Kompleksnost skrivenih Markovljevih modela

Rudman, Margareta

SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO–MATEMATIČKI FAKULTET
MATEMATIČKI ODSJEK

Margareta Rudman

KOMPLEKSNOST SKRIVENIH MARKOVLJEVIH MODELA

Diplomski rad

Voditelj rada:
doc. dr. sc. Pavle Goldstein

Zagreb, srpanj 2014.

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. _____

Mojim roditeljima jer su mi omogućili da studiram. Mojim sestrama jer sam učila od njih. Svim mojim prijateljima i kolegama koji su mi uljepšali studensko razdoblje. I Svenu.

Sadržaj

Uvod

Tema ovog diplomskog rada je bioinformatičke prirode. Bioinformatika je znanost koja se bavi analizom bioloških podataka o sljedovima, sadržaju i organizaciji genoma te predviđa 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. Popularnost koju su stekli skriveni Markovljevi modeli unatrag nekoliko desetljeća može se pripisati objavi radova *L. E. Bauma* i ostalih sedamdesetih godina 20. stoljeća u kojima su konstruirane efikasne metode za računanje s takvim modelima. Broj područja u kojima te metode i modeli nailaze 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.

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

U prvom poglavlju je dan pregled osnovnih pojmova iz teorije vjerojatnosti i Markovljevih lanaca. U drugom poglavlju formalno definiramo skrivene Markovljeve modele i dajemo primjer, dok u trećem opisujemo algoritme koje smo koristili. U zadnjem, četvrtom poglavlju, opisujemo postupak za procjenu parametara skrivenih Markovljevih modela i njegovu implementaciju te promatramo metode za procjenu kompleksnosti modela. Prezentiramo rezultate i zaključke.

Poglavlje 1

Osnovni pojmovi

1.1 Vjerojatnost

Definicija 1.1.1. ***Slučajni pokus*** *ili* ***slučajni eksperiment*** *je takav pokus čiji ishodi, tj. rezultati nisu jednoznačno određeni uvjetima u kojima izvodimo pokus.*

Tipični primjeri slučajnog pokusa su bacanje igraće kocke i bacanje novčića.

Definicija 1.1.2. ***Prostor elementarnih događaja*** *je neprazan skup Ω koji reprezentira skup svih ishoda slučajnog pokusa. Elemente od Ω označavamo s ω i zovemo* ***elementarni događaji****.*

Definicija 1.1.3. *Neprazna familija $\mathcal{F}$ podskupova od Ω zove se* ***σ-algebra*** *ako vrijedi:*

1. $\emptyset \in \mathcal{F}$
2. *Ako je* $A \in \mathcal{F}$, *tada je* $A^c = \Omega \backslash A \in \mathcal{F}$
3. *Ako je* $A_n \in \mathcal{F}$, $n \in \mathbb{N}$, *tada je* $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$

Drugim riječima, ***σ-algebra*** *je neprazna familija podskupova od Ω koja sadrži nemoguć događaj $\emptyset$, zatvorena je na komplementiranje i prebrojivu uniju.*

Definicija 1.1.4. *Neka je $\mathcal{F}$ σ-algebra na skupu Ω. Uređeni par $(\Omega, \mathcal{F})$ zove se* ***izmjeriv prostor****.*

Sada konačno možemo dati definiciju vjerojatnosti.

Definicija 1.1.5. *Neka je $(\Omega, \mathcal{F})$ izmjeriv prostor. Funkcija $\mathbb{P}\colon \mathcal{F} \to \mathbb{R}$ jest* ***vjerojatnost*** *ako vrijedi:*

1. $\mathbb{P}(\Omega) = 1$ *(normiranost)*
2. $\mathbb{P}(A) \geq 0 \ \ \forall A \in \mathcal{F}$ *(nenegativnost)*
3. *Za svaki niz* $(A_n,\ n \in \mathbb{N})$, $A_n \in \mathcal{F}$, *takav da je* $A_m \bigcap A_n = \emptyset$ *za* $m \neq n$, *vrijedi*

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n) \text{ } (\sigma \text{ -aditivnost ili prebrojiva aditivnost})$$

Definicija 1.1.6. *Uređena trojka* $(\Omega, \mathcal{F}, \mathbb{P})$, *gdje je* $\mathcal{F}$ σ*-algebra na* Ω *i* $\mathbb{P}$ *vjerojatnost na* $\mathcal{F}$, *zove se* ***vjerojatnosni prostor****. Ako je* Ω *konačan ili prebrojiv skup* $(\Omega, \mathcal{F}, \mathbb{P})$ *zovemo* ***diskretni vjerojatnosni prostor****.*

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

Definicija 1.1.7. *Neka je* $(\Omega, \mathcal{F}, \mathbb{P})$ *proizvoljan vjerojatnosni prostor i* $A \in \mathcal{F}$ *takav da je* $\mathbb{P}(A) > 0$. *Definirajmo funkciju* $\mathbb{P}_A \colon \mathcal{F} \to [0, 1]$*:*

$$\mathbb{P}_A(B) = \mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \bigcap B)}{\mathbb{P}(A)},\ B \in \mathcal{F}. \tag{1.1}$$

$\mathbb{P}_A$ *je vjerojatnost na* $\mathcal{F}$ *i zovemo je* ***uvjetna vjerojatnost uz uvjet A****. Broj* $\mathbb{P}(B \mid A)$ *zovemo* ***vjerojatnost od B uz uvjet A****.*

Definicija 1.1.8. *Konačna ili prebrojiva familija* $(H_i, i = 1, 2, \ldots)$ *događaja u vjerojatnosnom prostoru* $(\Omega, \mathcal{F}, \mathbb{P})$ *jest* ***potpun sistem događaja*** *ako je* $H_i \neq \emptyset$ *za svako* i, $H_i \bigcap H_j = \emptyset$ *za* $i \neq j$ *(tj. događaji se uzajamno isključuju) i* $\bigcup_i H_i = \Omega$.

