Poznámka
Toto nejsou úplné zápisky z přednášky, toto je jen moje příprava k zápočtovému testu a později ke zkoušce.

1 Markovovy řetězce

Definice 1.1 (Markovův řetězec)

Nechť S je nejvýše spočetná množina. Posloupnost $(X_t)_{t=0}^{\infty}$ náhodných veličin s oborem hodnot v S je Markovův řetězec (s diskrétním časem, s diskrétním prostorem a časově homogenní) pokud pro každé $t \geqslant 0$ a každé $a_0, \ldots, a_{t+1} \in S$ platí

$$P(X_{t+1} = a_{t+1} | X_t = a_t \wedge \ldots \wedge X_0 = a_0) = P(X_{t+1} = a_{t+1} | X_t = a_t) = p_{a_t, a_{t+1}},$$

pokaždé, když $P(X_t = a_t \wedge \ldots \wedge X_0 = a_0) > 0$.

Množině S se říká stavy, budeme předpokládat, že jsou nějak (pevně) očíslované přirozenými čísly (resp. přirozenými čísly s 0). $p_{a_t, a_{t+1}}$ je pravděpodobnost přechodu ze stavu a_t do stavu a_{t+1}

1.1 Přechody

Definice 1.2 (Přechodová matice)

Matice P, jejíž prvek $p_{i,j}$ vyjadřuje pravděpodobnost přechodu ze stavu i do stavu j.

Důsledek
Každý řádek přechodové matice má součet jeho prvků roven 1. Tj. $P \cdot (1, \ldots, 1)^T = (1, \ldots, 1)^T$.

Definice 1.3 (Přechodový graf/diagram)

Přechodový graf je ohodnocený orientovaný graf se smyčkami, jehož množina vrcholů je S. Hrana mezi vrcholy $i, j \in S$ vede právě tehdy, když $p_{i,j} > 0$ a má váhu $p_{i,j}$.

Definice 1.4 (Pravděpodobnostní rozdělení X)

Nechť $(X_t)_{t=0}$ je Markovův řetězec. Pravděpodobnostní rozdělení X_t budeme značit $\pi_i^{(t)} = P(X_t = i)$ pro každý stav $i \in S$, $t \in \mathbb{N}_0$. $\pi^{(t)}$ pak značí řádkový vektor hodnot $\pi_i^{(t)}$.

Věta 1.1

Pro libovolný Markovův řetězec s pravděpodobnostním rozdělením π a přechodovou maticí P a libovolné $k \geqslant 0$

$$\pi^{(k)} = \pi^{(0)} \cdot P^k.$$

Dokonce obecněji $\pi^{(t+k)} = \pi^{(t)} P^k$.

Důkaz

$$\forall m \in \mathbb{N} : P(X_m = j) = \sum_{i \in S} P(X_{m-1} = i) \cdot P(X_m = j | X_{m-1} = i),$$
$$\pi_j^{(m)} = \sum_{i \in S} \pi_i^{(m-1)} \cdot P_{i,j},$$
$$\pi^{(m)} = \pi^{(m-1)} \cdot P.$$

□

Definice 1.5 (k-krokový přechod)

$$r_{i,j}(k) = P(\text{přechod z } i \text{ do } j \text{ za } k \text{ kroků}) = P(X_k = j | X_0 = i).$$

Důsledek

$$r_{i,j}(k) = P(X_{t+k} = j | X_t = i).$$

Věta 1.2 (Chapman-Kolmogorov)

Pro libovolný Markovův řetězec a libovolné $k, l \in \mathbb{N}_0$ *platí*

- $r_{i,j}(k) = \left(P^{(k)}\right)_{i,j}$;
- $r_{i,j}(k+l) = \sum_{u \in S} r_{i,u}(k) r_{u,j}(l)$;
- $r_{i,j}(k+1) = \sum_{u \in S} r_{i,u}(k) p_{u,j}$.

1.2 Klasifikace stavů

Definice 1.6 (Dosažitelný stav)

Pro stavy i, j Markovova řetězce říkáme, že j je dosažitelný z i (píšeme $j \in A(i)$ nebo $i \to j$), pokud je nenulová pravděpodobnost, že začínaje v i dosáhneme j v konečném čase. Tedy
$$j \in A(i) \equiv \exists t \in \mathbb{N}_0 : P(X_t = j | X_0 = i) > 0.$$

Poznámka

Nevím, jestli na přednášce bylo $\exists t : P \ldots$ nebo $P(\exists t : \ldots) > 0$. Pokud se nepletu, je to ekvivalentní.

Důsledek
$j \in A(i)$ odpovídá existenci orientované cesty z i do j v přechodovém grafu.

Definice 1.7 (Komutující stavy)

Říkáme, že stavy i, j Markovova řetězce komutují, pokud $i \in A(j)$ a $j \in A(i)$. Píšeme $i \leftrightarrow j$.

Věta 1.3

Pro libovolný Markovův řetězec je relace $\leftrightarrow$ (na S) ekvivalence.

Definice 1.8 (Ireducibilní Markovův řetězec)

Markovův řetězec se nazývá ireducibilní, pokud $\forall i, j \in S : i \leftrightarrow j$.

Definice 1.9 (Rekurentní stav)

Stav $i \in S$ Markovova řetězce se nazývá rekurentní, pokud $\forall j \in A(i) : i \in A(j)$.

Definice 1.10 (Transientní stav)

Stav $i \in S$ Markovova řetězce se nazývá transientní (význam: dočasný, přechodný, pomíjivý), pokud není rekurentní.

Věta 1.4

Pro stav $i \in S$ Markovova řetězce označme $f_{ii} = P(\exists t \in \mathbb{N} : X_t = i | X_0 = i)$. Ať $|S| < \infty$. Potom, když $f_{ii} = 1$, tak je stav rekurentní, pokud $f_{ii} < 1$, tak je transientní.

