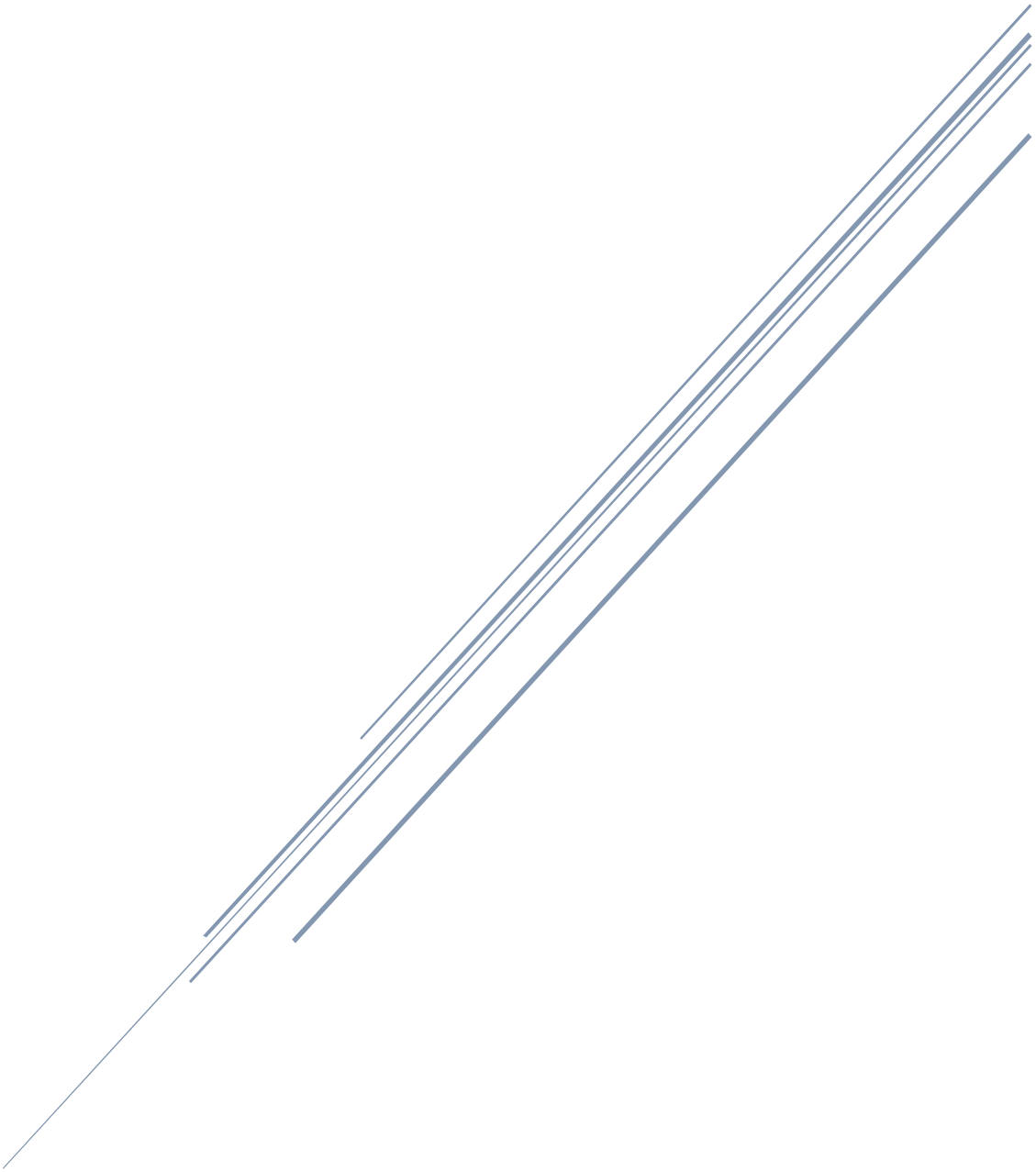
UKÁZKY APLIKACÍ MATEMATIKY

Google vyhľadávanie



MFF UK
Patrik Hric

Úvod

Matematika je veda, ktorá je aplikovaná v toľkých oblastiach ľudského života, že si to človek ani neuvedomuje. V tejto práci sa budem zaoberať jednou z nich, s ktorou sa takmer každý z nás stretáva dennodenne a tou je internet, respektíve vyhľadávanie informácii na internete. Zadá sa informácia do internetového prehliadača, ktorý nám zobrazí množstvo stránok, ktoré by nám mali pomôcť pri získaní informácie. Ale stránky nie sú zoradené len tak náhodne, ale tak, aby človek získal danú informáciu, čo najrýchlejšie a najefektívnejšie. Na to slúži algoritmus, ktorý sa nazýva Google PageRank a využívajú sa tu poznatky aj z matematiky. A to konkrétne Markovské reťazce a Skrytý Markovský model z oblasti stochastiky. Google tak kalkuluje skóre jednotlivých stránok a zoradí ich do poradia, ktoré sa zobrazí pri vyhľadávaní. Nazačiatok si zadefinujeme jednotlivé pojmy, uvediem nejaké príklady pre lepšie pochopenie a nakoniec popíšeme daný proces.

Markovské reťazce

Stochastické procesy

Predtým ako sa začnem venovať konkrétne Markovským reťazcom vám priblížim trošku teóriu stochastických procesov pomocou definície a nejakých príkladov, pretože Markovské reťazce sú špeciálnym typom stochastických procesov.

Definícia:[1] Nech (Ω, A) je merateľný priestor, R množina reálnych čísel, T= ∅ neprázdna množina (najčastejšie jej budeme prisudzovať význam času). Nech zobrazenie X: Ω×T $-\rightarrow$ R má tieto dve vlastnosti:

- ∀ $t \in$ T je X(., t) náhodná veličina vzhľadom k javovému poľu A . Značí sa X_t .
- ∀ω ∈ Ω je X(ω, .) prvkom množiny všetkých reálnych funkcií definovaných na T.

Zobrazenie X s týmito dvoma vlastnosťami nazývame stochastickým procesom definovaným na T. Značí sa { X_t ; $t \in$ T}.

