Kompleksnost skrivenih Markovljevih modela

Tepić, Martina

SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO–MATEMATIČKI FAKULTET
MATEMATIČKI ODSJEK

Martina Tepić

KOMPLEKSNOST SKRIVENIH MARKOVLJEVIH MODELA

Diplomski rad

Voditelj rada:
doc. dr. sc. Pavle Goldstein

Zagreb, veljača 2015.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Hvala mojim roditeljima jer su mi omogućili studiranje...
Hvala bratu i sestri jer su vjerovali u mene i bodrili me...
I jedno veliko hvala mentoru doc. dr. sc. Pavlu Goldsteinu na strpljenju, posvećenom vremenu i najvažnijoj lekciji koju me naučio: Neuspjeh u rješenju bilo kojeg zadatka je samo rezultat nedovoljnog broja pokušaja...

Sadržaj

Uvod

Bioinformatika je znanost koja se bavi analizom bioloških nizova, sadržaja i organizacije genoma, te predviđanjem strukture i funkcije makromolekula uz pomoć tehnika iz primijenjene matematike, statistike i računarstva.

U ovom diplomskom radu bavimo se skrivenim Markovljevim modelima - statističkim alatom za modeliranje nizova koje generira neki skriveni proces. Broj područja u kojima te metode i modeli nailaze na primjenu osiguravaju im i danas etiketu aktualne teme u stohastičkom modeliranju. Neke od najpoznatijih primjena su prepoznavanje govora (*speech recognition*), rukopisa (*handwriting recognition*) i gesta (*gesture recognition*), računalno prevođenje (*machine translation*), analiza vremenskih nizova te bioinformatika.

U ovom radu smo detaljno obradili jednu od metoda za procjenu parametara i dali primjer njene implementacije.

U prvom poglavlju su dani osnovni pojmovi iz teorije vjerojatnosti, Markovljevih lanaca i statistike. Formalna definicija skrivenih Markovljevih modela i primjer su dani u drugom poglavlju, dok se u trećem poglavlju nalaze algoritmi koje smo koristili za procjenu parametara modela i optimizaciju vjerodostojnosti. U četvrtom poglavlju su prezentirani rezultati i zaključci rada.

Poglavlje 1

Osnovni pojmovi

1.1 Vjerojatnost

Definicija 1.1.1. *Pod **slučajnim pokusom** podrazumijevamo takav pokus čiji **ishodi**, odnosno **rezultati** nisu jednoznačno određeni uvjetima u kojima izvodimo pokus. Rezultate slučajnog pokusa nazivamo **događajima**.*

Definicija 1.1.2. *Neka je A događaj vezan uz neki slučajni pokus. Pretpostavimo da smo taj pokus ponovili n puta i da se u tih n ponavljanja događaj A pojavio točno n_A puta. Tada broj n_A zovemo **frekvencija** događaja A, a broj $\frac{n_A}{n}$ **relativna frekvencija** događaja A.*

Definicija 1.1.3. *Osnovni objekt u teoriji vjerojatnosti jest neprazan skup Ω koji zovemo **prostor elementarnih događaja** i koji reprezentira skup svih ishoda slučajnih pokusa. Ako je Ω konačan ili prebrojiv, govorimo o **diskretnom** prostoru elementarnih događaja. Prostor elementarnih događaja je **kontinuiran** ako je Ω neprebrojiv skup.*
*Točke ω iz skupa Ω zvat ćemo **elementarni događaji***

Označimo sa $\mathcal{P}(\Omega)$ partitivni skup od Ω.

Definicija 1.1.4. *Familija $\mathcal{F}$ podskupova od Ω $(\mathcal{F} \subset \mathcal{P}(\Omega))$ jest $\sigma-$**algebra skupova** (na Ω) ako je:*

F1. $\emptyset \in \mathcal{F}$

F2. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$

F3. $A_i \in \mathcal{F},\ i \in \mathbb{N} \Rightarrow \cup_{i=1}^{\infty} A_i \in \mathcal{F}$

Definicija 1.1.5. *Neka je $\mathcal{F}$ σ-algebra na skupu Ω. Uređen par $(\Omega, \mathcal{F})$ se zove* ***izmjeriv prostor***

Sad možemo definirati vjerojatnost.

Definicija 1.1.6. *Neka je $(\Omega, \mathcal{F})$ izmjeriv prostor. Funkcija $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ jest* ***vjerojatnost*** *ako vrijedi:*

P1. $\mathbb{P}(\Omega) = 1$ *(normiranost vjerojatnosti)*

P2. $\mathbb{P}(A) \geq 0,\ A \in \mathcal{F}$ *(nenegativnost vjerojatnosti)*

P3. $A_i \in \mathcal{F},\ i \in \mathbb{N}\ \ i\ \ A_i \cap A_j = \emptyset\ \ za\ \ i \neq j \Rightarrow \mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ *(prebrojiva ili σ - aditivnost vjerojatnosti)*

Definicija 1.1.7. *Uređena trojka $(\Omega, \mathcal{F}, \mathbb{P})$ gdje je $\mathcal{F}$ $\sigma-$algebra na Ω i $\mathbb{P}$ vjerojatnost na $\mathcal{F}$, zove se* ***vjerojatnosni prostor****.*

Definicija 1.1.8. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Elemente $\sigma-$algebre zovemo* ***događaji****, a broj $\mathbb{P}(A),\ A \in \mathcal{F}$ se zove* ***vjerojatnost događaja*** *A.*

Budući da radimo sa slučajnim varijablama, potrebno je definirati otvoreni skup.

Definicija 1.1.9. *Neka je $x \in \mathbb{R}^n$ i $r > 0$. Skup*

$$K(x, r) = \{y \in \mathbb{R}^n : d(x, y) < r\} = \{y \in \mathbb{R}^n : \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} < r\}$$

nazivamo ***otvorena kugla oko x radijusa r****. Skup $A \subset \mathbb{R}^n$ je* ***otvoren*** *ako vrijedi*

$$\forall x \in A, \exists r > 0, K(x, r) \subset A.$$

Otvorena okolina *točke $x \in \mathbb{R}^n$ je svaki otvoreni skup koji sadrži točku x.*

Definicija 1.1.10. *Označimo sa $\mathcal{B}$ $\sigma-$algebru generiranu familijom svih otvorenih skupova na skupu realnih brojeva $\mathbb{R}$. $\mathcal{B}$ zovemo $\sigma-$* ***algebra skupova na*** *$\mathbb{R}$, a elemente $\sigma-$algebre $\mathcal{B}$ zovemo* ***Borelovi skupovi****.*

Budući da je svaki otvoreni skup na $\mathbb{R}$ prebrojiva unija intervala, lako je dokazati da vrijedi

$$\mathcal{B} = \sigma\{(a, b);\ a, b \in \mathbb{R},\ a < b\}$$

Definicija 1.1.11. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Funkcija $X : \Omega \to \mathbb{R}$ jest* ***slučajna varijabla*** *(na Ω) ako je $X^{-1}(B) \in \mathcal{F}$ za proizvoljno $B \in \mathcal{B}$, odnosno $X^{-1}(\mathcal{B}) \subset \mathcal{F}$.*

Definicija 1.1.12. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ proizvoljan vjerojatnosni prostor i $A \in \mathcal{F}$ takav da je $\mathbb{P}(A) > 0$. Definiramo funkciju $P_A : \mathcal{F} \to [0,1]$ ovako:*

$$P_A(B) = P(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}, \ B \in \mathcal{F}. \tag{1.1}$$

Lako je provjeriti da je P_A vjerojatnost na $\mathcal{F}$ i nju zovemo ***vjerojatnost od B uz uvjet A****.*

Definicija 1.1.13. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor i $A_i \in \mathcal{F}$, $i \in I$ proizvoljna familija događaja. Kažemo da je to* ***familija nezavisnih događaja*** *ako za svaki konačan podskup različitih indeksa $i_1, i_2, ..., i_k$ vrijedi*

$$\mathbb{P}(\cap_{i=1}^{k} A_{i_j}) = \prod_{j=1}^{k} \mathbb{P}(A_{i_j}). \tag{1.2}$$

Neka je X slučajna varijabla na diskretnom vjerojatnosnom prostoru $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ i neka je

$$X = \begin{pmatrix} a_1 & a_2 & \dots \\ p_1 & p_2 & \dots \end{pmatrix}$$

njena distribucija, odnosno vrijedi $\mathbb{P}(a_i) = p_i$.

