Introduktion till Markovkedjor

Jeffrey E. Steif och Jan Westhall

1 Definition av Markovkedjor

En Markovkedja är en synnerligen viktig typ av stokastisk process som spelar mycket stor roll för tillämpningar inom sannolikhetsläran.

Vi ska för enkelhets skull först behandla Markovkedjor i diskret tid och därefter i kontinuerlig. Den senare typen får vi användning för i behandlingen av kösystem.

Här kommer en definition vars innebörd sedan diskuteras.

Definition: *En stokastisk process* $\{X_n\}_{n\geq 0}$ *med värden i en ändlig eller uppräkneligt oändlig mängd* S *sägs vara en Markovkedja i diskret tid om:*

$Prob\{X_n = i_n | X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}, \ldots, X_0 = i_0\} =$
$Prob\{X_n = i_n | X_{n-1} = i_{n-1}\}$.

Idén bakom den här definitionen är att den betingade fördelningen för X_n givet var processen har varit förut bara beror på det sista tillståndet. Det är en ganska naturlig förutsättning som med andra ord säger att givet nuet beror vad som ska hända i framtiden (slumpmässigt) endast på nuet och ej det förflutna.

Så snart man definerat begreppet ovan (Markovegenskapen) är det naturligt att införa en viss matris som kallas övergångsmatrisen och är definerad enligt följande.

En övergångsmatris är en $k \times k$ matris P (där k är antalet element i S, som ju kan vara uppräkneligt oändlig) med element, sk övergångssannolikheter,

$$P(i,j) = Prob\{X_n = j | X_{n-1} = i\}.$$

Om övergångssannolikheterna inte beror på n kallas kedjan tidshomogen, och det är den enda typ av Markovkedjor som vi ska syssla med. Observera att $P(i,j) \geq 0$ för alla i och j och att

$$\sum_{j \in S} P(i,j) = 1$$

för alla i.

Vilken matris som helst som uppfyller dessa båda villkor kan användas för att konstruera en Markovkedja.

Exampel 1: Slumpvandring på heltal:

Vi tänker oss en berusad man som bor i en endimensionell värld (reella linjen) och som irrar fram och tilbaka mellan heltalen på ett slumpmässigt sätt ett steg i sänder. Tillståndsrummet är därför $\mathbf{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$.

Beteckna med X_n tillståndet där mannen befinner sig vid tiden n. Vi antar att han startar i noll, dvs $X_0 = 0$, och sedan varje tidsenhet tar ett steg åt höger eller ett åt vänster, med sannolikheten $1/2$ i båda fallen, oberorende av sin tidigare framfart. Inför varje nytt steg kastar han så att säga ett symmetriskt mynt och tar ett steg åt höger eller vänster beroende på vilken sidan av myntet som kommer upp.

Låt därför $\{Z_n\}_{n \geq 1}$ vara en följd av oberoende och likafördelade stokastiska variabler med

$$Prob(Z_n = 1) = Prob(Z_n = -1) = 1/2$$

och sätt $X_n = \sum_{k=1}^{n} Z_k$.

Det borde vara ganska uppenbart att X_n så definerad faktiskt är en Markovkedja. Om mannen befinner sig i tillstånd k vid tiden n, kommer han ju antingen att vara i $k+1$ eller $k-1$ vid tiden $n+1$ med sannolikheter som ej beror på var han varit före tiden n.

En Markovkedja är nästan bestämd av sin övergångsmatris. Den enda vi behöver veta är var kedjan ska starta. Om vi anger detta och övergångsmatrisen är alltså processen bestämd. Vi kan därför säga att en Markovkedja är precis samma sak som en övergångsmatris tillsammans med ett begynnelsetillstånd eller mer allmänt en övergångsmatris tillsammans med en begynnelsefördelning (där vi väljer begynnelsetillståndet slumpmässigt i stället för deterministiskt).

