

ALS1 – Přednáška 1

Pravděpodobnost, náhodná proměnná, očekávaná hodnota náhodné proměnné

Prostor elementárních jevů S je množina, jejíž prvky se nazývají *elementární jev*. *Jev* je podmnožina A ; A je *jistý jev* pokud $A = S$; A je *nemožný jev* pokud $A = \emptyset$. Dva jevy A, B jsou *vzájemně neslučitelné jevy* pokud $A \cap B = \emptyset$.

Axiomy pravděpodobnosti

Rozdělení pravděpodobnosti $\Pr\{\}$ na S je zobrazení $S \rightarrow \mathbb{R}$ t.ž.

1. $\Pr\{A\} \geq 0$ pro každý jev A .
2. $\Pr\{S\} = 1$.
3. $\Pr\{A \cup B\} = \Pr\{A\} + \Pr\{B\}$ pro každé dva vzájemně neslučitelné jevy A a B .

$\Pr(A)$ se nazývá *pravděpodobnost* jevu A .

Podmíněná pravděpodobnost a nezávislost

Podmíněná pravděpodobnost jevu A za podmínky, že nastal jev B je definována jako

$$\Pr\{A \mid B\} = \frac{\Pr\{A \cap B\}}{\Pr\{B\}}$$

pokud $\Pr\{B\} \neq 0$.

Dva jevy jsou *nezávislé*, pokud

$$\Pr\{A \cap B\} = \Pr\{A\} \cdot \Pr\{B\},$$

což je ekvivalentní, pokud $\Pr\{B\} \neq 0$, podmínce

$$\Pr\{A \mid B\} = \Pr\{A\}.$$

Náhodné proměnné

Diskrétní náhodná proměnná X je funkce z konečného nebo spočetně nekonečného prostoru vzorů do množiny reálných čísel.

Pro náhodnou proměnnou X a reálné číslo x definujeme událost $X = x$ jako $\{s \in S \mid X(s) = x\}$; tedy

$$\Pr\{X = x\} = \sum_{\{s \in S \mid X(s) = x\}} \Pr\{s\}.$$

Funkce $f(x) = \Pr\{X = x\}$ je funkce hustoty pravděpodobnosti náhodné proměnné X . Z axiomů pravděpodobnosti máme $\Pr\{X = x\} \geq 0$ a $\sum_x \Pr\{X = x\} = 1$.

Ríkáme, že dvě náhodné proměnné X a Y jsou *nezávislé*, pokud pro všechna x a y platí

$$\Pr\{X = x \text{ and } Y = y\} = \Pr\{X = x\} \cdot \Pr\{Y = y\}.$$

Očekávaná hodnota náhodné proměnné

Očekávaná hodnota diskrétní náhodné proměnné X je

$$E[X] = \sum_x \Pr\{X = x\} \cdot x,$$

což je dobře definováno pokud je ta suma konečná nebo konverguje absolutně.

Linearity očekávání

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

pokud jsou $E[X]$ a $E[Y]$ definované.

Pokud X je náhodná proměnná, libovolná funkce $g(x)$ definuje nová náhodná proměnná $g(X)$. Pokud je očekávaná hodnota $g(X)$ definovaná, pak

$$E[g(X)] = \sum_x g(x) \cdot \Pr\{X = x\}.$$

Pokud $g(x) = ax$ (a je konstanta), pak

$$E[aX] = aE[X].$$

Pokud jsou dvě náhodné proměnné X a Y , které jsou nezávislé a mají definovanou očekávanou hodnotu,

$$E[XY] = \sum_x \sum_y \Pr\{X = x \text{ and } Y = y\}$$

Pokud náhodná proměnná X nabývá hodnot z množiny přirozených čísel, platí:

$$E[X] = \sum_{i=0}^{\infty} i \Pr\{X = i\} = \sum_{i=0}^{\infty} i(\Pr\{X \geq i\} - \Pr\{X \geq i+1\}) = \sum_{i=0}^{\infty} \Pr\{X \geq i\}.$$

Jensenova nerovnost: platí $E[f(X)] \geq f(E[X])$ pro konvexní funkci $f(X)$.