Definicija 1.1.14. ***Funkcija gustoće vjerojatnosti*** *od X ili, kraće,* ***gustoća*** *od X jest funkcija $f_X = f : \mathbb{R} \to \mathbb{R}_+$ definirana sa*

$$f(x) = \mathbb{P}\{X = x\} = \begin{cases} 0, & x \neq a_i \\ p_i, & x = a_i \end{cases}, x \in \mathbb{R}$$

Definicija 1.1.15. ***Funkcija distribucije slučajne varijable*** *X jest funkcija $F_X = F : \mathbb{R} \to [0,1]$ definirana sa*

$$F(x) = \mathbb{P}\{X \leq x\} = \mathbb{P}\{\omega; X(\omega) \leq x\}, \ x \in \mathbb{R}.$$

1.2 Markovljevi lanci

Definicija 1.2.1. *Neka je S skup.* ***Slučajan proces*** *s diskretnim vremenom i prostorom stanja S je familija $X = (X_n : n \geq 0)$ slučajnih varijabli definiranih na nekom vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ s vrijednostima u S.*
Dakle, za svaki $n \geq 0$ je $X_n : \Omega \to S$ slučajna varijabla.

Definicija 1.2.2. *Neka je S prebrojiv skup. Slučajni proces $X = (X_n : n \geq 0)$ definiran na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ s vrijednostima u skupu S je* ***Markovljev lanac prvog reda*** *ako vrijedi*

$$\mathbb{P}(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, ..., X_0 = i_0) = \mathbb{P}(X_{n+1} = j | X_n = i) \tag{1.3}$$

za svaki $n \geq 0$ i za sve $i_0, ..., i_{n-1}, i, j \in S$ za koje su obje uvjetne vjerojatnosti dobro definirane.

Svojstvo u relaciji (1.3) naziva se *Markovljevim svojstvom.*

Definicija 1.2.3. *Označimo sa $p_{ij} = \mathbb{P}(X_{t+1} = j \mid X_t = i)$ vjerojatnost da slučajna varijabla X prijeđe u stanje j u trenutku $t + 1$, ako je u trenutku t bila u stanju i. Vrijednost p_{ij} nazivamo* ***prijelazna (tranzicijska) vjerojatnost****.*

Markovljev lanac zajedno sa zadanim prijelaznim vjerojatnostima nazivamo **Markovljevim modelom**.

1.3 Statistika

Definicija 1.3.1. *Za model $\mathcal{T} = \{f(\cdot; \theta) : \theta \in \Theta\}, f(\cdot; \theta) : \mathbb{R} \to [0, +\infty >, \Theta \subset \mathbb{R}$ kažemo da je regularan ako su zadovoljeni sljedeći uvjeti:*

i) $\sup f(\cdot; \theta) = \{x \in \mathbb{R} : f(x; \theta) > 0\}$ *ne ovisi o $\theta \in \Theta$*

ii) Θ *je otvoreni interval u $\mathbb{R}$*

iii) $\forall x \in \mathbb{R}$, $\theta \to f(x; \theta)$ *je diferencijabilna na Θ*

iv) Za slučajnu varijablu X kojoj je f funkcija gustoće vrijedi:

$$0 < I(\theta) := \mathbb{E}_\theta[(\frac{\partial}{\partial \theta} \log f(x; \theta))^2] < \infty$$

Broj $I(\theta)$ se zove ***Fisherova informacija****.*

v) $\forall \theta \in \Theta$, $\frac{d}{d\theta} \int_{\mathbb{R}} f(x; \theta) dx = \int_{\mathbb{R}} \frac{\partial}{\partial \theta} f(x; \theta) dx = 0$, *ako se radi o neprekidnoj slučajnoj varijabli, odnosno*
$\forall \theta \in \Theta$, $\frac{d}{d\theta} \sum_x f(x; \theta) = \sum_x \frac{\partial}{\partial \theta} f(x; \theta) = 0$, *ako je riječ o diskretnoj slučajnoj varijabli*[1].

[1]Prisjetimo se: slučajna varijabla je diskretna ako je definirana na diskretnom vjerojatnosnom prostoru, a neprekidna ukoliko joj je funkcija gustoće nenegativna realna funkcija

Definicija 1.3.2. *Neka je $(\Omega, \mathcal{F})$ izmjeriv prostor i $\mathcal{P}$ familija vjerojatnosnih mjera na $(\Omega, \mathcal{F})$. Uređena trojka $(\Omega, \mathcal{F}, \mathcal{P})$ se zove* ***statistička struktura****.*

Definicija 1.3.3. *n-dimenzionalni* ***slučajni uzorak*** *na statističkoj strukturi $(\Omega, \mathcal{F}, \mathcal{P})$ je niz $(X_1, ..., X_n)$ slučajnih varijabli na izmjerivom prostoru $(\Omega, \mathcal{F})$ takav da su slučajne varijable $X_1, \ldots, X_n$ nezavisne i jednako distribuirane $\forall \mathbb{P} \in \mathcal{P}$.*

Definicija 1.3.4. *Neka je $X = (X_1, ..., X_n)$ slučajan uzorak iz modela $\mathcal{P}$, $\mathcal{P} = \{f(\cdot; \theta) : \theta \in \Theta\}, \Theta \subset \mathbb{R}^m$. Ako je $X = (X_1, ..., X_n)$ jedna realizacija od $\mathbb{X}$, tada je* ***vjerodostojnost*** *funkcija $L : \Theta \to \mathbb{R}$*

$$L(\theta) = L(\theta|\mathbb{X}) := \prod_{i=1}^{n} f(X_i; \theta)$$

*Statistika $\hat{\theta} = \hat{\theta}(\mathbb{X})$ je procjenjitelj maksimalne vjerodostojnosti (****MLE****) ako vrijedi*

$$L(\hat{\theta}|\mathbb{X}) = \max_{\theta \in \Theta} L(\theta|\mathbb{X})$$

Definicija 1.3.5. *Za opaženu vrijednost x od $\mathbb{X}_n$, $l : \Theta \to \mathbb{R}$,*

$$l(\theta) = l(\theta|\mathbb{X}) = \log L(\theta|\mathbb{X}) = \sum_{i=1}^{n} \log f(x_i : \theta)$$

je ***log-vjerodostojnost****.*

Definicija 1.3.6. *Procjenitelj $T = t(X)$ za $\tau(\theta) \in \mathbb{R}$ je* ***nepristran*** *ako vrijedi*

$$\forall \theta \in \Theta,\ \mathbb{E}_\theta(T) = \tau(\theta)$$

. Procjenitelj koji nije nepristran je ***pristran****.*

Definicija 1.3.7. *T je* ***efikasan*** *procjenitelj za $\tau(\theta)$ ako je nepristran i vrijedi*

