} \Pr[W = i] \leq 2 \cdot \frac{\sqrt{n}}{2} \cdot \sqrt{\frac{2}{\pi n}} = \sqrt{\frac{2}{\pi}} \approx 0.8$$

כלומר

$$\Pr \left[\left| W - \frac{n}{2} \right| \geq \frac{\sqrt{n}}{2} \right] \geq 0.2$$

מ.ש.ל. (ההוכחה הזאת הייתה פשטנית. בפועל ההסתברות יותר גבוהה מזה)

בנוסף, לפי סימטריה סביב $\frac{n}{2}$, משפט האנטי-ריכוז הזה מראה שמתקיים:

$$\Pr \left[W \geq \frac{n}{2} + \frac{\sqrt{n}}{2} \right] \geq 0.1$$

הוכחת משפט 1:

תחילה נשים לב שלפי הבחירה של ההתפלגות \mathcal{D} מתקיים:

$$\text{error}_{\mathcal{D}}(f) = \Pr_{(x,y) \sim \mathcal{D}} [f(x) = y] = 0.5$$

כעת ננתח את $\text{error}_S(f)$ ונראה שבהסתברות גבוהה הוא יהיה הרבה יותר קטן. לצורך כך, נחשוב על תהליך שקול להגרלת המדגם S :

- קודם נגריל את העמודה $\vec{y} = (y_1, y_2, \dots, y_n)$
- אח"כ נגריל מי יהיו הקואורדינטות P "שיחזו טוב" את y
- ולבסוף נגריל את העמודות \vec{x}_i עבור קואו' $i \in P$ ועבור קואו' $i \notin P$, לפי ההסתברויות המותנות המתאימות.

נשים לב שבהינתן P , \vec{y} מתקיים שהעמודות $\{\vec{x}_i : i \in P\}$ הן בלתי תלויות. נקבע \vec{y} כלשהו ונקבע $P \subseteq \{1, 2, \dots, d\}$ מתקיים:

$$\mathbb{E}_{\vec{x}_1, \dots, \vec{x}_d} [\text{error}_S(f)] = \mathbb{E}_{\vec{x}_1, \dots, \vec{x}_d} \left[\Pr_{(x,y) \sim S} [f(x) \neq y] \right]$$

$$\begin{aligned}
&= \mathbb{E}_{\vec{x}_1, \dots, \vec{x}_d} \left[\mathbb{E}_{(x,y) \sim S} [\mathbb{1}\{f(x) \neq y\}] \right] = \mathbb{E}_{\vec{x}_1, \dots, \vec{x}_d} [\mathbb{1}\{f(x) \neq y\}] \\
&= \Pr_{\substack{\vec{x}_1, \dots, \vec{x}_d \\ (x,y) \sim S}} [f(x) \neq y] = \Pr_{\substack{\vec{x}_1, \dots, \vec{x}_d \\ (x,y) \sim S}} \left[\sum_{i \in P} \mathbb{1}\{x_i = y\} < \frac{|P|}{2} \right] = ((1))
\end{aligned}$$

כאשר השיויון האחרון נובע מהעובדה שלפי הגדרת f , עבור $(x, y) \in S$ מתקיים ש- $f(x) = y$ אם ורק אם $\sum_{i \in P} \mathbb{1}\{x_i = y\} \geq \frac{|P|}{2}$.

ננתח את התוחלת של $\sum_{i \in P} \mathbb{1}\{x_i = y\}$ לפי הגדרת P , לכל קביעה אפשרית של $\vec{x}_1, \dots, \vec{x}_d$ (אחרי שקבענו את (P, \vec{y}) עבור כל קואו' $i \in P$ מתקיים

$$\frac{1}{n} \sum_{(x,y) \in S} \mathbb{1}\{x_i = y\} \geq \frac{1}{2} + \frac{1}{\sqrt{n}}$$

כלומר,

$$\Pr_{(x,y) \sim S} [x_i = y] = \mathbb{E}_{(x,y) \sim S} [\mathbb{1}\{x_i = y\}] \geq \frac{1}{2} + \frac{1}{\sqrt{n}}$$

ולכן

$$\mathbb{E}_{(x,y) \sim S} \left[\sum_{i \in P} \mathbb{1}\{x_i = y\} \right] = \sum_{i \in P} \mathbb{E}_{(x,y) \sim S} [\mathbb{1}\{x_i = y\}] \geq \frac{|P|}{2} + \frac{|P|}{\sqrt{n}}$$

ולכן גם

$$\mathbb{E}_{\substack{\vec{x}_1, \dots, \vec{x}_d \\ (x,y) \sim S}} \left[\sum_{i \in P} \mathbb{1}\{x_i = y\} \right] \geq \frac{|P|}{2} + \frac{|P|}{\sqrt{n}}$$

