Markov-kæder og deres anvendelse på rangordning af knuder i netværk

Uli Fahrenberg

16 april 2008

Wikipedia/en: "A Markov chain is a sequence of random variables . . . with the . . . property . . . that, given the present state, the future and past states are independent. Formally,

$$\Pr(X_{n+1} = x \mid X_n = x_n, \ldots, X_1 = x_1) = \Pr(X_{n+1} = x \mid X_n = x_n)\text{ "}$$

Markov-kæder

Wikipedia/en: "A Markov chain is a sequence of random variables . . . with the . . . property . . . that, given the present state, the future and past states are independent. Formally,

$$\Pr(X_{n+1} = x \mid X_n = x_n, \ldots, X_1 = x_1) = \Pr(X_{n+1} = x \mid X_n = x_n)\text{ "}$$

Wikipedia/en: “A Markov chain is a sequence of random variables . . . with the . . . property . . . that, given the present state, the future and past states are independent. Formally,

$$\Pr(X_{n+1} = x \mid X_n = x_n, \dots, X_1 = x_1) = \Pr(X_{n+1} = x \mid X_n = x_n) \text{ ”}$$

Warning: All or part of this article may be confusing or unclear.

Og det passer sgu . . . Bedre ikke at starte her.

Operationel definition: En Markov-kæde er en endelig digraf (V, E) sammen med en kantvægtning $p : E \to [0, 1]$ som har den egenskab at

$$\sum_{e \in \text{ud}(v)} p(e) = 1$$

for alle $v \in V$.

Operationel definition: En Markov-kæde er en endelig digraf (V, E) sammen med en kantvægtning $p : E \to [0, 1]$ som har den egenskab at

$$\sum_{e \in \text{ud}(v)} p(e) = 1$$

for alle $v \in V$.

- en endelig digraf (V, E) : en endelig mængde V og en delmængde $E \subseteq V \times V$
- en kantvægtning $p : E \to [0, 1]$: en funktion der tillægger et reelt tal mellem 0 og 1 til enhver kant
- $\text{ud}(v)$, for $v \in V$: mængden af alle fra v udgående kanter; $\text{ud}(v) = \{(v, w) \in E \mid w \in V\} = (\{v\} \times V) \cap E$
- $\sum_{e \in \text{ud}(v)} p(e) = 1$: summen af vægtene af alle kanter der udgår fra v er 1

En stokastisk vektor (eller sandsynlighedsvektor) er en vektor $\vec{u} = (u_1, \dots, u_r) \in \mathbb{R}^r$ for hvilken

- $u_i \in [0, 1]$ for alle $i = 1, \dots, r$ og
- $u_1 + \dots + u_r = 1$.

En stokastisk vektor (eller sandsynlighedsvektor) er en vektor $\vec{u} = (u_1, \ldots, u_r) \in \mathbb{R}^r$ for hvilken

- $u_i \in [0, 1]$ for alle $i = 1, \ldots, r$ og
- $u_1 + \cdots + u_r = 1$.

Givet en Markov-kæde (V, E, p) med r tilstande, $V = \{s_1, \ldots, s_r\}$, og en sandsynlighedsvektor $\vec{u} = (u_1, \ldots, u_r) \in \mathbb{R}^r$, da “betyder” $\vec{u}$ at Markov-kæden er i en tilstand s_i med sandsynlighed u_i

En stokastisk vektor (eller sandsynlighedsvektor) er en vektor $\vec{u} = (u_1, \ldots, u_r) \in \mathbb{R}^r$ for hvilken

- $u_i \in [0, 1]$ for alle $i = 1, \ldots, r$ og
- $u_1 + \cdots + u_r = 1$.