$$Var_\theta = \frac{[\tau'(\theta)]^2}{nI(\theta)}, \forall \theta \in \Theta$$

Definicija 1.3.8. *Niz procjenitelja $(T_n : n \in \mathbb{N})$ je* ***konzistentan*** *procjenitelj za θ ako za proizvoljni $\epsilon > 0$ vrijedi*

$$\lim_{n \to \infty} \mathbb{P}_\theta\{|T_n - \theta| \geq \epsilon\} = 0$$

Teorem 1.3.9. *Neka je $\mathbb{X}_n = (X_1, ..., X_n)$ slučajan uzorak iz regularnog modela $\mathcal{P}$, uz dodatnu pretpostavku da je $\theta \to f(x;\theta)$ neprekidno diferencijabilna. Tada jednadžba vjerodostojnosti*

$$\frac{\partial}{\partial\theta} l(\theta|\mathbb{X}_n) = 0$$

na događaju čija vjerojatnost teži ka 1 za $n \to \infty$ ima korjen $\hat{\theta}_n = \hat{\theta}_n(X_n)$ takav da je $\hat{\theta}_n \xrightarrow{P_\theta} \theta$, za $n \to \infty$.

Napomena 1.3.10. *Ako jednadžba vjerodostojnosti ima jedinstvenu stacionarnu točku $\hat{\theta}_n \xrightarrow{\mathbb{P}_{\theta_0}} \theta_0$, tada Teorem 1.3.9 tvrdi da ona mora biti konzistentan procjenitelj za θ_0. Ako je MLE jedinstvena stacionarna točka kao točka lokalnog maksimuma, onda je MLE konzistentan procjenitelj za θ.*

Lema 1.3.11. *Neka je $X \tilde{} B(n,\theta)$ gdje je θ vjerojatnost uspjeha. Tada je procjenitelj maksimalne vjerodostojnosti za θ relativna frekvencija uspjeha.*

Dokaz. Označimo sa n broj pokušaja, a sa k broj uspjeha. Tada je vjerojatnost da smo imali točno k uspjeha dana s

$$f(\theta) = P(X = k) = \binom{n}{k}\theta^k(1-\theta)^{n-k}, k = 0, 1, 2, ..., n$$

Nađimo stacionarne točke koje su kandidati za lokalni maksimum:

$$\begin{aligned} f'(\theta) &= \binom{n}{k}[kp^{k-1}(1-p)^{n-k} - p^k(n-k)(1-p)^{n-k-1}] \\ &= \binom{n}{k}[p^{k-1}(1-p)^{n-k-1}(k(1-p) - (n-k)p)] \\ &= 0 \end{aligned}$$

$\Rightarrow k - kp - np + kp = 0$
$\Rightarrow np = k$
$\Rightarrow p = \frac{k}{n}$ □

1.4 Shannonova entropija

Definicija 1.4.1. *Entropija je mjera prosječne neizvjesnosti ishoda. Za danu slučajnu varijablu X sa vjerojatnostima $\mathbb{P}(x_i)$ za diskretan skup događaja $x_1, ..., x_K$ Shannonova entropija je definirana s*

$$H(X) = -\sum_{i=1}^{K} \mathbb{P}(x_i)\log(\mathbb{P}(x_i)) \tag{1.4}$$

Da bismo intuitivno shvatili o čemu je riječ razmotrimo primjer bacanja novčića: U ovom slučaju, imamo dva moguća simbola (K = 2), i oba se pojavljuju s vjerojatnošću $p(x_i) = \frac{1}{2}$.

Jednostavnim uvrštavanjem u formulu entropije dobivamo $H(X) = 1$ bit/simbol. Dakle, vrijednost entropije u ovisnosti o vjerojatnosti pojave pisma/glave kod bacanja novčića je 1 bit/simbol, odnosno srednji sadržaj informacije poruke koja se sastoji od uzastopnih rezultata bacanja novčića je 1 bit po simbolu.
Za slučaj "nepoštenog" novčića koji uvijek daje pismo, imamo $p(x_1) = 1,\ p(x_2) = 0$, dobivamo očekivano H(X) = 0 bit/simbol ($0 \log 0 = 0$, jer vrijedi $x \log x \to 0$ kada $x \to 0$). Uvrštavanjem svih mogućih vjerojatnosti pojave pisma u formulu entropije, dobivamo graf ovisnosti vrijednosti entropije o toj vjerojatnosti (1.1). Maksimum (1 bit/simbol) je postignut kada je vjerojatnost pisma jednaka vjerojatnosti glave ($p = \frac{1}{2}$) – dakle kada je najveća nesigurnost pojave jednog ili drugog. Primijetimo simetriju ovog grafa. Svejedno je pojavljuje li se s većom vjerojatnošću pismo ili glava. Zamjenom njihovih uloga situacija se s informacijskog gledišta ne mijenja.

Pretpostavimo da su zadane dvije funkcije više varijabli $f, \varphi : \mathcal{D} \to \mathbb{R}$ definirane na skupu $\mathcal{D} \subseteq \mathbb{R}^k$. Funkciji φ pridružimo implicitnu jednadžbu $\varphi(y_1, ..., y_k) = 0$ i pripadajući skup $S \subseteq \mathcal{D}$ definiran tom jednadžbom $S = \{(y_1, ..., y_k) \in \mathcal{D} \mid \varphi(y_1, ..., y_k) = 0\}$.

Definicija 1.4.2. *Ako za točku $T_0 = (x_{10}, ..., x_{k0}) \in S$ postoji okolina $K(T_0, \delta) \subseteq \mathcal{D}$ tako da je*

$$f(x_1, ..., x_k) < f(x_{10}, ..., x_{k0}),\ \forall (x_1, ..., x_k) \in S \cap K(T_0, \delta) \setminus \{T_0\}$$

onda kažemo da funkcija f u točki T_0 ima uvjetni lokalni maksimum uz uvjet $\varphi(x_1, ..., x_k) = 0$.

Problem uvjetnog lokalnog maksimuma

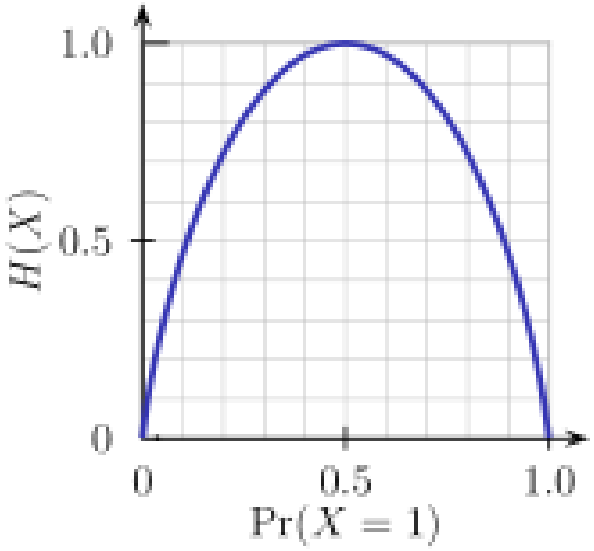
$$\begin{cases} z = f(x_1, ..., x_k) \to \max \\ \varphi(x_1, ..., x_k) = 0 \end{cases}$$

često rješavamo uvođenjem Lagrangeove funkcije $L(x_1, ..., x_k, \lambda)$:

$$L(x_1, .., x_k, \lambda) = f(x_1, ..., x_k) + \lambda \varphi(x_1, ..., x_k),\ (x_1, ..., x_k) \in \mathcal{D},\ \lambda \in \mathbb{R}.$$

Parametar λ zove se **Lagrangeov multiplikator**.