Důkaz (Transientní)
Označme j to $j \in A(i)$, pro které $i \notin A(j)$. Potom $P(\exists t \in \mathbb{N} : X_t = j | X_0 = i) \neq 0$ a zřejmě $P(\exists t \in \mathbb{N}\ \forall 0 < t_1 < t : X_t = j \wedge X_{t_1} \neq i | X_0 = i) \neq 0$ a $P(\exists t_2 > t : X_{t_2} = i | X_t = j) = 0$, tedy $f_{ii} \neq 1$. □

Důkaz (Rekurentní, na přednášce nebyl celý, moje vize:)
Nechť $m = \min_{j \in A(i)} P(\exists \tilde{t} < t : X_{\tilde{t}} = i | X_0 = j)$. Pro dostatečně velké t (maximum přes všechny časy z definice rekurentního stavu) je $m > 0$. To znamená, že $\sum_{j \in A(i), j \neq i} \pi_j^{(t)} \leqslant (1 - m) \cdot \sum_{j \in A(i), j \neq i} \pi_j^{(0)}$ (předpokládaje, že $p_{i,i} = 1$, protože při libovolném navštívení i jsme vyhráli). Tedy (stále předpokládaje $p_{i,i} = 1$)

$$\lim_{n \to \infty} \sum_{j \in A(i), j \neq i} \pi^{(n \cdot t)} \leqslant \lim_{n \to \infty} (1 - m)^n \cdot \sum_{j \in A(i), j \neq i} \pi^{(0)} = 0 \cdot \ldots = 0.$$

Ale pokud jsme začínali uvnitř $A(i)$ (což po rozutečení se z i rozhodně), tak $\sum_{j \in S \setminus A(i)} \pi_j^{(\cdot)} = 0$, tedy $\pi_1^{(\cdot)} \to 1$. □

Definice 1.11 (Počet návštěv)

Pro stav $i \in S$ Markovova řetězce označme náhodnou veličinu V_i s oborem hodnot v $\mathbb{N}_0^*$ počet návštěv i, tedy $V_i = |\{t | X_t = i\}|$.

Věta 1.5

Stav $i \in S$ Markovova řetězce je rekurentní $\implies P(V_i = \infty | X_0 = i) = 1$. i je transientní, pokud $V_i|_{X_0=i} \sim Geo(1 - f_{ii})$.

Definice 1.12 (Stacionární rozložení)

Nechť π je pravděpodobnostní rozložení na stavech S Markovova řetězce. Řekneme, že π je stacionární rozložení, pokud $\pi \cdot P = \pi$, kde π považujeme za řádkový vektor.

Důsledek
Pokud $\pi^{(0)}$ je stacionární rozložení, pak $\forall k \in \mathbb{N}_0 : \pi^{(k)} = \pi^{(0)}$.

Definice 1.13 (Periodický stav, periodický Markovův řetězec, aperiodický ...)

Stav $i \in S$ Markovova řetězce je periodický, pokud $\exists \Delta \in \mathbb{N} \setminus \{1\}$:

$$P(X_t = i | X_0 = i) > 0 \implies \Delta | t.$$

Markovův řetězec se nazývá periodický, pokud jsou všechny jeho stavy periodické.

Stav nebo Markovův řetězec se nazývá aperiodický, pokud není periodický.

Věta 1.6

Buď $(X_t)_{t=0}^{\infty}$ Markovův řetězec, který je ireducibilní, aperiodický a $|S| < \infty$. Potom $\exists \pi$ stacionární rozložení a

$$\forall j \ \forall i \lim_{k \to \infty} r_{i,j}(k) = \pi_j;$$

navíc π je jednoznačné řešení $\pi \cdot P = \pi$ a $\pi \cdot (1, \ldots, 1)^T = 1$.

Definice 1.14 (Absorbující stav)

Stav $a \in S$ Markovova řetězce je absorbující, pokud $p_{a,a} = 1$.

Definice 1.15 (Čas absorbování)

Předpokládejme $A \subseteq S$ neprázdnou množinu absorbujících stavů Markovova řetězce a BÚNO $0 \in A$. Pro každý stav $i \in S$ definujeme μ_i jako střední hodnotu času absorbování z i, tedy

$$\mu_i = \mathbb{E}(T | X_0 = i), \qquad T = \min \{t : X_t \in A\}.$$

Dále a_i buď pravděpodobnost, že začínaje ve stavu i skončíme v stavu 0.

$$a_i = P(\exists t : X_t = 0 | X_0 = i).$$

Věta 1.7

Pravděpodobnosti a_i jsou jednoznačné řešení

$$a_0 = 1, \qquad a_i = 0, \quad 0 \neq i \in A, \qquad a_i = \sum_{j \in S} p_{i,j} a_j, \quad i \in (S \backslash A) \cup \{0\}.$$

Důkaz
TODO? Jednoduchý, větou o úplné pravděpodobnosti. □

Věta 1.8

Střední hodnoty času (μ_i) jsou jednoznačné řešení

$$\mu_i = 0, \quad i \in A, \qquad \mu_i = 1 + \sum_{j \in S} p_{i,j} \mu_j, \quad i \in S \backslash A.$$

Důkaz
TODO? Jednoduchý, větou o úplné střední hodnotě. □

2 SAT

Definice 2.1 (k-SAT)

Je konjukce (φ) l klauzulí (= disjunkce nejvýše k literálů = proměnná nebo její negace) splnitelná (vhodným dosazením ano/ne za proměnné)? (Proměnné označme $x_1, \ldots, x_n$.)

Definice 2.2 (Algoritmus řešení 2-SAT)

Začneme z libovolně přiřazenými proměnnými (např. všechny ne). Následně $2 \cdot m \cdot n^2$-krát (pro zvolené $m \in \mathbb{N}$) zopakujeme: pokud je vše splněno, vyhráli jsme; jinak zvolíme libovolně nesplněnou klauzuli a z ní změníme náhodně proměnnou a znegujeme jí (tím jsme danou klauzuli splnili). Pokud po $2 \cdot m \cdot n^2$ krocích není hotovo, pak vrátíme ne.

