MODELOWANIE RZECZYWISTOŚCI

Daniel Wójcik
Instytut Biologii Doświadczalnej PAN

d.wojcik@nencki.gov.pl

tel. 5892 424

http://www.neuroinf.pl/Members/danek/swps/

Podręcznik

Iwo Białynicki-Birula
Iwona Białynicka-Birula

ISBN: 83-7255-103-0
Data wydania: 6 maja 2002

Metoda Monte Carlo

- Metoda Monte Carlo polega na wykonaniu wielu eksperymentów losowych w celu oszacowania wyniku.
- Program **Ulam**

Definicja prawdopodobieństwa Laplace'a

- Prawdopodobieństwo PA jest równe stosunkowi liczby przypadków sprzyjających wystąpieniu zdarzenia A do wszystkich możliwych przypadków
- Jak na podstawie genetyki otrzymać stosunek liczby urodzeń chłopców do liczby wszystkich urodzeń?

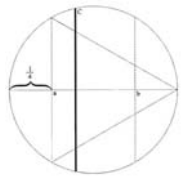
Nieskończona liczba przypadków

- Jakie jest prawdopodobieństwo zastania na stacji stojącego pociągu, jeżeli wiemy, że pociągi wjeżdżają na stację co 10 minut i stoją na niej minutę?
- Paradoks Bertranda:

  jakie jest prawdopodobieństwo tego, że na chybił trafił wybrana cięciwa koła będzie dłuższa od ramienia trójkąta równobocznego wpisanego w to koło?
- Program **Bertrand**

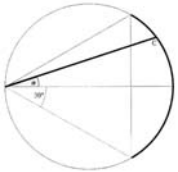
Paradoks Bertranda – rozwiązanie 1

Ustalmy kierunek cięciwy, np. pionowy. Przesuwając cięciwę od lewa do prawa widzimy,że tylko pomiędzy punktami a i b długość cięciwy jest większa od połowy średnicy. Długość ab jest równa połowie średnicy, zatem prawdopodobieństwo jest 1/2.



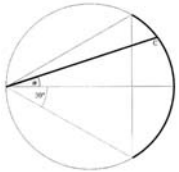
Paradoks Bertranda – rozwiązanie 2

Rozważmy cięciwę zaczepioną w jednym punkcie. Zmieniając jej kąt nachylenia względem średnicy w tym punkcie od -90 do +90 stopni dostajemy wszystkie możliwe cięciwy. Te z nich, które tworzą kąt od -30 do +30 stopni ze średnicą są dłuższe od połowy średnicy. Zatem szukane prawdopodobieństwo wynosi 60/180 = 1/3



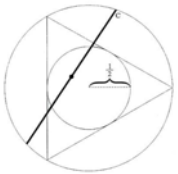
Paradoks Bertranda – rozwiązanie 3

Wybierzmy przypadkowo dwa punkty na okręgu łącząc je cięciwą. Pierwszy punkt jest dowolny, drugi musi leżeć na jednej trzeciej okręgu naprzeciwko pierwszego punktu, by cięciwa była odpowiednio długa. Zatem prawdopodobieństwo wynosi również 1/3.



Paradoks Bertranda – rozwiązanie 4

Środek każdej cięciwy leży wewnątrz koła i wyznacza jednoznacznie jej położenie. Stosunek części koła w której leżą środki cięciw spełniających zadany warunek do pola całego koła wynosi 1/4; ta część koła jest również kołem o promieniu równym połowie długości promienia dużego koła.



Ciągi liczb losowych

- Czy ciąg 0,1,0,1,0,1,... jest losowy?
- Napisz na kartce ciąg zer i jedynek o długości 100 znaków
- Policz ile jest w tym ciągu podciągów złożonych z trzech, czterech i pięciu jedynek
- Porównaj z wynikami wygenerowanymi przez program **Bernoulli**

Generatory liczb losowych

- Komputery generują liczby pseudolosowe, nie losowe
- Wygenerowane liczby powtarzają się po pewnym czasie. Im dłuższy okres, tym lepszy generator. Im lepieje „potasowane” liczby, tym lepszy generator.

