

## הרצאה 1: מבוא

Source: Lecture notes by  
Aaron Roth and Adam Smith

מרצה: אורי שטמר

השאלה המרכזית בקורס:

**איך אפשר ללמוד על התפלגות לא ידועה  $\mathcal{D}$  כאשר כל מה שיש לנו זה אוסף של דגימות מתוך  $\mathcal{D}$ ?**

הדגימות כמובן מכילות מידע שימושי על ההתפלגות  $\mathcal{D}$ , אבל הסכנה היא שאם לא נזהר אנחנו עלולים "לגלות" תובנות שאולי מתקיימות עבור המדגם הספציפי שיש לנו, אבל לא נכונות באופן כללי בהתפלגות  $\mathcal{D}$ . התופעה הזאת נקראת "overfitting" או "false discovery".

יש לנו הבנה דיי טובה של בעיית ה overfitting וכלים טובים להתמודד איתה במקרה הלא-אדפטיבי. כלומר במקרה בו אנחנו מתחייבים מראש על הבדיקות שנעשה לדטה שנאסף, אח"כ אוספים דטה, ואח"כ מבצעים בדיקת את הבדיקות שהתחייבנו אליהם (ושום דבר מעבר). המצב הרבה פחות ברור במקרה האדפטיבי, שבו החוקר מחליט על הבדיקות שיבצע לאחר שאסף (וראה) את הנתונים. הקורס שלנו יעסוק במקרה האדפטיבי. אבל היום, נתחיל מסקירה בסיסית של המקרה הלא-אדפטיבי.

### ניתוח מידע לא-אדפטיבי

נתחיל מהבעיה הפשוטה הבאה. נניח שיש לנו מטבע לא הוגן, כלומר ההסתברות לקבל "עץ" לא שווה להסתברות לקבל "פלי". אנחנו יכולים להטיל את המטבע כמה פעמים שנרצה ולראות את התוצאה. איך נוכל ללמוד מהי ההסתברות לקבל "עץ" בהטלת המטבע הזו?

#### פורמלית:

ישנה התפלגות Bernoulli( $p$ ) עבור פרמטר  $p$  לא ידוע לנו. כלומר אם  $X \sim \text{Bernoulli}(p)$  אזי  $\Pr[X = 1] = p$  ו-  $\Pr[X_i = 0] = 1 - p$ . יש לנו "דטהבייס"  $x = (x_1, x_2, \dots, x_n)$  המכיל  $n$  דגימות בלתי תלויות מההתפלגות הזאת (כלומר מכיל את התוצאות של  $n$  הטלות מטבע). איך נוכל להשתמש ב  $x$  כדי ללמוד מהו  $p$ ?

אפשרות אחת היא להשתמש בממוצע האמפירי  $\hat{p} = \frac{1}{n} \sum_{i=1}^n x_i$  בתור ההערכה שלנו ל-  $p$ .

כמה זה מדוייק? איך בכלל ניתן לכמת את "רמת הדיוק" כאן?

#### דוגמה:

נניח ש-  $p = 1/2$  ונניח ש-  $n = 1000$ . למרות שהתוחלת של  $\hat{p}$  שווה בדיוק ל-  $p$ , ההסתברות שבאמת נקבל  $\hat{p} = 1/2$  בדיוק היא יחסי קטנה, משהו כמו  $\sqrt{2/\pi n}$  (יכול להתחכם ולבחור את  $p$  להיות אי-רציונלי, ובמקרה זה ההסתברות ש-  $\hat{p} = p$  היא אפס...). כלומר, הטענה הבאה שגויה: ככל ש-  $n$  גדל, ההסתברות ש-  $\hat{p}$  יפגע בדיוק ב-  $p$  עולה. מה כן אפשר להגיד?

- אנחנו יכולים להיות "דיי בטוחים" ש-  $\hat{p}$  יהיה בין 0.45 ל- 0.55 (ההסתברות זה לא קורה היא  $\geq 0.02$ )
- אנחנו יכולים להיות אפילו "יותר בטוחים" שזה יהיה בין 0.44 ל- 0.56 (ההסתברות שזה לא קורה היא לכל היותר 0.0015)

טיעונים כאלה נקראים high probability bounds או confidence intervals.

עכשיו אנחנו מעוניינים להבין איך אפשר להוכיח טענות כאלה, כלומר איך אפשר להוכיח אמירות מהצורה

ההסתברות ש  $|p - \hat{p}|$  יהיה יותר מ-  $A$  היא לכל היותר  $B$

כאשר ההסתברות היא מעל הטלת המטבעות, כלומר מעל הגרלת "הדטהבייס" שלנו  $x$ , וכאשר נהייה מעוניינים ש-  $A, B$  יהיו קטנים ככל האפשר.