Typy: [2] Nech {X_t ; $t \in$T} je stochastický proces,

- Ak je množina T spočetná a lineárne usporiadaná, tj. $t_0 < t_1 < ...$, ide o stochastický proces s diskrétnym časom (tj. o časovou radu).
- Ak je množina T interval, ide o stochastický proces so spojitým časom (tj. o náhodnou funkcii).
- Ak pre ∀ $t \in$ T je náhodná veličina X_t diskrétna, ide o SP s diskrétnymi stavy.
- Ak pre ∀ $t \in$T je náhodná veličina X_t spojitá, ide o SP so spojitými stavy.

Príklady:

1. *(Stochastický proces s diskrétnym časom a spojitými stavmi)* V prevádzke používajú obrábacie nože a po každej výrobnej operácii kontrolujú stav opotrebenia. Vždy po n operáciách nôž vymenia. Máme teda stochastický proces {X_t ; $t \in$ T}, kde t je poradové číslo výrobnej operácie, T = {1, 2, ..., n} je množina výrobných operácii a množina stavov J = {$x \in$ R; $0 \leq x \leq a$} nadobúda hodnoty, ktoré ukazujú ako je nôž opotrebený pričom a je maximálne opotrebenie noža. – *Čas t je diskrétny, pretože opotrebenie sa kontroluje po každej operácii.*

2. *(Stochastický proces s diskrétnym časom a stavmi)* Dvaja hráči P a M vložia do hry spolu 10€, z toho hráč P vložil 7€ a hráč M 3€. Hráč P hádže opakovane mincou. Ak padne hlava, vyhráva 1€, keď padne znak, prehrá 1€. Hráč P hádže mincou, kým jeden z hráčov neprehral všetky peniaze. Zavedieme si stochastický proces $\{X_t ; t \in T\}$, kde t = 1, 2, ...je poradové číslo hodu mincou a $X_t = j$, keď hráč P má po *t*-tom hode *j* €, teda $J = \{0, 1, ..., 9, 10\}$. – *Čas je diskrétny z dôvodu, že počet €, ktoré má hráč P k dispozícii sa kontroluje vždy po hode kockou.*

Markovovské reťazce

Markovské reťazce sú pomenované po ruskom matematikovi Andrejovi Markovi, celým menom Andrej Andrejevič Markov. Pracoval v teórii čísel, pravdepodobnosti a matematickej analýze a dodnes je známy vďaka svojim teóriám stochastických (náhodných) procesov.

