SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO–MATEMATIČKI FAKULTET
MATEMATIČKI ODSJEK

Irma Valčić

ANALIZA KOMPLEKSNOSTI SKRIVENIH MARKOVLJEVIH MODELA

Diplomski rad

Voditelj rada:
doc. dr. sc. Pavle Goldstein

Zagreb, srpanj 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. _____

Zahvaljujem mentoru doc. dr. sc. Pavlu Goldsteinu na strpljenju, savjetima, redovnim dozama kofeina i životnim mudrostima tijekom izrade ovog rada.

Isto tako, hvala mojim roditeljima, sestri i mojoj baki na svoj podršci koju su mi nesebično pružali ne samo tijekom izrade rada, nego i tijekom cijelog školovanja.

Konačno, hvala svim mojim prijateljima i kolegama koji su bili uz mene i bez čije ogromne podrške i puno strpljenja ovog rada danas ne bi bilo. Puno hvala svima!

Sadržaj

Uvod

Svrha ovog diplomskog rada jest opis i primjena nekih statističkih alata na skrivene Markovljeve modele. Skriveni Markovljevi modeli danas nalaze primjenu u mnogim područjima: prepoznavanju govora, prepoznavanju rukopisa, računalnom prevođenju, analizi vremenskih nizova itd. Veliku primjenu također pronalaze u bioinformatici, znanosti koja se bavi analizom bioloških nizova, sadržaja i organizacije genoma te predviđanjem strukture i funkcije makromolekula primjenom tehnika iz matematike, statistike i računarstva. Specijalno, model povremeno nepoštene kockarnice korišten u ovom diplomskom radu u bioinformatici se često primjenjuje za modeliranje genoma.

Ovim radom želimo dati kratki pregled teorije skrivenih Markovljevih modela, dati primjer njihove implementacije i pojasniti postupke za procjenu parametara modela. K tome promatramo i neke metode za procjenu kompleksnosti skrivenih Markovljevih modela.

U prvom poglavlju definiramo osnovne pojmove iz teorije vjerojatnosti i statistike koji će nam biti potrebni u kasnijoj analizi. U drugom poglavlju dajemo primjer i formalnu definiciju skrivenih Markovljevih modela. U trećem poglavlju opisujemo nekoliko algoritama za procjenu parametara skrivenog Markovljevog modela, dok u posljednjem, četvrtom, poglavlju predstavljamo rezultate i zaključke rada.

Poglavlje 1

Osnovni pojmovi

1.1 Vjerojatnost

Definicija 1.1.1. *Pod* ***slučajnim pokusom*** *podrazumijevamo takav pokus čiji ishodi odn. 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 puta. Tada broj N zovemo* ***frekvencija*** *događaja A, a broj $\frac{N}{n}$* ***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* ***neprekidan*** *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* σ***-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. Uređeni par $(\Omega, \mathcal{F})$ zove se* ***izmjeriv prostor****.*

Sada 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 $(\Omega, \mathcal{F}, \mathbb{P})$ pri čemu 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 $\mathbb{P}(A),\ A \in \mathcal{F}$ zove se* ***vjerojatnost događaja*** *A.*

Definicija 1.1.9. *Označimo sa $\mathcal{B}$ $\sigma-$algebru generiranu familijom svih otvorenih skupova na skupu realnih brojeva $\mathbb{R}$. $\mathcal{B}$ zovemo σ-**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 1.1.10. *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 1.1.11. *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 1.1.12. *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 1.1.13. ***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 1.1.14. ***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 Statistika

Definicija 1.2.1. *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 $I(\theta)$ zove se ***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.*

Definicija 1.2.2. *Neka je $(\Omega, \mathcal{F})$ izmjeriv prostor i $\mathcal{P}$ familija vjerojatnosnih mjera na $(\Omega, \mathcal{F})$. Uređena trojka $(\Omega, \mathcal{F}, \mathcal{P})$ zove se* ***statistička struktura****.*

Definicija 1.2.3. *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 1.2.4. *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 procjenitelj maksimalne vjerodostojnosti (****MLE****) ako vrijedi*

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

Definicija 1.2.5. *Za opaženu vrijednost* x *od* $\mathbb{X}_n$, $l : \Theta \to \mathbb{R}$,

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

zovemo ***log-vjerodostojnost****.*

Definicija 1.2.6. *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 1.2.7. *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 1.2.8. *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 korijen* $\hat{\theta}_n = \hat{\theta}_n(X_n)$ *takav da je* $\hat{\theta}_n \xrightarrow{P_\theta} \theta$, *za* $n \to \infty$.

