מבני נתונים 15 ~ מבוא להסתברות

שחר פרץ

9 ביוני 2025

מרצה: עמית וינשטיין

מבוא להסתברות

אנחנו מתעניינים בהסתברות כאשר הלאוריתמים/מבני הנתונים אינם דטרמינסטיים. באלגו' הסתברותי, נתעיין בזמן הריצה

כבר היום נשתמש בזה בשביל לנתח באופן יותר מדויק quicksort. אלגוריתם לא דטרמיניסטי – "מטיל מטבעות", ואז מה שרוצים לשאול זה עבור כל פלט, מה אנחנו "מצפים" (ולא רק מה ה־worst case). ה"צפוי" עבור הקלט, על פני הטלת המטבעות של האלג/מבנה.

נגדיר הגדרות של מבוא להסתברות:

הגדרה 1. ניסוי (Experiment): תהליך שבו יש חוסר וודאות לגבי התוצאה

הגדרה 2. מרחב המדגם (Sample Space): קבוצת התוצאות האפשרויות בניסוי. מסומן ב־S או ב־Σ.

דוגמה. עבור קובייה, $[6]$ מהווה את מרחב המדגם. עבור מטבע, {פלי, עץ}. שתי הדוגמאות האלו דיסקרטיות, ויש אוסף סופי של אפשרויות. לצורך העניין, הממשיים $\mathbb{R}$ עבור קבוצת כל הזמנים עד שהמנורה תכבה. לצורכינו, נדבר על מרחבי מדגם סופיים.

הערה. זה קצת מוזר להגדיר דיסקרטי, כי $\mathbb{N}$ לרוב נחשב דיסקרטי, אך $\mathbb{Q}$ לא (אז עוצמות לא הגדרה טובה). זה לא משנה לצורכינו כי נעבוד רק עם מרחבי מדגם סופיים.

הגדרה 3. אירוע (Event): ת"ק של מרחב המדגם.

דוגמאות

1. יצא ערך זוגי
2. הטלנו 2 קוביות ייצא אותו הערך
3. הטעלנו 2 קוביות והסכום ≤ 4
4. הטעלנו קוביה ויצא 4.

הגדרה 4. אירוע פשוט: ת"ק בגודל 1 ממרחב המדרגם

בתור קבוצות, נרצה לחתוך, לאחד, לבדוק זרות/הכלה, וכן למצא משלים (כאשר עולם הדיון הוא תמיד מרחב המדגם). לצורך הדוגמה, ביחס לדוגמאות לעיל:

$$B \cap C = \{(1,1),(2,2)\},\ \ A^C = \{1,3,5\} \quad \text{etc.}$$

הגדרה 5. פונקציית ההסתברות P: היא פונקציה $P\colon \mathcal{P}(S) \to \mathbb{R}$, כאשר הממשיים ההסתברות שלהם. היא מקיימת את התכונות הבאות:

1. $\forall E \subseteq S\colon 0 \leq P(E) \leq 1$
2. $P(S) = 1$
3. $\forall E, F \subseteq S\colon E \cap F = \varnothing \implies P(E \cap F) = P(E) + P(F)$

אינטואיטית, אנחנו נותנים הסתברות לכל אחת מהמאורעות שיכולים לראות.

הגדרה 6. הסתברות מותנה (Conditional Probability): תסומן $P(E \mid F)$ ותוגדר להיות:

$$P(E \mid F) = \frac{P(E \cap F)}{P(F)}$$

ומוגדרת כאשר $P(F) > 0$.

הערה. נוכל לשחק עם המשוואה לעיל, נקבל: $P(E \cap F) = P(F) \cdot P(E \mid F)$

הגיונית, $P(E \cap F)$ מתאר את ההתסבורת של E בהינתן שקרה F.

דוגמה. כד עם 8 כדורים אדומים ו־4 כחולים. מוציאים 22 כדורים ללא חזרות. נגידר: E =הכדור הראשון יצא אדום, F = הכדור השני היה אדום. אז $P(E \cap F) = P(E) \cdot P(F \mid E) = \frac{8}{12} \cdot \frac{7}{11}$.

