מבני נתונים - תרגול 3
מבוא קצר להסתברות

גלעד אשרוב

13 במרץ 2013

תקציר

בתרגול זה נלמד מבוא קצרצר להסתברות, ונלמד מספר הגדרות חשובות כגון משתנה מקרי, מרחב ההסתברות ופונקציית הסתברות. בנוסף, נלמד את מושג התוחלת. הגדרות אלו נועדו לשמש ככלי עזר לניתוח אלגוריתמים המשתמשים באקראיות.

1 דוגמאות

מכיוון שהסתברות הוא נושא מאוד אינטואיטיבי - נתחיל דווקא מדוגמאות, ולאחר מכן נעבור להגדרות.

דוגמא - הטלת מטבע הוגן. הטלת מטבע הוא תהליך הסתברותי. כאשר מטילים מטבע, יש סיכוי שניפול על "עץ" או על "פלי". בחצי מהמקרים - יוצא "עץ", בחצי מהמקרים - "פלי".
האפשרויות:

- "עץ" - בהסתברות $1/2$.
- "פלי" - בהסתברות $1/2$.

נשים לב שסכום ההסתברויות מסתכם ל - 1.

דוגמא - מטילים 2 מטבעות. אוסף האפשרויות:

- עץ, עץ - בהסתברות $1/4$.
- עץ, פלי - בהסתברות $1/4$.
- פלי, עץ - בהסתברות $1/4$.
- פלי, פלי - בהסתברות $1/4$.

שוב, סכום כל ההסתברויות הוא 1. נשים לב שניתן לשאול שאלות נוספות מאשר מהן ההטלות: ניתן לחשב לדוגמא את ההסתברות שבהטלה הראשונה יצא לנו "עץ". זהו למעשה איחוד המקרים: "עץ, עץ" ו, "עץ פלי", וקורה בסה"כ בהסתברות $1/2$.

דוגמא - הטלת 100 מטבעות. מה הסיכוי שיהיו בדיוק 50 פעמים עץ? כעת, מספר האפשרויות הכולל הוא עצום.. (2^{100} מקרים).

- ע,ע,ע,...,ע,ע - בהסתברות 2^{-100}.
- ע,ע,ע,...,ע,פ - בהסתברות 2^{-100}.
- ...
- פ,פ,פ,...,פ,פ - בהסתברות 2^{-100}.

נתבונן כעת בכל הסדרות שבהם יש בדיוק 50 פעמים עץ. כמה סדרות כאלו יש? $\binom{100}{50}$ (אנחנו בוחרים את 50 המקומות שבהם יהיה עץ, והם כבר מגדירים את 50 המקומות בהם יש פלי). כל רצף כזה קורה בהסתברות 2^{-100}. לכן, בסה"כ:

$$\Pr\left[50 \text{ times H}\right] = \frac{\binom{100}{50}}{2^{100}}$$

דוגמא: לוטו. בהגרלה זו, יש בהס"כ 49 אפשרויות, ומגרילים 6 מספרים מתוך הקבוצה (ללא חזרות). נשים לב שמבחינתנו, המאורע $1,2,3,4,5,6$ שקול למאורע $6,5,4,3,2,1$ (לא אכפת לנו הסדר). שוב, מרחב ההסתברות שלנו הוא גדול:

- $1,2,3,4,5,6$
- $1,2,3,4,5,7$
- $\ldots$

יש בסה"כ $\binom{49}{6}$ רצפים כאלה. כולם שווה הסתברות. לכן, נקבל כי ההסתברות שהמספרים הם $1,2,4,5,6$ (ללא חשבות לסדר)?

$$\Pr\left[\text{the chosen numbers are} 1,2,3,4,5,6\right] = \frac{6}{49}\cdot\frac{5}{48}\cdot\frac{4}{47}\cdot\frac{3}{46}\cdot\frac{2}{45}\cdot\frac{1}{44} = \frac{1}{\binom{49}{6}}$$