2 Tvåstegssannolikheter

Övergångsmatrisen beskriver sannolikhetmässigt vad som kan hända i ett steg. Om man vill veta vad som kan hända på två steg, är det naturligt att betrakta

$$Prob\{X_n = j | X_{n-2} = i\}.$$

Det visar sig vara väldigt enkelt att beräkna sådana sannolikheter med hjälp av övergångsmatrisen P. Sannolikheten ovan blir nämligen $P^2(i,j)$ där P^2 är P multiplicerad med sig själv. Det kan man se så här:

$$Prob\{X_n = j | X_{n-2} = i\} = \frac{Prob\{X_n = j, X_{n-2} = i\}}{Prob\{X_{n-2} = i\}} =$$

$$\frac{\sum_{k \in S} Prob\{X_n = j, X_{n-1} = k, X_{n-2} = i\}}{Prob\{X_{n-2} = i\}} =$$

$$\frac{\sum_{k\in S} Prob\{X_{n-2}=i\}Prob\{X_{n-1}=k|X_{n-2}=i\}Prob\{X_n=j|X_{n-1}=k,X_{n-2}=i\}}{Prob\{X_{n-2}=i\}} =$$

$$\sum_{k\in S} Prob\{X_{n-1}=k|X_{n-2}=i\}Prob\{X_n=j|X_{n-1}=k,X_{n-2}=i\} =$$

(enligt Markovegenskapen)

$$\sum_{k\in S} Prob\{X_{n-1}=k|X_{n-2}=i\}Prob\{X_n=j|X_{n-1}=k\} =$$

$$\sum_{k\in S} P(i,k)P(k,j) =$$

$$P^2(i,j)$$

enligt definitionen av matrismultiplikation. Mera allmänt är det så (och läsaren uppmanas bevisa det själv) att $Prob\{X_n=j|X_{n-k}=i\}=P^k(i,j)$. Lägg märka till att P^k också är en övergångsmatris. Här är en enkel illustration:

Exempel 2:

Låt S ha endast 3 tillstånd, $\{1,2,3\}$, och låt $P(i,j)$ ges av

$$\begin{pmatrix} 1/3 & 1/6 & 1/2 \\ 1/4 & 1/3 & 5/12 \\ 1/2 & 1/4 & 1/4 \end{pmatrix}.$$

En enkel beräkning visar att

$$P^2 = \begin{pmatrix} 29/72 & 17/72 & 26/72 \\ 9/24 & 37/144 & 53/144 \\ 17/48 & 11/48 & 20/48 \end{pmatrix}.$$

Vi kan därför dra slutsatsen att t ex $Prob\{X_n=2|X_{n-2}=3\}=11/48$.

3 Stationära fördelningar

Ett viktigt begrepp i samband med Markovkedjor är stationära fördelningar. En sådan har egenskapen att om vi starta med den, ska den gälla hela tiden sedan också (dvs för varje X_n, $n \geq 1$).

Inledningsvis nöjer vi oss med att beräkna sannolikhetfördelningen i tiden 1 om vi börjar med en godtycklig sannolikhetfördelning π på S ($\pi(s) \geq 0$ för alla $s \in S$ och $\sum_{s \in S} \pi(s) = 1$.) Vi tänker oss att vi sätter igång processen genom att slumpmässigt välja begynnelsetillstånd enligt $\pi(s)$. Frågan är vilken sannolikhetfördelning vi då får efter ett steg.

$$Prob(X_1 = n) =$$

$$\sum_{k \in S} Prob(X_0 = k) Prob(X_1 = n | X_0 = k) = \sum_{k \in S} \pi(k) P(k, n).$$

Således: Om P är Markovkedjans övergångsmatris och π dess begynnelsefördelning, får vi fördelningen efter ett steg genom att multiplikera π (betraktad som vektor) med P.

OBS: Matrisen måste stå till höger vid multiplikationen.

I exempel 2 var $S = \{1, 2, 3\}$ och

$$P = \begin{pmatrix} 1/3 & 1/6 & 1/2 \\ 1/4 & 1/3 & 5/12 \\ 1/2 & 1/4 & 1/4 \end{pmatrix}.$$

Antag att begynnelsefördelningen är $\pi = (1/10, 2/5, 1/2)$. Då är fördelningen efter ett steg $\pi P = (46/120, 33/120, 41/120)$.

Vi kommer nu till ett mycket viktigt begrepp.

