



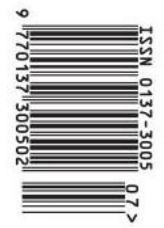
deltam

MATEMATYKA – FIZYKA – ASTRONOMIA – INFORMATYKA

$\ln=6.63 \cdot 10^{-34}$ | h | $\pi=3,141592\dots$

NR 7 (590) 2023
CENA 6 ZŁ VAT 8%
PL ISSN 0137-3005 | NR IND 35 550 X
MIESIĘCZNIK
www.deltami.edu.pl




Statystyczny
numer specjalny



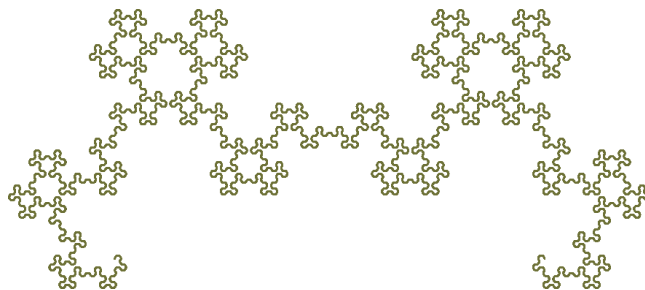
Nakład: 2800 egz.



SPIS TREŚCI NUMERU 7 (590)

O statystyce, braku precyzji i serach <i>Przemysław Grzegorzewski</i>	str. 1
Algorytm magicznych piątek <i>Oskar Skibski</i>	str. 6
 Wirus znaczy jad <i>Magdalena Fikus</i>	str. 8
Skoro wszystkie modele są błędne, to które są użyteczne? <i>Przemysław Biecek</i>	str. 9
Chłopiec czy dziewczynka? <i>Lukasz Rajkowski</i>	str. 12
Aktualności Nowe odpowiedzi, nowe pytania	str. 14
Poszukiwanie porządku <i>Andrzej Dąbrowski</i>	str. 15
 Zadania	str. 19
Klub 44	str. 20
Czy tu jest równomiernie? <i>Radosław Poleski</i>	str. 21
Prosto z nieba: Drewniany zegar	str. 23
Niebo w lipcu	str. 23
 Łańcuchy Markowa – część 1 <i>Bartłomiej Bzdęga</i>	str. 25

W następnym numerze:
Fraktalny świat papierowej tasiemki.



Miesięcznik *Delta* – matematyka, fizyka, astronomia, informatyka jest wydawany przez Uniwersytet Warszawski przy współpracy towarzystw naukowych: Polskiego Towarzystwa Matematycznego, Polskiego Towarzystwa Fizycznego, Polskiego Towarzystwa Astronomicznego i Polskiego Towarzystwa Informatycznego.

Komitet Redakcyjny: dr Waldemar Berej, doc. dr Piotr Chrzastowski-Wachtel, dr Krzysztof Ciesielski, prof. UJ – przewodniczący, prof. dr hab. Bożena Czerny, dr Andrzej Dąbrowski, dr Tomasz Greczyło, prof. UW, dr Adam Gregosiewicz, dr Andrzej Grzesik, prof. dr hab. Agnieszka Janiuk, dr hab. Artur Jeż, prof. UW, dr hab. Bartosz Klin, prof. dr hab. Andrzej Majhofer – wiceprzewodniczący, dr Adam Michalec, prof. dr hab. Damian Niwiński, prof. dr hab. Krzysztof Oleszkiewicz, dr hab. Krzysztof Pawłowski, prof. PAN, dr Milena Ratajczak, dr hab. Radosław Smolec, prof. PAN, prof. dr hab. Paweł Strzelecki, prof. dr hab. Andrzej Wymolek.

Redaguje kolegium w składzie: Wiktor Bartol, Michał Bejger, Szymon Charzyński – red. nac., Agnieszka Chudek, Anna Durkalec, Marta Gródek, Katarzyna Małek, Michał Miśkiewicz, Wojciech Przybyszewski, Lukasz Rajkowski – z-ca red. nac., Anna Rudnik, Krzysztof Rudnik, Oskar Skibski, Marzanna Wawro – sekr. red.

Adres do korespondencji:
Redakcja *Delty*, ul. Banacha 2, pokój 4020, 02-097 Warszawa
e-mail: delta@mimuw.edu.pl tel. 22-55-44-402.

Okładki i ilustracje:
Anna Ludwicka Graphic Design & Serigrafia.
Skład systemem L^AT_EX wykonała Redakcja.
Druk: Poligrafia NOT poligrafianot.pl

Prenumerata:
Garmond Press: www.garmondpress.pl
Kolporter: www.kolporter.com.pl (tylko instytucje)
RUCH S.A.: www.prenumerata.ruch.com.pl

Numery archiwalne (od 1987 r.) można nabyć w Redakcji osobiście lub zamówić przez e-mail.

Cena 1 egzemplarza: z ostatnich 12 miesięcy 6 zł; wcześniejsze egzemplarze 3 zł



Strona internetowa (w tym artykuły archiwalne, linki itd.):
deltami.edu.pl

Można nas też znaleźć na
facebook.com/Delta.czasopismo

Wydawca: Uniwersytet Warszawski

Wstęp do numeru specjalnego *Delty*

Dane to jeden z surowców, na których oparta jest współczesna cywilizacja. Jak niemal każdy surowiec, wymagają pewnego przetworzenia, zanim staną się w pełni użyteczne. Intelktualną rafinerią, którą wypracowała ludzkość, by radzić sobie z bogatymi złożami wydobywanej z wielką determinacją informacji, jest szeroko rozumiana statystyka. Stanowi ona również wspólny mianownik artykułów zamieszczonych w tym wydaniu *Delty*. Mamy nadzieję, że każdy Czytelnik (nie tylko *Statystyczny*) znajdzie w tym zeszycie coś odpowiadającego jego naukowym zainteresowaniom. Życzymy pouczającej lektury!

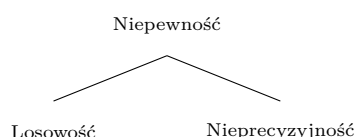
Powstanie niniejszego numeru jest skorelowane z odbywającą się w dniach 3–7 lipca 2023 roku w Warszawie konferencją *The European Meeting of Statisticians*, organizowaną przez Polskie Towarzystwo Matematyczne, Politechnikę Warszawską i Uniwersytet Warszawski. Uczestnicy konferencji otrzymają angielską wersję wydania, którą można również odnaleźć (i polecać wśród swoich niepolskojęzycznych znajomych) na stronie del.tami.edu.pl.

Redakcja

O statystyce, braku precyzji i serach

Przemysław GRZEGORZEWSKI*

* Wydział Matematyki i Nauk Informatycznych, Politechnika Warszawska



Statystyka może być postrzegana jako sztuka podejmowania decyzji w warunkach niepewności. Dostarcza narzędzi do opisu, wyjaśniania i prognozowania zjawisk i procesów zachodzących w realnym świecie oraz weryfikowania hipotez ich dotyczących. Przez długi czas niepewność była utożsamiana z losowością, a w konsekwencji rachunek prawdopodobieństwa był uznawany za jedyną solidnie ugruntowaną teorię pozwalającą opisać niepewność. W ciągu ostatnich pięćdziesięciu lat pojawiły się jednak pewne podejścia rozszerzające klasyczną teorię prawdopodobieństwa lub do niej ortogonalne. Ich wspólną cechą jest próba złagodzenia założeń standardowych metod, aby mogły radzić sobie z innymi rodzajami niepewności, takimi jak nieprecyzyjność.

Należy pamiętać, że sama nieprecyzyjność nie jest (*nomen omen*) precyzyjnie zdefiniowana. Dość często wyniki eksperymentów są nieprecyzyjne z powodu niedokładności aparatury lub błędów popełnionych przez osoby dokonujące pomiarów. Czasami pożądany pomiar jest tak trudny do przeprowadzenia, że jego wynik, z reguły, powinien być traktowany jako bardzo niepewny. Może również zdarzyć się, że dokładna wartość zmiennej jest celowo ukrywana ze względu na wrażliwy charakter gromadzonych danych. We wszystkich tych sytuacjach obserwacje często są rejestrowane jako zbiory (np. przedziały) zawierające ich dokładne wartości, nie dające jednak pełnej informacji o należącej do nich punktowej obserwacji x . Każdy taki zbiór reprezentuje pewien **epistemiczny** stan obserwacji. Istnieją jednak sytuacje, gdy dane eksperymentalne są nieprecyzyjne w samej swej istocie. Typowym przykładem są ludzkie odczucia, które nie mają obiektywnej wartości (jak smak czy nastrój). Z tego typu sytuacją mamy również do czynienia, gdy dane dotyczą zjawiska ze swej natury zmiennego, z rozmytymi lub zmieniającymi się granicami, elastycznymi przedziałami czasowymi lub skalami ocen itp. Każda taka obserwacja, nawet jeśli nie jest precyzyjna, odpowiada pewnemu istniejącemu stanowi rzeczywistości, a zatem ma charakter nieprecyzyjności **ontycznej**.

Wygodną metodę matematycznego reprezentowania nieprecyzyjności zaproponował Lotfi A. Zadeh (1921–2017), który wprowadził **teorię zbiorów rozmytych** stanowiącą rozszerzenie klasycznej teorii zbiorów. Zadeh, jeden z najwybitniejszych myślicieli współczesności, zdał sobie sprawę, że chociaż jesteśmy przyzwyczajeni do dzielenia wszystkiego na „tak” i „nie” lub na czarne i białe, to nasz świat rysuje się w odcieniach szarości. Jego słynna teza, że *wszystko jest kwestią stopnia*, stała się główną myślą stojącą za logiką rozmytą i jej imponującymi zastosowaniami. Warto w tym miejscu wyjaśnić, że logika rozmyta w rzeczywistości nie jest „rozmytą” czy „niedokładną” logiką, ale logiką, która opisuje i ujarzma nieprecyzyjność.

L.A. Zadeh, *Fuzzy sets*, Information and Control 8 (1965), 338–353.

Zauważmy, że w teorii mnogości funkcje $f: X \rightarrow Y$ są zazwyczaj definiowane jako zbiory; dokładniej, podzbiory $X \times Y$ takie, że dla każdego $x \in X$ istnieje dokładnie jedno $y \in Y$ spełniające $(x, y) \in f$. W tym sensie $f(x)$ to jedynie konwencja notacyjna oznaczająca jedyne y , dla którego $(x, y) \in f$.

Podstawowym pojęciem wprowadzonym przez Zadeha jest **zbiór rozmyty**. Niech \mathbb{U} będzie ustaloną przestrzenią rozważań. **Zbiór rozmyty** A w \mathbb{U} jest utożsamiany z odwzorowaniem $A: \mathbb{U} \rightarrow [0, 1]$, nazywanym **funkcją przynależności**, przypisującym każdemu obiektowi $x \in \mathbb{U}$ liczbę rzeczywistą $A(x)$ z przedziału $[0, 1]$, wskazującą stopień przynależności x do A . Zatem zbiór rozmyty A może być postrzegany jako (standardowy) podzbiór $\mathbb{U} \times [0, 1]$, tj.:

$$A = \{(x, A(x)) : x \in \mathbb{U}, A(x) \in [0, 1]\}.$$

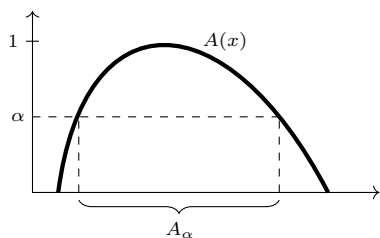
Interpretacja funkcji przynależności jest naturalna: jeśli $A(x) = 1$, to jesteśmy pewni, że element x należy do A , podczas gdy $A(x) = 0$ oznacza, że x nie należy do A . We wszystkich innych przypadkach, tj. gdy $A(x) \in (0, 1)$, mamy częściową przynależność do A . Oznacza to, że jeśli $A(x)$ jest bliskie 1, to stopień przynależności x do A jest wysoki, a jeśli $A(x)$ jest bliskie 0, to stopień przynależności x do A jest niski. Jeśli $A(x) \in \{0, 1\}$ dla wszystkich $x \in \mathbb{U}$, to A jest zbiorem w klasycznym rozumieniu (i odwrotnie, każdy „zwykły” zbiór może być traktowany jako zbiór rozmyty, którego funkcją przynależności jest funkcja charakterystyczna).

Innym ważnym pojęciem związanym ze zbiorami rozmytymi jest tzw. α -**cięcie**. Dla każdego $\alpha \in [0, 1]$ α -cięcie zbioru rozmytego A , oznaczane jako A_α , jest zdefiniowane jako

$$A_\alpha = \begin{cases} \{x \in \mathbb{U} : A(x) \geq \alpha\} & \text{dla } \alpha \in (0, 1], \\ \text{cl}\{x \in \mathbb{U} : A(x) > 0\} & \text{dla } \alpha = 0, \end{cases}$$

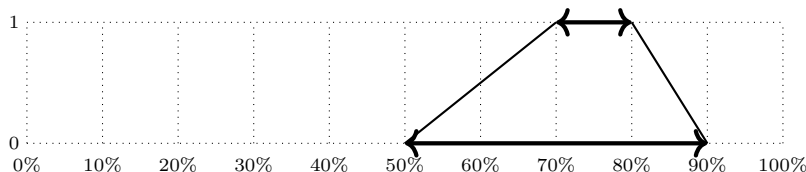
gdzie cl oznacza operację domknięcia (od tej pory zakładamy, że taka operacja jest zdefiniowana na \mathbb{U}). Innymi słowy, α -cięcie to „zwykły” podzbiór tych elementów \mathbb{U} , których stopień przynależności do A jest nie mniejszy niż α . Można pokazać, że każdy zbiór rozmyty jest jednoznacznie wyznaczony przez rodzinę wszystkich swoich α -cięć $\{A_\alpha\}_{\alpha \in [0, 1]}$. Dwa α -cięcia zasługują na szczególną uwagę: A_1 nazywane **jądrem**, które zawiera wszystkie wartości w pełni zgodne z opisywanym przez A pojęciem, oraz A_0 nazywane **nośnikiem**, zawierające elementy zgodne z tym pojęciem w jakimś stopniu.

Ważną podrodziną zbiorów rozmytych są liczby rozmyte. Mówimy, że A jest **liczbą rozmytą**, jeśli $A: \mathbb{R} \rightarrow [0, 1]$ i wszystkie jego α -cięcia są niepustymi przedziałami domkniętymi. Przykład liczby rozmytej jest przedstawiony na rysunku 1.



Rys. 1. Wykres przykładowej funkcji przynależności $A(x)$ liczby rozmytej A

Przykład. Ser Gamonedo to rodzaj sera pleśniowego produkowanego w Asturii (północna Hiszpania). Jest on wędzony, a następnie pozostawiany do dojrzewania w jaskiniach lub innych suchych miejscach. W trosce o zapewnienie odpowiedniej jakości produktu eksperci (degustatorzy) oceniają różne cechy sera: wizualne (kształt, skórka, wygląd), teksturę (twardość i kruchość), zapachowo-smakowe (intensywność i jakość), a także ogólne o nim wrażenie. Ostatnio zapostulowano, aby do opisu swych subiektywnych ocen degustatorzy posłużyli się trapezoidalnymi liczbami rozmytymi. Ten właśnie typ liczb rozmytych jest często stosowany ze względu na łatwość interpretacji oraz przetwarzania. Ocena sera dokonywana jest w skali od 0% (najniższa jakość) do 100% (najwyższa jakość). Na poziomie 0 (por. rys. 2) degustator zaznacza zakres wartości, który jest w jakimś stopniu zgodny z jego opinią (tzn. jakość rozważanej cechy z pewnością nie wychodzi poza ten zakres). Natomiast na poziomie 1 ekspert zaznacza zakres wartości, który w pełni odzwierciedla jego opinię odnośnie jakości. Przykład z rysunku 2 ilustruje sytuację, w której tester uważa, że dany ser spełnia wymagania jakościowe pod względem badanej cechy w 70–80%. Jednocześnie jest on przekonany, że wymagania jakościowe są spełnione w stopniu nie niższym niż 50%, ale też nie wyższym niż 90%.

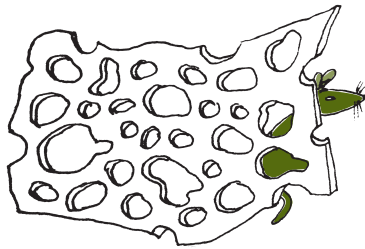


Rys. 2. Przykład eksperckiej opinii reprezentowanej przez trapezoidalny zbiór rozmyty

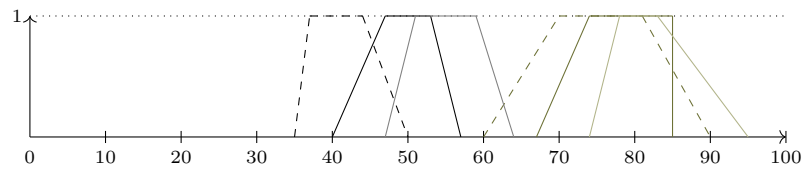


Rozwiązanie zadania M 1750.

Rozważmy dowolne rozmieszczenie liczb. Zauważmy, że liczby od 1 do 2022 reprezentują co najwyżej 2022 wiersze i 2022 kolumny. Istnieją zatem wiersz i kolumna zawierające wyłącznie liczby większe od 2022. Iloczyn dowolnych dwóch z nich jest równy co najmniej $2023 \cdot 2024$, a to więcej niż dowolna liczba wpisana w tablicę. Oznacza to, że nie istnieje rozmieszczenie liczb opisane w zadaniu.



Zakresy wartości oceny na pozostałych poziomach (poza 0 i 1) są następnie interpolowane liniowo, tak aby otrzymać trapezoidalną liczbę rozmytą (por. rys. 2). Przykład ocen wyrażonych przez sześciu degustatorów, które zostały zamodelowane przy użyciu trapezoidalnych liczb rozmytych, przedstawiono na rysunku 3.



Rys. 3. Próbkę sześciu eksperckich opinii reprezentowanych przez trapezoidalne zbiory rozmyte

Bardziej formalnie, mówimy, że A jest **trapezoidalną** liczbą rozmytą, jeśli jej funkcja przynależności jest dana wzorem

$$(1) \quad A(x) = \begin{cases} \frac{x-a_1}{a_2-a_1} & \text{gdy } a_1 \leq x < a_2, \\ 1 & \text{gdy } a_2 \leq x \leq a_3, \\ \frac{a_4-x}{a_4-a_3} & \text{gdy } a_3 < x \leq a_4, \\ 0 & \text{w przeciwnym przypadku,} \end{cases}$$

gdzie $a_1, a_2, a_3, a_4 \in \mathbb{R}$ spełniają $a_1 \leq a_2 \leq a_3 \leq a_4$. Ponieważ trapezoidalna liczba rozmyta (1) jest charakteryzowana przez cztery liczby rzeczywiste, często oznacza się ją przez $A = (a_1, a_2, a_3, a_4)_T$.

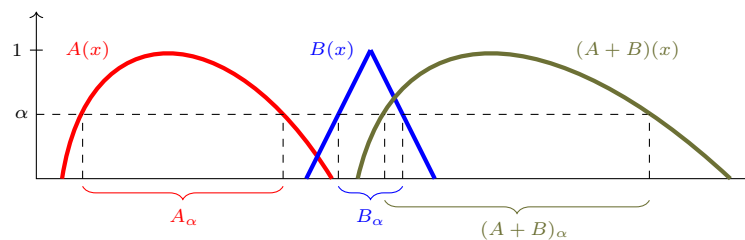
Zdefiniujemy teraz podstawowe operacje arytmetyczne na liczbach rozmytych. Chociaż można je wprowadzić, operując bezpośrednio na funkcjach przynależności, łatwiejsza wydaje się definicja wykorzystująca α -cięcia. Suma dwóch liczb rozmytych A i B jest określona przez tzw. *sumę Minkowskiego* α -cięć składników (patrz rys. 4), tj. dla każdego $\alpha \in [0, 1]$

$$(2) \quad (A + B)_\alpha = [\inf A_\alpha + \inf B_\alpha, \sup A_\alpha + \sup B_\alpha],$$

gdzie $\inf I$ oraz $\sup I$ oznaczają, odpowiednio, lewy i prawy koniec przedziału I .

