

Big Data

Lista zadań

Jacek Cichoń, WiT, PWr, 2025/26

1 Wstęp

Zadanie 1 — Rozważamy n wymiarową kostkę $K_n = [0, 1]^n$. Generujemy N losowych punktów $X_1, \dots, X_N$ z K_n zgodnie z rozkładem jednostajnym. Niech

$$f(x) = \mathbb{E} [|\{i : X_i \in [0, x]^n\}|]$$

Niech r_n i r_n^* będą takimi liczbami, że $f(r_n) = 1$ oraz $f(r_n^*) = N/2$.

1. Wyznacz liczby r_n i r_n^* .
2. Oblicz r_n^*/r_n dla $N = 10^4$ oraz $n = 1, 2, 3, 4$.
3. Ustalmy liczbę N . Oblicz granicę $\lim_{n \rightarrow \infty} r_n^*/r_n$.

Zadanie 2 — Znajdź swoją datę urodzin w liczbach π , e oraz $\sqrt{2}$.

Zadanie 3 — Niech $V_n(r)$ oznacza objętość kuli o promieniu r w przestrzeni $\mathbb{R}^n$.

1. Pokaż, że $V_n(r) = \frac{2\pi}{n} r^2 \cdot V_{n-2}$ dla $n \geq 3$.
2. Wyprowadź wzór

$$V_n(r) = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2} + 1)} r^n,$$

$$\text{gdzie } \Gamma(n+1) = n! \text{ oraz } \Gamma(n + \frac{1}{2}) = (n - \frac{1}{2})(n - \frac{3}{2}) \cdots \frac{1}{2} \sqrt{\pi}$$

Zadanie 4 — Narysuj wykres przedstawiający objętości $V_n(r)$ kul o promieniu r w przestrzeni $\mathbb{R}^n$ dla $r = 0.5, 1, 2$ oraz $n = 1, \dots, 50$.

Zadanie 5 — Losujemy zgodnie z jednostajnym rozkładem punkt X z kuli jednostkowej $B_n = \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$. Jaka jest wariancja zmiennej losowej $\|X\|_2$?

Zadanie 6 — Rozważmy następującą procedurę generowania punktu z kuli B_2 : (1) generujemy niezależnie dwie liczby losowe x, y z odcinka $[0, 1]$ zgodnie z rozkładem jednostajnym (2) zwracamy punkt $(\sqrt{x} \cos(2\pi y), \sqrt{x} \sin(2\pi y))$.

1. Pokaż, że metoda ta generuje losowy punkt z B_2 z rozkładem jednostajnym.
2. Jaki rozkład otrzymamy gdy "zapomnimy" o pierwiastku?

Zadanie 7 — Oblicz $\max\{d(X, Y) : X, Y \in [0, 1]^n\}$, gdzie d oznacza standardową odległość euklidesową.

Zadanie 8 — Wyznacz objętość n -wymiarowego sześcianu wpisanego w n -wymiarową kulę. Oblicz stosunek tej objętości do objętości kuli.

Zadanie 9 — Ustalmy parametr $n \geq 1$ oraz $k \geq 1$. Napisz procedurę, która generuje zbiór X złożony z k losowych punktów z przestrzeni $[0, 1]^n$ zgodnie z rozkładem jednostajnym, następnie wyznacza dwie liczby $d_{\max}(X) = \max\{d(P, Q) : P, Q \in X\}$ oraz $d_{\min}(X) = \min\{d(P, Q) : P, Q \in X \wedge P \neq Q\}$ i zwraca liczbę $d_{\max}/d_{\min}$. Przeanalizuj wyniki tego algorytmu dla $k = 50$ oraz $n = 1, 10, 100, 1000, 10000$.

2 Trick medianowy

Zadanie 10 — Pokaż, że jeśli chcesz jednoznacznie wyreprezentować każdą z liczb ze zbioru $\{0, 1, \dots, n\}$ za pomocą b bitów to $b \geq \lceil \log_2(n+1) \rceil$.

Zadanie 11 — Pokaż, że jeśli $x = \sum_{k=0}^s a_k 2^k$, gdzie $a_i \in \{0, 1\}$ oraz $a_s = 1$ to $s = \lfloor \log_2(x+1) \rfloor$

Zadanie 12 — Niech C_n będzie wartością klasycznego licznika Morris'a po n krotnym wywołaniu funkcji `onInc()`.

1. Pokaż, że $\text{var}[2^{C_n}] = \frac{1}{2}n(n-1)$.
2. Skorzystaj z nierówności Jensena dla wartości oczekiwanej zmiennej losowej do pokazania, że $E[C_n] \leq \log_2(n+1)$. **Wskazówka:** Funkcja $f(x) = 2^x$ jest wypukła.

Zadanie 13 — Rozważmy następującą modyfikację licznika Morrisa: ustalamy liczbę $\alpha > 0$ oraz rozważamy tak oprogramowany licznik:

```
init :: C = 0
onInc :: if (random() < (1/(1+alpha))^C) then C = C+1
onGet :: return (?????)
```

Niech C_n oznacza wartość zmiennej C po n wywołaniach metody `onInc`.

1. Wyznacz $E[(1+\alpha)^{C_n}]$
2. Uzupełnij funkcję `onGet` tak aby otrzymać nieobciążony estymator liczby użyć metody `onInc`.

Zadanie 14 — Wyznacz nierówność Chernoffa dla $\Pr[X \geq a]$ dla zmiennej losowej X o rozkładzie Poissona z parametrem λ . Znajdź ograniczenie górne $\Pr[X \geq 2\lambda]$.

Zadanie 15 — Niech $x_1, \dots, x_n$ będzie ciągiem liczb rzeczywistych. Rozważamy dwie funkcje $f(x) = \sum_{i=1}^n |x_i - x|$ oraz $g(x) = \sum_{i=1}^n (x_i - x)^2$

1. Pokaż, że funkcja g osiąga minimum w średniej arytmetycznej liczb $x_1, \dots, x_n$
2. Pokaż, że funkcja h osiąga minimum w medianie ciągu $x_1, \dots, x_n$.
Wskazówka: Możesz założyć, że $x_1 \leq x_x \leq \dots \leq x_n$. Przyjrzy się najpierw pomocniczej funkcji $\phi(x) = |x_1 - x| + |x_n - x|$. $\acute{e}$

Zadanie 16 — **(One sided Chebyshev inequalities)** Pokaż, że dla dowolnej zmiennej losowej X o wartości oczekiwanej μ , wariancji σ^2 oraz dowolnej liczby $a > 0$ mamy

$$\Pr(X \geq \mu + a) \leq \frac{\sigma^2}{\sigma^2 + a^2}, \quad \Pr(X \leq \mu - a) \leq \frac{\sigma^2}{\sigma^2 + a^2}.$$

Wskazówka: Rozważ zmienną losową $Y = (X + t)^2$ dla $t > -(\mu + a)$.

Zadanie 17 — Pokaż, że $|\mu - m| \leq \sigma$, gdzie μ oznacza wartość oczekiwaną, m medianę zaś σ odchylenie standardowe zmiennej losowej X . **Wskazówka:** Zastosuj poprzednie zadanie dla $a = \sigma$.

c.d.n.
Powodzenia,
Jacek Cichoń