

Laboratorijska vježba za skrivene Markovljeve modele

U sklopu ove laboratorijske vježbe provodi se eksperiment sa skrivenim Markovljevim modelom, te se izračuni provode korištenjem HMM Toolbox-a za Matlab. Prije samostalne izrade vježbe nužno je proučiti „Upute za rad sa HMM modelima u Matlabu“, posebno one dijelove koji se odnose na konkretne primjere rada s navedenim Toolbox-om. Zadatci iz vježbe sadržajno pratite strukturu poglavlja iz uputa.

Personalizacija zadataka za laboratorijsku vježbu, način unosa rješenja za automatsku provjeru i priprema završnog izvješća vježbe

Radi ostvarenja mogućnosti automatizirane provjere i bodovanja vaših rezultata izvođenja ove laboratorijske vježbe u Matlabu, koristi se Moodle. Iako svi studenti za vježbu rješavaju kasnije u tekstu opisani HMM model, svaki student će u kroz Moodle pristupiti svojoj personaliziranoj inačici zadatka.

Osim upisa traženih numeričkih rješenja u Moodle za svaki pod-zadatak u svrhu automatizirane provjere vaših rezultata, cijeli postupak rješavanja ove laboratorijske vježbe potrebno je dokumentirati kroz izvještaj i Matlab skriptu. Izvještaj se predaje kroz Moodle okruženje u .pdf formatu, a odvojeno kao prilog izvještaju u .zip formatu predaje se Matlab skripta. Navedene je potrebno izraditi na sljedeći način.

Prilikom izrade izvještaja za svaki pod-zadatak potrebno je:

- 1) naznačiti o kojem se pod-zadatku radi (odgovarajuća numeracija),
- 2) kopirati personalizirani tekst zadatka iz Moodle okruženja,
- 3) kopirati iz Matlab-a isječak programskog koda korištenog za rješavanje pod-zadatka,
- 4) istaknuti rezultate izvršavanja koda,
- 5) odgovoriti na pitanja, komentirati rezultate i sl. (oznaka **IZVJEŠTAJ**) zavisno o tekstu pod-zadatka.

Matlab skripta treba sadržavati:

- 1) odgovarajuće inicijalizacije svih potrebnih varijabli, koje će ovisiti o zadanim parametrima modela, odnosno o konkretnim/specifičnim pitanjima u individualiziranim Moodle pod-zadacima,
- 2) sve korištene Matlab naredbe za izračunavanje rješenja pojedinih pod-zadataka,
- 3) detaljne komentare odsječaka korištenog koda.

Kao primjer pravilno komentirane Matlab skripte, u nastavku je prikazan njen zajednički početak koji obuhvaća inicijalizaciju matrice vjerojatnosti prijelaza i vektora početne vjerojatnosti stanja u skladu s opisanim eksperimentom:

```
clear;
pack;
addpath(genpath('HMMall')) % Dodaj put na biblioteku funkcija

% =====
% 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)
% određen 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= ...; % Ovdje definirajte M iz vaseg 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;

```

Opis slučajnog eksperimenta korištenog u vježbi

Odabrani stohastički eksperiment je temeljen na bacanju tri pristrane igrake kocke s mogućim ishodima bacanja od 1 do 6. Ove kocke su „podešene“ na način da će prva kocka u prosjeku, u pola svih bacanja dati broj „1“, da će druga kocka također u pola svih bacanja dati broj „3“ i da će treća na jednak način u pola svih bacanja dati broj „5“. Vjerojatnosti preostalih ishoda bacanja bitno su manje, ali nisu jednake, nego su ovisne o pojedinoj kocki za svaki od preostalih 5 mogućih brojeva. Ishod bacanja kocke vidljiv je gledateljima, dok ono što im nije vidljivo jest informacija koja od tri pristrane kocke je korištena pri pojedinom bacanju, pa stoga indeks korištene kocke (1., 2. ili 3. kocka) predstavlja skriveno stanje ovog modela.

Promjene skrivenih stanja definirane su prijelaznom matricom modela, pri čemu neki prijelazi u ovom eksperimentu nisu dozvoljeni, tj. imaju nultu vjerojatnost. Konkretno, ako je u aktualnom bacanju korištena prva kocka, tada će u narednom bacanju biti korištena ta ista kocka, ili druga kocka, ali ne može biti odabrana treća kocka. Analogno, ako je model u stanju bacanja druge kocke, može ostati u tom stanju, ili prijeći u stanje treće kocke, ali se ne može vratiti u stanje prve kocke. Konačno kada uđe u stanje treće kocke, može ciklički prijeći u stanje prve kocke, ili ostati u postojećem stanju, ali se ne može vratiti u stanje druge kocke. Zaključno, ovaj model nužno ciklički prolazi kroz skrivena stanja 1->2->3->1, uz mogućnost zadržavanja u aktualnom stanju. Vjerojatnost zadržavanja istog stanja iznosi $(M-1)/M$, dok

vjerojatnost prijelaza u naredno cikličko stanje iznosi $1/M$.