Lema 1.4.3. *Uniformno distribuirani parametri imaju maksimalnu entropiju.*



Slika 1.1: Vrijednost entropije u ovisnosti o vjerojatnosti pojave pisma kod bacanja novčića

Prije samog dokaza prisjetimo se Bolzano-Weierstrassova i Rolleova teorema:

Teorem 1.4.4. *(Bolzano-Weierstrass): Neka je funkcija* $f : [a, b] \to \mathbb{R}$ *neprekidna na segmentu* $[a, b] \subset \mathbb{R}$. *Tada je* $f([a, b]) = [m, M]$ *takoder segment.*

Napomena 1.4.5. *Tvrdnja teorema može se razdvojiti na tri dijela:*

1. f *je ograničena na* $[a, b]$, *odnosno postoje* $m = \inf_{[a,b]} f$ *i* $M = \sup_{[a,b]} f$.
2. *funkcija* f *postiže svoj minimum i maksimum na* $[a, b]$, *odnosno postoje* $x_m, x_M \in [a, b]$ *takvi da vrijedi* $f(x_m) = m$ *i* $f(x_M) = M$.
3. *za svaki* $C \in (m, M)$, *postoji* $c \in [a, b]$ *takav da je* $f(c) = C$.

Teorem 1.4.6. *(Rolle): Neka je* $f : I \to \mathbb{R}$, *diferencijabilna na otvorenom intervalu* $I \subset \mathbb{R}$ *i neka za* $a, b \in I$, $a < b$, *vrijedi* $f(a) = f(b) = 0$. *Tada postoji* $c \in (a, b)$ *takav da je* $f'(c) = 0$

Dokaz. (Lema (1.4.3)): Definiramo funkcije $f : [0, 1]^k \to \mathbb{R}$ i $\varphi : [0, 1]^k \to \mathbb{R}$ s

$$f(p_1, ..., p_k) = -\sum_{i=1}^{k} p_i \log p_i$$

$$\varphi(p_1, ..., p_k) = \sum_{i=1}^{k} p_i - 1.$$

Neka je λ Lagrangeov multiplikator. Definiramo funkciju $g : \mathbb{R}^k \to \mathbb{R}$ sa

$$g(p_1, ..., p_k) = f(p_1, ..., p_k) + \lambda\varphi(p_1, ..., p_k)$$

Funkcija g je klase C^∞ na zatvorenom skupu $[0, 1]^k$, znači da je ujedno i neprekidna pa prema *Bolzano-Weierstrassovom teoremu* poprima minimum m i maksimum M na tom skupu. Budući da funkcija g nije konstantna funkcija na $[0, 1]^k$ barem jedna od te dvije vrijednosti se nalazi unutar otvorenog skupa $(0, 1)^k$.
Funkcija g je strogo pozitivna na $(0, 1)^k$, u rubovima je jednaka 0, stoga će prema *Rollovom teoremu* stacionarna točka biti maksimum.

Tražimo stacionarne točke te funkcije.

$$\frac{dg}{dp_i} = -\log p_i - 1 + \lambda = 0$$

$$\log p_i = \lambda - 1$$

$$p_i = \exp(\lambda - 1)$$

$$\sum_{i=1}^{k} p_i = 1 \;\Rightarrow\; k\exp(\lambda - 1) = 1$$

Slijedi da funkcija g postiže maksimum u točki $p_M = (p_1, ..., p_k)$

$$p_i = \frac{1}{k},\ i = 1, ..., k$$

. □

Poglavlje 2

Skriveni Markovljev model

2.1 HMM

Kod skrivenog Markovljevog modela, imamo niz stanja i niz simbola. Svaki simbol ovisi jedino o trenutnom stanju u kojem se proces nalazi. Zato generiranje simbola iz stanja modeliramo **Markovljevim lancem nultog reda** što je upravo *niz nezavisnih događaja.*
Niz stanja skrivenog Markovljevog modela modeliran je Markovljevim lancem prvog reda, tj. vjerojatnost da se nalazimo u nekom stanju ovisi samo o prethodnom stanju. Formalno rečeno:

Definicija 2.1.1. ***Skriveni Markovljev model prvog reda*** *(eng. Hidden Markov model, HMM) je skup slučajnih varijabli koji se sastoji od dva podskupa, Q i O:*

- $Q = Q_1, ..., Q_N$ *- skup slučajnih varijabli koje poprimaju diskretne vrijednosti*
- $O = O_1, ..., O_N$ *- skup slučajnih varijabli koje poprimaju diskretne ili kontinuirane vrijednosti.*

Te varijable zadovoljavaju sljedeće uvjete:

1.

$$P(Q_t|Q_{t-1}, O_{t-1}, ..., Q_1, O_1) = \mathbb{P}(Q_t|Q_{t-1}) \tag{2.1}$$

2.

$$\mathbb{P}(O_t|Q_T, O_T, ..., Q_{t+1}, O_{t+1}, Q_t, Q_{t-1}, O_{t-1}, ..., Q_1, O_1) = P(O_t|Q_t) \tag{2.2}$$

Uočili smo da, osim niza stanja kroz koja proces prolazi (označili smo ih sa Q_i), promatramo i niz opažanja (simbola, označili smo ih sa O_i).
Da pojasnimo, relacija (2.1) predstavlja vjerojatnost da smo, za neko $t \in \{1, 2, ..., N\}$, u stanju Q_t uz uvjet da su se dogodila sva prethodna stanja $Q_1, ..., Q_{t-1}$ i emitirali simboli $O_1, ..., O_{t-1}$ jednaka **tranzicijskoj vjerojatnosti** iz stanja Q_{t-1} u stanje Q_t. Relacija (2.2) povlači da realizacija nekog opažanja u sadašnjem stanju ovisi samo o tom stanju. Vjerojatnosti iz relacije (2.2) nazivamo **emisijska vjerojatnost** i kažemo da neko stanje Q_t **emitira** simbol O_t.

Skriveni Markovljev model zadan je sljedećim parametrima:

- N - broj stanja u kojima se proces može nalaziti

$$S = \{1, ..., N\} \tag{2.3}$$

  S - skup svih stanja procesa

- M - broj mogućih opažanja

$$B = \{b_1, ..., b_M\} \tag{2.4}$$

  B - skup svih opaženih vrijednosti

- L - duljina opaženog niza

$$X = (x_1, ..., x_L) \tag{2.5}$$

  X - opaženi niz

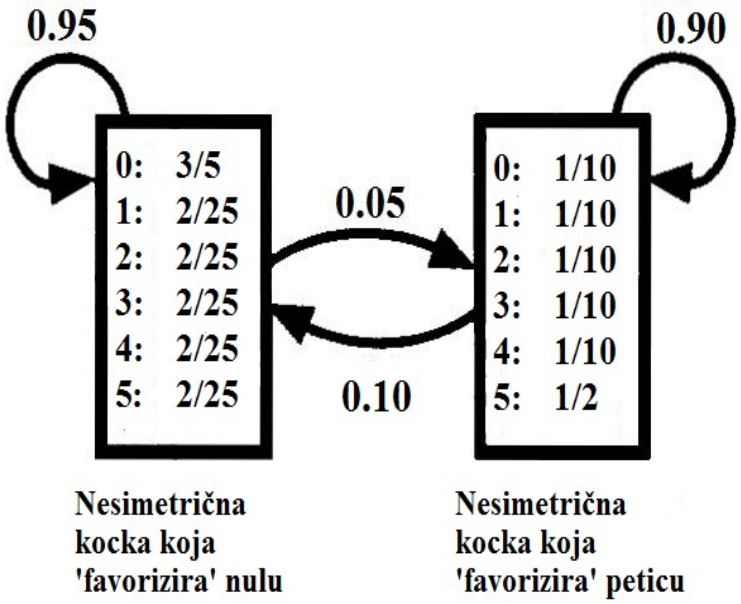
- A - matrica tranzicijskih vjerojatnosti

$$A = \{a_{ij}\}, a_{ij} = \mathbb{P}(Q_{t+1} = j | Q_t = i), 1 \leq i, j \leq N \tag{2.6}$$

