Webinarium przeprowadzone w ramach projektu "Mistrzostwa w Algorytmice i Programowaniu - Uczniowie", **finansowanego przez:**

Piotr Chrząstowski-Wachtel
Uniwersytet Warszawski

Rachunek prawdopodobieństwa

Rachunek prawdopodobieństwa

- Jedna z najbardziej intuicyjnych dziedzin matematycznych
- Pełna fałszywych intuicji
- Przez lata uprawiana na zasadzie machania rękami
- Zaksjomatyzowana dopiero w 1930 roku.

Typowe błędy

- Prawdopodobieństwo tego, że wypadnie orzeł w rzucie monetą wynosi 1/2 i wynika to z aksjomatów Kołmogorowa.
- Można wylosować dowolną liczbę całkowitą z równym prawdopodobieństwem.
- Można obliczyć, jakie jest prawdopodobieństwo tego, że wojna na Ukrainie skończy się do końca sierpnia.
- Istnieje coś takiego, jak losowa cięciwa okręgu.

Mimo tego

- Rachunek prawdopodobieństwa jest jedną z najbardziej użytecznych dziedzin matematycznych.
- Takie sektory gospodarki, jak bankowość, ubezpieczenia, biorą specjalistów na pniu.
- W informatyce są dwa duże zastosowania:
  - w algorytmice
    - do szacowania średnich kosztów algorytmów
    - do tworzenia i analizy algorytmów randomizowanych
  - w sztucznej inteligencji do analizy danych (korelacje, przyczynowość,...)

Zmienna losowa

- Jedno z najważniejszych pojęć matematycznych
- Kodowanie wydarzeń za pomocą liczb i określanie ich własności ilościowo.
- Za zmienną losową możemy uznać dowolną funkcję ze zbioru zdarzeń w zbiór liczb rzeczywistych.

Rozkład zmiennej losowej

- Wartościom zmiennych losowych przypisujemy prawdopodobieństwa.
- W przypadku dyskretnym jest to w miarę proste – po prostu określamy dla każdej wartości zmiennej losowej jej prawdopodobieństwo.
- W przypadku ciągłym używamy przedziałów wartości zmiennej losowej – pojedyncza wartość ma zazwyczaj prawdopodobieństwo równe zero (krzywa Gaussa to nie jest funkcja wartości prawdopodobieństw!).

Wartość oczekiwana zmiennej losowej

- Mylona często z wartością średnią.
- Jest sumą (lub całką) wartości zmiennej losowej przemnożonej przez funkcję gęstości jej rozkładu w każdym punkcie.

Niezależność zdarzeń

- Jedno z podstawowych pojęć.
- Formalnie dwa zdarzenia są niezależne, jeśli $P(A \cap B)=P(A)P(B)$
- Tak naprawdę chodzi o to, żeby $P(A)=P(A|B)$, gdzie przez $P(A|B)$ rozumiemy prawdopodobieństwo tego, że zaszło zdarzenie A pod warunkiem że zaszło zdarzenie B.

Przykład – dwójka dzieci

- Na potrzeby przykładu przyjmijmy dwa założenia (oba w rzeczywistości fałszywe).
  - prawdobodobieństwo urodzenia dziecka danej płci jest równe 1/2
  - płeć kolejnych dzieci danej pary jest niezależna od poprzednich.

Przykład – dwójka dzieci

- Spotyka się dwóch kolegów szkolnych po latach.
  - Co u ciebie słychać?
  - Ożeniłem się, mam dwójkę dzieci.
  - A masz córeczkę?
  - Mam
  - To pewnie też masz synka?
  - Skąd wiesz?
  - Obliczyłem!

Przykład – dwójka dzieci

- Wygląda na absurd: jak można obliczyć coś, co ma prawdopodobieństwo 1/2?
- Jeśli przyjrzymy się tej sytuacji bliżej, okaże się że ten drugi pan miał rację: prawdopodobieństwo posiadania syna, gdy wiadomo, że jednym z dzieci jest córka jest równe 2/3.

