Modellezés magasabb rendű Markov láncokkal

Egyed Tünde
(Témavezető: Csiszár Villő)

Eötvös Loránd Tudományegyetem
Természettudományi Kar

Önálló projekt 3
2023. június, Budapest

Bevezetés

- Elsőrendű modell helyett magasabbrendű modellek
- Cél: Markov lánc optimális rendjének megtalálása
- LAMP modell
- A rend vizsgálata egy valós mintán különböző módszerekkel

LAMP modell

- n: az állapottér elemszáma
- i_t: a minta t-edik eleme
- ψ: egy eloszlás az $\{1, 2 \ldots, k\}$ halmazon
- a^μ: a μ-edik átmenetmátrix,
- $a^\mu(i_t|i_{t-\mu})$ az $i_{t-\mu}$ i_t átmenetvalószínűsége a μ-lépéses átmenetmátrix szerint
- Ekkor a LAMP modell szerint az átmenetmátrix:

$$P(i_t|i_{t-1}, i_{t-2}, \ldots, i_{t-k}) = \sum_{\mu=1}^{k} \psi(\mu) a^\mu(i_t|i_{t-\mu})$$

Paraméterek becslése

A ψ és a paraméterek becslése EM alogritmussal történik:

$$\psi(\mu) = \frac{\sum_t P(x_t = \mu | I)}{\sum_{t,\nu} P(x_t = \nu | I)}$$

$$a^{\mu}(i'|i) = \frac{\sum_t P(x_t = \mu, i_{t-\mu} = i, i_t = i' | I)}{\sum_t P(x_t = \mu, i_{t-\mu} = i | I)}$$

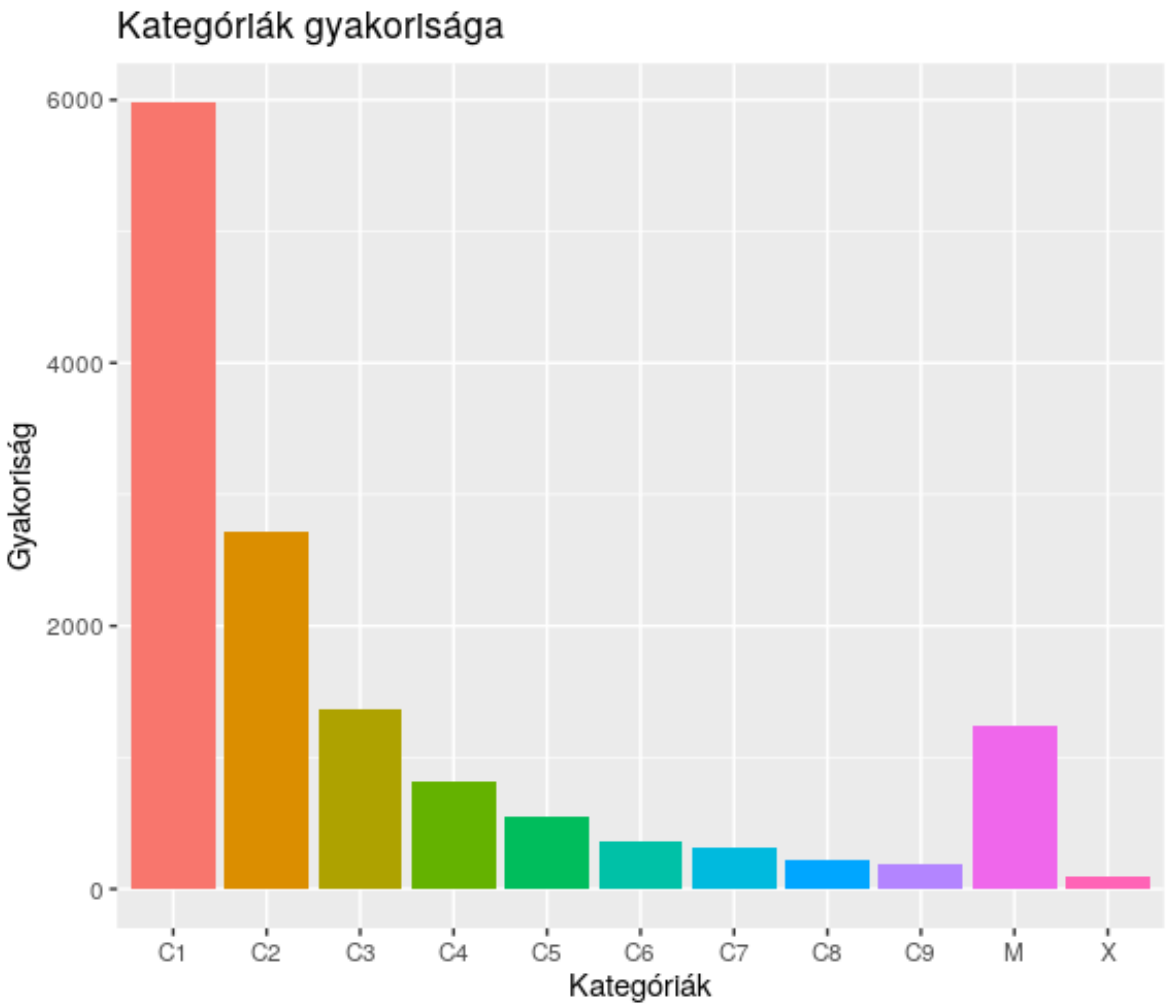
A fenti képleteket a következő összefüggés segítségével számolhatjuk ki:

