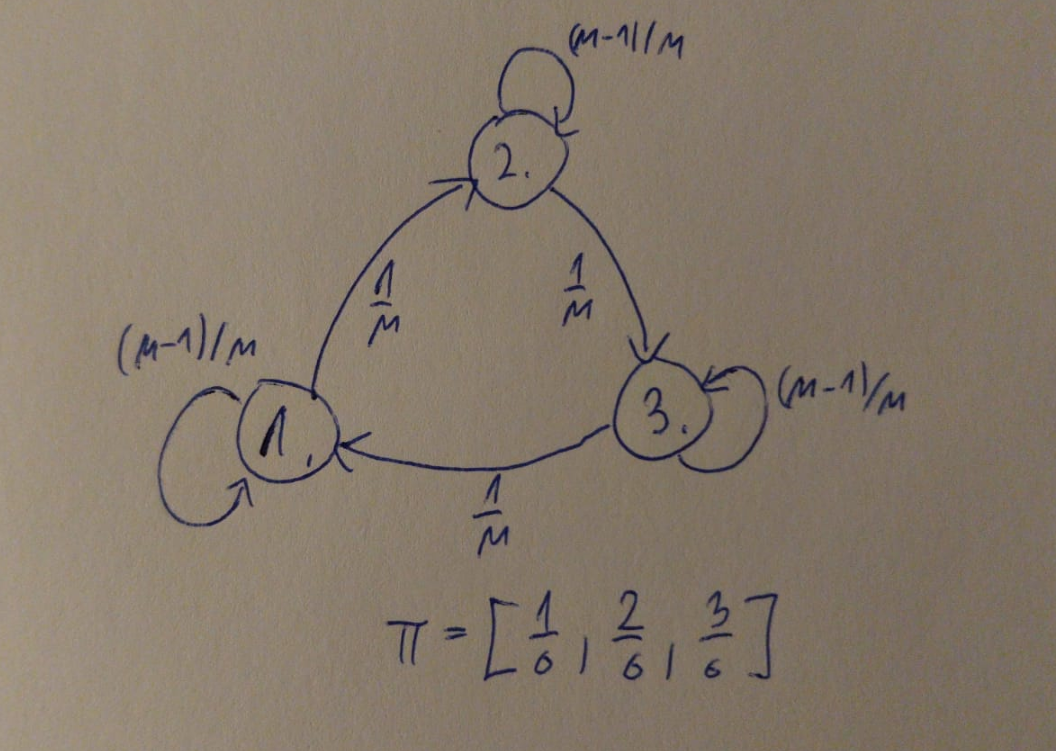
Obrada informacija – Treća laboratorijska vježba

Fran Galić, JMBAG: 0036546889

Opis:

U ovoj laboratorijskoj vježbi istražuje se primjena skrivenih Markovljevih modela (HMM) kroz praktične zadatke u MATLAB-u koristeći HMM Toolbox. Eksperiment uključuje proučavanje uputa za rad s HMM modelima, s naglaskom na konkretne primjere korištenja alata. Zadatci su strukturirani prema odgovarajućim poglavljima iz uputa, omogućujući postupno usvajanje znanja i primjenu u stvarnim scenarijima.

IZVJEŠTAJ: Skicirajte dijagram stanja za ovako opisan model.



**Pod-zadatak 1 - Cjelovito definiranje HMM modela u Matlabu**

Temeljem zadanih ucestalosti pojedinih ishoda bacanja pristranih kocki i temeljem zadanog parametra M u vasem Moodle zadatku, potrebno je dopuniti predlozak Matlab skripte kako bi cjelovito opisali zadani HMM model ovog eksperimenta ukljucujuci i matricu vjerojatnosti osmatranja izlaznih simbola.

% ===============================================================

% Oznacavanje stanja HMM modela

% Imamo tri pristrane kocke od kojih uvijek bacamo jednu odabranu

% Stanja modela su indeksi koristene pristrane kocke

% Vektor inicijalne vjerojatnosti stanja (za t=1)

% odredjen bacanjem nepristrane kocke:

prior0=[

1 % Prva kocka (ako je palo '1')

2 % Druga kocka (ako je palo '2' ili '3')

3 % Treca kocka (ako je palo '4', '5' ili '6')

]/6;

% Broj stanja HMM modela

Q=size(prior0,1);

% ---------------------------------------------------------------

% Matrica vjerojatnosti promjena stanja

%

% a11 a12 a13

% a21 a22 a23

% a31 a32 a33

% Za eksperiment sa stohastickom izmjenom stanja, parametar

% M se koristi za definiranje vjerojatnosi prijelaza u

% novo stanje u matrici prijelaza A, pri cemu se stanja nuzno

% mijenjaju ciklicki radi forsirane strukture tranzicijske matrice.

M= 7; % Ovdje definirate M iz vaseg personaliziranog zadatka.

