Łańcuchy Markowa

Paweł Rychlikowski

Instytut Informatyki UWr

3 grudnia 2018

Kasyno

- Na pewnym stole w kasynie gra polega na obstawianiu wyników rzutu kością.
- Krupier (kostera?) jest nieuczciwy i ma dwa egzemplarze kości:
  - Standardowy: $(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$
  -
  - Oszukany: $(\frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{2})$ (czyli dużo większa szansa wypadnięcia **szóstki**)

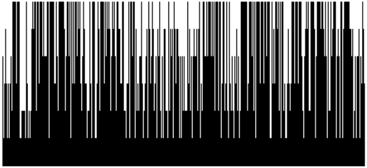
  Podmiana kości jest ryzykowna, zatem robi się ją z niewielkimi prawdopodobieństwami (równymi $\mathbf{p_0}$ i $\mathbf{p_1}$) (większość rzutów jest poprzednio użytą kością)

Zadanie

Widząc wyniki rzutów powiedzieć, kiedy gra jest uczciwa, a kiedy oszukana.

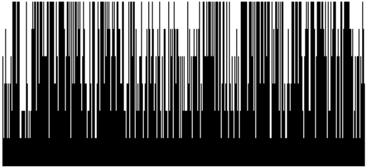
Kasyno (2)

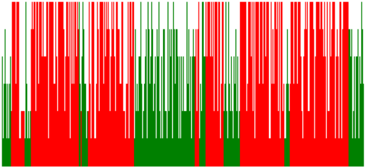
Popatrzmy na wyniki 300 rzutów kością na feralnym stoliku (wraz z wyjaśnieniem, czyli z informacją, jaka kość była użyta)



Kasyno (2)

Popatrzmy na wyniki 300 rzutów kością na feralnym stoliku (wraz z wyjaśnieniem, czyli z informacją, jaka kość była użyta)



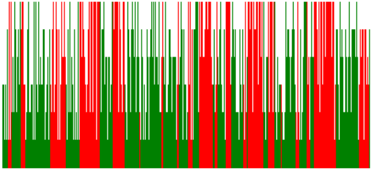


Kasyno (3)

Analogiczne wyniki, dwa razy większe prawdopodobieństwa zmiany kości (0.1 i 0.08)

Kasyno (3)

Analogiczne wyniki, dwa razy większe prawdopodobieństwa zmiany kości (0.1 i 0.08)



Kasyno. Kilka oczywistych spostrzeżeń.

1. Widzimy jedynie wyniki **obserwacji**, natomiast **stan** krupiera jest ukryty.
2. Wnioskować możemy tylko probabilistycznie (każdy wynik może być wynikiem obu kostek)
3. Łatwo zaproponować heurystyczne rozwiązanie (zielone wtedy, jak dużo szóstkek w niewielkim przedziale) – ale czy będzie ono optymalne (i co to znaczy „optymalne")?

Pytania do detektywa

1. Jakie jest p_0 i p_1?
2. Jaką kością rzucał krupier w $t = 146$ (wtedy nasz klient przegrał milion)
3. Jaki ciąg stanów (kolorów) najlepiej wyjaśnia sekwencję obserwacji

Łańcuch Markowa

Sytuacja w kasynie jest szczególnym przypadkiem Ukrytych Łańcuchów Markowa (Hidden Markov Model, HMM)

Uwaga

Tagowanie bigramowwe, o którym wcześniej mówiliśmy również

Formalnie: mamy sekwencję zmiennych losowych $X_1, \dots, X_T$ (rodzajów kości) przyjmujących wartości ze skończonego zbioru. Własności Markowa są dwie:

- Ograniczonego Horyzontu:

$$P(X_{t+1} = s | X_1, \dots, X_t) = P(X_{t+1} = s | X_t) = \dots$$

- Stacjonarności

$$\dots = P(X_2 = s | X_1)$$

Łańcuch Markowa (2)

Powiedzmy, że zbiór stanów to $\{s_1, \ldots, s_k\}$. Aby zadać łańcuch Markowa musimy

- znać tablicę „przejść” między stanami, A, taką że:

$$a_{ij} = P(X_{t+1} = s_j | X_t = s_i)$$

- Dodatkowo od jakiegoś stanu powinniśmy wyruszyć, czyli potrzebujemy

$$\pi_i = P(X_1 = s_i)$$

Łańcuchy Markowa używamy wtedy, gdy mamy „liniową sekwencję zdarzeń”. A wiele sytuacji językowych jest czymś takim.