Definition: En fördelning på mängden S kallas stationär för en Markovkedja om man får igen den i steg ett när man använder den som begynnelsefördelning.

Av det ovan sagda borde framgå att definitionen innebär att om vi tar begynnelsefördelningen och multiplicerar den med övergångsmatrisen (med matrisen till höger) får vi tillbaka samma vektor: $\pi P = \pi$. Med andra ord: π är en (vänster) egenvektor med egenvärdet 1.

Hur ska vi beräkna den stationära fördelningen i Exempel 2? Vi måste hitta $\pi = (\pi(1), \pi(2), \pi(3))$ så att

$$1/3\pi(1) + 1/4\pi(2) + 1/2\pi(3) = \pi(1)$$

$$1/6\pi(1) + 1/3\pi(2) + 1/4\pi(3) = \pi(2)$$

$$1/2\pi(1) + 5/12\pi(2) + 1/4\pi(3) = \pi(3)$$

och

$$\pi(1) + \pi(2) + \pi(3) = 1.$$

Vi lämnar åt läsaren att lösa problemet och därvid lägga märka till att i detta fallet är det entydigt lösbart.

Om stationära fördelningar finns några viktiga satser, som vi nu presenterar och diskuterar men avstår från att bevisa.

Sats 1: Om S är ändlig finns alltid en stationär fördelning.

Det brukar bara finnas en enda men helt allmänt gäller ej detta resultat.

Definition: En Markovkedja kallas irreducibel om för varje x och y i S, $P^n(x, y) > 0$ för något n.

Detta är bara ett elegant sätt att säga att för varje par av tillstånd x och y ska man kunna komma från det ena till det andra förr eller senare.

Sats 2: Om S är ändlig och Markovkedjan är irreducibel, så finns bara en enda stationär fördelning.

(Detta var alltså fallet i Exempel 2.)

Om S är uppräkneligt oändlig gäller en motsvarande sats.

Sats 3: Om S är uppräkneligt oändlig och Markovkedjan irreducibel, så finns högst en stationär fördelning.

OBS: När S är uppräkneligt oändlig behöver det ej finnas någon stationär fördelning. Man kan visa att slumpvandringen som diskuterades tidigare inte har någon stationär fördelning.

Det är naturligt att hoppas att om en Markovkedja är irreducibel och har en stationär fördelning $\pi = \{\pi(s)\}_{s\in S}$, så är π också en asymptotisk fördelning på så vis att för alla x,

$$\lim_{n\to\infty} Prob\{X_n = y|X_0 = x\} = \pi(y)$$

för alla y. Tyvärr är det inte alltid så.

Example 3: Antag att $S = \{1, 2\}$ och

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Det är lätt att se att denna Markovkedja är irreducibel och att den entydigt bestämda stationära fördelningen är $(1/2, 1/2)$. Däremot gäller ej att

$$\lim_{n\to\infty} Prob\{X_n = y|X_0 = x\} = 1/2$$

eftersom $Prob\{X_n = y|X_0 = x\}$ alltid är 0 eller 1. Vår "punkt" flyttar sig bara deterministiskt fram och åter mellan de två tillstånden. Detta fenomen kallas periodicitet och måste tas speciell hänsyn till.

Emellertid, kan man i vissa fall dra slutsatsen att den stationära fördelningen också är asymptotisk.

Sats 4: Antag att vi har en Markovkedja som är irreducibel och har en stationär fördelningen $\{\pi(s)\}_{s\in S}$. Antag vidare att för alla x och y finns ett tal N så att $P^n(x,y) > 0$ för alla $n > N$. Då gäller för varje x att

$$\lim_{n\to\infty} Prob\{X_n = y|X_0 = x\} = \pi(y)$$

för alla y.

Den andra förutsättningen verker ganska komplicerad, men kan förenklas om S är en ändlig mängd. Då räcker det att kräva att för något n ska matrisen P^n ej ha något element som är noll.