IZVJEŠTAJ: Skicirajte dijagram stanja za ovako opisan model.

U svrhu definiranja početnog stanja ovog modela, dakle samo za prvo bacanje, koristi se dodatna nepristrana kocka na ovaj način: ako je ishod bacanja te nepristrane kocke „1“ za prvo bacanje će se koristiti prva pristrana kocka, za ishode „2“ ili „3“ koristit će se druga kocka, dok se u slučaju preostala tri ishoda, „4“, „5“ i „6“, kao prva baca treća pristrana kocka.

U nastavku su detaljno opisani pojedini pod-zadaci koje morate riješiti za vaš konkretni primjer uz dodatne napomene vezane uz pripremu izvještaja. Ovaj opis je općenit, dok su specifične vrijednosti i konkretni traženi rezultati definirani u individualiziranim Moodle zadacima svakog studenta. Kao primjer, za svakog studenta definirane su drugačije izlazne vjerojatnosti osmatranja svih mogućih ishoda za sve tri pristrane kocke, odnosno različit će biti i parametar M koji određuje vjerojatnost prijelaza modela u novo stanje. Također, svaki student će dobiti i individualizirana pod-pitanja zadatka ove laboratorijske vježbe, na koja će morati odgovoriti upisom točne numeričke vrijednosti rješenja u odgovarajuće polje, a sve će detaljnije biti objašnjeno u nastavku ovog dokumenta.

Pod-zadatak 1 - Cjelovito definiranje HMM modela

Temeljem zadanih učestalosti pojedinih ishoda bacanja pristranih kocki i temeljem zadanog parametra M u vašem Moodle zadatku, potrebno je dopuniti gornju Matlab skriptu kako bi cjelovito opisali zadani HMM model ovog eksperimenta uključujući i matricu vjerojatnosti osmatranja izlaznih simbola.

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

U drugom pod-zadatku u Moodle-u zadana su dva konkretna osmotrena niza izlaznih simbola (ishodi bacanja pristranih kocki). Ubacite ove nizove u Matlab skriptu kao inicijalizaciju odgovarajućeg polja s dva retka (kojeg ćete koristiti i u narednim pod-zadacima), te korištenjem odgovarajuće funkcije izračunajte log-izvjesnosti osmatranja oba zadana niza.

Upišite dobivene log-izvjesnosti prvog i drugog niza u tražena polja u Moodle.

IZVJEŠTAJ: Možete li usporedbom zadanih nizova obrazložiti razlog zbog kojeg je drugi niz manje izvjestan od prvog? Opišite riječima.

Izračunajte i upišite u Moodle koliko puta je drugi niz manje izvjestan od prvog.

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

Samo za prvi osmotreni niz iz drugog pod-zadatka potrebno je izračunati unaprijednu (alfa) i unazadnu (beta) vjerojatnost za sva tri skrivena stanja i za sve trenutke osmatranja. U svrhu provjere vašeg rješenja, u individualiziranom pod-zadatku u Moodleu traži se da upišete izračunate vjerojatnosti alfa i beta za neke slučajno odabrane vremenske trenutke i neka slučajno odabrana stanja modela, pa upišite tražene vjerojatnosti koje ste odredili funkcijom *fwdback* u odgovarajuća polja u eksponencijalnom formatu zapisa broja. **Važno:** 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.

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.

Pod-zadatak 4 - Dekodiranje skrivenih stanja pomoću Viterbi algoritma

Potrebno je primjenom funkcije *viterbi_path* odrediti najizvjesniji niz skrivenih stanja modela također samo za prvi osmotreni niz iz drugog pod-zadatka. U svrhu verifikacije dobivenog rješenja, potrebno je u Moodle upisati dekodirana stanja modela za prva tri i za zadnja tri vremenska koraka prve opservacije.

Pod-zadatak 5 - Određivanje log-izvjesnosti osmatranja uzduž dekodiranih Viterbi puteva

Ponovite određivanje Viterbi niza stanja i za drugi osmotreni niz iz pod-zadatka 2, te za oba niza izračunajte log-izvjesnosti osmatranja, ali samo uzduž dekodiranih „optimalnih“ Viterbi puteva.

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.

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.

