6.12.2009

תכנון וניתוח אלגוריתמים
הרצאה 7

אלגוריתמים הסתברותיים

המרצה: בועז פת-שמיר
רשימות: יונתן נתיב

בעיית המזכירה *(Secretary problem)*

מקרה לדוגמא:
נניח שאנו רוצים לקחת אדם לעבודה. בחדר ההמתנה יש N מועמדים.
תוך כדי הראיון אנו קובעים את ציון ההתאמה של המועמד.

התנאים:
1) הציונים יכולים להיות כל מספר – לא חסום, בלי התפלגות א-פריורית.
2) בסוף הראיון אנו מחוייבים להגיד לאדם אם הוא התקבל (לא מראינים את השאר), או לא התקבל (אין אפשרות להחזיר אותו).
3) הצלחה מוגדרת אם בחרנו את המועמד הטוב ביותר.
בעיה זו נקראת בעיית המזכירה.

הגדרת הבעיה:
- קלט: n מועמדים
- שיטה: קוראים לכל מועמד, מוצאים את ערכו, מחליטים האם לקבלו או לא. אם המועמד התקבל נגמר המשחק.
- המטרה: לשכור את הטוב ביותר.

טענה: אין אלגוריתם דטרמיניסטי אשר תמיד פותר את בעיית המזכירה בהצלחה.

הוכחה:
בהינתן אלג' A כלשהו, נראה כי קיים קלט עליו A נכשל.
נקרא ערכים ל n-1 מועמדים ראשונים כלשהם.
אם A בחר לשכור מועמד מבין n-1 המועמדים הראשונים נגדיר את המועמד ה n כבעל ציון התאמה גבוה מכל שאר המועמדים. אם A לא בחר מועמד עד כה ניתן למועמד ה n ציון נמוך מהמקסימום עד כה.
בכל מקרה האלג' נכשל.

ראינו שאלגוריתם דטרמיניסטי לא עובד. הדבר המפתיע שיש אלגוריתם שכן עובד, ע"י איסוף סטטיסטיקה על הראיונות שבוצעו.

אלגוריתם פשוט שמצליח בהסתברות ¼ לבחור את המועמד הטוב ביותר:

1. האלגוריתם מסדר המועמדים בפרמוטציה אקראית.
2. מראיין את החצי הראשון של המועמדים וזוכר את הציון הגבוה ביותר. נסמנו h.
3. מראיין את החצי השני ושוכר את הראשון שציונו גבוה מ h.

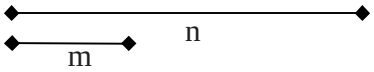
ניתוח האלגוריתם:

- בהסתברות ½ , הטוב ביותר בחצי השני
- בהסתברות $\frac{n/2}{n-1}$ השני הכי טוב בחצי הראשון
- כששני התנאים מתקיימים (בהסתברות 1/4~), האלגוריתם הצליח.

נראה כיצד ניתן לשפר את האלגוריתם הפשוט

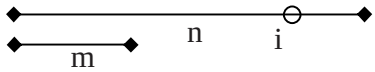
אופטימיזציה:

1. כמקודם.
2. מראיין m מועמדים, זוכר את הטוב.
3. כמקודם.



בפשטות, אנחנו הופכים את הבעיה ליותר כללית, מראינים m מועמדים במקום $\frac{n}{2}$.

השאלה היא מה ה-m שכדאי לקחת, ואת זה נחליט ע"י ניתוח מדוייק של הסתברות ההצלחה.



נבצע ניתוח בהנחה שהמקסימום נמצא במקום ה I (נשים לב שהצלחה תתקבל אם האיבר השני בגובהו מתוך I האיברים נמצא ב m האיברים הראשונים).

$$\Pr[\text{success}] \overset{\text{bayes}}{=} \sum_{\underbrace{i=m+1}_{\text{no chance for success when } i<m}}^{n} \Pr[success \mid best\, in\, position\ i] * \Pr[best\, in\, position\ i] =$$

$$\sum_{i=m+1}^{n} \Pr[\text{second best of } 1..i \text{ in pos } 1..m] \cdot \Pr[best\, in\, position\ i]$$

$$\Pr[best\, in\, position\ i] = \frac{1}{n}$$

$$\Pr[\text{second best of } 1..i \text{ in pos } 1..m] = \frac{m}{i-1}$$

$$\Pr[\text{success}] = \frac{1}{n}\sum_{i=m+1}^{n}\frac{1}{n}\frac{m}{i-1} = \frac{1}{n}\sum_{i=m+1}^{n}\frac{m}{i-1} = \frac{m}{n}\sum_{i=m+1}^{n}\frac{1}{i-1} =$$

$$= \frac{m}{n}\left(H_n - H_m\right) \approx \frac{m}{n}\left(\ln(n) - \ln(m)\right) = \frac{m}{n}\ln\left(\frac{n}{m}\right)$$