Definícia: [3] Markovov reťazec sa nazýva náhodný proces s diskrétnym časom pre ktorý platí:

a) $S = \{0, 1, 2, ... \}$
b) $T = \{0, 1, 2, ... \}$
c) *Markovská vlastnosť = pravdepodobnosť dosiahnutia výsledku j v budúcom čase n+1 je ovplyvnená iba výsledkom prítomným v čase n ,výsledky z minulosti nemajú na pravdepodobnosť vplyv.*
$\forall\, i_0, i_1, ...i_{n-1}, i, j \in S\ n \in T$: $P(x_{n+1}=j \mid x_1=i_1, x_2=i_2, ..., x_{n-1}=i_{n-1}, x_n=i) = P(x_{n+1}=j \mid x_n=i)$.

Definícia: [4] Nech $\{x_t , t \in T\}$ je Markovov reťazec,

- $P(x_t = j \mid x_s = i) = p_{i,j}(s, t),\ 0 \le s < t;\ s, t \in T;\ i, j \in S$, nazývame podmienené pravdepodobnosti prechodu Markovovho reťazca zo stavu *i* v čase *s* do stavu *j* v čase *t*.
- Nepodmienené pravdepodobnosti $P(x_t = j) = p_j(t),\ t \in T,\ j \in S$, nazývame absolútne pravdepodobnosti stavov Markovovho reťazca.

Definícia: [5] Matica $\tilde{P}(s,t) = (p_{i,j}(s, t))_{i,j \in S}$ sa nazýva matica pravdepodobnosti prechodu Markovovho reťazca. Ak táto matica spĺňa:

- $p_{i,j}(s, t) \ge 0$ *(všetky prvky matice sú nezáporné),*
- $\sum_{j \in S} p_{i,j}(s,t) = 1$ *(súčet prvkov v riadku je 1),*

sa nazýva stochastická matica a ak navyše aj platí:

- $\sum_{i \in S} p_{i,j}(s,t) = 1$

sa nazýva dvojne stochastická.

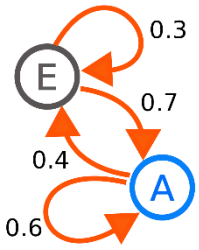
Veta: Matica pravdepodobností prechodov má nasledujúce vlastnosti:

- $\forall i,j \in S : p_{i,j} \ge 0$
- $\forall i \in S : \sum_j p_{i,j} = 1$ *t.j. súčet pravdepodobností v riadku je 1*

Definícia: [6] Diagram (graf) prechodu Markovovho reťazca je orientovaný graf, kde vrcholy reprezentujú stavy Markovovho reťazca a hrany znázornené šípkami označujú možné prechody medzi stavmi.

Zhrnutie: Markovov reťazec znázorňuje náhodný proces, kedy pravdepodobnosť z jedného bodu do druhého, kedy závisí len na súčasnom stave a nie na predchádzajúcich. Môžeme si to znázorniť na jednoduchom obrázku. Na obrázku (orientovanom grafe) máme dva body

označené písmenami a šípky, ktoré vedú z jedného bodu do druhého alebo do toho istého bodu. Keď sa nachádzame v bode A tak pravdepodobnosť zostatia v tom bode je 0,6 a pravdepodobnosť prechodu do bodu E je 0,4. Tieto pravdepodobnosti vôbec nezávisia na tom ako sme sa dostali do bodu A . To isté platí aj pre bod E. Teda tento graf závisí len na aktuálnom stave a nie na predchádzajúcom. Markovské reťazce majú veľa využití v reálnom živote ako napríklad v chémii, ekonómii, informatike a aj pri vyhľadávaní stránok, ktoré ovplyvňuje algoritmus PageRank. [7]