דוגמא: טוטו. במשחק זה, ישנם סה"כ 16 משחקים, ושלוש אפשרויות שונות לכל משחק $(1,X,2)$. מותר לנו לבחור את אותה התוצאה מספר פעמים (עם חזרות). שוב, רשימת כל האפשרויות היא גדולה:

- $1,1,1,\ldots,1,1$.
- $1,1,1,\ldots,1,X$
- $1,1,1,\ldots,1,2$
- $\ldots.$

בהנחה שכל משחק מקבל הסתברות שווה לכל אחד מהאפשרויות ("הקבוצות הן שווה כוחות"), ותחת ההנחה שכל המשחקים הם בלתי תלויים (משחקים באותו הזמן?), מהי ההסתברות שכל התוצאות הן 1?

בפשטות, כל ערך מקבל 1 בהסתברות $1/3$. לכן, ההסתברות שכל התוצאות הן 1:

$$\Pr\left[\text{all is } 1\right] = \left(\frac{1}{3}\right)^{16}$$

2 הגדרות

מרחב מדגם. מרחב מדגם של ניסוי הוא אוסף כל התוצאות האפשריות. אנו נבחן מרחבי מדגם בדידים. נסמן את מרחב המדגם ב- Ω.

הגדרה 1. (פונקציית הסתברות) בהינתן מרחב מדגם Ω, פונקציית הסתברות היא פונקציה $\Pr : \Omega \rightarrow [0,1]$ המקיימת: $\sum_{\omega\in\Omega}\Pr(\omega) = 1$.

נשים לב שעבור אותו מרחב מדגם, ניתן להגדיר המון פונקציות הסתברות.

מאורע אלמנטרי. כל איבר $\omega \in \Omega$ ייקרא "מאורע אלמנטרי". דוגמא - בהטלת 100 מטבעות, מאורע אלמנטרי הוא למשל $0\ldots00000$.

מאורע. נשים לב שניתן לשאול שאלות נוספות מלבד מאורעות אלמנטרים. לדוגמא, קובייה. מרחב המדגם שלנו הוא - $\Omega = \{1,2,3,4,5,6\}$. אבל - נוכל לשאול למשל, האם תוצאת ההטלה היא זוגית. כלומר, אנחנו שואלים על המאורע $E = \{2,4,6\}$.

אם כן, מאורע E הוא קבוצה של מאורעות אלמנטריים. נגדיר:

$$\Pr[E] \stackrel{\text{def}}{=} \sum_{\omega\in E}\Pr[\omega]$$

מכאן, נסיק מספר תכונות:

1. מאורעות משלימים. יהי Ω מרחב הסתברות, ויהי A מאורע. נגדיר את המאורע המשלים של A להיות: $\bar{A} = \Omega \setminus A$. קל לראות כי:
$$\Pr\left[\overline{A}\right] = 1 - \Pr[A]$$
לדוגמא, נתבונן על הטלת קובייה. נגדיר את A להיות הסיכוי שייצא פחות מ- 2 (כולל). כלומר- $A = \{1,2\}$. המאורע המשלים $\bar{A} = \{3,4,5,6\}$. מתקיים:
$$\Pr[A] = \frac{1}{3} \qquad \text{and} \qquad \Pr\left[\overline{A}\right] = \frac{2}{3}$$

2. איחוד מאורעות זרים. יהיו A, B מאורעות. לפי הגדרת הסתברות, אם A, B מאורעות זרים, כלומר $A \cap B = \emptyset$, מתקיים:
$$\Pr[A\cup B] = \sum_{\omega\in A\cup B}\Pr[\omega] = \sum_{\omega\in A}\Pr[\omega] + \sum_{\omega\in B}\Pr[\omega] = \Pr[A] + \Pr[B]$$

3. איחוד מאורעות לא זרים. המקרה הקודם הוא למעשה מקרה פרטי של מקרה זה. נניח A, B מאורעות לא זרים, כלומר, $A \cap B \neq \emptyset$. מתקיים $A \cup B = A + B - (A\cap B)$, ולכן:
$$\Pr[A\cup B] = \Pr[A] + \Pr[B] - \Pr[A\cap B] \leq \Pr[A] + \Pr[B]$$

4. חסם האיחוד (union bound): יהיו n מאורעות $A_1,\dots,A_n$. אזי:

$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] \le \sum_{i=1}^{n} \Pr\left[A_i\right]$$

ההוכחה ניתנת כשיעורי בית.