Prawdopodobieństwo wylosowania ciągów samych jedynek

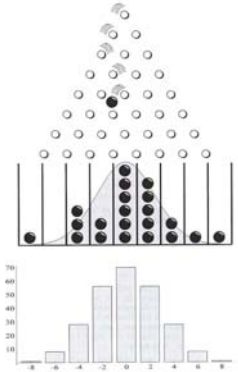
- Aby uzyskać samotną jedynkę w binarnym ciągu trzeba wyrzucić po kolei 0,1,0 – prawdopodobieństwo ½ * ½ * ½ = 1/8
- Aby uzyskać n jedynek w binarnym ciągu trzeba wyrzucić po kolei 0,1, ... ,1,0 – to daje prawdopodobieństwo

$$\frac{1}{2}*\underbrace{\frac{1}{2}*\ldots*\frac{1}{2}}_{n\,razy}*\frac{1}{2}=\frac{1}{2}^{(n+2)}$$

- Zatem prawdopodobieństwo wyrzucenia 6,7,8 jedynek pod rząd to odpowiednio 1/256, 1/512 i 1/1024

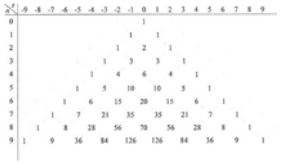
Deska Galtona

Pochylona deska z wbitymi gwoździami ułożonymi w trójkąt. Można jej użyć do wizualizacji wielokrotnego rzucania monetą



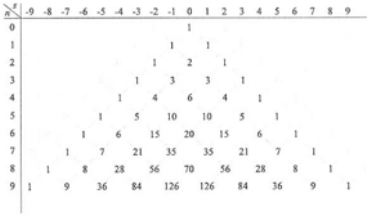
Deska Galtona

- Jeżeli prawdopodobieństwo skoku w prawo lub w lewo na każdym gwoździu jest takie samo, to prawdopodobieństwo rozkładu na ostatnim poziomie dane jest przez trójkąt Pascala
- Program **Galton**



Trójkąt Pascala

Trójkąt Pascala powstaje przy obliczaniu n-tej potęgi dwumianu. Każdy współczynnik w trójkącie Pascala równy jest liczbie dróg jakimi można do niego dojść



Współczynniki dwumianu

- Liczby w n-tym wierszu trójkąta Pascala, zwane współczynnikami dwumianowymi, oznaczane są symbolem Newtona i dane są wzorem:

$$\binom{n}{k}=\frac{n!}{k!(n-k)!}$$

- Zatem prawdopodobieństwa trafienia do odpowiedniej przegródki wynoszą

$$p(n,k)=\frac{n!}{2^n k!(n-k)!}$$

- Zachodzi

$$\sum_{k=0}^{n} p(n,k)=1$$

Błądzenie przypadkowe

- Ruch punktu na prostej lub w przestrzeni o dowolnym wymiarze polegający na wykonywaniu kroków o stałej długości losowo w jednym z kilku wybranych kierunków.
- Deska Galtona jest równoważna błądzeniu przypadkowemu na prostej.

Inne źródła liczb losowych

- Tabele liczb losowych, np. rozwinięcia liczb normalnych
- Pomiary fizyczne odpowiednich układów

Spacery losowe – model dyfuzji

- Błądzenie przypadkowe (spacer losowy) w przestrzeni stanowi model dyfuzji i ruchów Browna – rozprzestrzeniania się cząsteczek w danym środowisku.
- Średnia odległość od punktu początkowego rośnie z czasem t jak $\sqrt{t}$
- Program **Smoluchowski**

Liczby normalne ponownie

- Liczba normalna – liczba, w której rozwinięciu w danym układzie każdy blok cyfr jest tak samo prawdopodobny jak każdy inny blok tej samej długości
- Rozwijają liczbę normalną w bazie o podstawie 36 możemy generować losowe teksty: 10 cyfr + 26 liter łacińskich, albo 35 polskich liter i spacja.
- Przykład rozwinięcia liczby pi po polsku.

Prawdopodobieństwo wystąpienia dowolnego tekstu

- Trzydziestotomowa encyklopedia Brytannica zawiera około 30 milionów znaków. Prawdopodobieństwo wystąpienia jej tekstu w przypadkowym tekście wynosi
$$\left(\frac{1}{36}\right)^{30\,000\,000} \approx \left(\frac{1}{10}\right)^{45\,000\,000}$$
- Prawdopodobieństwo wystąpienia krótkich słów, jak „Budda” wynosi
$$\left(\frac{1}{36}\right)^{5} \approx \left(\frac{1}{60\,000\,000}\right)$$
czyli w tekście długości 60 mln znaków oba te słowa powinny wystąpić przynajmniej raz

Kodowanie tekstów w rozwinięciach

- Znajdowanie tekstów w rozwinięciach liczb normalnych nie tylko zależy od podstawy ale i od kodowania, tj. od tego co przypiszemy danemu symbolowi w rozwinięciu

Informacja i niepewność

- Matematyczna teoria informacji zajmuje się pojemnością kanału transmisji informacji, zupełnie abstrahuje od znaczenia, wartości i sensu przekazywanej informacji.
- Informacja i niepewność to dwie strony tego samego medalu: zdobywając informację usuwamy niepewność i na odwrót, tracąc informację powiększamy niepewność.
- Im większa niepewność co do poszukiwanego wyniku, tym więcej informacji zdobywamy poznając ten wynik.