Skrytý Markovov model (HMM)

Informácie: Ide o Markovov proces so skrytými (nepozorovanými) stavmi. Tento model vyvinul Leonard E. Baum spolu so svojim tímom. Rozdiel od jednoduchších Markovovych modelov (ako sú napr. Markovovské reťazce) je ten, že pri skrytých modeloch je viditeľný iba výstup a nie stav, od ktorého je výstup závislý. Model sa nazýva skrytý aj napriek tomu, že sú parametre známe a presne zadané.

Formálne: [8] HMM je 5-tica *(K,O,π,A,B), kde:*

- *K* je konečná množina stavov
- *O* je výstup
- *π* je vektor počiatočných pravdepodobností jednotlivých stavov
- *A* je matica prechodov jednotlivých stavov
- *B* je matica výstupných pravdepodobností
- nástroj, ktorý nájde skryté procesy danej postupnosti

Tento model poskytuje tri typy pravdepodobnostných informácií:

1. Pravdepodobnosť systému na začiatku
2. Ak p je jeden stav a q druhý tak prechod zo stavu p do stavu q je označený pravdepodobnostnou hodnotou. Ak je táto hodnota rovná nule tak neexistuje prechod zo stavu p do stavu q
3. Každý výstup zo stavu q je taktiež označený pravdepodobnostnou hodnotu, pravdepodobnosťou dosiahnutia finálneho stavu (výstupu)

Tri problémy, pre ktoré ponúka riešenie: [9]

- Stanovenie pravdepodobnosti pozorovania – zistiť aká je pravdepodobnosť, že daný model generoval daný výstup
- Určenie postupnosti stavov – zistiť aká je najpravdepodobnejšia ,,cesta,, cez model, ktorá produkuje danú postupnosť
- Trénovanie modelu – aké by mali byť parametre model, aby bola vysoká pravdepodobnosť vytvárania postupností

Na riešenie prvého problému sa využíva Forward algorithm, na riešenie druhého Viterbi algorithm a nakoniec Baum – Welch algorithm.

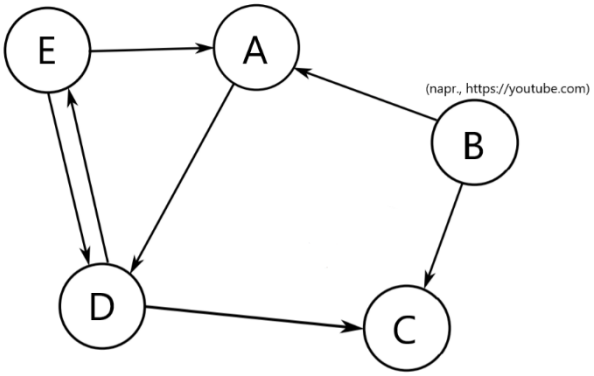
Príklad: [10] Pred miestnosťou, v ktorej je žena sa nachádza muž. Muž do danej miestnosti nevidí. Miestnosť obsahuje priehradky x_1, x_2, x_3, ... V každej priehradke sa nachádza známy počet hodiniek. Hodinky sú označené y_1, y_2, y_3, ... Žena náhodne vyberie jednu z priehradiek a vytiahne z nich hodinky a odovzdá ich mužovi za dverami. Tento proces sa opakuje. Muž vie v akom poradí sa k nemu dostali hodinky, ale nevie, z ktorých priehradok ich žena vytiahla.

Výber priehradok pre vybratie n-tých hodiniek závisí na náhodnom čísle a na vybratí priehradky pre vytiahnutie n-1ých hodiniek. Tento proces, ktorý sa týkal ženy bol Markovov proces. Proces, ktorý sa týkal už iba muža, ktorý mal iba informácie iba o výstupe.

