

Markovkedjor och kortblandningar

Johan Jonasson

October 2020

Markovkedja: En glömsk groda som hoppar mellan näckrosblad.
En stokastisk process

$$X_0, X_1, X_2, \dots$$

där X_t är tillståndet (näckrosbladet) vid tidpunkt t . Det gäller att $X_t \in S$, där S är mängden av möjliga tillstånd (mängden av näckrosblad). Vi antar här att S är ändlig.

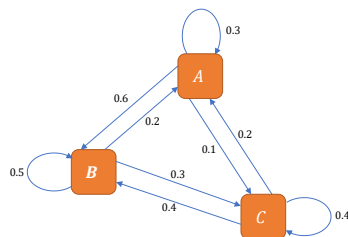
Sannolikheten att hoppa till ett visst tillstånd beror bara på var man är nu och inte hur man kom dit (grodan är glömsk). Formellt definieras denna egenskap, den s.k. Markovegenskapen, som att

$$\mathbb{P}(X_{t+1} = s | X_t = s_t, X_{t-1} = s_{t-1}, \dots, X_0 = s_0) = \mathbb{P}(X_{t+1} = s | X_t = s_t),$$

för alla $t, s, s_t, \dots, s_0$. Om man inte säger annat så brukar man anta att Markovkedjan är tidshomogen, dvs

$$\mathbb{P}(X_{t+1} = s' | X_t = s) = \mathbb{P}(X_1 = s' | X_0 = s) \text{ för alla } t, s, s'.$$

Exempel.



Antag att X_0 väljs slumpmässigt så att $\mathbb{P}(X_0 = s) = p_0(s)$, $s \in \{A, B, C\}$. Då gäller att

$$p_1(A) = \mathbb{P}(X_1 = A) = 0.3p_0(A) + 0.2p_0(B) + 0.2p_0(C)$$

$$p_1(B) = \mathbb{P}(X_1 = B) = 0.6p_0(A) + 0.5p_0(B) + 0.4p_0(C)$$

$$p_1(C) = \mathbb{P}(X_1 = C) = 0.1p_0(A) + 0.3p_0(B) + 0.4p_0(C).$$

Kan vi bestämma $p_0(s)$:en så att $p_1(s) = p_0(s)$ för $s \in \{A, B, C\}$?
Javisst, byt bara ut $p_1(s)$ mot $p_0(s)$ i ekvationssystemet och lös.
Man får då

$$p_0(A) \approx 0.222, p_0(B) \approx 0.494, p_0(C) \approx 0.284.$$

Om man väljer $p_0(s)$ enligt de siffrorna kommer det att gälla att $p_t(s) = p_0(s)$ för alla t och s (tack vare tidshomogenitet).
Om man inte väljer $p_0(s)$ så, så gäller ändå att $\mathbb{P}(X_t = s)$ konvergerar mot dessa tal (dvs 0.222, 0.494 respektive 0.284). Man brukar skriva

$$\pi(s) = \lim_{t \rightarrow \infty} p_t(s) = \lim_{t \rightarrow \infty} \mathbb{P}(X_t = s).$$

Generellt: Antag att S innehåller n element och kalla dessa $1, 2, \dots, n$. Skriv

$$\mathbf{p}_t = [p_t(1), p_t(2), \dots, p_t(n)].$$

Låt $P = [p_{ij}]$, där

$$p_{ij} = \mathbb{P}(X_1 = j | X_0 = i).$$

Då gäller

$$\mathbf{p}_{t+1} = \mathbf{p}_t P.$$

Förutom i speciella fall gäller att det finns en vektor $\pi = [\pi(1), \dots, \pi(n)]$ sådan att

$$\lim_{t \rightarrow \infty} \mathbb{P}(X_t = s) = \pi(s),$$

för alla s , oavsett hur man startar. Vektorn π ges då av motsvarigheten till ekvationssystemet ovan:

$$\pi P = \pi.$$

Kortblandningar: En kortlek med n kort blandas. S = mängden av alla $n!$ permutationer av de n korten. Uppenbarligen en Markovkedja; ordningen efter nästa steg i blandningen beror bara av hur korten ligger nu.