I Exempel 2 är förutsättningarna i Sats 4 uppenbarligen uppfyllda. Vi får därför t ex att

$$\lim_{n\to\infty} Prob\{X_n = 2|X_0 = 1\} = \pi(2).$$

Några övningar:

1. Betrakta en Markovkedja med övergångsmatris

$$P = \begin{pmatrix} 2/3 & 1/3 \\ 2/3 & 1/3 \end{pmatrix}.$$

a) Beräkna P^2.

b) Vad är $Prob(X_7 = 2|X_5 = 2)$?

c) Beräkna den stationära fördelningen.

d) Beräkna $\lim_{n\to\infty} Prob(X_n = 2|X_0 = 1)$.

2. Betrakta en Markovkedja med övergångsmatris

$$\begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 2/3 & 1/3 & 0 \\ 1/3 & 0 & 2/3 \end{pmatrix}.$$

a) Är den irreducibel?

b) Beräkna den stationära fördelningen.

3. Betrakta en Markovkedja med övergångsmatris

$$\begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

a) Beräkna den stationära fördelningen.

b) Är den entydigt bestämd?

4 Kontinuerlig Tid

Vi betraktar nu en process i kontinuerlig tid men ska fortfarande anta att tillståndsrummet är ändligt eller uppräkneligt oändligt. En sådan process kallas en Markovkedja (i kontinuerlig tid) om följande motsvarighet till formeln i definitionen på sidan 1 gäller:

$$Prob\{X_t = j | X_s = i, X_r, 0 \le r \le s\} = Prob\{X_t = j | X_s = i\}.$$

Idén bakom den här definitionen är densamma som för diskret tid. Vad som ska hända i framtiden (slumpmässigt) beror givet nuet bara på detta och inte på det förflutna.

Det är lite svårare att representera sådana här processer med matriser. I det diskreta fallet, behövde vi bara en matris. Här behöver vi en familj av matriser, $P^t(x, y)$ för alla $t \ge 0$.

På ungefär samma sätt som i det diskreta fallet sätte vi $P^t(i, j) = Prob\{X_{s+t} = j | X_s = i\}$, dvs övergångssannolikheterna beror ej på s: även här betraktar vi tidshomogena processer.

Det finns ett annat och för oss mera användbart sätt att beskriva Markovkedjor i kontinuerlig tid.

Givet är dels en mängd ickenegativa reella tal, $\{q_i\}_{i \in S}$, ett för varje tillstånd, dels en övergångsmatris $P(i,j)$. Vi använder detta på följande sättet för att beskriva kedjan.

Om kedjan är i tillstånd i stannar den där en slumpmässig tid, som är exponentialfördelad med parameter q_i (lägg märke till att intensiteten beror på vilket tillstånd man befinner sig i). När den slumpmässiga tiden tillbringats i tillståndet i, sker ett hopp till j med sannolikhet $P(i,j)$.

Det är mycket enklare och naturligare att beskriva Markovkedjor i kontinuerlig tid på detta vis, som kan kopplas till matriserna P^t genom att man konstruerar en matris $Q(i,j)$ med diagonalelement $q_{ii} = -q_i$ där q_i gavs ovan. Sätt sedan $q_{ij} = q_i P(i,j)$ för $i \neq j$.

Q kallas intensitetsmatris. Man ska tänka på q_i som intensiteten för hur snabbt man lämnar tillstånd i och q_{ij} som intensiteten för hur snabbt man går från tillstånd i till tillstånd j. En annan förklaring till att q_{ij} kallas intensitet är att

$$Prob\{X_{t+h} = j | X_t = i\} = q_{ij} h + o(h)$$

där $\lim_{h \to 0} o(h)/h = 0$.

Det visar sig (något vi inte kommer att bevisa) att P^t är lika med e^{tQ}, där vi tolkar e^{tQ} med hjälp av serieutveckling, dvs $e^{tQ} = \sum_{n=0}^{\infty} \frac{(tQ)^n}{n!}$.

Stationära fördelningar definieras likadant som i det diskreta fallet. Givet någon begynnelsefördelning får man fördelningen vid tidpunkten t genom att (höger) multiplicera begynnelsefördelningen-uppfattad som en vektor-med matrisen P^t.

Definition: En fördelning på mängden S sägs vara stationär om den som begynnelsefördelning ger samma fördelning för processen i varje tidpunkt i framtiden.