Drugim riječima, potpun sistem događaja konačna je ili prebrojiva particija skupa Ω *s tim da su elementi particije događaji.*

Teorem 1.1.9. *(Formula potpune vjerojatnosti) Neka je* $(H_i, i = 1, 2, \ldots)$ *potpun sistem događaja u vjerojatnosnom prostoru* $(\Omega, \mathcal{F}, \mathbb{P})$. *Tada za proizvoljno* $A \in \mathcal{F}$ *vrijedi*

$$\mathbb{P}(A) = \sum_i \mathbb{P}(H_i)\mathbb{P}(A \mid H_i). \tag{1.2}$$

Teorem 1.1.10. *(Bayesova formula) Neka je* $(H_i, i = 1, 2, \ldots)$ *potpun sistem događaja u vjerojatnosnom prostoru* $(\Omega, \mathcal{F}, \mathbb{P})$ *i* $A \in \mathcal{F}$ *takav da je* $\mathbb{P}(A) > 0$. *Tada za svako* i *vrijedi:*

$$\mathbb{P}(H_i \mid A) = \frac{\mathbb{P}(H_i)\mathbb{P}(A \mid H_i)}{\sum_j \mathbb{P}(H_j)\mathbb{P}(A \mid H_j)}. \tag{1.3}$$

Definicija 1.1.11. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Ako je $(\Omega, \mathcal{F}, \mathbb{P})$ diskretan vjerojatnosni prostor,* ***slučajna varijabla*** *jest prozvoljna realna funkcija na Ω. Ako je $(\Omega, \mathcal{F}, \mathbb{P})$ opći vjerojatnosni prostor,* ***slučajna varijabla*** *na Ω jest funkcija $X\colon \Omega \to \mathbb{R}$ ako je $X^{-1}(B) \in \mathcal{F}$ za prozvoljno $B \in \mathcal{B}$, tj. $X^{-1}(\mathcal{B}) \subset \mathcal{F}$, gdje je $\mathcal{B}$ σ-algebra generirana familijom svih otvorenih skupova na $\mathbb{R}$.*

1.2 Markovljevi lanci

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

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

$$\begin{aligned} &\mathbb{P}(X_t = i \mid X_{t-1} = i_{t-1}, X_{t-2} = i_{t-2}, \ldots, X_0 = i_0) = \\ &= \mathbb{P}(X_t = i \mid X_{t-1} = i_{t-1}, X_{t-2} = i_{t-2}, \ldots, X_{t-n} = i_{t-n}) \end{aligned} \tag{1.4}$$

za svaki $t \geq 0$ i za sve $i_0, i_1, \ldots, i_{t-1}, i \in S$ za koje su obje uvjetne vjerojatnosti definirane.

Svojstvo u relaciji (1.4) naziva se ***Markovljevim svojstvom***.

Definicija 1.2.3. *Vjerojatnost prijelaza iz stanja* i *u stanje* j *u trenutku* t *jest vjerojatnost da slučajna varijabla X_{t+1} poprimi vrijednost* j *ako je slučajna varijabla X_t poprimila vrijednost* i*. Tu vrijednost nazivamo* ***prijelazna (tranzicijska) vjerojatnost*** *i označavamo:*

$$a_{ij} = \mathbb{P}(X_{t+1} = j \mid X_t = i) \tag{1.5}$$

Markovljev lanac zajedno sa zadanim prijelaznim vjerojatnostima nazivamo **Markovljev model**.

Poglavlje 2

Skriveni Markovljev model

2.1 Skriveni Markovljev model (HMM)

Kod "običnog" Markovljevog modela ***niz stanja*** koji emitira neki niz opažanja nam je uvijek poznat.

Skriveni Markovljev model je Markovljev model kod kojeg su stanja "skrivena", tj. ne znamo ih pri emitiranju nekog niza vrijednosti, ali nam je taj niz vrijednosti poznat i pomoću njega možemo nešto zaključniti o nizu stanja koji odgovara emitiranom nizu vrijednosti.

Niz stanja skrivenog Markovljevog modela modeliran je Markovljevim lancem 1. reda, tj. vjerojatnost da se nalazimo u nekom stanju ovisi samo o prethodnom stanju.

Formalno:

Definicija 2.1.1. ***Skriveni Markovljev model prvog reda*** *(eng. hidden Markov model ili kraće HMM) zadajemo sa dva niza, Q i O:*

- $Q = Q_1,Q_2,\ldots,Q_N$ *- niz slučajnih varijabli koje poprimaju diskretne vrijednosti*
- $O = O_1,O_2,\ldots,O_N$ *- niz slučajnih varijabli koje poprimaju diskretne ili kontinuirane vrijednosti,*

gdje te varijable zadovoljavaju sljedeće uvjete:

1.

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

2.

$$\mathbb{P}(O_t \mid Q_N,O_N,Q_{N-1},O_{N-1},\ldots,Q_{t+1},O_{t+1},Q_t,Q_{t-1},O_{t-1},\ldots,Q_1,O_1)=\mathbb{P}(O_t \mid Q_t) \tag{2.2}$$

Skriveni Markovljev model (HMM) sa diskretnim opažanjima možemo opisno zadati sljedećim parametrima:

- N - broj stanja u kojima se proces može nalaziti

$$\mathcal{S} = \{1, \ldots, N\} \tag{2.3}$$

  $\mathcal{S}$ - skup svih stanja procesa

- M - broj mogućih opažanja

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

  B - skup svih opažanja

- L - duljina opaženog niza

$$x = (x_1, \ldots, x_L) \tag{2.5}$$

  x - opaženi niz

- A - matrica tranzicijskih vjerojatnosti

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

- E - matrica emisijskih vjerojatnosti

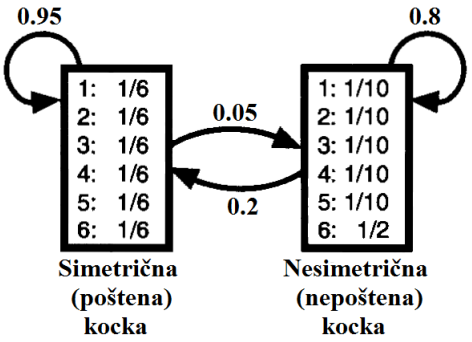
$$E = \{e_j(k)\},\ e_j(k) = \mathbb{P}(O_t = b_k \mid Q_t = j), 1 \leq j \leq N, 1 \leq k \leq M \tag{2.7}$$

Upravo je upotreba skrivenog Markovljevog modela u radu s ljudskim genomom u članku "Discovery and characterization of chromatin states for systematic annotation of the human genome" autora Jasona Ernsta i Manolisa Kellisa [3] bila motivacija za ovaj diplomski rad.

2.2 Primjer HMM-a: povremeno nepoštena kockarnica

U nekoj kockarnici koriste se dvije kocke: simetrična igrača kocka ("poštena" kocka) kod koje je vjerojatnost da padne bilo koji od brojeva iz skupa {1,2,3,4,5,6} jednaka $\frac{1}{6}$ i nesimetrična ("nepoštena" kocka) kod koje je vjerojatnost da padne šestica $\frac{1}{2}$, dok je vjerojatnost da padne bilo koji drugi broj iz skupa {1,2,3,4,5} jednaka $\frac{1}{10}$.

Ako je bačena simetrična kocka vjerojatnost da će ponovno biti bačena simetrična je 95% dok je vjerojatnost da će biti zamijenjena nesimetričnom 5%. U 80% slučajeva



Slika 2.1: Dijagram povremeno nepoštene kockarnice

kockarnica će, nakon što je zamijenila simetričnu kocku nesimetričnom i izvela bacanje njome, nastaviti izvoditi bacanja nesimetričnom kockom, dok će u 20% slučajeva zamijeniti nesimetičnu kocku simetričnom.

U prethodno iznesenoj notaciji za HMM dani primjer zapisujemo na sljedeći način:

- $N = 2$

$$\mathcal{S} = \{S\ ,\ N\}$$

  S - simetrična kocka, N - nesimetrična kocka

- M=6

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

- Matrica tranzicijskih vjerojatnosti:

$$tranz2 = \begin{pmatrix} 0.95 & 0.05 \\ 0.2 & 0.8 \end{pmatrix}$$

  gdje je $a_{11} = \mathbb{P}(S \mid S)$ - vjerojatnost da je nakon bacanja simetrične kocke opet bacana simetrična kocka, $a_{12} = \mathbb{P}(N \mid S)$ - vjerojatnost da je nakon bacanja simetrične kocke bacana nesimetrična kocka, $a_{21} = \mathbb{P}(S \mid N)$ - vjerojatnost da je nakon bacanja nesimetrične kocke bacana simetrična kocka i $a_{22} = \mathbb{P}(N \mid N)$ - vjerojatnost da je nakon bacanja nesimetrične kocke opet bacana nesimetrična kocka.

- Matrica emisijskih vjerojatnosti:

$$emp2 = \begin{pmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \\ \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{2} \end{pmatrix}$$

  gdje se u prvom retku matrice nalaze emisijske vjerojatnosti elemenata iz B (tim redosljedom) u stanju S, a u drugom retku su emisijske vjerojatnosti elemenata iz B u stanju N.

Primijetimo da je proces koji modelira baca li se simetrična ili nesimetrična kocka zapravo Markovljev proces prvog reda sa stanjima u $\mathcal{S}$. Vidimo da su emisijske vjerojatnosti simbola iz B u svakom od stanja različite i ne ovise o prijašnjim stanjima. Sada je jasno da je primjer povremeno nepoštene kockarnice primjer *skrivenog Markovljevog modela prvog reda.*

Postavlja se pitanje: *Što je tu "skriveno" ?*

Ako nam je poznat niz opažanja, tj. niz dobiven bacanjem kocaka, npr. $X =$ (1 3 5 6 6 6) ne znamo pri kojem bacanju je korištena simetrična a pri kojem nesimetrična kocka, to je poznato samo kockarnici. Dakle, *niz stanja* je nepoznat, skriven. Međutim, iako nam je niz stanja nepoznat, pomoću niza opažanja moguće je odrediti sljedeće:

- *Najvjerojatniji niz stanja za dobiveni niz opažanja.* U svrhu rješavanja ovog problema koristi se **Viterbijev algoritam** o kojem će biti riječi u daljnjem tekstu rada.

- *Vjerojatnost niza opažanja u odnosu na neki zadani model.* Ovaj problem rješava se **forward** ili **backward algoritmom** koji ćemo kasnije pojasniti.

- *Parametri modela: emisijske i tranzicijske vjerojatnosti.* Za procjenu ovih parametara koristimo **Baum-Welch algoritam** koji također kasnije objašnjavamo.

Da bi mogli primijeniti HMM na ljudski genom, Ernst i Kellis (2010.) [3] ga najprije dijele na disjunktne intervale od po 200 nukleotida. U svakom od tih intervala gledaju nalazi li se pojedini znak od njih 41 ukupno, ako se nalazi to zabilježe s '1', ako se ne nalazi s '0'. Na taj način se dobije specifična kombinacija nula i jedinica duljine 41 za svaki od intervala i pomoću te varijable izračuna se vjerodostojnost za dane parametre (broj stanja, tranzicijske i emisijske vjerojatnosti).