- E - matrica emisijskih vjerojatnosti

$$E = \{e_j(k)\}, e_j(k) = \mathbb{P}(O_t = b_k | Q_t = j), 1 \leq j \leq N, 1 \leq k \leq M \tag{2.7}$$

Primijetimo da nam je kod Markovljevog modela nultog reda **niz stanja** koji emitira neki niz simbola poznat.
Skriveni Markovljev model je Markovljev model kod kojeg su stanja "skrivena", odnosno ne znamo ih pri emitiranju nekog niza vrijednosti ili simbola. Međutim, taj niz vrijednosti nam je poznat i pomoću njega možemo donijeti neke zaključke o nizu stanja koji odgovara emitiranom nizu vrijednosti.



Slika 2.1: Primjer skrivenog Markovljevog modela s dvije kocke

2.2 Primjer skrivenog Markovljevog modela

Imamo dvije nepoštene igraće kocke. Jedna kocka, koju označavamo K_0, ima vjerojatnost da dobijemo nulu $\frac{3}{5}$, a vjerojatnost preostalih ishoda je $\frac{2}{25}$, dok druga kocka, u oznaci K_5 ima vjerojatnost da padne petica $\frac{1}{2}$, a vjerojatnost preostalih ishoda je $\frac{1}{10}$.
Pretpostavimo da počinjemo sa K_0. Vjerojatnost da ćemo ponovo koristiti K_0 je 95%, dok je vjerojatnost da ćemo je zamijeniti sa K_5 5%. Kad smo jednom K_0 zamijenili sa K_5, u 90% slučajeva ćemo je i nastaviti koristiti. Vjerojatnost da je zamijenimo sa K_0 je 10%.

Koristimo li notaciju za HMM, naš model zapisujemo na sljedeći način:

- N=2
$$\mathcal{S} = \{K_0, K_5\}$$
- M=6
$$B = \{0, 1, 2, 3, 4, 5\}$$

- Matrica tranzicijskih vrijednosti je dana s:

$$A = \begin{pmatrix} 0.95 & 0.05 \\ 0.1 & 0.9 \end{pmatrix}$$

  gdje je $a_{11} = \mathbb{P}(K_0|K_0)$ - vjerojatnost da je nakon K_0 ponovo bačena K_0, $a_{12} = \mathbb{P}(K_5|K_0)$ - vjerojatnost bacanja K_5, ako je prethodno bačena K_0, $a_{21} = \mathbb{P}(K_0|K_5)$ - vjerojatnost da je nakon bacanja K_5 bačena K_0 i $a_{22} = \mathbb{P}(K_5|K_5)$ - vjerojatnost da je nakon K_5 opet bačena K_5.

- Matrica emisijskih vjerojatnosti je:

$$E = \begin{pmatrix} 0.6 & 0.08 & 0.08 & 0.08 & 0.08 & 0.08 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.5 \end{pmatrix}$$

  Prvi redak čine emisijske vjerojatnosti elemenata iz B u stanju K_0, a drugi redak emisijske vjerojatnosti elemenata iz B u stanju K_5

Proces koji modelira izbor kocki je Markovljev proces prvog reda sa stanjima u $\mathcal{S}$. Kocke su stanja i prijelaz iz jedne kocke u drugu se može opisati Markovljevim lancem. Emisijske vjerojatnosti simbola iz B su u svakom od stanja različite i ne ovise o prijašnjim stanjima.
Možemo reći da smo dali primjer *skrivenog Markovljevog modela prvog reda.*
Ako imamo niz simbola, odnosno opaženih vrijednosti, primjerice $X = (1, 2, 5, 0, 0, 4, 3)$ ne znamo koja kocka stoji iza pojedine opažene vrijednosti. Dakle, *niz stanja* je skriven.
Ipak, iako je niz stanja nepoznat, pomoću niza simbola moguće je:

- odrediti *najvjerojatniji niz stanja* zadani niz simbola. U tu svrhu koristimo **Viterbijev algoritam** o kojem ćemo govoriti u sljedećem poglavlju
- procijeniti *parametre uvjetne maksimalne vjerodostojnosti* koristeći **Viterbijevo treniranje** koje ćemo pojasniti kasnije
- procijeniti parametre maksimalne vjerodostojnosti modela koristeći **determinističko kaljenje** koje će kasnije biti detaljno objašnjeno.

Poglavlje 3

Algoritmi za HMM

U ovom poglavlju objašnjeni su algoritmi[1] koje smo spomenuli u prethodnom poglavlju. Ponovimo, u skrivenom Markovljevom modelu je niz stanja nepoznat, ali pomoću niza emitiranih vrijednosti možemo nešto zaključiti o nizu stanja.
Niz emitiranih simbola ćemo označiti s $x = (x_1, ..., x_n)$, a pripadajući niz skrivenih stanja s $\pi = (\pi_1, ..., \pi_n)$.
Za tranzicijske i emisijske vjerojatnosti koristimo već spomenute oznake a_{kl}, odnosno $e_k(b)$.

3.1 Viterbijev algoritam

Prije nego opišemo Viterbijev algoritam moramo objasniti razliku između niza stanja i niza simbola. Niz stanja nazivamo stazom π. Sama staza slijedi Markovljev lanac, tako da vjerojatnost stanja ovisi o prethodnom stanju. Lanac je karakteriziran parametrima

$$a_{kl} = \mathbb{P}(\pi_i = l | \pi_{i-1} = k).$$

Tranzicijska vjerojatnost a_{0k} se može smatrati vjerojatnošću da počnemo u stanju k. Budući da smo razdvojili simbole b od stanja k, moramo uvesti novi skup parametara za model, $e_k(b)$. Općenito, stanje može proizvesti simbol iz distribucije preko svih mogućih simbola. Stoga definiramo

$$e_k(b) = \mathbb{P}(x_i = b | \pi_i = k),$$

vjerojatnost da je simbol b vidljiv u stanju k. Ove vjerojatnosti su poznate i kao emisijske vjerojatnosti. Zajednička distribucija niza opservacija X i niza stanja π je

[1] Preciznije, radi se o jednom algoritmu i njegovim modifikacijama

definirana s

$$P(X,\pi) = a_{0\pi_1}\prod_{i=1}^{L} e_{\pi_i}(X_i)a_{\pi_i\pi_{i+1}},\ \pi_{L+1}=0 \tag{3.1}$$

Viterbijev algoritam

Dekodiranje je postupak pronalaska "značenja" niza simbola, odnosno pridruživanje nekog niza stanja danom nizu simbola. Najčešći pristup dekodiranju je algoritam dinamičkog programiranja, Viterbijev algoritam.
Ukoliko smijemo izabrati samo jednu stazu za naše predviđanje, trebali bismo izabrati onu s najvećom vjerojatnosti,

$$\pi^* = \max_{\pi} \mathbb{P}(X,\pi).$$

Najvjerojatnija staza π^* se može naći rekurzivno. Pretpostavimo da je za sva stanja k poznata vjerojatnost najvjerojatnije staze koja u stanju k završava s opservacijom i, $v_k(i)$. Onda se ove vjerojatnosti mogu izračunati za sva opažanja X_{i+1} kao

$$v_l(i+1) = e_l(X_{i+1})\max_k(v_k(i)a_{kl}).$$

Svi nizovi moraju početi u stanju 0, tako da je početni uvjet $v_0(0)=1$.
Viterbijev algoritam se sastoji od četiri koraka:

1. **Inicijalizacija** (i=0):
$$v_0(0)=1,\ v_k(0)=0,\ k>0$$
2. **Rekurzija** (i=1,...,L):
$$v_l(i) = e_l(X_i)\max_k(v_k(i-1)a_{kl})$$
$$ptr_i(l) = argmax_k(v_k(i-1)a_{kl})$$
3. **Kraj**:
$$P(X,\pi^*) = \max_k v_k(L)a_{k0}$$
$$\pi_L^* = argmax_k(v_k(L)a_{k0})$$
4. **Povratak unazad** (i=L,...,1):
$$\pi_{i-1}^* = ptr_i(\pi_i^*)$$

Uočimo da je završno stanje pretpostavljeno i zbog toga je a_{k0} u završnom koraku.
Najteži problem u praksi je što množenje mnogo malih vjerojatnosti uvijek vodi malim brojevima, a to daje underflow greške na računalu. Zato se Viterbijev algoritam uvijek treba izvoditi u log-prostoru, odnosno, trebamo računati $\log(v_l(i))$. Tako će se produkti pretvoriti u sume i brojevi će ostati razumni.

Viterbijevo treniranje

Funkcija cilja u procjeni maksimalne vjerodostojnosti je maksimizacija relacije (3.1) preko svih parametara π za dani niz simbola X. Viterbijevo treniranje je iterativan proces koji garantira monotoni rast vjerodostojnosti kroz skup ponovo procjenjenih parametara.
Neka je zadan fiksan model M, te neki inicijalni parametri θ. Viterbijevim algoritmom pronađemo najbolji put π kroz model M. Na taj način svakom simbolu od X pridružimo stanje. Sad možemo odrediti emisijske i tranzicijske frekvencije. Prema lemi 1.3.11, relativne frekvencije su procjenitelji maksimalne vjerodostojnosti, one su novi parametri modela i proces se ponovo iterira.
Kako su parametri potpuno određeni stazom, u trenutku kad se staza prestane mijenjati, prestaju se mijenjati i parametri modela.
Opisana procedura pronalazi vrijednost θ koja maksimizira doprinos najvjerojatnije staze za sve nizove vjerodostojnosti $\mathbb{P}(X_1, ..., X_n|\theta, \pi^*(X_1), ..., \pi^*(X_n))$.
Viterbijevo treniranje je metoda sa širokom primjenom i može se koristiti kad je primarna svrha HMMova proizvesti dekodiranje preko Viterbi poravnanja.

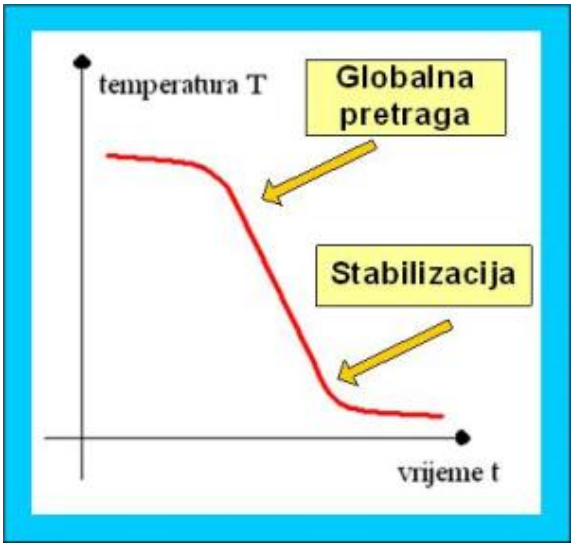
3.2 Deterministično kaljenje

Značajno poboljšanje naspram standardnim nadziranim i nenadziranim metodama učenja je primjena metode deterministčkog kaljenja. Ova metoda ima dva važna svojstva:

1. mogućnost izbjegavanja lokalnih optimuma
2. primjenjivost na mnogo različitih struktura

Simulirano kaljenje je generalna vjerojatnosna metaheuristika[2] za optimizacijski problem lociranja dobre aproksimacije za globalni optimum zadane funkcije.
Često se koristi kad tražimo optimum na diskretnom prostoru. Za neke probleme je simulirano kaljenje učinkovitije od iscrpnog nabrajanja - pod uvjetom da je cilj samo pronaći prihvatljivo najbolje rješenje u određenom vremenskom razdoblju, a ne najbolje moguće rješenje.
Ime i motivacija dolaze od kaljenja u metalurgiji. To je tehnika koja uključuje grijanje i kontrolirano hlađenje materijala da bi se povećala veličina kristala u materijalu i smanjili njegovi nedostaci. Svojstva materijala ovise o termodinamičkoj slobodnoj

[2]tehnika rješavanja koja se primjenjuje na širok skup optimizacijskih problema velike složenosti koji nisu rješivi korištenjem tradicionalnih pristupa



Slika 3.1: Globalna pretraga i stabilizacija kod simuliranog kaljenja

energiji. Grijanje i hlađenje materijala utječe na temperaturu i termodinamičku slobodnu energiju.
Kako se povećava hlađenje, proporcionalno se smanjuje temperatura, dok su temperatura i termodinamička energija proporcionalne vrijedosti.
Pojam sporog hlađenja je implementiran u algoritam simuliranog kaljenja u obliku sporog smanjenja vjerojatnosti prihvaćanja lošeg rješenja dok istražujemo prostor rješenja. Prihvaćanje lošijeg rješenja je osnovno svojstvo metaheuristike jer omogućava opsežniju potragu za optimalnim rješenjem.

Kod deterministčkog kaljenja, proces je deterministički što znači da ne želimo "slučajno" lutati po prostoru parametara dok postepeno napredujemo u maksimizaciji vjerodostojnosti. S druge strane, to je još uvijek metoda kaljenja koja teži globalnom optimumu, umjesto da pohlepno zapne u obližnjem lokalnom optimumu. U deterministčkom kaljenju pristup se formalno temelji na principima teorije informacija i teorije vjerojatnosti.

U početku zadajemo visoku vrijednost parametra kaljenja γ (temperaturu) na nekom intervalu i kaljenje se provodi na konveksnoj kombinaciji parametara maksimalne entropije i deterministički izračunatih parametara (u simuliranom kaljenju parametri su odabrani na slučajan način). U početku je doprinos deterministički određenih parametara jako mali, a što se više približavamo cilju, omjer se mijenja u njihovu korist.

Dodavanje parametara maksimalne entropije služi izbjegavanju lokalnih optimuma. Konkretno, u prvoj iteraciji zadamo inicijalne parametre modela. Ulazni parametri za Viterbijevo treniranje su konveksne kombinacije parametara maksimalne entropije i inicijalnih parametara. Kad je izračunat najvjerojatniji put kroz model, računamo tranzicijske i emisijske relativne frekvencije. U svakom sljedećem koraku su ulazni parametri konveksne kombinacije parametara maksimalne entropije i relativnih frekvencija izračunatih u prethodnom koraku.