Givet en Markov-kæde (V, E, p) med r tilstande, $V = \{s_1, \ldots, s_r\}$, og en sandsynlighedsvektor $\vec{u} = (u_1, \ldots, u_r) \in \mathbb{R}^r$, da “betyder” $\vec{u}$ at Markov-kæden er i en tilstand s_i med sandsynlighed u_i

Notation: Fra nu af har enhver Markov-kæde r tilstande som vi skriver $V = \{s_1, \ldots, s_r\}$

Kantvægtnings-funktionen $p : E \to [0, 1]$ fra en Markov-kæde (V, E, p) kan beskrives ved en $r \times r$-matriks

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1r} \\ p_{21} & p_{22} & \cdots & p_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ p_{r1} & p_{r2} & \cdots & p_{rr} \end{bmatrix}$$

hvor p_{ij} er vægten af kanten fra s_i til s_j, eller 0 hvis der ingen kant er;

$$p_{ij} = \begin{cases} p(s_i, s_j) & \text{hvis } (s_i, s_j) \in E \\ 0 & \text{hvis } (s_i, s_j) \notin E \end{cases}$$

En $r \times r$-matriks $M = [m_{ij}]$ kaldes rækkestokastisk hvis

- $m_{ij} \in [0, 1]$ for alle $i, j = 1, \ldots, r$ og
- $m_{i1} + m_{i2} + \cdots + m_{ir} = 1$ for alle i (dvs. summen af indgangene i hver række er 1)

En $r \times r$-matriks $M = [m_{ij}]$ kaldes rækkestokastisk hvis

- $m_{ij} \in [0, 1]$ for alle $i, j = 1, \dots, r$ og
- $m_{i1} + m_{i2} + \cdots + m_{ir} = 1$ for alle i (dvs. summen af indgangene i hver række er 1)

(Den kaldes søjlestokastisk hvis

- $m_{ij} \in [0, 1]$ for alle $i, j = 1, \dots, r$ og
- $m_{1j} + m_{2j} + \cdots + m_{rj} = 1$ for alle j (dvs. summen af indgangene i hver søjle er 1)

og stokastisk hvis den er både række- og søjlestokastisk)

En $r \times r$-matriks $M = [m_{ij}]$ kaldes rækkestokastisk hvis

- $m_{ij} \in [0, 1]$ for alle $i, j = 1, \dots, r$ og
- $m_{i1} + m_{i2} + \cdots + m_{ir} = 1$ for alle i (dvs. summen af indgangene i hver række er 1)

(Den kaldes søjlestokastisk hvis

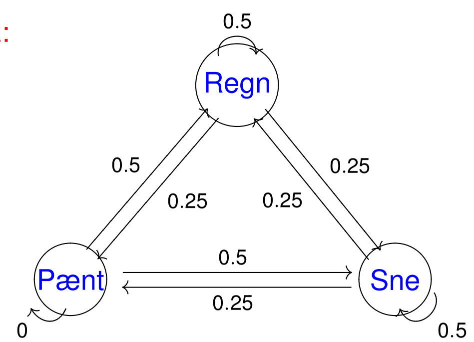
- $m_{ij} \in [0, 1]$ for alle $i, j = 1, \dots, r$ og
- $m_{1j} + m_{2j} + \cdots + m_{rj} = 1$ for alle j (dvs. summen af indgangene i hver søjle er 1)

og stokastisk hvis den er både række- og søjlestokastisk)

Overgangsmatricen til en Markov-kæde er rækkestokastisk.

- En Markov-kæde er en endelig digraf “med stokastiske kantvægte”
- Punkterne (knuderne) i digrafen kaldes tilstande (*states*), kanterne kaldes overgange (transitioner; *transitions*)
- Kantvægtene kan samles i en matriks, kaldet overgangsmatricen. Den er rækkestokastisk
- En stokastisk vektor kan bruges til at angive sandsynlighederne for at Markov-kæden er i en af dens tilstande

Vejret i landet Oz:



Transitionsmatricen:
$$\begin{array}{c} \\ R \\ P \\ S \end{array} \begin{bmatrix} R & P & S \\ 0.5 & 0.25 & 0.25 \\ 0.5 & 0 & 0.5 \\ 0.25 & 0.25 & 0.5 \end{bmatrix}$$

Definition: En Markov-kæde er en uendelig følge af stokastiske variable $(X_i)_{i \in \mathbb{N}}$ over et endeligt tilstandsrum $V = \{s_1, \ldots, s_r\}$ med den egenskab at sandsynligheden for at X_{k+1} er i en given tilstand kun afhænger af hvilken tilstand X_k var i:

$$\Pr(X_{k+1} = s_{i_{k+1}} \mid X_k = s_{i_k}, \ldots, X_0 = s_{i_0}) = \Pr(X_{k+1} = s_{i_{k+1}} \mid X_k = s_{i_k})$$

for alle $k \in \mathbb{N}$.