Warto podkreślić, że suma trapezoidalnych liczb rozmytych jest trapezoidalną liczbą rozmytą, tzn. jeśli $A = (a_1, a_2, a_3, a_4)_T$ i $B = (b_1, b_2, b_3, b_4)_T$, to

$$A + B = (a_1 + b_1, a_2 + b_2, a_3 + b_3, a_4 + b_4)_T.$$



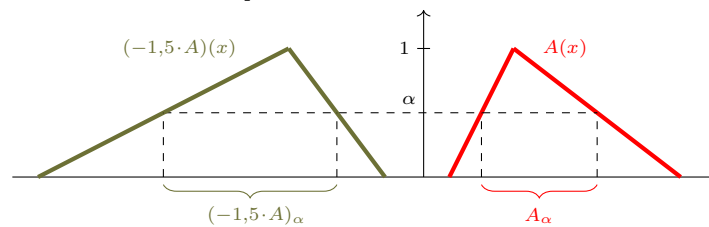
Rys. 4. Suma liczb rozmytych A i B

Podobnie, mnożenie liczby rozmytej A przez skalarną wartość $\theta \in \mathbb{R}$ jest określone poprzez tzw. *iloczyn Minkowskiego* α -cięć (patrz rys. 5), tzn. dla każdego $\alpha \in [0, 1]$

$$(3) \quad (\theta \cdot A)_\alpha = [\min\{\theta \inf A_\alpha, \theta \sup A_\alpha\}, \max\{\theta \inf A_\alpha, \theta \sup A_\alpha\}].$$

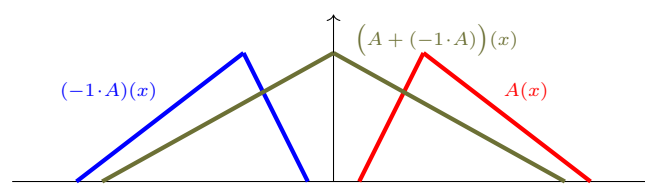
Pomnożenie trapezoidalnej liczby rozmytej $A = (a_1, a_2, a_3, a_4)_T$ przez skalar θ daje trapezoidalną liczbę rozmytą:

$$\theta \cdot A = \begin{cases} (\theta a_1, \theta a_2, \theta a_3, \theta a_4)_T & \text{gdy } \theta \geq 0, \\ (\theta a_4, \theta a_3, \theta a_2, \theta a_1)_T & \text{gdy } \theta < 0. \end{cases}$$



Rys. 5. Mnożenie liczby rozmytej A przez skalar

Niestety, w ogólności $A + (-1 \cdot A) \neq \mathbb{1}_0$ (patrz rys. 6). W konsekwencji, dla różnicy opartej na operacji Minkowskiego nie zawsze zachodzi $(A + (-1 \cdot B)) + B = A$.



Rys. 6. Ilustracja problemów z odejmowaniem liczb rozmytych

Aby sprostać niektórym trudnościom związanym z brakiem „porządnej” operacji odejmowania liczb rozmytych, zwłaszcza w kontekście konstruowania narzędzi do wnioskowania statystycznego, często rozważa się *odległość* między liczbami rozmytymi. Tak jak poprzednio, wygodnie będzie do jej opisu wykorzystać α -cięcia. Odległość między dwoma odcinkami $[a, b]$, $[c, d]$ możemy określić jako odległość między punktami (a, b) , (c, d) na płaszczyźnie (czyli $\sqrt{(a-c)^2 + (b-d)^2}$). Następnie odległość między liczbami rozmytymi A, B zdefiniujemy jako średnią kwadratową z odległości między A_α i B_α po wszystkich α , co prowadzi do następującego wzoru:

$$(4) \quad D(A, B) = \sqrt{\int_0^1 [(\inf A_\alpha - \inf B_\alpha)^2 + (\sup A_\alpha - \sup B_\alpha)^2] d\alpha}.$$

Odnotujmy, że rzeczywiście jest to odległość, czyli *metryka* (patrz np. artykuł Jarosława Górnickiego z Δ_{21}^5). Oczywiście jest własność $D(A, B) \geq 0$, jak i to, że $D(A, B) = 0$ tylko w przypadku $A = B$. Nieco trudniejsza w uzasadnieniu jest nierówność trójkąta $D(A, B) + D(B, C) \geq D(A, C)$, którą pozostawiamy jako ćwiczenie.

Załóżmy teraz, że mamy do czynienia z niezależnymi, rozmytymi obserwacjami dwóch populacji; n obserwacji z populacji pierwszej, $\mathbf{x} = (x_1, \dots, x_n)$, oraz m obserwacji z drugiej, $\mathbf{y} = (y_1, \dots, y_m)$ (wszystkie x_i oraz y_i są liczbami rozmytymi). Chcemy przekonać się, czy występuje znacząca różnica między tymi populacjami. W tym celu mierzymy odległość (4) między średnimi arytmetycznymi wyznaczonymi dla obu próbek. Zauważmy, że faktycznie potrafimy już obliczać średnią arytmetyczną liczb rozmytych (która sama jest liczbą rozmytą), gdyż mamy narzędzia do ich dodawania i mnożenia przez liczby rzeczywiste. Pozostaje pytanie, czy konkretna odległość między średnimi, np. 3,14, jest duża, czy mała? W tym miejscu na scenę wkracza statystyka matematyczna!

W żargonie statystyków naszym celem jest weryfikacja hipotezy zerowej H_0 , mówiącej, że obie próbki pochodzą tak naprawdę z tej samej populacji, przeciw hipotezie alternatywnej, że populacje się różnią. Jeśli hipoteza zerowa jest prawdziwa, oczekujemy, że średnie z obu próbek nie będą się zbytnio różnić. Natomiast znacząca różnica między średnimi wskazuje na to, że badane próbki pochodzą z różnych populacji.

Aby zdecydować, czy mierzona odległość jest wystarczająco duża, by uznać ją za znaczącą, statystycy często odwołują się do pojęcia *p-wartości*. W naszym przypadku będzie to prawdopodobieństwo (przy założeniu, że hipoteza zerowa jest prawdziwa) otrzymania co najmniej tak dużej odległości między średnimi, jak ta zaobserwowana. Zgodnie z intuicją, jeśli to prawdopodobieństwo jest niskie, mamy dobry powód, by odrzucić hipotezę zerową. Problem polega na tym, że w naszym przypadku nie możemy dokładnie wyznaczyć tego prawdopodobieństwa, ponieważ hipoteza zerowa orzeka jedynie, że dwie populacje można traktować jak jedną, ale nie podaje nam konkretnego opisu probabilistycznego tej populacji. Musimy zatem odwołać się do innego, sprytnego pomysłu.

Niech \mathbf{v} będzie *złączeniem* dwóch próbek, tj. $v_i = x_i$ dla $1 \leq i \leq n$ oraz $v_i = y_{i-n}$ dla $n+1 \leq i \leq N$, gdzie $N = n+m$. Niech teraz \mathbf{v}^* oznacza permutację początkowego zbioru danych \mathbf{v} . Z pierwszych n elementów ciągu \mathbf{v}^* tworzymy próbkę \mathbf{x}^* , natomiast z pozostałych m elementów, próbkę \mathbf{y}^* . Innymi słowy, dokonujemy losowego podziału elementów \mathbf{v} na dwie próbki o rozmiarach, odpowiednio, n i m . Zauważmy, że jeśli hipoteza H_0 jest spełniona, tj. obie próbki pochodzą z tego samego rozkładu, to taka operacja nie zmienia losowości stojącej za obserwowanymi próbkami (z punktu widzenia losowości dowolny taki podział jest w pewnym sensie równoprawny). W konsekwencji możemy *oszacować* prawdziwą *p*-wartość jako frakcję wszystkich możliwych permutacji \mathbf{v}^* , które prowadzą do większej różnicy między średnimi \mathbf{x}^* i \mathbf{y}^* niż ta zaobserwowana.



Rozwiązanie zadania M 1752.

Zauważmy, że dla dowolnych $M, a, b \in [0, 1]$ takich, że $M \geq a, b$, zachodzi nierówność:

$$(M-a)(M-b)(1-bM) \geq 0,$$

co po przekształceniach daje

$$(1-bM+b^2)(1-aM+M^2) \geq 1-ab+b^2.$$

Aby udowodnić nierówność z zadania dla $n=2$, wystarczy w powyższej nierówności wziąć $a=b=x_1$ oraz $M=x_2$.

Załóżmy, że teza zadania zachodzi dla pewnego n ; wywnioskujemy stąd jej słuszność dla $n+1$. Bez straty ogólności założmy, że $x_{n+1} = \max\{x_1, \dots, x_{n+1}\}$.

Używając powyższej nierówności dla $b=x_n$, $M=x_{n+1}$ oraz $a=x_1$, otrzymujemy:

$$(1-x_n x_{n+1} + x_n^2)(1-x_{n+1} x_1 + x_{n+1}^2) \geq (1-x_n x_1 + x_n^2).$$

Dlatego

$$\prod_{\text{cykl}}^{n+1} (1-x_i x_{i+1} + x_i^2) \geq \prod_{\text{cykl}}^n (1-x_i x_{i+1} + x_i^2) \geq 1,$$

gdzie przez „ \prod_{cykl} ” rozumiemy „zacyklony” iloczyn. Powołanie się na zasadę indukcji kończy dowód.

Równanie (5) można traktować jako definicję *prawdziwej* p-wartości, pod warunkiem zdarzenia, że nasze próbki sumują się (jako zbiory swoich wyrazów) do \mathbf{v} . Takie podejście jest dość standardowe przy projektowaniu tzw. *testów nieparametrycznych*, np. *testu serii* opisanego szczegółowo w artykule o tej samej nazwie w Δ_{17}^9 .

Poniższa tabela przedstawia opinie dwóch ekspertów dotyczące ogólnego wrażenia na temat łącznej liczby 78 próbek sera Gamonedo, por. Ramos-Guajardo i in. (2019). Każdy ekspert oceniał inne próbki.

Ekspert 1	Ekspert 2
(65, 75, 85, 85)	(50, 50, 63, 75)
(35, 37, 44, 50)	(39, 47, 52, 60)
(66, 70, 75, 80)	(60, 70, 85, 90)
(70, 74, 80, 84)	(50, 56, 64, 74)
(65, 70, 75, 80)	(39, 45, 53, 57)
(45, 50, 57, 65)	(55, 60, 70, 76)
(60, 66, 70, 75)	(50, 50, 57, 67)
(65, 65, 70, 76)	(65, 67, 80, 87)
(60, 65, 75, 80)	(50, 50, 65, 75)
(55, 60, 66, 70)	(50, 55, 64, 70)
(60, 65, 70, 74)	(39, 46, 53, 56)
(30, 46, 44, 54)	(19, 29, 41, 50)
(60, 65, 75, 75)	(40, 47, 52, 56)
(70, 75, 85, 85)	(54, 55, 65, 76)
(44, 45, 50, 56)	(59, 65, 75, 85)
(51, 56, 64, 70)	(50, 52, 57, 60)
(40, 46, 54, 60)	(60, 60, 70, 80)
(55, 60, 65, 70)	(50, 54, 61, 67)
(80, 85, 90, 94)	(40, 46, 50, 50)
(80, 84, 90, 90)	(44, 50, 56, 66)
(65, 70, 76, 80)	(60, 64, 75, 85)
(75, 80, 86, 90)	(54, 56, 64, 75)
(65, 70, 73, 80)	(50, 50, 60, 66)
(70, 80, 84, 84)	(44, 46, 55, 57)
(55, 64, 70, 70)	(59, 63, 74, 80)
(64, 73, 80, 84)	(49, 50, 54, 58)
(50, 56, 64, 70)	(55, 60, 70, 75)
(55, 55, 60, 70)	(44, 47, 53, 60)
(60, 70, 75, 80)	(19, 20, 30, 41)
(64, 71, 80, 80)	(40, 44, 50, 60)
(50, 50, 55, 65)	(50, 50, 59, 66)
(50, 54, 60, 65)	(50, 53, 60, 66)
(65, 75, 80, 86)	(50, 52, 58, 61)
(50, 55, 60, 66)	(60, 65, 72, 80)
(40, 44, 50, 50)	(50, 50, 55, 60)
(70, 76, 85, 85)	(30, 34, 43, 47)
(44, 50, 53, 60)	(19, 25, 36, 46)
(34, 40, 46, 46)	(53, 63, 74, 80)
(40, 45, 51, 60)	
(84, 90, 95, 95)	

Formalnie można wyrazić tę ideę wzorem:

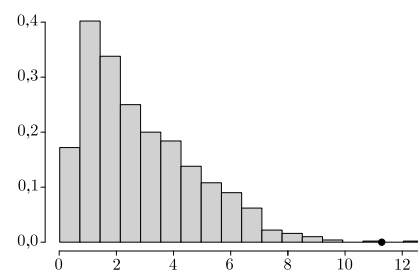
$$(5) \quad \text{p-wartość} \simeq \frac{1}{N!} \sum_{\mathbf{v}^*} \mathbb{1}(T(\mathbf{v}^*) \geq t_0),$$

w którym suma obejmuje wszystkie permutacje \mathbf{v}^* zbioru \mathbf{v} , $T(\mathbf{v}^*)$ oznacza odległość między średnimi z próbek \mathbf{x}^* i \mathbf{y}^* , zaś t_0 jest obserwowaną odległością między średnimi z próbek \mathbf{x} i \mathbf{y} . Wartość wyrażenia $\mathbb{1}(\text{warunek})$ jest równa 1, jeśli dany warunek jest spełniony, a 0 w przeciwnym przypadku.

Wzór (5) może być dalej uproszczony dzięki wykorzystaniu faktu, że permutacje można podzielić na grupy rozmiaru $n!m!$, w których permutacje dają te same średnie z próbek \mathbf{x}^* i \mathbf{y}^* (są to permutacje, z których każdą można uzyskać z innej poprzez permutowanie pierwszych n i ostatnich m obserwacji). Jednak nawet po dokonaniu takiego uproszczenia liczba składników rozważanej sumy, wynosząca wówczas $\binom{N}{n}$, rośnie wykładniczo wraz z N (utrzymując stosunek n/N na ustalonym poziomie). Dlatego, zamiast uwzględniać wszystkie możliwe permutacje, ograniczamy się do średniej obliczonej na podstawie (możliwie wielu) losowo wybranych permutacji. Konkretniej, niech $\mathbf{v}_1^*, \mathbf{v}_2^*, \dots, \mathbf{v}_K^*$ będą pewnymi losowymi permutacjami \mathbf{v} (gdzie K jest odpowiednio duże, np. większe niż 1000). Wtedy przybliżenie p-wartości naszego testu jest określone przez:

$$(6) \quad \text{p-wartość} \simeq \frac{1}{K} \sum_{k=1}^K \mathbb{1}(T(\mathbf{v}_k^*) \geq t_0).$$

Przykład. Wykorzystamy teraz dane z pracy Ramos-Guajardo i in. (2019), aby porównać ogólne wrażenia dwóch degustatorów na temat serów Gamonedo. Trapezoidalne liczby rozmyte odpowiadające 40 opiniom wydanym przez pierwszego eksperta oraz 38 opiniom drugiego są zebrane w tabeli na marginesie. Liczby w nawiasach odpowiadają notacji używanej do opisu trapezoidalnych liczb rozmytych, np. $x_1 = (65, 75, 85, 85)_T$, $y_1 = (50, 50, 63, 75)_T$ itd. Chcemy sprawdzić, czy istnieje zgodność między tymi dwoma ekspertami co do ogólnego wrażenia, jakie robią oceniane sery. Aby osiągnąć ten cel, weryfikujemy hipotezę zerową H_0 głoszącą, że nie ma istotnej różnicy między opiniami ekspertów, przeciwko temu, że ich opinie na temat jakości sera różnią się. Na podstawie danych z tabeli można wyznaczyć średnie wartości $\bar{x} = (57,65; 63,20; 69,18; 73,48)_T$ i $\bar{y} = (47,34; 51,21; 59,87; 66,84)_T$. Podstawiając te wyniki do (4), otrzymujemy wartość statystyki testowej $t_0 = D(\bar{x}, \bar{y}) = 11,26$. Następnie, po złączeniu próbek i wygenerowaniu $K = 1000$ losowych permutacji, stosując (6), otrzymujemy przybliżenie p-wartości równe 0,002. Jej interpretacja jest pokazana na rysunku 7, przedstawiającym histogram wszystkich uzyskanych symulacyjnie różnic $D(\mathbf{x}^*, \mathbf{y}^*)$. Czarny punkt oznacza obserwowaną wartość t_0 statystyki testowej. Słabo widoczny szary obszar na prawo od tego punktu odpowiada prawdopodobieństwu uzyskania odległości między średnimi \mathbf{x}^* a \mathbf{y}^* nie mniejszej niż t_0 . Dlatego też możemy bez większego wahania odrzucić hipotezę zerową i wnioskować, że nie ma ogólnej zgody między opiniami rozważanych ekspertów na temat sera Gamonedo.



Rys. 7

Przedstawiony permutacyjny test zgodności dwóch próbek zawierających nieprecyzyjne informacje jest jednym z wielu przykładów, jak modelowanie rozmyte może być połączone z klasycznym wnioskowaniem statystycznym. Pomimo początkowego sceptycyzmu wobec prób połączenia obu teorii z czasem stało się jasne, że zarówno statystyka, jak i teoria zbiorów rozmytych nie powinny być uważane za konkurencyjne i mogą skutecznie się uzupełniać. Co więcej, rozszerzenie statystyki o analizę zbiorów rozmytych nie tylko pomaga rozwiązać pewne problemy, ale jednocześnie stawia nowe pytania badawcze. W szczególności, rozróżnienie między tzw. zbiorami ontologicznymi i epistemicznymi prowadzi do różnych definicji nawet tak podstawowych pojęć jak wariancja, a co za tym idzie – różnych narzędzi wnioskowania.

Algorytm magicznych piątek

Oskar SKIBSKI*

* Wydział Matematyki, Informatyki i Mechaniki, Uniwersytet Warszawski

† Zawarte w artykule treści są jedynie (błędny) poglądami autora i nie muszą odzwierciedlać (i nie odzwierciedlają) poglądów redakcji (przyp. red.).

Kiedy piszemy, że czas działania $f(n)$ to $O(g(n))$, to mamy na myśli, że f jest co najwyżej rzędu g , czyli dla dużych wartości n funkcja $f(n)$ rośnie nie szybciej niż $g(n)$. Formalnie: istnieją takie stałe c oraz n_0 , że dla $n \geq n_0$ zachodzi $f(n) \leq c \cdot g(n)$. Jest to zatem oszacowanie górne, ale zwykle nie najlepsze, jakie da się znaleźć.

Oczytany Czytelnik może mieć (bardzo trafne) skojarzenie z algorytmem QuickSort. W algorytmie tym, aby posortować zbiór liczb, dzielimy go na elementy mniejsze i większe od pewnego elementu, a potem rekurencyjnie sortujemy obie części. Nie jest przypadkiem, że autorem algorytmu QuickSort jest właśnie Tony Hoare.



Wszyscy wiemy, że magia istnieje. Naukowcy starają się tłumaczyć większość zjawisk matematycznymi równaniami lub wymyślają nowe definicje, aby udawać, że to, czego nie da się opisać równaniami, też ma sens. Wiemy jednak dobrze, że udaje im się przekonać tylko tych już przekonanych. Jeżeli królika nie było w kapeluszu, a magik go z niego wyciągnął, to nie ma równania, które by to opisało. Jeden to jeden, a zero to zero.†

Jako że w *Delcie* konsekwentnie udajemy, że magia nie istnieje, w tym artykule omówimy algorytm, który wydaje się magiczny, a magiczny nie jest. Możemy porównać go do sztuczki magicznej, w której magik po prostu wyciągnął królika z kapelusza, ale wcześniej nie pokazał widowni, że kapelusz jest pusty. Przez chwilę może nas to zaskoczyć („O! Królik!”). Wnikliwa analiza i logiczne wnioskowanie pozwalają jednak wytłumaczyć tę sztuczkę: królik musiał siedzieć magikowi na głowie przez całe przedstawienie, a przy zdejmowaniu kapelusza magik zręcznie podważył mu łapki i dzięki temu miał co z kapelusza wyciągnąć.

Sztuczka magiczna: Mamy dany kapelusz z n króli. . . nie, no bądźmy choć trochę poważni! Mamy dany zbiór A zawierający n liczb (liczby mogą się powtarzać). Jeden z widzów podaje dowolną liczbę k od 1 do n . Magik w *czasie liniowym* odpowiada, jaka jest k -ta co do wielkości liczba w zbiorze A .