Tvrzení 2.1

Pravděpodobnost špatného výsledku je menší než $\frac{1}{nm}$.

Důkaz
Předpokládejme, že $\varphi(s_1, \dots, s_n)$ je pravdivá a položme $X_t = |\{x_i^t = s_i\}$. Tedy pokud $X_t = n$, tak jsme našli splnění φ.

Pokud $X_t = 0$, pak $X_{t+1} = 1$. Pokud $0 < X_t < n$, pak ve vybrané klauzuli máme minimálně jednu ze dvou proměnných špatně ($x_i \neq s_i$). Když změníme správnou, tak $X_{t+1} = X_t + 1$. Pokud zvolíme druhou, tak ona mohla být také správně, takže $X_{t+1} = X_t \pm 1$.

Tím dostáváme Markovův řetězec tvaru $n+1$ dlouhé cesty, kde pravděpodobnost cesty doprava je alespoň 1/2. Tento řetězec jsme (prý) už analyzovali, vyjde nám, že střední hodnota příchodu do posledního vrcholu je menší než n^2.

Tím nám z Markovovy nerovnosti vychází, že $P(T > 2mn^2) \leqslant \frac{1}{2m}$, kde T je nejmenší tak, že $X_T = n$. □

Tvrzení 2.2

Pravděpodobnost špatného výsledku je menší než $\frac{1}{2^m}$.

Důkaz
m-krát zopakujeme postup pro „$m = 1$“ (začátek volíme libovolně, takže je nám jedno, že předchozí iterace nám dala nějaký stav, ze kterého pokračujeme). □

Poznámka
Když toto aplikujeme na 3-SAT, tak budeme mít problém s tím, že pravděpodobněji půjdeme doleva místo doprava. Tudíž musíme něco zlepšit.

Definice 2.3 (Algoritmus pro řešení 3-SAT)

Zopakujeme $2 \cdot 2 \cdot 3^{n/2}$ krát: náhodně zvolíme začátek a $n/2$-krát zopakujeme krok z 2-SATu.

Tvrzení 2.3

Špatnou odpověď dá tento algoritmus s pravděpodobností $\frac{1}{2}$.

Důkaz
V každém z $2 \cdot 2 \cdot 3^{n/2}$ kroků (začínáme náhodně, tedy X_0 má binomické rozdělení)

$$P(win) = P(X_0 \geqslant n/2) \cdot P(win|X_0 \geqslant n/2) \geqslant \frac{1}{2}3^{-n/2}.$$

Tedy střední hodnota opakování vnějšího cyklu je $\frac{1}{p} = 2 \cdot 3^{n/2}$. A my víme, že $P(T > 2 \cdot 2 \cdot 3^n) \leqslant \frac{\mathbb{E}T}{2 \cdot 2 \cdot 3^n} \leqslant \frac{1}{2}$. □

3 Bayesovská statistika

3.1 Postup

Definice 3.1 (Parametr hledaného rozdělení)

Hledáme rozdělení s parametrem Θ, který budeme považovat za náhodnou veličinu.

Definice 3.2 (Apriorní rozdělení)

Nejprve vybereme apriorní rozdělení s pmf (probability mass function) $p_\Theta(\vartheta)$ nebo pdf (probability density function) $f_\Theta(\vartheta)$ náhodné veličiny Θ nezávisle na datech.

Definice 3.3 (Statistický model)

Potom zvolíme statistický model $p_{X|\Theta}(x|\vartheta)$ (nebo $f_{X|\Theta}(x|\vartheta)$), který popisuje jak jsou (věříme, že jsou) rozděleny data, pokud je Θ rovno nějakému konkrétnímu ϑ.

Definice 3.4 (Posteriorní rozdělení)

Poté, co pozorujeme $X = x$ (více měření považujeme za pozorování jednoho $X = x$ z vícedimenzionálního rozdělení) spočítáme posteriorní rozdělení $f_{\Theta|X}(\vartheta|x)$.

Poznámka

Nakonec najdeme, co potřebujeme vědět, například a, b tak, aby $P(a \leqslant \Theta \leqslant b|X = x) = \int_a^b f_{(\Theta|X)}(\vartheta|x)d\vartheta \geqslant 1 - \alpha$.

3.2 Bayesova věta

Věta 3.1 (Bayesova pro obě diskrétní)

Nechť X, Θ jsou diskrétní náhodné veličiny, pak

$$p_{\Theta|X}(\vartheta|x) = \frac{p_{X|\Theta}(x|\vartheta)p_\Theta(\vartheta)}{\sum_{\vartheta' \in \operatorname{Im}\Theta \setminus \{p_\Theta(\vartheta')=0\}} p_{X|\Theta}(x|\vartheta')p_\Theta(\vartheta')}.$$

Věta 3.2 (Bayesova pro obě spojitá)

Nechť X, Θ jsou spojité náhodné veličiny, pak

$$f_{\Theta|X}(\vartheta|x) = \frac{f_{X|\Theta}(x|\vartheta)f_\Theta(\vartheta)}{\int_{\vartheta' \in \operatorname{Im}\Theta \setminus \{f_\Theta(\vartheta')=0\}} f_{X|\Theta}(x|\vartheta')f_\Theta(\vartheta')}.$$

Věta 3.3 (Bayesova pro diskrétní a spojité)

Nechť X je diskrétní a Θ spojitá náhodná veličina, pak

$$f_{\Theta|X}(\vartheta|x) = \frac{p_{X|\Theta}(x|\vartheta) f_\Theta(\vartheta)}{\int_{\vartheta' \in \operatorname{Im}\Theta \setminus \{f_\Theta(\vartheta')=0\}} p_{X|\Theta}(x|\vartheta') f_\Theta(\vartheta')}.$$

3.3 Bodové odhady

Definice 3.5 (MAP – maximum a-posteriori)

Zvolíme modus Θ.

