

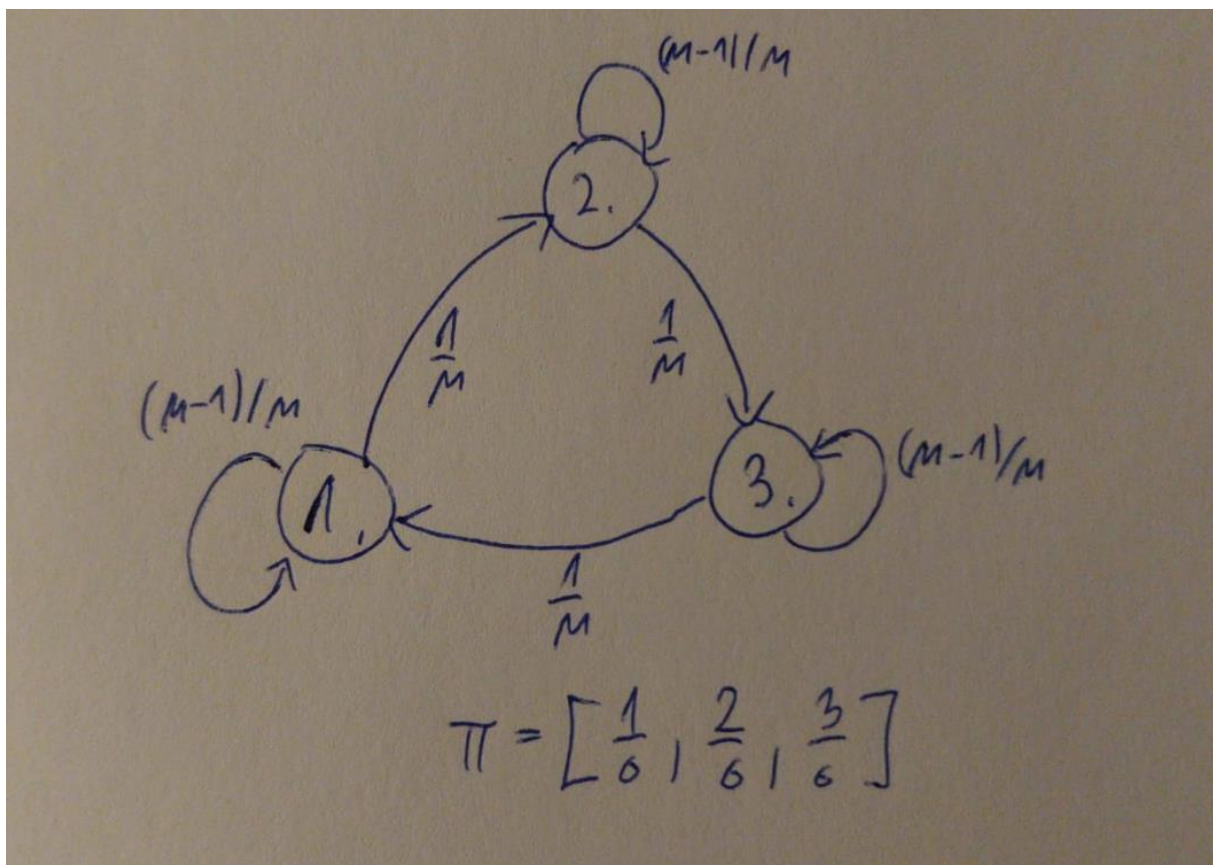
Obrada informacija – Treća laboratorijska vježba

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 učestalosti pojedinih ishoda bacanja pristranih kocki i temeljem zadanog parametra M u vašem Moodle zadatku, potrebno je dopuniti predložak Matlab skripte kako bi cjelovito opisali zadani HMM model ovog eksperimenta uključujući 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 % Treća 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 vjerojatnosti prijelaza u
% novo stanje u matrici prijelaza A, pri čemu se stanja nužno
% mijenjaju ciklički radi forsirane strukture tranzicijske matrice.
M= 7; % Ovdje definirajte M iz vašeg personaliziranog zadatka.
% Formiraj matricu vjerojatnosti prijelaza stanja
% (uz cikličku 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 mogućoj 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:

Workspace	
Name ▲	Value
B_count	3x6 double
M	7
num_rolls	40
O	6
obsmat0	3x6 double
prior0	[0.1667;0.3333;0.5000]
Q	3
transmat0	[0.8571,0.1429,0;0,0.8...

