Workshop - PageRank

I denne Workshop betragter vi en person som surfer rundt på internettet. For at simplificere det en smule kigger vi kun på 5 mulige internetsider givet i mængden $W = \{w_1, w_2, w_3, w_4, w_5\}$. Vi antager, at personen via links på siderne går fra en af hjemmesiderne i mængden W til en anden. Dette illustrerer vi ved de stokastiske variabler $W^{(t)}$. Udfaldet af $W^{(t)}$ er nummeret på den internetside personen er på på tidspunktet t. Lad nu sandsynlighederne for første udfald være

$$P(W^{(0)} = 1) = \frac{1}{2}, \qquad P(W^{(0)} = i) = \frac{1}{8}, \text{ hvor } i \in \{2, 3, 4, 5\}. \tag{1}$$

Og sandsynlighederne for de resterende være

$$P(W^{(t+1)} = 1 | W^{(t)} = 1) = 0, \qquad P(W^{(t+1)} = i | W^{(t)} = 1) = \frac{1}{4}$$
$$P(W^{(t+1)} = 1 | W^{(t)} = i) = \frac{1}{2}, \qquad P(W^{(t+1)} = i | W^{(t)} = j) = \frac{1}{8}$$

for $i, j \in \{2, 3, 4, 5\}$. Dette er illustreret i Figur 1.

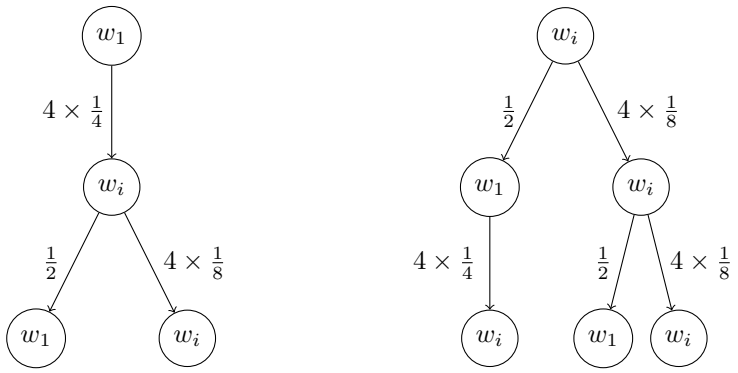

Figure 1: Sandsynligheder. Til venstre ses, hvordan det vil se ud, hvis udfaldet i første eksperiment er w_1, hvilket sker med sandsynlighed $\frac{1}{2}$, og til højre ses hvordan det vil se ud, hvis udfaldet af første eksperiment er en af de andre w_i, $i = 2, \ldots, 5$, hvilket sker med sandsynlighed $4 \times \frac{1}{8}$ (dvs, sandsyndlighed $\frac{1}{8}$ for hver $i = 2, \ldots, 5$).

Lad nu X være en stokastisk variabel som tæller hvor mange ganger internetside 1 besøges i tidsintervallet $t = 0, 1, 2$. Dvs

$$X = \text{antallet af ganger } W^{(t)} = 1 \text{ er opfyldt, } t = 0, 1, 2.$$

Delopgave 1

1. Vi lader udfaldsrummet være

   $$S = \{w_i w_j w_k \mid i, j, k \in \{1, 2, 3, 4, 5\}\},$$

   hvor hvert udfald svarer til de 3 internetsider besøgte i løbet af tidsintervalet $t = 0, 1, 2$. Husk, at X er en funktion fra S til $\mathbb{R}$ og man kan derfor snakke om urbilledet af $y \in \mathbb{R}$, hvilket er givet ved $\{s \in S \mid X(s) = y\}$. Forklar, hvorfor dette urbillede svarer til en hændelse.

2. For hvilke $y \in \mathbb{R}$ har vi et ikke-tomt urbillede? Og hvad er $p(X = y)$ for disse y?

3. Hvad er middelværdien og variansen for X?

4. Hvis $W^{(t)}$ i stedet havde fulgt en Bernoullifordeling for alle t med sandsynlighed $\frac{3}{8}$ for at få w_1 (hvilket vi betegner som vores "succes"), hvilken fordeling havde X så fulgt og hvad havde middelværdien og variansen været? Sammenlign med det foregående eksempel.

Bemærk, at $W^{(t)}$ kan beskrives som en Markov-kæde med stokastisk matrix P givet ved

$$P = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \\ \frac{1}{4} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \\ \frac{1}{4} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \\ \frac{1}{4} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \end{bmatrix}$$

I Delopgave 1 betragtede vi kun $t = 0, 1, 2$. Fra nu af begrænser vi os ikke til dette, men kan kigge på et vilkårligt $t \in \mathbb{N}$.

