

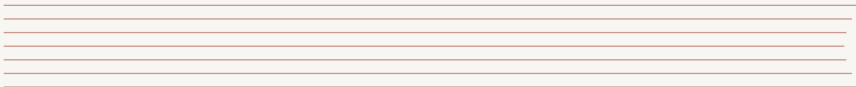
GENERATIVNO MODELIRANJE Z MPS NA TABELARIČNIH PODATKIH

Barbara Makovec

Mentor: doc. dr. Bojan Žunkovič

Fakulteta za računalništvo in informatiko & Fakulteta za matematiko in fiziko

23. julij 2026



UNIVERZA
V LJUBLJANI

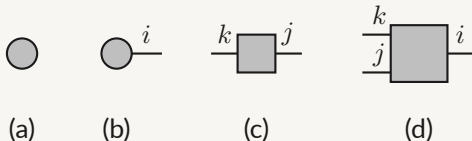


SECTION 1

Tenzorji in tenzorske mreže

KAJ SO TENZORJI?

Tenzor je večdimenzionalni seznam števil. Predstavimo ga kot neko geometrijsko obliko, ki ima r nog, kjer je r red tenzorja.



Skalar: 0 nog, vektor: 1 noga, matrika: 2 nogi, tenzor reda tri: 3 noge

TENZORSKI PRODUKT

Tenzorski produkt je produkt vrednosti tenzorjev po komponentah. V diagramu ga predstavimo kot dva tenzorja drug zraven drugega.

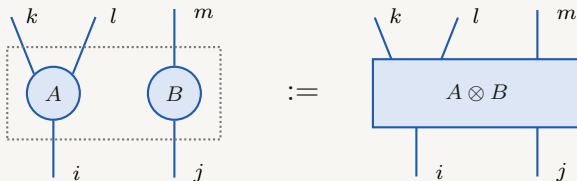


Diagram predstavlja tenzor: $T_{ijklm} = A_{kli}B_{jlm}$.

Tenzorje združujemo in razdružujemo z namenom pretvorbe v matrike ali vektorje, da lahko na njih izvajamo že znane matrične operacije.

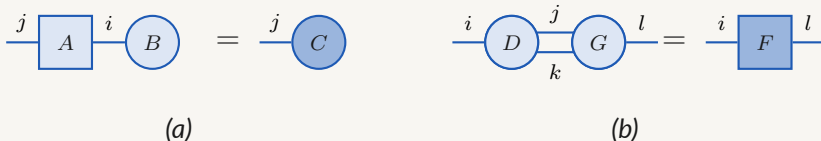


SKRČITEV TENZORJEV

Skrčitev tenzorjev je seštevanje po skupnem indeksu. Če imata tenzorja A in B skupen indeks α z enako dimenzijo, je njuna skrčitev tenzor:

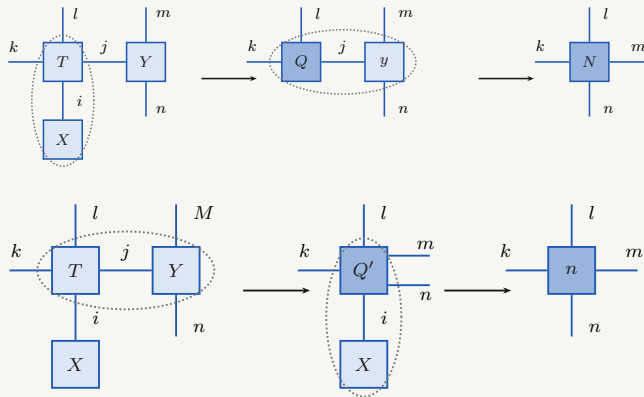
$$C_{i_1 \dots r j_1 \dots j_s} = \sum_{\alpha} A_{i_1 \dots r \alpha} B_{\alpha j_1 \dots j_s}$$

Množenje matrik in množenje matrike z vektorjem sta posebna primera skrčitve.



V notaciji označimo skrčitev kot združitev pripadajočih nog. Skrčena indeksa izgineta, red rezultata je za 2 manjši od vsote redov vhodnih tenzorjev.

Skrčitev tenzorske mreže je posplošitev skrčitve dveh tenzorjev.

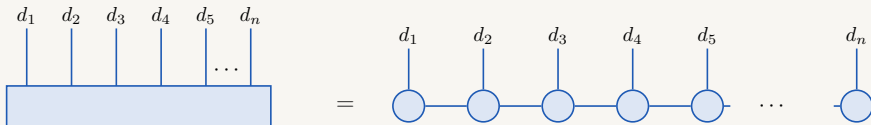


Vrstni red skrčitev ne vpliva na vrednost tenzorske mreže, vpliva pa na kompleksnost izračuna (vmesni tenzor Q' ima višji red - zasede več pomnilnika & več računske moči).