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:

$O = [o_1 \dots o_T] =$

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

(2a) [1 bod] Izracunajte log-izvjesnosti osmatranja ova dva niza uz zadane parametre HMM modela te ih upisite u naredna dva polja:

--	--

```
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 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:

```
l11 =  
-62.4252  
  
l12 =  
-95.7412
```

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 više pojavljivanja broja 6 koji nije toliko vjerojatan.

(2b) [1 bod] Izračunajte i upisite u Moodle koliko puta je drugi niz manje izvjestan od prvog u eksponencijalnom zapisu:

Računamo po formuli:

$$\text{Omjer} = e^{\log L_1 - \log L_2}$$

I dobivamo: $2.944116e14$

Pod-zadatak 3 - Izračunavanje vjerojatnosti unaprijed i unazad za sva skrivena stanja modela i sve vremenske trenutke osmatranja

(3a) [1 bod] Za prvu sekvencu iz pod-zadatka 2 potrebno je primijeniti algoritme "Unaprijed" i "Unazad" i izračunati unaprijedne vjerojatnosti $\alpha_t(\text{stanje})$ i unazadne vjerojatnosti $\beta_t(\text{stanje})$ za sve trenutke osmatranja $t=1 \dots T$ za zadani model L.

Vazno: pri pozivu funkcije ne smijete aktivirati skaliranje vjerojatnosti, tj. u pozivu funkcije morate definirati ..., 'scaled', 0); kao što je učinjeno i u primjeru u uputama.

Upisite koji iznos unaprijedne vjerojatnosti ste dobili za $\alpha_t(3)$ za $t=25$ u prvo polje, odnosno iznos unazadne vjerojatnosti za $\beta_t(3)$ za $t=12$ u drugo polje u eksponencijalnom zapisu.

```
[alpha1, beta1, gamma1, l11] = ...  
fwdback(prior0, transmat0, obslik01, 'scaled', 0);  
[alpha2, beta2, gamma2, l12] = ...  
fwdback(prior0, transmat0, obslik02, 'scaled', 0);  
  
alpha1(3, 25)
```

```
beta1(3, 12)
```

Rezultat:

```
ans =  
  
5.1748e-17  
  
ans =  
  
5.2388e-19
```

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.

Pod-zadatak 4 - Dekodiranje skrivenih stanja pomocu Viterbi algoritma

(4a) [1 bod] Potrebno je primjenom Viterbi algoritma odrediti najizvjesniji niz skrivenih stanja modela za prvi osmotreni niz iz drugog pod-zadatka. U narednih šest polja upisite dekodirana stanja modela za prva tri i za zadnja tri vremenska koraka prve opservacije:

--	--	--	--	--	--

```
% Najizvjesniji put  
vpath1 = viterbi_path(prior0, transmat0, obslik01)
```

Rezultat:

```
vpath1 =  
  
Columns 1 through 21  
2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3  
  
Columns 22 through 41  
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1  
fr ~
```

Pod-zadatak 5 - Odredjivanje log-izvjesnosti osmatranja uzduz dekodiranih Viterbi puteva

(5a) [1 bod] Ponovite odredjivanje Viterbi niza stanja i za drugi osmotreni niz iz pod-zadatka 2, te za oba niza izracunajte log-izvjesnosti osmatranja ali samo uzduz dekodiranih ? optimalnih? Viterbi puteva. Usporedite dobivene rezultate s onima iz pod-zadatka 2 gdje je izracunata ukupna log-izvjesnost za sve moguće puteve skrivenih stanja. U naredna dva polja upisite razliku log-izvjesnosti preko svih puteva i log-izvjesnosti uzduz Viterbi puta za oba osmotrena niza:

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)

[l111, p11] = dhmm_logprob_path(prior0, transmat0, obslik01, vpath1)

[l112, p22] = dhmm_logprob_path(prior0, transmat0, obslik02, vpath2)

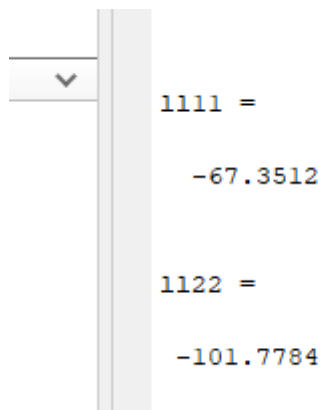
l111

l112

l11 - l111

l12 - l112

Rezultat:



```
l111 =  
-67.3512  
  
l112 =  
-101.7784
```

Rješenje:

```
ans =  
  
4.9259  
  
ans =  
  
6.0372
```

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.

Pod-zadatak 6 - Određivanje izvjesnosti osmatranja za skraćeni niz i najizvjesniji pojedinačni putevi stanja

(6a) [1 bod] Za prvi osmotreni niz iz pod-zadatka 2 potrebno je odrediti ukupnu izvjesnosti osmatranja skraćenog niza, tj. samo za prva četiri osmotrena izlazna simbola o1, o2, o3 i o4. U tu svrhu trebate iskoristiti ranije rjesenje iz trećeg pod-zadatka u kojem ste odredili sve vjerojatnosti modela, ali za cjelovit niz. Upisite u eksponencijalnom zapisu koliko iznosi izvjesnost (ne log-izvjesnost!) osmatranja prva četiri izlazna simbola:

%%

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 vrijednosti za sva stanja u trenutku 4.

(6b) [1 bod] Ponovno odredite Viterbi put, ali sada za ovu skracenu opservacijsku sekvencu, te izracunajte i u naredno polje upisite koji udio izvjesnosti osmatranja (normirano na 1) se ostvaruje uzduz Viterbi puta u odnosu na sve moguće puteve stanja ovog modela:

```
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 što je nešto Viterbijev put za cijeli niz ne znači da će također biti i Viterbijev put za neki podniz.

(6c) [1 bod] Upisite nadjeni Viterbi put stanja za prva četiri osmotrena simbola prvog niza:

2	2	2	2
---	---	---	---

vpath3 =

2 2 2 2

(6d) [1 bod] Izračunajte izvjesnosti osmatranja prva četiri izlazna simbola, ali uzduž svih mogućih pojedinačnih puteva resetke stanja, prema primjeru iz uputa. Koliko ukupno ima ovih pojedinačnih puteva stanja?

```
%%  
% 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;
```

(6e) [1 bod] Temeljem izračunatih izvjesnosti pojedinačnih puteva stanja, odredite koliko puteva od svih njih uopće nisu mogući, pa upišite broj puteva koji imaju nultu izvjesnost osmatranja skraćenog niza:

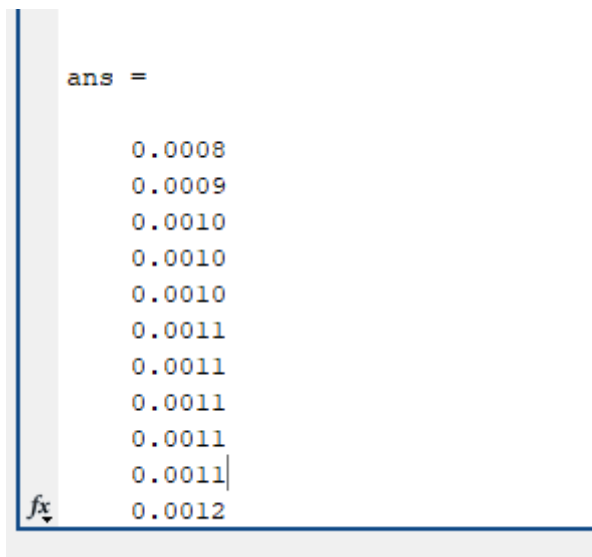
```
count_inf = sum(isinf(llm) & llm < 0);  
count_inf
```

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.