% Formiraj matricu vjerojatnosti prijelaza stanja

% (uz ciklicku strukturu izmjene stanja, jer su

% prijelazi 1->3, 2->1 i 3->2 zabranjeni)

transmat0=[

M-1 1 0 % P(1|1) P(2|1) P(3|1)

0 M-1 1 % P(1|2) P(2|2) P(3|2)

1 0 M-1 % P(1|3) P(2|3) P(3|3)

]/M;

% Matrica emisijskih vjerojatnosti

% svaki redak odgovara jednom stanju, a

% svaki stupac jednoj mogucoj opservaciji

% Matrica učestalosti osmatranja (prema slici)

B\_count = [

20, 5, 5, 6, 2, 2; % Kocka 1

5, 5, 20, 5, 3, 2; % Kocka 2

6, 7, 3, 1, 20, 3 % Kocka 3

];

% Ukupan broj bacanja po kocki

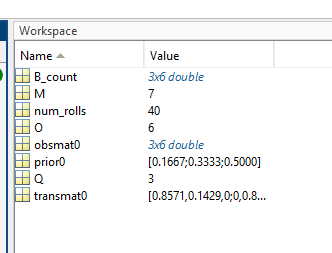
num\_rolls = 40;

% Izračun matrice emisijskih vjerojatnosti

obsmat0 = B\_count / num\_rolls;

O=size(obsmat0,2);

Rezultat:



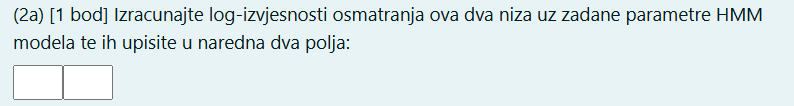
**Pod-zadatak 2 - Odredjivanje log-izvjesnosti osmatranja zadanog izlaznog niza simbola za zadani model**

Osmotrena su dva niza duljine T=41 simbola kojeg je generirao model L:

O = [o1 .. oT] =

[ 3 3 1 4 5 2 1 1 2 3 5 1 1 1 1 6 1 5 5 5 5 5 5 5 5 6 5 6 5 5 4 1 4 5 5 5 5 5 5 1 1]

[ 6 5 2 2 4 2 6 2 4 6 5 2 2 2 6 6 6 3 6 6 6 6 4 5 3 6 6 6 6 6 3 6 6 3 6 4 5 2 4 1 2]



data1=[ 3 3 1 4 5 2 1 1 2 3 5 1 1 1 1 6 1 5 5 5 5 5 5 5 5 6 5 6 5 5 4 1 4 5 5 5 5 5 5 1 1];

data2=[ 6 5 2 2 4 2 6 2 4 6 5 2 2 2 6 6 6 3 6 6 6 6 4 5 3 6 6 6 6 6 3 6 6 3 6 4 5 2 4 1 2];

if ~iscell(data1)

data1 = num2cell(data1, 2);

end

ncases1 = length(data1);

if ~iscell(data2)

data2 = num2cell(data2, 2);

end

ncases2 = length(data2);

loglik1 = 0;

errors1 = [];

for m=1:ncases1

obslik01 = multinomial\_prob(data1{m}, obsmat0);

[alpha1, beta1, gamma1, ll1] = ...

fwdback(prior0, transmat0, obslik01, 'scaled', 0);

if ll1==-inf

errors1 = [errors1 m];

end

loglik1 = loglik1 + ll1;

end

loglik2 = 0;

errors2 = [];

for m=1:ncases2

obslik02 = multinomial\_prob(data2{m}, obsmat0);

[alpha2, beta2, gamma2, ll2] = ...

fwdback(prior0, transmat0, obslik02, 'scaled', 0);

if ll2==-inf

errors2 = [errors2 m];

end

loglik2 = loglik2 + ll2;

end

ll1

ll2

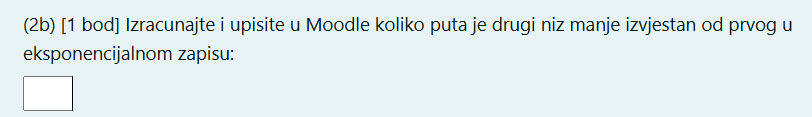
Rezultat:



IZVJEŠTAJ: Možete li usporedbom zadanih nizova obrazložiti razlog zbog kojeg je drugi niz

manje izvjestan od prvog? Opišite riječima.

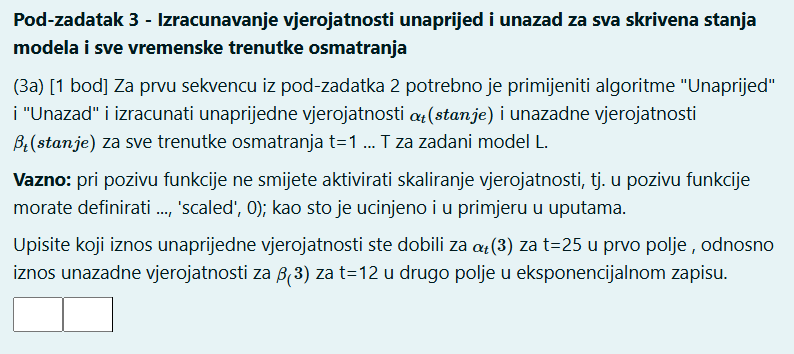
Odgovor: Drugi niz je manje vjerojatan jer sadrži više neuobičajenih prijelaza i ishoda koji su manje vjerojatni prema zadanim emisijskim vjerojatnostima i matrici prijelaza, npr. Sadrži vise pojavljivanja broja 6 koji nije toliko vjerojatan.