Hidden Markov Model

HMM określamy jako graf, w którym zapisujemy tabelkę a_{ij} oraz określamy na każdej krawędzi tego, co podczas tego przejścia jest emitowane, czyli

$$b_{ijk} = P(O_t = k | X_t = s_i, X_{t+1} = s_j)$$

(rysunek na tablicy, emisja w stanie, bądź na krawędzi)

Zastosowania HMM-ów

- Tagowanie: obserwacje to słowa, stany ukryte – tagi (uwaga: $b_{ijk} = b_{i*k}$)
- Korekta pisowni: obserwacje to słowa zniekształcone, stany to słowa prawdziwe.
- Rozpoznawanie mowy: obserwacje to przetworzony wave, stany to fonemy (lub ich części)
- Tłumaczenie: obserwacje to słowa w języku A, stany to słowa w języku B (jak poradzić sobie z różną liczbą słów?).
- W przyszłości zastanowimy się jeszcze nad rozbiorem zdań wykorzystującym HMM.

Trzy pytania dla HMMów

1. Mając dany model $\mu = (A, B, \pi)$, chcemy efektywnie obliczać $P(O|\mu)$.
2. Mamy obserwacje oraz μ. Pytanie: jaka sekwencja stanów najlepiej wyjaśnia tę obserwację
3. Mamy sekwencję obserwacji i przestrzeń modeli, interesuje nas najlepszy model.

Trzy pytania dla HMMów (2)

Te pytanie mają zastosowania praktyczne

1. Możemy wybrać pomiędzy dwoma modelami dla danej sekwencji obserwacji.
2. Najbardziej typowy scenariusz, czyli na przykład tagowanie, etc.
3. Tworzeni modeli HMM w przypadku braku otagowanego korpusu.

Wracając do pozytywistów... (1)

- Zadanie POS: Prus, Orzeszkowa, Sienkiewicz
- A może Sienkiewicz lubi coś takiego:

  Pan Wołodyjowski walnął, sieknął i dźgnął Tatara, a ów zaskowyczał, zadrgał i umarł.

- A może Orzeszkowa lubi coś takiego:

  Nad Niemnem rozpościerały się cudowne kwieciste łąki, kontrastujące z głęboką granatową wodą cicho płynącej chłodnej rzeki.

Wracając do pozytywistów... (1)

- Zadanie POS: Prus, Orzeszkowa, Sienkiewicz
- A może Sienkiewicz lubi coś takiego:

  Pan Wołodyjowski walnął, sieknął i dźgnął Tatara, a ów zaskowyczał, zadrgał i umarł.

- A może Orzeszkowa lubi coś takiego:

  Nad Niemnem rozpościerały się cudowne kwieciste łąki, kontrastujące z głęboką granatową wodą cicho płynącej chłodnej rzeki.

Prawdopodobieństwo w modelu i tagowanie

- Można wywołać algorytm tagowania, a następnie policzyć $P(T)$.
- Czy to jest poprawne rozwiązanie?

Niekoniecznie, bo

... wydaje się, że między $P(T_{\text{opt}})$, $P(W)$ oraz $P(W|T_{\text{opt}})$ mogą być jakieś różnice!
Wg którego powinniśmy wybierać?

Odpowiedź: **P(W)**
Przypominamy, że $P(W|\mu)$, gdzie $\mu \in \{\text{Prus}, \text{Sienkiewicz}, \text{Orzeszkowa}\}$ to prawdopodobieństwo ciągu słów, czyli w języku HMM-a ciągu obserwacji w danym modelu.