אחד המושגים החשובים ביותר בהסתברות הוא "משתנה מקרי":

הגדרה 2. (משתנה מקרי) בהינתן מרחב מדגם Ω ופונקציית הסתברות $\Pr$, משתנה מקרי X הוא פונקציה $X:\Omega\to\mathbb{R}$.

כלומר, אחנו מתאימים לכל ערך במרחב המדגם שלנו - ערך מספרי. נזכור שמרחב המדגם הוגדר להיות קבוצה כלשהי, ולכן יש צורך להגדיר הגדרה זו, כלומר - להעביר אותנו אל תוך המספרים. ניתן להסתכל על משתנה מקרי פשוט כאל "ערך שתלוי במזל". דוגמאות: בדוגמא עם הטוטו - משתנה מקרי אפשרי הוא מספר התוצאות שניחשתי נכון. משתנה אחר שניתן להגדיר על אותו המשחק - הוא הפרס שאותו אני מקבל עבור הטבלה / הטופס שמילאתי.

תוחלת. בהינתן משתנה מקרי כלשהו, ניתן להסתכל על הממוצע שהוא מקבל. מהו הערך אותו הוא מקבל - בממוצע. זהו למעשה "תוחלת". ניתן לחשוב על ערך התוחלת באופן הבא: אם נבצע את הניסוי שוב ושוב, ממוצע כל הערכים שהמשתנה המקרי שלנו קיבל - זהו התוחלת. זה הערך אותו אנו "מצפים לקבל".

הגדרה 3. (תוחלת - expected value) יהי X משתנה מקרי, המקבל ערכים $x_1,x_2,\dots$. אזי, תוחלת המשתנה X היא:

$$E(X)=\sum_i \Pr[X=x_i]\cdot x_i$$

זהו למעשה ממוצע משוקלל. לדוגמא, נניח אנו מטילים מטבע. בהסתברות $1/2$ הוא "עץ", ובהסתברות $1/2$ הוא "פלי". אנו משחקים במשחק הבא: אנו מטילים את המטבע שוב ושוב, עד אשר מקבלים "עץ". אם קיבלנו "עץ" - נפסיק את המשחק. אם קיבלנו "פלי" - נמשיך לעוד סיבוב. כמה הטלות אנו מצפים שנבצע עד שנקבל "עץ"? אינטואיטיבית, התשובה היא 2. נחשב:

- ההסתברות שנקבל עץ אחרי הטלה אחת היא $1/2$.
- ההסתברות שנקבל עץ אחרי 2 הטלות היא $1/4$ (פלי בהטלה הראשונה, ועץ בשנייה).
- ההסתברות שנקבל עץ אחרי 3 הטלות היא $1/8$ (פלי בשתי ההטלות הראשונות, ועץ בשלישית).
- ההסתברות שנקבל עץ אחרי i הטלות, היא $1/2^i$.

ולכן, תוחלת מספר הסיבובים של המשחק:

$$\begin{aligned}
E(\#\ of\ rounds) &= \sum_{i=1}^{\infty}\frac{1}{2^i}\cdot i=\frac{1}{2}+2\cdot\frac{1}{4}+3\cdot\frac{1}{8}+4\cdot\frac{1}{16}+5\cdot\frac{1}{32}+\dots\\
&= (\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{16}+\dots)\\
&+ \ (\frac{1}{4}+\frac{1}{8}+\frac{1}{16}+\dots)
\end{aligned}$$

$$+ \quad (\frac{1}{8} + \frac{1}{16} + \ldots) +$$
$$\vdots$$
$$= \quad 1 + \frac{1}{2} + \frac{1}{4} + \ldots = 2$$

ליניאריות התוחלת. לפי ההגדרה, קל לראות כי בהינתן שני משתנים מקריים, X, Y, מתקיים:

$$E[X+Y] = E[X] + E[Y]$$

3 תרגילים

שאלה 1. נניח n תאים, ואנו זורקים k כדורים באקראי אל תוך התאים (כל כדור נוחת בכל תא בהתפלגות אחידה).