נציב זאת ב- ((1)) ונקבל ש-

$$\mathbb{E}_{\vec{x}_1, \dots, \vec{x}_d} [\text{error}_S(f)] \leq \Pr_{\substack{\vec{x}_1, \dots, \vec{x}_d \\ (x,y) \sim S}} \left[\sum_{i \in P} \mathbb{1}\{x_i = y\} < \mathbb{E}_{\substack{\vec{x}_1, \dots, \vec{x}_d \\ (x,y) \sim S}} \left[\sum_{i \in P} \mathbb{1}\{x_i = y\} \right] - \frac{|P|}{\sqrt{n}} \right] \leq \exp\left(-\frac{2|P|}{n}\right)$$

כאשר אי-השיויון האחרון נובע מחסם צ'רנוף. אנחנו יכולים להשתמש פה בחסם צ'רנוף מכיוון שאחרי שקבענו את (P, \vec{y}) , מתקיים ש- $\sum_{i \in P} \mathbb{1}\{x_i = y\}$ הוא סכום של $|P|$ משתני 0/1 בלתי תלויים.

הביטוי האחרון הוא לכל היותר 1/2000 עבור $|P| \geq \frac{n \cdot \ln(2000)}{2}$.

במקרה כזה, לפי אי-שיויון מרקוב, נקבל ש-

$$\Pr_{\vec{x}_1, \dots, \vec{x}_d} \left[\text{error}_S(f) > \frac{1}{100} \right] \leq \frac{\mathbb{E}_{\vec{x}_1, \dots, \vec{x}_d} [\text{error}_S(f)]}{1/100} \leq \frac{1}{20}$$

זה נכון לכל קביעה של ϵ ולכל קביעה של P כך ש- $\Gamma \triangleq \frac{n \cdot \ln(2000)}{2}$. לפי טענת העזר, P כזה מתקבל בהסתברות לפחות 0.95 (כי טענת העזר אמרה שבהסת' לפחות 0.95 מתקיים ש- $|P|$ יהיה לנארי ב- d , ולכן אם $d = \Omega(n)$ אז אנחנו מקבלים את זה...).
 לכן בסה"כ נקבל

$$\begin{aligned} \Pr_{S \sim \mathcal{D}} \left[\text{error}_S(f) > \frac{1}{100} \right] &= \\ &= \Pr[|P| < \Gamma] \cdot \Pr \left[\text{error}_S(f) > \frac{1}{100} \mid P \right] + \Pr[|P| \geq \Gamma] \cdot \Pr \left[\text{error}_S(f) > \frac{1}{100} \mid P \right] \\ &\leq \frac{1}{20} \cdot \Pr \left[\text{error}_S(f) > \frac{1}{100} \mid P \right] + \Pr[|P| \geq \Gamma] \cdot \frac{1}{20} \leq \frac{1}{20} + \frac{1}{20} = \frac{1}{10} \end{aligned}$$

כלומר, בהסת' לפחות 0.9 מתקיים שהשגיאה האמפירית היא לכל היותר 1/100.

מ.ש.ל.

הערות:

1. בניגוד ל"מתקפה" הראשונה שראינו (עם ה low-order bits), האנליזה האחרונה לא דרשה לקבל תשובות עם דיוק גבוהה מידי והיא חסינה להרעשות קטנות בתשובות (מסדר גודל $o(1/\sqrt{n})$)
2. ראינו 2 "מתקפות" שהצליחו לעשות overfit למדגם. במובן מסויים, שתי המתקפות האלה קודם "למדו" את המדגם S בצורה כלשהי, ולאחר מכן השתמשו בידע על S כדי למצוא שאילתא שעושה overfit. המתקפה הראשונה שראינו עשתה את זה באופן מפורש, והמתקפה השנייה עשתה את זה באופן יותר סמוי. המסקנה כאן היא שכל שיטה שתבטיח הכללה במקרה האדפטיבי חייבת בצורה כלשהי למנוע מהתקוף ללמוד יותר מידי מידע על המדגם.

סיכום של הדברים שלמדנו עד כאן:

- במקרה הלא-אדפטיבי ניתן לענות על $\exp(n)$ שאילתות סטטיסטיות בעזרת מדגם בגודל n
- במקרה האדאפטיבי אנחנו חייבים להגביל את עצמו באיזושהי צורה. אחרת, אם האנליסט מקבל את כל המדגם ויכול לבחור כל שאילתא, אז אי אפשר להימנע מ overfitting
- החלטנו להגביל את הגישה של האנליסט למדגם
- איך נגביל אותו בדיוק? ניסינו לאפשר לו גישה לנתונים רק דרך שאילתות סטטיסטיות.
- חשבנו על משחק שבו בכל שלב האנליסט קובע שאילתא סטטיסטית ולומד את הערך האמפירי שלה על המדגם.
- ראינו שהגבלה כזאת לא מספיקה ושעדיין לא נוכל להבטיח שלא יהיה overfitting:
 - ראינו דוגמה לאנליסט שמתוך התשובה לשאילתא הסטטיסטית הראשונה לומד בדיוק את כל המדגם. לאחר מכן האנליסט יכול להגדיר שאילתא "רעה" עבורה הערך האמפירי שונה מאוד מהערך האמיתי (כלומר מהתוחלת על פני ההתפלגות שממנה הגיע המדגם).
 - ראינו דוגמה נוספת שבה גם אנליסט שנראה יותר "תמים" מגיע לשאילתא "רעה" כזאת.