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

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 rješenje iz trećeg pod-zadatka u kojem ste odredili sve vjerojatnosti modela, ali za cjelovit niz.

IZVJEŠTAJ: Objasnite kako ste dobili izvjesnost osmatranja skraćenog niza.

Upišite u Moodle u eksponencijalnom zapisu koliko iznosi izvjesnost (ne log-izvjesnost) osmatranja prva četiri izlazna simbola. Nakon toga ponovno odredite Viterbi put, ali za ovu skraćenu opservacijsku sekvencu, te izračunajte i u Moodle upišite koji udio izvjesnosti osmatranja (normirano na 1) se ostvaruje uzduž Viterbi puta u odnosu na sve moguće puteve stanja modela.

IZVJEŠTAJ: Jeste li za nalaženje ovog Viterbi rješenja skraćenog niza smjeli koristiti rješenje iz pod-zadataka 4 i 5? Obrazloži odgovor.

Za provjeru rezultata, u Moodle upišite nađeni Viterbi put stanja za ova prva četiri osmotrena simbola prvog niza.

Konačno, potrebno je izračunati izvjesnost osmatranja prva četiri izlazna simbola, ali uzduž svih mogućih pojedinačnih puteva rešetke stanja, prema primjeru iz uputa. Upišite u Moodle koliko ukupno ima ovih pojedinačnih puteva stanja. Temeljem izračunatih izvjesnosti pojedinačnih puteva stanja, odredite koliko puteva od svih njih uopće nisu mogući, tj. u Moodle upišite broj puteva koji imaju nultu izvjesnost osmatranja skraćenog niza.

IZVJEŠTAJ: Obrazložite razlog.

Sortirajte puteve od najizvjesnijih prema najmanje izvjesnima te u Moodle upišite koji udio ukupne izvjesnosti osmatranja (normirano na 1) se kumulativno ostvaruje uzduž prvih pet najizvjesnijih puteva ove sortirane liste.

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?

Pod-zadatak 7 - Generiranje opservacija za zadani model

Potrebno je generirati višestruke slučajne nizove osmotrenih izlaznih simbola (*nex* različitih nizova), pri čemu svaki niz treba biti duljine T vremenskih uzoraka. Za generiranje podataka koristiti funkciju *dhmm_sample* u skladu s uputama, uz parametre HMM modela iz vašeg individualnog pod-zadatka 1. U vašem individualiziranom Moodle pod-zadatku, zadani su željeni broj nizova (*nex*) i njihova duljina (T). Sačuvajte ovu matricu opservacija jer će biti intenzivno korištena i u narednim pod-zadacima.

Važno: kako bi algoritam automatizirane provjere mogao jednoznačno provjeriti vaše rezultate izvođenja, nužno je da se pri provjeri koriste identični generirani nizovi kao što ih generirate i koristite vi prilikom vašeg izvođenja. Svaki novi poziv ove funkcije *dhmm_sample*, generirat će potpuno novi skup slučajnih osmatranja, pri čemu ova funkcija koristi ugrađeni Matlab-ov generator pseudo-slučajnih brojeva. Kako bi osigurali istovjetnost generiranog niza s očekivanim, potrebno je inicijalizirati ovaj generator pseudo-slučajnih brojeva na identičnu početnu vrijednost. To svakako učinite naredbom *rng('default')* prije poziva ove funkcije za generiranje opservacija. Vaša uspješnost generiranja opservacija bit će provjerena u narednom pod-zadatku.

IZVJEŠTAJ: Dokumentirajte način generiranja opservacija.

Pod-zadatak 8 - Određivanje dugotrajne statistike osmotrenih simbola i usporedba s njihovim teorijskim očekivanjima

Za nizove koji su generirani u pod-zadatku 7, potrebno je eksperimentalno odrediti vjerojatnosti osmatranja svih izlaznih simbola korištenjem sličnih primjera iz uputa. Za prvu osmotrenu sekvencu iz pod-zadatka 7, koja je duljine T simbola, u Moodle upišite broj osmatranja svakog izlaznog simbola, od 1 do 6, kojeg ćete naći funkcijom *hist*.

Nakon toga, potrebno je odrediti teorijska očekivanja dugotrajnih vjerojatnosti osmatranja izlaznih simbola. Pri tome, prvo odredite stacionarnu distribuciju stanja (*pi_stac*) uzastopnim množenjem 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 očekivane stacionarne vjerojatnosti osmatranja svih izlaznih simbola (1 do 6), a sve sukladno primjeru iz uputa. Za provjeru točnosti vaših rješenja, u Moodle upišite dugotrajnu vjerojatnost određenog traženog stanja modela, kao i dugotrajnu vjerojatnost osmatranja određenog traženog izlaznog simbola. Traženo stanje i traženi simbol bit će navedeni u vašem individualiziranom pod-zadatku u Moodle-u.