1. מהי ההסתברות שכל הכדורים ייפלו בתא הימני ביותר?

   פתרון: אנחנו רוצים שכל הכדורים ייפלו בתא ספציפי. לכן, $(1/n)^k$.

2. מהי ההסתברות שכל הכדורים ייפלו באותו התא? (תא כלשהו)

   פתרון: זהו איחוד זר של n מאורעות דומים לסעיף הקודם. לדוגמא, המאורע בו כל הכדורים נופלים בתא הימני ביותר, המאורע בו כל הכדורים נופלים בתא שלידו, וכו'... בסה"כ, יש לנו איחוד של n מאורעות, כל אחד קורה בהסתברות $(1/n)^k$. לכן בסה"כ נקבל שהנ"ל קורה בהסתברות $(1/n)^{k-1}$.

   אפשרות נוספת היא להסתכל על תהליך הזריקה. זורקים כדור ראשון לתא כלשהו. כעת, אנחנו שואלים מהי ההסתברות שכל הכדורים האחרים ייפלו לאותה התא שבו נמצא כבר הגדור. זה קורה בהסתברות $(1/n)^{k-1}$.

3. מהי ההסתברות ששני כדורים נתונים ייפלו לאותו התא?

   פתרון: הכדור הראשון ייפול לתא כלשהו. הכדור השני ייפול לאותו התא בהסתברות $1/n$.

4. מהי ההסתברות ששני כדורים כלשהם ייפלו לאותו התא? כלומר, מהי ההסתברות שתהיה איזושהי "התנגשות"?

   פתרון: בסעיף הקודם שאלנו ־ מהי ההסתברות ששני כדורים כלשהם נופלים לאותו התא. כאן, אנחנו מסתכלים על סך האפשרויות - כלומר, על כל זוגות הכדורים. יש בסה"כ $\binom{k}{2}$ זוגות. ולכן:

$$\Pr\left[collision\right] = \Pr\left[\bigvee_{i \neq j} \text{balls i,j collide}\right] \leq \sum_{i,j} \Pr\left[\text{balls i,j collide}\right] = \binom{k}{2}\frac{1}{n} \leq \frac{k(k-1)}{2n}$$

שאלה 2. באוניברסיטה מסוימת במרכז הארץ, נלמד הקורס "מבני נתונים". הסטודנטים בקורס קיבלו תרגיל לשיעורי בית ובו רשימה של פונקציות שעליהם למיין לפי סדר גודל. בנוסף, הסטודנטים התבקשו לבחור מהרשימה 5 זוגות של פונקציות, ולהוכיח לגביהן טענה מסוימת.

1. בהינתן שברשימה היו סה"כ 16 פונקציות - כמה זוגות של פונקציות ניתן להרכיב בסה"כ?
2. תחת ההנחה שכל סטודנט בחר באקראי בהתפלגות אחידה ובצורה בלתי תלויה את חמשת הזוגות, מה ההסתברות ששני סטודנטים נתונים יבחרו **בדיוק** את אותם חמשת הזוגות?
3. אם ישנם 100 סטודנטים בקורס בסה"כ, מצא חסם עליון להסתברות שישנם שני סטודנטים כלשהם עם אותם בדיוק חמש זוגות?
4. **בונוס:** בהינתן התוצאות שלעיל, מה מתרגלי הקורס עלולים להסיק אם נמצאו שני סטודנטים כלשהם עם אותם חמש זוגות?