Poznámka
Tj. maximum $p_{\Theta|X}(\vartheta|x)$, resp $f_{\Theta|X}(\vartheta|x)$.

Definice 3.6 (LMS – least mean square)

Zvolíme střední hodnotu Θ, tedy $\mathbb{E}(\Theta|X = x)$.

Poznámka
Dostaneme nestranný bodový odhad, který minimalizuje $\mathbb{E}((\Theta - \cdot)^2|X = x)$.

Poznámka (Medián)
Obdobně, když vezmeme medián (tj. m tak, že $P(\Theta \leqslant m|X = x) = \frac{1}{2}$), tak minimalizujeme $\mathbb{E}((\Theta - \cdot)|X = x)$, tento přístup však nebudeme dále používat.

TODO? (Zbytek B. statistiky, speciálně Bayesův klasifikátor.)

4 Stochastické procesy

Poznámka
I Markovovy řetězce jsou vlastně stochastický proces.

4.1 Bernoulliho proces

Definice 4.1 (Bernoulliho proces)

Bernoulliho proces (s parametrem p), píšeme $Bp(p)$, je posloupnost nezávislých náhodných veličin $(X_t)_{t=1}^{\infty}$, kde $X_t \sim Ber(p)$, tedy $p(X_t = 1) = p$ a $p(X_t = 0) = 1 - p$, $\forall t \in \mathbb{N}$.

Důsledek

$$\{X_t\}_{t=1}^{\infty} \sim Bp(p) \implies \{X_t\}_{t=k}^{\infty} \sim Bp(p), \forall k \in \mathbb{N}.$$

$$\{X_t\}_{t=1}^{\infty} \sim Bp(p) \implies \{X_t\}_{t=N}^{\infty} \sim Bp(p),$$

kde N je náhodná veličina závisející pouze na minulosti.

Definice 4.2 (Čas prvního úspěchu, čas k-tého)

$$T := \min\left\{t | X_t = 1\right\}, \qquad T_k := \min\left\{t \middle| \sum_{s=1}^{t} X_s = k\right\}.$$

Důsledek

$$T \sim Geom(p), \qquad \mathbb{E}[T] = \frac{1}{p}, \qquad \operatorname{var} T = \frac{1-p}{p^2}.$$

Definice 4.3 (Doba čekání)

$$L_k := T_k - T_{k-1}, \qquad (T_0 = 0).$$

Důsledek

$$L_k \sim T \sim Geom(p).$$

Důkaz
Restartujeme Bernoulliho proces v T_{k-1}. □

Důsledek

$$T_k = \sum_{i=1}^{k} L_i.$$

$$\mathbb{E}[T_k] = \sum_{i=1}^{k} \mathbb{E} L_i = \frac{k}{p}, \qquad \operatorname{var} T_k = \sum_{i=1}^{k} \operatorname{var} L_i = k \cdot \frac{1-p}{p^2}.$$

$$p(T_k = t) = \binom{t-1}{k-1} \cdot p^k \cdot (1-p)^{t-k}, \qquad \chi(T_k = t) \sim Pas(p, k),$$

kde $Pas(p,k)$ je tzv. Pascalovo rozdělení (definované právě $p(T_k = t) = \dots$ výše), také nazývané negativní binomické.

Věta 4.1 (Spojování Bernoulliho procesů)

Mějme $\{X_t\}_{t=1}^{\infty} \sim Bp(p)$ a $\{Y_t\}_{t=1}^{\infty} \sim Bp(q)$, pak $\{X_t \vee Y_t\}_{t=1}^{\infty} \sim Bp(p+q-pq)$.

Věta 4.2 (Rozdělování Bernoulliho procesů)

Mějme $\{Z_t\}_{t=1}^{\infty} \sim Bp(p)$. Potom $\{Z_t \cdot Y_t\}_{t=1}^{\infty} \sim Bp(p \cdot q)$, kde $Y_t \sim Ber(q)$ jsou navzájem nezávislé (a nezávislé na Z_t).

4.2 Poissonův proces

Definice 4.4 (Poissonův proces)

Definujme časy příchodů jako reálná čísla: $0 < T_1 < T_2 < T_3 < \dots$. Po Poissonově procesu požadujeme:

1. Pro každou délku intervalu τ chceme, aby pravděpodobnost k příchodů v tomto intervalu byla stejná, označme ji $p(k,\tau)$.
2. Počet příchodů v intervalu $[a,b]$ je nezávislý na počtu příchodů v $[0,a]$.
3. $p(0,\tau) = 1 - \lambda\tau + o(\tau)$, $p(1,\tau) = \lambda\tau + o(\tau)$ ($\implies p(k,\tau) = o(\tau)$, $\forall k \geqslant 2$).

Poissonův proces je tedy posloupnost náhodných reálných veličin $0 < T_1 < T_2 < T_3 < \dots$, která splňuje tyto 3 body.