Znajdywanie prawdopodobieństwa obserwacji

To jest **pytanie 1**, które pasuje do zadania z Sienkiewiczem

Mamy ciąg stanów X.

$$P(O|X,\mu) = \prod_{t=1}^{T} P(O_t|X_t, X_{t+1}, \mu) = b_{x_1 x_2 o_1} b_{x_2 x_3 o_2} \dots b_{x_T x_{T+1} o_T}$$

ponadto

$$P(X|\mu) = \pi_{x_1} a_{x_1 x_2} a_{x_2 x_3} a_{x_T x_{T+1}}$$

Ze wzoru Bayesa mamy

$$P(O, X|\mu) = P(O|X,\mu)P(X|\mu)$$

Zatem sumując po wszystkich sekwecjach X otrzymamy

$$P(O|\mu) = \sum_{(X_1,\dots,X_{T+1})} \prod_{t=1}^{T} a_{x_t x_{t+1}} b_{x_t x_{t+1} o_t}$$

Znajdywanie prawdopodobieństwa obserwacji (2)

Wzór z poprzedniego slajdu:

$$P(O|\mu) = \sum_{(X_1,\ldots,X_{T+1})} \prod_{t=1}^{T} a_{x_t x_{t+1}} b_{x_t x_{t+1} o_t}$$

przypomina to to, co obliczaliśmy dla optymalnej sekwencji tagowania (tylko z sumą zamiast argmax.

Uwaga

argmax i sum to nie jest to samo i obliczając prawdopodobieństwo powinniśmy uzględniać wszystkie sekwencje, a nie tylko najlepszą

Znajdywanie prawdopodobieństwa obserwacji (3)

- Wzór z poprzedniego slajdu zakłada sumowanie po wykładniczo wielu sekwencjach.
- Oczywiście nie chcemy takiego algorytmu
- Obliczamy sekwencję α:

$$\alpha_i(t) = P(o_1 \ldots o_{t-1}, X_t = i|\mu)$$

Które zawierają prawdopodobieństwo, że w danym kroku będziemy w i-tym stanie i że zobaczyliśmy do tego czasu konkretną sekwencję obserwacji.

Algorytm obliczania prawdopodobieństwa obserwacji

Daje to wzory:

- Inicjalizacja:$\alpha_i(1) = \pi_i$
- Krok indukcyjny:

$$\alpha_j(t+1) = \Sigma_{i=1}^{N} \alpha_i(t) a_{ij} b_{ijo_t}$$

- Pozostaje wysumować po stanach:

$$P(O|\mu) = \sum_{i=1}^{N} \alpha_i(T+1)$$

Algorytm Viterbiego (revisited)

Znajduje sekwencję stanów, która najlepiej odpowiada obserwowanym danym

W zasadzie to już było, traktujmy więc najbliższe informacje jako przypomnienie.
Szukamy

$$\mathrm{argmax}_X P(X|O,\mu)$$

Ponieaż O jest ustalone, zatem szukamy

$$\mathrm{argmax}_X P(X,O,\mu)$$

Algorytm Viterbiego (cd)

Znowu będziemy używać algorytmu dynamicznego. Tym razem δ:

$$\delta_j(t) = \max_{X_1 \dots X_{t-1}} P(X_1 \dots X_{t-1}, o_1, \dots, o_{t-1}, X_t = j|\mu)$$

Czyli jakie jest prawdopodobieństwo najbardziej prawdopodobnej ścieżki, która nas tu doprowadziła.
Dodatkowo powinniśmy zapamiętać informacje o stanach, które znajdują się na optymalnej ścieżce.

Trzy etapy algorytmu

- $\delta_j(1) = \pi_j$
- Indukcja:

$$\delta_j(t+1) = \max_{i=1,\ldots,N} \delta_i(t) a_{ij} b_{ijo_t}$$

- Koniec + odtwarzanie ścieżki:

$$P(\hat{X}) = \max_i \delta_i(T+1)$$

Obliczenia wsteczne

Prawdopodobieństwo możemy też liczyć „z drugiej strony". Zmienne wsteczne β, określamy następująco:

$$\beta_i(t) = P(o_t \ldots o_T | X_t = i)$$

Czyli jest to prawdopodobieństwo tego, że zobaczymy resztę obserwacji, jeżeli w chwili t będziemy w stanie i.