Przykład – dwójka dzieci

- Przestrzeń zdarzeń elementarnych:
  - DD 1/4
  - DC 1/4
  - CD 1/4
  - CC 1/4
- Określmy teraz dwa zdarzenia:
  - Ch – jest wśród dzieci chłopiec
  - Dz – jest wśród dzieci dziewczynka
- Zatem Ch={DC,CD,CC},Dz={DD,DC.CD}; P(Ch)=P(Dz)=3/4
- Teraz $P(Ch|Dz)=P(Ch \cap Dz)/P(Dz)=P(\{CD,DC\})/P(Dz) = (1/2)/(3/4) = 2/3$

Przykład – dwójka dzieci – wersja 2

- Tym razem dialog urywa się wcześniej:
  - Co u ciebie słychać?
  - Ożeniłem się, mam dwójkę dzieci.
  - A masz córeczkę?
  - Mam
- Następnego dnia kolega dzwoni do taty na domowy telefon i słuchawkę podnosi dziewczynka. Prosi tatusia, a ten ze zdziwieniem wysłuchuje, że pewnie ma też synka. Okazuje się jednak, że drugim dzieckiem jest też córeczka.

Przykład – dwójka dzieci – wersja 2

- Przestrzeń zdarzeń elementarnych: tym razem kolejne litery oznaczają: pierwsze dziecko, drugie dziecko, kto odebrał telefon.
  - DDD 1/4
  - DDC 0
  - DCD 1/8
  - DCC 1/8
  - CDD 1/8
  - CDC 1/8
  - CCD 0
  - CCC 1/4

Przykład – dwójka dzieci – wersja 2

- Teraz
  - Ch={DCD,DCC,CDD,CDC,CCD,CCC}; P(Ch)=3/4
  - (T=Dz) = {DDD,DCD,CDD,CCD}; P(T=Dz)=1/2
  - Ch∩(T=Dz)={DCD,CDD,CCD}; P(Ch∩(T=Dz))=1/4
  - P(Ch|(T=Dz)) = (1/4)/(1/2)=1/2

Przykład – dwójka dzieci – wersja 2

- Widać, jak bardzo uważnie trzeba konstruować model probabilistyczny. Tutaj końcowe prawdopodobieństwo zależy od sposobu pozyskania informacji o płci jednego z dzieci.
- Gdyby kolega jakoś dowiedział się, że dziewczynką jest **starsze** z dzieci, to rzecz jasna w wersji 1 natychmiast prawdopodobieństwo tego, że drugim jest chłopiec spada do 1/2

Paradoks dwóch puszek

- Nasz bardzo bliski przyjaciel gra z nami w taką grę. Bierze dwie identyczne puszki, udaje się do sąsiedniego pokoju i tam rzuca monetą tak długo, aż wypadnie orzeł. W zależności od tego, ile wykonał rzutów, wkłada do puszek różne ilości pieniędzy. Dokładniej: przy n rzutach do jednej z puszek wkłada 3^{n-1}, a do drugiej 3^n zł.

  Następnie stawia przed nami te dwie puszki, pozwala otworzyć jedną z nich, przeliczyć pieniądze i zachować je lub zdecydować się na wzięcie pieniędzy z drugiej puszki, ale bez możliwości wycofania się z tej decyzji. Jaką strategię powinniśmy przyjąć?

Paradoks dwóch puszek

- Jasne, że jeśli w otworzonej puszce widzimy złotówkę (tak się stanie, jeśli przyjaciel od razu trafi orła a my akurat wybierzemy na chybił-trafił tę „gorszą” puszkę), to rzecz jasna powinniśmy zdecydować się na tę drugą, w której na pewno są 3 zł.
- Co jednak, jeśli widzimy jakąś większą sumę?

Paradoks dwóch puszek

- Z jakim prawdopodobieństwem dostajemy pary puszek?

  O P(O) =1/2 (1, 3)

  RO P(RO)=1/4 (3, 9)

  RRO P(RRO)=1/8 (9, 27)

  ...