נרצה לעשות אופטימיזציה – למקסם את התוצאה שקיבלנו. לשם כך נציב $x=\frac{m}{n}$,

$$f(x)=x\cdot\ln\left(\frac{1}{x}\right)$$

$$f'(x)=\ln\left(\frac{1}{x}\right)+x\cdot\frac{1}{1/x}\cdot\frac{-1}{x^2}=\ln\frac{1}{x}-1$$

$$f''(x)=\frac{1}{1/x}\frac{1}{x^2}<0\quad\forall x>0$$

נשווה נגזרת ל-0 למצוא מקסימום:

$$\ln\left(\frac{1}{x}\right)=1$$

$$x=\frac{1}{e}$$

ואז הסתברות ההצלחה: $\frac{1}{e}\ln\left(\frac{1}{e}\right)=\frac{1}{e}$ - <u>בערך 37%!</u>

<u>בעיית אספן הקלפים coupon collector</u>

- <u>קלט:</u> N סוגי קלפים
- <u>שיטה:</u> בכל ניסוי מקבלים קלף באקראי כשההסתברות לסוג i היא 1/n לכל i
- <u>מטרה:</u> צריך לאסוף כל הסוגים
- <u>שאלה</u> : כמה ניסויים יש לבצע עד אשר מחזיקים את כל הקלפים?

<u>הבעיה במילים</u>: קונים קורנפלקס ואוספים קלף כל פעם. כדי לקבל פרס, יש לאסוף את כל סוגי הקלפים. כמה קופסאות צריך לקנות כדי להשלים את הסדרה?

<u>תשובה I תוחלת:</u>

נגדיר משתנה אקראי X_j $1\le j\le n$ x_j – אורך הזמן מהרגע שיש j סוגים שונים ועד שיש j+1 סוגים שונים.

נשאל: מהי התוחלת $E\left[X_j\right]$?

הסיכוי שקלף חדש אינו אחד מ j הסוגים שכבר התקבלו הוא: $\frac{n-j}{n}$

זהו משתנה גיאומטרי עם $p=\frac{n-j}{n}$ ולכן התוחלת שלו $\frac{n}{n-j}$.

לפי ליניאריות של התוחלת : $E\begin{bmatrix} \#cards_to \\ complete_all \end{bmatrix} = E\left[\sum_{j=0}^{n-1} x_j\right] = \sum_{j=0}^{n-1} E\left[x_j\right]$

$$\sum_{j=0}^{n-1} E\left[x_j\right] = \sum_{j=0}^{n-1} \frac{n}{n-j} = nH_n$$

<u>תשובה II "הסתברות גבוהה":</u>

נרצה לאמר כעבור כמה זמן נסיים בהסתברות כלשהי להצלחה ← מה ההסתברות שלא נקבל קלף מסוג I מסוים ב t ניסיונות?

$$\left(1-\frac{1}{n}\right)^t = \left[\left(1-\frac{1}{n}\right)^n\right]^{\frac{t}{n}} < e^{-\frac{t}{n}}$$

נדרוש: $P\begin{bmatrix} card_\#i_missing \\ in_t_tries \end{bmatrix} < \frac{1}{n^c}$ עבור c > 0 כלשהו.

הסתברות פולינומים קטנה, מספיק:

$$e^{-\frac{t}{n}} < n^{-c}$$

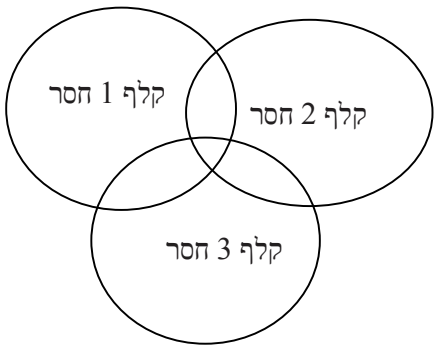
$$\frac{-t}{n} < -c\ln n$$

$$t > c \cdot n \cdot \ln n$$

$$P\begin{bmatrix} all_cards \\ in_t_tries \end{bmatrix} = 1 - P\begin{bmatrix} card_\#i_missing \\ in_t_tries \end{bmatrix}$$

נשתמש בחסם "ברברי" – חסם האיחוד (union bound):