Napomena 2.2.1. *Ernst i Kellis rade s multivarijatnim HMM-om pa im vjerodostojnost ima drugačiji oblik nego kod nas.*

Poglavlje 3

Algoritmi za HMM

U ovom poglavlju objašnjeni su algoritmi koje smo spomenuli u prošlom poglavlju. Ponovimo, iako je niz stanja u skrivenom Markovljevom modelu nepoznat, poznat nam je niz emitiranih vrijednosti i pomoću njega možemo nešto zaključniti o nizu stanja.

Označimo sa $x = (x_1,x_2,\ldots,x_n)$ niz emitiranih simbola i sa $\pi = (\pi_1,\pi_2,\ldots,\pi_n)$ pripadajući niz skrivenih stanja. Koristimo već spomenute oznake za tranzicijske vjerojatnosti (a_{kl} - vjerojatnost prelaska iz stanja π_k u stanje π_l) i emisijske vjerojatnosti ($e_k(b)$ - vjerojatnost emisije simbola b u stanju k).

3.1 Viterbijev algoritam

Viterbijev algoritam je algoritam dinamičkog programiranja za pronalaženje najvjerojatnijeg niza stanja π^* u skrivenom Markovljevom modelu koji emitira zadani niz simbola x. Niz stanja π^* dobivenih Viterbijevim algoritmom nazivamo *Viterbijev put* ili *Viterbijev prolaz.*

Dakle, tražimo

$$\pi^* = \arg\max_{\pi} \mathbb{P}(x,\pi) = \arg\max_{\pi} \mathbb{P}(\pi \mid x). \tag{3.1}$$

Pri tome $\mathbb{P}(x,\pi)$ definiramo kao:

$$\mathbb{P}(x,\pi) = a_{0\pi_1} \prod_{i=1}^{n} e_{\pi_i}(x_i)\, a_{\pi_i\pi_{i+1}} \tag{3.2}$$

gdje $a_{0\pi_1}$ i $a_{\pi_n\pi_{n+1}}$ koristimo za modeliranje početka i kraja te stavljamo $a_{0\pi_1} = a_{\pi_n\pi_{n+1}} = 1$.

Označimo sa $v_k(i)$ vjerojatnost najvjerojatnijeg prolaza π^* koji završava u stanju k pri čemu su emitirani simboli $x_1, \ldots, x_i$. Sada Viterbijev algoritam možemo zapisati u sljedeća četiri koraka:

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

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

2. **Rekurzija** $(i = 1, \ldots, n)$:

$$v_l(i) = e_l(x_i) \cdot \max_k(v_k(i-1)\ a_{kl})$$

$$ptr_i(l) = \arg\max_k(v_k(i-1)\ a_{kl})$$

3. **Završetak**:

$$\mathbb{P}(x, \pi^*) = \max_k(v_k(n)\ a_{k0})$$

$$\pi_n^* = \arg\max_k(v_k(n)\ a_{k0})$$

4. **Najvjerojatniji put** $(i = n \ldots 1)$:

$$\pi_{i-1}^* = ptr_i(\pi_i^*)$$

Napomena 3.1.1. *U praksi se računanje u Viterbijevom algoritmu (i ostalim algoritmima koje navodimo u nastavku) provodi s logaritmiranim vrijednostima jer množenje malih brojeva uzrokuje "underflow" na računalu.*

3.2 Forward algoritam

Forward algoritam računa ukupnu vjerojatnost niza u odnosu na model, tj. sumu po svim putevima vjerojatnosti niza po nekom putu kroz model. Označimo tu vjerojatnost sa $\mathbb{P}(x)$.

Ako su $x_1, \ldots, x_i$ opaženi simboli, π_i stanje modela koje emitira simbol na poziciji i u nizu x, e_l emisijska vjerojatnost u stanju l, a_{kl} tranzicijska vjerojatnost iz stanja k u stanje l, onda vjerojatnost da niz $x_1, \ldots, x_i$ završava u stanju k (u oznaci $f_k(i)$)

možemo raspisati rekurzivno na sljedeći način:

$$
\begin{aligned}
f_k(i) &= \mathbb{P}(x_1,\ldots,x_i \mid \pi_i = k) && (3.3)\\
&= \sum_{\pi_1,\ldots,\pi_{i-1}} \mathbb{P}(x_1,\ldots,x_i,\pi_1,\ldots,\pi_{i-1} \mid \pi_i = k) && (3.4)\\
&= \sum_{l} \sum_{\pi_1,\ldots,\pi_{i-2}} \mathbb{P}(x_1,\ldots,x_i,\pi_1,\ldots,\pi_{i-2},\pi_{i-1} = l \mid \pi_i = k) && (3.5)\\
&= e_k(x_i) \sum_{l} a_{lk} \sum_{\pi_1,\ldots,\pi_{i-2}} \mathbb{P}(x_1,\ldots,x_{i-1},\pi_1,\ldots,\pi_{i-2} \mid \pi_{i-1} = l) && (3.6)\\
&= e_k(x_i) \sum_{l} a_{lk} \mathbb{P}(x_1,\ldots,x_{i-1} \mid \pi_{i-1} = l) && (3.7)\\
&= e_k(x_i) \sum_{l} f_l(i-1) a_{lk} && (3.8)
\end{aligned}
$$

Sada forward algoritam možemo zapisati u tri koraka:

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

$$f_0(0) = 1,\ f_k(0) = 0\ \forall k > 0$$

2. **Rekurzija** $(i = 1,\ldots,n)$:

$$f_l(i) = e_l(x_i) \sum_k f_k(i-1) a_{kl}$$

3. **Završetak**:

$$\mathbb{P}(x) = \sum_k f_k(n) a_{k0}$$

3.3 Backward algoritam

Backward algoritam također računa ukupnu vjerojatnost niza u odnosu na model, jedina je razlika između forward i backward algoritama što backward algoritam tu vjerojatnost računa od kraja, i niza i modela. Konačni rezultati oba algoritma su jednaki.

