מבני נתונים ואלגוריתמים - תרגול 8

18 בדצמבר 2011

מיון טופולוגי

יש באתר הסבר איך עושים מיון טופולוגי עם DFS.

אנטרופיה - הגדרה

יהי x משתנה מקרי המקבל מספר סופי או בן מניה של ערכים $\{a_n\}$ ויהי $p_n = P\left(x = a_n\right)$ אז מגדירים את האנטרופיה של x להיות:

$$H\left(x\right) = -\sum_{n} p_n \lg p_n$$

אם x משתנה מקרי רציף עם פונק' צפיפות $f\left(x\right)$ אז האנטרופיה מוגדרת כך:

$$H\left(x\right) = -\int_{-\infty}^{\infty} f\left(x\right) \lg f\left(x\right)$$

תכונות

- $H\left(x\right) \geq 0$
- $H\left(x\right)$ לא תלויה בערכים שx מקבל אלא רק בהסתברויות $\{p_i\}$.
- אם x, y מ"מ בת"ל אז $H\left(x, y\right)$ היא האנטרופיה של המשתנה $\left(x, y\right)$ והיא:

$$H\left(x, y\right) = H\left(x\right) + H\left(y\right)$$

מוטיבציה

לא ניתן לקודד ערך של משתנה מקרי x בפחות מ$H\left(x\right)$ ביטים בממוצע

תרגיל

מטילים 3 מטבעות הוגנים. מה האנטרופיה של מס' הראשים?

פתרון

מס' הראשים	p_i
0	$\frac{1}{8}$
1	$\frac{3}{8} = \binom{3}{1}\frac{1}{8}$
2	$\frac{3}{8}$
3	$\frac{1}{8}$

נסמן את מס' הראשים בx אזי:

$$H\left(x\right) = -\left(\frac{1}{8}\lg\frac{1}{8} + \frac{3}{8}\lg\frac{3}{8} + \frac{3}{8}\lg\frac{3}{8} + \frac{1}{8}\lg\frac{1}{8}\right)$$

$$
\begin{aligned}
&= -\left(\frac{1}{4}\lg\frac{1}{8}-\frac{3}{4}\lg\frac{3}{8}\right)\\
&= -\frac{1}{4}\cdot(-3)-\frac{3}{4}(\lg 3-3)\\
&= \frac{3}{4}-\frac{3}{4}\lg 3+\frac{9}{4}=1.81
\end{aligned}
$$

עוד תכונה של אנטרופיה:

- אם x מקבל n ערכים שונים אז $H(x)\le\lg n$.
  שוויון מתקבל כאשר x מתפלג אחיד.

תרגיל

יהי x משתנה נורמלי $x\sim N(0,1)$. חשבו את $H(x)$.

פתרון

תזכורת:

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$$

מתקיים

$$\int_{-\infty}^{\infty}e^{-x^2}dx=\sqrt{\pi}$$

נחשב את האנטרופיה עם $\ln$ במקום עם $\lg$ ואח"כ נמיר את התוצאה ל$\lg$.

$$
\begin{aligned}
H(x) &= -\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\ln\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\right)dx\\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\frac{x^2}{2}}\ln\left(\sqrt{2\pi}e^{\frac{x^2}{2}}\right)dx\\
&= \frac{2}{\sqrt{2\pi}}\int_{0}^{\infty}e^{-\frac{x^2}{2}}\left[\ln\sqrt{2\pi}+\frac{x^2}{2}\right]dx\\
&= \frac{2}{\sqrt{2\pi}}\int_{0}^{\infty}e^{-\frac{x^2}{2}}\ln\sqrt{2\pi}dx+\frac{2}{\sqrt{2\pi}}\int_{0}^{\infty}e^{-\frac{x^2}{2}}\frac{x^2}{2}dx\\
&= 2\ln\sqrt{2\pi}\int_{0}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx+\frac{2}{\sqrt{2\pi}}\int_{0}^{\infty}e^{-\frac{x^2}{2}}\frac{x^2}{2}dx\\
&= 2\ln\sqrt{2\pi}\int_{0}^{\infty}f(x)\,dx+\frac{2}{\sqrt{2\pi}}\int_{0}^{\infty}e^{-\frac{x^2}{2}}\frac{x^2}{2}dx\\
&= 2\ln\sqrt{2\pi}\cdot\frac{1}{2}+\frac{2}{\sqrt{2\pi}}\int_{0}^{\infty}e^{-\frac{x^2}{2}}\frac{x^2}{2}dx\\
&= \ln\sqrt{2\pi}+\frac{2}{\sqrt{2\pi}}\int_{0}^{\infty}e^{-u^2}u^2\sqrt{2}du\\
&= \ln\sqrt{2\pi}+\frac{2}{\sqrt{\pi}}\int_{0}^{\infty}e^{-u^2}u^2du\\
&= \ln\sqrt{2\pi}+\frac{2}{\sqrt{\pi}}\left[\left[-\frac{1}{2}e^{-u^2}\cdot u\right]_0^{\infty}-\int_{0}^{\infty}-\frac{1}{2}e^{-u^2}du\right]\\
&= \ln\sqrt{2\pi}+\frac{2}{\sqrt{\pi}}\left[0+\frac{1}{2}\int_{0}^{\infty}e^{-u^2}\right]\\
&= \ln\sqrt{2\pi}+\frac{2}{\sqrt{\pi}}\cdot\frac{\sqrt{\pi}}{4}
\end{aligned}
$$