$$P(x_t = \mu | I) = \frac{\psi(\mu) a^{\mu}(i_t | i_{t-\mu})}{\sum_{\nu} \psi(\nu) a^{\nu}(i_t | i_{t-\nu})}$$

Minta

- A vizsgált minta napkitörések erősségét tartalmazza 2002. január és 2019. május között
- 13 870 napkitörés
- Eredetileg 211 állapotot tartalmazott, amit 11 állapotba vontam össze

Minta


Kategóriák gyakorisága
6000
4000
2000
0
Gyakoriság
C1
C2
C3
C4
C5
C6
C7
C8
C9
M
X
Kategóriák

Likelihood módszer

- Kiszámoljuk a loglikelihood értékeket
- Valószínűség-hányados próba a szignifikancia tesztelésére H_0 : a modell rendje k, H_1 : a modell rendje m hipotézisek mellett a próbastatisztika:

$$ {}_k\eta_m = -2(\log L(\theta_k, x) - \log L(\theta_m, x)) $$

- Ez H_0 esetén χ^2 eloszlású $(|S|^m - |S|^k)(|S| - 1)$ szabadságfokkal, ahol S a lehetséges állapotok halmaza

A másodrendű modell jobb, mint az elsőrendű, de a harmadrendű már nem javít jelentősen.

	$k = 1, m = 2$	$k = 1, m = 3$	$k = 2, m = 3$
${}_k\eta_m$	1 712	7 912	6 200
Szabadságfok	1 100	13 200	12 100
p-érték	0	1	1

Cross validation

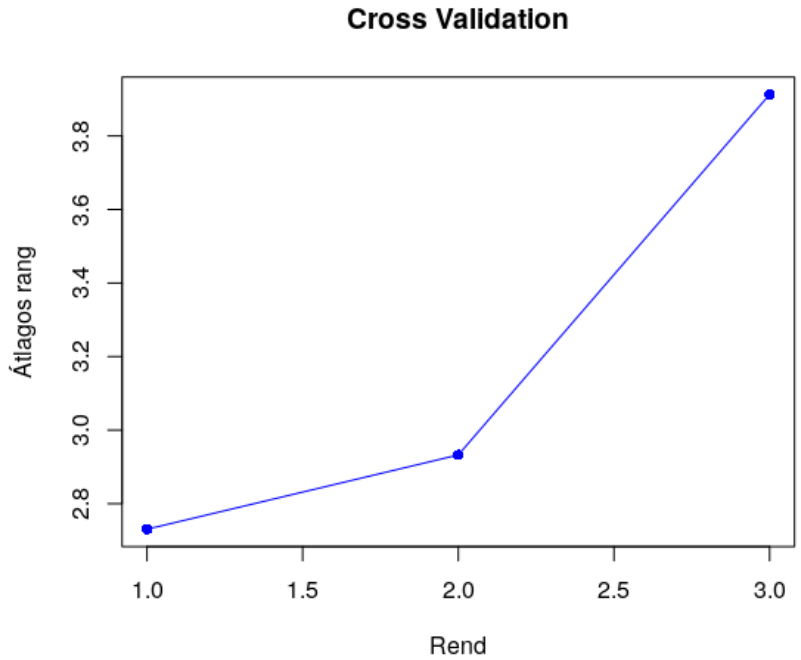
- A mintát két részre osztjuk, egy tanító halmazra és egy validáló halmazra
- A tanítóhalmazon megbecsüljük az átmenetvalószínűségeket
- r_{ij}: az x_i állapotból hányadik legvalószínűbb, hogy x_j-be kerülünk
- n_{ij}: az átmenetek száma x_i-ből x_j-be a validáló halmazon
- Kiszámoljuk az átlagos rangot:

$$\frac{\sum_i \sum_j n_{ij} r_{ij}}{\sum_i \sum_j n_{ij}}$$

- Minél kisebb az átlagos rang, annál jobb a modell

Cross validation

- A tanuló halmazt a minta első 10 000 eleme alkotta, a maradék 3 870 elem került a validáló halmazba.
- Ez a módszer az elsőrendű modellt javasolja.