Računamo po formuli:



I dobivamo: 2.944116e14



[alpha1, beta1, gamma1, ll1] = ...

fwdback(prior0, transmat0, obslik01, 'scaled', 0);

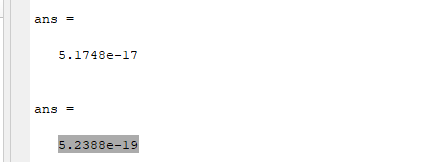
[alpha2, beta2, gamma2, ll2] = ...

fwdback(prior0, transmat0, obslik02, 'scaled', 0);

alpha1(3, 25)

beta1(3, 12)

Rezultat:



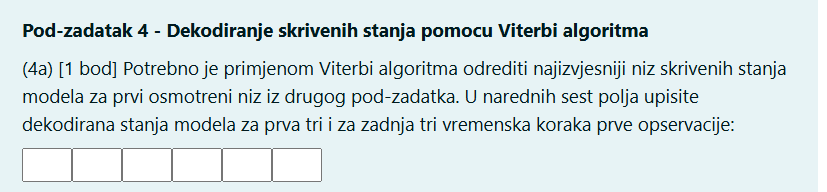
IZVJEŠTAJ: Obrazložite i prikažite kako možete iskoristiti vjerojatnosti alfa iz zadnjeg koraka u

svrhu određivanja log-izvjesnosti osmatranja cijelog niza, odnosno kako možete iskoristiti

izračunatu unazadnu vjerojatnost beta iz prvog vremenskog koraka u istu svrhu, te usporedite

tako dobivene rezultate s onim iz pod-zadatka 2 za prvi osmotreni niz.

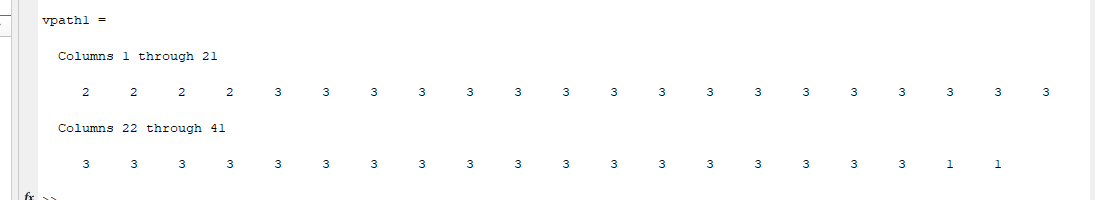
Odgovor: Log-izvjesnost niza može se izračunati korištenjem α vjerojatnosti iz zadnjeg koraka ili β vjerojatnosti iz prvog koraka. Kod α metode, sumira se ukupna vjerojatnost svih stanja na kraju niza, dok se kod β metode uzimaju početne vjerojatnosti, emisijske vjerojatnosti prvog simbola i unazade vjerojatnosti. Oba pristupa daju isti rezultat, što potvrđuje dosljednost modela. Ukoliko postoje nekakva odstupanja riječ je o numeričkim pogreškama.

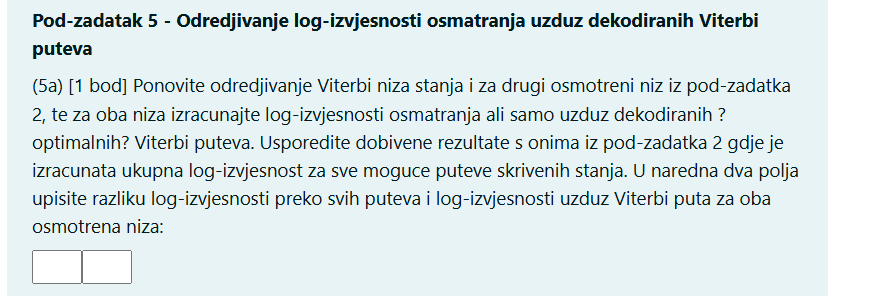


% Najizvjesniji put

vpath1 = viterbi\_path(prior0, transmat0, obslik01)

Rezultat:





IZVJEŠTAJ: Usporedite dobivene rezultate s onima iz pod-zadatka 2 gdje je izračunata ukupna

log-izvjesnost za sve moguće puteve skrivenih stanja.

U Moodle treba upisati razliku log-izvjesnosti preko svih puteva i log-izvjesnosti uzduž Viterbi

puta za oba osmotrena niza.