בשביל זה אנחנו צריכים להיזכר ברקע מהסתברות.

**משפט [אי-שוויון מרקוב]:** לכל משתנה מקרי אי שלילי  $Y$  ולכל  $a > 0$  מתקיים

$$\Pr[Y \geq a] \leq \frac{\mathbb{E}[Y]}{a}$$

**הוכחה:** נוכיח עבור משתנה מקרי  $Y$  בדיד:

$$\mathbb{E}[Y] = \sum_y y \cdot \Pr[Y = y] \geq \sum_{y \geq a} y \cdot \Pr[Y = y] \geq \sum_{y \geq a} a \cdot \Pr[Y = y] = a \cdot \sum_{y \geq a} \Pr[Y = y] = a \cdot \Pr[Y \geq a]$$

בדוגמה שלנו עם המטבע, אנחנו יודעים ש-  $\mathbb{E}[\hat{p}] = p$  ולכן אי-שוויון מרקוב נותן לנו איזשהו מידע (חלש) לגבי הקשר בין  $p$  ו-  $\hat{p}$ . אבל זה עוד לא מספיק כדי לקבל high probability bounds כמו שרצינו. עדיין, אי-שוויון מרקוב הוא כלי מאוד שימושי. בפרט, הוא מאפשר לנו להוכיח את אי-שוויון צ'בישב:

**משפט [אי-שוויון צ'בישב]:** לכל משתנה מקרי  $Y$  עם תוחלת  $\mu = \mathbb{E}[Y]$  ושונות  $\sigma^2 = \text{Var}(Y) = \mathbb{E}[(Y - \mu)^2]$  ולכל  $a > 0$  מתקיים:

$$\Pr[|Y - \mu| \geq a\sigma] \leq \frac{1}{a^2}$$

**הוכחה:**

$$\Pr[|Y - \mu| \geq a\sigma] = \Pr[(Y - \mu)^2 \geq a^2\sigma^2] \leq \frac{\mathbb{E}[(Y - \mu)^2]}{a^2\sigma^2} = \frac{1}{a^2}$$

כאשר אי השוויון נובע מאי-שוויון מרקוב.

זה כבר מספיק כדי לתת לנו איזשהו חסם הסתברותי לא טריוויאלי לגבי המטבע שלנו. כלומר אנחנו רוצים להשתמש באי-שוויון צ'בישב כדי לקבל חסם הסתברותי על  $|\hat{p} - p|$ . בשביל זה אנחנו צריכים לנתח את השונות של  $\hat{p}$ . בשביל זה אנחנו צריכים להזכר בתכונות של שונות ותוחלת.

**משפט:** עבור משתנים מקריים בלתי תלויים מתקיים שתוחלת של מכפלה שווה למכפלת התוחלות, כלומר

$$\mathbb{E}[X_1 \cdot X_2 \cdots X_n] = \mathbb{E}[X_1] \cdot \mathbb{E}[X_2] \cdots \mathbb{E}[X_n]$$

**הוכחה:** נוכיח עבור משתנים מקריים  $X, Y$  בדידים ובלתי תלויים

$$\begin{aligned} \mathbb{E}[X \cdot Y] &= \sum_{x,y} \Pr[X = x, Y = y] \cdot xy = \sum_{x,y} \Pr[X = x] \cdot \Pr[Y = y] \cdot xy \\ &= \left( \sum_x \Pr[X = x] \cdot x \right) \cdot \left( \sum_y \Pr[Y = y] \cdot y \right) = \mathbb{E}[X] \cdot \mathbb{E}[Y] \end{aligned}$$

**משפט:** עבור משתנה מקרי  $Y$  ועבור  $a > 0$  מתקיים

$$\text{Var}(aY) = a^2 \cdot \text{Var}(Y)$$

**הוכחה:** נסמן  $\mu = \mathbb{E}[Y]$  אזי

$$\text{Var}(aY) = \mathbb{E}[(aY - a\mu)^2] = a^2 \mathbb{E}[(Y - \mu)^2] = a^2 \cdot \text{Var}(Y)$$

**משפט:** עבור זוג משתנים מקריים בלתי תלויים  $Y_1, Y_2$  מתקיים

$$\text{Var}(Y_1 + Y_2) = \text{Var}(Y_1) + \text{Var}(Y_2)$$

**הוכחה:** נובע מהגדרת השונות ומהעובדה (שהוכחנו) שעבור משתנים מקריים בלתי תלויים מתקיים שתוחלת של מכפלה שווה למכפלת התוחלות:

$$\begin{aligned} \mathbb{E} \left[ (Y_1 + Y_2 - \mathbb{E}[Y_1 + Y_2])^2 \right] &= \mathbb{E} \left[ (Y_1 - \mathbb{E}[Y_1]) + (Y_2 - \mathbb{E}[Y_2]) \right]^2 \\ &= \mathbb{E} \left[ (Y_1 - \mathbb{E}[Y_1])^2 + (Y_2 - \mathbb{E}[Y_2])^2 + 2(Y_1 - \mathbb{E}[Y_1])(Y_2 - \mathbb{E}[Y_2]) \right] \\ &= \text{Var}(Y_1) + \text{Var}(Y_2) + 2 \mathbb{E} \left[ (Y_1 - \mathbb{E}[Y_1])(Y_2 - \mathbb{E}[Y_2]) \right] \\ &= \text{Var}(Y_1) + \text{Var}(Y_2) + 2 \mathbb{E} \left[ Y_1 Y_2 - Y_1 \mathbb{E}[Y_2] - Y_2 \mathbb{E}[Y_1] + \mathbb{E}[Y_1] \mathbb{E}[Y_2] \right] \\ &= \text{Var}(Y_1) + \text{Var}(Y_2) + 2 \left( \mathbb{E} \left[ Y_1 Y_2 \right] - \mathbb{E}[Y_1] \mathbb{E}[Y_2] - \mathbb{E}[Y_2] \mathbb{E}[Y_1] + \mathbb{E}[Y_1] \mathbb{E}[Y_2] \right) \\ &= \text{Var}(Y_1) + \text{Var}(Y_2) + 2 \left( \mathbb{E} \left[ Y_1 Y_2 \right] - \mathbb{E}[Y_1] \mathbb{E}[Y_2] \right) \stackrel{\text{אי תלות}}{=} \text{Var}(Y_1) + \text{Var}(Y_2) \end{aligned}$$

נחזור למטבע שלנו. אנחנו רוצים לנתח את השונות של  $\hat{p}$ . הזכרו כי הגדרנו  $\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i$  כאשר של  $X_i$  הוא משתנה מקרי ברנולי (עם פרמטר  $p$ ). עבור כל  $X_i$  כזה מתקיים ש-

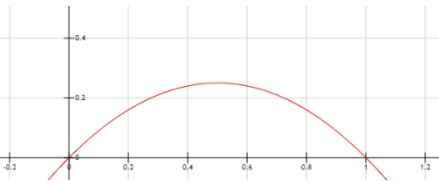
$$\text{Var}(X_i) = \mathbb{E}[(X_i - p)^2] = \underbrace{\mathbb{E}[X_i^2]}_{\substack{X_i \text{ is a bit} \\ \text{and so } X_i^2 = X_i \\ \text{and } \mathbb{E}[X_i^2] = \mathbb{E}[X_i] = p}} - 2p\mathbb{E}[X_i] + p^2 = p - 2p^2 + p^2 = p(1 - p)$$

עובדה (ראו גרף משמאל):

$$\text{Var}(X_i) \leq \frac{1}{4} \text{ לכן } p(1 - p) \leq \frac{1}{4} \text{ עבור } 0 \leq p \leq 1$$

**מסקנה:**

$$\text{Var}(\hat{p}) = \text{Var} \left( \frac{1}{n} \sum_{i=1}^n X_i \right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) \leq \frac{1}{4n}$$



לכן, נוכל להפעיל את אי-שוויון צ'בישב על  $\hat{p}$  ולקבל שלכל  $a > 0$  מתקיים

$$\Pr \left[ |\hat{p} - p| \geq \frac{a}{2\sqrt{n}} \right] \leq \frac{1}{a^2}$$

עבור פרמטר  $\beta > 0$  נסמן  $a = \sqrt{\frac{1}{\beta}}$  ונקבל

$$\Pr \left[ |\hat{p} - p| \geq \sqrt{\frac{1/\beta}{4n}} \right] \leq \beta$$

או במילים אחרות, בהסתברות לפחות  $1 - \beta$  מתקיים ש  $|\hat{p} - p| \leq \sqrt{\frac{1/\beta}{4n}}$

צורה אחרת להסתכל על זה: נניח שעבור פרמטרים  $\alpha, \beta > 0$  מסויימים, אני רוצה להבטיח שבהסתברות לפחות  $(1 - \beta)$  ההערכה שלנו  $\hat{p}$  תהייה קרובה ל- $p$  עד כדי טעות  $\alpha$ . מהו גודל המדגם  $n$  שאני צריך?

כדי לענות על השאלה הזאת בעזרת אי-שוויון האחרון, נדרוש  $|\hat{p} - p| \leq \sqrt{\frac{1/\beta}{4n}} \leq \alpha$ . נפתור עבור  $n$  ונקבל שמספיק לדרוש ש-

$$n \geq \frac{1/\beta}{4\alpha^2}$$

זה מצויין. זה נותן לנו high probability bound כמו שרצינו. אבל אנחנו לא רוצים להסתפק בזה. בדרישה הנ"ל מתקיים ש- $n$  צריך לגדול לנראית עם  $1/\beta$ , שזה לא כל כך טוב אם אנחנו רוצים ש- $\beta$  יהיה פצפון (כלומר אם אנחנו רוצים להיות "סופר בטוחים" בנכונות השערוך שלנו).