MPS - MATRIČNO PRODUKTNO STANJE

Horizontalne povezave predstavljajo dimenzije vezi D , vertikalne pa fizično dimenzijo d . Mejna tenzorja A_1 in A_N sta ranga 2, ostali pa so ranga 3.

$$\Psi_{d_1, d_2, \dots, d_N} = \sum_{\{D_i\}} A_{D_1, d_1}^{(1)} A_{D_1 d_2 D_2}^{(2)} \cdots A_{D_{j-1} d_j D_j}^{(j)}$$



Dimenzija vezi D je pomembna za določanje ekspresivnosti MPS.

SECTION 2

Generativno modeliranje z MPS

I Generativno modeliranje

- ◉ Cilj: modeliranje verjetnostne porazdelitve podatkov
- ◉ Z uporabo takšnega modela lahko generiramo nove vzorce podatkov
- ◉ Uporabno pri detekciji lastnosti in priporočilnih algoritmi

I MPS za generativno modeliranje

MPS je eksakten verjetnostni model, ki omogoča učinkovito vzorčenje in izračun verjetnosti poljubnega podatkovnega vzorca.

Z učinkovitim računanjem robnih in pogojnih porazdelitev omogoča tudi obravnavo manjkajočih vrednosti ter napovedovanje iz poljubne množice znanih atributov.

PODATKI

Podatkovna množica o raku dojk iz UKC Ljubljana. 286 vzorcev, kategorični nominalni in ordinalni atributi.

Atribut	Možne vrednosti
Ponovitev	Ni ponovitve; So ponovitve
Starost	10–19, 20–29, ..., 90–99
Menopavza	lt40, ge40, premeno
Velikost tumorja	0–4, 5–9, 10–14, ..., 55–59
Prizadete bezgavke	0–2, 3–5, 6–8, 9–11, ..., 36–39
Prebite bezgavke	Da; Ne
Stopnja malignosti	1, 2, 3
Dojka	Leva; Desna
Kvadrant dojke	levo zg., levo sp., desno zg., desno sp., sredina
Obsevanje	Da; Ne

Različni tipi atributov $\implies$ različne vložitve.

$X = (x_1, \dots, x_N)$ vhodni vektor. Iščemo vložitveno fun. $\Phi(X) = \phi(x_1) \otimes \dots \otimes \phi(x_N)$

eniško kodiranje

$$\phi(x_i) = e_{x_i}$$

Fourierjeva vložitev

$$t_c = \frac{c - 1}{K_i - 1} \in [0, 1]$$

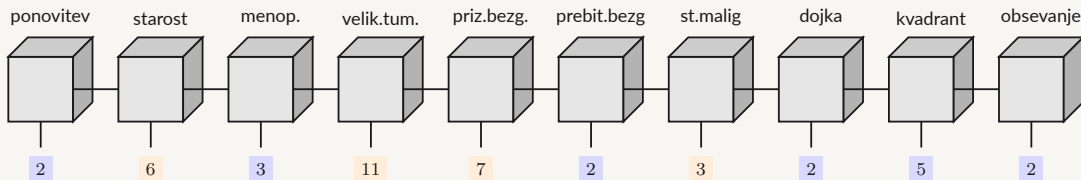
$$\phi_1(c) = 1$$

$$\phi_j(c) = \sqrt{2} \cos(\pi(j - 1)t_c), \quad j = 2, \dots, d_i$$

MPS omogoča, da vsakemu atributu priredimo svojo lokalno vložitev. Implementacija omogoča poljubno fizično dimenzijo d_k in s tem različno število kategorij za vsak stolpec.

UČENJE IN ZASNOVA MODELA

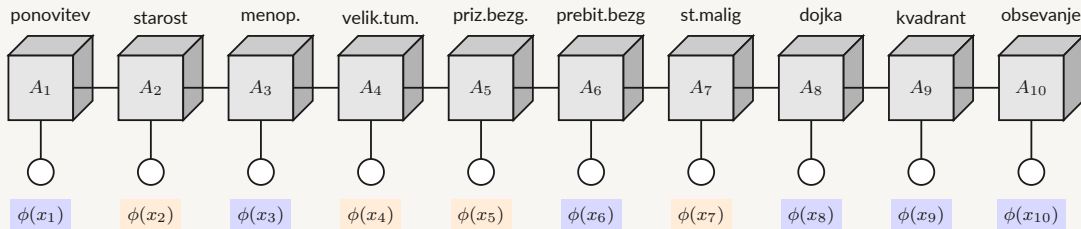
MPS je tenzor $\Psi_{d_1, d_2, \dots, d_N} = \sum_{D_i} A_{d_1, D_1}^{(1)} A_{D_1, d_2, D_2}^{(2)} \cdots A_{D_{N-1}, d_N}^{(N)}$, kjer je d_i fizična dimenzija in D_i dimenzija vezi.