Definice 4.5 (Počet příchodů do času t)

$$N_t := \max k | T_k \leqslant t$$

Věta 4.3

$$N_t \sim Pois(\lambda \cdot t), \qquad p(N_t = k) = e^{-\lambda \cdot t} \frac{(\lambda \cdot t)^k}{k!}.$$

Důkaz
Rozdělme si interval $[0,t]$ na l intervalů pro nějaké l velké. Pak délka jednoho intervalu je $\frac{t}{l}$, $p\left(1,\frac{t}{l}\right)=\frac{\lambda\cdot t}{l}+o\left(\frac{t}{l}\right)$ a $p\left(k,\frac{t}{l}\right)=o\left(\frac{t}{l}\right)$. $o\left(\frac{t}{l}\right)$ zanedbáme, tedy máme Binomické rozdělení s parametry l a $\frac{\lambda\cdot t}{l}$, což pro rostoucí l vede k Poissonovu rozdělení s parametrem $\lambda\cdot t$. Tedy

$$p(N_t=k)=e^{-\lambda\cdot t}\frac{(\lambda\cdot t)^k}{k!}.$$

□

Definice 4.6 (Čekání na další příchod)

$$L_k:=T_k-T_{k-1}.$$

Důsledek

$$p(L_k\geqslant t)=p(0,t)=e^{-\lambda\cdot t},\qquad p(L_k\leqslant t)=1-p(L_k\geqslant t)=1-e^{-\lambda\cdot t}.$$
$$L_k\sim Exp(\lambda).$$

Důsledek

$$\mathbb{E}T_k=\sum_{i=1}^{k}\mathbb{E}L_i=k\cdot\frac{1}{\lambda}.$$

$$\operatorname{var}T_k=\sum_{i=1}^{k}\operatorname{var}L_i=k\cdot\frac{1}{\lambda^2}.$$

$$f_{T_k}(t)=\frac{\lambda^k t^{k-1}e^{-\lambda\cdot t}}{(k-1)!}$$

Věta 4.4 (Rozdělování Poissonových procesů)

Mějme $0<T_1<T_2<\dots$ Poissonův proces s parametrem λ a každý příchod nezávisle s pravděpodobností p ponechejme. Pak nová $0<T_1'<T_2'<\dots$ jsou Poissonův proces s parametrem $\lambda\cdot p$. Odstraněné $0<\tilde{T}_1<\tilde{T}_2<\dots$ jsou Poissonův proces s parametrem $\lambda\cdot(1-p)$. A tyto procesy jsou na sobě nezávislé.

Důkaz

$$p_p(k,\tau) = \sum_{n=k}^{\infty} p(n,\tau) \cdot P(Bin(n,p) = k).$$

Následně se ověří podmínky Poissonova procesu (na přednášce ukázán trochu zjednodušený výpočet).

Nezávislé $\Leftrightarrow$ $P(X = k \wedge Y = l) = P(X = k) \cdot P(Y = l)$. Následně jsme ověřili dosazením. □

Věta 4.5 (Spojování Poissonových procesů)

Nechť $0 < T_1 < T_2 < \dots$ a $0 < S_1 < S_2 < \dots$ jsou Poissonovy procesy s parametry λ, $\varkappa$. Potom jejich sjednocením získáme Poissonův proces $0 < R_1 < R_2 < \dots$ s parametrem $\lambda + \varkappa$. (Případně můžeme spojovat i libovolně mnoho Poissonových procesů do Poissonova procesu s parametrem rovným součtu parametrů původních.)

Důkaz

$$p(R_1 > t) = P(T_1 > t \wedge S_1 > t) = P(T_1 > t) \cdot P(S_1 > t) = e^{-\lambda t} \cdot e^{-\varkappa t} = e^{-(\lambda+\varkappa)t}.$$

Následně restartujeme procesy v R_1 a začínáme nanovo :) □

5 Balls and bins

Definice 5.1 (Narozeninový paradox)

k lidí, jaká je pravděpodobnost, že dva lidé mají narozeniny ve stejný den?

Poznámka
Řešení:

$$P(\text{každý v jiný den}) = \left(1 - \frac{1}{365}\right) \cdot \left(1 - \frac{2}{365}\right) \cdot \ldots \cdot \left(1 - \frac{k-1}{365}\right) \approx$$

$$(e^{-x} \approx 1 - x) \qquad \approx \prod_{i=1}^{k-1} e^{-\frac{i}{365}} = e^{-\sum_{i=1}^{k-1} \frac{i}{365}} = e^{-\frac{k\cdot(k-1)}{2\cdot 365}}.$$

Definice 5.2 (Balls and bins)

Máme m kuliček, které rozdělíme do n příhrádek.

Například
Můžeme se ptát na:

- narozeninový paradox;
- # kuliček v první příhrádce ($\sim Bin(m, 1/n)$);
- první příhrádka prázdná ($(1 - 1/n)^m \approx e^{-m/n}$);
- # prázdných příhrádek ($\mathbb{E} = n \cdot (1 - 1/n)^m \approx n \cdot e^{-m/n}$);
- průměrný počet kuliček v příhrádce (m/n);
- maximální počet kuliček v příhrádce (následující věta);
- ...

Věta 5.1

Pokud $m = n$ je velké, $M := \frac{3 \log n}{\log \log n}$, pak

$$P(\textit{maximální počet kuliček} \geqslant M) < \frac{1}{n}.$$

Důkaz

$$P(\text{počet kuliček v příhrádce } 1 \geqslant M) \leqslant P(Bin(n, 1/n) = M) =$$
$$= \binom{n}{M} \frac{1}{n^M} \left(1 - \frac{1}{n}\right)^{n-M} < \binom{n}{M} \frac{1}{n^M} = \frac{n \cdot (n-1) \cdot \ldots \cdot (n - (M-1))}{M! \cdot n^M} < \frac{1}{M!} < \left(\frac{e}{M}\right)^M.$$

$$P(\#\text{ kuliček v nějaké příhrádce} \geqslant M) \leqslant \sum_{i=1}^{m} P(\#\text{ kuliček v příhrádce } i \geqslant M) = n \cdot \left(\frac{e}{M}\right)^M.$$

Chtěli bychom $n \cdot \left(\frac{e}{M}\right)^M < \frac{1}{n}$. Tedy přidáme logaritmus:

$$\log n + M \cdot (1 - \log M) < -\log n$$

$$2 \log n + \frac{3 \log n}{\log \log n}(1 - \log 3 - \log \log n + \log \log \log n) < 0$$

$$-(\log n) \cdot \left(1 - 3\frac{1 - \log 3}{\log \log n} - 3 \cdot \frac{\log \log \log n}{\log \log n}\right) < 0.$$

A jelikož $\frac{\log x}{x} \to 0$ pro $x \to \infty$, tak pro dostatečně velká n nerovnost platí. □

Důsledek (Bucketsort)
Chceme setřídit $n = 2^k$ l-bitových („náhodných“) čísel. Rozdělíme čísla na prvních k bitů ($b(x)$) a zbylých $l - k$ bitů. Potom za prvé roztřídíme čísla podle $b(x)$ do příhrádek $(1, \dots, 2^k)$. Následně setřídíme každou příhrádku (např. bubblesortem) v kvadratickém čase. Nakonec slijeme příhrádky dohromady.