Kétlépéses visszatérés

- Kétlépéses visszatérési valószínűségek becslése
- Stacionárius eloszlás kiszámítása
- A kétlépéses visszatérési valószínűségek összege a stacionárius eloszlással súlyozva

$$P(X_n = X_{n+2}) = \sum_i \pi(i) \sum_j P(X_{n+2} = i, X_{n+1} = j | X_n = i)$$

- Eredmények összevetése a mintában szereplő kétlépéses visszatérések relatív gyakoriságával

Kétlépéses visszatérés

A mintában annak relatív gyakorisága, hogy két lépés múlva ugyanabba az állapotba kerülünk $0,31$, a különböző rendű modellekkel lenti táblázatban szereplő értékeket kapjuk. Ez alapján a másdod- vagy harmadrendű modell tűnik a legjobbnak.

Order	Two-Step Return
1	0.27
2	0.31
3	0.31

Entrópia

- Az első rendű Markov folyamat betűnkénti entrópiája az alábbi képlettel számolandó:

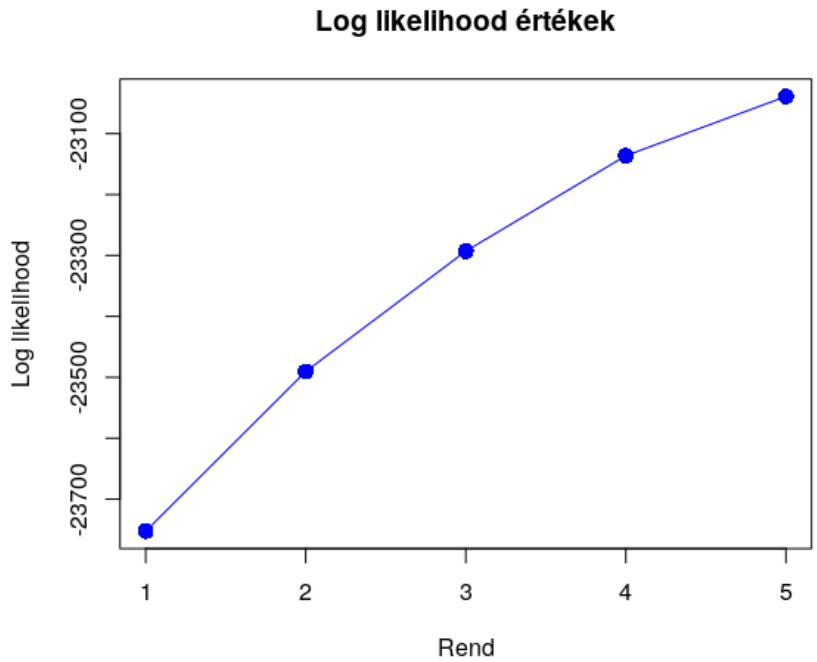
$$H(X_{t+1}|X_t) = -\sum_{j,k} \pi(j) p_{jk} \log(p_{jk})$$

  ahol π a stacionárius eloszlás.
- Az optimális rend megtalálását jelzi, ha a rend további növelésével már nem csökken tovább az entrópia.
- Ennél a módszernél azt látjuk, hogy még a harmadrendű modell esetén is jelentősen csökken az entrópia.

Order	Entropy Rate
1	0.34
2	0.14
3	0.07

LAMP loglikelihood

A LAMP modellel kapott loglikelohood értékek az alábbi ábrán láthatóak:



LAMP Informcációs kritériumokkal

- Akaike-féle információs kritérium:

$$AIC = -2 \log L + 2k,$$

ahol k a becsült paraméterek száma, L a modell likelihood függvényének maximum értéke.

- Bayes-féle információs kritérium:

$$BIC = -2 \log L + k \log n,$$

ahol n a minta elemszáma.

LAMP Informcációs kritériumokkal

Akaike-féle és Bayes-féle információs kritériumok:

Modell	Paraméterszám	Loglikelihood	AIC	BIC
Markov 1	110	-23 753	47 725	48 554
Markov 2	1 210	-22 897	50 320	57 334
Markov 3	13 310	-19 797	74 605	166 538
LAMP 1	110	-23 753	47 725	48 554
LAMP 2	221	-23 490	47 423	49 089
LAMP 3	332	-23 293	47 250	49 752
LAMP 4	443	-23 136	47 158	50 498
LAMP 5	554	-23 039	47 186	51 365

- Akaike-féle információs kritérium alapján a negyedrendű LAMP modell a legjobb
- Bayes-féle információs kritérium szerint az elsőrendű modell

Köszönöm a figyelmet!