מסתבר שבמקרה הזה אפשר לקבל הבטחות הרבה יותר טובות בעזרת מה שנקרא חסמי צ'רנוף/הופדינג.

### משפט (חסמי צ'רנוף והופדינג):

• יהיו  $X_1, X_2, \dots, X_n$  משתנים מקריים בלתי תלויים כאשר לכל  $i$  מתקיים  $\Pr[X_i=1] = p$  ו-  $\Pr[X_i=0] = 1 - p$ , עבור פרמטר  $0 < p < 1$ . תוחלת סכום המשתנים היא  $\mathbb{E}[\sum_{i=1}^n X_i] = p \cdot n$ . אזי מתקיים:

$$\Pr[\sum_{i=1}^n X_i \geq (1 + \alpha) \cdot pn] < \exp(-\alpha^2 pn/4) \quad \text{0} < \alpha < 1 \text{ מתקיים (א)}$$

$$\Pr[\sum_{i=1}^n X_i \leq (1 - \alpha) \cdot pn] < \exp(-\alpha^2 pn/2) \quad \text{0} < \alpha < 1 \text{ מתקיים (ב)}$$

• עבור  $B > A > 0$  יהיו  $X_1, X_2, \dots, X_n \in [A, B]$  משתנים מקריים בלתי תלויים ונסמן  $\mathbb{E}[X_i] = \mu$ . אזי:

$$\Pr[|(\sum_{i=1}^n X_i) - \mu| \geq \delta] \leq 2 \exp\left(-\frac{2\delta^2}{n \cdot (B-A)^2}\right) \quad \text{0} < \delta \text{ מתקיים (ג)}$$

לפני שנדבר על ההוכחה של המשפט הזה, נראה מה הוא נותן לנו עבור הדוגמה שלנו עם המטבע. למשל, נשתמש באי-שוויון (ג) עם  $\delta = n\alpha$  ונקבל

$$\Pr[|\hat{p} - p| \geq \alpha] \leq 2 \exp(-2\alpha^2 n)$$

כדי להבטיח ש- $2 \exp(-2\alpha^2 n)$  יהיה לכל היותר  $\beta$  (עבור פרמטר  $\beta > 0$  כלשהו), מספיק לדאוג ש  $n \geq \frac{\ln(\frac{2}{\beta})}{2\alpha^2}$ . שימו לב שעכשיו התלות של  $n$  ב  $1/\beta$  היא לוגריתמית. זה אומר "שבמחיר" נמוך יחסית מבחית גודל המדגם  $n$  אנחנו יכולים לדאוג ש- $\beta$  יהיה פצפון.

## הוכחת חסמי צ'רנוף/הופדינג

אנחנו נוכיח רק את (א). ההוכחות של (ב), (ג) דומות. יהיו  $X_1, X_2, \dots, X_n$  משתנים מקריים בלתי תלויים כאשר לכל  $i$  מתקיים  $\Pr[X_i = 1] = p$  ו-  $\Pr[X_i = 0] = 1 - p$  ויהי  $0 < \alpha < 1$ . עלינו להראות ש-

$$\Pr\left[\sum_{i=1}^n X_i \geq (1 + \alpha) \cdot pn\right] < \exp(-\alpha^2 pn/4)$$

נסמן  $t = (1 + \alpha)pn$  ונסמן  $c = \alpha/2$ . (נשים לב שמכיוון ש-  $0 < \alpha < 1$  אזי  $0 < c < 1/2$ ). נחשב:

$$\Pr[\sum X_i \geq t] = \Pr[c \cdot \sum X_i \geq c \cdot t] = \Pr[e^{c \cdot \sum X_i} \geq e^{c \cdot t}] = ((1))$$

כעת לפי אי-שוויון מרקוב נקבל ש

$$((1)) \leq e^{-c \cdot t} \cdot \mathbb{E}[e^{c \cdot \sum X_i}] = e^{-c \cdot t} \cdot \mathbb{E}[e^{c \cdot X_1} \cdot e^{c \cdot X_2} \dots e^{c \cdot X_n}] = ((2))$$

כעת מכיוון שהמשתנים  $X_1, \dots, X_n$  הם בלתי תלויים נקבל ש

$$((2)) = e^{-c \cdot t} \cdot \mathbb{E}[e^{c \cdot X_1}] \cdot \mathbb{E}[e^{c \cdot X_2}] \dots \mathbb{E}[e^{c \cdot X_n}] = ((3))$$

