DÚ č.3 - Markovské řetězce s diskrétním časem

Marek Nevole, Jan Novotný

ČVUT - FIT

{nevolmar, novot103}@fit.cvut.cz

15. března 2022

1 Úvod

Ve třetím úkolu z předmětu vybrané statistické metody jsme se zabývali Markovskými řetězci s diskrétním časem. Za reprezentanta byl zvolen Marek Nevole.

$$
\begin{aligned}
K &= 28 \\
L &= 6 \\
X &= ((23KL) \bmod 20) + 1 \\
X &= 5 \\
Y &= ((X + ((5K + 7L) \bmod 19)) \bmod 20) + 1 \\
Y &= 17
\end{aligned}
$$

Výsledkem těchto rovnic jsou názvy vybraných datových souborů. V našem případě budeme pracovat se soubory 005.txt a 017.txt.

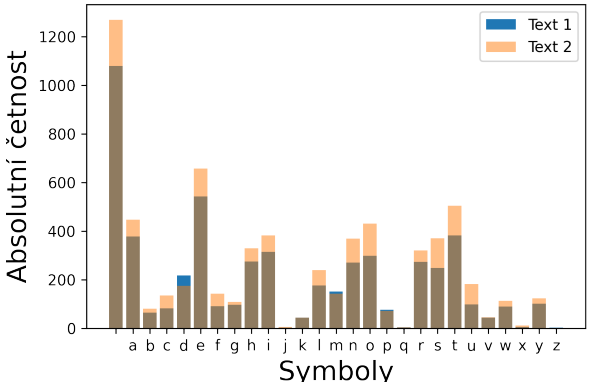
Úkol jsme vypracovali pomocí programovacího jazyka Python[1] v prostředí Jupyter Notebook[2] s volně dostupnými knihovnami SciPy[3], NumPy[4] a Matplotlib[5].

2 Text zadání

Z obou datových souborů načtěte texty k analýze. Pro každý text zvlášť zjistěte absolutní četnosti jednotlivých znaků (symbolů včetně mezery), které se v textech vyskytují. Dále předpokládejme, že první text je vygenerován z homogenního Markovského řetězce s diskrétním časem.
Absolutní četnosti lze pozorovat na obrázku 1.

3 Úloha č.1

*Za předpokladu výše odhadněte matici přechodu Markovského řetězce pro **první** text. Odhadnuté pravděpodobnosti přechodu vhodně graficky znázorněte, např. použitím heatmapy.*



Obrázek 1: Graf absolutních četností symbolů obou textů.

Dle předpokladu je první text vygenerován z homogenního Markovského řetězce s diskrétním časem. Tedy i-tý symbol textu je stav Markovského procesu v čase $i-1$. Dvojice sousedních symbolů znázorňují přechody tohoto procesu.