במקום לסכום ההסתברות על איחוד השטחים
(סיכמנו את השטחים החופפים מספר פעמים)



$$1 - P\begin{bmatrix} card_\#i_missing \\ in_t_tries \end{bmatrix} \geq 1 - \sum_{i=1}^{n} P\begin{bmatrix} card_\#i \\ is_missing \end{bmatrix} > 1 - ne^{-\frac{t}{n}} > 1 - n^{-c+1}$$

$$for: \quad t = cn\ln n$$

הגדרה: מאורע X קורה בהסתברות גבוהה אם P[x]=1-o(1) (ההסתברות לכישלון פחותה מקבוע).
בד"כ, לצרכי חסם האיחוד: $c > 0 \quad P[x] = 1 - O(n^{-c})$

הטלת מטבע הוגן עד לקבלת "עץ"

הגדרה: בהטלת מטבע הוגן נרצה בהסתברות גבוהה לקבל עץ.
מס' ההטלות הדרוש כדי שבהסתברות לפחות $1-\delta$ נקבל "עץ".

$$P\begin{bmatrix} no_pally \\ in_t_throws \end{bmatrix} = \left(\frac{1}{2}\right)^t$$

נדרוש $\left(\frac{1}{2}\right)^t < \delta$ ונקבל $\begin{matrix} t > -\log_2 \delta \\ t > \log_2 \frac{1}{\delta} \end{matrix}$

לקבלת $\delta = n^{-c}$ צריך להטיל $t > \log_2 n^c = c \log_2 n$

אם הסתברות הצלחה בנסיון בודד היא p אז $P\begin{bmatrix} t \\ failures \end{bmatrix} = (1-p)^t$ ואז $\delta = (1-p)^t$

$\log\delta < t\log(1-p)$ כלומר: $t > \frac{\log\delta}{\log(1-p)}$

- הניתוח מתבסס על חוסר תלות בניסויים.

(אם נשנה את הניסוי למטבע שכל הטלה חוזרת על תוצאת ההטלה הראשונה התוחלת לא תשתנה אך הבטחה לסיכוי הצלחה מסוים היא לא פשוטה).

MAX3SAT

קלט: נוסחת 3CNF : m פסוקיות, כ"א 3 ליטרלים (משתנה או שלילתו).
נסמן n – מספר משתנים.

מטרה: לספק כמה שיותר פסוקיות.

ננסה להציב לכל המשתנים True או לכל המשתנים False, אחת משתי ההצבות הללו בהכרח תספק יותר ממחצית הפסוקיות (כל הפסוקיות השליליות הופכות לחיוביות בהצבה הנגדית).

נבדוק השמה אקראית: ניתן F/T בהסתברות ½ לכל משתנה.
הסתברות שפסוקית מקבלת שקר – 1/8 (הסתברות חצי עבור כל ליטרל בפסוקית).
מלינאריות של תוחלת: מספר הפסוקיות המסופקות = $m \cdot \frac{7}{8}$.

אי שוויון מרקוב (מורחב):
אינטואיציה – אם תוחלת גובהו של אדם הוא 1.75, אז פחות מ ½ מהאנשים הם בגובה מעל 3.5 (אחרת התוחלת היתה מעל 1.75)

פורמלית:
יהי x משתנה אקראי דיסקרטי, וידוע ש $0 \le X \le \beta$.
רוצים לחשב p=Pr[x≤L] ל L נתון.

$$E[x] = \sum_{j=0}^{B} j \cdot \Pr[x = j] = \sum_{j=0}^{L} j \cdot \Pr[x = j] + \sum_{j=L+1}^{B} j \cdot \Pr[x = j]$$

$$\le \sum_{j=0}^{L} L \cdot \Pr[x = j] + \sum_{j=L+1}^{B} B \cdot \Pr[x = j]$$

$$p = \Pr[x \le L]$$

$$L \cdot p + B(1-p)$$

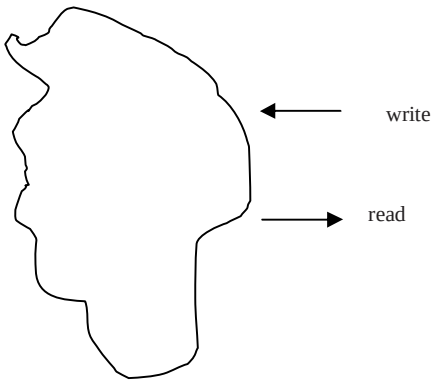
$$\Rightarrow p \le \frac{B - E[x]}{B - L}$$

מה ההסתברות שהשמה אקראית תספק לפחות $m \cdot \frac{7}{8}$ פסוקיות?