נוכל להכליל באינדוקציה: ,

$$P\left(\bigcap_{i=1}^{n} E_i\right) = P(E_1) + \prod_{i=2}^{n} P\left(E_i \mid \bigcap_{i=1}^{i-1} E_i\right) = P(E_1) \cdot P(E_2 \mid E_1) \cdots P(E_n \mid E_1 \cap E_2 \cdots \cap E_{n-1})$$

עבור חלוקה $S = F_1 \cup F_2 \cup F_3 \cup \cdots \cup F_n$ זרים בזוגות, היא חלוקה של מרחב המדגם, מתקיים בעבור אירוע E:

$$P(E) = \sum_{i=1}^{n} P(F_i) \cdot P(E \mid F_i) = \sum_{i=1}^{n} P(E \cap F_i)$$

מסקנה:

$$P(E) = P(E \mid F) \cdot P(F) + P(E \mid F^C) \cdot P(F^C) = P(E \cap F) + P(E \cap F^C)$$

הגדרה 7. מאורעות בלתי תלויים (Independent Events): יהיו E, F מאורעות, הם יקראו בלתי תלויים אם $P(E \cap F) = P(E) \cdot P(F)$. בניסוח שקול $P(E \mid F) = P(E)$.

כאשר אירועים זרים, הם ממש ממש תלויים. הסיבה: כי אם אחד קורה, השני בהכרח לא קורה. זה נותן עליו ידע. חוץ ממקרים מנוונים כמו $P(E) = 0$.

הגדרה 8. משתנה מקרי (Random Variable): פונקציה $S \to \mathbb{R}$. אוהבים להשתמש באותיות x, y עבור משתנים מקריים. בקיצור מ"מ.

סימון 1. עבור מ"מ X, נסמן ב־$X = x$ את האירוע שבו X קיבל את הערך x. ואז נובל לכתוב $P(X = x)$.

הערה: $X = x$לאו דווקא מאורע פשוט.

דוגמה. אם X הוא סכום הטלה של 2 קוביות, אז זהו משתנה מקרי. הוא הופך את מרחב המדגם לערך קונקרטי שאפשר להשתמש בו.

הגדרה 9. תוחלת (Expectation): הממוצע המשוקלל של משתנה מקרי, כאשר המשקלים הם ההסתברות. כלומר:

$$\mathbb{E}[x] = \sum_{x \in \operatorname{dom} X} x \cdot P(X = x)$$

הרעיון: להגיד מה הערך ה"ממוצע" של מ"מ. עבור X =ערך הטלת קובייה, מתקיים $\mathbb{E}[x] = 3.5$.

משפט 1. שתי תכונות חשובות:

1. אפשר לחשוב על Y לעיל כמו מ"מ חדש.

$$\mathbb{E}[\underbrace{a \cdot x + b}_{Y}] = a \cdot \mathbb{E}[X] + b$$

2. "יותר מעניינת", נכון עבור כל X, Y זוג מ"מ. (שימו לב, $X + Y$ חיבור פונקציות)

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

התכונה השנייה מאפשר לעשות משהו מאוד נחמד - אם המשתנה המקרי שלנו מסובך במיוחד, לפרק מ"ממים למ"ממים קטנים יותר שהוא שווה לסכומם. בהמשך, לדוגמה, ניקח את המ"מ הוא כמות ההשוואות ב־quicksort, ונפרק אותו לסכום של מ"מים יותר קטנים.

הגדרה 10. מ"מ קינדיקטור: מקבל ערך 1 אם קרא אירוע מסויים, ו־0 אחרת. יסומן ב־$\mathbb{I}_E$.

אזי:

$$\mathbb{E}[\mathbb{I}_E] = 1 \cdot P(E) + 0 \cdot P(E^C) = P(E)$$

"ועוד אפס כפול ההסתברות של אמאשלי" - עמית על קבוצת המשלים.

הגדרה 11. התפלגות אחידה: מרחב מדגם סופי בו כל התוצאות האפשריות בו, (בהסתברות זהה, כלומר התפלגות קבועה.