Jak wykonać sztuczkę: Dla niektórych wartości k zadanie to jest bardzo proste, np. dla $k = 1$ nasz problem sprowadza się do szukania najmniejszego elementu w zbiorze, co łatwo zrobić w czasie liniowym. W ogólności (np. jeżeli $k = \lceil n/2 \rceil$) nie jest jednak jasne, jak możemy to zrobić. Przypadek $k = \lceil n/2 \rceil$ jest zresztą bardzo ważny dla statystyków, gdyż dotyczy tzw. *mediany*, która potrafi mieć dla nich nawet więcej uroku niż średnia. Naturalnym pomysłem jest posortowanie wszystkich liczb i potem wskazanie tej z pozycji k -tej. Najszybsze algorytmy sortowania działają jednak w czasie $O(n \log n)$, a nasz problem wydaje się dużo prostszy niż problem sortowania – jak rozwiązać go w czasie liniowym?

Pomysł na rozwiązanie pożyczymy ze sztuczki magicznej polegającej na przepiłowaniu asystentki na pół, albo, jak kto woli, z *algorytmu Hoare’a*. Weźmy losowy element m ze zbioru A i podzielmy cały nasz zbiór na dwa niepuste zbiory tak, aby w pierwszym zbiorze znalazły się tylko elementy mniejsze bądź równe m (zbiór A_{\leq}), a w drugim większe bądź równe m (zbiór A_{\geq}). Aby zbiory były niepuste, możemy np. do pierwszego z nich wrzucić wszystkie elementy mniejsze bądź równe m i jeżeli drugi zbiór okaże się pusty, przerzucić do niego m . Teraz jeżeli A_{\leq} ma co najmniej k elementów, to szukany element musi być w A_{\leq} – rekurencyjnie szukamy go zatem w tym zbiorze. Jeżeli z kolei A_{\leq} ma mniej niż k elementów, to szukany element jest w zbiorze A_{\geq} : musimy zatem rekurencyjnie znaleźć tam element $k - |A_{\leq}|$ co do wielkości.

Pomysł jest prosty, jednak jego skuteczność nie musi być zbyt dobra – jeżeli będziemy pechowo losować zawsze najmniejszy lub największy element, to co krok nasz zbiór będzie się zmniejszał tylko o jedną liczbę. A skoro każdy krok zajmuje czas liniowy, to pesymistyczny czas działania naszego algorytmu będzie kwadratowy: $O(n^2)$. To gorzej, niż gdybyśmy użyli sortowania!

Nasz algorytm możemy jednak poprawić, zmieniając liczbę, względem której dzielimy zbiór. Aby znaleźć wskazaną przez widza kartę, magik zwykle nie zdaje się na los, ale starannie przekłada talię, tak aby kontrolować, gdzie ta karta jest. My zrobimy tak samo – trochę przetasujemy liczby i wyciągniemy taką, która zagwarantuje nam, że żadna z części nie będzie zbyt duża.

Tak właśnie działa *algorytm magicznych piątek*. W dowolny sposób podzielmy nasz zbiór na piątka elementów (ostatnia piątka może być niepełna) i dla każdej znajdziemy jej medianę. Teraz, korzystając rekurencyjnie z naszego algorytmu, znajdziemy medianę median: oznaczmy ją przez m . Podzielmy nasz zbiór na trzy części: elementy mniejsze od m (oznaczane $A_{<}$), elementy równe m (oznaczane $A_{=}$) oraz elementy większe od m (oznaczane $A_{>}$). Teraz jeżeli $|A_{<}| \geq k$, to rekurencyjnie szukamy w nim elementu k -tego co do wielkości.



Rozwiązanie zadania F 1076.

Jeśli zaniedbamy efekty brzegowe, to po naładowaniu okładek pojawi się między nimi jednorodne pole elektryczne prostopadłe do powierzchni okładek. Jednorodnie też będą gęstości ładunków. Przyjmijmy, że całkowita powierzchnia każdej z okładek wynosi S , wówczas gęstości ładunków wyniosą odpowiednio: $\sigma_1 = Q_1/S$ i $\sigma_2 = Q_2/S$. Na podstawie prawa Gaussa stwierdzamy, że wewnątrz kondensatora każda z okładek jest źródłem pola elektrycznego o natężeniu

$$E_i = \frac{\sigma_i}{2\epsilon_0}.$$

W powyższym wzorze ϵ_0 oznacza przenikalność dielektryczną próżni. Pole elektryczne skierowane jest „od okładki”, jeżeli jej ładunek jest dodatni, i „do okładki” – jeśli jest ujemny. Wypadkowe pole elektryczne wewnątrz kondensatora jest sumą pól od obu okładek i wynosi:

$$E = E_1 - E_2 = \frac{\sigma_1 - \sigma_2}{2\epsilon_0} = \frac{Q_1 - Q_2}{2\epsilon_0 S},$$

a wartość różnicy potencjałów:

$$U = Ed = (E_1 - E_2)d = \frac{(Q_1 - Q_2)d}{2\epsilon_0 S} = \frac{Q_1 - Q_2}{2C}.$$

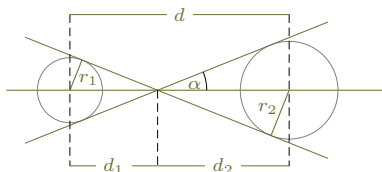


Rozwiązanie zadania F 1075.

Obserwacje z kierunków pomiędzy dwiema styczniymi do powierzchni gwiazd (rysunek) związane są z pojawianiem się zaćmień. Te stycznie przecinają się w punkcie dzielącym odcinek d na odcinki d_1 i d_2 , $d = d_1 + d_2$. Odpowiada to zakresowi kątów obserwacji α , dla których

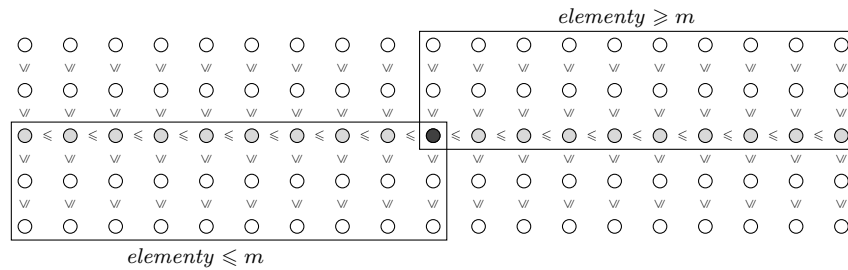
$$|\sin \alpha| \leq \frac{r_1}{d_1} = \frac{r_2}{d_2} = \frac{r_1 + r_2}{d}.$$

Rozważany w zadaniu model opisuje najprostszy przypadek układu podwójnego. W ogólnym przypadku wzajemne oddziaływania grawitacyjne mogą prowadzić do odkształceń gwiazd od kształtu kulistego do elipsoidalnego lub nawet do przepływu materii pomiędzy nimi.



Jeżeli $|A_{<}| < k$, ale $|A_{<}| + |A_{=}| \geq k$, to znaczy, że k -ty co do wielkości element to po prostu m . Jeśli zaś $|A_{<}| + |A_{=}| < k$, to naszego elementu musimy szukać rekurencyjnie w zbiorze $A_{>}$ – jest on tam $(k - |A_{<}| - |A_{=}|)$ -ty co do wielkości. Tadam!

Wyjaśnienie sztuczki: No dobrze, ale skąd pewność, że żadna z części nie będzie zbyt duża, a czas działania liniowy? Zauważmy, że w połowie piątek mediana jest mniejsza bądź równa m . W każdej takiej piątce co najmniej połowa elementów jest mniejsza bądź równa m . Wynika z tego, że co najmniej 1/4 wszystkich elementów jest mniejsza bądź równa m . To oznacza, że elementów większych od m jest nie więcej niż 3/4 wszystkich. Analogicznie, elementów mniejszych od m jest też nie więcej niż 3/4 wszystkich. Oznacza to, że nieważne, który przypadek wystąpi, rekurencyjnie wywołamy nasz algorytm na zbiorze zmniejszonym o co najmniej 25%.



Czy to wystarcza, aby nasz algorytm działał w czasie liniowym? Okazuje się, że tak! Oznaczmy czas wykonania naszego algorytmu przez $T(n)$. Podzielenie na piątki i wybranie z każdej mediany wykonujemy w czasie liniowym – $O(n)$. Wybranie mediany median zajmie nam czas $T(\lceil \frac{n}{5} \rceil)$. Ostatnim krokiem jest wywołanie rekurencyjne dla którejś z części. Jak już uzasadniliśmy, żadna z części nie jest większa niż 3/4 całości, a zatem czas to $T(\lceil \frac{3n}{4} \rceil)$. Dostajemy więc następujące oszacowanie:

$$T(n) \leq O(n) + T\left(\left\lceil \frac{n}{5} \right\rceil\right) + T\left(\left\lceil \frac{3n}{4} \right\rceil\right).$$

Jak się okazuje, z oszacowania tego wynika, że czas naszego algorytmu jest liniowy! Kluczowa jest tu nierówność $\frac{1}{5} + \frac{3}{4} = \frac{19}{20} < 1$, która zapewnia, że $T(n)$ nie rośnie zbyt szybko.

Żeby to zobaczyć, weźmy najpierw za n pewną potęgę liczby 20, aby ładnie dzieliła się przez 4 i 5. Wiemy, że czas potrzebny na podzielenie na piątki i wybranie median możemy oszacować z góry przez cn dla pewnego c naturalnego. Łatwo pokazać przez indukcję, że zachodzi: $T(n) \leq 20cn$:

$$T(n) \leq cn + T\left(\frac{n}{5}\right) + T\left(\frac{3n}{4}\right) \leq cn + 20cn\left(\frac{1}{5} + \frac{3}{4}\right) = 20cn.$$

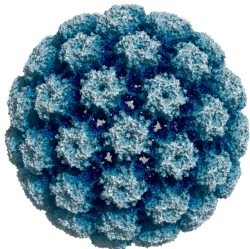
Z oszacowania dla potęg 20 dostajemy natychmiast oszacowanie dla innych liczb: Weźmy dowolne n i dobierzmy k tak, aby $20^k < n \leq 20^{k+1}$. Skoro T jest funkcją niemalejącą, to możemy oszacować $T(n)$ przez $T(20^{k+1})$. Ale wiemy, że $n > 20^k$, więc

$$T(n) \leq T(20^{k+1}) \leq 20c \cdot 20^{k+1} \leq 20^2 cn.$$

Skoro istnieje taka stała $20^2 c$, że dla dowolnego n zachodzi $T(n) \leq 20^2 c \cdot n$, to oznacza, że nasz algorytm działa w czasie $O(n)$.

Na koniec pozostaje zadać pytanie – czemu akurat piątki? Naturalne wydaje się wzięcie liczby nieparzystej (aby istniały wartości środkowe), ale czy nie mogliśmy wziąć trójek albo siódemek? Albo jedenastek?

Okazuje się, że trójek wziąć nie mogliśmy: musielibyśmy najpierw znaleźć medianę 1/3 wszystkich liczb, a problem zredukowalibyśmy znowu o 1/4. Ponieważ $1/3 + 3/4 > 1$, to odpowiednia rekurencja dałaby czas $O(n \log n)$. Moglibyśmy natomiast wziąć siódemki, dziewiątki itd.: Szukanie mediany wśród 1/7 czy 1/9 liczb byłoby nawet szybsze niż szukanie ich wśród 1/5 liczb. Wzrosłyby nam jednak koszt wyznaczania median (czyli to enigmatyczne c w powyższym dowodzie) i implementacja byłaby bardziej złożona. W algorytmie użyte są więc piątki, bo pięć jest najmniejszą liczbą nieparzystą większą niż cztery. Ot, cała magia.



Białko kapsydowe wirusa HPV typu 16. Udostępnione na licencji CC BY-SA 4.0 przez Wikimedia Commons, autor: Opabinia regalalis

Wirus znaczy jad

Odkrycie wirusów w ogóle – formy życia wymagającej jako gospodarza pełnej komórki bakteryjnej, zwierzęcej lub roślinnej – nie jest zbyt odległe. Ze zdziwieniem zauważył wirusa wścieklizny (jeszcze bez tej nazwy) Ludwik Pasteur w końcu XIX wieku. Odkrył, że można wyodrębnić czynnik zakaźny, przesączaając wyciągi tkanek przez porcelanowe filtry zatrzymujące bakterie. Był to pierwszy sygnał, że wirus wścieklizny musi być od bakterii mniejszy. Skonstruowanie po latach mikroskopu elektronowego spowodowało, że zobaczyliśmy różne wirusy, poznaliśmy ich kształty. Urządzenie to okazało się także niezwykle przydatne w badaniach podstawowych przemian dużych cząsteczek, bez których „nie ma życia”: białek, kwasów nukleinowych i ich wzajemnych zależności strukturalnych i metabolicznych. Słowo „wirus” znane jest powszechnie, choć wiedza szczegółowa o różnych wirusach jest dostępna już tylko dla specjalistów. Co pewien czas odnajdujemy w mediach wiadomości o wirusach i o naszych z nimi relacjach (pandemie!) – i wtedy przydatne jest posłuchać, co o nich wiedzą wirusolodzy.

W ostatnich dniach maja rozpoczęto w Polsce zapisy dzieci 12- i 13-letnich na nieobowiązkowe, bezpłatne szczepienia przeciw wirusom ludzkiego brodawczaka, HPV. Niecodzienna wiedza i niecodzienne osiągnięcia medyczne stoją za tym faktem. Wirusy brodawczaka wywołują morfologiczne zmiany w komórkach płaskiego nabłonka szyjki macicy, które z czasem przekształcają się w komórki raka płaskonabłonkowego. Niezwykle odkrycie uczonych, że choroba nowotworowa może być wywołana przez wirusy, uruchomiło poszukiwania i znalezienie!!! szczepionki. Konstrukcja szczepionki opisana została 17 lat temu. Kliniczne badania bezpieczeństwa szczepień przeprowadzono pod kontrolą Europejskiej Agencji Leków. Istnieją także dane o skuteczności szczepionki: według badania przeprowadzonego na ponad 2000 kobiet z krajów skandynawskich zapadalność na raka szyjki macicy obniżyła się o 90% w stosunku do liczby zachorowań w okresie, gdy nie było tych szczepień. Przez 8 lat trwania badania wśród uczestniczek nie odnotowano ani jednego nowego przypadku tej choroby, co świadczy o wysokiej skuteczności szczepionki. Powszechne szczepienia wprowadzono najwcześniej w Australii, w 2007 roku, w 2008 roku w Wielkiej Brytanii, do 2019 roku niemal wszystkie kraje Unii Europejskiej już takie bezpłatne szczepienia dla dziewcząt i chłopców wprowadziły. Na całym świecie są one stosowane w 125 krajach (według danych z maja 2023 r.).

Znanych jest ponad 150 typów HPV, w tym typy niskiego ryzyka, które wywołują m.in. niezłośliwe brodawki narządów płciowych (kłykciny) i brodawczaki, oraz typy o wysokim potencjale onkogenym odpowiedzialne za zmiany przedrakowe i za raka szyjki macicy, a także inne choroby nowotworowe. Do zakażenia HPV dochodzi drogą płciową, najczęściej w początkowym okresie po rozpoczęciu aktywności seksualnej. W ciągu swojego życia 80% aktywnych seksualnie kobiet i mężczyzn było lub będzie zakażonych HPV. Zakażenia HPV mogą prowadzić również do raka odbytu, pochwy, sromu, prącia, przetrzeni ustno-gardłowej oraz okolic głowy i szyi. Wysoce onkogenne są trzy najczęściej występujące typy: HPV-16, HPV-18 i HPV-45, uważa się je za odpowiedzialne za 80–90% różnych nowotworów, o których wspomniano wyżej. Na 1 mln kobiet zakażonych wirusem onkogenym na raka zapada 8 tysięcy; śmiertelność na raka szyjki macicy wynosi 50%.

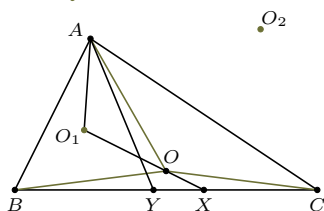
Konstruowane według takiej samej zasady szczepionki mają jako antygen wirusowe białko L1. Skierowane na różne typy wirusów szczepionki składają się z różnych (choć chemicznie podobnych) białek L1 specyficznych dla danego typu wirusa. Są to białka powłoki wirusów ubezpieczające przed zakażeniem danym wirusem. Białka te tworzą wirusopodobne puste otoczki, we wnętrzu nie ma żadnego materiału genetycznego, nie wywołują zakażeń. Uzyskiwane są dzięki laboratoryjnej metodzie z zakresu metod inżynierii genetycznej, nie z preparatyki prawdziwego wirusa.

Szczepionki przeciw HPV są **pierwszymi preparatami leczniczymi**, których głównym celem jest zapobieganie chorobom nowotworowym. Dla mnie: rewelacja!

Magdalena FIKUS (magda.fikus@gmail.com)



Rozwiązanie zadania M 1751.



Niech O_1 i O_2 będą środkami okręgów opisanych odpowiednio na trójkątach AOB i AOC . Oczywiście O_1 leży na prostej OX , dlatego

$$\sphericalangle AO_1X = 2\sphericalangle ABO = \pi - \sphericalangle AOB = \pi - 2\sphericalangle ACB = \sphericalangle AYX,$$

a zatem O_1 leży na okręgu opisanym na AXY . Analogicznie dowodzimy, że O_2 też na nim leży.

Skoro wszystkie modele są błędne, to które są użyteczne?

*MI2.AI, Uniwersytet Warszawski,
Politechnika Warszawska

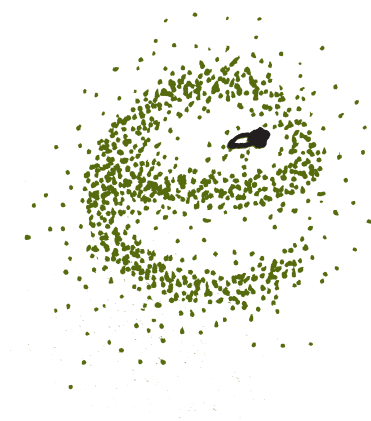
G. Box, „Science and statistics”, Journal
of the American Statistical Association
(1976).

Przemysław BIECEK*

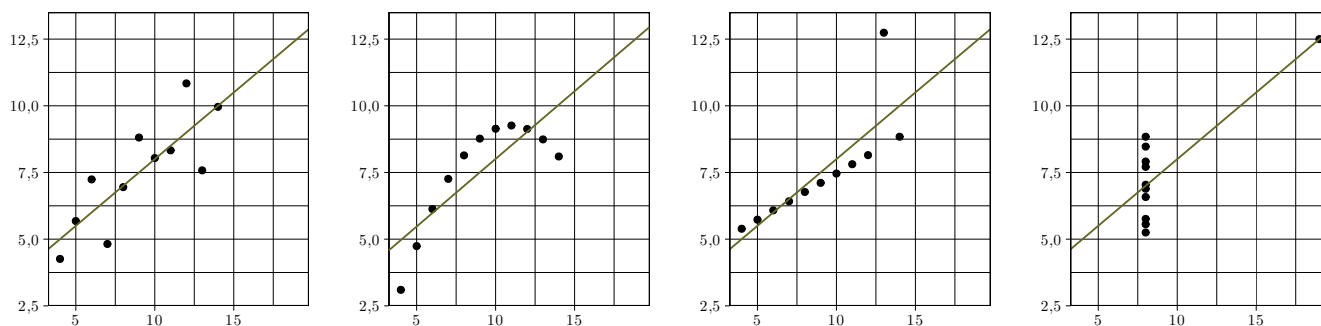
Jedno z najbardziej znanych powiedzeń związanych z modelowaniem statystycznym, ukute przez George'a Boxa, brzmi: „Wszystkie modele są błędne, ale niektóre są użyteczne”. To krótkie zdanie stanowi głęboką refleksję nad zastosowaniami statystyki. Z definicji modele stanowią uproszczony opis rzeczywistości. Gdy analizujemy złożone zjawiska (np. ekonomiczne, przyrodnicze, medyczne...), to niemożliwe jest stworzenie modelu uwzględniającego wszystkie istotne czynniki. Pomimo tego nawet niedoskonałe modele potrafią być przydatne. Używano ich przez dziesięciolecia, by formułować pytania i weryfikować hipotezy dotyczące otaczającego nas świata, takie jak: Czy analizowana terapia przynosi pozytywne efekty medyczne?, Czy inwestycje w edukację przekładają się na wyniki uczniów? To tylko dwa przykłady zagadnień badawczych, na które można odpowiedzieć za pomocą modeli statystycznych. Takie podejście jest podstawą wszystkich nauk empirycznych.