נשתמש באי שוויון מרקוב
X – מס' פסוקיות מסופקות
7/8m – E[x]
m – B
(7/8-e)m – L
ולכן:

$$\Pr[less then \frac{7}{8}m\ sat.] \le \frac{m-\frac{7}{8}m}{m-\left(\frac{7}{8}m-\varepsilon\right)} = \frac{1/8}{1/8+\varepsilon} = \frac{1}{1+8\varepsilon} = 1-\frac{8\varepsilon}{1+8\varepsilon}$$

מערכת קוורומים Quorum system



יש חווה של n שרתים לא אמינים.
השרתים נותנים זכרון משותף. יש פעולות כתיבה וקריאה
בכתיבה נכתוב יחד עם המידע את זמן הכתיבה. בקריאה ניקח את הערך המעודכן ביותר.

Quorum system:
קבוצה X של שרתים |X|=n

מערכת קוורומים: $Q=\{S_1..S_n\}$, $S_i \in X$
כך שלכל $S,S' \in Q$ מתקיים $S \cap S' \neq \phi$.

דוגמא טריויאלית:
- הקוורום בנוי משרת יחיד $Q=\{\{1\}\}$
- יש קוורום אחד בלבד
- המערכת פגיעה לנפילות

דוגמא:
2 קבוצות שכל אחת מכילה לפחות חצי מהאיברים: $Q=\left\{S:|S|>\frac{n}{2}\right\}$

החסרון: בכל קריאה/כתיבה, יש להגיע למספר לינארי של מעבדים, ולכן עושים הרבה עבודה. $\Omega(n)$ - חסם תחתון של עבודה לגישה.
בכל כתיבה, תופסים איזשהו קוורום וכותבים לכל המעבדים בקוורום זה עם timestamp. כדי לקרוא, בוחרים קוורום וקוראים מכל המעבדים בו.

בהנתן מערכת קוורומים (מ"ק) נגדיר:
עמידות: #מעבדים מינימאלי שבלעדיהם אין קוורום חוקי (עמידות לתקלות).
עומס: עומס לשרת (תחת בחירה אקראית) – כמה עבודה (מס' הגישות) כל שרת כזה עושה?
נניח $p \in \{S_1...S_k\}$ ו- |Q|=m.

אז העומס על שרת p הוא k/m. (ניגשים לכל המעבדים בקוורום והשרת נמצא ב k מתוך m הקוורומים)
עומס המערכת – עומס מקסימלי לשרת (אם בוחרים קוורום באקראי, מה תוחלת מקסימום הגישות למחשב).
בדוגמא השניה, עומס המערכת הוא בערך ½, כי בכל פעולה יש סיכוי ½ שניגש לשרת מסויים (שנמצא ב ½ מהקוורומים).

זמן: תוחלת גודל הקוורום (מכיוון שכותבים/קוראים מכולם).

דוגמאות:
Q={ {1} } עמידות =1, עומס=1, זמן=1.
Q={ {1,2,3} } עמידות =3, עומס=1, זמן=3.
Q={ S : |S|=(n+1)/2 } עמידות = n/2, עומס = ½, זמן= n/2.

נרצה לסדר את תתי הקבוצות באופן כזה שיביא למקסימום את העמידות ולמינימום את העומס וזמן.

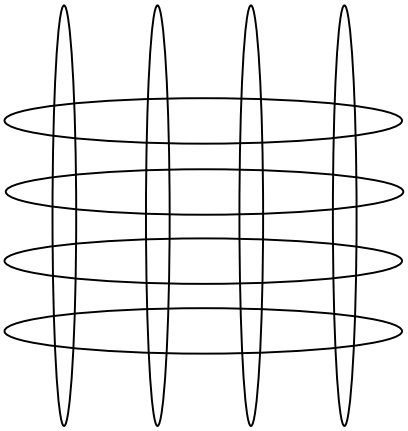
<u>תכונה:</u>
תהי Q מ"ק עם עומס L (התוחלת של הכי גרוע),
קוורום מינימאלי בגודל C, ו- n שרתים סה"כ.
אז $L \geq \max\left(\frac{1}{c}, \frac{c}{n}\right)$.