ומכיוון ש-  $X_1, \dots, X_n$  מתפלגים אותו הדבר מתקיים  $\mathbb{E}[e^{c \cdot X_1}] = \mathbb{E}[e^{c \cdot X_2}] = \dots = \mathbb{E}[e^{c \cdot X_n}]$  ולכן

$$((3)) = e^{-c \cdot t} \cdot \left(\mathbb{E}[e^{c \cdot X_1}]\right)^n$$

כלומר קיבלנו ש

$$\Pr[\sum X_i \geq t] \leq e^{-c \cdot t} \cdot \left(\mathbb{E}[e^{c \cdot X_1}]\right)^n$$

כעת נשים לב ש-

$$\mathbb{E}[e^{c \cdot X_1}] = p \cdot e^c + (1 - p)e^0 = p \cdot e^c + 1 - p \leq p(1 + c + c^2) + 1 - p = 1 + p(c + c^2) \leq e^{p(c+c^2)}$$

כאשר אי-השוויון הראשון הוא לפי הנוסחה  $e^z \leq 1 + z + z^2$  לכל  $z \leq 1$  וכאשר אי-השוויון השני הוא לפי הנוסחה  $1 + z \leq e^z$  לכל  $z \in \mathbb{R}$ .

נציב זאת בחשבון הקודם שעשינו ונקבל

$$\Pr[\sum X_i \geq t] \leq e^{-c \cdot t} \cdot \left(\mathbb{E}[e^{c \cdot X_1}]\right)^n \leq e^{-c \cdot t} \cdot \left(e^{p(c+c^2)}\right)^n = e^{-c \cdot t} \cdot e^{pn(c+c^2)} = ((4))$$

נזכור שבחרנו  $t = (1 + \alpha)pn$  ולכן נקבל

$$((4)) = e^{-c \cdot (1+\alpha)pn} \cdot e^{pn(c+c^2)} = e^{-c \cdot pn(1+\alpha-c-c^2)} = ((5))$$

נזכור שבחרנו  $c = \alpha/2$  ולכן

$$((5)) = e^{-\alpha^2 pn/4}$$

מ.ש.ל.

### מה קרה בהוכחה הזאת?

אי-שוויון מרקוב אומר שמשנתה מקרי אי שלילי לא יכול לסטות בהרבה מהתוחלת שלו. אבל בחסם צ'רנוף רצינו להראות ש- $\sum X_i$  לא יכול לסטות אפילו בקצת מהתוחלת שלו. לכן במקום להפעיל את אי-שוויון מרקוב ישירות על  $\sum X_i$  חשבנו על המשתנה המקרי  $\exp(\sum X_i)$ .

למה זה טוב? עכשיו מרקוב אומר לנו שהסתברות ש- $\exp(\sum X_i)$  יחרוג בהרבה מהתוחלת שלו היא קטנה וזאת אומרת שהסתברות ש- $\sum X_i$  יחרוג מהתוחלת שלו אפילו בקצת היא קטנה (כי אם  $\sum X_i$  חורג אפילו קצת אז  $\exp(\sum X_i)$  חורג הרבה...)

בנוסף, בזכות העובדה ש- $X_1, \dots, X_n$  הם בלתי תלויים יכולנו לנתח את התוחלת של  $\exp(\sum X_i)$  כי התוחלת הזאת התפצלה לנו למכפלה של תוחלות.

אז למדנו איך להעריך את התוחלת של מטבע לא ידוע. באופן אולי מתפיע, זה כלי סופר שימושי. למשל, נניח שהדטה שלנו מכיל נתונים רפואיים של אנשים, שנדגמו באקראי מאוכלוסיה מסויימת. בנוסף, נניח שיש לנו בדיקה מסויימת אשר בהינתן המידע הרפואי של אדם, אומרת אם הוא חולה במחלה מסויימת או לא. בעזרת אותה שיטה נוכל להעריך, מתוך הדטה, את אחוז החולים בכלל האוכלוסיה.

**הגדרה:** שאילתא סטטיסטית מעל דומיין  $X$  היא פונקציה  $q: X \rightarrow [0,1]$ . עבור התפלגות  $\mathcal{D}$  מעל הדומיין  $X$ , הערך של שאילתא  $q$  על  $\mathcal{D}$  הוא

$$q(\mathcal{D}) := \mathbb{E}_{x \sim \mathcal{D}}[q(x)]$$

עבור דטהבייס  $S \in X^n$ , הערך של  $q$  על  $S$  הוא

$$q(S) := \frac{1}{n} \sum_{x \in S} q(x)$$

בעזרת חסמי צ'רנוף/הופדינג נוכל להעריך את  $q(\mathcal{D})$  בעזרת  $q(S)$ :

**משפט:** תהי  $\mathcal{D}$  התפלגות מעל דומיין  $X$ , תהי  $q: X \rightarrow [0,1]$  שאילתא סטטיסטית כלשהי, ויהי  $\beta > 0$ . עבור מדגם  $S \sim \mathcal{D}^n$  המכיל  $n$  דגימות בלתי תלויות מ- $\mathcal{D}$  מתקיים

$$\Pr \left[ |q(S) - q(\mathcal{D})| \leq \sqrt{\frac{\ln\left(\frac{2}{\beta}\right)}{2n}} \right] \geq 1 - \beta$$

כאשר ההסתברות היא מעל הגרלת  $S$ .