נסתכל על משחק דומה לזה שראינו בפעם שעברה:

יהי M מכניזם שמקבל מדגם ועונה על שאילתות. עבור מדגם S ואנליסט A נגדיר את המשחק הבא שנקרא $AG_{n,k}(A, S, M)$, או בקיצור $AG_{n,k}(A, S, M)$

$AG_{n,k}(A, S, M)$
1. המכניזם M מקבל את המדגם S (האנליסט A לא מקבל את המדגם S)
2. עבור $i = 1, 2, \dots, k$: <ul style="list-style-type: none">• האנליסט A קובע שאילתא q_i• המכניזם M מקבל את השאילתא q_i ומחזיר תשובה a_i• האנליסט A מקבל את a_i
3. החזר את הטרנסקריפט של האינטראקציה בין האנליסט למכניזם: $T = (q_1, a_1, q_2, a_2, \dots, q_k, a_k)$

אנחנו מעוניינים לתכנן מכניזם M כך שכל התפלגות \mathcal{D} ולכל אנליסט A , עבור $S \sim \mathcal{D}^n$, בהסתברות גבוהה מתקיים: לאורך כל הריצה של המשחק הנ"ל, התשובות ש- M מחזיר קרובות לערך האמיתי של השאילתות. כלומר, לכל $1 \leq i \leq k$ מתקיים $a_i \approx q_i(\mathcal{D}) = \mathbb{E}_{x \sim \mathcal{D}}[q_i(x)]$.

הגדרה 1: מכניזם M הוא (α, β) -מדוייק-סטטיסטי עבור k שאילתות אדאפטיביות בהינתן מדגם בגודל n אם לכל התפלגות \mathcal{D} ולכל אנליסט A מתקיים

$$\Pr_{S \sim \mathcal{D}^n} [\exists i \text{ s.t. } |a_i - q_i(\mathcal{D})| > \alpha] \leq \beta$$

$AG_{n,k}(A, S, M)$

הערה: לפעמים נאמר בקיצור ש- M הוא (α, β) -מדוייק במקום (α, β) -מדוייק-סטטיסטי. **שימו לב:** בשביל שמכניזם יהיה (α, β) -מדוייק הוא צריך להבטיח שהתשובות שלו מדוייקות לכל אנליסט A . בפרט התשובות צריכות להיות מדוייקות גם עבור האנליסט "הגרוע ביותר" שמנסה לגרום למכניזם שלנו לטעות. לכן לפעמים נחשוב על האנליסט בתור "יריב".

שאלה מרכזית שנתעניין בה: מה צריך להיות n כפונקציה של k, α, β על מנת שנוכל להבטיח דיוק? כלומר, המטרה שלנו זה לתכנן מכניזם M שיעמוד בהגדרה הנ"ל בעזרת גודל מדגם $n = n(k, \alpha, \beta)$ קטן ככל האפשר.

מסקנה מתחילת השיעור: אם המכניזם M עונה על כל שאילתא בעזרת הממוצע האמפירי המדוייק, אז הוא לא עומד בהגדרה הנ"ל, אפילו לא עבור $k = 2$.

שאלה: בהגדרה הנ"ל, האם המכניזם M יודע את ההתפלגות \mathcal{D} ? האם האנליסט יודע?

איך נתכנן מכניזמים מדוייקים במקרה האדפטיבי (כלומר מדוייקים לפי הגדרה 1)?

אפשרות ראשונה: Sample Splitting

נתכנן מכניזם אשר מחלק את המדגם שלו ל- k חלקים זרים, ועונה על כל שאילתא בעזרת חלק אחר:

$\text{SampleSplitting}_{n,k,\alpha,\beta}(S)$
קלט: מדגם S בגודל n
1. חלק את S ל- k תתי מדגמים זרים: S_1, S_2, \dots, S_k
2. עבור $i = 1, 2, \dots, k$
– קבל את השאילתא הבאה q_i
– החזר $q_i(S_i) = \frac{1}{ S_i } \sum_{x \in S_i} q_i(x)$

שאלה: עבור איזה פרמטר $k = k(\alpha, \beta, n)$ נוכל להראות שהאלגוריתם הנ"ל הוא (α, β) -מדוייק עבור k שאילתות בהינתן מדגם בגודל n ?