%%

% Najizvjesniji put

vpath1 = viterbi\_path(prior0, transmat0, obslik01)

% Najizvjesniji put

vpath2 = viterbi\_path(prior0, transmat0, obslik02)

[ll11, p11] = dhmm\_logprob\_path(prior0, transmat0, obslik01, vpath1)

[ll22, p22] = dhmm\_logprob\_path(prior0, transmat0, obslik02, vpath2)

ll11

ll22

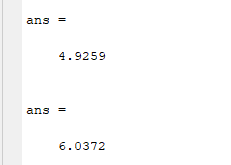
ll1 - ll11

ll2 - ll22

Rezultat:



Rješenje:



IZVJEŠTAJ: Što nam govori predznak ovih razlika? Diskutirajte dobivene rezultate u izvještaju.

Biste li mogli izračunati i izvjesnosti osmatranja cjelovitih zadanih osmotrenih nizova (u punoj

dužini) uzduž svih mogućih pojedinačnih puteva rešetke stanja, kao što je opisano u

dokumentu s uputama? Ako ne, zašto ne? Obrazložite.

**Odgovor:**

**Što nam govori predznak razlika?**

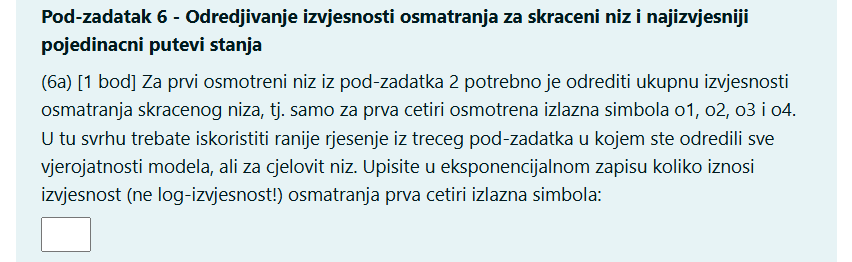
Predznak razlika je pozitivan, što znači da je log-izvjesnost uzduž Viterbi puta manja od ukupne log-izvjesnosti svih mogućih puteva. To ukazuje da, iako Viterbi algoritam odabire najvjerojatniji put, ukupna log-izvjesnost uzima u obzir sve puteve, uključujući manje vjerojatne, koji također doprinose ukupnoj vjerojatnosti osmatranja.

**Diskusija o rezultatima:**

Razlika u log-izvjesnosti između Viterbi puta i svih mogućih puteva potvrđuje očekivanja u kontekstu HMM-a. Viterbi algoritam odabire samo jedan optimalni put, dok ukupna log-izvjesnost uzima u obzir sve puteve, što povećava ukupnu vjerojatnost. Razlike od 4.926 i 6.0372 pokazuju da se s većim brojem puteva razlika u log-izvjesnosti povećava.

**Izračun izvjesnosti osmatranja svih puteva:**

Izračun izvjesnosti za sve moguće puteve je računarski zahtjevan, osobito za složene modele i duže nizove, jer bi to zahtijevalo previše računalnih resursa. Zbog toga se u praksi koristi Viterbi algoritam koji traži samo optimalni put.



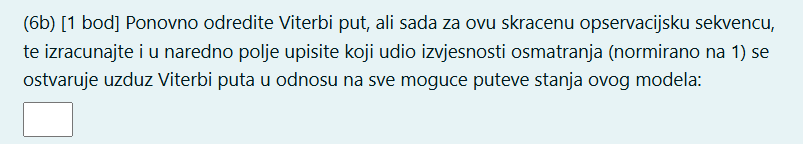
%%

alpha1(1,4) + alpha1(2,4) + alpha1(3,4)

Rezultat: 0.0012

IZVJEŠTAJ: Objasnite kako ste dobili izvjesnost osmatranja skraćenog niza.

Odgovor: Pozbrojio sam alpha vrijendosti za sva stanja u trenutku 4.



data3=[3 3 1 4];

if ~iscell(data3)

data3 = num2cell(data3, 2);

end

ncases3 = length(data3);

loglik3 = 0;

errors3 = [];

for m=1:ncases3

obslik03 = multinomial\_prob(data3{m}, obsmat0);

[alpha3, beta3, gamma3, ll3] = ...

fwdback(prior0, transmat0, obslik03, 'scaled', 0);

if ll3==-inf

errors3 = [errors3 m];

end

loglik3 = loglik3 + ll3;

end

vpath3 = viterbi\_path(prior0, transmat0, obslik03)

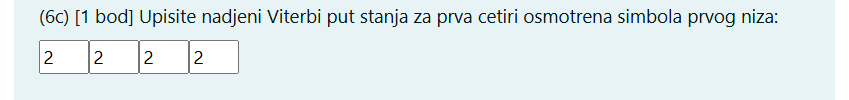
[ll33, p33] = dhmm\_logprob\_path(prior0, transmat0, obslik03, vpath3)

I onda uzmemo e^(ll33) / 0.0012 iz ranijeg podzadatka

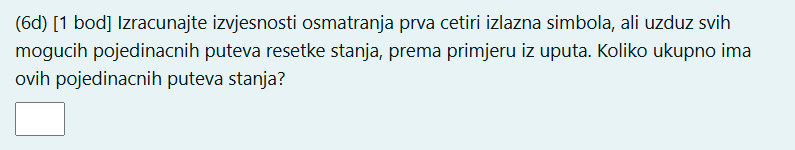
IZVJEŠTAJ: Jeste li za nalaženje ovog Viterbi rješenja skraćenog niza smjeli koristiti rješenje iz

pod-zadataka 4 i 5? Obrazloži odgovor.