Nawet jeżeli zgodzimy się z powiedzeniem George'a Boxa, to pozostaje ważne pytanie: jak stwierdzić, które modele są przydatne? *Niektóre*, ale nie wszystkie, więc wybór właściwego modelu jest kluczowy dla poprawnego wnioskowania. Standardowe podejście do tego zagadnienia polega na ustaleniu pewnego kryterium jakości dopasowania do analizowanych danych, a następnie wybraniu modelu, który ma najlepszy wynik w sensie wybranego kryterium. Istnieje kilka popularnych kryteriów używanych przez statystyków, w tym: RMSE, R^2 , AIC, BIC (wybacz, Drogi Czytelniku, że nie rozwijam skrótów ani nie zdradzam technicznych szczegółów, jednak dla tego tekstu nie mają one istotnego znaczenia); kilka innych stosowanych jest przez osoby zajmujące się uczeniem maszynowym. Ogólna procedura pozostaje taka sama: zaczynamy od grupy kandydujących modeli, wybieramy najlepszy według określonego kryterium i uznajemy go za najlepszy opis rzeczywistości.

Takie podejście może prowadzić do ciekawych niespodzianek, jak również paradoksów. Jednym z nich jest opracowany 50 lat temu (dokładnie!) *kwartet Anscombe'a*. Francis Anscombe przedstawił cztery sztucznie utworzone zestawy danych, każdy z nich składał się z jedenastu par liczb (można sobie wyobrazić, że były to wzrost i waga jedenastu noworodków). Dane były tak skonstruowane, że dla każdego zestawu najlepiej dopasowana liniowa zależność (w sensie kryterium R^2) była taka sama (i dawała tę samą wartość R^2). Jednak każdy z tych zestawów danych opowiada zupełnie inną historię! Aby to odkryć, najlepiej przedstawić dane w sposób graficzny z użyciem wykresu kropkowego, jak na rysunku 1.



F. Anscombe, „Graphs in Statistical
Analysis”, American Statistician (1973).



Rys 1. Kwartet Anscombe'a. Dla każdego zbioru danych model $y = x/2 + 3$ ma najlepsze wśród modeli liniowych dopasowanie do danych ze współczynnikiem determinacji $R^2 = 0,66$

Nawet najlepiej dopasowany do danych model liniowy może prowadzić do błędnych wniosków dotyczących występujących w nich zależności. Analiza wizualna może w takich przypadkach uzupełniać wnioskowanie statystyczne, dlatego też wielu znanych statystyków przez dekady proponowało nowe metody wizualizacji danych, które dziś określa się mianem narzędzi *eksploracyjnej analizy danych*.

J. Tukey, „Exploratory Data Analysis”,
Pearson (1977).

L. Breiman, „Statistical Modeling: The Two Cultures”, *Statistical Science* (2001).

Anscombe pokazał, że różne zestawy danych opisujące różne historie mogą dawać ten sam model. Czy jednak może być odwrotnie? Czy jeden zbiór danych może prowadzić do kilku różnych modeli przedstawiających różne historie, a jednak tak samo dopasowanych do danych? Twierdzącej odpowiedzi na to pytanie udzielił Leo Breiman w opiniotwórczym artykule „The Two Cultures”. Opisane przez niego zjawisko znane jest dzisiaj pod nazwą „perspektywa Rashomona” lub „mnogość dobrych modeli”. Nazwa „Rashomon” odnosi się do filmu Akiry Kurosawy z 1950 roku, opisującego pewne wydarzenie z perspektywy czterech świadków, z których każdy przedstawia inną relację tego, co się wydarzyło. Relacje są tak różne, że nie sposób odgadnąć, jak było naprawdę. Breiman użył tego terminu, aby opisać hipotetyczny scenariusz, w którym kilka modeli ma równie dobre dopasowanie do danych, ale modele w różny sposób je *objaśniają*. Taka sytuacja stawia pod znakiem zapytania wszelkie wnioski oparte na „jednym najlepszym” modelu. Na przykład, co należy zrobić w obliczu dwóch modeli, które skutkują różnymi wnioskami co do skuteczności danej terapii medycznej? Któremu z tych modeli należy zaufać, jeśli oba są równie dobrze dopasowane do danych?

P. Biecek, H. Baniecki, M. Krzyżiński, and D. Cook, „Performance is not enough: a story of the Rashomon’s quartet” arxiv (2023).

Aby zilustrować problem mnogości dobrych modeli, Breiman przedstawił przykład kilku modeli liniowych prowadzących do różnych wniosków na temat danych, które jednocześnie były do nich w tym samym stopniu dopasowane, a stopień tego dopasowania różnił się nieznacznie od najlepszego możliwego dopasowania. Aby uczynić to zjawisko jeszcze bardziej wyraźnym, skonstruowaliśmy niedawno *kwartet Rashomona*. W naszym artykule przedstawiamy drzewo regresji, las losowy, sieć neuronową i model liniowy (szczegóły dotyczące tych modeli nie są w tym tekście dla nas istotne). Wszystkie zostały, w miarę numerycznych możliwości, najlepiej dopasowane do tych samych danych i miały takie samo dopasowanie, a jednocześnie opisywały *całkowicie różne historie*.

J. Friedman, „Greedy Function Approximation: A Gradient Boosting Machine”, *Annals of Statistics* (2000).

... Chwileczkę, ale skąd mamy wiedzieć, jakie historie pokazują tak złożone modele jak sieć neuronowa czy las losowy złożony z setek drzew? Z pomocą przychodzą nam techniki wizualizacji rozwijane pod nazwą eXplainable Artificial Intelligence (XAI) lub Explanatory Model Analysis (EMA). Jedną z nich jest analiza zależności częściowej modelu (Partial Dependence, PD), technika zaproponowana przez Jerome’a Friedmana w jego słynnej pracy o metodzie *boosting*. Profile PD można wyznaczyć dla modeli o dowolnej strukturze. Ze względu na tę uniwersalność metoda ta szybko znalazła wiele zastosowań.

P. Biecek, T. Burzykowski, „Explanatory Model Analysis”, CRC Press (2021). <https://ema.drwhy.ai/>

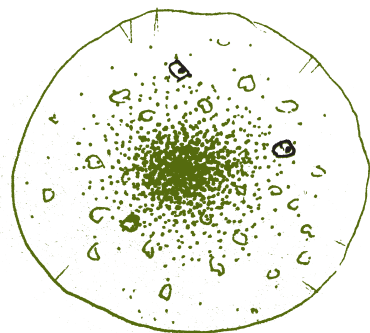
Przedstawmy intuicję stojącą za profilem zależności częściowej. Wartość PD dla wybranej zmiennej S i jej potencjalnej wartości t to średnie przewidywanie modelu dla dostępnych danych, w których wartość tej zmiennej została „na siłę” ustalona wszędzie jako t .

Aby opisać rzecz bardziej formalnie, załóżmy, że mamy N obserwacji p zmiennych wejściowych (np. dla N noworodków rejestrujemy wzrost, wagę, kolor oczu...), gdzie i -ta obserwacja to $(x_{i,1}, \dots, x_{i,p})$. Na podstawie tych zmiennych chcemy coś przewidywać (to coś zwykle się nazywa *zmienną docelową*) i w tym celu używamy funkcji $f(x_1, \dots, x_p)$, która jest naszym modelem. Zależność częściowa s -tej zmiennej to funkcja określona wzorem

$$PD^s(t) = \frac{1}{N} \sum_{i=1}^N f(x_{i,1}, \dots, x_{i,s-1}, t, x_{i,s+1}, \dots, x_{i,p}).$$

Profil PD mówi nam zatem, jak zmienia się przewidywana wartość zmiennej docelowej, gdy zmienia się wartość interesującej nas zmiennej wejściowej, przy jednoczesnym utrzymaniu wszystkich innych zmiennych wejściowych na wartościach obserwowanych w analizowanych danych.

Profile PD każdego modelu w kwartecie Rashomona przedstawiono na rysunku 2, sugerując wyraźne i różne zależności między zmiennymi, szczególnie dla zmiennych X_2 i X_3 . Na przykład zmienna X_2 jest nieistotna w pierwszym modelu (tzn. średnio jej zmiana nie wpływa na przewidywania modelu), ale ma dodatni efekt w pozostałych modelach (tzn. średnio jej zwiększenie powoduje wzrost przewidywań modelu). Z kolei zmienna X_3 jest nieistotna w pierwszym modelu, ma ujemny efekt w drugim modelu i dodatni w trzecim i czwartym. Jednak skoro

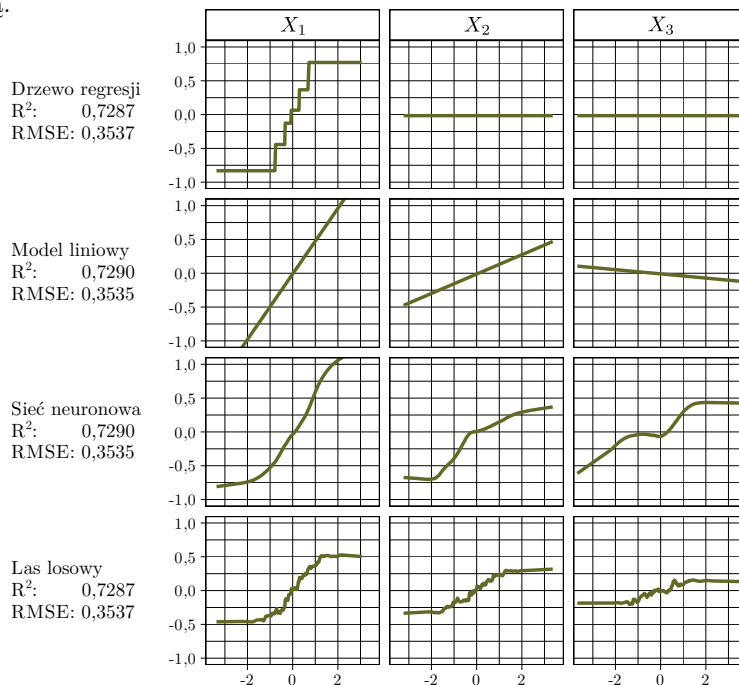




Rys. 2. Kwartet Rashomona. Każdy rząd opisuje inny model, a kolumny kolejne zmienne wejściowe. Panele pokazują profile zależności częściowej. Te cztery modele zostały dopasowane do jednego i tego samego zbioru danych. Przedstawione współczynniki dopasowania R^2 i RMSE pokazują, że wszystkie te modele są równie dobrze dopasowane do analizowanych danych

wszystkie modele mają takie samo dopasowanie do danych, to jak określić, który z tych opisów jest poprawny?

Któremu modelowi powinniśmy zaufać? A jeśli nie wiemy, to dlaczego mamy ufać któremukolwiek z nich? Ponieważ ze względu na losowe fluktuacje każdy z tych modeli potencjalnie mógłby zostać uznany za najlepszy, analizowanie tylko jednego, najlepszego modelu z pominięciem tych nieco gorszych może być złą strategią.



C. Rudin, Ch. Chen, Z. Chen, H. Huang, L. Semenova, Ch. Zhong, „Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges”, arxiv (2021).

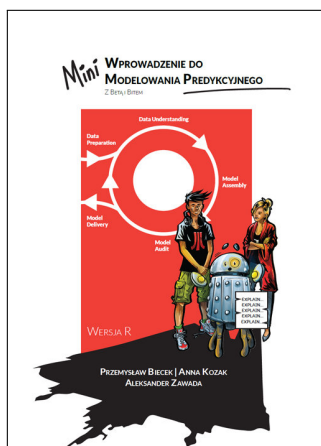
Profile PD mówią wiele o modelach, ale nie są jedynym narzędziem ich wizualizacji. Mimo swojej uniwersalności mają pewne ograniczenia, np. korelacje między zmiennymi lub interakcje mogą zniekształcić prezentowany obraz. Co więcej, analiza wpływu każdej zmiennej będzie trudna, jeśli model wykorzystuje setki lub tysiące zmiennych. Charakteryzacja zbioru najlepszych modeli jest wciąż otwartym problemem badawczym. Aby zrozumieć świat, nie możemy polegać wyłącznie na perspektywie jednego modelu, nawet jeśli ma on najlepsze wyniki w odpowiednich kryteriach. Musimy zestawiać ze sobą perspektywy różnych modeli, po to aby odróżnić hipotezy poparte danymi od hipotez, które są artefaktem wybranej struktury modelu.

Narzędzia do pełnego rozpoznawania sytuacji, w których możemy mieć do czynienia z perspektywą Rashomona, nie zostały jeszcze opracowane. Do czasu, aż to nastąpi, pozostaje nam eksploracja tych modeli z wykorzystaniem szeregu nowych technik wizualizacji. Kwartet Anscombe’a pokazał, że techniki wizualizacji danych są niezbędne do wnioskowania o naturze relacji między zmiennymi. Podobnie kwartet Rashomona pokazuje, że techniki wizualizacji modeli są równie przydatne.

Czytelniku, jeśli chcesz poznać kolejne metody wizualizacji i eksploracji modeli, sięgnij po komiks statystyczny „Mini Wprowadzenie do Modelowania Predykcynego” dostępny na stronie <https://betaandbit.github.io/MiniML/>.

Na przykładzie predykcji śmiertelności wirusa SARS-COV-2 omawia on proces trenowania modeli predykcynych, a także metody eksploracji tych modeli z wykorzystaniem profili zależności częściowej, wartości Shapleya czy profili Ceteris Paribus. Przedstawione powyżej przykłady oparte były na sztucznie wygenerowanych danych, jednak podobne wyzwania można napotkać, analizując problemy ze świata rzeczywistego.

Podsumowując, wszystkie modele są błędne, a my nie możemy być pewni, które z nich są przydatne. Możemy jednak przyjrzeć się wielu modelom jednocześnie, a spojrzenie na świat przez taką perspektywę jest niezbędne do oddzielenia zależności prawdziwie wspieranych przez dane od takich relacji, które są artefaktami wybranej techniki modelowania.



Chłopiec czy dziewczynka?

* Wydział Matematyki, Informatyki i Mechaniki, Uniwersytet Warszawski

Łukasz RAJKOWSKI*

Wyobraź sobie, Drogi Czytelniku, że wybierasz się w daleką podróż pociągiem. Kiedy już wygodnie rozsiadłeś się na swoim miejscu i zacząłeś czytać ulubione czasopismo, lektura została przerwana przez rubaszny okrzyk: „Hej, kogo ja widzę! Kopę lat, co?”. Podnosisz głowę i Twoim oczom ukazują się przyjacielska twarz, w której przy odrobinie wysiłku pamięciowego rozpoznajesz swojego dawno niewidzianego znajomego z podstawówki.

„Taaak, kopę lat. . .” – odpowiadasz, niechętnie zamykając świeżutki numer *Delty*.

„Co u Ciebie słychać?” – kurtuazyjnie dopytuje znajomy.

Cóż, następuje moment, w którym należy w jednym zdaniu streścić kilkanaście lat życia, w związku z czym odpowiadasz: „Dobrze! A co u Ciebie?”. Znajomy otwiera usta i już wiesz, że pytanie było błędem – on **faktycznie** chce Ci streścić ostatnich kilkanaście lat swojego życia, zdecydowanie nie w jednym zdaniu. Po jakiejś godzinie opowiadania o szkolno-zawodowych perypetiach w końcu przeszło na sprawy rodzinne. W pewnym momencie ograniczasz się do uprzejmego kiwania głową, uśmiechania się i niezawodnego „No tak. . .”, a tak naprawdę w swojej głowie rozwiązujesz deltowe Zadania z Myszką [w tym numerze na stronie 19 – przyp. red.]. W którymś momencie orientujesz się, że znajomy zaczął opowiadać o trudnościach w znalezieniu opieki dla swojego dziecka. Problem w tym, że umknął Ci moment, w którym wspominał, że je w ogóle ma.

Hm – myślisz sobie – nie usłyszałem, czy to chłopiec, czy dziewczynka. Lepiej w takim razie w przyszłości wypowiadać się o nim/niej bezpłciowo; inaczej mam około 50% szans na niezręczną pomyłkę.

W dalszej rozmowie okazuje się, że znajomy poszukiwał opieki nie dla jednego dziecka, a dla dwójki! Ponadto wychwyciłeś w monologu zdanie „Poszedłem z synkiem na spacer”, co ustala płć jednego z tych dzieci. Płć drugiego pozostaje dla Ciebie tajemnicą.

Ha, czytałem kiedyś coś o tym w Internecie! – pomyślałeś – W pierwszej chwili wydawałoby się, że szansa na to, że drugie dziecko to też chłopiec, jest równa 50%. Ale można na problem spojrzeć inaczej. Jeśli zapomnę o informacjach, które już posiadam, to szansa na to, że starsze dziecko mojego znajomego jest chłopcem, to 50%; podobnie szansa na to, że młodsze jest chłopcem, to 50%, a zatem szansa na dwóch chłopców to $50\% \cdot 50\% = 25\%$. Szansa na dwie dziewczynki to również 25%. Pozostały jeszcze dwie możliwości, każda z prawdopodobieństwem 25%: starsze to chłopiec, młodsze to dziewczynka i odwrotnie. Wiem już, że znajomy ma synka, a więc z tych czterech równoprawdopodobnych sytuacji mogę ograniczyć się do trzech (wykluczam dwie córeczki). A z tych trzech tylko jedna odpowiada sytuacji, że mój znajomy ma dwóch synków, a więc szansa na to jest równa $1/3$. Przyparty do muru powinienem zatem obstawić córeczkę!

Aby upewnić się o słuszności swojego rozumowania, dyskretnie wyciągnąłeś kawałek papieru oraz ołówek i naszkicowałeś przekonującą tabelkę 2×2 , której pola odpowiadały różnym kombinacjom płci starszego i młodszego dziecka (reprodukujemy ją na rys. 1). Tak jak wcześniej pomyślałeś, wszystko się zga. . .

(głos z offu) – NIE! Nic się nie zgadza! Tak naprawdę masz do czynienia z sytuacją, w której losowo napotkana osoba (co z tego, że to Twój znajomy) okazuje się mieć dwójkę dzieci oraz zaczyna Ci opowiadać o jednym z nich, które okazuje się chłopcem. Rozsądnie jest przyjąć, że Twój znajomy losowo wybrał jedno ze swoich dzieci (każde z równym prawdopodobieństwem). Jeżeli znajomy jest ojcem dwóch chłopców, to na pewno opowie o chłopcu. Jeżeli jest ojcem chłopca i dziewczynki, to ma szansę 50% na to, by opowiadać o chłopcu. Ujmując rzecz nieco inaczej, ze wszystkich ojców dwojga dzieci obojga płci tylko połowa zaczęłaby opowiadać Ci o swoim synku. Twoja tabelka powinna zatem odpowiadać tej na rysunku 2. Ostatecznie z Twojej perspektywy szanse na to, że znajomy ma dwóch synków, to niezmiennie 50%!



Podobnym trudnościom poświęcony jest utwór Łony i Webbera *Nie mam pojęcia*.

	Ch	Dz
Ch		
Dz		

Rys. 1. Wiersze odpowiadają płci starszego dziecka, a kolumny – młodszego. Zakreskowany obszar stanowi $1/3$ obszaru zamalowanego

	Ch	Dz
Ch	/	opowiada o chłopcu
Dz	opowiada o chłopcu	

Rys. 2. Zakreskowany obszar stanowi $\frac{1}{2}$ obszaru zamalowanego

	Ch							Dz						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Ch	1	2	3	4	5	6	7							
Dz														

Rys. 3. Liczby odpowiadają dniu tygodnia, w którym urodziło się dane dziecko. Zakreskowany obszar stanowi $\frac{13}{27}$ obszaru zamalowanego

	Ch							Dz						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Ch	1	2	3	4	5	6	7							
Dz														

Rys. 4. Zakreskowany obszar stanowi $\frac{1}{3}$ obszaru zamalowanego

Co innego, gdybyś sam poprosił znajomego: *Drogi kolego, proszę, zacznij mi opowiadać o jednym ze swoich synków, jeśli masz co najmniej jednego*. Zakładając, że wtedy znajomy istotnie rozpocznie jakąś opowieść (i nie postanowi zmienić miejsca siedzącego), przedstawione wcześniej rozumowanie oraz tabelka z rysunku 1 są słuszne i szansa na to, że ma on dwóch synków, faktycznie jest równa $\frac{1}{3}$. **Na potrzeby dalszej części historii założmy, że tak właśnie było.**

Monolog znajomego trwa. Znowu zająłeś się czymś innym. W którymś momencie do Twojej świadomości przebija się informacja, że wspomniany wcześniej synek urodził się w poniedziałek i od tamtego czasu poniedziałek jest ulubionym dniem tygodnia Twojego znajomego (szczęściarz!). Wiadomość ta skłoniła Cię do pewnej matematycznej refleksji.