  RR...RO $P(RR...RO)=1/2^n$ $(3^{n-1}, 3^n)$

  n rzutów n rzutów

  ...

- To jest znany rozkład geometryczny

Paradoks dwóch puszek

- Zauważmy, że jeśli widzimy 3^n zł w pierwszej puszce i n>0, to możliwe są tylko dwie sumy w drugiej: 3^{n-1} i 3^{n+1}
- Ta pierwsza wartość jest dwukrotnie bardziej prawdopodobna, niż druga.
- Jeśli zdecydujemy się na pozostanie przy pierwszej puszce, to średnio wygramy 3^n zł.
- Jeśli jednak zdecydujemy się na drugą puszkę, to będzie to średnio $(2/3)\ 3^{n-1} + (1/3)\ 3^{n+1}$ zł, a to jest więcej niż 3^n zł.

Paradoks dwóch puszek

- Wniosek: Zawsze się opłaca zmienić decyzję!
- Ale od którego momentu?
- Przecież na samym początku nie ma znaczenia, którą puszkę zaczniemy otwierać. Są identyczne.
- Otwieramy,... cały czas jest obojętne aż do momentu, w którym już widzimy, ile jest w niej złotych. W tym momencie (ale dopiero w tym!) opłaca się wziąć drugą puszkę!
- Absurd!
- A gdybyśmy od razu wybrai tę drugą puszkę, to też by się nam opłacało zmienić decyzję?
- Z powyższego rozumowania wynika że tak, ale dopiero, jak przeliczymy w niej kasę.

Wracamy do wyszukiwania binarnego

- Jest to tak ważny algorytm, że każde jego ulepszenie może być istotne.
- Zakładamy, że tablica A jest posortowana niemalejąco.
- Powiedzmy, że interesuje nas jakikolwiek indeks elementu x w posortowanej tablicy lub informacja, że x nie ma w tablicy.
- Może się zdarzyć, że takich elementów jest więcej, wtedy algorytm wyszukiwania binarnego od pewnego momentu będzie rozważał przedział, w któym są same x, a mógłby przerwać pętle wcześniej.
- Do tego jednak trzeba dodatkowego porównania: czy A[s] jest równe x. .

Wersja z dodatkowym porównaniem

```
int szukajx (int A[], int n int x)
{ int l,p,s;
l=0; p=n-1; s=(l+p)/2;
while (l<p && A[s]!=x)
{if (x>A[s]) l=s+1;
        else p=s;
 s=(l+p)/2);
}
if (A[s]==x) return s;
else return -1;  // nie ma x
}
```

Pytanie: czy się opłaca?

- Ogólnie zależy:
  - jeśli jest dużo wartości x, to tak, w szczególności jeśli jest ich więcej niż n/2, to w ogóle nie wejdziemy do pętli.
  - jeśli x nie ma w tablicy, to się nie opłaca – i tak musimy zejść do przedziału jednoelementowego i tylko niepotrzebnie będziemy robili dodatkowe sprawdzenie. `

A jeśli jest dokładnie jeden x?

- To warto taki przypadek wziąć pod lupę, szczególnie, że jest bardzo typowy.
- Dla uproszczenia przyjmijmy, że $\mathbf{n=2^k-1}$. W razie czego możemy (choćby wirtualnie) uzupełnić tablicę nieskończonościami do najbliższej takiej wartości. Nie jest to szczególnie bolesne ograniczenie:
  - po pierwsze po pierwszym strzale mamy na pewno co najwyżej tyle elementów niż oryginalnie. W obliczeniach. Pomylimy się nie więcej niż o 1.
  - testy pokazują że wyniki się uśrednią do uzyskanych dla tych właśnie szczególnych wartości.

Przypadek pierwszego algorytmu

- Tu nie ma dużego problemu: koszt będzie równy k, czyli sufit z $\log_2 n$; za każdym razem długość przedziału spada o połowę.