Odgovor: Ne jer samo zato sto je nešto Viterbijev put za cijeli niz ne znaci da će također biti i Viterbijev put za neki podniz.







%%

% Generiranje svih mogućih puteva za 4 stupca (vrijednosti od 1 do 3)

[grid1, grid2, grid3, grid4] = ndgrid(1:3, 1:3, 1:3, 1:3);

% Kombinacija svih vrijednosti u matricu

mpath = [grid1(:), grid2(:), grid3(:), grid4(:)];

% Prikaz rezultata

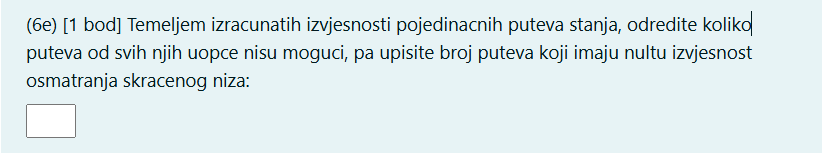
mpath

llm=zeros(81,1); % Stupac za log-izvjesnosti

for i=1:81,

[llm(i), p] = dhmm\_logprob\_path(prior0, transmat0, obslik03, mpath(i,:));

end;

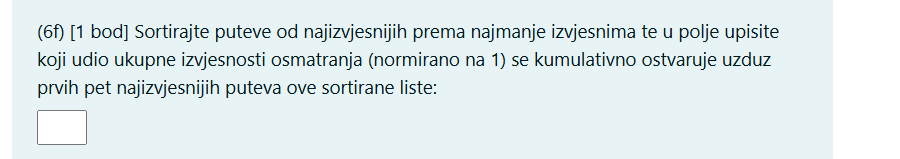


count\_inf = sum(isinf(llm) & llm < 0);

count\_inf

IZVJEŠTAJ: Obrazložite razlog.

U skrivenom Markovljevom modelu, neki putevi kroz rešetku stanja mogu biti nemogući zbog nulte ili izuzetno niske vjerojatnosti prijelaza između određenih stanja. To znači da, ako prijelazna matrica ne dopušta prijelaz između određenih stanja, ti putevi neće biti mogući. Također, početne vjerojatnosti mogu ograničiti koje sekvence stanja su uopće moguće.



%%

% Sortiraj puteve prema izvjesnosti osmatranja od najizvjesnijeg do

% najmanje izvjesnog

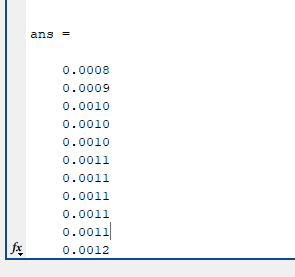
[sllm,illm]=sort(-llm);

% Kumulativno zbrojih izvjesnosti od samo jednog najizvjesnijeg

% puta, pa prva dva puta stanja po izvjesnosti, pa prva tri, ... i

% tako sve do sume izvjesnosti po svim mogucim putevima stanja

cumsum(exp(-sllm))



Rješenje: 0.0010 / 0.0012 = 0.833333333…

IZVJEŠTAJ: Navedite o kojim putevima stanja se radi. Nalazi se među njima i skraćeni Viterbi

put? Mogu li različiti putevi stanja imati istu izvjesnost? O čemu to ovisi?

**Odgovor:**

Putevi stanja sortirani od najizvjesnijih prema najmanje izvjesnima predstavljaju različite sekvence stanja koje mogu generirati promatrani niz. Skraćeni Viterbi put se obično nalazi među prvih nekoliko najizvjesnijih puteva jer Viterbi algoritam odabire put s najvećom vjerojatnošću.

Različiti putevi stanja mogu imati istu izvjesnost ako postoji više puteva koji imaju vrlo slične ili identične vjerojatnosti, što ovisi o samoj strukturi prijelaznih i emisijskih vjerojatnosti u modelu.

**Pod-zadatak 7 - Generiranje opservacija za zadani model**

(7a) [0 bodova] Generirajte visestruke slucajne nizove osmotrenih izlaznih simbola s nex=19 razlicitih nizova, pri cemu svaki niz treba biti duljine T=188 vremenskih uzoraka. Za generiranje podataka koristiti funkciju dhmm\_sample u skladu s uputama, uz parametre HMM modela iz vaseg individualnog pod-zadatka 1. Sacuvajte ovu matricu opservacija jer ce biti intenzivno koristena i u narednim pod-zadatcima. Prije poziva funkcije, svakako resetirajte generator slucajnih brojeva na pocetnu vrijednost naredbom rng('default'). Vase rjesenje ce biti provjereno i bodovano u narednom pod-zadatku.

IZVJEŠTAJ: Dokumentirajte način generiranja observacija.