Napomena 1.2.9. *Ako jednadžba vjerodostojnosti ima jedinstvenu stacionarnu točku* $\hat{\theta}_n \xrightarrow{\mathbb{P}_{\theta_0}} \theta_0$*, tada Teorem 1.2.8 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 1.2.10. *Neka je* $X \sim B(n, \theta)$ *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:

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

Odakle slijedi
$k - k\theta - n\theta + k\theta = 0$
$\Rightarrow n\theta = k$
$\Rightarrow \theta = \frac{k}{n}$

□

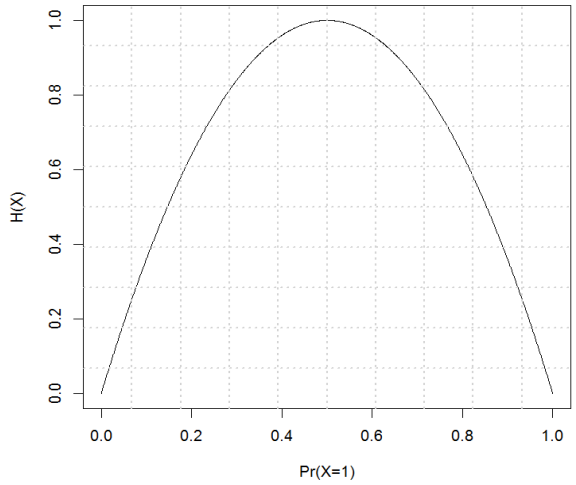
1.3 Entropija

Definicija 1.3.1. *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*

$$H(X) = -\sum_{i=1}^{K} \mathbb{P}(x_i)\log(\mathbb{P}(x_i)) \tag{1.3}$$

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

Jednostavnim uvrštavanjem u formulu entropije dobivamo $H(X) = -\log(0.5)$. Promatramo li logaritam u bazi 2, dobivamo $H(X) = 1$ bit po simbolu. Dakle, vrijednost entropije u ovisnosti o vjerojatnosti pojave pisma/glave kod bacanja novčića, odnosno srednji sadržaj informacije poruke koja se sastoji od uzastopnih rezultata bacanja novčića jest 1 bit po simbolu.
Za slučaj nepoštenog novčića koji uvijek daje pismo ($p(x_1) = 1,\ p(x_2) = 0$) očekivano dobivamo $H(X) = 0$. Uvrštavanjem svih mogućih vjerojatnosti pojave pisma u formulu entropije, dobivamo graf ovisnosti vrijednosti entropije o toj vjerojatnosti (1.1). Maksimum je postignut kada je vjerojatnost pisma jednaka vjerojatnosti glave ($p = \frac{1}{2}$) tj. 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.



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

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 1.3.2. *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 1.3.3. *Uniformno distribuirani parametri imaju maksimalnu entropiju.*

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

Teorem 1.3.4. *(Bolzano-Weierstrass): Neka je funkcija $f : [a, b] \to \mathbb{R}$ neprekidna na segmentu $[a, b] \subset \mathbb{R}$. Tada je $f([a, b]) = [m, M]$ također segment.*

Napomena 1.3.5. *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 1.3.6. *(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 (1.3.3)): 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 Rolleovom 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 $p_M = (p_1, ..., p_k)$

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

. □

1.4 AIC i BIC

Osim promatranja funkcije vjerodostojnosti ili omjera vjerodostojnosti, za odabir najboljeg statističkog modela mogu se koristiti još neki kriteriji. Dva često korištena kriterija su **AIC** (Akaike Information Criterion) i **BIC** (Bayesian Information Criterion). AIC i BIC oba se temelje na funkciji vjerodostojnosti, a mjere koliko „dobro" model opisuje podatke. Dani su sljedećim jednadžbama:

$$AIC = -2log(L) + 2k \tag{1.4}$$

$$BIC = -2log(L) + klog(m) \tag{1.5}$$

pri čemu je L maksimalna vjerodostojnost modela, m duljina niza, a k broj slobodnih parametara. Informacijski kriteriji temelje se na funkciji vjerodostojnosti no osim njene vrijednosti, također sadržavaju i penalizaciju za kompleksnost modela. Znamo, općenito, da pri određivanju modela isti možemo poboljšati dodajući mu parametre. No s druge strane, cilj nam je da model bude što jednostavniji tj. da ima što manje parametara. Primjećujemo da i AIC i BIC penaliziraju količinu parametara (BIC više nego AIC), a time ujedno mogu riješiti i problem *overfittinga* do kojeg može doći dodamo li previše parametara u model. Budući da AIC i BIC zapravo procjenjuju koliko je informacija izgubljeno modeliranjem podataka, najbolji model bit će onaj s najmanjim AIC-om (BIC-om).