Przypadek drugiego algorytmu

- Załóżmy, że nic nie wiemy o położeniu x. Innymi słowy zakładamy, że z równym prawdopodobieństwem może wystąpić na każdej pozycji.
- W związku z tym obliczmy, ile obrotów pętli wykonamy, gdy x jest na pierwszej od lewej pozycji, na drugiej, ..., na n-tej. Wysumujmy te wartości i podzielmy przez n. To będzie średnia liczba obrotów pętli – wartość oczekiwana zmiennej losowej, która jest równa właśnie liczbie obrotów pętli.

Przypadek drugiego algorytmu

- Możemy mieć szczęście i trafić x za pierwszym razem. Będzie tak jednak tylko w jednym przypadku na n: gdy x jest w połowie tablicy. Wykonamy wtedy jedno sprawdzenie warunku.
- Jeśli nie trafimy za pierwszym razem w x, to za drugim razem trafimy w dwóch przypadkach: kiedy x jest dokładnie w połowie lewego segmentu lub w połowie prawego segmentu. Mamy zatem dwie pozycje dla x skutkujące dwoma sprawdzeniami warunku.
- Jeśli nie trafimy ani za pierwszym ani za drugim razem, to mamy 4 pozycje (w połowach ćwiartek), które nam dadzą trzy sprawdzenia warunku.

Przypadek drugiego algorytmu

Dla n=7 mamy jeden element z jednym sprawdzeniem warunku,
dwa z dwoma i cztery z trzema.

Ogólnie mamy 2^{i-1} elementów z i sprawdzeniami warunku

Przypadek drugiego algorytmu

- Zatem z prawdopodobieństwem $2^{i-1}/n$ wykonamy i sprawdzeń warunku.
- Pozostaje zatem obliczyć sumę

$$\Sigma_{i=1}^{k} i2^{i-1}/n = (1/n) \Sigma_{i=1}^{k} i2^{i-1}$$

- Skupmy się na sumie

$$\Sigma_{i=1}^{k} i2^{i-1}$$

$\Sigma_{i=1}^{k} i2^{i-1}$

Utrudnijmy sobie najpierw zadanie i zdefiniujmy dla dowolnego rzeczywistego x funkcję $f(x) = \Sigma_{i=1}^{k} ix^{i-1}$

Chodzi nam o wyznaczenie wartości f(2). Zatem:

$f(x) = \Sigma_{i=1}^{k} ix^{i-1} = \Sigma_{i=1}^{k} (x^i)' = (\Sigma_{i=1}^{k} x^i)' = ((x^{k+1}-x)/(x-1))' =$

$= ((k+1)x^k-1)(x-1)-x^{k+1}+x)/(x-1)^2$

Wygląda nie najlepiej, ale dla x=2 sama przyjemność:

$f(2) = (k+1)2^k-1-2^{k+1}+2 = (k-1)2^k+1$

Pora na końcowe wnioski

Zatem porównujemy dwie wartości:

- z pierwszego algorytmu k,
- z drugiego $((k-1)2^k+1)/n$

- Ale ta druga wartość, to $((k-1)2^k+1)/(2k-1) > ((k-1)2^k)/(2^k)=k-1$

Zyskaliśmy zatem średnio niecałe jedno sprawdzenie warunku. Opłaca się?

Algorytmy probabilistyczne

- To algorytmy, w któych używamy generatora liczb losowych.
- Są dwa główne typy:
  - algorytmy Las Vegas
  - algorytmy Monte Carlo

Algorytmy Las Vegas

- Algorytmy, w których działamy „do skutku” i od losowania zależy złożoność obliczeniowa. Mamy zatem stuprocentową pewność otrzymania odpowiedniego wyniku
- Przykład: szukamy jakiejś wartości w nieuporządkowanej tablicy, wiedząc, że wypełnia ją w ponad połowie. Losujemy kolejne indeksy (żeby nie iść np. od lewej).