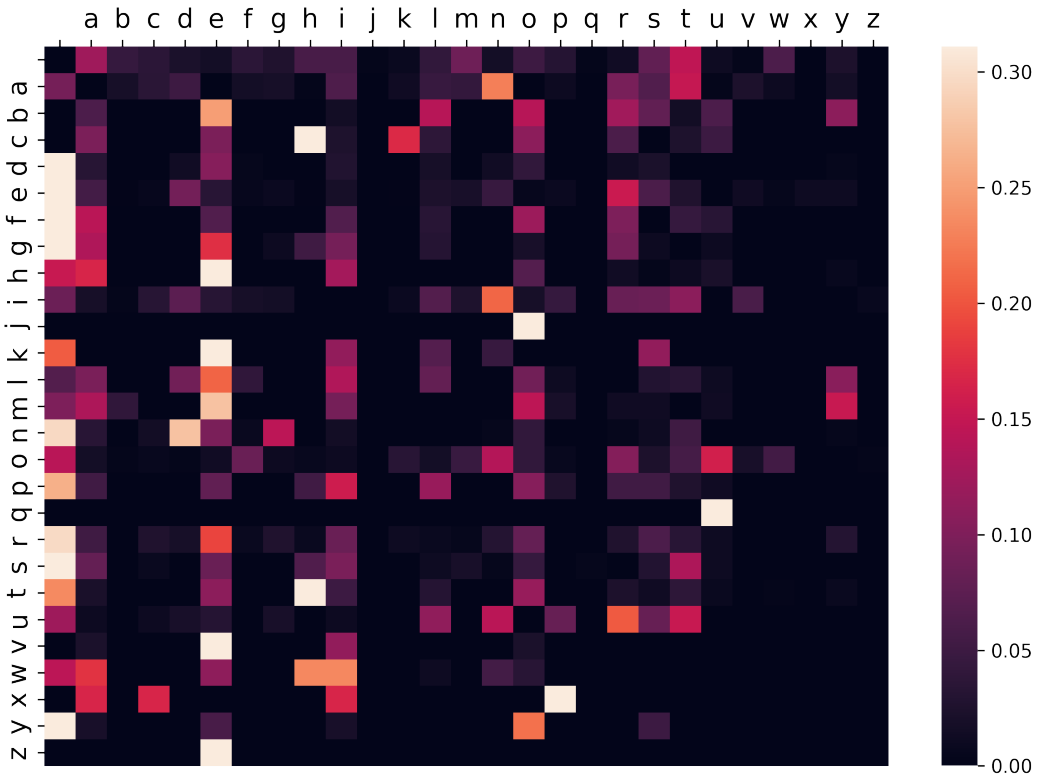
Maximálně věrohodným odhadem matice přechodu Markovského procesu je matice s prvky $\hat{\mathbf{p}}_{ij} = \frac{n_{ij}}{n_{i\bullet}}$, tedy matici přechodu jsem odhadli pomocí četností přechodů. Výslednou matici přechodu $\hat{\mathbf{p}}$ jsme graficky znázornili pomocí zmíněné heatmapy, kterou lze pozorovat na obrázku 2. Matice má rozměry 27×27. 26 za písmena anglické abecedy a 1 za symbol mezery.

4 Úloha č.2

*Na základě matice z předchozího bodu najděte stacionární rozdělení $\boldsymbol{\pi}$ tohoto řetězce pro **první** text.*

Dle typu Markovského procesu, tedy generace textu, jsme odhadli, že množina stavů je již nerozložitelná, tj. všechny dvojice symbolů jsou navzájem dosažitelné. Toto platí pro všechny stavy, tedy celý Markovský řetězec nazýváme nerozložitelný. Dále víme, že v řetězcích s konečným počtem stavů neexistují stavy trvalé nulové. Tedy platí, že všechny stavy v našem zkoumaném Markovském řetězci jsou trvalé nenulové. Také tvrdíme, že jsou aperiodické.

[1] python.org
[2] jupyter.org
[3] scipy.org
[4] numpy.org
[5] matplotlib.org



Obrázek 2: Heatmapa odhadu matice přechodu homogenního Markovského řetězce prvního textu.

Na základě předchozích vlastností a odhadů můžeme využít větu o existenci stacionárního rozdělení a konkrétně její druhý bod, který říká:
Jsou-li všechny stavy trvalé nenulové, stacionární rozdělení $\boldsymbol{\pi}$ *existuje a je jediné.*
Jsou-li navíc všechny stavy aperiodické, platí

$$\boldsymbol{\pi} = \lim_{n \to +\infty} \boldsymbol{p}(n) \text{ pro libovolné } \boldsymbol{p}(0).$$