Poglavlje 2

Skriveni Markovljevi modeli

2.1 Markovljevi lanci

Definicija 2.1.1. *Neka je S skup.* ***Slučajni proces*** *s diskretnim vremenom i prostorom stanja S jest 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 2.1.2. *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*

$$\mathbb{P}(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, ..., X_0 = i_0) = \mathbb{P}(X_{n+1} = j | X_n = i) \tag{2.1}$$

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 (2.1) naziva se *Markovljevim svojstvom.*

Definicija 2.1.3. *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****.*

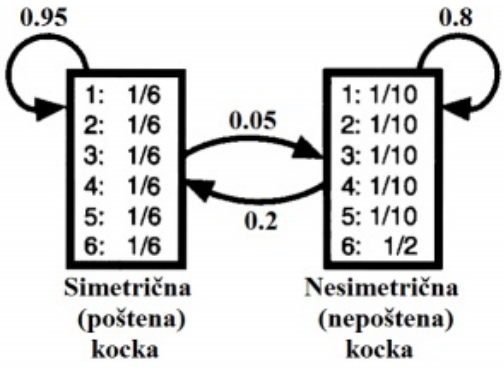
Markovljev lanac zajedno sa zadanim prijelaznim vjerojatnostima nazivamo **Markovljevim modelom**.

2.2 Skriveni Markovljevi modeli

Skriveni Markovljevi modeli (na engleskom hidden Markov model ili **HMM**) su statistički modeli koji nalaze široku primjenu u molekularnoj biologiji, prepoznavanju

govora i računalnom prevođenju. Radi lakšeg razumijevanja, u ovom ćemo dijelu prvo dati primjer jednog skrivenog Markovljevog modela, a potom iznijeti formalne definicije.

Primjer HMM-a: povremeno nepoštena kockarnica



Slika 2.1: Primjer skrivenog Markovljevog modela s dvije kocke

Imamo dvije igraće kocke, jednu poštenu i jednu nepoštenu. Na kocki K_P, poštenoj, vjerojatnosti da padne neki broj od 1 do 6 jesu jednake za svaki od brojeva i iznose $\frac{1}{6}$. Kocka K_N, nepoštena, češće pada na šesticu nego na bilo koji drugi broj. Vjerojatnost da padne šestica jest $\frac{1}{2}$, a vjerojatnost da padne bilo koji drugi broj jest $\frac{1}{10}$. Pretpostavimo da počinjemo bacanjem K_P. Vjerojatnost da ćemo ponovo koristiti K_P je 95%, dok je vjerojatnost da ćemo je zamijeniti sa K_N 5%. Kad smo jednom prešli iz K_P u K_N, u 80% slučajeva ćemo i nastaviti bacati K_N. Vjerojatnost da je zamijenimo sa K_P jest 20%.

Naš model zapisat ćemo na sljedeći način:

- Imamo dvije kocke tj. dva stanja.

$$\mathcal{S} = \{K_P, K_N\}$$

- Imamo 6 brojeva koji mogu pasti na kockama.

$$B = \{1, 2, 3, 4, 5, 6\}$$

- Matrica tranzicijskih vrijednosti dana je s:

$$A = \begin{pmatrix} 0.95 & 0.05 \\ 0.2 & 0.8 \end{pmatrix}$$

pri čemu je $a_{12} = \mathbb{P}(K_N|K_P)$ - vjerojatnost da je nakon K_P bačena K_N, $a_{21} = \mathbb{P}(K_P|K_N)$ - vjerojatnost bacanja K_P ako je prethodno bačena K_N i $a_{jj} = \mathbb{P}(K_j|K_j)$ - vjerojatnost da je nakon bacanja K_j ponovno bačena K_j, $j \in \{P, N\}$.

- Matrica emisijskih vjerojatnosti je:

$$E = \begin{pmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{2} \end{pmatrix}$$

Prvi redak čine emisijske vjerojatnosti elemenata iz B u stanju K_P, a drugi redak emisijske vjerojatnosti elemenata iz B u stanju K_N.

Proces koji modelira izbor kocaka jest 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*. Model povremeno nepoštene kockarnice osnova je za analizu genoma o čemu se više može pronaći u [2].