Algorytmy Monte Carlo

- Algorytmy, w których działamy przez pewną skończoną i zazwyczaj z góry określoną liczbę kroków.
- Nie gwarantują one poprawności wyniku, jednak prawdopodobieństwo pomyłki można zazwyczaj ograniczyć przez dowolnie małą stałą.
- Przykład: test pierwszości: algorytm Rabina-Millera

Algorytm Millera-Rabina

- Jest to algorytm, którego zasadniczą ideę wymyślił Miller, jednak jej poprawność zależy od tego, czy jest prawdziwa hipoteza Riemanna o funkcji zeta.
- Rabin ulosowił go otrzymując poprawny algorytm, jednak nie dający odpowiedzi ze stuprocentową pewnością. Mamy zatem dowolnie małą, ale niepewność co do prawdziwości odpowiedzi.

Algorytm Millera-Rabina

- Dane do algorytmu: liczba n, o której chcemy stwierdzić, czy jest pierwsza oraz parametr k, będący liczbą naturalną, który precyzuje dokładność z jaką chcemy znać odpowiedź – możemy w ten sposób wpływać na prawdopodobieństwo błędu.
- Wynik:
  - Odpowiedź NIE – wtedy wiemy na 100%, że liczba jest złożona lub
  - Odpowiedź TAK – wtedy wiemy z prawdopodobieństwem $1-1/4^k$, że ta liczba jest pierwsza.

Algorytm Millera-Rabina

- Kolejne kroki:
  - Wyznaczamy największą potęgę dwójki s taką, że $2^s<n$;
  - obliczamy nieparzyste $d=n/2^s$;
  - losujemy liczbę a z przedziału 1,...,n-1
    - Jeśli $\sim(a^d \equiv 1 \pmod n)$ oraz dla każdego $r=0,1,2,...,s-1$ mamy $\sim(a^{2^r d} \equiv (-1) \pmod n)$, to a jest złożona
    - W przeciwnym razie szansa na pierwszość wzrasta i powtarzamy ten krok aż osiągniemy albo dowód złożoności albo liczbę kroków k. Po k krokach negatywnych stwierdzamy, że n jest prawdopodobnie pierwsza.

Twierdzenie 1

- Jeśli n jest liczbą pierwszą, zaś a jest mniejsze od n. Niech tym razem $d=(n-1)/2^s$, gdzie d jest nieparzyste. Wówczas
  - albo $a^d \equiv 1 \pmod{n}$
  - albo istnieje r=0,1,...,s-1 takie że $a^{2^r d} \equiv (-1) \pmod{n}$
- Liczbę a, która nie spełnia powyższych warunków nazywa się **świadkiem złożoności** n.

Twierdzenie 2

- Jeśli n>3 jest nieparzystą liczbą złożoną, to w zbiorze 1,2,...n-1 jest co najwyżej (n-1)/4 liczb niebędących świadkami jej złożoności.
- Zatem każde kolejne wykonanie testu, związane z losowaniem a, zwiększa prawdopodobieństwo pierwszości czterokrotnie.

Rozwiązanie paradoksu dwóch puszek

Polecam moje dwa artykuły z Delty:

- http://www.deltami.edu.pl/temat/matematyka/rachunek_prawdopodobienstwa/2014/08/26/Tak_bardzo_oczekiwana_wartosc/
- http://www.deltami.edu.pl/temat/matematyka/rachunek_prawdopodobienstwa/2014/10/30/Wartosc_nieoczekiwana/

<Mistrzostwa
w Algorytmice
i Programowaniu>

Projekt „Mistrzostwa w Algorytmice i Programowaniu – Uczniowie” jest finansowany ze środków pochodzących z „Programu Rozwoju Talentów Informatycznych na lata 2019-2029”

Dofinansowanie Projektu: 4.887.850,50 zł

Całkowita wartość Projektu: 5.460.850,50 zł

Publikacja multimedialna wyraża jedynie poglądy autorów i nie może być utożsamiana z oficjalnym stanowiskiem Kancelarii Prezesa Rady Ministrów