Definition: En Markov-kæde er en uendelig følge af stokastiske variable $(X_i)_{i \in \mathbb{N}}$ over et endeligt tilstandsrum $V = \{s_1, \ldots, s_r\}$ med den egenskab at sandsynligheden for at X_{k+1} er i en given tilstand kun afhænger af hvilken tilstand X_k var i:

$$\Pr(X_{k+1} = s_{i_{k+1}} \mid X_k = s_{i_k}, \ldots, X_0 = s_{i_0}) = \Pr(X_{k+1} = s_{i_{k+1}} \mid X_k = s_{i_k})$$

for alle $k \in \mathbb{N}$.

(Hovsa, det lyder nu næsten som Wikipedia alligevel . . .)

Definition: En Markov-kæde er en uendelig følge af stokastiske variable $(X_i)_{i \in \mathbb{N}}$ over et endeligt tilstandsrum $V = \{s_1, \ldots, s_r\}$ med den egenskab at sandsynligheden for at X_{k+1} er i en given tilstand kun afhænger af hvilken tilstand X_k var i:

$$\Pr(X_{k+1} = s_{i_{k+1}} \mid X_k = s_{i_k}, \ldots, X_0 = s_{i_0}) = \Pr(X_{k+1} = s_{i_{k+1}} \mid X_k = s_{i_k})$$

for alle $k \in \mathbb{N}$.

(Hovsa, det lyder nu næsten som Wikipedia alligevel . . .)

Indfør en kantvægtet graf ved $E = V \times V$ og
$p(s_i, s_j) = \Pr(X_{k+1} = s_j \mid X_k = s_i)$
$\Rightarrow$ en Markov-kæde som vi kender den!

Definition: En Markov-kæde er en uendelig følge af stokastiske variable $(X_i)_{i \in \mathbb{N}}$ over et endeligt tilstandsrum $V = \{s_1, \ldots, s_r\}$ med den egenskab at sandsynligheden for at X_{k+1} er i en given tilstand kun afhænger af hvilken tilstand X_k var i:

$$\Pr(X_{k+1} = s_{i_{k+1}} \mid X_k = s_{i_k}, \ldots, X_0 = s_{i_0}) = \Pr(X_{k+1} = s_{i_{k+1}} \mid X_k = s_{i_k})$$

for alle $k \in \mathbb{N}$.

(Hovsa, det lyder nu næsten som Wikipedia alligevel . . .)

Indfør en kantvægtet graf ved $E = V \times V$ og
$p(s_i, s_j) = \Pr(X_{k+1} = s_j \mid X_k = s_i)$
$\Rightarrow$ en Markov-kæde som vi kender den!

Og en stokastisk vektor $\vec{u} = (u_1, \ldots, u_r) \in \mathbb{R}^r$ beskriver nu sandsynlighederne $u_i = \Pr(X_k = s_i)$ til et eller andet tidspunkt k.

Min præsentation lægger sig tæt op ad Grinstead, Snell: Introduction to Probability. The Chance Project, Juli 2006

Specielt afsnittene 11.1, 11.3 og 11.4 i den bog er god læsning.

Regulære Markov-kæder

Sætning 11.2 (ca.): Lad P være en Markov-kædes overgangsmatriks, og lad $\vec{u}^{(n)}$ betegne den stokastiske (række)vektor der angiver sandsynlighederne for at kæden er i de enkelte tilstande til tiden n. Da gælder $\vec{u}^{(n)} = \vec{u}^{(n-1)}P$.

Specielt er $\vec{u}^{(n)} = \vec{u}^{(0)}P^n$.