Om omordningen som sker av kortleken i ett blandningssteg inte beror av hur korten ligger nu, så får man alltid

$$\pi(s) = \frac{1}{n!}.$$

Om omordningen innehåller någon slump (dvs inte varannankortsblandning eller något liknande) gäller dessutom

$$\lim_{n \rightarrow \infty} \mathbb{P}(X_t = s) = \pi(s).$$

En kortlek är välblandad när $\mathbb{P}(X_t = s) \approx \pi(s)$ för (nästan) alla s .
Vad betyder det?

Definition

Totalvariansavståndet mellan fördelningen för X_t och π ges av

$$d(X_t, \pi) = \max_{U \subseteq S} |\mathbb{P}(X_t \in U) - \pi(U)| = \frac{1}{2} \sum_{s \in S} |\mathbb{P}(X_t = s) - \pi(s)|.$$

$d(X_t, \pi) < 1$ och $d(X_0, \pi)$ är väldigt nära 1 om X_0 är en fix startordning. Man brukar bestämma ett $\epsilon > 0$ och vara nöjd när t är så stort att $d(X_t, \pi) < \epsilon$.

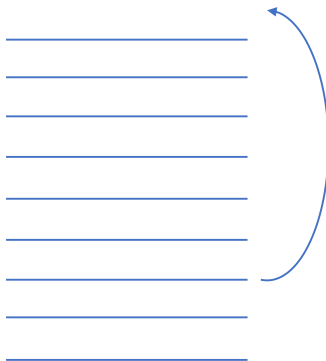
Hur stort t som behövs brukar man uttrycka i termer av n , dvs hur många kort som finns i leken. Oftast vill man ha en övre skattning på hur stort t behöver vara.

Koppling: Låt $Y_0, Y_1, \dots$ vara en annan kortlek som blandas på samma sätt som X -leken, med skillnaden att den är stationär redan från början: $\mathbb{P}(Y_0 = s) = 1/n!$ för alla s . Konstruera ett beroende mellan blandningarna för X -leken och Y -leken. Beroendet måste vara sådant att lekarna var för sig uppfyller den blandningsteknik som vi vill studera. Vi kräver också att $X_t = Y_t \Rightarrow X_{t'} = Y_{t'}$ för alla $t' > t$.

Låt $\tau = \min\{t : X_t = Y_t\}$. Då gäller

$$\begin{aligned} d(X_t, \pi) &= d(X_t, Y_t) = \max_U |\mathbb{P}(X_t \in U) - \mathbb{P}(Y_t \in U)| \\ &= \max_U |\mathbb{P}(X_t \in U, \tau \leq t) - \mathbb{P}(Y_t \in U, \tau \leq t) \\ &\quad + \mathbb{P}(X_t \in U, \tau > t) - \mathbb{P}(Y_t \in U, \tau > t)| \\ &\leq \mathbb{P}(\tau > t). \end{aligned}$$

Random-to-top shuffle: Ett steg i blandningen består av att ta ett kort på måfå och lägga det på toppen av leken.



Koppling: Välj nu hela tiden samma kort i X -leken och i Y -leken (dvs om man väljer ruter knekt i den ena så tar man ruter knekt i den andra också). Denna koppling uppfyller kraven.

Observera att så fort ruter knekt blivit vald hamnar den på samma position i de båda lekarna och kommer att förbli det (även om den rör sig). Därför gäller

$$\tau \leq T_0 + T_1 + \dots + T_{n-1}$$

där T_i = tiden från det att i olika kort valts tills $i + 1$ olika kort har valts.

När vi väntar på T_i har vi chansen $(n - i)/n$ vid varje blandningssteg att ta ett kort som inte tagits förut. Därför borde T_i i genomsnitt bli $n/(n - i)$.