(6f) [1 bod] Sortirajte puteve od najizvjesnijih prema najmanje izvjesnima te u polje upišite koji udio ukupne izvjesnosti osmatranja (normirano na 1) se kumulativno ostvaruje uzduž prvih pet najizvjesnijih puteva ove sortirane liste:

```
%%  
  
% 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 mogućim 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 višestruke slučajne nizove osmotrenih izlaznih simbola s $nex=19$ različitih nizova, pri čemu 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 će biti intenzivno korištena i u narednim pod-zadacima. Prije poziva funkcije, svakako resetirajte generator slučajnih brojeva na početnu vrijednost naredbom `rng('default')`. Vase rješenje će biti provjereno i bodovano u narednom pod-zadatku.

IZVJEŠTAJ: Dokumentirajte način generiranja opservacija.

```
%%
rng('default')
T = 188; % duljina svakog niza
nex = 19; % broj opservacijskih nizova
dataRG = dhmm_sample(prior0, transmat0, obsmat0, nex, T);
```

Pod-zadatak 8 - Odredjivanje dugotrajne statistike osmotrenih simbola i usporedba s njihovim teorijskim ocekivanjima

(8a) [1 bod] Za nizove koji su generirani u pod-zadatku 7, potrebno je eksperimentalno odrediti vjerojatnosti osmatranja svih izlaznih simbola koristenjem slicnih primjera iz uputa. Za prvu osmotrenu sekvencu iz proslog pod-zadatka upisite broj osmatranja svakog izlaznog simbola, od 1 do 6, kojeg cete naci funkcijom hist:

--	--	--	--	--	--

%%

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

Rezultat:

```

hm =

    51    49    43    42    56    63    56    40    52    37    47    41    62    44    60    49    37    39    52
    29    25    29    21    21    23    18    29    25    37    20    28    25    26    23    25    31    27    25
    52    40    51    49    38    34    48    43    46    60    56    47    49    33    39    55    39    59
    16    11    26    23    19    18    17    19    22    9    20    21    20    21    25    19    23    17    15
    33    55    27    47    41    36    42    45    33    47    31    33    22    38    34    39    34    51    29
    7     8    12     6    13    14     7    12    13    12    10     9    12    10    13    17     8    15     8
fx >>

```

(8b) [1 bod] Potrebno je odrediti teorijska ocekivanja dugotrajnih vjerojatnosti osmatranja izlaznih simbola. Pri tome, prvo odredite stacionarnu distribuciju stanja (π_{stac}) uzastopnim mnozenjem zadane prijelazne matrice A same sa sobom i to T puta, te zatim temeljem ove dugotrajne statistike vjerojatnosti stanja modela i matrice izlaznih vjerojatnosti osmatranja B, odredite ocekivane stacionarne vjerojatnosti osmatranja svih izlaznih simbola (1 do 6), a sve sukladno primjeru iz uputa. Za provjeru tocnosti vasih rjesenja, upisite dugotrajnu vjerojatnost stanja 3 modela, $p(q=3)$ kao i dugotrajnu vjerojatnost osmatranja izlaznog simbola 1, $p(o=1)$:

0.333	
-------	--

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

```
a0(1,:)*obsmat0
```

rezultat:

```

a0 =

    0.3333    0.3333    0.3333
    0.3333    0.3333    0.3333
    0.3333    0.3333    0.3333

ans =

    0.2583    0.1417    0.2333    0.1000    0.2083    0.0583

```

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 $p(q=3)=0.3333$ ($p(q=3)=0.3333$), pokazuju koliko će svaki od mogućih stanja biti zastupljen u dugoročnom smislu, dok dugotrajne vjerojatnosti izlaznih simbola, kao $p(o=1)=0.2583$ ($p(o=1)=0.2583$), 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.

(8c) [1 bod] Odredite empirijske dugotrajne vjerojatnosti osmatranja simbola (pomocu funkcije hist) i to usrednjavanjem broja pojava simbola preko svih nex eksperimenata, te ih usporedite s upravo izracunatim ocekivanim dugotrajnim statistikama izlaznih simbola. Upisite najveći apsolutni iznos razlike izmedju empirijskih i teorijskih vjerojatnosti izlaznih simbola maksimiziran preko svih 6 izlaznih simbola :

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:

```
ans =  
-0.0008  
-0.0053  
0.0133  
0.0011  
-0.0076  
-0.0007
```

Pod-zadatak 9 - Izracun log-izvjesnosti osmatranja pojedinačnih generiranih opservacija temeljem zadanog modela

(9a) [1 bod] Za svaki od slučajnih nizova koji su generirani u pod-zadatku 7 potrebno je izracunati log-izvjesnost osmatranja uz zadani model, tj. uz isti model koji je koristen za generiranje ovih osmatranja. Nakon toga izracunajte najveću, najmanju i srednju vrijednost log-izvjesnosti usrednjenu preko svih nex osmotrenih nizova, te upišite dobivene rezultate u naredna tri polja (max, min i mean):

```
% Izracunaj u petlji log-izvjesnosti svakog niza  
nex2=size(dataRG,1); % Broj eksperimenata  
l1m2=zeros(nex2,1); % Stupac log-izvjesnosti  
for i=1:nex2,  
l1m2(i)=dhmm_logprob(dataRG(i,:), prior0, transmat0, obsmat0);  
end;  
  
l1m2  
  
max_value = max(l1m2); % Najveća vrijednost  
min_value = min(l1m2); % Najmanja vrijednost  
mean_value = mean(l1m2); % 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.

Pod-zadatak 10 - Provedite postupak treniranja parametara HMM modela

(10a) [2 boda] Temeljem svih nizova osmatranja koji su generirani u pod-zadatku 7, potrebno je izracunati dva nova HMM modela primjenom funkcije `dhmm_em`. **Vazno:** u oba slucaja ogranicite broj iteracija EM postupka na najvise 200, a prag relativne promjene izvjesnosti u odnosu na proslu iteraciju za zavrsetak postupka postavite na $1E-6$.

Za prvi HMM model inicijalizacija parametara modela za pocetnu iteraciju EM postupka treba biti potpuno slucajna (prema uputama), uz prethodno **resetiranje** generatora pseudo-slucajnih brojeva na pocetnu vrijednost. Za drugi HMM model za inicijalizaciju EM postupka iskoristite parametre zadanog modela. Tocnost vasesg izracuna parametara modela verificirat ce se u narednom pod-zadatku.

Za brzu provjeru upisite broj iteracija koji je bio potreban za estimaciju parametara HMM modela EM postupkom za oba modela (prvi i drugi):

%%

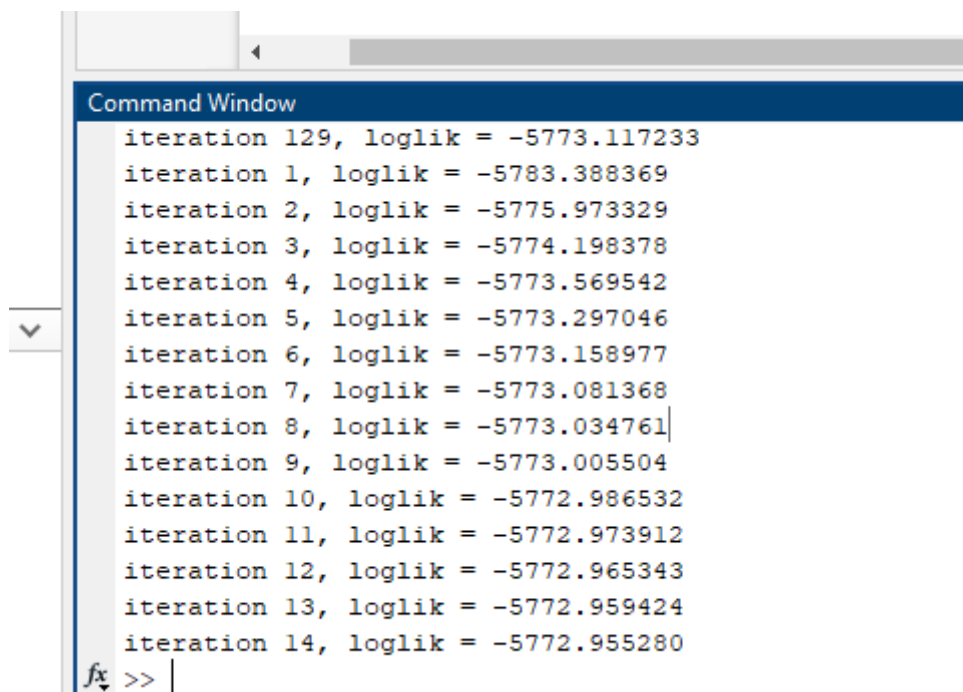
```
rng('default')
prior1 = normalise(rand(Q,1));
transmat1 = mk_stochastic(rand(Q,Q));
obsmat1 = mk_stochastic(rand(Q,O))
```

```
[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:



```
Command Window
iteration 129, loglik = -5773.117233
iteration 1, loglik = -5783.388369
iteration 2, loglik = -5775.973329
iteration 3, loglik = -5774.198378
iteration 4, loglik = -5773.569542
iteration 5, loglik = -5773.297046
iteration 6, loglik = -5773.158977
iteration 7, loglik = -5773.081368
iteration 8, loglik = -5773.034761
iteration 9, loglik = -5773.005504
iteration 10, loglik = -5772.986532
iteration 11, loglik = -5772.973912
iteration 12, loglik = -5772.965343
iteration 13, loglik = -5772.959424
iteration 14, loglik = -5772.955280
fx >>
```

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.

Pod-zadatak 11 - Usporedna evaluacija zadanog modela, slučajnog modela i treniranih modela na istim podacima koji su korišteni za trening

(11a) [2 boda] Potrebno je usporediti uspješnost modeliranja opservacijskih nizova generiranih u pod-zadatku 7 sa svim raspoloživim HMM modelima, izračunom log-izvjesnosti osmatranja svih generiranih nizova funkcijom `dhmm_logprob`. Kao "los" model za usporedbu, potrebno je koristiti HMM model s potpuno slučajnim parametrima, koji je korišten za inicijalizaciju prvog od dva nova "optimalna" HMM modela u proslom pod-zadatku (**Vazno:**, ... pazite da su parametri ovog slučajnog modela uistinu generirani odmah nakon inicijalizacije generatora pseudo-slučajnih brojeva).

U četiri polja upišite dobivene log-izvjesnosti osmatranja ovim redom: za zadani model, za "losi" slučajni model, za prvi novi HMM model sa slučajnom inicijalizacijom i konačno za drugi novi HMM model sa zadanom inicijalizacijom:

--	--	--	--

%%

```
l100=dhmm_logprob(dataRG, prior0, transmat0, obsmat0)
```

```
l111=dhmm_logprob(dataRG, prior1, transmat1, obsmat1)
```

```
l122=dhmm_logprob(dataRG, prior222, transmat222, obsmat222)
```

```
l133=dhmm_logprob(dataRG, prior333, transmat333, obsmat333)
```

Rezultat:

```
l100 =  
-5.7834e+03  
  
l111 =  
-6.1403e+03  
  
l122 =  
-5.7731e+03  
  
l133 =  
-5.7730e+03
```

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 , π) sa zadanim modelom. Kako objašnjavate razlike parametara ovih modela? Je li provjera estimiranog modela na istim podacima 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 podacima, pokazuju slične vrijednosti ukupne log-izvjesnosti kao zadani model, što ukazuje na dobru prilagodbu podacima.

Što se tiče razlika u parametrima (A , B , π), trenirani modeli s inicijalno slučajnim i zadanim 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 podacima 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 podacima kako bi se osigurala njegova generalizacija.