Imamo li niz simbola, odn. opaženih vrijednosti, primjerice $X = (1, 2, 5, 6, 6, 4, 3)$, 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 za 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 također pojasniti u sljedećem poglavlju

Formalna definicija HMM-a

Sada imamo intuitivnu predodžbu o tome što je skriveni Markovljev model. Kod skrivenog Markovljevog modela imamo niz stanja (kocke) i niz simbola (brojevi na kockama). Svaki simbol ovisi jedino o trenutnom stanju u kojem se proces nalazi pa zato generiranje simbola iz stanja modeliramo Markovljevim lancem nultog reda što je zapravo **niz nezavisnih događaja**. Formalno rečeno:

Definicija 2.2.1. ***Skriveni Markovljev model prvog reda*** *jest 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.
$$P(Q_t|Q_{t-1}, O_{t-1}, ..., Q_1, O_1) = \mathbb{P}(Q_t|Q_{t-1}) \tag{2.2}$$

2.
$$\mathbb{P}(O_t|Q_T, O_T, ..., Q_{t+1}, O_{t+1}, Q_t, Q_{t-1}, O_{t-1}, ..., Q_1, O_1) = P(O_t|Q_t) \tag{2.3}$$

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).

Relacija (2.2) kaže da je vjerojatnost da se, za neko $t \in \{1, 2, ..., N\}$, nalazimo u stanju Q_t uz uvjet da su se dogodila sva prethodna stanja $Q_1, ..., Q_{t-1}$ i da su emitirani simboli $O_1, ..., O_{t-1}$ jednaka **tranzicijskoj vjerojatnosti** iz stanja Q_{t-1} u stanje Q_t.

Relacija (2.3) povlači da realizacija nekog opažanja u sadašnjem stanju ovisi samo o tom stanju. Vjerojatnosti iz relacije (2.3) nazivamo **emisijske vjerojatnosti** i kažemo da stanje Q_t **emitira** simbol O_t.

Skriveni Markovljev model zadan je sljedećim parametrima:

- N - broj stanja u kojima se proces može nalaziti
$$S = \{1, ..., N\} \tag{2.4}$$
S - skup svih stanja procesa
- M - broj mogućih opažanja
$$B = \{b_1, ..., b_M\} \tag{2.5}$$
B - skup svih opaženih vrijednosti
- L - duljina opaženog niza
$$X = (x_1, ..., x_L) \tag{2.6}$$
X - opaženi niz

- A - matrica tranzicijskih vjerojatnosti

$$A = \{a_{ij}\}, a_{ij} = \mathbb{P}(Q_{t+1} = j | Q_t = i), 1 \leq i, j \leq N \tag{2.7}$$

- E - matrica emisijskih vjerojatnosti

$$E = \{e_j(k)\}, e_j(k) = \mathbb{P}(O_t = b_k | Q_t = j), 1 \leq j \leq N, 1 \leq k \leq M \tag{2.8}$$

Unatoč tome što nam je niz stanja u skrivenom Markovljevom modelu nepoznat, poznat nam je niz simbola i pomoću njega možemo donijeti neke zaključke o nizu stanja koji odgovara emitiranom nizu vrijednosti.

Poglavlje 3

Algoritmi za analizu HMM-ova

U ovom poglavlju objasnit ćemo algoritme koje smo spomenuli u prethodnom poglavlju, a koji su korišteni u izradi ovog rada, ali i neke koji nisu korišteni u ovom radu, ali su važni u primjeni na skrivene Markovljeve modele.

Uvedimo neke oznake. Niz emitiranih simbola označit ćemo s $x = (x_1, ..., x_n)$, a pripadajući niz skrivenih stanja s $\pi = (\pi_1, ..., \pi_n)$. Podsjetimo se, u skrivenom Markovljevom modelu niz stanja je nepoznat, ali o njemu možemo izvesti neke zaključke na temelju niza emitiranih vrijednosti. Za tranzicijske i emisijske vjerojatnosti koristit ćemo već spomenute oznake a_{ij} i $e_j(k)$.

Niz stanja (niz bačenih kocaka) nazivamo stazom π. Staza slijedi Markovljev lanac, tako da vjerojatnost stanja u trenutku t ovisi o prethodnom stanju, $t-1$. Lanac je karakteriziran parametrima

$$a_{ij} = \mathbb{P}(\pi_t = j | \pi_{t-1} = i).$$

Tranzicijsku vjerojatnost a_{0j} možemo smatrati vjerojatnošću da započnemo u stanju j. Budući da smo razdvojili simbole k od stanja j, moramo uvesti novi skup parametara za model, $e_j(k)$. Općenito, stanje može proizvesti bilo koji simbol po nekoj zadanoj distribuciji. Stoga definiramo

$$e_j(k) = \mathbb{P}(x_t = k | \pi_t = j),$$

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

$$P(X, \pi) = a_{0\pi_1} \prod_{t=1}^{L} e_{\pi_t}(X_t) a_{\pi_t \pi_{t+1}}, \ \pi_{L+1} = 0 \tag{3.1}$$

3.1 Viterbijev algoritam

Postupak pronalaska značenja niza simbola, odnosno pridruživanje nekog niza stanja danom nizu simbola, naziva se *dekodiranje*. Postoji nekoliko pristupa problemu dekodiranja, a jedan od najpoznatijih jest **Viterbijev algoritam**.