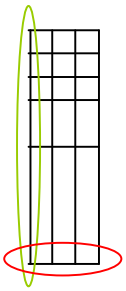
<u>הוכחה:</u>
כיוון שכל קוורום, גודלו ≤ C,
בכל גישה מתבצעות לפחות C פניות לשרתים, לכן $L \geq \frac{c}{n}$.
בנוסף נסתכל על קוורום בגודל מינימלי:
בכל גישה לפחות אחד משרתיו משתתף ==> עומס שרתיו הממוצע ≤ C. ■

<u>מסקנה:</u> העומס ≤ $\frac{1}{\sqrt{n}}$

בשביל לקבל מערכת אופטימאלית, נרצה ששני הביטויים יהיו זהים ולכן נבחר c=√n, ואז L≥1/√n.

<u>מערכת אופטימלית</u>





נניח $n=C^2$.

נסדר השרתים בריבוע CxC.

הקוורומים: שורות/עמודות.

שורות לכתיבה, עמודות לקריאה.

מערכת קוורומים ממש: כל קוורום: שורה + עמודה. כל פעם שנרצה לכתוב למשל, נבחר שורה ונכתוב בכולה.

$$L = \frac{2\sqrt{n}-1}{n} \approx \frac{2}{\sqrt{n}}$$

פגיעות: $\sqrt{n}$ (כל עמודה או שורה שנפלו מהמערכת).

עד כאן המערכת דטרמיניסטית..

<u>מערכת הסתברותית:</u>

נדרוש שלכל $S, S' \in Q$, $Pr[S \cap S' \neq \emptyset] > 1-\delta$, ל $\delta < 1$ נתון.

נבחר קוורומים אקראיים בגודל C.

$$\Pr[S \cap S' = \phi] = \left(\frac{n-|S|}{n}\right)^{|S'|} = \left(1-\frac{c}{n}\right)^{C} = \left[\left(1-\frac{c}{n}\right)^{\frac{n}{c}}\right]^{C^2/n} \approx e^{-C^2/n}$$

$$\delta > e^{-C^2/n}$$

$$\ln\delta > -\frac{C^2}{n}$$

$$\ln\frac{1}{\delta} < \frac{C^2}{n}$$

$$\sqrt{n\ \ln\frac{1}{\delta}} < C$$

אם הסתברות הכשלון פולינומית קטנה, אז מספיק $C > O(\sqrt{n}, \sqrt{\log n})$

פרדוקס יום ההולדת Birthday Paradox

מה הסיכוי שבקבוצה אקראית של t אנשים יש שנים עם אותו יום הולדת?
בשנה יש n ימים וכולם שווי הסתברות.

תוחלת: $E\begin{bmatrix} \#\,Pairs\,with \\ same\,birthday \end{bmatrix} = \sum_{i=1}^{t}\sum_{j=i+1}^{t} P[i,\,j\,have\,same\,BD] = \sum_{i=1}^{t}\sum_{j=i+1}^{t}\frac{1}{n} = \frac{t(t+1)}{2}\frac{1}{n} \approx \frac{t^2}{2n}$

כלומר אם $t=\sqrt{2n}$ התוחלת בערך 1.

שתי קבוצות אקראיות מתוך n איברים יחתכו בהסתברות קבועה כאשר גודלן $\theta(\sqrt{n})$.

למשל, אם יש קבוצה שלא יודעים מה גודלה אבל ניתן להוציא ממנה איברים. נוציא איברים עד שנקבל את אותו האיבר פעמיים. אם ניסינו T פעמים, גודל הקבוצה הוא כנראה כ- T^2.

מה תוחלת מספר ימי ההולדת המשותפים ל t אנשים?

באמצעות משתני אינדיקטור:

נגדיר $X_{ij} = \begin{cases} 1 - Person\ i\ \ and\ \ j\ have\ same\ birthday\ date \\ 0 - otherwise \end{cases}$

נבחר - $1 \le i \le j < t$

ואז $X = \sum_{ij} X_{ij}$

$Pr[X_{ij}=1]=1/365$

הסיכוי שאין התנגשות?

הראשון: 1

השני: $\frac{364}{365}$

השלישי: $\frac{363}{365}$

$$\Pr[no\,collision] = \prod_{i=0}^{t-1}\left(1-\frac{i}{365}\right) = \prod_{i=0}^{t-1}\left[\left(1-\frac{i}{365}\right)^{\frac{365}{i}}\right]^{\frac{i}{365}} \approx \prod_{i=0}^{t-1} e^{-\frac{i}{365}} = e^{\left(-\sum_{i=0}^{t-1}\frac{i}{365}\right)} = e^{\left(-\frac{t(t+1)}{2\cdot 365}\right)}$$

ולכן כאשר $t=\sqrt{30}$ הסיכוי קבוע.