Vjerojatnost niza $x_{i+1}, \ldots, x_n$ uz uvjet da je u trenutku i niz u stanju k (u oznaci $b_k(i)$) raspisujemo rekurzivno na sljdeći način:

$$\begin{aligned} b_k(i) =& \mathbb{P}(x_{i+1}, \ldots, x_n \mid \pi_i = k) && (3.9)\\ =& \sum_{\pi_{i+1},\ldots,\pi_n} \mathbb{P}(x_{i+1}, x_{i+2} \ldots, x_n, \pi_{i+1}, \ldots, \pi_n \mid \pi_i = k) && (3.10)\\ =& \sum_l \sum_{\pi_{i+2},\ldots,\pi_n} \mathbb{P}(x_{i+1}, x_{i+2} \ldots, x_n, \pi_{i+1} = l, \pi_{i+2}, \ldots, \pi_n \mid \pi_i = k) && (3.11)\\ =& \sum_l e_l(x_{i+1}) a_{kl} \sum_{\pi_{i+2},\ldots,\pi_n} \mathbb{P}(x_{i+2} \ldots, x_n, \pi_{i+2}, \ldots, \pi_n \mid \pi_{i+1} = l) && (3.12)\\ =& \sum_l e_l(x_{i+1}) a_{kl} \mathbb{P}(x_{i+2}, \ldots, x_n \mid \pi_{i+1} = l) && (3.13)\\ =& \sum_l e_l(x_{i+1}) a_{kl} b_l(i+1) && (3.14) \end{aligned}$$

Sada backward algoritam također možemo zapisati u tri koraka:

1. **Inicijalizacija** $(i = n)$:
$$b_k(n) = a_{k0}, \forall k$$
2. **Rekurzija** $(i = n-1, \ldots, 1)$:
$$b_k(i) = \sum_l a_{kl} e_l(x_{i+1}) b_l(i+1)$$
3. **Završetak**:
$$\mathbb{P}(x) = \sum_l a_{0l} e_l(x_1) b_l(1)$$

3.4 Baum-Welch algoritam

Ovaj algoritam je iterativni postupak za određivanje parametara modela na temelju niza opažanja. Parametri se procjenjuju tako da se maksimizira očekivanje promatranog niza s obzirom na odabir parametara.

Prisjetimo se primjera povremeno nepoštene kockarnice iz prethodnog poglavlja (2.2). Kako bismo mogli odrediti parametre tog modela? U slučaju da nam je put π poznat, mogli bismo jednostavno pebrojiti sve tranzicije i emisije koje se pojavljuju u nizu, te pomoću njih dobiti procjenitelje parametara modela

$$a_{ij} = \frac{A_{ij}}{\sum_k A_{ik}} \tag{3.15}$$

$$e_i(s) = \frac{E_i(s)}{\sum_{s'} E_i(s')} \tag{3.16}$$

gdje je A_{ij} broj tranzicija iz stanja i u stanje j, a $E_i(s)$ broj emisija simbola s iz stanja i. Budući da nam je put π najčešće nepoznat, ne možemo jednostavno prebrojiti tranzicije i emisije kako bismo dobili A_{ij} i $E_i(s)$, već računamo njihove očekivane vrijednosti sljedećim formulama

$$A_{ij} = \frac{1}{\mathbb{P}(x)} \sum_t f_i(t) a_{ij} e_j(x_{t+1}) b_j(t+1) \tag{3.17}$$

$$E_i(s) = \frac{1}{\mathbb{P}(x)} \sum_{\{t|x_t=s\}} f_i(t) b_i(t) \tag{3.18}$$

gdje je $f_i(t)$ forward varijabla definirana u (3.3) a $b_i(t)$ odgovarajuća backward varijabla definirana u (3.9).

Primijetimo da očekivane vrijednosti tranzicija i emisija ovise i o parametrima samog modela a_{ij} i $e_i(k)$ koje želimo procijeniti. Može se pokazati da iterativnim postupkom optimizacije seta parametara θ formulama (3.15) i (3.16), koristeći procjene broja tranzicija (3.17) i emisija (3.18) povećavamo vjerodostojnost modela, tj. za nove vrijednosti seta parametara θ' vrijedi

$$\mathbb{P}(x \mid \theta') \geq \mathbb{P}(x \mid \theta).$$

Ovakav postupak optimizacije parametara za skrivene Markovljeve modele zovemo Baum-Welch algoritam. Možemo ga zapisati u tri koraka:

1. **Inicijalizacija**: Proizvoljno odaberi parametre.

2. **Rekurzija**:
   a) Svim elementima matrica A i E dodijeli vrijednost 0.
   b) Izračunaj $f_k(i)$ pomoću forward algoritma.
   c) Izračunaj $b_k(i)$ pomoću backward algoritma.
   d) Izračunaj elemente matrice A po formuli

$$A_{kl} = \frac{1}{\mathbb{P}(x)} \sum_t f_k(t) a_{kl} e_l(x_{t+1}) b_l(t+1).$$

   Izračunaj elemente matrice E po formuli

$$E_k(s) = \frac{1}{\mathbb{P}(x)} \sum_{\{t|x_t=s\}} f_k(t) b_k(t).$$

e) Izračunaj nove parametre modela koristeći formule

$$a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}}$$

,

$$e_k(s) = \frac{E_k(s)}{\sum_{s'} E_k(s')}.$$

f) Izračunaj $L = \mathbb{P}(x \mid \theta)$.

3. **Završetak**: Stani ako je dosegnut maksimalan broj iteracija ili je $\Delta L < \epsilon$.