Bit

- Miarą informacji jest **bit –** skrót od binary digit. Jest to miara informacji otrzymanej w odpowiedzi na **elementarne pytanie**, to jest pytanie na które odpowiedź może brzmieć tylko „tak” lub „nie”.
- Większe jednostki to **bajt**, **kilobajt**, **megabajt**, itd.
- UWAGA:
kilometr to 1000=10^3 metrów
kilobajt to 1024 = 2^{10} bajtów

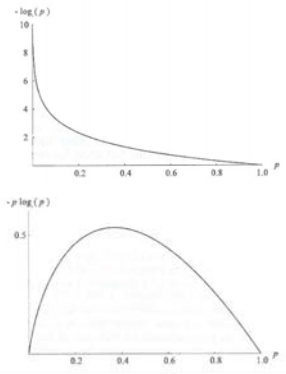
Wzór Shannona

- Zawartość informacyjna przekazu złożonego z *n* znaków wyrażona jest przez prawdopodobieństwa występowania tych znaków p_i

$$H = -\sum_{i=1}^{n} p_i \log_2 (p_i)$$

- Wielkość *H* oznacza informację mierzoną w bitach. Nazywa się ją **entropią informacyjną**.

P log(P)



Własności informacji

- Informacja, jaka może być zawarta w danym ciągu znaków jest proporcjonalna do długości tego ciągu. (Informacja jest wielkością **ekstensywną**)
- Przyjmijmy, że informacja jest zapisana w alfabecie binarnym (0,1)
- Słowem binarnym jest ciąg zer i jedynek o długości N. Liczba N mierzy objętość nośnika informacji. Informacja zawarta w słowie jest proporcjonalna do N.

Dlaczego logarytm

- Wybieramy jednostkę informacji tak, że

  Informacja H = Długość_słowa_binarnego

  Przy takim wyborze słowo złożone z jednego znaku niesie jeden bit informacji.

- Istnieje 2^N słów binarnych o długości N znaków. Zatem

  Długość_słowa = $\log_2$ (Liczba_słów)

Dlaczego logarytm cd

- Jeżeli prawdopodobieństwo wystąpienia każdego słowa jest takie samo, to wynosi ono

  p=1/Liczba_słów

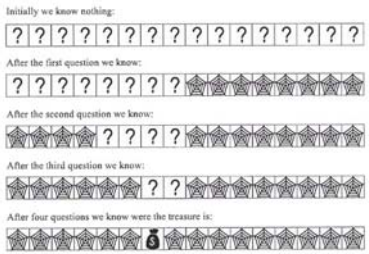
- Wobec tego informacja zawarta w pojedynczym słowie wynosi

  H = - $\log_2$ (p)

Gra w 20 pytań

- Miara informacji jest równa liczbie pytań potrzebnych do odgadnięcia słowa
- Rozważmy uproszczoną sytuację, kiedy jest 2^N równoprawdopodobnych słów o długości N. Ponumerujmy je wszystkie liczbami naturalnymi od 1 do 2^N.
- Mamy 2^N zakrytych komórek. W jednej z nich jest „skarb”. Znalezienie „skarbu” jest tym samym co odgadnięcie słowa.

Najprostsza strategia



20 pytań cd

- Liczba pytań potrzebna do uzyskania pełnej informacji równa jest początkowej niepewności
- Na ogół prawdopodobieństwa wystąpienia różnych słów nie są takie same.
- Kiedy mamy odgadnąć słowo postaci KU_A nie wiemy, czy jest to KUFA, KULA, KUMA, KUNA, KUPA czy KURA.
- Kiedy mamy słowo postaci ŚW_T, to nie ma problemu.