Střední hodnota složitosti tohoto algoritmu je lineární.

Důkaz
První krok je lineární v n, stejně tak třetí. Po prvním kroku bude # kuliček v i-té příhrádce $\sim Bin(n, 1/n)$. Tedy složitost (ve střední hodnotě) kroku dva bude (c je konstanta z bubblesortu)

$$\mathbb{E}\sum_{i=1}^{n} c \cdot X_i^2 = \sum_{i=1}^{n} c \cdot \mathbb{E}(X_i^2) = n \cdot c \cdot \big(\operatorname{var} X_i + (\mathbb{E}X)^2\big) < 2n \cdot c.$$

□

Důsledek (Hešování)
Chceme n objektů (např. řetězců) ukládat tak, aby šlo rychle hledat. Předpokládáme, že máme hashovací funkci (zobrazení z objektů do $[0, m-1] \cap \mathbb{N}$), která je „náhodná“.

Pokud je přibližně $n < \sqrt{m}$, potom pravděpodobnost kolize (2 objekty mají stejný hash) je přibližně $\frac{1}{2}$ z narozeninového paradoxu.

Pokud je $m = n$ dostatečně velké, pak pravděpodobnost, že maximální počet objektů v příhrádce překoná $M := \frac{3 \log n}{\log \log n}$ je menší než $\frac{1}{n}$ z předchozí věty.

Očekávaný čas na nalezení prvku je ve všech případech $\frac{n}{m}$, neboť očekávaný počet objektů v příhrádce je $\frac{n}{m}$.

Maximální čas nalezení bude pro $n = m$ dostatečně velká, nejvýše M s pravděpodobností větší než $1 - \frac{1}{n}$. (Moc lépe to nejde kvůli následující větě.)

Věta 5.2

Za předpokladu dostatečně velkého $m = n$ a $M_2 = \frac{\log n}{\log \log n}$ je

$$P(\textit{maximální počet kuliček} \geqslant M_2) < \frac{1}{n}.$$

Definice 5.3 (Značení)

$$X_i^{(m)} = \#\ \text{kuliček v } i\text{-té příhrádce}.$$

To znamená $(X_1^{(m)}, \ldots, X_n^{(m)})$ má multinomické rozdělení, tj. (pro $\sum k_i = m$, $0 \leqslant k_i \leqslant m$)

$$P\left(X_1^{(m)} = k_1, \ldots, X_n^{(m)} = k_n\right) = \binom{m}{k_1, \ldots, k_n} \cdot \frac{1}{n^m} = \frac{m!}{k_1! \cdot \ldots \cdot k_n!} \cdot \frac{1}{n^m}.$$

Také to znamená, že $X_i^{(m)}$ má rozdělení $Bin(m, 1/n)$, což je přibližně $Pois(m/n)$.

Věta 5.3

Nechť $m, n \in \mathbb{N}$, $Y_1^{(k)}, \ldots, Y_n^{(k)}$ jsou nezávislé stejně rozdělené veličiny s rozdělením $Pois(k/n)$ a $X_i^{(m)}$ jako v předchozím. Pak rozdělení $X_i^{(m)}$ je shodné s rozdělením $Y_i^{(k)}$, pokud $\sum_{i=1}^n Y_i^{(k)} = m$.

Důkaz

Mějme $k_1 + \ldots + k_n = m$ a $0 \leqslant k_i \leqslant m$, potom chceme

$$P\left(X_1^{(m)} = k_1, \ldots, X_n^{(m)} = k_n\right) = P_X = P_Y = P\left(Y_1^{(k)} = k_1, \ldots, Y_n^{(k)} = k_n \middle| \sum Y_i^{(k)} = m\right).$$

$$P_X = \binom{m}{k_1, \ldots, k_n} \cdot \frac{1}{n^m}. \qquad P_Y = \frac{P(\ldots|)}{P(|\ldots)} = \frac{A}{B}.$$

$$A = P(Y_1^{(k)} = k_1) \cdot \ldots \cdot P(Y_n^{(k)} = k_n) = e^{-\frac{k}{n}} \cdot \frac{\left(\frac{k}{n}\right)^{k_1}}{k_1!} \cdot \ldots \cdot e^{-\frac{k}{n}} \cdot \frac{\left(\frac{k}{n}\right)^{k_n}}{k_n!} = e^{-k} \cdot \left(\frac{k}{n}\right)^m \cdot \frac{1}{k_1! \cdot \ldots \cdot k_n!}.$$

$$\sum_{i=1}^n Y_i^{(k)} \sim Pois\left(\frac{k}{n} + \ldots + \frac{k}{n}\right) = Pois(k) \implies B = e^{-k}\frac{k^m}{m!}.$$

□

Věta 5.4

Buďte X, Y jako v předchozí větě a $f(x_1, \ldots, x_n) \geqslant 0$. Potom $\mathbb{E}f(X_1^{(m)}, \ldots, X_n^{(m)}) \leqslant \mathbb{E}f(Y_1^{(k)}, \ldots, Y_n^{(k)}) \cdot e \cdot \sqrt{k}$.

