Markow rządzi (nawet buldogami pod dywanem)

Rafał SZTENCEL

ORORR ROROR OROOO RRROO ORROR
OROOR ROROO ROORO RRORR RRROR
OOORR OOOOO RRORR ROORO RORRO
OOOOR ROROO OOORR RRORR ROORR

OOROR ROROO RRROO ROORR ORROR
ROORO RORRR ORROO RRORO RRORR
OOROR ROOOR RORRO RRORR OOROR
OOROR RRORO ORROR OOROR RORRO

Czy można rozpoznać, który z widniejących na marginesie ciągów orłów i reszek powstał w wyniku rzutów monetą, a który został napisany przez człowieka? Nie jest to trudne, bowiem ludzie mają na ogół mylne wyobrażenie o losowości. Nie każdy wie, że w ciągu 100 rzutów monetą bardzo często pojawia się seria pięciu jednakowych wyników. Jeszcze skuteczniejszej metody dostarczają łańcuchy Markowa. Jeśli potraktować pojedyncze symbole jako stany procesu, to w przypadku ciągu losowego wszystkie prawdopodobieństwa przejścia p_{OO}, p_{OR}, p_{RO} i p_{RR} powinny być równe $\frac{1}{2}$. Tymczasem zbadanie 250 ciągów o długości 50, wygenerowanych przez respondentów autora [2] daje

$$p_{\mathrm{OR}} \approx p_{\mathrm{RO}} \approx 0{,}57, \quad p_{\mathrm{OO}} \approx p_{\mathrm{RR}} \approx 0{,}43.$$

Już tu widać wzruszająco naiwne próby przywracania równowagi; a gdy za stany procesu przyjmiemy dwa sąsiednie symbole (np. dla ciągu OOROR kolejnymi stanami będą OO, OR, RO, OR), to szansa przejścia z OO do OO wyniesie tylko 36%, za to z RR do RO – 59%, a z OO do OR aż 64%.

Krótki spacer po Internecie i lektura znalezionej tam pracy [3] prowadzą do wniosku, że niewiele jest dziedzin działalności ludzkiej, gdzie nie stosuje się łańcuchów Markowa. Za pomocą tak zwanych ukrytych modeli Markowa można rozpoznawać znaki (IBM, 1967), twarze, zagrania w tenisie, a nawet przewidywać wydarzenia polityczne. My ograniczymy się do krótkiego przedstawienia modelu, który pozwala prognozować czas pomiędzy kolejnymi erupcjami gejzeru Old Faithful (drugiej, po misiu Yogi, atrakcji parku Yellowstone).

Errata. Prof. Wojciech Guzicki zwrócił mi uwagę na przykry i poważny błąd w artykule *Kod Huffmana*, (*Delta* 1/2007). Zdanie „Pierwszy sekretarz KC PZPR, Edward Gierek,...” powinno brzmieć „Pierwszy Sekretarz KC PZPR, towarzysz Edward Gierek,...”.

Przepraszam. RS

Azzalini i Bowman [1] stwierdzili, że zarówno czasy trwania erupcji, jak i odstępy pomiędzy nimi, dzielą się na „długie” i „krótkie” (odpowiednie histogramy mają dwa ostre i wyraźnie rozdzielone maksima). Co więcej, długość erupcji i czas oczekiwania na następną są silnie skorelowane.

Pierwszą składową modelu jest łańcuch Markowa o macierzy przejścia

$$\begin{bmatrix} 0{,}00 & 1{,}00 \\ 0{,}83 & 0{,}17 \end{bmatrix}$$

który opisuje wewnętrzny stan układu, bezpośrednio nieobserwowalny. Zauważmy, że po stanie „krótkim” musi nastąpić „długi”. Z formalnego punktu widzenia łańcuch Markowa jest ciągiem zmiennych losowych X_0, X_1, $X_2, \ldots$ o wartościach w zbiorze $\{d, k\}$. Jeśli znamy wartość zmiennej losowej X_n, to zmienną Y_n (także o wartościach w zbiorze $\{d, k\}$) otrzymujemy za pomocą losowania, tak by

$$P(Y_n = d | X_n = k) = 0{,}23, \quad P(Y_n = d | X_n = d) = 1.$$

Drugą składową modelu stanowi zatem wektor warunkowych prawdopodobieństw wylosowania stanu d: $[0{,}23; 1]$. Zmienne losowe Y_n są obserwowane bezpośrednio i służą do estymacji parametrów modelu, podczas gdy zmienne X_n opisują tytułową walkę buldogów pod dywanem.

Prognoza zachowania się gejzeru jest niezwykle ważna z punktu widzenia turystów. Pod `www.nps.gov/archive/yell/oldfaithfulcam.htm` można obejrzeć bieżący stan gejzeru. Od 20 grudnia 2006 miała być dostępna prognoza czasu najbliższej erupcji.

Łańcuch Markowa można krótko opisać jako proces, w którym przyszłość zależy tylko od teraźniejszości, a nie od przeszłości. W języku zmiennych losowych o wartościach w zbiorze stanów $\{s_0, s_1, s_2, \ldots\}$ wyraża się to tak: dla każdego $n = 1, 2, \ldots$

$$P(X_n = s_n | X_{n-1} = s_{n-1}) = P(X_n = s_n | X_{n-1} = s_{n-1}, X_{n-2} = s_{n-2}, \ldots, X_0 = s_0),$$

jeśli tylko prawdopodobieństwo warunkowe po prawej stronie jest dobrze określone. Trudno nie zadać pytania: czy ciąg (Y_n) jest łańcuchem Markowa?

Literatura

[1] A. Azzalini, A. W. Bowman, *A look at some data on the Old Faithful geyser*, Appl. Statist. 39 (1990), 357–365.

[2] P. Pacewicz, *Jak odróżnić ciąg losowy od nielosowego? Przykład zastosowania sieci neuronowej*, Praca magisterska, WMiM UW, Warszawa 2002.

[3] P. W. Zwiernik, *Ukryte modele Markowa w analizie danych dotyczących koniunktury gospodarczej*, Praca magisterska, SGH, Warszawa 2003.