דוגמה. X מתפלג אחיד בין $a, a+1 \ldots b$ אז $\forall a \le x \le b\colon P(X = x) = \frac{1}{b-a+1},\ \mathbb{E}[x] = \frac{a+b}{2}$

הגדרה 12. התפלגות כיאומטרית: חוזרים על אירוע שמצליחים בהסתברות P כמה פעמים עד ההצלחה הראשונה, והשמתנה המקרי הוא כמות הניסויים.

בעבור X התפלגות גיאומטרית:

$$\forall k > 0\colon P(X = k) = (1-p)^{k+1} \cdot p,\ P(X \ge k) = (1 - p^{k+1})$$

כאשר $(1-p)$ מתאר כשלונות, ו־p את ההסתברות להצלחת E. אז התוחלת:

$$\mathbb{E}[X] = \sum_{\infty}^{j=1} k \cdot P(X = k) = \sum_{i=1}^{\infty} P(X \ge x) = \sum_{k=1}^{\infty} (1-p)^{k-1} = \sum_{k=0}^{]inf} (1-p)^x = \frac{1}{p}$$

QuickSort

נסמן ב־X להיות מ"מ הוא מספר ההשוואות שהאלגו' מבצע על הקלט $x_1x_2\dots x_n$ כאשר $x_i \neq x_j$. אחרי המיון, המערך יראה:

$$z_1 < z_2 < z_3 < \cdots < z_i < \cdots < z_n$$

יהי לנו יותר קל לנתח כשנגע מה נמצא איפה. זכרו שלא משנה מה הקלט עצמו כי הבחירות של האלגו' הם הדבר החשוב. נשים לב ש־z_i, z_j עלולים להיות משווים לכל היותר פעם אחת – כאשר הראשון מבינהם נבחר להיות ה־pivot. בזכות התובנה הזו, נוכל להגדיר $X = \sum X_{i,j}$ כאשר $X_{i,j}$ מתאר האם אנחנו משווי סאת z_i ו־z_j. נוכל לדחוף $\mathbb{E}$ על שני אגפי המשוואה ומאדטיביות רק לחפש את $\mathbb{E}[X_{i,j}]$:

- עור $j = i + 1$, מתקיים $P(i,j) = 1$.
- עבור $i \ll j$ מתקיים $P(i,j)$ אינטואיטיבית קטן
- עבור $i = 1, j = n$ נקבל $P(i,j) = \frac{2}{n}$
- נראה ש־$P(i,j) = \frac{2}{j-i+1}$ נראה הגיוני כי עובד במקרים לעיל. נוכיח את הנוסחה הזו.

הוכחה. נוכיח את הנוסחה באינדוקציה על אורך המערך הנוכחי.

בסיס: $n = 2$, אז $i = 1, j = 2$ בה"כ ואכן מתקיים:

$$P(1,2) = \frac{2}{2-1+1} = \frac{2}{2} = 1$$

צעד: נניח נוכנות עד $n-1$, ונוכיח עבור n. נניח שה־pivot שנבחר הוא z_k. נחלק למקרים:

- אם $i < j < k$, אז z_i, z_j יהיו במערך השמאלי עם אותם האינדקסים ומה.א. נקבל $P(i,j) = \frac{2}{j-i+1}$
- אם $k < i < j$ אז z_i, z_j יהיו במערך השמאלי עם אינדקסים $i-k, j-k$ ולכן:

$$P(i,j) = \frac{2}{(j-k)-(i-k)+1} = \frac{2}{j-i+1}$$

- אם $i \le k \le j$, נשווה רק כאשר $k = i$, או $k = j$, ואחרת לא נשווה אותם בעתיד. לכן $P(i,j) = \frac{2}{j-i+1}$.

■

משפט 2. $\mathbb{E}[x] = O(n \log n)$

הוכחה.

$$\mathbb{E}[X] = \sum_{i<j} P(i,j) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{2}{j-i+1} \overset{k=j-i+1}{=} \sum_{i=1}^{n-1} \sum_{k=2}^{n-i+1} \frac{2}{k} \le 2 \cdot \sum_{i=1}^{n-1} \underbrace{\sum_{k=1}^{n} \frac{1}{k}}_{H_n}$$

ידוע שהטור ההרמוני H_n מקיים $H_n \le \ln(n+1)$. לכן:

$$\le 2 \cdot n \cdot (\ln n + 1) = O(n \log n)$$

■

. . . .