Navíc pokud je pravá strana monotónní v m, pak můžeme $e \cdot \sqrt{k}$ nahradit 2.

Důkaz

$$\mathbb{E}f(Y_1^{(k)},\dots,Y_n^{(k)}) = \sum_{i=0}^{\infty}\mathbb{E}\left(f(Y_1^{(k)},\dots,Y_n^{(k)})|\sum_{j=1}^{n}Y_j^{(k)}=i\right)\cdot P\left(\sum_{j=1}^{n}Y_j^{(k)}=i\right)\geqslant$$

$$\geqslant\left(f(Y_1^{(k)},\dots,Y_n^{(k)})|\sum_{j=1}^{n}Y_j^{(k)}=m\right)\cdot P\left(\sum_{j=1}^{n}Y_j^{(k)}=m\right)=$$

$$=\mathbb{E}f(X_1^{(m)},\dots,X_n^{(m)})\cdot P\left(\sum_{j=1}^{n}Y_j^{(k)}=m\right)=\mathbb{E}f(X_1^{(m)},\dots,X_n^{(m)})\cdot e^{-k}\frac{k^m}{m!}\geqslant$$

$$\geqslant\mathbb{E}f(X_1^{(m)},\dots,X_n^{(m)})\cdot\frac{1}{e\sqrt{k}}.$$

TODO!!!

Totéž provedeme pro monotónní pravou stranu, jen budeme odhadovat lepší pravděpodobností? (TODO?) □

Důkaz (Předpředchozí věty)
Z předchozí věty (aplikované na $f(x_1,\dots,x_n)=(\max\{x_1,\dots,x_n\}<M))$ nám stačí dokázat, že

$$P(\max\left\{Y_1^{(k)},\dots,Y_n^{(k)})\right\}<M)\leqslant\frac{1}{e\cdot\sqrt{k}\cdot n}.$$

$$P(\dots)=P(Y_1^{(k)}<M)\cdot\ldots\cdot P(Y_n^{(k)}<M)\leqslant(1-P(Y_1^{(k)}=M))\cdot\ldots\cdot(1-P(Y_n^{(k)}=M))=$$

$$=\left(1-e^{-1}\frac{1^M}{M!}\right)^n\approx\left(e^{-\frac{1}{e\cdot M!}}\right)^n\leqslant e^{-\frac{n}{e\cdot M!}}<\frac{1}{n^2},$$

neboť to je totéž jako

$$\frac{1}{e\cdot M!}>2\log n,$$

což spočítáme pomocí odhadu $M!\leqslant M\cdot(M/e)^M$. □

6 Neparametrická statistika

Definice 6.1 (Neparametrická statistika)

Nemáme model (rozdělení závisející na parametru).

TODO (Permutační test)

Definice 6.2 (Permutační test)

Mějme data $x_1, \dots, x_n$ a $y_1, \dots, y_m$ (např. testovací a kontrolní vzorek). Dále mějme f, které rozhoduje, zda dané $z_1, \dots, z_{m+n}$ splňuje nulovou hypotézu.

$$\mathcal{F} := \{f(\pi(z))\}_{\pi \in S_{n+m}}$$

p-hodnota je podíl prvků souboru $\mathcal{F}$, které splňují nulovou hypotézu. Nulovou hypotézu zamítneme, pokud je tento podíl menší než α.

(Požadujeme, aby za nulové hypotézy byla pravděpodobnost každého prvku $\mathbb{F}$ stejná.)

Definice 6.3 (Permutační test ++)

Pokud nemůžeme počítat f pro všechny $\pi \in S_{n+m}$, nasamplujeme $\mathcal{F}^* \subset \mathcal{F}$.

Definice 6.4 (Znamínkový test)

$X_1, \dots, X_n$ nezávislé náhodné veličiny z neznámého spojitého rozdělení symetrické podle střední hodnoty. Nulová hypotéza je, že střední hodnota je 0.

Nechť $Y_i = \operatorname{sgn}(X_i) = +1$ nebo 0 (pozor, ne -1). Potom při předpokladu nulové hypotézy $Y = \sum_{i=1}^n Y_i \sim Binom(n, \frac{1}{2})$. Tedy nulovou hypotézu zamítneme, pokud $Y \leqslant Y_{\alpha/2}$ nebo $Y > Y_{1-\alpha/2}$, kde $P(Binom(n, \frac{1}{2}) < Y_x) = x$.

Definice 6.5 (Pair test)

Mějme data, která jsou přirozeně v párech (např. hodnota před a po vylepšení algoritmu) a mějme nějakou hypotézu, kterou můžeme testovat po prvcích (např. jestli se průměr nových a starých hodnot shoduje, což můžeme testovat jako „jestli je průměr rozdílů hodnot 0“). Potom se můžeme na pár dívat jako na jeden prvek.

Definice 6.6 (Wilcoxonův test znamínka hodnosti)

$X_1, \dots, X_n$ nezávislé náhodné veličiny z neznámého spojitého rozdělení symetrické podle střední hodnoty. Nulová hypotéza je, že střední hodnota je 0.

Hodnost (rank, r_i) je pořadí v seřazení $|X_i|$ (místo sdíleného pořadí vezmeme průměr sdílených míst, to se ve skutečnosti v spojitém rozdělení nemůže stát). Definujeme

$$T := (W :=) \sum_{i=1}^n r_i \cdot \operatorname{sgn}(X_i) = T^+ - T^-.$$

Zamítneme nulovou hypotézu, pokud T je moc velké nebo moc malé, tj. $T < Y_{\alpha/2}$ nebo $T > Y_{1-\alpha/2}$ ve správném (TODO?) rozdělení.