Det bör vara klart att man också kan uttrycka detta så att om vi (höger) multiplicerar begynnelsefördelningen π med P^t får vi tillbaka samma vektor (fördelning): $\pi P^t = \pi$. Med andra ord: π är en (vänster) egenvektor till P^t med egenvärdet 1 för alla t.

Alla satser från det diskreta fallet gäller även här.

Sats 1: Om S är ändlig finns alltid en stationär fördelning.

Definition: En Markovkedja kallas irreducibel om för varje x och y i S, $P^t(x,y) > 0$ för något t.

(Man kan bevisa att om det finns ett t med $P^t(x,y) > 0$, så gäller detta för alla $t > 0$.)

Sats 2: Om S är ändlig och Markovkedjan är irreducibel, så finns bara en stationär fördelning.

Sats 3: Om S är uppräkneligt oändlig och Markovkedjan irreducibel, så finns högst en stationär fördelning.

OBS: Som i det diskreta fallet behöver det ej finnas någon stationär fördelning om S är uppräkneligt oändlig.

Kontinuerlig tid har en fördel framför diskret och det är att fenomenet periodicitet inte förekommer.

Sats 4: Om en Markovkedja är irreducibel och har en stationär fördelning $\{\pi(s)\}_{s \in S}$, så gäller för varje x att
$\lim_{t \to \infty} Prob\{X_t = y | X_0 = x\} = \pi(y)$ för alla y.

För att få exempel på markovkedjor i kontinuerlig tid behöver vi bara titta i boken, där alla processer vi hittills undersökt är av den typen.

Ta t ex födelse- och dödsprocessen med två möjliga tillstånd. Om vi tittar på sidan 94, ser vi att $-q_{00} = q_0 = \lambda$, $-q_{11} = q_1 = \mu$ och $q_{01} = \lambda$,

$q_{10} = \mu$. Den stationära fördelningen beräknas i boken.

Födelse- och dödsprocesserna i avsnitt 2.4 är också Markovkedjor med $-q_{kk} = q_k = \lambda_k + \mu_k$, $q_{kk+1} = \lambda_k$ och $q_{kk-1} = \mu_k$ (utom naturligtvis för $k = 0$, där man får något annat).

Som framgår av boken kan det vara väldigt svårt att beräkna P^t. Emellertid är det Q som är mest användbar för oss.

Det finns två sätt att beräkna den stationära fördelningen med hjälp av Q. Ett är att försöka hitta en (vänster) egenvektor till Q med egenvärdet 0. Läsaren uppmanas försöka inse att det egentligen är samma metod som används i boken.

Det andra sättet att hitta den stationära fördelningen leder inte alltid till något svar. Ett tillräckligt men inte nödvändigt villkor för $\pi = \{\pi(s)\}_{s \in S}$ att vara en stationär fördelning är, visar det sig, att $\sum_{s \in S} \pi(s) = 1$ (naturligtvis) och att $\pi(x)q_{xy} = \pi(y)q_{yx}$ för alla x och y. Som sagt kan det hända att det inte finns någon π–vektor med denna egenskap, men hittar man en sådan är ju saken klar, så metoden är bra som första steg i undersökningen.

Kontrollera gärna att den stationära fördelningen för födelse- och dödsprocesserna i avsnitt 2.4 kan fås med denna metod.

Några övningar:

1. Sätt $S = \{1, 2, 3\}$, $(q_1, q_2, q_3) = (2, 3, 4)$ och

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \end{pmatrix}.$$

a) Beräkna Q–matrisen.

b) Beräkna den stationära fördelningen genom att lösa $\pi_i q_{ij} = \pi_j q_{ji}$.

c) Beräkna $\lim_{t \to \infty} Prob(X_t = 3 | X_0 = 2)$.

2. Sätt $S = \{1, 2, 3, 4\}$, $(q_1, q_2, q_3, q_4) = (1, 1, 1, 4)$ och

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

a) Kan man lösa $\pi_i q_{ij} = \pi_j q_{ji}$?

b) Beräkna den stationära fördelningen.

c) Beräkna $\lim_{t \to \infty} Prob(X_t = 4 | X_0 = 1)$.