Deterministično kaljenje se provodi na sljedeći način:
Inicijalizacija:

T_u= zadani tranzicijski parametri
E_u= zadani emisijski parametri
flT i flE su tranzicijski odnosno emisijski parametri maksimalne entropije

Petlja:
Dok je brojač manji od ukupnog broja iteracija radi:

1. $$\begin{cases} T = \gamma flT + (1-\gamma)T_u \\ E = \gamma flE + (1-\gamma)E_u \end{cases}$$

   γ - *parametar kaljenja*

2. Viterbijevim algoritmom računamo najvjerojatniju stazu kroz model i maksimalnu vjerodostojnost

3. Računamo relativne tranzicijske i emisijske frekvencije

4. $$\begin{cases} T_u = \textit{relativne tranzicijske frekvencije izračunate u koraku 3.} \\ E_u = \textit{relativne emisijske frekvencije izračunate u koraku 3.} \end{cases}$$

5. Provjera uvjeta petlje

Kraj:
Imamo matrice relativnih tranzicijskih i emisijskih parametara i maksimalnu vjerodostojnost koju smo dobili u koraku 2.

Poglavlje 4

Rezultati

4.1 Maksimizacija vjerodostojnosti

Nakon što smo zadali skriveni Markovljev model, moramo maksimizirati vjerodostojnost.
Budući da smo simulacijom dobili niz simbola, Viterbijevim algoritmom računamo najvjerojatniji niz stanja koji je emitirao dani niz simbola. Sad imamo niz stanja i niz simbola, pa računamo relativne tranzicijske i emisijske frekvencije. Dobivene vrijednosti su novi parametri modela i s tim parametrima ponovo iteriramo proces.
Rezultati Viterbijevog treniranja ovise o zadanim inicijalnim parametrima modela i što su ti parametri udaljeniji od onih s kojima smo simulirali niz, procjena je lošija. Viterbijevo treniranje "pogodi" koji broj na kocki ima najveću vjerojatnost.

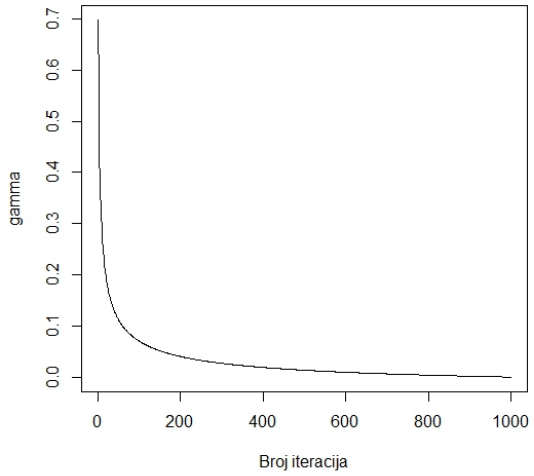
Kako nam je cilj procjena parametara modela, koristimo deterministično kaljenje, tehniku modifikacije Viterbijevog treniranja. Dodavanjem buke u model pokušavamo zaobići lokalne maksimume. Parametar kaljenja γ smo definirali kao padajuću funkciju koja ovisi o broju iteracija, $\gamma = f(it)$. U samom početku procesa dodajemo jako puno buke, a kako raste broj iteracija, tako smanjujemo buku.

$$T = \gamma * flT + (1 - \gamma) * T_u$$

$$E = \gamma * flE + (1 - \gamma) * E_u.$$

Pronaći parametar kaljenja nije nimalo lagan posao. Kandidate smo tražili među linearnim funkcijama, potencijama i korjenima koristeći početne uvjete:

$$\begin{cases} f(0) = 0.999 \\ f(br.iteracija - 1) = 0.001 \end{cases}$$



Slika 4.1: Graf parametra γ

Ispostavilo se da rješenju našeg problema vodi funkcija

$$\gamma = f(it) = \frac{\alpha}{\sqrt{(it+1)}} + \beta \tag{4.1}$$

Bukom nazivamo tranzicijsku i emisijsku matricu u kojoj su vjerojatnosti uniformno distribuirane. To su redom:

$$flT = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$$

$$flE = \begin{pmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}$$

Prema lemi (1.4.3) flT i flE su parametri maksimalne entropije.
U prvoj iteraciji su T_u, odnosno E_u zadani inicijalni parametri, a kasnije se u te varijable "spremaju" relativne frekvencije modela. Svaki korak se sastoji od računanja najvjerojatnijeg niza stanja Viterbijevim algoritmom i, jednom kad to imamo, računamo tranzicijske i emisijske relativne frekvencije za koje smo već rekli da su parametri maksimalne vjerodostojnosti. Imamo nove parametre modela i ponovo iteriramo proces. Promatrajući vjerodostojnost parametara s kojima smo simulirali niz simbola, uočavamo da je ona manja od one koju smo dobili za parametre koji su rezultat deterministčkog

kaljenja. Zaključujemo da smo simulirali model s bukom.
Deterministično kaljenje na našem primjeru uspješno zaobilazi lokalne maksimume i daje maksimalnu vjerodostojnost modela.

4.2 Simulacija i optimizacija

U radu je korišten programski jezik Python.

Simulacija i Viterbijevo treniranje

Simulirali smo niz duljine 60 000 koristeći dvije nesimetrične kocke. Prva kocka ima vjerojatnost da padne nula 0.6, a ostale vjerojatnosti su 0.08, dok druga kocka ima vjerojatnost da padne petica 0.5, a vjerojatnosti preostalih simbola su 0.1. Dakle, model ima stanja

$$S = \{K_0, K_5\}$$

Matrica tranzicijskih vjerojatnosti je dijagonalno dominantna i izgleda ovako:

$$T = \begin{pmatrix} 0.95 & 0.05 \\ 0.1 & 0.9 \end{pmatrix}$$

dok je matrica emisijskih vrijednosti

$$E = \begin{pmatrix} 0.6 & 0.08 & 0.08 & 0.08 & 0.08 & 0.08 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.5 \end{pmatrix}$$

Uočavate da smo niz emisijskih vjerojatnosti pomakli ulijevo, odnosno, umjesto skupa vrijednosti igraćih kocki {1,2,3,4,5,6}, koristimo skup {0,1,2,3,4,5}. Dakle, zadali smo skriveni Markovljev model.

Za procjenu parametara modela, koristili smo Viterbijevo treniranje, algoritam koji pronalazi skup parametara θ koji maksimizira vjerodostojnost najvjerojatnijeg niza skrivenih stanja. Inicijalne parametre s kojima pokrećemo Viterbijev algoritam smo prikazali kao uređene 16-orke (tranzicijske vjerojatnosti prve kocke, tranzicijske vjerojatnosti druge kocke, emisijske vjerojatnosti prve kocke, emisijske vjerojatnosti druge kocke) te kroz 20 iteracija pokušali doći do parametara s kojima smo simulirali model i maksimizirati vjerodostojnost. Sukladno očekivanju, Viterbijevo treniranje povećava vjerodostojnost s brojem iteracija, no što su inicijalni parametri udaljeniji od onih s kojima smo simulirali niz, to dobivamo slabije procjene parametara.
Razlog tomu je što Viterbijev algoritam zapne u lokalnom maksimumu.

Za inicijalne parametre

$$T = \begin{pmatrix} 0.55 & 0.45 \\ 0.45 & 0.55 \end{pmatrix}$$

$$E = \begin{pmatrix} 0.25 & 0.15 & 0.15 & 0.15 & 0.15 & 0.15 \\ 0.15 & 0.15 & 0.15 & 0.15 & 0.15 & 0.25 \end{pmatrix}$$

Viterbijevim treniranjem smo dobili sljedeće rezultate:

$$T_{res} = \begin{pmatrix} 0.85 & 0.15 \\ 0.33 & 0.67 \end{pmatrix}$$

$$E_{res} = \begin{pmatrix} 0.546 & 0.107 & 0.106 & 0.108 & 0.107 & 0.025 \\ 0.026 & 0.077 & 0.077 & 0.076 & 0.076 & 0.668 \end{pmatrix}$$

Maksimalna vjerodostojnost nakon dvadesete iteracije iznosi: -95887.615193.