Definice 6.7 (Mannův–Whitneyho U-test)

Máme dvě množiny $X_1, \dots, X_n$ a $Y_1, \dots, Y_m$.

$$U := \sum_{i=1}^{n} \sum_{j=1}^{m} S(X_i, Y_i), \qquad S(X,Y) := \begin{cases} 0, & X > Y, \\ \frac{1}{2}, & X = Y, \\ 1, & X < Y. \end{cases}$$

Nulová hypotéza je $P(X < Y) = P(Y < X)$.

TODO!!! (Simpson paradox)

7 Moment generating function

Definice 7.1 (Moment generating function (MGF))

Pokud X je náhodná veličina a $s \in \mathbb{R}$, potom $M_X(s) := \mathbb{E}(e^{sX})$.

Věta 7.1

$$M_X(s) = \sum_{k=0}^{\infty} \mathbb{E}(X^k) \frac{s^k}{k!}. \qquad \textit{(Pro } s \textit{ z intervalu, kde je } M_X(s) \textit{ definováno.)}$$

Důkaz

$$\mathbb{E}(e^{s \cdot X}) = \mathbb{E}\left(\sum_{k=0}^{\infty} \frac{(s \cdot X)^k}{k!} \right) = \sum_{k=0}^{\infty} \mathbb{E}(X^k) \frac{s^k}{k!}.$$

□

Věta 7.2

$$M_{a \cdot X + b} = e^{b \cdot s} M_X(a \cdot s).$$

Důkaz

$$\mathbb{E}\left(e^{s \cdot (a \cdot X + b)}\right) = \mathbb{E}\left(e^{a \cdot s \cdot X} \cdot e^{b \cdot s}\right) = e^{b \cdot s} M_X(a \cdot s).$$

□

Věta 7.3

X a Y nezávislé $\implies M_{X+Y} = M_X \cdot M_Y$.

Důkaz

$$M_{X+Y}(s) = \mathbb{E}\left(e^{s\cdot(X+Y)}\right) = \mathbb{E}\left(e^{s\cdot X} \cdot e^{s\cdot Y}\right) = \mathbb{E}\left(e^{s\cdot X}\right) \cdot \mathbb{E}\left(e^{s\cdot Y}\right) = M_X(s) \cdot M_Y(s).$$

□

Věta 7.4

Pokud $\exists \varepsilon > 0\ \forall s \in (-\varepsilon, \varepsilon) : M_X(s) = M_Y(s) \in \mathbb{R}$, pak $F_X(t) = F_Y(t)\ \forall t \in \mathbb{R}$.

Důkaz

Bez důkazu. □

Věta 7.5

Pokud $\exists \varepsilon > 0\ \forall s \in (-\varepsilon, \varepsilon) : M_{Y_n}(s) \to M_Z(s) \in \mathbb{R}$ a F_Z je spojité, pak $F_{Y_n}(t) \to F_Z(t)$ $\forall t \in \mathbb{R}$ ($Y_n \xrightarrow{D} Z$).

Důkaz

Bez důkazu. □

Věta 7.6 (Centrální limitní věta)

$X_1, X_2, \ldots$ nezávislé stejně rozdělené veličiny, $\mathbb{E}X_i = \mu$, $\operatorname{var} X_i = \sigma^2$, potom

$$Y_n = \frac{X_1 + \ldots + X_n - n \cdot \mu}{\sigma\sqrt{n}}.$$

Potom $Y_n \xrightarrow{D} N(0,1)$.

Důkaz

Použijeme předchozí větu, kde $Z \sim N(0,1)$, $M_Z = e^{\frac{s^2}{2}}$, zřejmě F_Z je spojitá. Můžeme předpokládat, že $\mu = 0$. Také předpokládejme, že $M_{X_i}(s)$ existuje. Potom $Y_n = \frac{X_1+\ldots+X_n}{\sigma\sqrt{n}}$. Tedy

$$M_{Y_n}(s) = M_{X_1+\ldots+X_n}\left(\frac{s}{\sigma\sqrt{n}}\right) = \left(M_{X_1}\left(\frac{s}{\sigma\sqrt{n}}\right)\right)^n = \left(1 + \sigma^2\frac{s^2}{2\sigma^2 \cdot n} + o\left(\frac{s^2}{\sigma^2 \cdot n}\right)\right)^n \approx$$

$$\approx \left(1 + \frac{s^2}{2n}\right)^n \to e^{s^2/2} = M_Z(s).$$

$\approx$ je trochu podvod, ale dokáže se jednoduše zlogaritmováním. □

Věta 7.7 (Chernoffova)

$X_1, \ldots, X_n \sim 1 - 2 \cdot Ber\left(\frac{1}{2}\right)$ jsou nezávislé stejně rozdělené veličiny, $X = X_1 + \ldots + X_n$, $\sigma^2 = \operatorname{var} X = n$, $t > 0$, potom

$$P(X \leqslant t) = P(X \geqslant t) \leqslant e^{-\frac{t^2}{2n}}.$$

Důkaz

Pro libovolné s máme

$$P(X \geqslant t) = P(e^{s\cdot X} \geqslant e^{s\cdot t}) \leqslant \frac{\mathbb{E}e^{s\cdot X}}{e^{s\cdot t}} = \frac{M_X(s)}{e^{s\cdot t}} = \frac{(M_{X_1}(s))^n}{e^{s\cdot t}} = \frac{(e^s + e^{-s})^n}{2\cdot e^{s\cdot t}} =$$

$$= \frac{\left(\sum_{k=0}^{\infty} \frac{s^{2k}}{(2k)!}\right)^n}{e^{s\cdot t}} \leqslant \frac{\left(\sum_{k=0}^{\infty} \frac{(s^2/2)^k}{k!}\right)^n}{e^{s\cdot t}} = \frac{e^{n\cdot s^2/2}}{e^{s\cdot t}} = e^{\frac{n\cdot s^2}{2} - s\cdot t}.$$

Následně dosadíme $s = \frac{t}{n}$. □

TODO(Shannon's coding theorem)