Viterbijev algoritam jest algoritam dinamičkog programiranja koji se koristi za pronalaženje optimalnog niza stanja u skrivenom Markovljevom modelu, a za zadani niz opažanja. Pretpostavimo da je za sva stanja k poznata vjerojatnost najvjerojatnijeg prolaza koji završava u stanju k simbolom (opažanjem) i, $v_k(i) = \mathbb{P}(x_1, ..., x_i|\pi_t = k)$. Tada se vjerojatnost optimalnog puta kroz model gdje su emitirani $x_1, ..., x_{i+1}$ i koji završava u stanju j može izraziti preko sljedeće rekurzije:

$$v_j(i+1) = e_j(x_{i+1}) \max_k (v_k(i)a_{kj}).$$

Viterbijev algoritam sastoji se od četiri koraka:

1. **Inicijalizacija** (i=0):

$$v_0(0) = 1,\ v_k(0) = 0,\ k > 0$$

2. **Rekurzija** (i=1,..., L):

$$v_j(i) = e_j(X_i) \max_k (v_k(i-1)a_{kj})$$

$$ptr_i(j) = argmax_k(v_k(i-1)a_{kj})$$

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):

$$\pi^*_{i-1} = ptr_i(\pi^*_i)$$

Napomena 3.1.1. *U praksi se računanje Viterbijevog algoritma izvodi u log-prostoru tj. računa se $log(v_j(i))$. Razlog tome jest taj da množenje velikog broja malih vjerojatnosti uvijek vodi veoma malim brojevima te nerijetko dolazi do underflowa pri računu. Prelaskom u log-prostor množenje tih malih vjerojatnosti pretvara se u zbrajanje i time se izbjegavaju računalne greške.*

Viterbijevo treniranje

Viterbijevo treniranje jedan je od pristupa procjeni optimalnih parametara skrivenih Markovljevih modela. Neka je zadan model M i neki inicijalni parametri θ. Viterbijevim algoritmom pronađemo najbolji put $\pi^* = \max_\pi \mathbb{P}(X, \pi)$ kroz model M i time svakom simbolu od X (broju na kocki) pridružimo stanje (kocku). Budući da imamo jedinstveno zadani put, time su tada zadane i emisijske i tranzicijske frekvencije. Poznat je rezultat (Lema 1.2.10) da su relativne frekvencije procjenitelji maksimalne vjerodostojnosti te njih uzimamo kao nove parametre modela i iteriramo proces. Budući da parametri koje dobivamo Viterbijevim treniranjem direktno ovise o putu, a broj puteva je konačan, ta metoda konvergira. Opisana procedura pronalazi vrijednost θ koja maksimizira doprinos najvjerojatnijeg puta.

3.2 Baum-Welchov algoritam

Uz Viterbijevo treniranje, jedan od najpoznatijih postupaka za treniranje modela jest **Baum-Welchov algoritam**. Kao i Viterbijevo treniranje, to je iterativni postupak za određivanje parametara modela na temelju niza opažanja. No za razliku od Viterbijevog treniranja, Baum-Welchov algoritam parametre procjenjuje tako da maksimizira očekivanje promatranog niza uzimajući u obzir sve puteve, a ne samo najbolji. Budući da nam je put stanja π nepoznat, ne možemo jednostavno prebrojati tranzicije i emisije nego moramo izračunati njihove očekivane vrijednosti. U svakoj iteraciji dobivamo novi skup vrijednosti parametara iz očekivanog broja emisija i tranzicija, uzimajući u obzir sve moguće puteve. Algoritam obično zaustavljamo nakon određenog broja iteracija ili kad promjena logaritma funkcije maksimalne vjerodostojnosti postane dovoljno mala. Može se pokazati da Baum-Welchov algoritam konvergira u lokalni optimum. Uobičajeno je da sustav ima mnogo lokalnih ekstrema pa konvergencija u jedan od njih uvelike ovisi o zadanim početnim parametrima.

Iako Baum-Welchov algoritam u općenitom slučaju daje bolje rezultate nego Viterbijevo treniranje, u ovom smo se radu odlučili na korištenje Viterbijevog treniranja zbog njegove manje složenosti i bitno jednostavnije implementacije.