Og P^ns indgange $p_{ij}^{(n)}$ angiver sandsynlighederne for at kæden i n skridt kommer fra s_i til s_j.

Lad $P = \begin{bmatrix} & R & P & S \\ R & 0.5 & 0.25 & 0.25 \\ P & 0.5 & 0 & 0.5 \\ S & 0.25 & 0.25 & 0.5 \end{bmatrix}$ ligesom sidst, og start med en solskinsdag: $\vec{u}^{(0)} = (0, 1, 0)$. Så har vi

$$\begin{aligned}
\vec{u}^{(0)} &= (0, 1, 0) \\
\vec{u}^{(1)} &= (0.5, 0, 0.5) \\
\vec{u}^{(2)} &= (0.375, 0.25, 0.375) \\
\vec{u}^{(3)} &= (0.406, 0.188, 0.406) \\
\vec{u}^{(4)} &= (0.398, 0.203, 0.398) \\
\vec{u}^{(5)} &= (0.400, 0.199, 0.400) \\
\vec{u}^{(6)} &= (0.400, 0.200, 0.400)
\end{aligned}$$

osv. (det konvergerer!). Så efter en uge er der 20% sandsynlighed for at solen skinner en tilfældig dag.

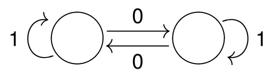
Definition 11.4: En Markov-kæde kaldes ergodisk (eller irreducibel) hvis enhver tilstand på et tidspunkt kan nås fra enhver anden tilstand.

Dvs. hvis der for ethvert par af tilstande s_i, s_j findes et n så $\Pr(X_n = s_j \mid X_0 = s_i) > 0$.

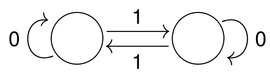
Definition 11.5: En Markov-kæde kaldes regulær hvis der findes et n (fælles for alle i, j) således at der for ethvert par af tilstande s_i, s_j gælder at $\Pr(X_n = s_j \mid X_0 = s_i) > 0$.

Dvs. hvis der findes n så P^n ikke indeholder nogen 0er.

Den her kæde er ikke ergodisk:



Den her kæde er ergodisk, men ikke regulær:



Og vejret i landet Oz er regulært (heldigvis).

Definition 11.6: En stokastisk vektor $\vec{w}$ kaldes en stabil fordeling (eller ligevægtsfordeling) for en Markov-kæde med overgangsmatriks P hvis $\vec{w}P = \vec{w}$.

Ikke alle Markov-kæder har en stabil fordeling, og de kan godt have flere end én. Men:

Sætning 11.10: Enhver ergodisk Markov-kæde har præcist én stabil fordeling.

Sætning 11.9: Hvis P er overgangsmatricen til en regulær Markov-kæde og $\vec{u}$ en vilkårlig stokastisk vektor, da gælder $\lim_{n\to\infty} \vec{u}P^n = \vec{w}$, hvor $\vec{w}$ er kædens stabile fordeling.

Lad (V, E) være internettet og numerér $V = \{s_1, \ldots, s_r\}$.

Lad H være hyperlink-matricen givet ved

$$h_{ij} = \begin{cases} \dfrac{1}{|\mathrm{ud}(s_i)|} & \text{hvis der er et link fra } s_i \text{ til } s_j \\ 0 & \text{ellers} \end{cases}$$

Lad (V, E) være internettet og numerér $V = \{s_1, \dots, s_r\}$.

Lad H være hyperlink-matricen givet ved

$$h_{ij} = \begin{cases} \dfrac{1}{|\mathrm{ud}(s_i)|} & \text{hvis der er et link fra } s_i \text{ til } s_j \\ 0 & \text{ellers} \end{cases}$$

H er række-substokastisk: $h_{ij} \in [0, 1]$ for alle i, j, og rækkesummerne er højst 1.
(Faktisk er rækkesummerne enten 0 eller 1 i det her tilfælde.)
(Bevis?)

Lad S være den matriks der fremkommer fra H ved at erstatte alle indgange i alle nul-rækker med $\frac{1}{r}$.