Czy informacja, którą właśnie uzyskałem, powinna zmienić moje oszacowanie szans na to, że mój znajomy ma dwóch synków? Wydawać by się mogło, że w żadnym wypadku – przecież dzień tygodnia, w którym urodził się wspomniany przez niego synek, nie może wpływać w jakikolwiek sposób na płeć pozostałego potomstwa. Jednak z drugiej strony ta wiedza sprawia, że powinienem zmodyfikować naszkicowaną wcześniej tabelkę. Teraz dla każdego dziecka, starszego i młodszego, powinienem odnotować nie tylko jego płeć, lecz również dzień tygodnia, w którym się urodziło. Zakładając, że każdy dzień tygodnia jest równoprawdopodobnym dniem urodzin, daje mi to 14 równoprawdopodobnych konfiguracji par płeć i dzień tygodnia dla każdego z dzieci, czyli łącznie $14^2 = 196$ konfiguracji dla obojga dzieci. Tym razem 27 z nich odpowiada występowaniu chłopca urodzonego w poniedziałek, a z tego w 13 przypadkach pozostałe dziecko też jest chłopcem. Tak więc niniejszym moje oszacowanie prawdopodobieństwa tego, że znajomy ma dwóch chłopców, powinno się zwiększyć z $\frac{1}{3}$ do $\frac{13}{27}$.

Ponieważ sytuacja Cię frapowała, przerysowałeś swoją tabelkę, uzyskując coś zbliżonego do rysunku 3 na marginesie. I chociaż z doświadczenia wiesz, że z tabelkami się nie dyskutuje, konkluzje, do których doszedłeś, wciąż Cię drażniły.

(głos z offu) – I słusznie, gdyż takie rozumowanie jest powieleniem opisanego wcześniej błędu. Tym razem należałoby uwzględnić fakt, że jeśli Twój znajomy ma dwóch chłopców, z których dokładnie jeden urodził się w poniedziałek, to zaczął opowiadać właśnie o nim, a mógł o „tym drugim” – co dwukrotnie zmniejsza prawdopodobieństwo takiej konfiguracji z Twojej perspektywy. Uwzględnienie tego w rozumowaniu kończy się odpowiedzią: $\frac{1}{3}$ (taką jak wcześniej), o czym można się przekonać, analizując rysunek 4. Znowu – gdybyś to Ty dopytał, czy znajomy ma synka urodzonego w poniedziałek, to odpowiedź twierdząca pozwoliłaby zmienić oszacowanie szans na dwóch chłopców do $\frac{13}{27}$, zgodnie z przedstawionym wcześniej rozumowaniem (a odpowiedź przecząca kazałaby zmniejszyć to oszacowanie do $\frac{3}{10}$, o czym polecamy przekonać się samodzielnie).

Po tej informacji temat dzieci się urwał. Niebawem, po wysłuchaniu jeszcze kilku przejmujących historii z życia Twojego znajomego, pociąg dojechał na miejsce – nie zaskoczyło Cię, gdy wyszło na jaw, iż obaj wysiadacie na tej samej stacji. Na peronie czekała już żona znajomego wraz z rozczochranym urwisem.

Pytanie 1. Jak teraz oceniasz szanse na to, że znajomy ma dwóch synków?

I już zaczęliście się żegnać, gdy za plecami Twojego szkolnego znajomego pojawił się nie kto inny, jak... Twoja niewidziana od dawien dawna koleżanka z przedszkola! Zanim zdążyła cokolwiek powiedzieć, krzyknęłaś do niej:

„Hej tam, nie widzieliśmy się całe wieki! Czy przypadkiem nie masz dwójki dzieci, z których co najmniej jedno jest chłopcem urodzonym w poniedziałek?”

Twoja koleżanka zaniemówiła, a po chwili uśmiechnęła się chytrze i odpowiedziała:

„Owszem, mam dwójkę dzieci, w tym co najmniej jednego chłopca, przy czym żadne z moich dzieci nie urodziło się w poniedziałek”.

Pytanie 2. Jak oceniasz szanse na to, że koleżanka ma dwóch synków?

(Odpowiedzi na powyższe pytania można znaleźć na stronie 14).

Nowe odpowiedzi, nowe pytania

Jedenaście lat temu na lipcowej konferencji prasowej w CERN-ie ogłoszono odkrycie bozonu Higgsa. Błogosławieństwo wieku średniego pozwala mi być zarazem dość starym, by rozumieć wówczas znaczenie tego odkrycia, i wystarczająco młodym, by pamiętać towarzyszące mu emocje. Jednak z ważnymi odkryciami cząstek elementarnych jest trochę tak jak z filmami kryminalnymi – publiczności pokazuje się (czasem efektowne) morderstwo, po czym do akcji już poza zasięgiem kamer i scenariusza wkraczają specjaliści, dokumentują zdarzenie i starają się lepiej zrozumieć, co się właściwie wydarzyło.

Takim właśnie badaniom poddawany był bozon Higgsa przez przeszło dekadę. Poszczególne wyniki, niekiedy dokumentowane na tych łamach, z coraz większą dokładnością potwierdzały, że doświadczalnie wyznaczone własności tej cząstki są zgodne z przewidywaniami Modelu Standardowego cząstek elementarnych. A fizycy teoretycy cierpliwie czekali na pojawienie się jakichś niezgodności, które mogłyby świadczyć o tym, że Model Standardowy trzeba rozszerzyć.

Badanie bozonu Higgsa nie jest łatwe. Jest on cząstką pozbawioną ładunku i oddziałującą jedynie słabo, co oznacza, że prawdopodobieństwo jego produkcji jest małe. Potrzeba naprawdę dużych energii i wielkich liczb zderzających się cząstek, aby zebrać jakąkolwiek statystycznie istotną próbkę. Nawet w tak wielkim zderzaczach jak LHC produkuje się zaledwie jedna taka cząstka na dwie sekundy działania maszyny. Nic zatem dziwnego, że wyznaczenie charakterystyk niektórych rzadkich procesów z udziałem bozonu Higgsa trwa całymi latami.

Przykładem takiego rzadkiego procesu jest rozpad bozonu Higgsa na bozon Z, masywną, neutralną cząstkę przenoszącą oddziaływania słabe, i znany z oddziaływań elektromagnetycznych foton. Model Standardowy przewiduje, że zaledwie w 0,15% przypadków rozpad bozonu Higgsa przebiega właśnie w ten sposób. Co gorsza, pojawienie się bozonu Z trzeba też jakoś wykryć. Nie jest to łatwe, bo jest on też cząstką nietrwałą, żyjącą zbyt krótko, by można ją było bezpośrednio zaobserwować. Identyfikuje się ją za pomocą produktów jej rozpadu – w parę elektron-pozyton lub mion-antymion, co zdarza się w przypadku 6,6% rozpadów. Oznacza to, że przy założeniu stuprocentowej wydajności detektorów – a jest to przecież bezsensownie optymistyczne założenie! – zaledwie jeden na dziesięć tysięcy rozpadających się bozonów Higgsa wyprodukuje interesujący nas sygnał, który moglibyśmy zaobserwować.

Znaczenie rozpadu bozonu Higgsa na bozon Z i foton polega na tym, że – zgodnie z Modelem Standardowym – nie następuje on bezpośrednio, tylko w pewnym sensie asystują mu inne cząstki obecne w przyrodzie, być może również te, które nie zostały jeszcze wykryte. Zatem, jeśli zmierzone prawdopodobieństwo tego rozpadu byłoby istotnie różne od przewidywań Modelu Standardowego, stanowiłoby to argument za tym, że takie nowe cząstki istnieją, a ich masy są zbliżone do skali energii osiągniętych przez cząstki zderzane w LHC.

Aby wyznaczyć doświadczalnie szukane prawdopodobieństwo, konieczne było połączenie sił zespołów fizyków przy detektorach ATLAS i CMS. Używając metod sztucznej inteligencji do analizy danych, badacze stwierdzili, że wynosi ono 0,34% z niepewnością pomiarową 0,11%. Z jednej strony może to wzbudzać pewien niepokój, bo wyznaczona wartość jest 2,2 raza większa od przewidywań, ale, z drugiej strony, niepewności pomiarowe są jeszcze na tyle duże, że zgodność z Modelem Standardowym nie jest jeszcze wykluczona, chociaż prawdopodobieństwo przypadkowej fluktuacji statystycznej takiego rzędu nie jest wcale bardzo duże i wynosi tylko 6%. Na rozwiązanie zagadki, co właściwie zostało zmierzone, przyjdzie nam zatem jeszcze zapewne poczekać kilka lat...

Krzysztof TURZYŃSKI

Wydział Fizyki, Uniwersytet Warszawski

Odpowiedzi na pytania ze strony 13:

Pytanie 1. W tej sytuacji należałoby uwzględnić to, że gdyby znajomy miał synka i córeczkę, to ich mama miałaby tylko 50% szans na przyprowadzenie synka. W związku z czym tabelkę z rysunku 4 w artykule należałoby zmodyfikować poprzez dwukrotne zwięźlenie zamalowanych, niezakreskowanych prostokątów. Daje to odpowiedź 50%.

Pytanie 2. Interpretacja tej sytuacji wydaje się do pewnego stopnia uznaniowa, jednak w opinii autora artykułu należałoby przyjąć, że znajoma nie odpowiedziała na nasze pytanie (*Czy co najmniej jedno dziecko jest chłopcem urodzonym w poniedziałek?*), lecz na dwa osobne: *Czy masz co najmniej jednego synka?* i *Czy masz co najmniej jedno dziecko urodzone w poniedziałek?* Gdybyśmy faktycznie takie jej zadali (w jakimś sensie to zrobiliśmy, gdyż to my wybraliśmy *poniedziałek*), to szansę na dwóch chłopców szacowalibyśmy jako $\frac{1}{3}$ (warto wyobrazić sobie graficzne przedstawienie tej sytuacji na rysunku 3 z artykułu).

The ATLAS and CMS Collaborations. 2023. Evidence for the Higgs boson decay to a Z boson and a photon at the LHC. ATLAS-CONF-2023-025

Poszukiwanie porządku

Andrzej DĄBROWSKI*

*Wydział Matematyki i Informatyki,
Uniwersytet Wrocławski

Pełna nazwa to *Seminarium z Matematyki Stosowanej Sekcji Ogólnej Wydziału Zastosowań Państwowego Instytutu Matematycznego* [1].

Czaszki z Ngandong są grupą skamieniałości ludzkich czaszek odkrytych w 1931 roku w dolinie rzeki Solo na Jawie w Indonezji. Czaszki te są datowane na około 100-200 tysięcy lat i są uważane za pozostałości hominidów z gatunku *Homo erectus*.

Jan Czekanowski (1882–1965) był antropologiem i etnografem. Studiował na Uniwersytecie w Zurychu antropologię, anatomię, etnografię i matematykę. Był uczestnikiem *Seminarium* we Wrocławiu (czterokrotnie, według kroniki *Seminarium*).

W roku 1950 jedno z jesiennych posiedzeń wrocławskiego *Seminarium z Matematyki Stosowanej* poświęcone było czaszkom z Ngandong. Antropolodzy oczekiwali od matematyków uporządkowania chronologicznego tych czaszek na podstawie tablic Czekanowskiego *przysłanych nam łaskawie przez prof. dra T. Henzla* (cytat z kroniki *Seminarium*).

Porządek, rozumiany jako liniowy, jest oczekiwany w wielu dziedzinach. Tworzy się różnego rodzaju rankingi (dotyczące szkół, uczelni, poparcia polityków). Trudno o widowiska sportowe bez rankingów – opartych na mierzonym czasie, dystansie czy punktach (pięciobój). Aby porównać często trudne do zestawienia obiekty, buduje się indeksy. W ekonomii stany giełdy ocenia się przez indeksy giełdowe (np. Dow Jones, WIG), a poziom cen – przez wskaźnik inflacji. Stan zdrowia populacji ocenia się przez oczekiwany średni czas życia czy wskaźnik śmiertelności niemowląt. Metody porównania oparte na rankingach i indeksach polegają na przypisaniu obiektom jakiejś liczby. A liczby, jak wiadomo, łatwo uporządkować.

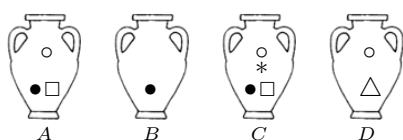
Gorzej, kiedy obiekt można opisać zespołem wielu liczb. Jeszcze gorzej, gdy trudno przypisać mu jakiegokolwiek liczby. Taki problem powstaje, gdy trzeba uporządkować znalezione na starożytnym śmietniku skorupy z różnymi motywami (jest to tzw. zagadnienie seriacji w archeologii). Z podobnym problemem – uporządkowania w czasie czaszek z Ngandong – zwrócili się antropolodzy do matematyków.

Tablice Czekanowskiego, którymi dysponował zespół profesora Steinhausa (Kazimierz Florek, Józef Łukaszewicz, Julian Perkal i Stefan Zubrzycki), zawierały informacje o stopniu zróżnicowania każdej pary czaszek. Stopień ten był wyrażony przez odległość euklidesową w siedmiowymiarowej przestrzeni parametrów reprezentujących długości charakterystycznych odcinków na czaszce.

Uogólniając, taką tablicę można potraktować jako *funkcję rozbieżności*. Jest to dowolna funkcja $d(x, y)$ określona na parach elementów zadanego zbioru X , która spełnia warunki:

$$d(x, y) \geq 0, \quad d(x, y) = d(y, x) \text{ oraz } d(x, y) = 0 \Leftrightarrow x = y.$$

Szczególnym przypadkiem funkcji rozbieżności są funkcje odległości – metryki. Metryka to funkcja rozbieżności, która spełnia dodatkowo nierówność trójkąta: $d(x, z) \leq d(x, y) + d(y, z)$ dla każdych x, y, z .



Rys. 1

Przykład 1. Wazy. Odkryto 4 wazy z ozdobami. Każdą wazę traktujemy jako zbiór namalowanych na niej ornamentów. Funkcją rozbieżności może być tzw. rozbieżność Kulczyńskiego określona dla dwóch zbiorów ornamentów R i S wzorem

$$d(R, S) = 1 - \frac{1}{2} \left(\frac{|R \cap S|}{|R|} + \frac{|R \cap S|}{|S|} \right).$$

Nietrudno sprawdzić, że taka funkcja faktycznie jest funkcją rozbieżności. Jest ona też dość intuicyjna, gdyż $\frac{1}{2} \left(\frac{|R \cap S|}{|R|} + \frac{|R \cap S|}{|S|} \right)$ to średnia arytmetyczna frakcji wspólnych elementów R i S , zawartych w zbiorze R i zawartych w zbiorze S .

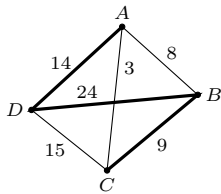
Wynikająca z tego wzoru macierz rozbieżności przedstawiona jest na marginesie. Zauważmy, że funkcja Kulczyńskiego nie jest metryką, gdyż $d(B, D) = 1 > \frac{11}{12} = d(B, A) + d(A, D)$.

Stanisław Kulczyński (1895–1975), zoolog, arachnolog, tatarnik. Rektor uniwersytetów we Lwowie i Wrocławiu.

	A	B	C	D
A	0	$\frac{1}{3}$	$\frac{1}{8}$	$\frac{7}{12}$
B	$\frac{1}{3}$	0	$\frac{3}{8}$	1
C	$\frac{1}{8}$	$\frac{3}{8}$	0	$\frac{5}{8}$
D	$\frac{7}{12}$	1	$\frac{5}{8}$	0

Macierz rozbieżności Kulczyńskiego

Funkcja rozbieżności a porządek liniowy. Jak wprowadzić porządek wśród obiektów, na których mamy zadaną macierz rozbieżności? Rozważmy najpierw szczególny przypadek. Załóżmy, że obiekty to punkty rozmieszczone na prostej, a za funkcję rozbieżności przyjmujemy odległości między nimi. Czy z samej macierzy rozbieżności można odtworzyć uporządkowanie punktów na prostej? Można – wystarczy wybrać takie uporządkowanie, które minimalizuje sumę odległości między kolejnymi punktami.

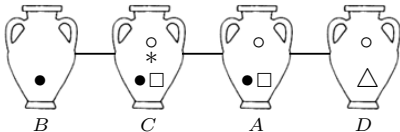


Rys. 2. Graf z rozbieżnościami Kulczyńskiego (pomnożonymi przez 24) ze ścieżką $ADBC$ o długości 47

Dla wygody przyjmijmy, że obiekty O_1, O_2, \dots, O_n z rozbieżnościami $d_{ij} = d(O_i, O_j)$ będziemy reprezentować przez nieskierowany graf pełny G o n wierzchołkach O_1, O_2, \dots, O_n , którego krawędź $O_i O_j$ ma wagę d_{ij} . Długością ścieżki $O_{i_1} O_{i_2} \dots O_{i_n}$ nazwiemy liczbę $d_{i_1 i_2} + d_{i_2 i_3} + \dots + d_{i_{n-1} i_n}$. Na rysunku 2 widać przykład takiego grafu dla 4 waz.

Wcześniejsza obserwacja sugeruje, aby porządek obiektów O_1, \dots, O_n wyznaczała ścieżka w grafie G o minimalnej długości, przechodząca przez wszystkie jego wierzchołki (ścieżka optymalna).

Przykład 2. Pięć stolic. Poniższa tablica przedstawia odległości drogowe (w kilometrach) między Amsterdamem, Berlinem, Paryżem, Rzymem i Warszawą.



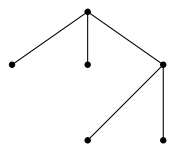
Rys. 3. Ścieżka optymalna $BCAD$ w grafie z wazami

	A	B	P	R	W
A	0	650	510	1650	1140
B	650	0	1040	1460	570
P	510	1040	0	1430	1550
R	1650	1460	1430	0	1730
W	1140	570	1550	1730	0

Spośród 60 możliwych ścieżek najkrótsza wiedzie z Warszawy przez Berlin, Amsterdam, Paryż aż do Rzymu i ma długość 3160 km.

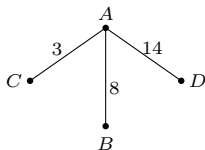
Dendryty. Uporządkowanie liniowe jest często niewystarczające, a nawet nieadekwatne. Przytoczmy tu obserwację Juliana Perkala: *Jak zauważyłem [...] liniowe porządkowanie jest w wielu przypadkach nienaturalne, na przykład linia genealogiczna właśnie często się rozgałęzia.* [2]

Strukturą, która umożliwia zbliżone do liniowego uporządkowanie wierzchołków grafu, jest dendryt, szerzej znany jako drzewo – z powodów historycznych będziemy tu używać pierwszego określenia. Dendryt to graf bez cykli, który jest spójny. Te dwa warunki razem oznaczają, że każde dwa wierzchołki są połączone jednoznacznie wyznaczoną ścieżką (rys. 4). Sama ścieżka też jest zresztą szczególnym przypadkiem dendrytu.



Rys. 4

Długość dendrytu określa suma wag jego krawędzi. Inspirowani wcześniejszymi rozważaniami przyjmijmy, że dendryt o minimalnej długości (dendryt optymalny) najlepiej odtworzy porządek n wierzchołków grafu (rys. 5).



Rys. 5. Optymalny dendryt dla problemu 4 waz. Jego długość 25 jest mniejsza od długości optymalnej ścieżki $BCAD$ o długości 26

Taksonomia wrocławska. W przykładzie z wazami nietrudno było wskazać optymalny dendryt. Wraz ze wzrostem liczby wierzchołków grafu powstaje pytanie o algorytmiczne jego wyznaczanie. W informatyce zagadnienie to jest klasyczne i znane pod nazwą poszukiwania *minimalnego drzewa rozpinającego*. Według [3] najwcześniejsze opublikowane rozwiązanie tego problemu (1926) pochodzi od czeskiego matematyka Otakara Borůvki, który zajmował się nim w kontekście opracowywania optymalnej sieci elektrycznej na Morawach. Znane zapewne uczestnikom olimpiad informatycznych algorytmy Kruskala i Prima zostały opublikowane w latach 1956 i 1957. Wydaje się, że matematycy wrocławscy jako pierwsi zaczęli zajmować się tym problemem w kontekście metod porządkowania i klasyfikowania (tzn. *taksonomii*) w antropologii, biologii czy językoznawstwie. Ich metoda, opublikowana w 1951 roku [4], jest co do zasady algorytmem Borůvki (z którego istnienia nie zdawali sobie sprawy), ale w polskiej literaturze statystycznej nosi nazwę *taksonomii wrocławskiej*.

W ramach tych prac pierwszy krok w kierunku ogólnej konstrukcji optymalnego dendrytu zrobił w roku 1949 Kazimierz Florek. Zwrócił on uwagę na fakt, że w optymalnym dendrycie powinny się znaleźć odcinki łączące najbliższe obiekty – to znaczy łączące pewien obiekt z jego najbliższym sąsiadem. Odcinki te nazwano *połączeniami rzędu I*.