Skalarni produkt med MPS in vložitvijo podatkov:

$$\Psi(x_1, x_2, \dots, x_N) = A_1[x_1]A_2[x_2] \cdots A_N[x_N]$$

Vsak $A_j[x_j]$ je matrika dimenzij $D \times D$, skrajni rezini $A_1[x_1]$ in $A_N[x_N]$ pa so vektorji dimenzij $1 \times D$ in $D \times 1$.



Skalar $\Psi(X)$ dobimo s skrčitvijo MPS z vložitvijo. A_j so parametri modela, ki jih učimo.

Ker so vrednosti $\Psi(X)$ negativne, to ni veljavna verjetnostna porazdelitev $\implies$ kvadiramo. Kvadrat norme MPS in njegove robne verjetnost lahko **učinkovito** in **točno** izračunamo.

$$P(x) = \frac{\Psi(x)^2}{Z}$$

$$Z = \sum P(x) = \sum \left| \begin{array}{c} \text{[cube]} \text{---} \text{[cube]} \text{---} \dots \text{---} \text{[cube]} \\ | \\ \text{[circle]} \quad \text{[circle]} \quad \dots \quad \text{[circle]} \end{array} \right|^2 = \sum \begin{array}{c} \text{[cube]} \text{---} \text{[cube]} \text{---} \dots \text{---} \text{[cube]} \\ | \quad | \quad \dots \quad | \\ \text{[circle]} \quad \text{[circle]} \quad \dots \quad \text{[circle]} \\ | \quad | \quad \dots \quad | \\ \text{[cube]} \text{---} \text{[cube]} \text{---} \dots \text{---} \text{[cube]} \end{array}$$

Normalizacijska konstanta Z pri drugih generativnih modelih navadno ni učinkovito izračunljiva, pri MPS jo lahko izračunamo v $O(NdD^3)$.

I Vzorčenje

- ◉ Z naučenim Bornovim strojem generiramo sintetične vzorce s predniškim vzorčenjem
- ◉ $P(x_1, \dots, x_N) = P(x_1)P(x_2 \mid x_1)P(x_3 \mid x_1, x_2) \dots$
- ◉ Za predniško vzorčenje potrebujemo normalizirane pogojne verjetnosti - v splošnih generativnih modelih računsko neizvedljivo
- ◉ v MPS je to mogoče zaradi učinkovitega izračuna Z

I Ocenjevanje uspešnosti modela

Naučeni model ocenimo na testni množici (podatkih, ki jih model še ni videl).

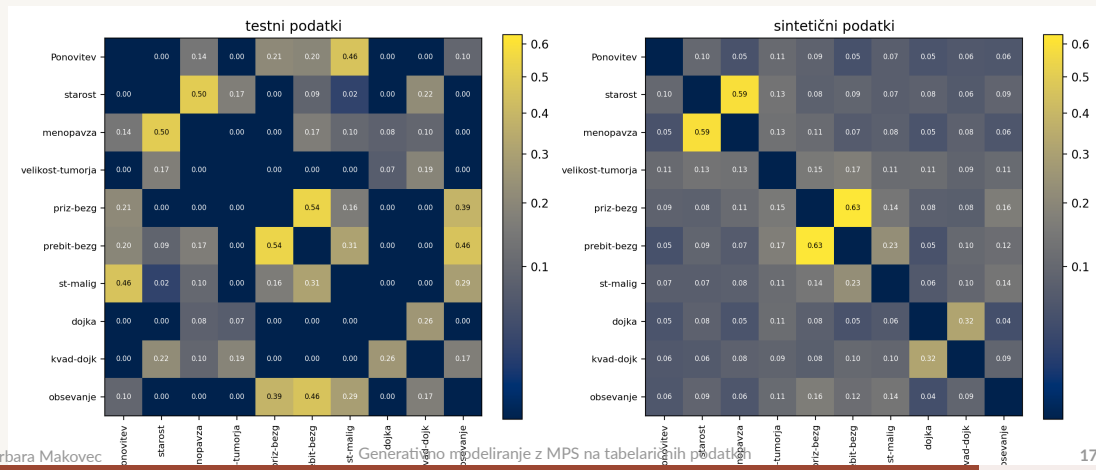
- ◉ Ali model določi visoko verjetnost vzorcem iz testne množice?
- ◉ Ali so sintetični podatki, ki jih generira, podobni resničnim?