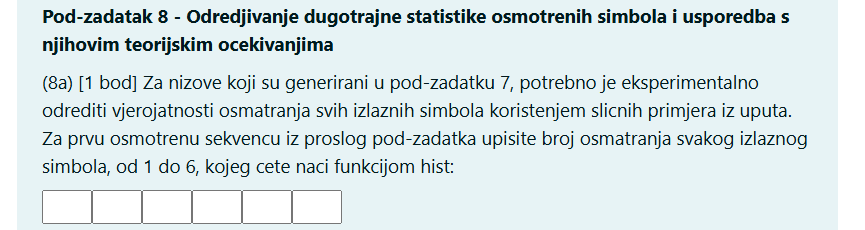
%%

rng('default')

T = 188; % duljina svakog niza

nex = 19; % broj opservacijskih nizova

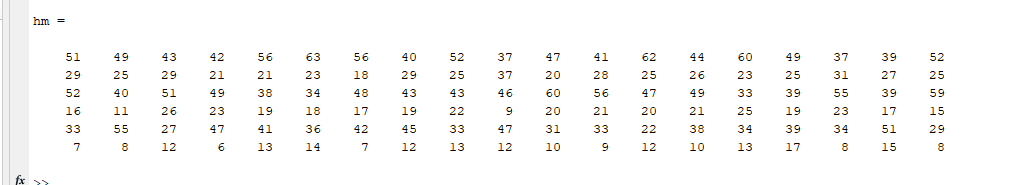
dataRG = dhmm\_sample(prior0, transmat0, obsmat0, nex, T);

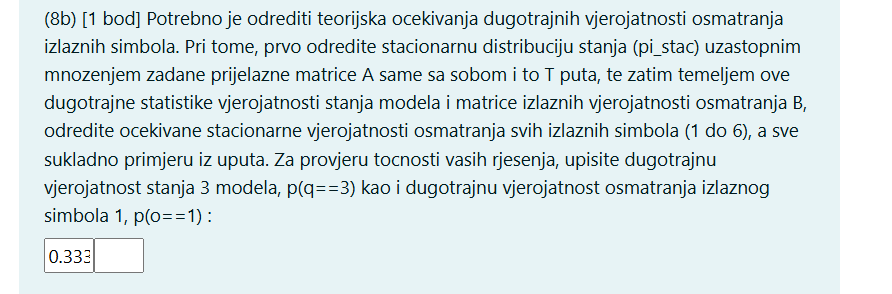


%%

hm=hist(dataRG',[1 2 3 4 5 6])

Rezultat:



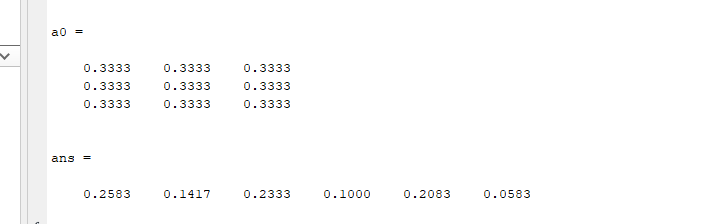


a0=transmat0; for i=1:188, a0=a0\*transmat0; end;

a0

a0(1,:)\*obsmat0

rezultat:



IZVJEŠTAJ: Diskutirajte dobivene dugotrajne vjerojatnosti pojedinih stanja, odnosno izlaznih

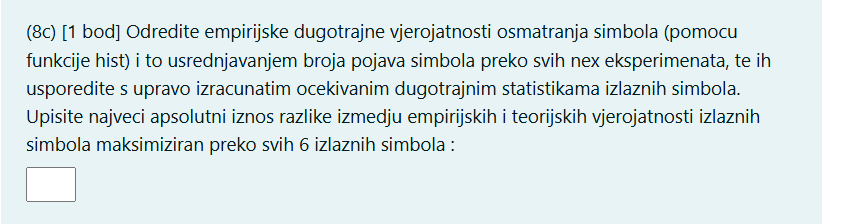
simbola. Kako bi izgledao degenerirani HMM model s jednakim dugotrajnim statistikama

opservacija izlaznih simbola?

**Odgovor:**

Dobivene dugotrajne vjerojatnosti stanja i izlaznih simbola ukazuju na raspodjelu vjerojatnosti unutar modela. Dugotrajne vjerojatnosti stanja, poput p(q=3)=0.3333p(q = 3) = 0.3333p(q=3)=0.3333, pokazuju koliko će svaki od mogućih stanja biti zastupljen u dugoročnom smislu, dok dugotrajne vjerojatnosti izlaznih simbola, kao p(o=1)=0.2583p(o = 1) = 0.2583p(o=1)=0.2583, pokazuju vjerojatnost da će određeni izlazni simbol biti generiran na temelju stanja modela.

Degenerirani HMM model s jednakim dugotrajnim statistikama opservacija izlaznih simbola bio bi model u kojem su prijelazne vjerojatnosti svih stanja jednake (npr. sve vjerojatnosti prijelaza su 1), čime bi svaki izlazni simbol imao istu dugoročnu vjerojatnost. Takav model ne bi bio sposoban za diskriminaciju između različitih stanja jer bi u svakoj situaciji imao istu vjerojatnost za sve izlazne simbole, što znači da bi gubio informaciju potrebnu za razlikovanje između različitih sekvenci stanja.