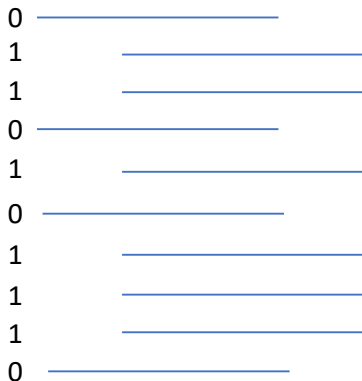
Alltså

$$\sum_{i=0}^{n-1} T_i \approx n \sum_{i=1}^n \frac{1}{i} \approx n \ln n.$$

Man kan visa att för $\delta > 0$ gäller

$$\lim_{n \rightarrow \infty} \mathbb{P}((1 - \delta)n \ln n < \tau < (1 + \delta)n \ln n) = 1.$$

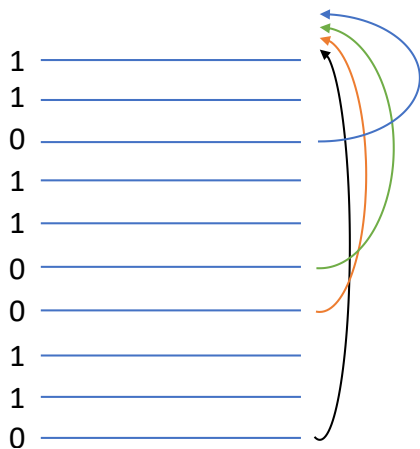
Detta är samma som idolbilsproblemet. Man får köpa $n \ln n$ idolbilder innan man fått alla.

The Riffle shuffle:

Man kan visa att om man kör en blandning "baklänges", så går det precis lika fort att blanda.

I det här fallet betyder det att dra isär korten, så att varje kort går uppåt med sannolikhet $1/2$ och nedåt med sannolikhet $1/2$ oberoende av andra kort.

För att modellera detta, ge varje kort märkning 0 m.s. $1/2$ och märkning 1 m.s. $1/2$. Dra sedan ut korten med märkning 0 och lägg dem ovanpå korten med märkning 1 utan att ändra deras inbördes ordning.



Varje kort får på det här sättet en märkning, 0 eller 1, i varje blandningssteg. Vid varje tidpunkt gäller att

- Kort som har 0 som sista märkning ligger över kort med 1 som sista märkning.
- För kort som har samma sista märkning gäller att kort som har 0 som näst sista märkning ligger över kort som 1 som näst sista märkning.
- För kort som har samma märkning de två sista stegen gäller att kort med 0 som tredje sista märkning ligger över kort som har 1 som tredje sista märkning, ...

Detta betyder att när alla kort har en unik följd av märkningar, kan vi läsa av ordningen på kortleken utan att ha någon information om hur de låg från början.

Koppling: Vid varje blandningssteg, låt varje kort få samma märkning i Y -leken som det får i X -leken. När alla kort har sin unika märkning kommer X -leken och Y -leken ligga i samma ordning.

Om man fixerar ett par (t.ex. ruter knekt och spader 10) av kort, så är sannolikheten att de har samma märkningsföljd efter t blandningssteg 2^{-t} . Det finns $n(n-1)/2$ par av kort, så sannolikheten att det finns två par som har samma märkningsföljd efter t steg är högst

$$\frac{n(n-1)}{2} \frac{1}{2^t}.$$

Detta uttryck går mot 0 då $n \rightarrow \infty$ om $t = (2 + \delta) \log_2 n$. Alltså

$$d(X_t, \pi) \leq \mathbb{P}(\tau > t) \rightarrow 0$$

då $n \rightarrow \infty$ om $t = (2 + \delta) \log_2 n$.

Bayer och Diaconis (1992) visade att i själva verket räcker $(3/2) \log_2 n$ steg.