I Parametri

- ◉ Skupno število vzorcev: 277
- ◉ Učna in testno množica: 80:20

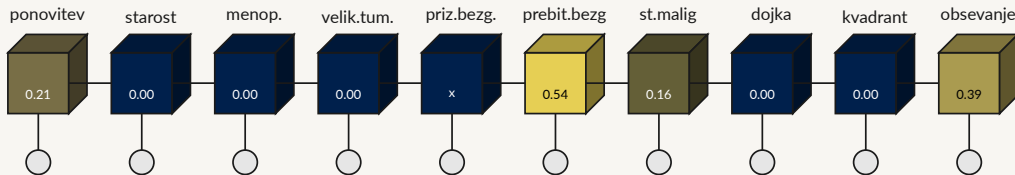
EVALVACIJA - KORELACIJA MED ATRIBUTI (IZVORNI VRSTNI RED)

Merili smo korelacijo med atributi. Ali se model uspe naučiti ne samo robnih porazdelitev, ampak tudi odvisnosti med atributi. Primerjamo matriko povezanosti atributov v sintetičnih in testnih podatkih.

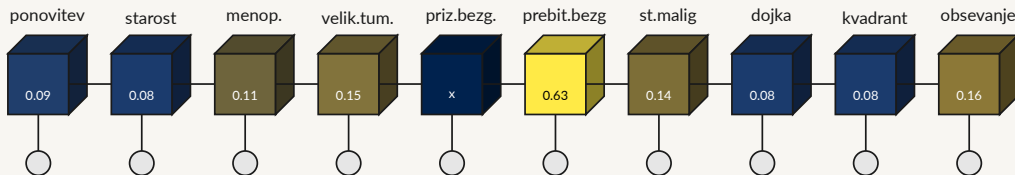


STOLPEC "PRIZADETE BEZGAVKE" - PRIMERJAVA PORAZDELITEV

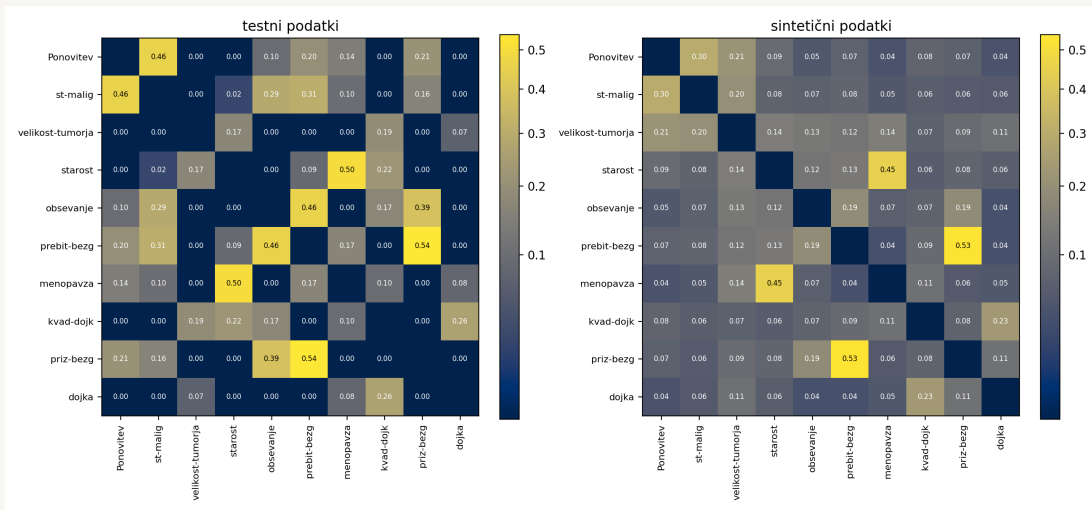
testni:



synthetčni:



EVALVACIJA - KORELACIJA MED ATRIBUTI (PERMUTIRANI VRSTNI RED)



GENERATIVNO MODELIRANJE Z MPS NA TABELARIČNIH PODATKIH

- Tenzorske mreže predstavljajo razcep visokodimenzionalnih tenzorjev v manjše tenzorje.
- Posebna oblika tenzorske mreže: MPS
- V enem MPS lahko uporabimo različne tipe vložitev.
- Z MPS lahko modeliramo tabelarične podatke.

MPS je eksakten verjetnostni model. V nasprotju z ostalimi generativnimi modeli lahko zaradi strukture omogoča učinkovit izračun robnih in pogojnih porazdelitev. Uporabimo ga lahko za generiranje podatkov.