IZVJEŠTAJ: Usporedite ih s upravo izračunatim očekivanim dugotrajnim statistikama izlaznih

simbola.

hm=hist(dataRG',[1 2 3 4 5 6])

a0=transmat0; for i=1:188, a0=a0\*transmat0; end;

a0

teorijska\_vjerovatnost = a0(1,:)\*obsmat0

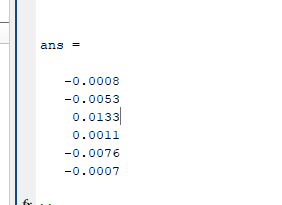
empiriskeVjerovatnosti = hm /188

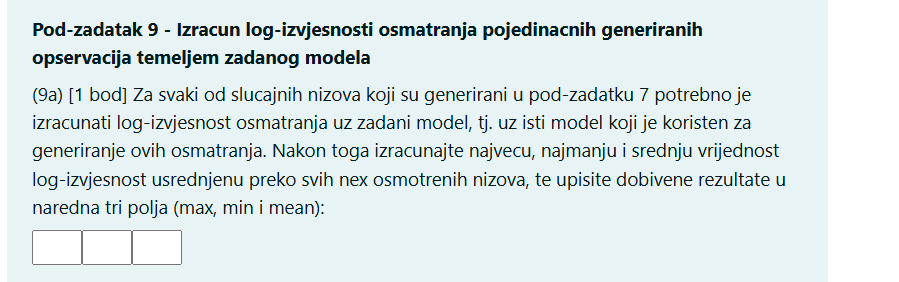
row\_means = mean(empiriskeVjerovatnosti, 2);

row\_means

row\_means - teorijska\_vjerovatnost'

Rezultat:





% Izracunaj u petlji log-izvjesnosti svakog niza

nex2=size(dataRG,1); % Broj eksperimenata

llm2=zeros(nex2,1); % Stupac log-izvjesnosti

for i=1:nex2,

llm2(i)=dhmm\_logprob(dataRG(i,:), prior0, transmat0, obsmat0);

end;

llm2

max\_value = max(llm2); % Najveća vrijednost

min\_value = min(llm2); % Najmanja vrijednost

mean\_value = mean(llm2); % Srednja vrijednost

max\_value

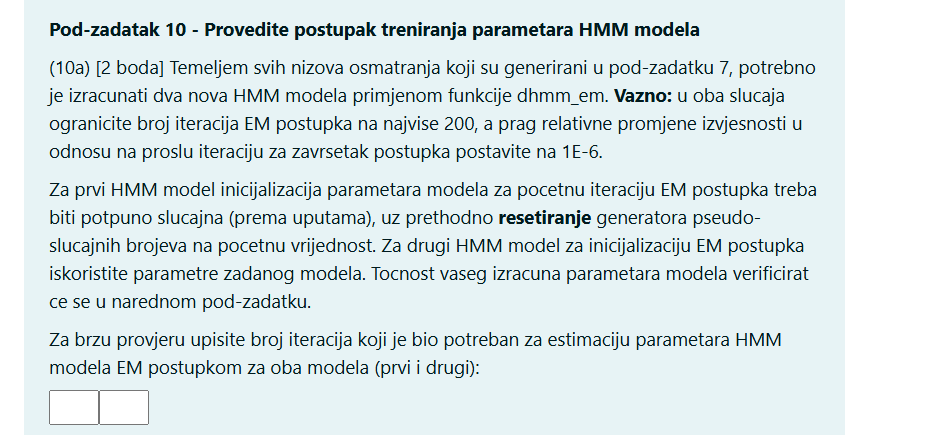
min\_value

mean\_value

IZVJEŠTAJ: Zašto se izvjesnosti pojedinih nizova razlikuju?

Odgovor:

Izvjesnosti pojedinih nizova se razlikuju jer su nizovi generirani kao slučajni procesi prema zadanim vjerojatnostima modela. Razlike u izvjesnostima ovise o tome koliko su generirani nizovi usklađeni s parametrima modela, poput prijelaznih i emisijskih vjerojatnosti. Nizovi koji sadrže simbole češće očekivane prema modelu imaju veću izvjesnost, dok oni s manje vjerojatnim simbolima imaju manju izvjesnost.