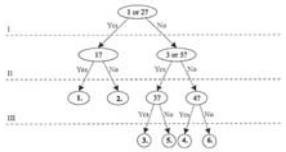
Rozkład 1

- Rozważmy skarb ukryty w jednej z 4 komórek, z prawdopodobieństwami $p_1 = 0.5$, $p_2 = 0.25$, $p_3 = 0.125$, $p_4 = 0.125$
- Pierwotna strategia daje średnio 2 pytania do osiągnięcia sukcesu
- Lepsza strategia:
  - Czy skarb jest w pierwszej komórce?
  - Czy skarb jest w drugiej komórce?
  - Czy skarb jest w trzeciej komórce?
- Średnio prowadzi ona do
  $1*0.5+2*0.25+3*0.25=7/4 < 2$ pytań
  zatem jest lepsza od bisekcji.

Rozkład 2

- Rozważmy rozkład 6 komórkowy:
  $p_1 = 1/3$, $p_2 = 1/5$, $p_3 = 1/5$,
  $p_4 = 2/15$, $p_5 = 1/15$, $p_6 = 1/15$

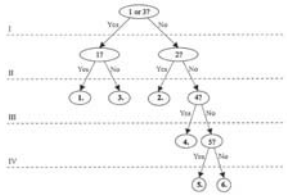
Średnia liczba pytań wynosi tu 37/15



Rozkład 2 – druga strategia

- Rozważmy rozkład 6 komórkowy:
  $p_1 = 1/3$, $p_2 = 1/5$, $p_3 = 1/5$,
  $p_4 = 2/15$, $p_5 = 1/15$, $p_6 = 1/15$

Średnia liczba pytań wynosi tu 36/15



Optymalna strategia – algorytm Huffmana

- Z początkowego rozkładu $p^0_1, p^0_2, \ldots, p^0_n$ wybieramy dwa najmniej prawdopodobne zdarzenia p^0_i oraz p^0_j.
- Łączymy je w jedno o prawdopodobieństwie p^1_k, mamy nowy rozkład $p^1_1, p^1_2, \ldots, p^1_{n-1}$
- Powtarzamy procedurę n-1 razy
- W ten sposób, od dołu, powstaje optymalne drzewo pytań.

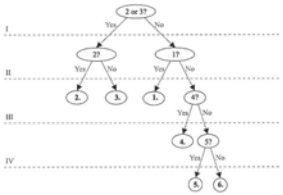
Przykład działania strategii

- Zaczynamy od rozkładu p^0_1=1/3, p^0_2=1/5, p^0_3=1/5, p^0_4=2/15, p^0_5=1/15, p^0_6=1/15
- Łączymy p^0_5 i p^0_6 w p^1_5. Dostajemy: p^1_1=1/3, p^1_2=1/5, p^1_3=1/5, p^1_4=2/15, p^1_5=2/15
- Łączymy p^1_4 i p^1_5 w p^2_4. Dostajemy: p^2_1=1/3, p^2_2=1/5, p^2_3=1/5, p^2_4=4/15
- Łączymy p^2_2 i p^2_3 w p^3_3. Dostajemy: p^3_1=1/3, p^3_2=4/15, p^3_3=6/15
- Łączymy p^3_1 i p^3_2 w p^4_2. Dostajemy: p^4_1=9/15, p^4_2=6/15

Rozkład 2 – trzecia strategia

- Rozważmy rozkład 6 komórkowy:
  p_1 = 1/3, p_2 = 1/5, p_3 = 1/5,
  p_4 = 2/15, p_5 = 1/15, p_6 = 1/15

Średnia liczba pytań wynosi tu również 36/15



Porównanie dwóch ostatnich strategii

- Rozważmy rozkład 6 komórkowy:
  p_1 = 1/3, p_2 = 1/5, p_3 = 1/5,
  p_4 = 2/15, p_5 = 1/15, p_6 = 1/15

Średnia liczba pytań wynosi dla obu strategii 36/15

Średnia informacja

- Nasze rozważania pokazują, że entropia Shannona mierzy średnią informację obliczoną w przypadku, gdy znane są wszystkie prawdopodobieństwa elementarne.
- Ogólne twierdzenie o bezszumowym kodowaniu możemy sformułować tak:

  Nie istnieje strategia o średnio mniejszej liczbie pytań niż entropia Shannona

Doświadczenia Hymana

- R. Hyman pokazał, że czas reakcji na bodźce o określonej zawartości informacji jest proporcjonalny do entropii Shannona.
- Program **Hyman**

Użyte programy

- Ulam
- Bertrand
- Bernoulli
- Galton
- Poe
- Smoluchowski
- Shannon
- Huffman
- Hyman