Binární vyhledávací stromy (binary search tree, BST)

Máme lineárně uspořádanou množinu klíčů. BST je strom, kde

- uzly jsou označeny klíčem,
- každý uzel má nejvýše dva potomky (nejvýše jednoho levého a nejvýše jednoho pravého).
- pokud má uzel s klíčem k levého potomka s klíčem l , platí $l < k$.
- pokud má uzel s klíčem k pravého potomka s klíčem p , platí $k < p$.

Theorem 1. Očekávaný počet porovnání při vyhledávání v BST o N klíčích je asi

$$2 \ln N \approx 1.386 \log_2 N.$$

To platí za předpokladu, že se jedná o náhodně postavený strom (randomly built BST; RBBST):

- pravděpodobnost, že N klíčů bylo vloženo v každém z $N!$ pořadí je stejná.
- ze stromu se nemazalo.

Označme:

- C_N – průměrný počet porovnání při úspěšném hledání v BST s N klíči.
- C'_N – průměrný počet porovnání při neúspěšném hledání v BST s N klíči.

$$C_N = 1 + \frac{C'_0 + C'_1 + \dots + C'_{N-1}}{N} \quad (1)$$

Rozšířený binární strom – přidáme zvláštní uzly tam, kde měl původní strom prázdný podstrom.

délka vnější cesty (E – external path length): Součet vzdáleností kořene od všech vnějších uzlů.

délka vnitřní cesty (I – internal path length): Součet vzdáleností kořene od všech vnitřních uzlů.

Theorem 2. Pro BST o N vnitřních uzlech platí, že

$$E = I + 2N. \quad (2)$$

Pokud předpokládáme, že každý klíč je vyhledáván se stejnou pravděpodobností a že každý z $N+1$ intervalů mezi klíči a vně extrémních hodnot klíčů je stejně pravděpodobný, dostáváme:

$$C_N = 1 + \frac{I}{N} \quad \text{and} \quad C'_N = \frac{E}{N+1}.$$

S použitím (2) dostáváme

$$C_N = \left(1 + \frac{1}{N}\right) C'_N - 1. \quad (3)$$

Z (1) a (3) dostáváme

$$(N+1)C'_N = 2N + C'_0 + C'_1 + \dots + C'_{N-1}. \quad (4)$$

Zbavíme se rekurence – odečteme od (5) rovnici

$$NC'_{N-1} = 2(N-1) + C'_0 + C'_1 + \dots + C'_{N-2}. \quad (5)$$

Dostaneme

$$(N+1)C'_N - NC'_{N-1} = 2 + C'_{N-1}.$$

Po úpravě

$$C'_N = C'_{N-1} + \frac{2}{N+1}.$$

Intermezzo: Harmonická čísla

$$H_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} = \sum_{k=1}^n \frac{1}{k} = \ln n + \mathcal{O}(1)$$

poslední krok se dá ukázat aproximací integrály.

Když se dá sumace vyjádřit jako $\sum_{k=m}^n f(k)$, kde $f(k)$ je monotónně klesající funkce, můžeme ji aproximovat:

$$\int_m^{n+1} f(x)dx \leq \sum_{k=m}^n f(k) \leq \int_{m-1}^n f(x)dx$$

Dolní hranice:

$$\sum_{k=1}^n \frac{1}{k} \geq \int_1^{n+1} \frac{1}{x} dx = \ln(n+1)$$

Horní hranice:

$$\sum_{k=2}^n \frac{1}{k} \leq \int_1^n \frac{1}{x} dx = \ln n \quad \text{a tedy} \quad \sum_{k=1}^n \frac{1}{k} \leq \ln n + 1$$

Harmonická čísla rekurzivně:

$$H_n = \begin{cases} 0 & \text{pro } n = 0, \\ H_{n-1} + \frac{1}{n} & \text{jinak.} \end{cases}$$

Takže $C'_N = 2H_{N+1} - 2$.

konec intermezza

Po dosazení do (3) a zjednodušení dostáváme

$$C_N = 2 \left(1 + \frac{1}{N} \right) H_N - 3 \approx 2 \ln N.$$

Theorem 3. *Očekávaná výška RBBST o N uzlech je $\mathcal{O}(\log(N))$.*

Nadefinujeme si tyto tři náhodné proměnné:

- výška RBBST o N prvcích X_N ,
- exponenciální výška $Y_N = 2^{X_N}$
- klíč v kořeni RBBST R_N .