IZVJEŠTAJ: Diskutirajte dobivene dugotrajne vjerojatnosti pojedinih stanja, odnosno izlaznih simbola. Kako bi izgledao degenerirani HMM model s jednakim dugotrajnim statistikama opservacija izlaznih simbola?

U konačnici, odredite empirijske dugotrajne vjerojatnosti osmatranja simbola (pomoću funkcije *hist*) i to usrednjavanjem broja pojava preko svih *nex* eksperimenata.

IZVJEŠTAJ: Usporedite ih s upravo izračunatim očekivanim dugotrajnim statistikama izlaznih simbola.

U Moodle upišite najveći apsolutni iznos razlike između empirijskih i teorijskih vjerojatnosti izlaznih simbola maksimiziran preko svih 6 izlaznih simbola.

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

Za svaki od slučajnih nizova koji su generirani u pod-zadatku 7 potrebno je izračunati log-izvjesnost osmatranja uz zadani model, tj. uz isti model koji je korišten za generiranje ovih osmatranja. Nakon toga izračunajte najveću, najmanju i srednju vrijednost log-izvjesnosti usrednjenu preko svih n_{ex} osmotrenih nizova, te upišite dobivene rezultate u Moodle.

IZVJEŠTAJ: Zašto se izvjesnosti pojedinih nizova razlikuju?

Pod-zadatak 10 - Provedite postupak treniranja parametara HMM modela

Temeljem svih nizova osmatranja koji su generirani u pod-zadatku 7, potrebno je izračunati dva nova HMM modela primjenom funkcije `dhmm_em`. U oba slučaja ograničite broj iteracija EM postupka na najviše 200, a prag relativne promjene izvjesnosti u odnosu na prošlu iteraciju za završetak postupka postavite na $1E-6$ dodavanjem ovih opcija u pozive funkcije `dhmm_em`:
`..., 'max_iter', 200, 'thresh', 1E-6);`

Za prvi HMM model inicijalizacija parametara modela za početnu iteraciju EM postupka treba biti potpuno slučajna, kao što je objašnjeno u uputama u programskom odsječku koji započinje s komentarom „% initial guess of parameters“. **Važno:** kako bi i automatizirana verifikacija vaših rezultata koristila identične slučajne parametre inicijalnog modela (`prior1`, `transmat1` i `obsmat1`), prije izvođenja ovog bloka naredbi ponovno resetirajte generator pseudo-slučajnih brojeva na početnu vrijednost naredbom `rng('default')`.

Za drugi HMM model za inicijalizaciju EM postupka iskoristite parametre zadanog modela (istog koji je korišten za generiranje osmatranja za treniranje), kako bi tako novi „fino-prilagođen“ model još bolje opisivao baš ove konkretne zadane opservacije.

U Moodle upišite broj iteracija koji je bio potreban za estimaciju parametara HMM modela EM postupkom za oba modela.

IZVJEŠTAJ: Obrazložite razliku broja iteracija.

Sama točnost vašeg izračuna parametara modela verificirat će se u narednom pod-zadatku.

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

Jednako kao u uputama, 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`. Za zadani model to je već učinjeno u pod-zadatku 9, izračunavanjem log-izvjesnosti svakog pojedinog generiranog niza. Sada to želimo ponoviti za zadani model kao i dva nova trenirana HMM modela iz pod-zadatka 10, koji se međusobno razlikuju samo po načinu inicijalizacije postupka treniranja, ali za sve generirane nizove iz pod-zadatka 7 zajedno. Kao zadnji model za usporedbu, potrebno je koristiti HMM model s potpuno slučajnim parametrima, koji je opisan s vjerojatnostima (`prior1`, `transmat1` i `obsmat1`), iz istog pod-zadatka 10. To je zapravo slučajni model koji je korišten za inicijalizaciju prvog od dva nova „optimalna“ HMM modela.

Važno: pazite da su parametri ovog slučajnog modela uistinu generirani odmah nakon inicijalizacije generatora pseudo-slučajnih brojeva kako bi bili istovjetni onima korištenim u automatiziranoj provjeri vaših rezultata.

U Moodle upišite dobivene log-izvjesnosti osmatranja za zadani model, za slučajni model i za dva nova HMM modela određena u pod-zadatku 10: optimalni model sa slučajnom inicijalizacijom i optimalni model sa zadanom inicijalizacijom.

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?