Obliczenia wsteczne (2)

Między α i β jest różnica:

$$\alpha_i(t) = P(o_1 \dots o_{t-1}, X_t = i)$$

$$\beta_i(t) = P(o_t \dots o_T | X_t = i)$$

Obliczenia wsteczne (3)

$\beta_i(t) = P(o_t \dots o_T | X_t = i)$

Daje to nam następujące wzory:

- $\beta_i(T+1) = 1$ (bo pusta sekwencja obserwacji i pusta koniunkcja)
- $$\beta_i(t) = \sum_{j=1}^{N} a_{ij} b_{ijo_t} \beta_j(t+1)$$
- W sumie mamy: $P(O|\mu) = \sum \pi_i \beta_i(1)$

Obliczenia wsteczne (4)

Współczynniki α i β można połączyć i otrzymamy:

$$\alpha_i(t)\beta_i(t) = P(o_1 \dots o_{t-1}, X_t = i)P(o_t \dots o_T | X_t = i)$$

A zatem

$$P(O|\mu) = \sum_{i=1}^{N} \alpha_i(t)\beta_i(t)$$

dla każdego t.

$P(O|\mu)$ umieliśmy policzyć już wcześniej. Ale współczynniki α i β się przydadzą również do innych celów.

Najlepszy ciąg stanów

To tak naprawdę są dwa różne zadania. Może nas interesować

- ciąg najlepszych stanów
- najlepszy ciąg stanów (o tym mówiliśmy przy tagowaniu)

Które zadanie jest tym właściwym?

Uwaga

Częściej interesuje nas wyjaśnienie **całej historii**, niż skupienie się na jednym jej momencie (milionowa strata w punkcie T=146).

Przypominamy:

$$\alpha_i(t) = P(o_1 \dots o_{t-1}, X_t = i|\mu)$$

$$\beta_i(t) = P(o_t \dots o_T|X_t = i|\mu)$$

Szacujemy prawdopodobieństwo bycia w stanie i w czasie t:

$$\begin{aligned}
\gamma_i(t) &= P(X_t = i|O, \mu) && (1)\\
&= \frac{P(X_t = i, O|\mu)}{P(O|\mu)} && (2)\\
&= \frac{\alpha_i(t)\beta_i(t)}{\sum_{i=1}^{N} \alpha_i(t)\beta_i(t)} && (3)\\
& && (4)
\end{aligned}$$

Ciąg najlepszych stanów

Możemy wybrać stan jako

$$\hat{X}_t = \mathrm{argmax}_i \gamma_i(t)$$

W ten sposób maksymalizujemy oczekiwaną liczbę prawidłowo zgadniętych stanów.
Ale sekwencja traktowana jako całość będzie (być może) taka sobie.

Tworzenie modelu od podstaw

- Obliczamy $\hat{a}_{ij}$ jako:

$$\frac{\text{oczekiwana liczba przejść ze stanu i do stanu j}}{\text{oczekiwana liczba przejść ze stanu i}}$$

- Liczymy prawdopodobieństwo przejścia w każdym momencie i wyciągamy średnią.
- Będziemy obliczać $\xi_t(i,j) = P(q_t = i, q_{t+1} = j | O, \mu)$
- Prawie-$\xi_t(i,j) = P(q_t = i, q_{t+1} = j, O | \mu) =$

$$\alpha_t(i) a_{ij} b_{io_{t+1}} \beta_{t+1}(j)$$

Tworzenie modelu od podstaw

Prawdziwe ξ_i:

$$\xi_t(i,j) = \frac{\text{Prawie-}\xi_t(i,j)}{P(O|\mu)}$$

Współczynniki a są średnią ξ po czasie, czyli:

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \sum_{j=1}^{N} \xi_t(i,j)}$$

Uwaga

To też jest algorytm **EM**! (liczymy lepsze a przy założeniu starych a). Czym się różni od tego z poprzedniego wykładu?

Różne oblicza EM

- W poprzednim wykładzie znajdywaliśmy optymalną sekwencję tagów i dla niej liczyli nowe statystyki.
- Tu obliczenia są w pewnym sensie **rozmyte** (nie decydujemy się na najlepszą, lecz sprawdzamy wszystkie, przypisując im odpowiednie prawdopodobieństwa)