פתרון:

1. ניתן להרכיב $\binom{16}{2} = 16 \cdot 15/2 = 120$ זוגות בסה"כ.
2. אנחנו קובעים את הבחירה של הסטודנט הראשון, ושואלים מה ההסתברות שהסטודנט השני בוחר בדיוק אותן הפונקציות. כלומר - אנחנו שואלים מה ההסתברות שהסטודנט השני בחר חמש זוגות ספציפיים מתוך רשימה של 120 זוגות. נקבל:

$$\Pr\left[S_2 \text{ chooses the same functions as } S_1\right] = \frac{1}{\binom{120}{5}} = \frac{5!}{120 \cdot 119 \cdot 118 \cdot 117 \cdot 116} = 5.24 \cdot 10^{-9}$$

3. עבור שני סטודנטים נתונים, ההסתברות ששניהם בחרו בדיוק אותן שתי פונקציות היא $5.24 \cdot 10^{-9}$. ההסתברות שישנם שני סטודנטים כלשהם שבחרו את אותן הפונקציות חסומה ע"י:

$$\begin{aligned}&\Pr\left[\exists S_1, S_2 \text{ that chose the same 5 functions}\right]\\ \leq\ &\binom{100}{2} \Pr\left[S_2 \text{ chooses the same functions as } S_1\right] = 4950 \cdot 5.24 \cdot 10^{-9} = 2.59 \cdot 10^{-5}\end{aligned}$$

שאלה 3. נניח מבחן אמריקאי ובו 10 שאלות, ו־4 תשובות אפשריות לכל שאלה, כאשר רק אפשרות אחת היא נכונה. בכל הסעיפים אנו מניחים שהסטודנט מנחש באקראי את התשובות הנכונות.

1. מהי ההסתברות שסטודנט יקבל 100 בבחינה?

פתרון: ישנה רק תשובה אחת לכל שאלה. הסטודנט מנחש תשובה נכונה לכל שאלה בהסתברות $1/4$. נקבל בסה"כ:

$$\Pr\left[\text{grade} = 100\right] = \left(\frac{1}{4}\right)^{10} = \frac{1}{2^{20}} = \frac{1}{1048576}$$

2. מהי ההסתברות שהציון הוא בדיוק 90?

פתרון: נפתור בשני שלבים. ראשית נשאל ־ מההי ההסתברות שתשעת השאלות הראשונות הן נכונות, והאחרונה אינה נכונה. ההסתברות שתשעת השאלות הראשונות הן נכונות היא $(1/4)^9$ ו־שהאחרונה נכונה היא $(1/4)$. לכן, בסה"כ נקבל: $3 \cdot \left(\frac{1}{4}\right)^{10}$.
כעת, ייתכן שתשעת התשובות הנכונות הן 8 הראשונות, יחד עם האחרונה (כלומר, התשובה שאינה נכונה היא התשובה התשיעית). גם כן ההסתברות שמקרה זה ייקרה היא $3 \cdot \left(\frac{1}{4}\right)^{10}$.
באופן כללי, אנחנו צריכים לקבוע את הסדר של התשובה הנכונות ושל אלו שאינן נכונות. ישנם $\binom{10}{9}$ אפשרויות לבחור את התשובות הנכונות. נקבל:

$$\Pr\left[\text{grade} = 90\right] = \binom{10}{9}\left(\frac{1}{4}\right)^{9} \cdot \left(\frac{3}{4}\right)^{1}$$

3. במקרה הכללי, מהי ההסתברות לקבל $c \cdot 10$ (עבור איזשהו $0 \leq c \leq 10$)?

פתרון: בדיוק כמו המקרה הקודם, אנחנו צריכים קודם לבחור את c התשובות הנכונות, ואת $10 - c$ התשובות הלא נכונות. אחרי שבחרנו את הסדר, אנחנו פשוט שואלים מהי ההסתברות שבאמת זה המקרה. נקבל:

$$\Pr\left[\text{grade} = c \cdot 10\right] = \binom{10}{c}\left(\frac{1}{4}\right)^{c} \cdot \left(\frac{3}{4}\right)^{10-c}$$