Ovaj postupak garantira konvergenciju algoritma u lokalni maksimum. Uobičajeno je da sustav ima mnogo lokalnih ekstrema, te konvergencija u neki od njih bitno ovisi o inicijalnom izboru parametara modela. Baum-Welch algoritam je specijalan slučaj općenitog postupka za procjenu parametara vjerojatnosnih modela metodom maksimalne vjerodostojnosti - EM (*Expectation Maximization*) algoritma.

Poglavlje 4

Rezultati

4.1 Simulacija i optimizacija

U radu je korišten programski jezik Python.

Simulirali smo 2 niza duljine 20 000, prvi niz je bio simulacija bacanja 2 kocke, simetrične i nesimetrične s vjerojatnošću pada šestice $\frac{1}{2}$, kao što je opisano u drugom poglavlju (2.2 Primjer HMM-a: povremeno nepoštena kockarnica). Drugi niz je simulacija bacanja 3 kocke, prve dvije kao kod nepoštene kockarnice a treća je bila kocka s "težom" jedinicom, tj. vjerojatnost da padne 1 na kocki je $\frac{1}{2}$, a neki od brojeva iz skupa {2,3,4,5,6} je $\frac{1}{10}$. Dakle, drugi model ima 3 stanja

$$\mathcal{S} = \{S\ ,\ N_6\ ,\ N_1\}$$

S - simetrična kocka, N_6 - nesimetrična kocka s "težom" šesticom, N_1 - nesimetrična kocka s "težom" jedinicom. Matrica emisijskih vjerojatnosti izgleda ovako:

$$emp3 = \begin{pmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} \end{pmatrix}$$

Tranzicijska matrica i u ovom slučaju je dijagonalno dominantna, oblika:

$$tranz3 = \begin{pmatrix} 0.95 & 0.025 & 0.025 \\ 0.1 & 0.8 & 0.1 \\ 0.05 & 0.05 & 0.9 \end{pmatrix}$$

Dakle, zadali smo 2 skrivena Markovljeva modela sa različitim brojem stanja i različitim tranzicijskih i emisijskim vjerojatnostima. Za svaki od modela simulirali smo nizove te pokušali rekonstruirati parametre modela Baum-Welch algoritmom. U tu svrhu smo za svaki od simuliranih nizova optimizirali parametre s 2, 3, 4 i 5 stanja (kocki) i raznim inicijalnim setovima tranzicijskih i emisijskih parametara.

Kao što smo spomenuli ranije, očekuje se da s povećanjem broja iteracija u Baum-Welch-u raste i vjerodostojnost modela no u praksi je stanje drugačije jer postoji više lokalnih maksimuma. U koji od njih će proces konvergirati ovisi o izboru inicijalnih parametara. Zato smo implementirali algoritam tako da smo iteriranje zaustavili kada je vjerodostojnost prvi puta počela padati ili je razlika u vjerodostojnosti između dvije uzastopne iteracije postala manja od $\epsilon = 0.01$. Maksimalan broj iteracija postavili smo na 400.

Viterbijevo treniranje

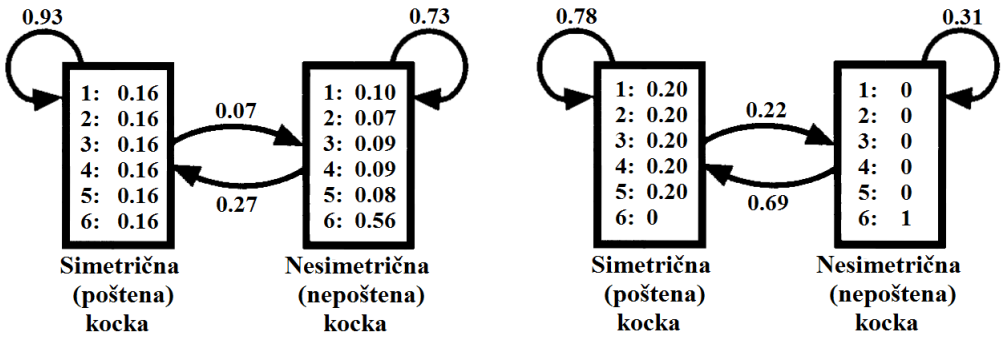
Kao alternativni način procjene parametara modela koristili smo i **Viterbijevo treniranje**. Umjesto maksimizacije vjerodostojnosti opaženih podataka $x = (x_1, x_2, \ldots, x_n)$ uz dani set parametara θ, tj.

$$\mathbb{P}(x_1, x_2, \ldots, x_n \mid \theta), \tag{4.1}$$

Viterbijevo treniranje pronalazi set parametara θ koji maksimizira vjerodostojnost najvjerojatnijeg niza skrivenih stanja

$$\mathbb{P}(x_1, x_2, \ldots, x_n, \pi^*(x_1), \pi^*(x_2), \ldots, \pi^*(x_n) \mid \theta). \tag{4.2}$$

U tom pristupu najvjerojatniji putevi za opaženi niz se određuju pomoću Viterbijevog algoritma koji smo opisali u prethodnom poglavlju. I ovdje se postupak iterira nakon što se odrede novi parametri modela. Napravili smo 30 iteracija za niz duljine 20 000 i za svaku kombinaciju inicijalnih parametara kojima smo optimizirali za svaki od modela dobili smo lošiju procjenu nego kod Baum-Welcha. Čest rezultat je bio da kod Viterbija neka stanja uopće ne emitiraju simbole ili da se neka kocka "vrti sama u sebi" tj. nema tranzicija u drugu (druge) kocke. Pogledajmo na sljedećem primjeru usporedbu procjena parametara Baum-Welchom i Viterbijem.