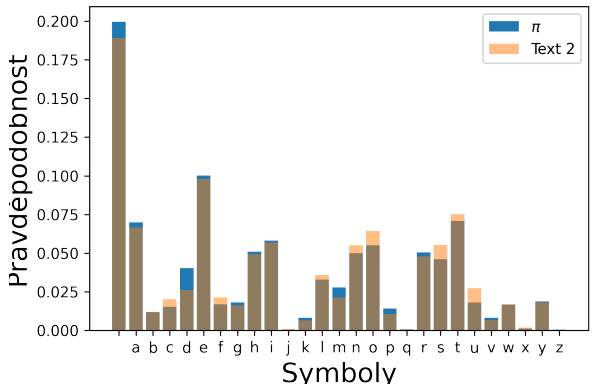
Na základě vlastností homogenního Markovského řetězce a vícekrokových pravděpodobností přechodu platí rovnost $\boldsymbol{p}(n) = \boldsymbol{p}(0) \cdot \mathbf{P}^n$. Tedy pro libovolné $\boldsymbol{p}(0)$, splňující podmínky stacionárního rozdělení, získáme rozdělení $\boldsymbol{\pi}$ následovně $\boldsymbol{\pi} = \boldsymbol{p}(0) \cdot \mathbf{P}^n$. Limitně pro $n \to +\infty$ obsahuje matice $\mathbf{P}^n$ požadované rozdělení $\boldsymbol{\pi}$ ve svých řádcích. Numericky je toto iterativní metoda, tedy provádíme výpočet do té doby, dokud nedosáhneme požadované přesnosti. Pro $n = 10000$ jsme dostali rozdělení, pro které platí $\|\boldsymbol{\pi}\mathbf{P} - \boldsymbol{\pi}\| < 10^{-15}$.

Druhou metodou je využití lineární algebry, vlastností vlastní čísel a vektorů. $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}/^T$ po transponování celé rovnice dostáváme $\mathbf{P}^T\boldsymbol{\pi}^T = \boldsymbol{\pi}^T$, což je velice podobné rovnici vlastních čísel $\mathbf{A}\boldsymbol{x} = \lambda\boldsymbol{x}$, kde $\lambda = 1$. Tedy hledaným rozdělením $\boldsymbol{\pi}$ je vlastní vektor vlastního čísla rovného(nebo velice blízkého k) 1 přeškálovaný, aby platily podmínky $\Sigma_{i \in S}\boldsymbol{\pi}_i = 1$ a $\forall i \in S : \boldsymbol{\pi}_i \geq 0$. Takto získané rozdělení mělo normu od nulového vektoru srovnatelně malou jako rozdělení získané iterativní metodou, tedy $\|\boldsymbol{\pi}\mathbf{P} - \boldsymbol{\pi}\| < 10^{-15}$.

5 Úloha č.3

Porovnejte rozdělení znaků ***druhého*** *textu se stacionárním rozdělením* $\boldsymbol{\pi}$*, tj. na hladině významnosti 5% otestujte hypotézu, že rozdělení znaků* ***druhého*** *testu se rovná rozdělení* $\boldsymbol{\pi}$ *z předchozího bodu.*

Stacionární rozdělení $\boldsymbol{\pi}$ a rozdělení znaků druhého textu lze pozorovat na obrázku 3. Úlohou bylo otestovat, zda se tato dvě diskrétní multinomická rozdělení rovnají. K tomuto jsme využili test dobré shody, v tomto případě se známými parametry. Za nulovou hypotézu jsme postavili rovnost pravděpodobností $H_0 : \boldsymbol{p} = \boldsymbol{\pi}$, alternativní hypotézou byla nerovnost $H_A : \boldsymbol{p} \neq \boldsymbol{\pi}$. Testovou statistikou byla χ^2 statistika. Knihovna SciPy nabízí funkci *chisquare*, která na vstupu bere naměřené a teore-



Obrázek 3: Graf stacionárního rozdělení $\boldsymbol{\pi}$ a relativních četností symbolů druhého textu.

tické četnosti a vrátí testovou statistiku společně s p-hodnotou. Jelikož data nesplňují všude podmínku teoretických četností $np_i \geq 5$, tak jsme si testovou statistiku naprogramovali sami, a pro každý symbol jsme pozorovali, jak tyto četnosti přispívají do výsledné hodnoty. Největšími nárůsty byly symboly d a u, které teoretické četnosti splňovaly, samy zvýšily hodnotu χ^2 o 64.6, což stačí k překročení kritické hodnoty, tedy na základě tohoto pozorování považujeme výsledek za správný i přes nedostatek dat. Dle p-hodnoty zamítáme nulovou hypotézu H_0 ve prospěch alternativní hypotézy H_A, což je silným výsledkem a s velikou pravděpodobností můžeme věřit tomu, že se tato rozdělení nerovnají.

$$\begin{aligned} \chi^2 &= 145.2 \\ \hat{p} &= 1.5 \times 10^{-}18 \\ \chi^2_{0.05,26} &= 38.8 \end{aligned}$$