**שאלה:** מה קורה אם יש לנו  $k$  שאילתות סטטיסטיות ואנחנו מעוניינים לקבל הערכה לתוחלת של כל אחת מהן?

**משפט:** תהי  $\mathcal{D}$  התפלגות מעל דומיין  $X$ , יהיו  $q_1, q_2, \dots, q_k: X \rightarrow [0,1]$  שאילתות סטטיסטיות כלשהן, ויהי  $\beta > 0$ . עבור מדגם  $S \sim \mathcal{D}^n$  המכיל  $n$  דגימות בלתי תלויות מ- $\mathcal{D}$  מתקיים

$$\Pr \left[ \max_{i \in \{1, 2, \dots, k\}} |q_i(S) - q_i(\mathcal{D})| \leq \sqrt{\frac{\ln(2k/\beta)}{2n}} \right] \geq 1 - \beta$$

כאשר ההסתברות היא מעל הגרלת  $S$ .

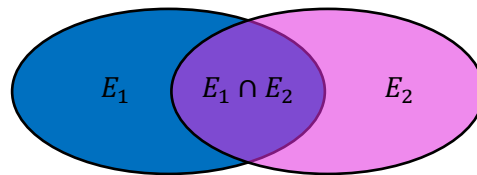
כלומר, עבור פרמטר  $\beta$  וגודל מדגם  $n$  נתונים, השגיאה המקסימלית שלנו, כשאנחנו משערכים את התוחלת של  $k$  שאילתות, גדלה רק לוגריתמית עם  $k$ . בניסוח אחר, נניח שיש טעות מסויימת  $\alpha$  (למשל  $\alpha = 1/100$ ) שאנחנו מוכנים לספוג. אז אנחנו יכולים לשערך את התוחלת של  $k = \frac{2}{\beta} \cdot e^{2\alpha^2 n}$  שאילתות (מספר אקספוננציאלי ב- $n$ ).

## הוכחה:

היזכרו בחסם האיחוד: לכל אוסף של  $k$  מאורעות  $E_1, E_2, \dots, E_k$  (באותו מרחב הסתברות) מתקיים

$$\Pr \left[ \bigcup_{j=1,2,\dots,k} E_j \right] \leq \sum_{j=1}^k \Pr[E_j]$$

בציר:



יהי  $\beta > 0$  כפי שראינו, לפי ההופדינג, עבור כל שאילתא סטטיסטית בודדת  $q_i$  מתקיים

$$\Pr \left[ |q_i(S) - q_i(\mathcal{D})| > \sqrt{\frac{\ln(2k/\beta)}{2n}} \right] \leq \frac{\beta}{k}$$

לכן, לפי חסם האיחוד,

$$\begin{aligned} & \Pr \left[ \max_{i \in \{1,2,\dots,k\}} |q_i(S) - q_i(\mathcal{D})| > \sqrt{\frac{\ln(2k/\beta)}{2n}} \right] = \\ & = \Pr \left[ \left\{ |q_1(S) - q_1(\mathcal{D})| > \sqrt{\frac{\ln(2k/\beta)}{2n}} \right\} \text{ OR } \dots \text{ OR } \left\{ |q_k(S) - q_k(\mathcal{D})| > \sqrt{\frac{\ln(2k/\beta)}{2n}} \right\} \right] \\ & \leq \Pr \left[ |q_1(S) - q_1(\mathcal{D})| > \sqrt{\frac{\ln(2k/\beta)}{2n}} \right] + \dots + \Pr \left[ |q_k(S) - q_k(\mathcal{D})| > \sqrt{\frac{\ln(2k/\beta)}{2n}} \right] \leq \beta \end{aligned}$$

מ.ש.ל.

מצויין. אז במקרה הלא-אדפטיבי אנחנו יכולים לשערך את התוחלת של המון שאילתות בעזרת מדגם קטן יחסית.

מה "לא-אדפטיבי" כאן? השאילתות  $q_1, q_2, \dots, q_k$  נקבעות לפני שמגרילים את המדגם  $S$ .