Slika 4.1: Procjena parametara modela sa 2 stanja Baum-Welch algoritmom (lijevi dijagram) i Viterbijevim treniranjem (desni dijagram), sa inicijalnim parametrima emp201 i tranz203 (konkretne vrijednosti parametara u Dodatku)

Budući da su procjene parametara dobivene Baum-Welch algoritmom bile bliže stvarnim vrijednostima iz kojih smo simulirali model nego procjene dobivene Viterbijevim treniranjem, odlučili smo se nadalje koncentrirati na procjenu parametara Baum-Welchom.

Numerička stabilnost

Kod implementacije skrivenih Markovljevih modela i pripadajućih algoritama "underflow" predstavlja značajan problem. Forward i backward algoritmi računaju vjerojatnosti parcijalnih nizova, a te vjerojatnosti mogu biti vrlo male za velike duljine niza. Postoje dva načina za izbjegavanje "underflow" grešaka:

1. **Skaliranje** - najčešće primjenjivano rješenje je korištenje skalirajućih koeficijenata za vjerojatnosti tako da vrijednosti ostanu u rasponu koji je prikaziv na računalu

2. **Log- space** - umjesto da računamo vjerojatnosti radimo sa logaritmiranim vjerojatnostima. Na taj način produkti iz korištenih algoritama postaju sume i tako brojevi ostaju u rasponu od 10^{-308} do 10^{308} tj. ostaju "dovoljno veliki" da ih računalo može memorirati. Za logaritam zbroja brojeva $a, b \in \mathbb{R}$ vrijedi sljedeća jednakost

$$log(a+b) = log(a) + log(1 + e^{log(b)-log(a)}), a > b \tag{4.3}$$

gdje je $log = log_e$.

Budući da je prirodni logaritam strogo rastuća funkcija i uz uvjet $a > b$, $log(b) - log(a)$ je uvijek manje od nule odnosno $e^{log(b)-log(a)}$ je iz intervala $< 0, 1 >$. To osigurava da je drugi sumand s desne strane jednakosti (4.3) nenegativan pa je promatrani logaritam moguće izračunati i za brojeve izvan spomenutog raspona prikazivih brojeva na računalu.

U ovom radu koristi se *log - space*.

4.2 Informacijski kriteriji: AIC i BIC

Vrijednosti tranzicijskih (*tranz*) i emisijskih (*emp*) matrica kojima smo optimizirali parametre modela mogu se naći u sljedećoj cjelini tj. Dodatku.

Param. za simulaciju	**Baum-Welch L**	**AIC**	**BIC**
emp2, tranz2	-35388.4984350	70800.9968701	70895.8387207
Param. za optimizaciju	**Baum-Welc L**	**AIC**	**BIC**
emp201, tranz203	-35390.9470592	70805.8941184	70900.7359690
emp302, tranz309	-35394.6666406	70831.3332813	70997.3065199
emp402, tranz405	-35387.9944289	70839.9888578	71092.9004595
emp503, tranz502	-35437.5086267	70965.0172533	71320.6741932

Tablica 4.1: Najbolja kombinacija parametara za optimizaciju s 2, 3, 4 i 5 stanja za niz simuliran iz 2 stanja

Param. za simulaciju	**Baum-Welc L**	**AIC**	**BIC**
emp3, tranz3	-34814.5681421	69671.1362842	69837.1095228
Param. za optimizaciju	**Baum-Welc L**	**AIC**	**BIC**
emp202, tranz208	-34995.0247212	70014.0494425	70108.8912931
emp302, tranz309	-34819.544737	69681.089474	69847.0627126
emp402, tranz404	-34626.9349903	69317.8699807	69570.7815823
emp503, tranz501	-34793.813134	69677.626268	70033.2832079

Tablica 4.2: Najbolja kombinacija parametara za optimizaciju s 2, 3, 4 i 5 stanja za niz simuliran iz 3 stanja

Izuzev optimizacija s 3 i 5 stanja u tablici 4.1 i optimizacije s 5 stanja u tablici 4.2, vjerodostojnost L raste s povećanjem broja parametara. Međutim, nas je zanimao penal za broj parametara u modelu pa smo u tu svrhu računali informacijske kriterije

koji se baziraju na vjerodostojnosti: **AIC** (*Akaike information criterion*) i **BIC** (*Bayesian information criterion*). Informacijski kriteriji su kriteriji za izbor modela, tj. oni mjere koliko "dobro" model opisuje podatke. AIC i BIC su dani sljedećim jednadžbama

$$AIC = -2log(L) + 2k \tag{4.4}$$

$$BIC = -2log(L) + klog(n) \tag{4.5}$$

gdje je L maksimalna vjerodostojnost modela, n duljina niza, a k broj slobodnih parametara modela.

Informacijski kriteriji se općenito minimiziraju pa su nam i ovdje od interesa modeli s najmanjim vrijednostima AIC-a i BIC-a. Općenito, BIC penalizira slobodne parametre jače, rigoroznije nego AIC, odnosno penal je veći u BIC-u nego u AIC-u. To je vidljivo iz priloženih rezultata. Iz tablica 4.1 i 4.2 vidimo da i AIC i BIC upućuju na isti rezultat: model s 2 stanja najbolje opisuje simulaciju iz 2 stanja, a model s 4 stanja najbolje opisuje niz simuliran iz 3 stanja. Dakle, niti informacijski kriteriji nam ne daju zadovoljavajuće rezultate.

Možemo zaključiti da niti jedan od načina za procjenu parametara koji smo proučavali nije dao zadovoljavajuće rezultate, Viterbijevo treniranje zbog loše procjenjenih parametara a Baum-Welch algoritam zbog problema s lokalnim ekstremima i inicijanim parametrima. Informacijski kriteriji AIC i BIC su se također pokazali nekorisnima jer se ti kriteriji koriste uz pretpostavku da je vjerodostojnost u Baum-Welch algoritmu dobro izračunata a to ne vrijedi.