3.3 Determinističko kaljenje

Pronalazak optimalnog rješenja za neke optimizacijske procese može biti iznimno težak zadatak. Problem često nastaje kada skup mogućih rješenja koji pretražujemo postane prevelik da bismo ga cijelog ispitali u nekom razumnom vremenu i u takvim

slučajevima često posežemo za procesima koji nam mogu dati neko rješenje „blizu" optimalnom. Jedan od takvih procesa jest **simulirano kaljenje**. Simulirano kaljenje jest stohastička tehnika za optimizacijski problem pronalaska dobre aproksimacije globalnog optimuma neke zadane funkcije. Ideja za ovu tehniku potječe iz kaljenja u metalurgiji: grijanje i hlađenje materijala u svrhu promjene njegovih fizičkih svojstava. Kako se metal počinje hladiti, njegova struktura se fiksira i počinje zadržavati nova svojstva. Preneseno u algoritam simuliranog kaljenja, imamo parametar kaljenja γ (temperaturu) kojem u početku zadamo visoku vrijednost i potom ga postupno smanjujemo. Kad je temperatura visoka, algoritam svako rješenje smatra gotovo jednako dobrim te će zbog toga prihvatiti i ona rješenja koja bi kasnije bila ocjenjena lošijima od našeg trenutnog. To dozvoljava algoritmu da iskoči iz lokalnih optimuma koje je možda pronašao u početku te u kojima bi inače možda bio zapeo. Kako smanjujemo temperaturu, smanjuje se i vjerojatnost prihvaćanja lošijih rješenja čime dozvoljavamo algoritmu da se postepeno koncentrira u jednom području i pronađe rješenje blisko optimalnom.

Determinističko kaljenje pokušava objediniti najbolje iz oba područja. Ono je deterministično — dakle nemamo „slučajnog" lutanja po prostoru parametara kao u simuliranom kaljenju — no još uvijek je riječ o metodi kaljenja koja pokušava pronaći globalni optimum umjesto da pohlepno ode u lokalni optimum. Kaljenje provodimo na konveksnoj kombinaciji parametara maksimalne entropije i deterministički izračunatih parametara. 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.

Determinističko 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

Glavno pitanje koje smo si u ovom radu postavili bilo je: *možemo li iz dovoljno velike količine podataka otkriti broj stanja?* Jednostavnije rečeno, možemo li isključivo na temelju niza dobivenih brojeva odrediti koliko smo kocaka bacali?

4.1 Simulacija i optimizacija

U radu je za simulaciju i optimizaciju korišten programski jezik Python, dok je za grafičku usporedbu korišten programski jezik R.

Za procjenu parametara maksimalne vjerodostojnosti modela moglo se koristiti nekoliko algoritama (v. Poglavlje 3) od kojih je u ovom radu korišteno Viterbijevo treniranje modificirano tehnikom determinističkog kaljenja.

Iz definicije entropije (1.3) vidimo da entropiju lako možemo dovesti u vezu s $-\log$-vjerodostojnosti. Budući da procesi optimizacije povećavaju vjerodostojnost, iz definicije entropije slijedi da zapravo smanjuju entropiju. Zbog toga smo za potrebe ovog rada simulirali podatke iz parametara male entropije tj. tako da je svaka kocka imala jedan dominantan simbol.

Tranzicijskom matricom T:

$$T = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}$$

i emisijskom matricom E:

$$E = \begin{pmatrix} 0.8 & 0.04 & 0.04 & 0.04 & 0.04 & 0.04 \\ 0.04 & 0.8 & 0.04 & 0.04 & 0.04 & 0.04 \\ 0.04 & 0.04 & 0.8 & 0.04 & 0.04 & 0.04 \end{pmatrix}$$

zadali smo skriveni Markovljev model kojeg interpretiramo kao bacanje triju asimetričnih kocaka. Potom smo iz tog modela simulirali niz duljine 60 000.

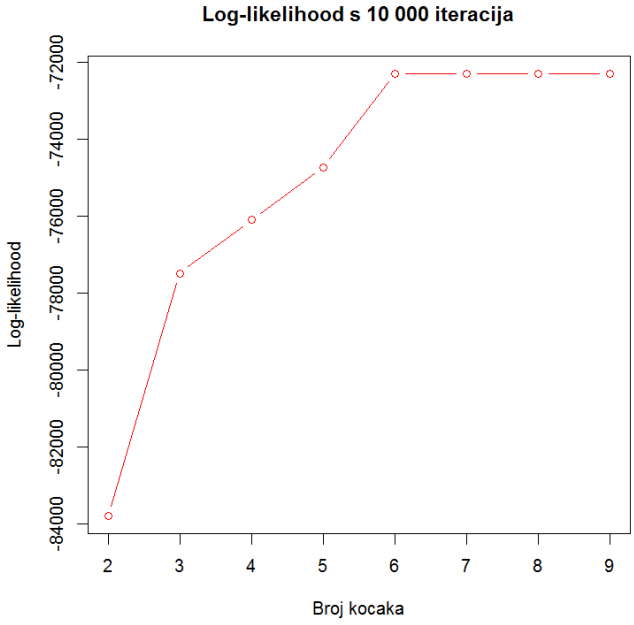
Sada smo isključivo na temelju simuliranog niza pokušali procijeniti parametre i odrediti maksimalnu vjerodostojnost modela pomoću Viterbijevog treniranja modificiranog deterministčkim kaljenjem. Broj iteracija postavljen je na 10 000.