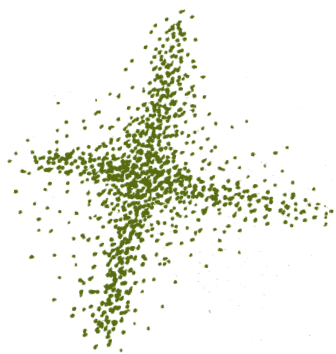


Rys. 6

W przykładzie z wazami takie postępowanie rozwiązuje sprawę – połączenia rzędu I tworzą dendryt. Nie zawsze tak jest. Połączenia rzędu I dla 5 stolic tworzą niespójny podgraf, przedstawiony na marginesie. Zauważmy jednak, że

w grafie połączeń rzędu I nie mogą występować cykle – taki graf nazywamy *lasem* i w naturalny sposób rozbija się on na dendryty.

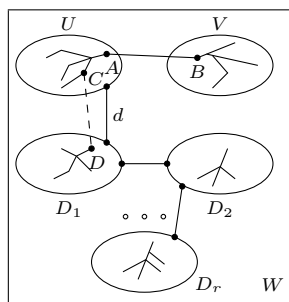
Twórcy taksonomii zaproponowali w 1950 roku konstrukcję optymalnego dendrytu, będącą iteracyjną wersją pomysłu Florka:



Metoda W

1. Zbudować połączenia rzędu I dla grafu G . Jeśli tworzą dendryt, konstrukcja jest skończona. Jeśli nie, przechodzi się do punktu 2.
2. Utworzyć nowy graf G_1 . Wierzchołkami grafu G_1 są poddendryty utworzone z połączeń rzędu I grafu G . Rozbieżność między poddendrytami A i B jest rozbieżnością najbliższych sąsiadów: $d(A, B) := \min\{d(P, Q) : P \in A, Q \in B\}$. Utworzyć połączenia rzędu I dla grafu G_1 (są to połączenia rzędu II dla G). Jeśli tworzą dendryt, konstrukcja jest skończona. Jeśli nie, powtarza się postępowanie z punktu 2 dla grafu G_1 .
3. W ten sposób utworzymy kolejno grafy G_1, G_2, G_3, \dots i dla każdego z nich połączenia kolejnego rzędu. Iteracje te muszą się skończyć, gdyż liczba wierzchołków w kolejnych grafach G_i maleje.
4. Konstrukcję ostatecznego dendrytu kończy się, łącząc dendryty kolejnych rzędów krawędziami między obiektami realizującymi rozbieżność najbliższego sąsiada.

Zobaczmy, jak działa metoda W dla przykładu 5 stolic. Połączenia I rzędu tworzą poddendryty $O_1 = \{A, P, R\}$ i $O_2 = \{B, W\}$. Graf G_1 ma wierzchołki O_1 i O_2 . Minimum odległości z B do zbioru punktów $\{A, P, R\}$ wynosi 650 km, minimum z W do $\{A, P, R\}$ – 1140 km, zatem rozbieżność najbliższego sąsiada między O_1 i O_2 jest równa 650 km, i jest odległością między Berlinem a Amsterdamem. G_1 z połączeniem rzędu II między tymi stolicami jest już dendrytem. Kończy to konstrukcję, tworząc ostateczny optymalny dendryt, który okazał się ścieżką $WBAPR$.



Rys. 7

Dowód optymalności metody W. Bez straty ogólności można założyć, że wszystkie niezerowe rozbieżności w grafie G są różne. W razie potrzeby wystarczy bowiem zamienić rozbieżności d_{ij} na $d'_{ij} = d_{ij} + \varepsilon_{ij}$ (dla $i < j$), $d'_{ji} = d'_{ij}$ i $\varepsilon_{ij} > 0$ są to dowolnie wybrane różne liczby mniejsze niż $\min d_{ij}$. Po zakończeniu konstrukcji należy powrócić do rozbieżności d_{ij} .

Niech W będzie dendrytem sporządzonym metodą W . Przypuśćmy, że W nie jest dendrytem optymalnym, więc jest różny od pewnego dendrytu optymalnego M . W dendrycie W istnieje zatem krawędź, która nie występuje w dendrycie M – powiedzmy, że łączy wierzchołki A i B .

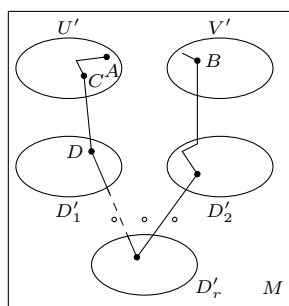
Przyjmijmy, że krawędź AB ma wagę a i rząd k w dendrycie W . Łączy ona dendryty U i V rzędu $k - 1$ o zbiorach wierzchołków U' i V' . Pozostałe dendryty rzędu $k - 1$, które oznaczymy przez D_1, D_2, \dots, D_r , mają zbiory wierzchołków D'_1, D'_2, \dots, D'_r . Zgodnie z konstrukcją W zachodzi jeden z dwóch przypadków:

1. a jest mniejsze niż każda z rozbieżności U z D_1, \dots, D_r ,
2. a jest mniejsze niż każda z rozbieżności V z D_1, \dots, D_r .

Bez straty ogólności przyjmijmy, że zachodzi przypadek 1.

Wierzchołki A i B dendrytu M łączy ścieżka niezawierająca krawędzi AB , która musi zawierać krawędź CD taką, że C należy do U' , a D nie należy. Przyjmijmy, że D należy do D'_1 . Waga krawędzi CD jest nie mniejsza niż odległość d między dendrytami U i D_1 w dendrycie W . Z kolei $a < d$, bo zachodzi przypadek 1. Można więc stworzyć nowy dendryt M' zawierający wszystkie krawędzie M – z wyjątkiem CD zastąpionej przez krawędź AB . Długość M' jest mniejsza od długości M . Jest to sprzeczne z założeniem, że M jest dendrytem optymalnym.

Podział na grupy. Zbiór obiektów może być niejednorodny: skorupy znalezione na starożytnym śmietniku czy czaszki znalezione na badanym terytorium mogą pochodzić z kilku odrębnych okresów. Jak podzielić dane, tak aby dendryty odpowiadające podziałowi wskazały na istotne różnice w tych grupach?



Rys. 8

Rodzina k dendrytów D_1, D_2, \dots, D_k o zbiorach wierzchołków Z_1, Z_2, \dots, Z_k jest *podziałem* grafu pełnego G o wierzchołkach Z , gdy

$$Z = Z_1 \cup Z_2 \cup \dots \cup Z_k \quad \text{oraz} \quad Z_i \cap Z_j = \emptyset \quad \text{dla} \quad i \neq j.$$

Długość podziału jest sumą długości $l(D_i)$ składowych dendrytów. *Podział jest optymalny*, gdy długość podziału jest minimalna. Z definicji tej natychmiast wynika, że składowe podziału optymalnego muszą być optymalne, możemy więc przyjąć $D_i = W(Z_i)$.

Wróćmy do przykładu czterech waz. Można je podzielić na dwie grupy na siedem sposobów:

Z_1	$\{A, B\}$	$\{A, C\}$	$\{A, D\}$	$\{A, B, C\}$	$\{A, B, D\}$	$\{A, C, D\}$	$\{B, C, D\}$
Z_2	$\{C, D\}$	$\{B, D\}$	$\{B, C\}$	$\{D\}$	$\{C\}$	$\{B\}$	$\{A\}$
$W(Z_1)$	$A-B$	$A-C$	$A-D$	$B-A-C$	$B-A-D$	$C-A-D$	$B-C-D$
$W(Z_2)$	$C-D$	$B-D$	$B-C$	D	C	B	A
$l(W(Z_1)) + l(W(Z_2))$	23	27	23	11	22	17	24

Optymalny podział $D_1 = C-A-B$, $D_2 = D$ jest podgrafem dendrytu uzyskanego metodą W (rys. 5). Okazuje się, że tak jest zawsze.

Twierdzenie. *Jeżeli $\{W(Z_1), W(Z_2), \dots, W(Z_k)\}$ jest optymalnym podziałem grafu G o wierzchołkach Z , to*

$$W(Z_1) \cup W(Z_2) \cup \dots \cup W(Z_k) \subset W(Z).$$

Dowód. Zakładamy, jak poprzednio, bez straty ogólności, że wszystkie niezerowe rozbieżności są różne. Przypuśćmy, że w $W(Z_1) \cup W(Z_2) \cup \dots \cup W(Z_k)$ istnieje krawędź AB nienależąca do $W(Z)$. Istnieje w $W(Z)$ ścieżka s_{AB} łącząca te punkty.

Z rozłączności Z_i wynika, że istnieje r takie, że krawędź AB należy do $W(Z_r)$. Elementy Z_r można podzielić na podzbiory U i V w następujący sposób: U składa się z wierzchołka A i wszystkich wierzchołków Z_r , które są połączone ścieżką w dendrycie $W(Z_r)$ wychodzącą z A i niezawierającą krawędzi AB ; V składa się z wierzchołka B i wszystkich wierzchołków Z_r , które są połączone ścieżką w dendrycie $W(Z_r)$ wychodzącą z B i niezawierającą krawędzi AB .

Ścieżka s_{AB} musi zawierać krawędź CD taką, że $C \in U$, $D \notin U$. Niech $d(A, B) = x$, $d(C, D) = y$. Zachodzi nierówność $y < x$, w przeciwnym przypadku, zamieniając CD z AB w dendrycie $W(Z)$, zmniejszylibyśmy jego długość, a to byłaby sprzeczność.

Zachodzą dwa przypadki: $D \in V$ i $D \in Z_s$ dla $s \neq r$.

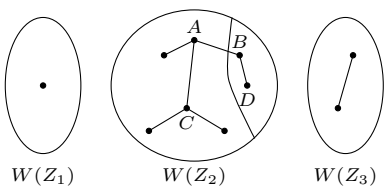
1. $D \in V$ (rys. 9). Wtedy, zastępując AB przez CD w $W(Z_r)$, uzyskujemy, że dendryt $W(Z_r)$ nie jest optymalny – sprzeczność.

2. $D \in Z_s$ dla $s \neq r$ (rys. 10). Zastępujemy dendryt $W(Z_r)$ przez dendryt $W(U) \cup CD \cup W(Z_s)$ o długości $l(W(U)) + y + l(W(Z_s))$, zaś dendryt $W(Z_s)$ przez $W(V)$ o długości $l(W(V))$. Suma długości k dendrytów po zmianie jest mniejsza niż suma ich długości przed zmianą, co przeczy założeniu, że podział jest optymalny. To kończy dowód.

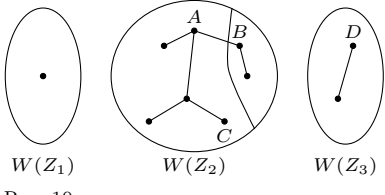
Z powyższego twierdzenia wynika użyteczny w kontekście szukania optymalnego grupowania

Wniosek. *Optymalny podział grafu G na k poddendrytów polega na usunięciu z dendrytu $W(Z)$ $k - 1$ krawędzi o największej rozbieżności.*

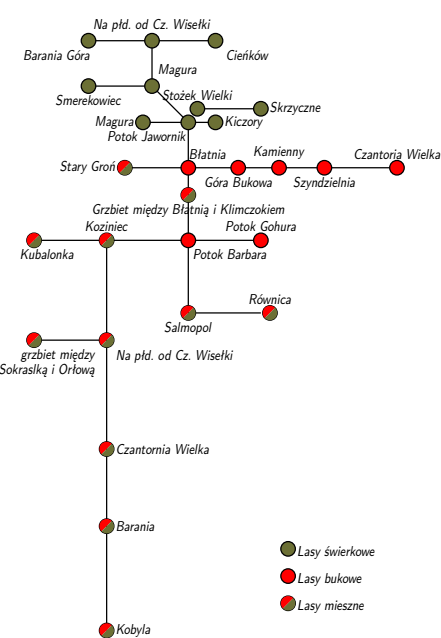
O zastosowaniach parę słów. *Seminarium z Matematyki Stosowanej* bardzo aktywnie propagowało pomysł taksonomii, stosując ją w różnych dziedzinach. W *Kalejdoskopie Matematycznym* Steinhausa można przeczytać o taksonomii wątrobowców (*Hepaticae*) w Beskidzie Śląskim. Dendryt odpowiadający częstości występowania różnych gatunków wątrobowców okazał się powiązany z typem lasu – co było ciekawym odkryciem (rys. 11).



Rys. 9



Rys. 10



Rys. 11

Literatura

- [1] Szczotka, W., *Fenomen Steinhausowskiego Seminarium Matematyki Stosowanej*, *Antiquitates Mathematicae*, (2018) vol. 12(1), 197–228, (2019) vol. 13(1), 177–233.
- [2] Perkal, J. (1953). *Taksonomia wrocławska*, *Przegląd Antropologiczny*.
- [3] Graham, R. L., Hell, P. (1985). *On the history of the minimum spanning tree problem*, *Annals of the History of Computing*, 7(1), 43–57.
- [4] Florek K., Łukaszewicz J., Perkal J., Steinhaus H., Zubrzycki S., *Sur la liaison et la division des points d'un ensemble fini*, *Colloquium Mathematicum* (1951), vol. 2, nr 3–4 str. 282–285.
- [5] Zubrzycki, F. (1953). *O łańcuszkach gwiazdnych*. *Applications Mathematicae*, 1(3), 197–205.

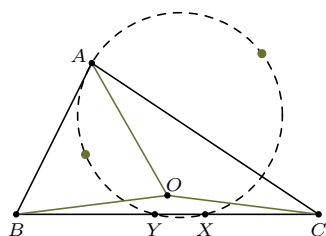
Charakterystycznym rysem pracy *Seminarium* było podejmowanie każdego, nawet najbardziej nietypowych problemów. Na posiedzeniu w styczniu 1952 roku Julian Perkal zakomunikował, że zrobił dendryt pieśni ludowych dla córki prof. Czekanowskiego (cytat ze starannie prowadzonej księgi protokołów *Seminarium*). Klasyfikacja pieśni ludowych metodą taksonomii wrocławskiej stała się jednym z ważnych narzędzi badawczych Anny Czekanowskiej-Kuklińskiej, profesorki Uniwersytetu Warszawskiego (zm. 2021) i kierowniczki stworzonego przez nią Zakładu Etnomuzykologii.

Stefan Zubrzycki opublikował w 1953 roku pracę [5] wykorzystującą taksonomię wrocławską, w której odpowiada na pytanie astronoma Włodzimierza Zonna, czy gwiazdy układają się w nieprzypadkowe konstelacje (zwane przez autorów *łańcuszkami*), czy też rozmieszczone są losowo na sferze niebieskiej. Wykazał, że układają się losowo, co potwierdza, że gwiazdozbiory są tylko mnemotechniczną metodą zapamiętania położenia gwiazd.

Julian Perkal kończy pracę o taksonomii (op. cit.) ostrzeżeniem: [...] można skonstruować maszynę do robienia dendrytów. Stwarza to niebezpieczeństwo mechanicznego podejścia do przedmiotów przyrodniczych i żyrowania metodami matematycznymi fałszywych niekiedy weksli przyrodniczych. Warto o tym pamiętać.



Zadania



Przygotował Dominik BUREK

M 1750. Czy liczby od 1 do 2023^2 mogą zostać w taki sposób wpisane w pola kwadratowej planszy 2023×2023 , by dla dowolnego wyboru pewnego wiersza i pewnej kolumny można było znaleźć na nich trzy liczby takie, że jedna z nich jest iloczynem dwóch pozostałych?

Rozwiązanie na str. 2

M 1751. Niech O będzie środkiem okręgu opisanego na trójkącie ABC . Punkty X i Y leżą na boku BC i spełniają równości $AX = BX$ oraz $AY = CY$. Wykazać, że okrąg opisany na trójkącie AXY przechodzi przez środki okręgów opisanych na trójkątach AOB i AOC .

Rozwiązanie na str. 8

M 1752. Niech $x_1, \dots, x_n \in [0, 1]$. Udowodnić, że

$$(1 - x_1x_2 + x_1^2) \cdot (1 - x_2x_3 + x_2^2) \cdot \dots \cdot (1 - x_{n-1}x_n + x_{n-1}^2) \cdot (1 - x_nx_1 + x_n^2) \geq 1.$$

Rozwiązanie na str. 4

Przygotował Andrzej MAJHOFER

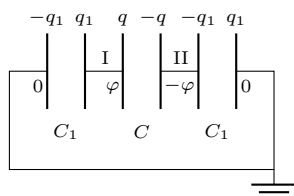
F 1075. Fizycznie podwójny, zaćmieniowy układ gwiazd o promieniach r_1 i r_2 obserwowany jest z Ziemi pod kątem α do płaszczyzny orbity wzajemnego ruchu gwiazd. Jaki warunek wiąże kąt α , promienie r_1 i r_2 oraz średnicę d orbity? Zakładamy, że orbita jest kołowa.

Rozwiązanie na str. 7

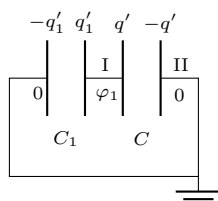
F 1076. Na jedną z okładek kondensatora płaskiego o pojemności C naniesiono ładunek Q_1 , a na drugą ładunek Q_2 . Ile wynosi różnica potencjałów między okładkami?

Wskazówka: Jak zwykle w zadaniach tego typu, zaniedbujemy efekty brzegowe.
Rozwiązanie na str. 7

Klub 44 F



Rys. 1



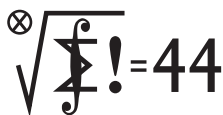
Rys. 2

Czołówka ligi zadaniowej **Klub 44 F**
po uwzględnieniu ocen rozwiązań zadań
752 ($WT = 2,03$), 753 ($WT = 2,73$)
z numeru 2/2023

Jan Zambrzycki	Białystok	4-44+3,03
Tomasz Rudny	Poznań	43,41
Marian Łupieżowiec	Gliwice	2-38,81
Jacek Konieczny	Poznań	36,51
Tomasz Wietecha	Tarnów	16-23,66
Ryszard Baniewicz	Wrocław	1-20,97
Konrad Kapcia	Poznań	2-20,00
Andrzej Nowogrodzki	Chocianów	3-19,70

Regulamin Ligi znajduje się na naszej stronie: https://deltami.edu.pl/delta/redakcja/regulamin_klubu_44/

Klub 44 M



Czołówka ligi zadaniowej **Klub 44 M**
po uwzględnieniu ocen rozwiązań zadań
851 ($WT = 2,97$) i 852 ($WT = 1,17$)
z numeru 12/2022

Norbert Porwol	Essen	40,25
Paweł Najman	Kraków	39,93
Radosław Kujawa	Wrocław	39,13
Adam Woryna	Ruda Śl.	38,27
Marcin Kasperski	Warszawa	37,65
Szymon Tur		35,35
Piotr Kumor	Olsztyn	30,62

Pierwsze równanie (1) pokazuje, że $x^2 \geq a^2 + 4$, co w połączeniu z nierównością kwadratową (2) daje dwustronne oszacowanie:

$$(3) \quad \sqrt{a^2 + 4} \leq x \leq 2 + \sqrt{a^2 + 4}.$$

Nietrudno sprawdzić, że liczba $a + 2$ leży w przedziale $[\sqrt{a^2 + 4}, 2 + \sqrt{a^2 + 4}]$ (długości 2). Gdy $a \geq 1$, jest to jedyna w tym przedziale liczba całkowita tej samej

Redaguje Elżbieta ZAWISTOWSKA

Rozwiązania zadań z numeru 3/2023

Przypominamy treść zadań:

754. Okładki kondensatora płaskiego o pojemności C naładowano do potencjałów φ i $(-\varphi)$ względem ziemi. Każda z okładek tworzy z ziemią kondensator o pojemności C_1 . Znaleźć stosunek natężenia pola elektrycznego między okładkami kondensatora o pojemności C na początku i po uziemieniu jednej z okładek.

755. Zamknięte naczynie całkowicie wypełnione jest wodą. Tuż nad dnem naczynia znajduje się pęcherzyk powietrza. Jak zmieni się ciśnienie na poziomie dna, gdy pęcherzyk wypłynie?