4.3 Dodatak: popis parametara za optimizaciju

$$emp201 = \begin{pmatrix} 0.166 & 0.166 & 0.166 & 0.166 & 0.166 & 0.166 \\ 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.75 \end{pmatrix}$$

$$tranz203 = \begin{pmatrix} 0.95 & 0.05 \\ 0.1 & 0.9 \end{pmatrix}$$

$$emp302 = \begin{pmatrix} 0.166 & 0.166 & 0.166 & 0.166 & 0.166 & 0.166 \\ 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.75 \\ 0.5 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \end{pmatrix}$$

$$tranz309 = \begin{pmatrix} 0.95 & 0.025 & 0.025 \\ 0.05 & 0.9 & 0.05 \\ 0.05 & 0.05 & 0.9 \end{pmatrix}$$

$$emp402 = \begin{pmatrix} 0.166 & 0.166 & 0.166 & 0.166 & 0.166 & 0.166 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.5 \\ 0.5 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.166 & 0.166 & 0.166 & 0.166 & 0.166 & 0.166 \end{pmatrix}$$

$$tranz405 = \begin{pmatrix} 0.8 & 0.066 & 0.066 & 0.066 \\ 0.033 & 0.9 & 0.033 & 0.033 \\ 0.8 & 0.066 & 0.066 & 0.066 \\ 0.066 & 0.066 & 0.066 & 0.8 \end{pmatrix}$$

$$emp503 = \begin{pmatrix} 0.166 & 0.166 & 0.166 & 0.166 & 0.166 & 0.166 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.5 \\ 0.5 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.166 & 0.166 & 0.166 & 0.166 & 0.166 & 0.166 \\ 0.166 & 0.166 & 0.166 & 0.166 & 0.166 & 0.166 \end{pmatrix}$$

$$tranz502 = \begin{pmatrix} 0.8 & 0.05 & 0.05 & 0.05 & 0.05 \\ 0.025 & 0.9 & 0.025 & 0.025 & 0.025 \\ 0.8 & 0.05 & 0.05 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.05 & 0.8 & 0.05 \\ 0.05 & 0.05 & 0.05 & 0.05 & 0.8 \end{pmatrix}$$

$$emp202 = \begin{pmatrix} 0.166 & 0.166 & 0.166 & 0.166 & 0.166 & 0.166 \\ 0.03 & 0.03 & 0.03 & 0.03 & 0.03 & 0.85 \end{pmatrix}$$

$$tranz208 = \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{pmatrix}$$

$$tranz404 = \begin{pmatrix} 0.95 & 0.016 & 0.016 & 0.016 \\ 0.033 & 0.9 & 0.033 & 0.033 \\ 0.033 & 0.033 & 0.9 & 0.033 \\ 0.033 & 0.033 & 0.033 & 0.9 \end{pmatrix}$$

$$tranz501 = \begin{pmatrix} 0.95 & 0.0125 & 0.0125 & 0.0125 & 0.0125 \\ 0.025 & 0.9 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.9 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.9 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 & 0.9 \end{pmatrix}$$

Bibliografija

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

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

[3] J. Ernst, M. Kellis, *Discovery and characterization of chromatin states for systematic annotation of the human genome*, Nature Biotechnology, 28 (2010.), 817-827

[4] D. Kežman, *Kompleksnost skrivenih Markovljevih modela*, diplomski rad, PMF-MO, Zagreb, 2013.

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

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

[7] http://en.wikipedia.org/wiki/Hidden_Markov_model

[8] http://en.wikipedia.org/wiki/Bayesian_information_criterion

Sažetak

U ovom radu bavimo se statističkim modelom, skrivenim Markovljevim modelom (eng. hidden Markov model ili kraće HMM) za što nas je motivirala njegova primjena u radu s ljudskom genomom znanstvenika J. Ernsta i M. Kellisa.

U radu dajemo formalnu definiciju skriveniog Markovljevog modela, dajemo primjere takvih modela te opisujemo algoritme koje koristimo, između ostalih i forward, backward te Viterbijev algoritam. Nakon toga implementiramo i analiziramo postupke za procjenu parametara (emisijske i tranzicijske vjerojatnosti) HMM-a: Baum-Welch algoritam i Viterbijevo treniranje. Baum-Welch algoritam se pokazao efikasniji na primjerima koje navodimo ali unatoč tome zaključujemo da ne radi u praksi zbog niza problema kao što su lokalni ekstemi i izbor inicijalnih parametara. Također, kriteriji za procjenu kompleksnosti modela koje smo promatrali, AIC i BIC, pokazali su se nekorisni.

Summary

This thesis is concerned with a statistical model called hidden Markov model (HMM). We were led to this topic by the paper by J. Ernst and M. Kellis where they applay HMMs to the study of human genome.

We give a formal definition of the HMM, give examples of such models and describe algorithms that are used, among others, the forward, the backward and the Viterbi algorithm. Furthermore, we implement and analyze methods for parameter estimation (emission and transition probabilities) for HMMs: the Baum-Welch algorithm and Viterbi training. The Baum-Welch algorithm turned out to be more effective than Viterbi training when tested on simulated examples that we present. However, the Baum.Welch algorithm does not work in practise because of the issues like the local maxima and the choice of the initial parameters. Finally, we applay information criteria AIC and BIC to study complexity of our models.

Životopis

- Rođena sam u Zagrebu 24.3.1988.
- Od 1995. do 2003. pohađam OŠ Većeslava Holjevca u Zagrebu
- Od 2003. do 2007. pohađam XI. gimnaziju u Zagrebu
- Od 2007. do 2011. pohađam preddiplomski sveučilišni studij Matematika na PMF-MO u Zagrebu
- 2011. upisujem diplomski sveučilišni studij Matematička statistika na PMF-MO u Zagrebu