**שאלה:** האם תוצאה דומה תקפה גם כאשר השאלות נבחרות אחר שהמדגם  $S$  מוגרל, על ידי אנליסט שרואה את המדגם  $S$ ?

**תשובה:** באופן כללי, לא!

**דוגמה:** נניח שהדומיין שלנו הוא  $X = \{0,1\}^d$  ושההתפלגות  $\mathcal{D}$  היא אחידה על פני הדומיין הזה. אחרי שהמדגם  $S$  הוגרל (ואחרי שראינו אותו) נוכל להגדיר את השאלתא הבאה:

$$q_S(x) = \begin{cases} 1, & \text{if } x \in S \\ 0, & \text{if } x \notin S \end{cases}$$

לפי הגדרה מתקיים ש-  $q_S(S) = \frac{1}{n} \sum_{x \in S} q_S(x) = 1$ , אבל  $q_S(\mathcal{D}) = \mathbb{E}_{x \sim \mathcal{D}}[q_S(x)] \leq \frac{n}{2^d}$ .

כלומר, אם לאחר ש-  $S$  נקבע נגדיר את  $q_S$  באופן הזה, אז בהסתברות 1 נקבל ש-

$$|q_S(S) - q_S(\mathcal{D})| \geq 1 - \frac{n}{2^d} \approx 1$$

(בהנחה ש-  $n$  קטן ביחס ל-  $2^d$ )

**שימו לב:** בדוגמה האחרונה נתנו לאנליסט גישה בלתי מוגבלת למדגם  $S$ , מה שאפשר לו להגדיר שאלתא סטטיסטית שעושה overfit חזק: השאלתא מחזירה 1 בדיוק על כל איברי המדגם ומחזירה 0 מחוץ למדגם. המסקנה היא שאם אנחנו רוצים להבטיח משהו עבור המקרה האדפטיבי, אז אנחנו צריכים להגביל איכשהו את המצב שלנו. ישנן 2 גישות עיקריות לכך:

**גישה 1:** נגביל את האנליסט באיזושהי צורה. כלומר, האנליסט מקבל את המדגם  $S$ , אבל אנחנו מניחים שהאנליסט מוגבל באיזושהו אופן במה שהוא יכול לעשות לאחר מכן. למשל,

- אנליסט שמוגבל לבחור שאלות רק ממשפחה מסויימת, כלומר לא כל שאלתא היא "אפשרית".
- כאשר האנליסט מעוניין לבצע סדרת פעולות ספציפית עם המדגם (למשל, קודם לבחור משתנים ואח"כ לבצע רגרסיה על המשתנים האלה), אז לפעמים אפשר לנתח את זה ולהראות שאין סכנה ל overfitting.

**גישה 2:** נגביל את הגישה של האנליסט למדגם  $S$  (אבל מעבר לכך לא נגביל את האנליסט). כלומר, לא ניתן לאנליסט את המדגם  $S$  עצמו, אלא נרשה לו רק גישה מוגבלת באיזושהו אופן ל-  $S$ . מעבר לכך לא נגביל את האנליסט ולא נניח שום דבר על אופן הפעולה שלו.

בקורס שלנו אנחנו נתמקד בגישה 2. היתרונות של גישה 2 הם: (1) זאת גישה יותר כללית (2) יותר קל לאכוף אותה (3) לא צריך להניח שום דבר על מה שהאנליסט יעשה עם הנתונים. היתרון של גישה 1 הוא שהיא מאפשרת פתרונות "תפורים" למקרים ספציפיים, ולכן לפעמים משיגה תוצאות טובות יותר.

**שאלה:** איך אפשר/הגייוני להגביל את הגישה של האנליסט למדגם?

**הצעה:** נאפשר לאנליסט לגשת לנתונים רק על ידי בדיקת הערך האמפירי של שאלות סטטיסטיות לבחירתו.

באופן יותר פורמלי, נחשוב על המשחק הבא עבור התפלגות  $\mathcal{D}$  ואנליסט  $A$

1. הגרל מדגם  $S \sim \mathcal{D}^n$  (האנליסט  $A$  לא מקבל את המדגם  $S$ )

2. עבור  $i = 1, 2, \dots, k$ :

• האנליסט  $A$  קובע שאלתא סטטיסטית  $q_i$

• האנליסט  $A$  מקבל את הערך האמפירי של השאלתא  $q_i(S) = \frac{1}{n} \sum_{x \in S} q_i(x)$