Hodnota R_N je se stejnou pravděpodobností kterýkoli z prvků $\{1, \dots, N\}$. Pokud $R_N = i$, pak

- levý podstrom je RBBST o $i - 1$ prvcích,
- pravý podstrom je RBBST o $N - i$ prvcích,
- $Y_N = 2 \cdot \max(Y_{i-1}, Y_{N-i})$

Jako krajní případy Y_N máme $Y_1 = 1, Y_0 = 0$.

Dále definujeme náhodné proměnné $Z_{N,1}, Z_{N,2}, \dots, Z_{N,N}$, kde $Z_{N,i} = I\{R_N = i\}$.

Máme $\Pr\{R_N = i\} = \frac{1}{n}$ pro $i = 1, 2, \dots, n$, a tedy

$$E[Z_{N,i}] = \frac{1}{n} \text{ pro } i = 1, 2, \dots, n.$$

Protože právě jedna hodnota $Z_{N,i} = 1$ a všechny ostatní jsou 0, máme taky

$$Y_N = \sum_{i=1}^N Z_{N,i} (2 \cdot \max(Y_{i-1}, Y_{N-i})).$$

Ukážeme, že $E[Y_N]$ je polynomická v N , z toho pak vyplýne, že $E[X_N] = \mathcal{O}(\log N)$.

Náhodná proměnná $Z_{N,i}$ je nezávislá na Y_{i-1} a Y_{N-i} .

Když vybereme $R_N = i$, levý podstrom (s exp. výškou Y_{i-1}) je náhodně postaven z $i - 1$ klíčů, které jsou menší než i . Tento podstrom je jako jakýkoli jiný podstrom postavený z $i - 1$ prvků:

- pouze počet prvků v něm je závislý na volbě R_N ,

– jeho struktura není nijak závislá na volbě R_N . stejně tak pro pravý podstrom.

$$\begin{aligned}
 E[Y_N] &= E \left[\sum_{i=1}^N Z_{N,i} (2 \cdot \max(Y_{i-1}, Y_{N-i})) \right] = \sum_{i=1}^N E [Z_{N,i} (2 \cdot \max(Y_{i-1}, Y_{N-i}))] \\
 &= \sum_{i=1}^N E [Z_{N,i}] \cdot E [(2 \cdot \max(Y_{i-1}, Y_{N-i}))] = \sum_{i=1}^N \frac{1}{N} \cdot E [(2 \cdot \max(Y_{i-1}, Y_{N-i}))] \\
 &= \frac{2}{N} \sum_{i=1}^N E [\max(Y_{i-1}, Y_{N-i})] \leq \frac{2}{N} \sum_{i=1}^N E [Y_{i-1}] + E [Y_{N-i}]
 \end{aligned}$$

V posledním výrazu se každý term $E[Y_0], E[Y_1], \dots, E[Y_{N-1}]$ vyskytuje dvakrát – můžeme ho zjednodušit na

$$E[Y_N] \leq \frac{4}{N} \sum_{i=0}^{N-1} E[Y_i].$$

Ukážeme, že pro všechna kladná N je to ekvivalentní s

$$E[Y_N] \leq \frac{1}{4} \binom{N+3}{3}.$$

Pro $N = 1$ toto platí

$$1 = Y_1 = E[Y_1] \leq \frac{1}{4} \binom{1+3}{3} = 1$$

Dále

$$\begin{aligned}
 E[Y_N] &\leq \frac{4}{N} \sum_{i=0}^{N-1} E[Y_i] = \frac{4}{N} \sum_{i=0}^{N-1} \frac{1}{4} \binom{i+3}{3} = \frac{1}{N} \sum_{i=0}^{N-1} \binom{i+3}{3} \\
 &= \frac{1}{N} \binom{N+3}{4} = \frac{1}{N} \cdot \frac{(N+3)!}{4!(N-1)!} = \frac{1}{4} \cdot \frac{(N+3)!}{3! \cdot N!} = \frac{1}{4} \cdot \binom{N+3}{3}
 \end{aligned}$$

Platí, že $2^{E[X_N]} \leq E[2^{X_N}] = E[Y_N]$.

Z toho dostáváme, že

$$2^{E[X_N]} \leq \frac{1}{4} \binom{N+3}{3} = \frac{(N+3)(N+2)(N+1)}{24}.$$