754. Układ równoważny do opisanego w zadaniu przedstawia rysunek 1.

Kondensator o pojemności C naładowany jest ładunkiem $q = 2\varphi C$, ładunki na okładkach kondensatorów o pojemnościach C_1 wynoszą $q_1 = \varphi C_1$. Całkowity ładunek na lewej płytce jest równy

$$(1) \quad Q = q + q_1 = (2C + C_1)\varphi.$$

Po uziemieniu prawej płytki równoważny układ przedstawia rysunek 2. Ładunek na okładkach kondensatora C wynosi $q' = \varphi_1 C$, przez φ_1 oznaczyliśmy potencjał nieuziemionej okładki. Ładunek kondensatora o pojemności C_1 wynosi $q'_1 = \varphi_1 C_1$, a całkowity ładunek lewej płytki to

$$(2) \quad Q' = q' + q'_1 = \varphi_1(C + C_1) = Q.$$

Uwzględniając (1), otrzymujemy

$$\varphi_1 = \frac{\varphi(2C + C_1)}{C + C_1}.$$

Natężenie pola elektrycznego między okładkami kondensatora jest ilorazem

napięcia i odległości między nimi, zatem szukany stosunek tych natężeń wynosi

$$\frac{E}{E_1} = \frac{2\varphi}{\varphi_1} = \frac{2(C + C_1)}{2C + C_1}.$$

755. W stanie początkowym ciśnienie powietrza w pęcherzyku jest takie samo, jak ciśnienie wody nad dnem naczynia. Różnica ciśnień między dnem naczynia a górą wynosi $\Delta p = \rho gh$, gdzie ρ to gęstość wody, h – wysokość naczynia. Podczas wznoszenia się pęcherzyka jego objętość nie zmienia się, bo ciecz jest praktycznie nieściśliwa, nie zmienia się więc również ciśnienie powietrza w pęcherzyku. Gdy pęcherzyk znajduje się u góry naczynia, to ciśnienie wody pod pokrywą równe jest ciśnieniu na dnie w położeniu początkowym, zatem ciśnienie na dole wzrosło o wielkość Δp .

Redaguje Marcin E. KUCZMA

Rozwiązania zadań z numeru 3/2023

Przypominamy treść zadań:

857. Znaleźć wszystkie pary dodatnich liczb całkowitych x, y , dla których liczby $x^2 - 4y$ oraz $y^2 - 4x$ są kwadratami liczb całkowitych.

858. W przestrzeni znajduje się trójkąt równoboczny ABC o boku długości 1 oraz odcinek DE długości 1, mający punkt wspólny z trójkątem ABC . Udowodnić, że pewien z punktów A, B, C, D, E jest w odległości nie większej niż 1 od czterech pozostałych.

857. Szukamy rozwiązań układu równań

$$(1) \quad x^2 = a^2 + 4y, \quad y^2 = b^2 + 4x$$

w liczbach całkowitych $x, y \geq 1$, $a, b \geq 0$. Widać, że liczby x, a muszą mieć jednakową parzystość (tak samo liczby y, b).

Z uwagi na symetrię wystarczy rozważyć sytuację, gdy $x \geq y$. Wówczas $x^2 = a^2 + 4y \leq a^2 + 4x$, czyli

$$(2) \quad x^2 - 4x - a^2 \leq 0.$$

parzystości co a ; zatem $x = a + 2$. Układ równań (1) wymusza równość

$$y = a + 1 \text{ oraz } (a + 1)^2 = b^2 + 4(a + 2).$$

Tę ostatnią przepisujemy w postaci $(a - 1)^2 = b^2 + 8$, czyli

$$(a - 1 - b)(a - 1 + b) = 8,$$

z jedynym rozwiązaniem (w liczbach naturalnych)

$a = 4, b = 1$. Stąd $x = 6, y = 5$. Czwórka $(a, b, x, y) = (4, 1, 6, 5)$ spełnia wyjściowy układ (1).

Gdy natomiast $a = 0$, nierówność podwójna (3) jest spełniona przez dwie liczby parzyste: $x = 2$ i $x = 4$. Po wstawieniu do równań (1) pierwsza z tych wartości daje sprzeczność, zaś druga szybko prowadzi do rozwiązania: $(a, b, x, y) = (0, 0, 4, 4)$.

Uwzględniając symetrię (czyli odrzucając założenie $x \geq y$), dostajemy odpowiedź: szukane pary (x, y) to $(4, 4), (6, 5), (5, 6)$.

858. Tworzymy macierz zerowejedynkową:

$$\begin{bmatrix} i(A, D) & i(B, D) & i(C, D) \\ i(A, E) & i(B, E) & i(C, E) \end{bmatrix}$$

gdzie

$$i(X, Y) = \begin{cases} 0 & \text{gdy } XY \leq 1 \\ 1 & \text{gdy } XY > 1 \end{cases}$$

(XY oznacza długość odcinka o końcach X, Y).

Ponieważ $AB = AC = BC = DE = 1$, zadanie sprowadza się do wykazania, że w tej macierzy któryś wiersz lub któraś kolumna składa się z samych zer. Przypuśćmy więc, że tak nie jest. Wówczas – po ewentualnej permutacji symboli A, B, C i/lub D, E (co nie wpływa na warunki zadania) – pojawia się w macierzy układ jedynek $\begin{bmatrix} 1 & 1 \\ & 1 \end{bmatrix}$; w terminach geometrycznych (przy oznaczeniach przyjętych wcześniej):

$$AD > 1, \quad BD > 1, \quad CE > 1;$$

czyli

$$AD > AC, \quad BD > BC, \quad CE > DE.$$

Niech π będzie płaszczyzną symetralną odcinka CD . Uzyskane nierówności oznaczają, że punkty A i B leżą po tej samej stronie płaszczyzny π co punkt C ; zaś punkt E leży po tej stronie π co punkt D (i żaden z nich nie leży na niej). Tak więc płaszczyzna π oddziela trójkąt ABC od odcinka DE – wbrew założeniu, że te figury mają punkt wspólny. Sprzeczność dowodzi tezy zadania.

Czy tu jest równomiernie? ... czyli o metodzie V/V_{\max}

Radostaw POLESKI*

*Obserwatorium Astronomiczne Uniwersytetu Warszawskiego (rpoleski@astrouw.edu.pl)

W XXI wieku astronomowie pozyskują ogromne ilości danych obserwacyjnych o różnym stopniu szczegółowości. Oczywiście wyekstrahowanie wiedzy z takiego kosmosu danych wymaga analizy statystycznej. Te analizy mogą być bardzo różnego rodzaju: od przetwarzania pikseli ze zdjęcia nieba na pomiary jasności i pozycji obiektów, przez analizę wielu obserwacji w celu znalezienia okresu, z jakim zachodzi jakieś zjawisko (np. zaćmienia w układzie podwójnym gwiazd), po analizę danych z różnych źródeł w celu określenia parametrów wybranego obiektu (np. jaka jest odległość do centrum Galaktyki lub jaką część masy we Wszechświecie stanowi materia barionowa).

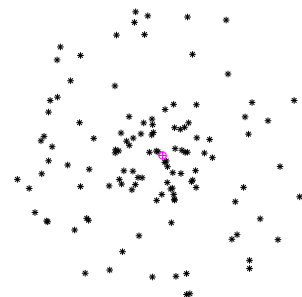
Warto tutaj zwrócić uwagę na pewną istotną różnicę w pozyskiwaniu danych między astronomią a fizyką. Otóż ta pierwsza bazuje w znacznej mierze na obserwacjach zjawisk, w które nie ingerujemy, natomiast ta druga opiera się przede wszystkim na wynikach kontrolowanych doświadczeń. Typowo obserwacyjny charakter badań astronomicznych może stanowić istotną trudność – zwiększenie próbki badanych obiektów może być bardzo kosztowne, jak również może wymagać czasu dłuższego niż przewidywany czas życia badacza. Z reguły mamy zatem do czynienia z próbką dalece niekompletną (choć bardzo chcielibyśmy, by było inaczej). Są jeszcze inne przeszkody, jak na przykład to, że informacje o badanych obiektach mogą pochodzić z obserwacji wykonywanych w różnych warunkach i w różnym czasie itp. Uwzględnienie w analizie statystycznej tych subtelności nie jest łatwe, ale konieczne.

W niniejszym artykule chciałbym przedstawić pewną typowo astronomiczną metodę o egzotycznej

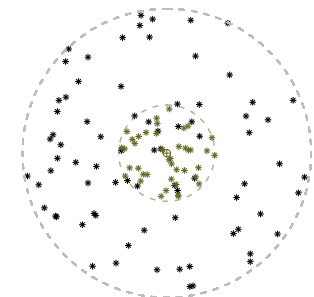
nazwie „ V/V_{\max} ”. Została ona wymyślona pod koniec lat 60., by rozwiązać następujący problem: mamy do dyspozycji katalog kwazarów o znanych jasnościach oraz przesunięciach ku czerwieni i chcemy sprawdzić, czy kwazary te są rozmieszczone równomiernie w przestrzeni.

Kwazar (z ang. *quasi-stellar object*) jest to rodzaj galaktyki aktywnej, która emituje niezwykle jasne promieniowanie.

Zwróćmy uwagę, że dla kwazarów emitujących najwięcej światła nasza próbka może być traktowana jako kompletna dla dużych odległości, jednak dla kwazarów emitujących mniej światła próbka jest kompletna jedynie dla małych odległości. Wynika stąd, że choć na pierwszy rzut oka może wydawać się, że koncentracja kwazarów spada wraz ze wzrostem ich odległości od Ziemi, to może to być artefakt zależności, że im bliżej Ziemi jesteśmy, tym więcej kwazarów jesteśmy w stanie zaobserwować (por. rys. 1 i 2).



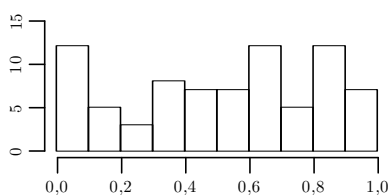
Rys. 1. Rozmieszczenie 150 kwazarów zaobserwowanych przez pewnego dwuwymiarowego astronoma w dwuwymiarowym Wszechświecie. Kwazary wykazują pozorną koncentrację wokół Ziemi (oznaczonej symbolem \oplus)



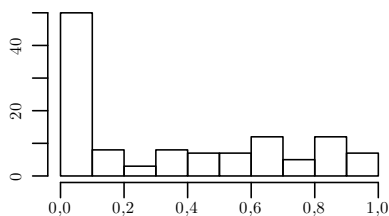
Rys. 2. Rozmieszczenie kwazarów identyczne z tym z rysunku 1, z zaznaczeniem, do której z dwóch grup o ustalonych jasności absolutnych należy każdy obiekt. W ramach każdej grupy kwazary są rozłożone równomiernie w całym obszarze, w jakim można je zaobserwować

Zacznijmy od pewnych prostych rozważań. Załóżmy, że wszystkie kwazary mają tę samą jasność absolutną (czyli taką, jaką badany obiekt miałby, gdyby był widziany z odległości 10 parseków). Gdyby wszystkie kwazary były równomiernie rozłożone w przestrzeni, to *obserwowane* kwazary byłyby równomiernie rozmieszczone w obszarze (kuli), w którym mamy szansę je zaobserwować (por. czarne kropki na rys. 2). Niech V_{\max} będzie objętością tej kuli. Dla każdego

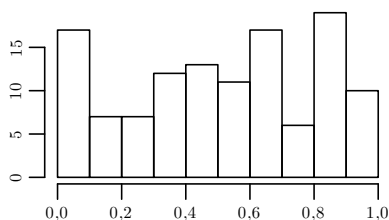
kwazara K wyznaczamy jego odległość od nas i obliczamy objętość kuli V_K o promieniu takim jak ta odległość. Jeśli założenie o równomiernym rozkładzie jest spełnione, to rozkład wartości V/V_{\max} powinien być jednostajny w przedziale $[0, 1]$ (rys. 3), zatem średnia wartość V/V_{\max} powinna być w przybliżeniu równa $1/2$. Jeśli rozkład V/V_{\max} nie posiada tej cechy, to hipotezę o równomiernym rozmieszczeniu kwazarów w przestrzeni należy raczej odrzucić.



Rys. 3. Histogram wartości V/V_{\max} dla „czarnych kwazarów” z rysunku 2, o stałej wartości V_{\max} . Wysokość słupka odpowiada liczbie kwazarów o wartości V/V_{\max} leżących w przedziale wyznaczonym przez podstawę słupka



Rys. 4. Histogram wartości V/V_{\max} dla wszystkich kwazarów z rysunku 2 z założeniem stałej wartości V_{\max}



Rys. 5. Histogram wartości V/V_{\max} dla wszystkich kwazarów z rysunku 2 z V_{\max} wyznaczoną oddzielnie dla każdego kwazara

A co, jeśli kwazary będą miały różne jasności absolutne (czyli tak jak w rzeczywistości)? Najpierw założmy (dla prostoty), że dla każdego kwazara jasność absolutna przyjmuje jedną z dwóch wartości (czyli tak jak na rys. 1 i 2). Jak już podkreśliliśmy, wpłynie to na próbkę obserwowanych kwazarów, bo te świecące jaśniej będą widoczne z odległości większych niż te świecące słabiej. Jeśli użyjemy takiej samej wartości V_{\max} dla każdego kwazara, to zobaczymy histogram taki jak na rysunku 4, czyli znacząco inny niż ten z rysunku 3. Musimy zmienić nasze obliczenia i wartość V_{\max} wyznaczać oddzielnie dla każdego kwazara, w zależności od jego jasności absolutnej (dla kwazara K oznaczmy ją teraz jako $V_{\max,K}$). Tym niemniej, jeśli wszystkie kwazary są rozmieszczone równomiernie w przestrzeni, to dla każdego kwazara wartość $V_K/V_{\max,K}$ możemy traktować jak jednostajnie wylosowaną z odcinka $[0, 1]$, niezależnie od pozostałych. Innymi słowy, rozkład wartości V/V_{\max} wciąż jest jednorodny na odcinku $[0, 1]$ (por. rys. 4 i 5), a jego średnia wartość jest bliska $1/2$. Te cechy mają zarówno wszystkie kwazary z rysunku 2, jak i tylko czarne lub tylko kolorowe.

Jesteśmy teraz gotowi na uogólnienie – każdy kwazar ma inną jasność absolutną. Dla każdego kwazara możemy obliczyć jego indywidualne wartości V_K i $V_{\max,K}$. Znowu oczekujemy, że dla równomiernego rozkładu kwazarów rozkład V/V_{\max} wciąż jest jednorodny na odcinku $[0, 1]$ i ma średnią bliską $1/2$. Jeśli mamy dane obserwacyjne, to wystarczy wyznaczyć V/V_{\max} , sprawdzić, czy wynik jest zgodny z oczekiwanym i... gotowe!

Metoda V/V_{\max} doczekała się różnych rozwinięć. Dla Ciekawskich Czytelników (i jednocześnie tych, którzy są odrobinę bardziej obeznani z tą tematyką) chciałbym szkicowo przedstawić jedno z nich, moim zdaniem najważniejsze. Jest to możliwość analizy, jak zmieniają się parametry badanej populacji w zależności od czasu. Wyobraźmy sobie Wszechświat w wieku odpowiadającym przesunięciu ku czerwieni z . Niech $\rho(z)$ będzie stosunkiem gęstości przestrzennej badanych obiektów wtedy i obecnie (tj. $z = 0$). Zamiast objętości V rozważamy uogólnioną objętość: $V'(z) = \int_0^z \rho(z') dV(z')$, gdzie $V(z)$ to objętość we współrzędnych współporuszających się, która jest wyliczana na podstawie założonego modelu kosmologicznego. Dla każdego kwazara znamy jego przesunięcie ku czerwieni z_K , więc możemy obliczyć odpowiednie $V'_K(z_K)$ i $V'_{\max,K}(z_K)$. Jeśli założymy, że kwazary są rozmieszczone równomiernie, i założymy funkcję $\rho(z)$ bliską prawdziwej, to rozkład V'/V'_{\max} powinien być... (chyba nie będę się powtarzał). Skąd mamy wiedzieć, jakiego $\rho(z)$ użyć? Najprościej jest spróbować wielu różnych możliwości, dla każdej z nich obliczyć V'/V'_{\max} , a na koniec wybrać zbiór tych, które dają oczekiwane wyniki.

Warto podkreślić, że metoda V/V_{\max} i jej warianty wciąż pojawiają się w publikacjach naukowych. Dla przykładu, kilka lat temu była zastosowana do zupełnie innego zagadnienia – badania częstości występowania planet pozasłonecznych $f(q)$ w zależności od stosunku masy planety do masy gwiazdy q dla bardzo małych wartości tych stosunków, tzn. tam, gdzie nasza wiedza jest dosyć uboga ([1], [2]). Problem polegał na tym, że uwzględniono planety znalezione w danych zbieranych w niejednorodny sposób. Zamiast $\rho(z)$ poszukiwana była funkcja $f(q)$, a zamiast $V(z)$ używane było prawdopodobieństwo znalezienia w danym układzie planety o innym stosunku mas. Dokładny opis tych badań, a także wyników, to jednak jest już temat na inną opowieść...

Literatura

- [1] Udalski i in. *OGLE-2017-BLG-1434Lb: Eighth $q < 1 \times 10^{-4}$ Mass-Ratio Microlens Planet Confirms Turnover in Planet Mass-Ratio Function*. Acta Astronomica 68.1 (2018): 1-42.
- [2] Jung i in. *KMT-2017-BLG-0165Lb: A Super-Neptune-mass Planet Orbiting a Sun-like Host Star*. The Astronomical Journal 157.2 (2019): 72.

Prosto z nieba: Drewniany zegar

Na podstawie ilości radioaktywnego pierwiastka obecnego w badanych próbkach archeolodzy określają ich wiek. Węgiel-14, jako nietrwały izotop, zamienia się z powrotem na azot, z czasem półrozpadu około 5750 lat. Dzięki ciągłemu pobieraniu i wydalaniu materii organizmy żywe utrzymują mniej więcej stały poziom radioaktywnych pierwiastków, natomiast w organizmach martwych z czasem jest ich coraz mniej.

w przeszłości sięgającej nawet tysięcy lat. Zawartość węgla-14 zależy też od innych czynników, np. od pól magnetycznych Ziemi i Słońca, które osłaniają powierzchnię Ziemi przed promieniowaniem kosmicznym pochodzącym spoza Układu Słonecznego (więcej cząstek dociera do Ziemi, gdy te pola są słabsze, a mniej, gdy są silniejsze). Zmiany poziomu węgla-14 rejestrowane w słojach pni drzew przechowują historię zmian magnetyzacji Ziemi. Są też dowodem na 11-letni cykl dynamo słonecznego, które jest związane z polem magnetycznym Słońca.

Drewno zawiera w sobie także takie dane, których nie potrafimy wyjaśnić. W 2012 roku japońska fizyczka Fusa Miyake odkryła znaczny skok zawartości węgla-14 w słojach drzew z 774 roku. Tak duży skok musiało wywołać promieniowanie kosmiczne wielokrotnie większe od przeciętnego. Kolejne „zdarzenia Miyake” to lata 993 naszej ery i 663 przed naszą erą, a także jeszcze wcześniejsze – z 5259, 5410 i 7176 przed naszą erą. Dobrze zlokalizowane w drewnie (i w czasie) zdarzenia pozwalają na precyzyjne określenie daty konkretnych wydarzeń z dokładnością co do roku. I tak zdarzenie z 993 roku pozwoliło na stwierdzenie momentu powstania pierwszej europejskiej osady w Ameryce – założonej w roku 1021 wioski wikingów w Nowej Fundlandii.

Jak dochodzi do tak ogromnego i krótkotrwałego promieniowania? Wśród „podejrzanych” są pobliskie supernowe, błyski promieniowania gamma, emisja z namagnesowanych gwiazd neutronowych, a nawet komety. Aktualnie najlepszym wyjaśnieniem jest to,

“Modelling cosmic radiation events in the tree-ring radiocarbon record”, Qingyuan Zhang i inni, Proc. R. Soc. A. 478 2022.0497, 2022.

Niebo w lipcu

Przez cały miesiąc Słońce obniży wysokość swojego górowania o ponad $4,5^\circ$, skracając przy tym czas przebywania nad horyzontem w środkowej Polsce o ponad godzinę. 24 lipca Słońce przetnie równoleżnik $+20^\circ$ deklinacji i tym samym skończy się okres najdłuższych dni i najkrótszych nocy. Jak co roku, na początku lipca Ziemia znajduje się w aphelium swojej orbity, co oznacza, że Słońce ma najmniejszą tarczę w ciągu roku. Dlatego łatwiej jest ją zasłonić Księżycowi podczas ewentualnego zaćmienia – i takie zaćmienia trwają też dłużej.

W lipcu najjaśniejsze planety Układu Słonecznego w większości są widoczne słabo. **Mercury** zacznie miesiąc od górnego złączenia ze Słońcem i podąży ku maksymalnej elongacji wschodniej, którą osiągnie

Kiedy wysokoenergetyczne promieniowanie kosmiczne zderza się z górnymi warstwami ziemskiej atmosfery, niektóre kolizje prowadzą do powstania neutronów, co prowadzi do reakcji neutron-proton, $n-p$, czyli zamiany jąder azotu $^{14}_7\text{N}$ w radioaktywny węgiel-14, $^{14}_6\text{C}$: $^{14}_7\text{N} + n \rightarrow ^{14}_6\text{C} + p$. Węgiel-14 opada na powierzchnię Ziemi i bierze udział w zwykłej biochemii żywych organizmów, w tym zostaje związany podczas wzrostu drzew.