Dvs. $s_{ij} = \begin{cases} \frac{1}{r} & \text{hvis } h_{i1} = h_{i2} = \cdots = h_{ir} = 0 \\ h_{ij} & \text{ellers} \end{cases}$

Eller på en mere fancy måde: $S = H + \vec{a}\frac{1}{r}\vec{e}^T$, hvor $\vec{e}^T$ er en rækkevektor med r 1-taller, og $\vec{a}$ en søjlevektor med $a_i = 1$ hvis $|\mathrm{ud}(s_i)| = 0$ og $a_i = 0$ ellers.

Lad S være den matriks der fremkommer fra H ved at erstatte alle indgange i alle nul-rækker med $\frac{1}{r}$.

Dvs. $s_{ij} = \begin{cases} \frac{1}{r} & \text{hvis } h_{i1} = h_{i2} = \cdots = h_{ir} = 0 \\ h_{ij} & \text{ellers} \end{cases}$

Eller på en mere fancy måde: $S = H + \vec{a}\frac{1}{r}\vec{e}^T$, hvor $\vec{e}^T$ er en rækkevektor med r 1-taller, og $\vec{a}$ en søjlevektor med $a_i = 1$ hvis $|\mathrm{ud}(s_i)| = 0$ og $a_i = 0$ ellers.

S er rækkestokastisk (Bevis?) og derfor overgangsmatriks for en Markov-kæde.
(Men den Markov-kæde er ikke nødvendigvis regulær, hvilket vi gerne ville have.)

Vælg et reelt tal α med $0 < \alpha < 1$ og definer en ny matriks G ved $G = \alpha S + (1 - \alpha)\frac{1}{r}\vec{e}\vec{e}^T$.

α kaldes dæmpefaktoren, og ligningen ovenover kan i koordinater skrives som $g_{ij} = \alpha s_{ij} + (1 - \alpha)\frac{1}{r}$.

(Det kan den fordi $\vec{e}\vec{e}^T$ er en $r \times r$-matriks med lutter 1-taller.)

Vælg et reelt tal α med $0 < \alpha < 1$ og definer en ny matriks G ved $G = \alpha S + (1-\alpha)\frac{1}{r}\vec{e}\vec{e}^T$.

α kaldes dæmpefaktoren, og ligningen ovenover kan i koordinater skrives som $g_{ij} = \alpha s_{ij} + (1-\alpha)\frac{1}{r}$.

(Det kan den fordi $\vec{e}\vec{e}^T$ er en $r \times r$-matriks med lutter 1-taller.)

G er stadig rækkestokastisk (Bevis?), og alle indgange i G er > 0 (Bevis). Derfor er G overgangsmatricen for en regulær Markov-kæde.

Vælg et reelt tal α med $0 < \alpha < 1$ og definer en ny matriks G ved $G = \alpha S + (1-\alpha)\frac{1}{r}\vec{e}\vec{e}^T$.

α kaldes dæmpefaktoren, og ligningen ovenover kan i koordinater skrives som $g_{ij} = \alpha s_{ij} + (1-\alpha)\frac{1}{r}$.

(Det kan den fordi $\vec{e}\vec{e}^T$ er en $r \times r$-matriks med lutter 1-taller.)

G er stadig rækkestokastisk (Bevis?), og alle indgange i G er > 0 (Bevis). Derfor er G overgangsmatricen for en regulær Markov-kæde.

Den stabile fordeling for Markov-kæden givet ved overgangsmatricen G kaldes for internettets pagerank-vektor og skulle efter sigende give en rimelig god rangordning af internettets knuder. Med sætning 11.9 kan den beregnes som en grænseværdi $\lim_{n\to\infty} \vec{u}G^n$, hvor $\vec{u}$ er en vilkårlig stokastisk startvektor.

- *skrå: engelsk standard*, blå: Grinstead / Snell, rød: Langville / Meyer
- rækkestokastisk – *right stochastic* – stochastic
- ergodisk – irreducibel – irreducible – ergodic
- regulær – *regular* – primitive
- (skal der flere ord på her? Forslag?)