Pri korištenju modificiranog Viterbijevog treniranja, primijetili smo pad vrijednosti funkcije vjerodostojnosti nakon određenog broja iteracija. Valja naglasiti da se on u teoriji ne bi smio događati jer je Viterbijevo treniranje modificirano metodom determinističkog kaljenja optimizacijski proces maksimizacije uvjetne vjerodostojnosti te bi vjerodostojnost trebala rasti sa svakom iteracijom. Ipak, do greške dolazi zbog postavljanja pseudozbroja relativnih frekvencija na inicijalnu vrijednost 1 (umjesto da krenemo od 0). Kako iteriramo proces, entropija sustava se smanjuje. No, u određenom trenutku entropija postane toliko mala da čak i inicijalne vrijednosti postavljene na 1 postanu veće od onoga što je algoritam izračunao te zbog toga vjerodostojnost tada počinje opadati. Smanjimo li inicijalne vrijednosti na nešto manje od 1, samo „odgađamo" trenutak pada vjerodostojnosti tj. iteraciju u kojem će se on dogoditi. Problem smo u ovom radu riješili tako da smo maksimizirali vjerodostojnost koliko god je to bilo moguće, a kada smo primijetili da je počela padati, prekinuli smo proces.

Pokušavajući odrediti broj stanja na temelju dobivenog niza, početni problem modelirali smo s $i = 1, 2, ..., 9$, $i = 50$ i $i = 100$ kocaka i dobili sljedeće rezultate:

$$
\begin{array}{cc}
i = 1 & L = -87880.9966977 \\
i = 2 & L = -83801.3822719 \\
i = 3 & L = -77489.7562758 \\
i = 4 & L = -76106.7141355 \\
i = 5 & L = -74736.3189648 \\
i = 6 & L = -72302.2190836 \\
i = 7 & L = -72306.5878387 \\
i = 8 & L = -72310.9525160 \\
i = 9 & L = -72315.3194397 \\
i = 50 & L = -72492.6390580 \\
i = 100 & L = -72705.5578453
\end{array}
$$

Na temelju ovih rezultata odmah smo odbacili $i = 50$ i $i = 100$ jer nam dobivena vjerodostojnost ne opravdava uvođenje tolikog broja stanja. Preostale rezultate zornije prikažimo grafom.

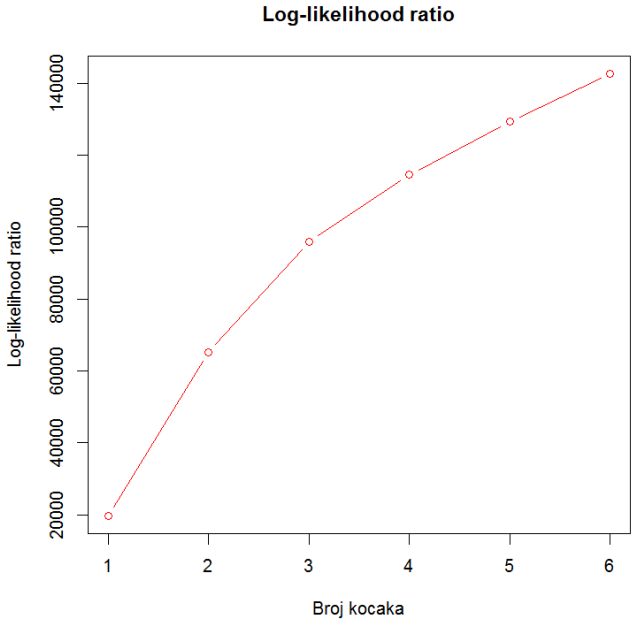


Iz grafa vidimo da čak i za $i = 7, 8, 9$ nemamo značajnog porasta vjerodostojnosti tako da smo se fokusirali na $i = 1, ..., 6$. Uočimo da funkcija vjerodostojnosti jako raste kada smo prešli s modela s 2 kocke na model sa 3 kocke (rast se smanjuje kada prijeđemo sa 3 na 4 kocke), te pri prijelazu s modela s 5 kocaka na model sa 6 (nakon 6 kocaka, funkcija više gotovo uopće ne raste).