Badając np. pnie drzew, a w szczególności różnice w zawartości węgla-14 w poszczególnych słojach, można określić ilość początkową radioaktywnego materiału, czyli poniekąd zbadać ewolucję ilości promieniowania kosmicznego

że zdarzenia Miyake są związane ze słonecznymi superburzami. Te (hipotetyczne) erupcje ze Słońca są 50–100 razy bardziej energetyczne niż największa zarejestrowana w erze nowożytnej burza słoneczna, obserwowana przez Richarda C. Carringtona i Richarda Hodgsona w 1859 roku.

W swojej pracy Qingyuan Zhang i współpracownicy analizują dostępne materiały pochodzące ze słoju drzew, znajdując dowody na to, że zdarzenia mogą pojawić się w każdym momencie 11-letniego cyklu aktywności Słońca (o którym pisaliśmy np. w Δ_{21}^1). Z drugiej strony, rozbłyski słoneczne mają tendencję do występowania w okolicach szczytu cyklu. Kilka z zarejestrowanych skoków radioaktywności wydaje się trwać dłużej niż wskazywałby na to model pojedynczej superburzy słonecznej. Sugeruje to, że czasami zdarzenia mogą trwać dłużej niż rok, co nie jest oczekiwane w przypadku jednego gigantycznego rozbłysku słonecznego, czyli że mieliśmy do czynienia z długotrwałą burzliwą pogodą słoneczną.

Gdyby takie zdarzenie zaistniało dzisiaj, zniszczyłoby sieci energetyczne, telekomunikacyjne i większość satelitów. Jeśli takie zdarzenia występują losowo, na przykład raz na tysiąc lat, to prawdopodobieństwo wystąpienia burzy słonecznej w ciągu następnej dekady wynosi około 1%. To niezaniebawalne duże prawdopodobieństwo!

Michał BEJGER

Centrum Astronomiczne im. Mikołaja Kopernika PAN, Istituto Nazionale di Fisica Nucleare (INFN), Sezione di Ferrara, Włochy

w pierwszej dekadzie sierpnia, oddalając się wtedy na bardzo duże 27° od Słońca. Niestety o tej porze roku i doby ekliptyka jest nachylona niekorzystnie do widnokregu, przez co planeta zachodzi mniej niż godzinę po Słońcu, i z dużych szerokości geograficznych jest niewidoczna. Szczególnie że z każdą kolejną dobą jasność planety staje się coraz mniejsza, od $-0,4^m$ 19 lipca do $+0,1^m$ 31 lipca. A szkoda, bo 28 lipca planeta przejdzie mniej niż $20'$ od Regulusa, najjaśniejszej gwiazdy Lwa.

Warunki obserwacyjne **Venus** są jeszcze gorsze. Druga planeta od Słońca po czerwcowej maksymalnej elongacji szybko dąży do sierpniowej koniunktacji dolnej ze Słońcem. Oznacza to, że planeta prędko zbliża się do nas, zwiększając przy tym rozmiary

tarczy i zmniejszając fazę. Na początku lipca Wenus zaprezentuje tarczę o jasności $-4,4^m$, średnicy $34''$ i fazie 31%. Ostatniego dnia miesiąca odpowiednie wielkości wyniosą: $-4,2^m$, $53''$ i 5%. Niestety planeta zanurkuje przy tym kilka stopni pod słabo nachyloną ekliptykę, ginąc w zorzy wieczornej pod koniec miesiąca. Tarcza Wenus w lipcu jest zatem atrakcyjnym celem obserwacyjnym dla posiadaczy nawet niedużych lornetek i teleskopów, jednakże nie dla tych przebywających daleko na północ od równika.

Na początku lipca po zmierzchu można próbować dostrzec także **Marsa**. Czerwona Planeta pierwotnie przebywa niecałe 4° od Wenus, ale potem Mars powędruje dalej na południowy wschód, a Wenus zawróci w kierunku Słońca, więc do 20 lipca dystans między planetami zwiększy się ponad dwukrotnie. 10 lipca Mars minie wspomnianego już Regulusa w odległości niewiele przekraczającej $0,5^\circ$. Marsowi najtrudniej jest przebić się przez zorzę wieczorną, gdyż jest on już daleko od Ziemi i świeci z jasnością $+1,7^m$, czyli niewiele większą od pobliskiej najjaśniejszej gwiazdy Lwa.

Wszystkie te planety w trzeciej dekadzie miesiąca odwiedzi Księżyc w fazie cienkiego sierpa. Nów Srebrnego Globu przypada 17 lipca i w następnych dniach przeniesie się on na niebo wieczorne, ale też ucierpi wskutek nisko położonej ekliptyki. Jego sytuację poprawi nieco fakt, że prawie cały czas, do przypadającej 25 lipca I kwadry, spędzi on na północ od niej. Już 18 lipca można spróbować dostrzec bardzo cienki sierp Księżyca o zmierzchu, jednak jest to trudne zadanie, wymagające bardzo czystej atmosfery i nisko odsłoniętego widnokręgu. 30 minut po zachodzie Słońca tarcza Księżyca w fazie zaledwie 1% zajmie pozycję na wysokości 3° . Tymczasem 7° na lewo od Księżyca znajdzie się planeta Merkury, kolejne zaś 15° dalej – planeta Wenus. Z kolei $3,5^\circ$ nad Wenus pokaże się Regulus, 5° od Regulusa, na godzinie 10. względem niego znajdzie się planeta Mars. Księżyc, Merkury i Wenus zajdą bardzo szybko, Regulus i Mars nieco później, w odszukaniu jednak wszystkich tych ciał niebieskich niezbędna może się okazać lornetka. 19 lipca Księżyc w fazie 4% przejdzie 4° nad Merkurym, dobę później natomiast, przy fazie zwiększonej do 8%, Srebrny Glob przejdzie 3° nad Regulusem i jednocześnie 7° nad Wenus. 6° na lewo od Księżyca pokaże się Mars. 21 lipca faza księżycowej tarczy urośnie do 14%, Marsa zaś należy wtedy szukać w odległości 6° na godzinie 4. względem Księżyca.

Po minięciu planet naturalny satelita Ziemi w okolicach I kwadry, 24 i 25 lipca, spotka się ze Spiką, najjaśniejszą gwiazdą Panny. Warto jeszcze wspomnieć o bardzo bliskim spotkaniu Księżyca z Antaresem, 28 lipca. W momencie zachodu Słońca tarcza Księżyca w fazie 79% pokaże się $0,5^\circ$ od najjaśniejszej gwiazdy Skorpiona. Do końca miesiąca Księżyc pozostanie na południe od ekliptyki, wędrując nisko nad horyzontem. Im bliżej jednak końca lipca, tym księżycowa tarcza

pełniejsza, gdyż 1 sierpnia wieczorem naszego czasu Księżyc przejdzie przez pełnię.

Pod wpływem blasku tarczy Srebrnego Globu upłynie również początek miesiąca. Lipcowa pełnia przypada 3 lipca w gwiazdozbiornie Strzelca. Przed nią w pierwszych dwóch nocach miesiąca Księżyc odwiedzi Skorpiona, świecąc najpierw 6° na prawo, a potem 8° na lewo od Antaresa. 7 lipca, prezentując tarczę oświetloną w 83%, naturalny satelita Ziemi zbliży się na 5° do Saturna. Planeta z pierścieniami w sierpniu przejdzie przez opozycję i będzie poruszać się na południowy zachód mniej niż 1° od gwiazdy 5. wielkości σ Aqr. Tarcza planety przekroczy średnicę $18''$, świecąc z jasnością $+0,7^m$. Saturn będzie górować o świcie, wznosząc się na wysokość ponad 25° .

Srebrny Glob spotka się z **Neptunem** 9 lipca, zbliżając się doń na odległość 5° . Planeta we wrześniu przejdzie przez opozycję względem Słońca i także poruszać się będzie ruchem wstecznym. W tym roku zakreśli pętlę na pograniczu gwiazdozbioru Ryb i Wodnika, niedaleko charakterystycznego, przypominającego miniaturowy Mały Wóz układu gwiazd 5. i 6. wielkości, który tworzą gwiazdy 30, 33, 27, 29, 24 i 20 Psc. W lipcu Neptun znajduje się około 1° na północ od 24 Psc, świeci z jasnością $+7,8^m$ i około godziny 2 wznosi się na wysokość powyżej 20° ponad południowo-wschodni widnokrąg.

Księżyc przejdzie przez ostatnią kwadrę 10 lipca, a następnie podąży ku słabo widocznym planetom **Jowisz** i **Uran**. 12 lipca nad ranem faza księżycowej tarczy spadnie poniżej 30% i wędruje on tuż po północy $2,5^\circ$ od Jowisza i jednocześnie 9° na zachód od Urana. Obie planety w nadchodzącym sezonie obserwacyjnym zakreślą pętle na tle gwiazdozbioru Barana w odległości kilku-kilkunastu stopni od siebie i w listopadzie obie przejdą przez opozycję względem Słońca. Na razie dzieli je mniej więcej 10° i około godziny 2 wznoszą się na 15° ponad wschodnią część nieboskłonu. Jowisz świeci z jasnością $-2,3^m$, prezentując tarczę o średnicy $38''$, więc z jego dostrzeżeniem nie ma kłopotu. Co innego Uran, którego jasność wynosi $+5,8^m$ i ze względu na małą wysokość na ciemnym niebie jego obraz zależy od stanu atmosfery.

Nów Srebrnego Globu przypada 17 lipca wieczorem naszego czasu i dzięki temu, że jego orbita jest teraz prawie maksymalnie wychylona na północ od ekliptyki, jego cienki sierp wraz z tzw. światłem popielatym pozostanie widoczny przez kolejne 4 dni. 13 dnia miesiąca sierp Księżyca w fazie 20% zbliży się na 3° do Plejad, dobę później jego faza spadnie do 13% i przejdzie on 8° nad Aldebaranem, 15 lipca zaś, w fazie 7%, minie w odległości nieco ponad 2° El Nath, dwie jasne gwiazdy Byka. 16 lipca o świcie Księżyc w fazie zaledwie 3% pokaże się na wysokości 7° , ponad 20° pod Capellą.

Ariel MAJCHER



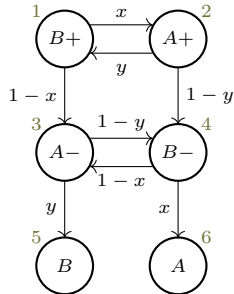
Łańcuchy Markowa – część 1

Bartłomiej BZDEGA

Uniwersytet im. A. Mickiewicza w Poznaniu

Rozważmy następujące zadanie:

Dwóch zawodników – A i B – gra w szachy, przy czym A rozpoczyna. Każdy ruch może być dobry lub słaby. Wygrywa ten, kto jako pierwszy odpowie dobrym ruchem na słaby ruch przeciwnika. Zawodnik A wykonuje dobry ruch z prawdopodobieństwem x , a słaby z prawdopodobieństwem $1 - x$. Analogicznie dla zawodnika B z wartościami y i $1 - y$. Zakładamy, że $0 < x < 1$ i $0 < y < 1$. Dla jakich x i y zwycięstwa A i B są równo prawdopodobne?



Rozwiązanie. Zwróćmy uwagę, że póki nikt nie wygrał, sytuacja w rozgrywce zależy tylko od ostatniego ruchu. Są cztery możliwości: po dobrym/słabym ruchu zawodnika A/B . Oznaczmy je przez $A+$, $A-$, $B+$, $B-$. Moment, w którym zawodnik A rozpoczyna, możemy rozważyć oddzielnie, ale nie ma takiej konieczności, bo jest on równoważny sytuacji $B+$. Możemy więc przyjąć $B+$ jako punkt wyjścia. Do tego dołożymy jeszcze: A (A wygrał) i B (B wygrał). Na rysunku pokazano wszystkie sześć sytuacji, przyporządkowano im liczby od 1 do 6 oraz połączono je strzałkami opisującymi prawdopodobieństwa przejść.

Taki obiekt nazywamy skończonym łańcuchem Markowa. Jest on opisany przez pewien zbiór stanów S_1, S_2, \dots, S_n oraz prawdopodobieństwa $p_{i,j}$ przejścia w jednym kroku ze stanu S_i do stanu S_j dla wszystkich $i, j \in \{1, 2, \dots, n\}$. Oznaczamy to, że będąc w stanie S_i łańcuch z prawdopodobieństwem $p_{i,1}$ przejdzie do stanu S_1 , z prawdopodobieństwem $p_{i,2}$ do stanu S_2 i tak dalej. Dla każdego i zachowana jest równość $p_{i,1} + p_{i,2} + \dots + p_{i,n} = 1$. Stany S_i , z których nie można wyjść ($p_{i,i} = 1$), nazywamy pochłaniającymi. Interesuje nas prawdopodobieństwo zwycięstwa zawodnika A , czyli że proces zakończy się w stanie S_6 , pod warunkiem, że rozpoczął się w S_1 . Na ten problem trzeba spojrzeć ogólniej – niech q_i oznacza prawdopodobieństwo, że proces zakończy się

w stanie S_6 , pod warunkiem, że rozpoczął się w stanie S_i . Zwróćmy przy tym uwagę, że $q_6 = 1$ i $q_5 = 0$. Ze stanu S_1 można przejść do S_2 (z prawdopodobieństwem x) lub S_3 (z prawdopodobieństwem $1 - x$). Wynika z tego, że $q_1 = xq_2 + (1 - x)q_3$. Analogicznie otrzymujemy równości: $q_2 = yq_1 + (1 - y)q_4$, $q_3 = (1 - y)q_4$, $q_4 = (1 - x)q_3 + x$. Po rozwiązaniu układu czterech równań z niewiadomymi q_1, q_2, q_3, q_4 otrzymamy: $\frac{1}{2} = q_1 = \frac{x(1-y)}{(1-xy)(x+y-xy)} \iff x - y = xy(1-x)(1-y)$, więc dla x i y spełniających ostatnią równość obaj gracze mają jednakową szansę na zwycięstwo. Zauważmy, że w szczególności musi zachodzić $x > y$, gdyż prawa strona ostatniej równości jest dodatnia.

Zadania

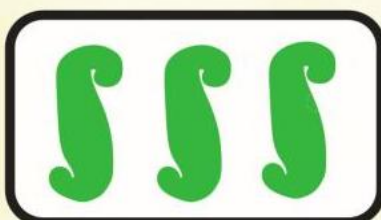
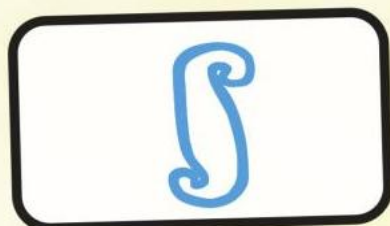
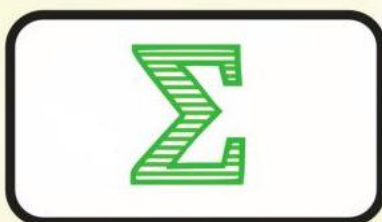
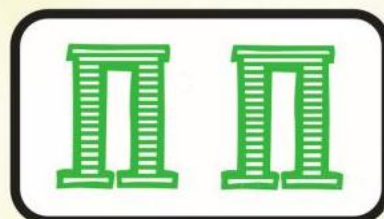
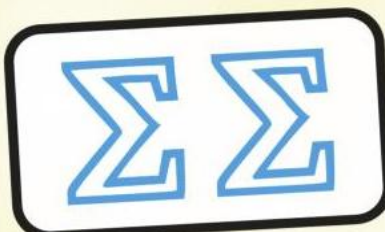
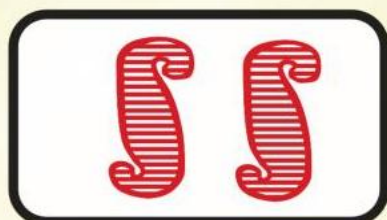
- Kotek wędruje między domem, przedszkolem, ogrodem, polem i lasem. Zaczyna w ogrodzie – tam zawsze wybiera jedno z pozostałych czterech miejsc. W domu zawsze się bawi, a następnie idzie na pole lub do ogrodu. Po spacerze w lesie kot zawsze idzie albo do ogrodu, albo do przedszkola. Jeśli kot trafi do przedszkola, już stamtąd nie wychodzi (ach, te dzieci...). Zaś na polu kot łowi mysz i kończy wędrowkę. Obliczyć prawdopodobieństwo zakończenia wędrowki w polu. (Zakładamy, że wybory kota są losowe i zawsze jednakowo prawdopodobne).
- Student chce kupić swój ulubiony napój energetyczny, który kosztuje 5 złotych. Niestety, student ma tylko 2 złote. Postanawia w takim razie pójść do kasyna, w którym za postawioną złotówkę z prawdopodobieństwem p można wygrać 3 złote albo stracić postawioną złotówkę z prawdopodobieństwem $1 - p$. Student przestaje grać, gdy uzyska kwotę pozwalającą kupić napój albo gdy zostanie bez pieniędzy. W zależności od p wyznaczyć szansę na to, że student osiągnie swój cel. *Uwaga. Autor kącika zdecydowanie nie promuje napojów energetycznych ani tym bardziej hazardu.*
- Jaś i Małgosia rzucają uczciwą monetą. Jeśli w trzech kolejnych rzutach pojawi się konfiguracja ORR (O oznacza orła, a R reszkę), gra kończy się wygraną Jasia, a jeśli OOR – wygraną Małgosi. Wyznaczyć prawdopodobieństwo wygranej Jasia.
- Siedmioro dzieci stoi w kręgu i bawi się piłką. Każde z nich, mając w danej chwili piłkę, rzuca ją do dziecka stojącego bezpośrednio z lewej strony (z prawdopodobieństwem $p < \frac{1}{2}$) lub bezpośrednio z prawej strony (również z prawdopodobieństwem p) albo zabiera piłkę i wraca do domu (z prawdopodobieństwem $1 - 2p$). W zależności od p wyznaczyć prawdopodobieństwo, że do domu z piłką wróci to samo dziecko, które ją przyniosło.

Wskazówki do zadań

- Niech S_1, S_2, \dots, S_5 oznaczają kolejne pobyt kota w domu, przedszkolu, ogrodzie, polu i w lesie. Prawdopodobieństwa oznaczamy $p_{i,j}$ i q_i . Oczwiaszczenie $q_4 = 1$ i $q_5 = 0$. Wartość q_3 wyznaczamy, rozwiązując układ równań: $q_1 = \frac{1}{2}q_2 + \frac{1}{2}q_3$, $q_2 = \frac{1}{2}q_1 + \frac{1}{2}q_4$, $q_3 = \frac{1}{2}q_2$.
- Stan S_i dla $i = 0, 1, 2, 3, 4$ można utozsamnić z kwotą, którą ma student, a stan S_5 – to ten, w którym student ma 5 lub 6 złotych, czyli stać go na napój.
- Jako stany można w tym zadaniu wyróżnić start, O i R (wynik pierwszego rzutu oraz tych następujących po reszce), ORR (wygrana Małgosi) i OOR (wygrana Jasia). Szansa na wygraną Jasia to $\frac{1}{2}$.
- Ciekawostka.** Dla każdego wyboru konfiguracji trzech kolejnych wyników skutkujących wygraną Jasia Małgosia może wybrać inną konfigurację, która daje jej przewagę w takiej rozgrywce (RR oznacza O Reszka i OOR).
- Pomnijmy dzieci od 1 do 7 według kolejności stania w kręgu. Niech S_i będzie stanem, w którym i -te dziecko ma piłkę, a S_{7+i} – stanem, w którym i -te dziecko zabiera piłkę do domu. Założymy, że pierwsze dziecko przyszedł do z piłką, więc stanem początkowym jest S_1 , a interesuje nas q_8 . Rachunki znacznie ułatwiają równości $q_2 = q_7$ i $q_4 = q_5$, które wynikają z symetrii.

66. SZKOŁA MATEMATYKI POGLĄDOWEJ

PODOBIENSTWA I RÓŻNICE



TERMIN: 25-28 SIERPANIA 2023

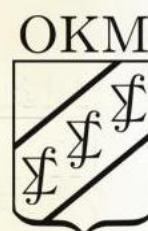
MIEJSCE: SIEDLCE

ORGANIZATOR:

OŚRODEK

KULTURY

MATEMATYCZNEJ



smp.uph.edu.pl