Big Data
Lista zadań

Jacek Cichoń, WiT, PWr, 2024/25

1 Wstęp

Zadanie 1 — Znajdź swoją datę urodzin w liczbach π, e oraz $\sqrt{2}$.

Zadanie 2 — 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 $n \geqslant 3$.
2. Wyprowadź wzór
$$V_n(r) = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)} r^n \ ,$$
gdzie $\Gamma(n+1) = n!$ oraz $\Gamma(n+\frac{1}{2}) = (n-\frac{1}{2})(n-\frac{3}{2})\cdots\frac{1}{2}\sqrt{\pi}$

Zadanie 3 — Narysuj wykres przedstawiający objętości $V_n(r)$ kul o promieniu r w przestrzeni $\mathbb{R}^n$ dla $r = 0.5, 1, 2$ oraz $n = 1, \ldots 50$.

Zadanie 4 — Losujemy zgodnie z jednostajnym rozkładem punkt X z kuli jednostkowej $B_n = \{x \in \mathbb{R}^n : \|x\|_2 <= 1\}$. Jaka jest wariancja zmiennej losowej $\|X\|_2$?

Zadanie 5 — 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 6 — Oblicz $\max\{d(X, Y) : X, Y \in [0, 1]^n\}$, gdzie d oznacza standardową odległość euklidesową.

Zadanie 7 — Wyznacz objętość n-wymiarowego sześcianu wpisanego w n-wymiarową kulę. Oblicz stosunek tej objętości do objętości kuli.

Zadanie 8 — Ustalmy parametr $n \geqslant 1$ oraz $k \geqslant 1$. 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 $d_{min}(X) = \max\{d(P, Q) : P, Q \in X \wedge P \neq Q\}$ i zwraca liczbę d_{max}/d_{min}. Przeanalizuj wyniki tego algorytmu dla $k = 50$ oraz $n = 1, 10, 100, 1000, 10000$.

2 "Hello World" dla Big Data

Zadanie 9 — Pobierz plik z kilkoma dramatami Szekspira ze strony wykładu. Wybierz jeden z dramatów.

1. Oczyść wybrany plik. Podziel go na słowa.
2. Usuń z niego "Stop Words" i usuń z niego słowa o długości mniejszej lub równej 2.
3. Zbuduj chmurę wyrazów (word cloud) z otrzymanej listy. Możesz skorzystać np. z serwisu `http://www.wordclouds.com/`

Celem tego zadania jest wygenerowanie mniej więcej takiego obrazka (dla poematu "Pan Tadeusz"):

Zadanie 10 — To jest kontynuacja poprzedniego zadania.

1. Zastosuj część funkcji które napisałeś do realizacji poprzedniego zadania do wyznaczenia indeksów TF.IDF dla wszystkich wyrazów z dramatów Szekspira znajdujących się w pliku ze strony wykładu.

2. Zbuduj chmury wyrazów oparte o TF.IDF dla wszystkich rozważanych dramatów.

3 Trick medianowy

Zadanie 11 — Niech C_n będzie wartością klasycznego licznika Morris'a po n krotnym wywołaniu funkcji `onInc()`.

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

Zadanie 12 — Rozważmy następującą modyfikację licznika Morrisa: ustalamy liczbę $\alpha > 0$ oraz rozważamy tak oprogramowany licznik:

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

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

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

Zadanie 13 — Wyznacz nierówność Chernoffa dla $\Pr[X \geqslant a]$ dla zmiennej losowej X o rozkładzie Poissona z parametrem λ. Znajdź ograniczenie górne $\Pr[X \geqslant 2\lambda]$.

Zadanie 14 — Niech $x_1, \ldots, x_n$ 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 $x_1, \ldots, x_n$
2. Pokaż, że funkcja h osiąga minimum w medianie ciągu $x_1, \ldots, x_n$.
   Wskazówka: Możesz założyć, że $x_1 \leqslant x_x \leqslant \cdots \leqslant x_n$. Przyjrzy się najpierw pomocniczej funkcji $\phi(x) = |x_1 - x| + |x_n - x|$. ć

Zadanie 15 — **(One sided Chebyshev inequalities)** Pokaż, że dla dowolnej zmiennej losowej X o wartości oczekiwanej μ, wariancji σ^2 oraz dowolnej liczby $a > 0$ mamy

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

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

Zadanie 16 — Pokaż, że $|\mu - m| \leqslant \sigma$, gdzie μ oznacza wartość oczekiwaną, m medianę zaś σ odchylenie standardowe zmiennej losowej X. Wskazówka: Zastosuj poprzednie zadanie dla $a = \sigma$.

4 Zliczanie

Zadanie 17 — Niech $X_{k:n}$ oznacza k-tą statystykę pozycyjną dla rozkładu jednostajnego w odcinku $[0, 1]$.

1. Wyznacz $\mathrm{E}[X_{k:n}]$ oraz $\mathrm{var}[X_{k:n}]$.
2. Niech $Y_n = \frac{k-1}{X_{k:n}}$. Wyznacz $\mathrm{E}[Y_{k:n}]$ oraz $\mathrm{var}[Y_{k:n}]$.

Zadanie 18 — Zaimplementuj algorytm zliczania unikalnych elementów w strumieniu danych oparty o k-tą statystykę pozycyjną. Przetestuj dokładność tego algorytmu dla $k = 400$.

Zadanie 19 — **("'Rozgrzewka"')** Rozważamy następujący wariant licznika probabilistycznego. Mamy dwie zmienne L_0 i L_1.Podczas inicjalizacji ustawiamy $L_0 = 0$ oraz $L_1 = 0$. W procedurze `onTick()` generujemy najpierw liczbą losową $i \in \{0, 1\}$ zgodnie z rozkładem jednostajnym i następnie z prawdopodobieństwem $\frac{1}{2}$ zwiększamy zawartość zmiennej L_i. Funkcja `onGet()` zwraca liczbę $L_0 + L_1$.

1. Wyznacz rozkład zmiennej losowej $L_0 + L_1$ po n wywołaniach procedury `onTick()`.
2. Jeśli poprzedni podpunkt zrealizowałeś za pomocą rachunków, to znajdź proste wytłumaczenie zauważonego zjawiska.

Zadanie 20 — Rozważamy wariant poprzedniego zadania dla liczników Morrisa: tym razem L_0 i L_1 są licznikami Morrisa. Inkrementacja każdego L_i odbywa się z prawdopodobieństwem 2^{-L_i}. Funkcja `onGet()` zwraca liczbę $L_0 L_1$.

1. Wyznacz $\mathrm{E}[L_0 L_1]$ i $\mathrm{var}[L_0 L_1]$ po n wywołaniach procedury `onTick()`.
2. Rozważ k krotne iteracyjne zagłębienie tej procedury.

Zadanie 21 — Rozważamy wariant Zadania 19 dla algorytmu zliczania liczby unikalnych elementów w strumieniu danych opartego na statystykach pozycyjnych. Tym razem L_0 i L_1 są $k/2$-tymi elementami z obserwowanych danych (niech k będzie liczbą parzystą). Funkcja `onGet()` zwraca liczbę $L_0 + L_1$.

1. Wyznacz $\mathrm{E}[L_0 + L_1]$ i $\mathrm{var}[L_0 + L_1]$ po n wywołaniach procedury `onTick()`.
2. Rozważ k krotne iteracyjne zagłębienie tej procedury. Jak prościej możesz opisać otrzymany algorytm.
3. Zaimplementuj algorytm powyższy algorytm i sprawdź jego skuteczność.

c.d.n.
Jacek Cichoń