בדיקה: נשים לב שסכום ההסתברויות לקבל כל אחד מהציונים אמור להסתכם ל־1. נוכל לבדוק אם החישוב שעשינו הוא הגיוני:

$$\begin{aligned}\Pr\left[0 \leq \text{grade} \leq 100\right] &= \Pr\left[\bigvee_{c=0}^{10} \text{grade} = c \cdot 10\right] = \sum_{c=0}^{10} \Pr\left[\text{grade} = c \cdot 10\right] \\ &= \sum_{c=0}^{10}\binom{10}{c}\left(\frac{1}{4}\right)^{c} \cdot \left(\frac{3}{4}\right)^{10-c} = \left(\frac{1}{4} + \frac{3}{4}\right)^{10} = 1^{10} = 1.\end{aligned}$$

כאשר השלב האחרון הוא נכון לפי נוסחאת הבינום (נזכור: $(a+b)^n = \sum_{k=0}^{n}\binom{n}{k}a^k b^{n-k}$).

4. מהי תוחלת הציון?

פתרון: סתם כך, בלי חישוב - נראה שהתשובה היא 25 (אם סה"כ הסתברות לצדוק בכל שאלה היא $1/4$, ויש בסה"כ 100 נקודות, מצפים לקבל ממוצע של 25). פורמלית צריך לחשב:

$$\begin{aligned}E[\text{grade}] &= \sum_{c=0}^{10} \Pr\left[\text{grade} = 10 \cdot c\right] \cdot (10 \cdot c) \\ &= \sum_{c=0}^{10}\binom{10}{c}\left(\frac{1}{4}\right)^{c} \cdot \left(\frac{3}{4}\right)^{10-c} \cdot (10 \cdot c) \\ &= 10\sum_{c=1}^{10} c \cdot \binom{10}{c}\left(\frac{1}{4}\right)^{c} \cdot \left(\frac{3}{4}\right)^{10-c} \qquad (1)\end{aligned}$$

$$
\begin{aligned}
&= 100 \cdot \frac{1}{4} \sum_{c=1}^{10} \binom{9}{c-1} \left(\frac{1}{4}\right)^{c-1} \cdot \left(\frac{3}{4}\right)^{10-c} \qquad (2) \\
&= 25 \cdot \sum_{c=0}^{9} \binom{9}{c} \left(\frac{1}{4}\right)^{c} \cdot \left(\frac{3}{4}\right)^{9-c} = 25 \cdot \left(\frac{1}{4} + \frac{3}{4}\right)^{9} = 25
\end{aligned}
$$

כאשר שלב (1) נכון מכיוון ש:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n \cdot (n-1)!}{k \cdot (k-1)! \cdot (n-k)!} = \frac{n}{k} \cdot \frac{(n-1)!}{(k-1)! \cdot (n-k)!} = \frac{n}{k} \cdot \binom{n-1}{k-1}$$

והשלב האחרון כאן נכון מכיוון:

$$\binom{n-1}{k-1} = \frac{(n-1)!}{(k-1)!(n-1-(k-1))!} = \frac{(n-1)!}{(k-1)!(n-k)!}$$

את כל החישוב המסובך הנ"ל ניתן לחסוך בקלות בזכות לינאריות התוחלת. נסמן ב - X את תוחלת הציון. נסמן ב - X_i את תוחלת מספר הנקודות שהסטודנט מקבל עבור שאלה i. אזי, ברור כי: $X = \sum_{i=1}^{n} X_i$. כמו־כן, ברור כי: $E[X_i] = \frac{1}{4} \cdot 10 + \frac{3}{4} \cdot 0 = 2.5$. לכן:

$$E[X] = E\left[\sum_{i=1}^{10} X_i\right] = \sum_{i=1}^{10} E[X_i] = \sum_{i=1}^{10} 2.5 = 25$$

4 לקריאה נוספת

חלקים מהסיכום מבוססים על:
http://moodle.cs.huji.ac.il/cs10/mod/resource/view.php?id=4839.
בנוסף, ניתן להסתכל בספר - "מבוא לאלגוריתמים", פרק 6 (אם כי, ממש לא כיסינו כאן את כל הפרק).

פסח כשר ושמח!