לסיכום:

$$H(x)=\ln\left(\sqrt{2\pi}\right)+\frac{1}{2}=\ln\sqrt{2\pi e}$$

כדי להמיר ל$\lg$ נזכור ש:

$$\lg t = \frac{\ln t}{\ln 2}$$

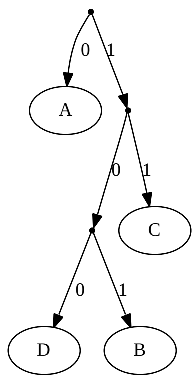
לכן ב$\lg$ נקבל:

$$H\left(x\right) = \frac{\ln\sqrt{2\pi e}}{\ln 2} = \lg\sqrt{2\pi e}$$

אלגוריתם הופמן

מקודדים מחרוזת ע"י עץ קידוד.

דוגמה



נקודד את $ACBBD$:

$$011101101100$$

אפשר גם לפענח - את 10110100011 נפענח כ:

$$BBAAAC$$

תרגיל

נניח שיש התפלגות על אותיות בשפה:

אות	הסתברות
A	0.13
B	0.14
C	0.15
D	0.16
E	0.17
F	0.18
G	0.07

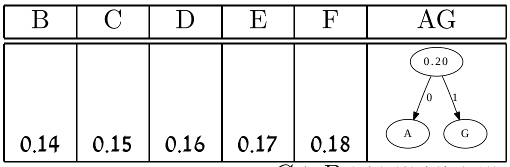
מה עץ הקידוד האופטימלי? מה שיעור הדחיסה שלו לעומת האופטימום (האנטרופיה)?

פתרון

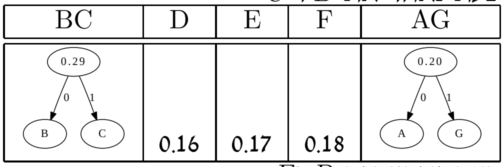
נרשום בטבלה:

A	B	C	D	E	F	G
0.13	0.14	0.15	0.16	0.17	0.18	0.07

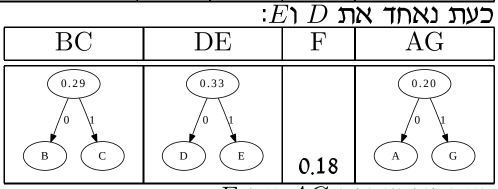
נאחד את הקטנים ביותר, A וG, ונקבל:



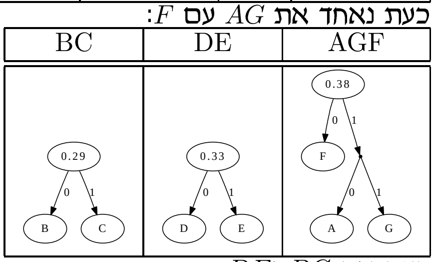
כעת נאחד את B ו C:



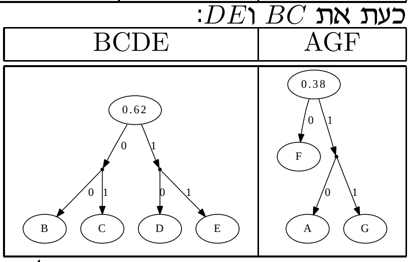
כעת נאחד את D וE:



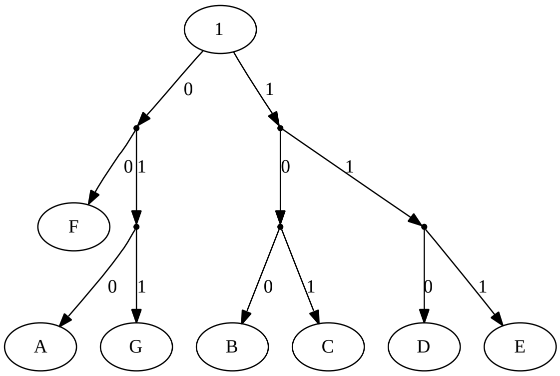
כעת נאחד את AG עם F:



כעת את BC וDE:



כעת נחבר את שניהם ביחד ונקבל את העץ הסופי:



תוחלת כמות הביטית הדרושה לייצוג אות בשפה:

$$\begin{aligned} E(x) &= (0.13+0.14+0.15+0.16+0.17+0.07)\cdot 3+0.18\cdot 2 \\ &= 2.82 \end{aligned}$$

כמות הביטים המינימלית הדרושה לייצוג אות בשפה:

$$H(\text{random letter}) = -0.13\lg 0.13 - 0.14\lg 0.14 - ...$$

זה יוצא פחות מ2.82 אך במעט.

תרגיל

1. יהי x מ"מ המקבל ערכים $1, ..., n$ בהסתברויות $p_1, ..., p_n$ בהתאמה ונניח ש$2p_i < p_{i+1}$.
איך יראה עץ קידוד אופטימלי של ערכי המשתנה x?

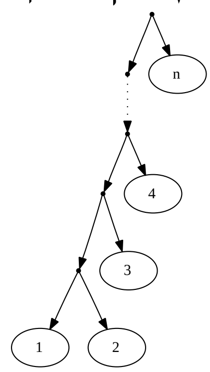
2. נניח שx מקבל את הערכים 1,2 בהסתברויות $\frac{1}{4}, \frac{3}{4}$ בהתאמה.
מה ניתן לעשות כדי לדחוס ביעילות גדולה יותר סדרת ערכים של x?

פתרון

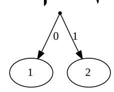
1. נשים לב ש:

$$p_1 + ... + p_k < p_{k+1}$$

לכן באיטרציה ה$k-1$ של אלגוריתם הופמן יהיה עץ עם $1, ..., k$ והוא יצורף ל$k+1$.
העץ שנקבל בסוף הוא:



2. עץ הקידוד האופטימלי הוא:



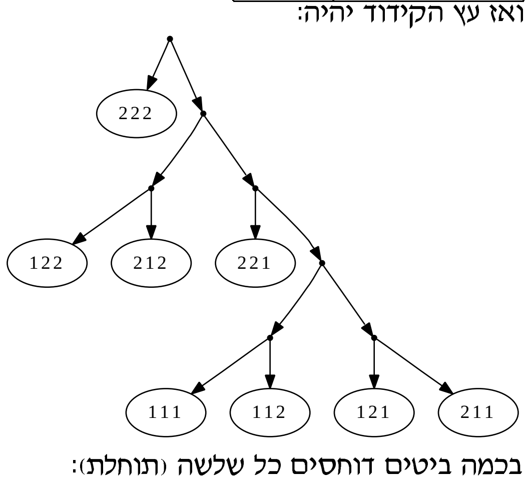
כמות הביטים הממוצעת לאות היא 1, אבל האנטרופיה היא:

$$-\frac{1}{4}\lg\frac{1}{4} - \frac{3}{4}\lg\frac{3}{4} = \frac{1}{2} - \frac{3}{4}\lg\frac{3}{4} = 0.81$$

אפשר להתגבר על הבעיה (כלומר לדחוס ע"י עץ קידוד ביעילות גדולה יותר) ע"י דחיסה זוגות או שלשות של ערכי המשתנה x.
לדוגמה, נדחוס שלשות:

הסתברות * 64	שלשה
1	111
3	112
3	121
3	211
9	122
9	212
9	221
27	222

ואז עץ הקידוד יהיה:



בכמה ביטים דוחסים כל שלשה (תוחלת):

$$E = \frac{1}{64}[27 \cdot 1 + 9 \cdot 3 \cdot 3 + 1 \cdot 5 + 3 \cdot 3 \cdot 5]$$

$$= \quad \frac{158}{64} = 2.46$$

אנחנו מקודדים שלשת ערכים ב2.46 ביטים בממוצע לכן אנו מקודדים ערך בודד בכ-0.82 ביטים (והאנטרופיה היא 0.81).