Delopgave 2

1. Hvad er sandsynligheden for at være i tilstand w_1 til tiden $t = 5$, hvis vi bruger startfordelingen fra (1) og den stokastiske matrix P?
2. Har Markov-kæden en stationær fordeling? Hvis ja, bestem sådan en.

Bemærk at vi ovenfor antog, at vi kunne komme til alle andre internetsider via links uanset fra hvilken internetside vi kom fra. Det er måske fair nok i eksemplet ovenfor hvor vi kun betragter 5 sider, men ville nok ikke holde hvis vi betragtede flere sider. Bemærk, at vi stadig kan modellere det med en Markov-kæde, men at det blot vil betyde, at indgang i, j i matricen for Markov-kæden er 0, hvis der ikke er et link fra w_j til w_i.

Vi kigger nu på et lidt større eksempel med 7 hjemmesider, hvor der er links mellem siderne som illustreret på Figur 2. Vi vil med dette som eksempel kigge nærmere på Google's PageRank algoritme.

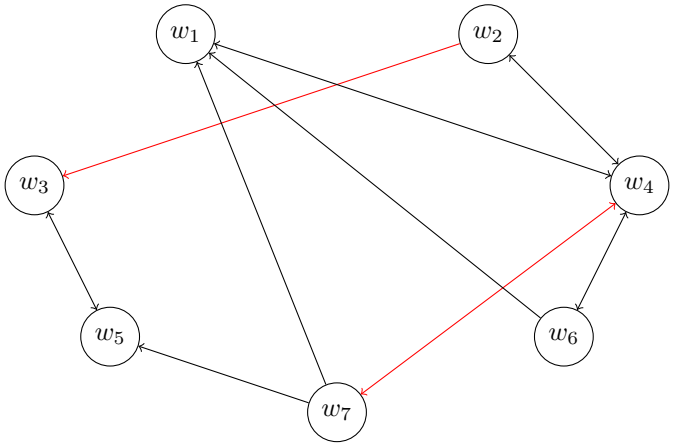

Figure 2: En pil fra en internetside til en anden betyder, at der er et link på den første til den anden side.

I Google's PageRank algoritme betragter vi netop en person som surfer rundt på nettet. Vi bruger denne tænkte situation til at "ranke" internetsiderne. Den første tanke går på, at når surferen står på en side vil han vælge blandt de mulige links på siden med lige stor sandsynlighed. Derfor bliver den stokastiske matrix som illustrerer denne Markov kæde følgende (hvor både de røde og sorte pile tælles med)

$$P_1 = \begin{bmatrix} 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{2} & \frac{1}{3} \\ 0 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 1 & 0 & 0 \\ 1 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{3} \\ 0 & 0 & 1 & 0 & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 \end{bmatrix} \tag{2}$$

Idéen i PageRank er mere eller mindre at finde sandsynligheden for at surferen er på en bestemt side efter "tilstrækkelig mange" klik. Det vi søger er altså en stationær fordeling.

Delopgave 3

1. Find stationær fordeling for Markov-kæden beskrevet ud fra den stokastiske matrix P_1.
2. Find stationær fordeling for Markov-kæden beskrevet ud fra Figur 2 hvis vi fjerner de røde pile. Derudover siger vi, at sandsynligheden for at gå fra en tilstand til den næste bliver fordelt ligeligt mellem de kanter der går ud fra tilstanden. (F.eks. går der to sorte pile væk fra w_7, hvorfor indgangene her bliver $\frac{1}{2}$).
3. Er der en entydig stationær fordeling i de to tilfælde? Kan I forklare hvorfor der er/ikke er?

Problemet med denne første tanke er, at vi ikke altid vil have en entydig stationær fordeling. For at sikre os, at vi altid opnår dette, tilføjer PageRank en sandsynlighed for at surferen starter et helt nyt sted (altså blot indtaster en ny internetadresse i stedet for at følge de links der er på siden).

Derfor benyttes i stedet den stokastiske matrix

$$P = \alpha P_1 + (1 - \alpha)P_2 \tag{3}$$

hvor P_1 er sandsynligheden via links (som vist i et eksempel i (2)) og P_2 er en matrix hvor alle indgange er 1 over antallet af tilstande/internetsider (i vores eksempel $\frac{1}{7}$). α er en variabel som kan justeres alt efter hvor sandsynligt, vi mener det er, at surferen følger links eller skifter til en ny side.

Delopgave 4

1. Hvorfor vil en Markov-kæde med stokastisk matrix P som i (3) altid have en entydig stationær fordeling hvis $0 < \alpha < 1$?

2. Forklar hvad et lille/stort α betyder for vores vurdering af surferens adfærd.

3. Beregn den entydige stationære fordeling for de to eksempler i Delopgave 3 hvis vi sætter $\alpha = 0.85$, som typisk er blevet brugt i literaturen. Ud fra resultaterne, hvordan vil I så "ranke" internetsiderne i de to eksempler?