PageRank

Informácie: [11] PageRank má dva významy. Prvým je číselné ohodnotenie nejakej stránky a druhým je algoritmus, ktorý priradí dané číslo tejto stránke. Nás bude samozrejme zaujímať ten algoritmus. Vznikol na Stanfordovej univerzite Larrym Pagom, po ktorom je aj pomenovaný dodnes. Samozrejme nepracoval na tom sám, preto je dôležité spomenúť aj ďalších jeho kolegov, ktorí sa podieľali na tomto projekte. Sergey Mikhaylovich Brin, Rajeev Motwani a Terry Allen Winograd.

Princíp: Predpokladajme, že web obsahuje *n* stránok a každá stránka je označená číslom *k, 1 ≤ k ≤ n.* A šípky znázorňujú, že napr. stránka D obsahuje hypertextový odkaz na stránku E. Ideou algoritmu je priradiť skóre (významnosť) na základe liniek, ktoré smerujú na túto stránku. Predstavme si, že spojenie dvoch stránok ako odporúčanie (samozrejme čím väčší počet odporúčaní tým je stránka dôležitejšia). Zároveň je dôležitý aj odporúčateľ. Je rozdiel, keď odporučí stránku Mark Zuckerberg ako nejaký predavač, ale zároveň to závisí aj od počtu odporučených stránok. Je rozdiel keď je odporučených 30 000 stránok alebo 10. Teda dôležitosť danej stránky získame sčítaním dôležitosti stránok, ktoré odkazujú na danú stránku a vydelíme ich počtom liniek, čo z nich vychádzajú. Dôležitosť stránky znázorňuje číslo, ktoré je nezáporne, ale môže byť aj nula. Vtedy má daná stránka najmenšie možné ohodnotenie.



Príklad: Web obsahuje 5 stránok a pre každú stránku dokážeme vypočítať dôležitosť pomocou rovníc.

$x_1 = x_2/3 + x_3/2$

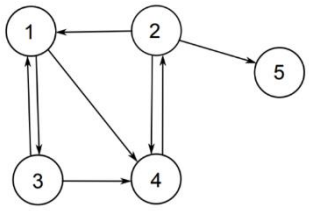
$x_2 = x_4$

$x_3 = x_1/2$

$x_4 = x_1/2 + x_2/3 + x_3/2$

$x_5 = x_2/3$

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & 0 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$



Máme 5 rovníc kde neznáme tvoria popularitu (dôležitosť) danej stránky. Predpokladajme, že na začiatku každá stránka má popularitu *1/5,* kde *5* znázorňuje počet všetkých stránok. Teda pre web s *5* stránkami máme:$x_0 = (1/5, 1/5, \ldots, 1/5)$.Po vynásobení x_0 maticou liniek dostávame $x_1 = x_0A$, $x_2 = x_1A = x_0A^2$, ..., $x_5 = x_4A = x_0A^5$

Vzorec: [12]

- **pôvodný:** $PR(P_i) = (1-d) + d.\sum (PR(P_j)/C(P_j))$
- **nový:** $PR(P_i) = (1-d)/N + d.\sum (PR(P_j)/C(P_j)$
  - $PR(P_i)$ – PageRank i-tej stránky
  - d = damping faktor menší ako 1 a väčší ako nula, cca 0,8-0,9
  - $P_1, P_2, \ldots, P_n$ sú všetky stránky v indexe a teda N je počet stránok
  - $\sum$ = suma pre všetky P_j z $M(P_i)$, kde $M(P_i)$ je množina všetkých stránok odkazujúcich na danú stránku i
  - $C(P_j)$ = počet odkazov na j-tej stránke

Google:

Keď človek zadá nejaký kľúčový pojem do vyhľadávača zobrazia sa mu stránky zoradené podľa dôležitosti jednotlivých stránok. Dôležitosť stránky určuje dané skóre, ktoré je určené podľa PageRankingu, Markovských reťazcov a Skrytých Markovovych modelov. Jednotlivé pojmy boli popísané už vyššie, ale bolo by fajn ich dať dokopy ako fungujú. PageRanking je algoritmus, ktorý priraďuje skóre jednotlivým stránkam. Jeho úlohou je teda zabezpečiť to, že keď hľadáme nejaký pojem na internete nájdeme ich na prvých desiatich stránkach, ktoré nám Google zobrazí. Uznajte, kto z vás niekedy vyhľadával informácie na druhej strane. Tento systém musí brať aj do úvahy správanie nás ľudí. A na to využíva Markovov model, konkrétne Markovské reťazce. Ide tu o predvídanie nášho správania, kedy sa chceme dostať z jednej stránky na druhú. Nachádzame sa na nejakej stránke a chceme kliknúť na inú. A tento pohyb závisí na aktuálnom stave, čiže na stránke, na ktorej sa nachádzame a nie na minulosti, teda udalostiach, ktoré nastali predtým. A nakoniec Skrytý Markove model (HMM), ktorý na základe štúdií môže mať aj daný internetový vyhľadávač pre MapReduce funkciu alebo pre triedenie distribuovaných dát. Funkcia map sa stará o to, že na základe vstupu vygeneruje medzivýsledok. Vstup funkcie map je kľúč a hodnota na výstupe môže byť niekoľko takýchto dvojíc. Medzivýsledok, ktorý vyprodukuje funkcia Map, je vstupom pre funkciu Reduce. Táto funkcia na základe kľúčov z medzivýsledku spracuje príslušné hodnoty a vypočíta výsledok.

Záver:

Matematika má veľmi široké pole pôsobnosti v bežných veciach, s ktorými sa dennodenne stretávame ako napríklad aj Google vyhľadávanie. Každý z nás, keď potrebuje nájsť nejaké informácie, adresy alebo recepty na prípravu obyčajného sladkého koláča si okamžite zapne svoje mobilné zariadenie alebo počítač a zadá do Googlu daný kľúčový pojem, ktorý nám zobrazí x – strániek , ktoré nám danú informáciu poskytnú behom pár desiatok sekúnd. Ale málo kto sa niekedy zamyslel nad tým, v akom poradí sa tie stránky zobrazia a ako efektívne toto zoradenie je. Toto funguje na základe Google PageRanking, Markovho reťazca a Skrytého Markovho modelu , z ktorých posledné dva sú neodmysliteľnou súčasťou matematiky. S matematikou sa nestretávame iba na základnej škole, keď sa učíme počítať alebo na strednej škole, keď riešime rovnice, ale každý deň počas celého nášho života preto je taká dôležitá.

Zdroje:

[1] Definícia je prevzatá z https://is.muni.cz/th/cmdet/bakalarka.pdf konkrétne z 1.kapitoly 12.strany

[2] Typy a aj príklady sú z https://is.muni.cz/th/cmdet/bakalarka.pdf , 1.kapitola 13. – 14. strana

[3] Definícia je prevzatá z http://fpedas.utc.sk/~stacho/01THO_2012.pdf , 9.slide

[4] Definícia z https://unibook.upjs.sk/img/cms/2018/pf/nahodne-procesy-web.pdf strana 21

[5] Definícia, veta z https://unibook.upjs.sk/img/cms/2018/pf/nahodne-procesy-web.pdf strana 21,22

[6] Definícia z https://unibook.upjs.sk/img/cms/2018/pf/nahodne-procesy-web.pdf strana 23

[7] Obrázok https://cs.wikipedia.org/wiki/Markov%C5%AFv_%C5%99et%C4%9Bzec

[8] Z https://pdfs.semanticscholar.org/3eb3/40d9605a6e46d49302f0be945afcff43da1b.pdf 12.strana

[9] Z https://pdfs.semanticscholar.org/3eb3/40d9605a6e46d49302f0be945afcff43da1b.pdf 12.strana

[10] Príklad z https://cs.wikipedia.org/wiki/Skryt%C3%BD_Markov%C5%AFv_model

[11] Informácie z https://sk.wikipedia.org/wiki/PageRank

[12] Vzorec z https://www.seochat.sk/google-pagerank