**שאלה:** האם גישה כזאת לנתונים מבטיחה להימנע מ overfitting בהסתברות גבוהה? כלומר, האם זה מבטיח שלא משנה מה האנליסט יעשה, כל עוד הוא ניגש לנתונים בצורה כזאת, אז בהסתברות גבוהה, לכל  $q_i$  כנ"ל מתקיים

$$q_i(S) \approx q_i(\mathcal{D})$$

**תשובה:** זה לא מבטיח את זה. האסטרטגיה הזאת יכולה להיכשל אפילו עבור  $k = 2$ .

**דוגמה:** נניח שהדומיין הוא  $X = \{1, 2, \dots, 2n\}$  ונניח שההתפלגות  $\mathcal{D}$  היא אחידה על פני  $X$ . נראה אנליסט ששואל שאילתא אחת  $q_1$  ולאחר מכן מוצא שאילתא  $q_2$  עבורה  $q_2(S) \gg q_2(\mathcal{D})$ :

השאילתא הראשונה  $q_1$  מוגדרת באופן הבא:

$$q_1(x) = \underbrace{0.000000 \dots 01}_{\text{\#zeroes} = x \cdot \log n}$$

$$\begin{aligned} &0.0001 \\ &+ 0.0000000001 \\ &+ 0.0000000000001 \\ &+ 0.00000000000001 \\ &+ 0.000000000000001 \end{aligned}$$

$$\hline 0.0001000001011$$

שימו לב: מתוך הערך האמפירי  $q_1(S)$  האנליסט לומד את כל המדגם  $S$ , כי הוא "מקודד" בביטים הנמוכים של התשובה. למשל, אם  $S = (1, 3, 4, 4, 4)$  אזי  $n \cdot q_1(S) = \sum_{x \in S} q_1(x)$  הוא:

לאחר שהאנליסט למד את  $S$ , הוא יכול להגדיר את השאילתא  $q_2$  כמו שראינו קודם:

$$q_2(x) = \begin{cases} 1, & \text{if } x \in S \\ 0, & \text{if } x \notin S \end{cases}$$

$$q_S(\mathcal{D}) = \mathbb{E}_{x \sim \mathcal{D}}[q_S(x)] \leq \frac{n}{|X|} = \frac{1}{2} \quad \text{אבל} \quad q_S(S) = \frac{1}{n} \sum_{x \in S} q_S(x) = 1 \quad \text{שוב, מתקיים ש-}$$

### דיון:

1. הדוגמה האחרונה הסתמכה על כך שאנחנו עובדים עם דיוק מאוד גבוה, ונותנים לאנליסט את הערך האמפירי של השאילתא **במדויק**. הדוגמה הזאת נשברת אם נגביל את עצמנו לעבוד עם מספר קטן של ספרות אחרי הנקודה. אולי זה יפתור את הבעיה באופן כללי?

2. האנליסט מהדוגמה האחרונה ניסה בצורה מפורשת לעשות overfit. אולי הבעיה לא קיימת עבור אנליסטים "תמימים"?

### תרגיל בית למחשבה לקראת השיעור הבא – רשות בלבד – לא להגשה:

נניח שהדומיין 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 "התיג" של  $x$ . בהינתן מדגם  $S$  (שנדגם מתוך התפלגות לא ידועה  $\mathcal{D}$  מעל הדומיין  $X$ ), המטרה שלנו היא למצוא פונקציית תיג  $f: X \rightarrow \{0, 1\}$  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-  $\text{error}_D(f)$  קוראים "שגיאת ההכללה" של  $f$

איך נוכל למצוא פונקציה  $f$  כזאת? בתרגיל זה נבחן את האלגוריתם הבא:

**קלט:**  $S = ((x_1, y_1), \dots, (x_n, y_n))$  כאשר  $x_j \in \{0,1\}^d$  וכאשר  $y_j \in \{0,1\}$ .

1. לכל קואורדינטה  $1 \leq i \leq d$ , חשב  $c_i = \frac{1}{n} \sum_{(x,y) \in S} \mathbb{1}\{x[i] = y\}$  כלומר, לכל קואו'  $i$  אנחנו בודקים כמה "טוב" היא חוזה את התיוג

2. נאמר שקואו'  $i$  "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

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

ממשו את האלגוריתם ונתחו את הביצועים שלו. ספציפית, עבור ערכים שונים של  $n, d$ :

- צרו מדגם אקראי  $S = ((x_1, y_1), \dots, (x_n, y_n))$  כאשר כל  $x_j$  נדגם באופן אחיד (ובלתי תלוי) מתוך  $\{0,1\}^d$  וכאשר כל  $y_j$  נדגם באופן אחיד (ובלתי תלוי) מתוך  $\{0,1\}$ .
- הריצו את האלגוריתם על  $S$  ובחנו את  $|P|$  ואת  $\text{error}_S(f)$  וכל פרמטר אחר שנראה לכם רלוונטי.
- מה הקשר בין הביצועים של כלל התחזית  $f$  על המדגם  $S$  והביצועים שלו על ההתפלגות שממנה המדגם הגיע? נסו לחשוב למה