Sljedeći pokušaj pronalaska najboljeg modela bio je test (log-)omjera vjerodostojnosti.

$$LLR = log\left[\frac{\mathbb{P}(X|M)}{\mathbb{P}(X|R)}\right]$$

pri čemu je LLR oznaka za log-likelihood ratio (logaritam omjera vjerodostojnosti), $\mathbb{P}(X|M)$ jest vjerodostojnost za svaki od modela (s 1, 2,..., 9 kocaka) izračunata Viterbijevim treniranjem modificiranim deterministčkim kaljenjem, a $\mathbb{P}(X|R) = \left(\frac{1}{6}\right)^m \left(\frac{1}{i}\right)^{m-1}$, pri čemu je m duljina niza, a i broj kocaka u modelu.



Ponovno vidimo da u trojci imamo maleni „lakat" na grafu.

Osim promatranja same vjerodostojnosti te omjera vjerodostojnosti, odlučili smo izračunati AIC i BIC kako bismo pokušali odrediti najbolji model za naš zadani niz. Rezultate prikazujemo sljedećom tablicom.

broj kocaka	**AIC**	**BIC**
1	175772.0	175817.0
2	167628.8	167745.8
3	155025.5	155232.6
4	152283.4	152598.5
5	149570.7	150011.8
6	**144734.4**	**145319.5**

Vidimo da se najmanje vrijednosti AIC-a i BIC-a postižu na modelu sa 6 kocaka. Budući da su naši podaci simulirani sa 3 kocke, zaključujemo da AIC i BIC nisu dobri kriteriji u našem slučaju. O neprikladnosti primjene AIC-a i BIC-a na skrivenim Markovljevim modelima više se može pronaći u [4, str. 106-107].

Bibliografija

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

[2] J. Ernst, M. Kellis, *Discovery and characterization of chromatin states for systematic annotation of the human genome*, http://www.nature.com/nbt/journal/v28/n8/full/nbt.1662.html (lipanj 2015.)

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

[4] I. L. MacDonald, W. Zucchini, *Hidden Markov and Other Models for Discrete-Valued Time Series*, Chapman & Hall, New York, 1997.

[5] I. S. Pandžić, *Uvod u teoriju informacije i kodiranje*, Element, Zagreb, 2009.

[6] M. Rudman, *Kompleksnost skrivenih Markovljevih modela (diplomski rad)*, Prirodoslovno-matematički fakultet, 2014.

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

[8] M. Tepić, *Kompleksnost skrivenih Markovljevih modela (diplomski rad)*, Prirodoslovno-matematički fakultet, 2015.

[9] Z. Vondraček, *Markovljevi lanci (skripta iz kolegija)*, http://web.math.pmf.unizg.hr/ vondra/ml12-predavanja.html (svibanj 2015.)

Sažetak

U ovom radu analizirali smo skrivene Markovljeve modele, statistički alat koji danas nalazi sve veću primjenu u različitim područjima. Dali smo njihovu formalnu definiciju, opisali neke algoritme za rad sa skrivenim Markovljevim modelima i implementirali ih u programskom jeziku Python. Konstruirali smo i primjer povremeno nepoštene kockarnice kao osnovu za kompliciranije primjene u bioinformatici (npr. modeliranje genoma), proveli simulaciju te pokušali pronaći najbolji model koji bi opisao tako dobivene podatke. Koristili smo nekoliko statističkih metoda za odabir najboljeg modela — maksimizaciju vjerodostojnosti, omjer vjerodostojnosti, AIC i BIC — no niti jedna od tih metoda nije dala dobar rezultat. To ukazuje na kompleksnost skrivenih Markovljevih modela unatoč tome što intuitivno ne djeluju kao nešto iznimno složeno.

Summary

This thesis explores hidden Markov models, a powerful statistical tool applied in various scientific fields. We give the formal definition of a hidden Markov model, describe several algorithms traditionally used in their analysis and present their implementation in Python. We also construct an example of an occasionally dishonest casino as the basis for more complicated applications in bioinformatics (e.g. genome analysis), simulate data and attempt to find the best model for it. Several statistical methods are used — likelihood maximization, log-likelihood ratio, AIC and BIC — but none of them yield satisfactory results. This indicates the complexity of hidden Markov models even though initially they may not appear particularly difficult.

Životopis

Rođena sam 1989. godine u Zagrebu. Od 1995. do 2003. pohadala sam osnovnu školu Vladimira Nazora, a 2007. godine završavam srednjoškolsko obrazovanje u zagrebačkoj V. gimnaziji. Iste godine upisujem preddiplomski studij matematike na Prirodoslovno-matematičkom fakultetu u Zagrebu. Akademske godine 2011/2012 provodim semestar u Beču na Universität Wien u sklopu programa studentske razmjene Erasmus, a 2013. godine nastavljam svoje obrazovanje diplomskim studijem na Prirodoslovno-matematičkom fakultetu, smjer Matematička statistika.