Za inicijalne parametre

$$T = \begin{pmatrix} 0.95 & 0.05 \\ 0.1 & 0.9 \end{pmatrix}$$

$$E = \begin{pmatrix} 0.6 & 0.08 & 0.08 & 0.08 & 0.08 & 0.08 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.5 \end{pmatrix}$$

Viterbijevim treniranjem smo dobili sljedeće rezultate:

$$T_{res} = \begin{pmatrix} 0.95 & 0.05 \\ 0.08 & 0.92 \end{pmatrix}$$

$$E_{res} = \begin{pmatrix} 0.515 & 0.092 & 0.092 & 0.094 & 0.093 & 0.115 \\ 0.084 & 0.115 & 0.113 & 0.112 & 0.111 & 0.464 \end{pmatrix}$$

Maksimalna vjerodostojnost iznosi nakon dvadesete iteracije iznosi: -91777.320040.

Kako smo odredili da se vjerodostojnost svake iteracije upiše u niz, mogli smo promatrati kako se ponaša kroz iteracije. Za zadanih 37 kombinacija različitih od parametara s kojima smo simulirali model uočili smo da vjerodostojnost raste do određene iteracije, a onda se više ne mijenja.
Međutim, promatrajući Viterbijevo treniranje na parametrima s kojima smo simulirali model, uočili smo da je najveća vjerodostojnost u prvoj iteraciji. Potom, između druge i osme iteracije naizmjenice raste i pada, da bi se od devete iteracije stabilizirala i u konačnici iznosila -91777.320040. Razlog tomu je ubacivanje pseudozbroja od 3%. Naime, tvrdimo da je vrijednost svake tranzicije, odnosno emisije barem 0.03.

Deterministično kaljenje

Budući da rezultati Viterbijevog treniranja ovise o početnim parametrima modela, modificirali smo algoritam i uveli deterministično kaljenje. Parametar kaljenja $\gamma \in (0, 1)$ zadan je s

$$\gamma = \frac{\alpha}{\sqrt{it+1}} + \beta$$

Tranzicijska matrica T dobivena je kao konveksna kombinacija matrice s "flat" parametrima i matrice s inicijalnim parametrima, odnosno

$$T = \gamma \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix} + (1-\gamma) \begin{pmatrix} 0.55 & 0.45 \\ 0.45 & 0.55 \end{pmatrix}, \ \gamma \in (0.001, 0.999)$$

Analogno, matricu emisijskih vjerojatnosti definiramo s

$$E = \gamma \begin{pmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix} + (1-\gamma) \begin{pmatrix} \frac{1}{4} & \frac{3}{20} & \frac{3}{20} & \frac{3}{20} & \frac{3}{20} & \frac{3}{20} \\ \frac{3}{20} & \frac{3}{20} & \frac{3}{20} & \frac{3}{20} & \frac{3}{20} & \frac{1}{4} \end{pmatrix}$$

Broj iteracija je postavljen na 1000.

Postupak se iterira nakon što se odrede novi parametri modela. Dok u Viterbijevom treniranju bez modifikacije vjerodostojnost prestane rasti nakon nekoliko iteracija, u treniranju s determinističkim kaljenjem, ona neprestano raste.

Nakon 1000 iteracija smo dobili sljedeće rezultate:

$$T_{res} = \begin{pmatrix} 0.871 & 0.129 \\ 0.329 & 0.671 \end{pmatrix}$$

$$E_{res} = \begin{pmatrix} 0.608 & 0.097 & 0.097 & 0.099 & 0.098 & 2.32 * 10^5 \\ 11.79 * 10^5 & 0.057 & 0.056 & 0.055 & 0.055 & 0.777 \end{pmatrix}$$

Maksimalna vjerodostojnost iznosi: -93747.3445166.

Za inicijalne parametre

$$T = \begin{pmatrix} 0.95 & 0.05 \\ 0.1 & 0.9 \end{pmatrix}$$

$$E = \begin{pmatrix} 0.6 & 0.08 & 0.08 & 0.08 & 0.08 & 0.08 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.5 \end{pmatrix}$$

deterministično kaljenje je dalo sljedeće rezultate:

$$T_{res} = \begin{pmatrix} 0.983 & 0.017 \\ 0.047 & 0.953 \end{pmatrix}$$

$$E_{res} = \begin{pmatrix} 0.567 & 0.079 & 0.079 & 0.081 & 0.081 & 0.114 \\ 0.072 & 0.106 & 0.103 & 0.102 & 0.101 & 0.516 \end{pmatrix}$$

Maksimalna vjeroodostojnost nakon 1000 iteracija iznosi: -90748.6250036.

Bibliografija

[1] A. Allahverdyan, A. Galstyan,*Comparative Analysis of Viterbi Training and Maximum Likelihood Estimation for HMMs*, USC Information Sciences Institute, USA

[2] M. Bujanović, *Predikcija suprasekundarne strukture proteina i HMM*, diplomski rad, PMF-MO, Zagreb, 2012.

[3] R. Durbin, S. Eddy, A. Krogh, G. Mitchinson, *Biological sequence analysis*, Cambridge University Press, 1998.

[4] B. Guljaš, *Matematička analiza I & II*, PMF-MO predavanja, 2014.

[5] M. Huzak, *Matematička statistika*, PMF-MO predavanja, 2012.

[6] K. Rose, *Deterministic Annealing for Clustering, Compresion, Classification, Regression and Related Optimization Problems*, IEEE, (1998.), 2210-2239

[7] M. Rudman, *Kompleksnost skrivenih Markovljevih modela*, diplomski rad, PMF-MO, Zagreb, 2014.

[8] N. Sarapa, *Teorija vjerojatnosti*, Školska knjiga, Zagreb, 2002.

[9] Z. Vondraček, *Markovljevi lanci*, PMF-MO skripta, 2008.

[10] https://element.hr/artikli/file/1357

[11] http://en.wikipedia.org/wiki/Simulated_annealing

Sažetak

U ovom diplomskom radu smo se bavili skrivenim Markovljevim modelima, moćnim statističkim alatom koji je namijenjen modeliranju nizova koje generira neki skriveni proces.
Dali smo formalnu definiciju skrivenog Markovljevog modela, opisali varijante algoritma koje koristimo i implementirali ih u programskom jeziku Python.
Na primjeru dvije nesimetrične kocke konstruirali smo Markovljev model i za procjenu parametara i povećanje vjerodostojnosti iskoristili Viterbijevo treniranje sa i bez determinističkog kaljenja. Uočili smo da nam Viterbijevo treniranje sa determinističkim kaljenjem radi bolje od treniranja bez kaljenja.

Summary

This thesis is concerned with a statistical model called hidden Markov model (HMM), powerful statistical tool designed for modelling sequences generated by hidden processes.
We gave a formal definition of the hidden Markov model, described various algorithms and implemented them in Python.
Using the example of two asymmetric dies, we designed an HMM and used Viterbi training with and without deterministic annealing to optimize the parameters. We have noticed that Viterbi training with deterministic annealing works better than the one without annealing.

Životopis

- rođena sam 23. listopada 1989. u Šibeniku
- od 1996. do 2004. pohađam OŠ Petra Krešimira IV. u Šibeniku
- od 2004. do 2008. pohađam Turističko-ugostiteljsku školu u Šibeniku
- od 2008. do 2012. pohađam preddiplomski studij Matematika na PMF-MO u Zagrebu
- 2012. upisujem diplomski sveučilišni studij Matematička statistika na PMF-MO u Zagrebu