%%

rng('default')

prior1 = normalise(rand(Q,1));

transmat1 = mk\_stochastic(rand(Q,Q));

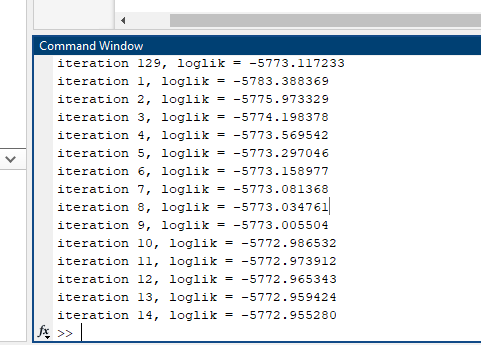
obsmat1 = mk\_stochastic(rand(Q,O))

[LL222, prior222, transmat222, obsmat222] = dhmm\_em(dataRG, prior1, transmat1, obsmat1, 'max\_iter', 200, 'thresh', 1e-6);

%%

[LL333, prior333, transmat333, obsmat333] = dhmm\_em(dataRG, prior0, transmat0, obsmat0, 'max\_iter', 200, 'thresh', 1e-6);

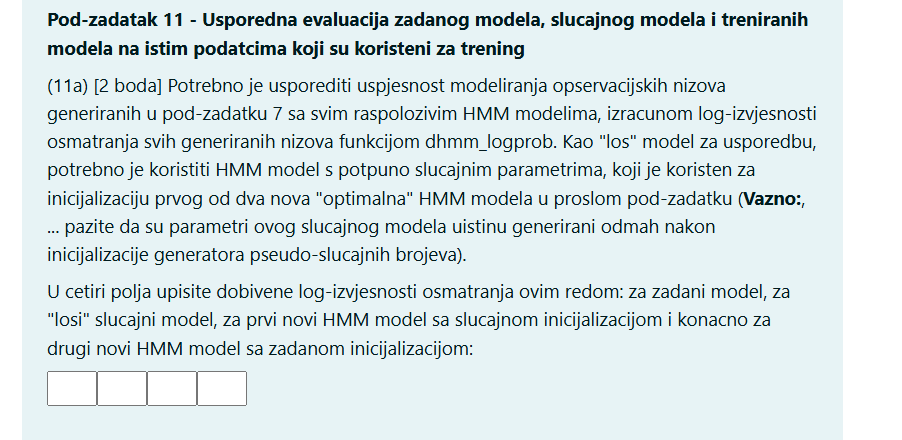
Rezultat:



IZVJEŠTAJ: Obrazložite razliku broja iteracija.

Odgovor:

Razlika u broju iteracija između dva slučaja nastaje zbog početnih uvjeta za EM postupak. U prvom slučaju, gdje su parametri inicijalizirani potpuno slučajno, algoritam treba više iteracija da konvergira jer kreće od nepovoljnog početnog stanja. U drugom slučaju, gdje su za inicijalizaciju korišteni već poznati parametri zadanog modela, algoritam počinje bliže optimalnom rješenju, pa konvergencija zahtijeva manje iteracija.



%%

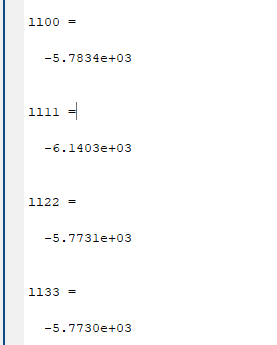
ll00=dhmm\_logprob(dataRG, prior0, transmat0, obsmat0)

ll11=dhmm\_logprob(dataRG, prior1, transmat1, obsmat1)

ll22=dhmm\_logprob(dataRG, prior222, transmat222, obsmat222)

ll33=dhmm\_logprob(dataRG, prior333, transmat333, obsmat333)

Rezultat:



IZVJEŠTAJ: Objasnite koji je odnos log-izvjesnosti pojedinih nizova iz pod-zadataka 9 s upravo

određenom ukupnom log-izvjesnosti svih nizova za zadani model. Diskutirajte dobivene

rezultate novih modela u usporedbi s log-izvjesnosti osmatranja istih nizova za zadani model.

Prikažite i usporedite estimirane vrijednosti parametara novih treniranih modela (matrice A,

B, pi) sa zadanim modelom. Kako objašnjavate razlike parametara ovih modela? Je li provjera

estimiranog modela na istim podatcima koji su korišteni za treniranje primjeren postupak?

Kako bi se trebao provesti pravi postupak treniranja i validacije modela?

**Odgovor**:

Dobiveni rezultati pokazuju da ukupna log-izvjesnost svih nizova za zadani model (iz pod-zadatka 9) predstavlja sumu pojedinačnih log-izvjesnosti osmatranih nizova. Usporedba s novim modelima otkriva da trenirani modeli, iako bazirani na istim podatcima, pokazuju slične vrijednosti ukupne log-izvjesnosti kao zadani model, što ukazuje na dobru prilagodbu podacima.

Što se tiče razlika u parametrima (A, B, pi), trenirani modeli s inicijalno slučajnim i zadanima parametrima pokazali su blisku konvergenciju, ali zbog različitih početnih uvjeta moguće su male varijacije. Takve razlike odražavaju ovisnost EM algoritma o početnoj točki, ali i sposobnost algoritma da pronađe lokalno optimalna rješenja.

Provjera modela na istim podatcima korištenima za treniranje nije idealna, jer može dovesti do overfittinga. Pravi postupak zahtijevao bi podjelu podataka na trening i test skup te validaciju modela na neovisnim podatcima kako bi se osigurala njegova generalizacija.