SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

.

Skriveni Markovljevi modeli

Otkrivanje znanja u skupovima podataka

Tessa Bauman

Zagreb, kolovoz 2023.

SADRŽAJ

1. Uvod

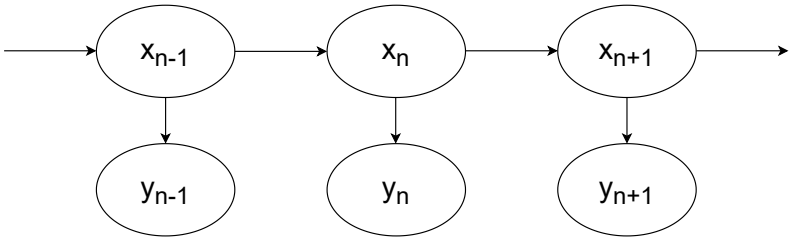
Statistički modeli su važan alat za proučavanje i predviđanje stvarnih procesa. U ovom seminaru bit će obrađen skrivenim Markovljevim modelom (Hidden Markov model - HMM). Ovaj model pretpostavlja postojanje skrivenog procesa koji se opisuje Markovljevim lancem i izravno utječe na niz podataka koje promatramo. Opaženi podaci tumače se kao posljedica kretanja tog lanca. Ono što izdvaja skriveni Markovljev model od drugih statističkih modela je to što pretpostavlja postojanje uzročnika, ali informacije o njemu nisu izravno dostupne. Cilj ovog seminara je objasniti model i uvesti glavne probleme istog.

Markovljev model je koncept koji je formaliziran sredinom 19. stoljeća u radovima Andreija Markova o Markovljevim lancima. Ključni napredak u razvoju HMM-a dogodio se 1960-ih i 1970-ih godina (1; 2) s konstrukcijom forward, Viterbi i Baum-Welch algoritama. Za učinkovitu primjenu modela, neophodno je pronaći parametre koji najbolje odgovaraju promatranim podacima i otkriti skriveni niz koji ih generira, što su problemi koje su uspješno riješili spomenuti algoritmi. Danas, HMM ima široku primjenu u obradi prirodnog jezika, prepoznavanju govora, biomedicinskoj analizi i drugim područjima. Nastavlja se istraživanje kako unaprijediti performanse i primjenu ovog modela u raznim dinamičkim procesima.

U seminaru će se obraditi teoretski koncepti modela, glavni problemi te rješenja istih. Nadalje, u seminaru će biti pokazana primjena modela u financijskom problemu detekcije režima. Financijski podaci često imaju složenu strukturu s promjenjivim režimima ponašanja, što čini HMM dobrim kandidatom za analizu takvih podataka.

2. Skriveni Markovljevi modeli

Skriveni Markovljevi modeli definirani su pomoću 2 procesa, od kojih je jedan skriveni. Pretpostavka modela je da taj skriveni lanac generira opažanja koja su nam dostupna. Slika 2.1 prikazuje shematski prikaz HMM-a. Proces sa skrivenim stanjima X u svakom koraku n emitira opažanje Y.



Slika 2.1: Shematski prikaz modela

HMM pretpostavlja da je skriveni lanac Markovljev te je za definiciju modela prvo potrebno objasniti što to znači.

2.1. Markovljevi lanci

Definicija 2.1.1 *Slučajni proces $X = (X_n : n \geq 0)$ s vrijednostima u skupu S je Markovljev lanac ako vrijedi*

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

za svaki $n \in \mathbf{N}$ i za sve i_0,...,i_{n-1}, i, j $\in$ S.

Svojstvo (2.1) zove se Markovljevo svojstvo (3) te ono govori da vjerojatnost da je proces u trenutku $n+1$ u proizvoljnom stanju ovisi o trenutnom stanju u trenutku n, ali ne i o prošlim stanjima.

Neka su

- $p_{ij} := \mathbf{P}(X_{n+1} = j \,|\, X_n = i)$, za svaki $n \geq 0,\ i, j \in S$,
- $\pi_i := \mathbf{P}(X_0 = i)$, za svaki $i \in S$.

Matrica $P = (p_{ij} : i, j \in S)$ naziva se *prijelazna matrica*, a vjerojatnosna distribucija $\Pi = (\pi_i : i \in S)$ *početna distribucija* Markovljevog lanca $X = (X_n : n \geq 0)$.

2.2. Skriveni Markovljevi modeli

Definicija 2.2.1 *Neka je* $X = (X_n : n \geq 0)$ *Markovljev lanac na skupu stanja S, s matricom prijelaza P = (*p_{ij} *: i, j* $\in$ *S) i početnom distribucijom* Π *= (*π_i *: i* $\in$ *S). Slučajan proces* $(X, Y) = \{(X_n, Y_n) : n \geq 0)\}$ *zovemo skriveni Markovljev model (HMM) ako vrijedi:*

- $\mathbf{P}\left(Y_n \in A \mid X_1 = x_1, \ldots, X_n = x_n\right) = \mathbf{P}\left(Y_n \in A \mid X_n = x_n\right)$*, za svako* $n \geq 1$*,* $x_1, \ldots, x_n$ *i svaki skup A.*

Stanja procesa X_n nazivaju se skrivena stanja, a $\mathbf{P}\left(Y_n \in A \mid X_n = x_n\right)$ naziva se vjerojatnost emisije. U standardnom tipu skrivenog Markovljevog modela koji se ovdje razmatra, prostor stanja skrivenih varijabli je diskretan, dok distribucija opaženog procesa Y s obzirom na skriveno stanje X, tj. distribucija $Y|X$, može biti diskretna ili kontinuirana (obično iz Gaussove distribucije) (4).

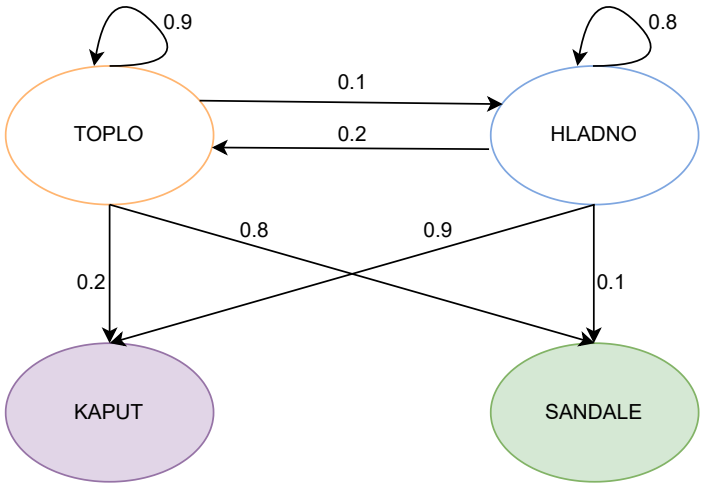
Skriveni Markovljev model možemo okarakterizirati s uređenom petorkom H=(S, O, P, Ψ, Π) gdje S, P i Π predstavljaju parametre Markovljevog lanca $X = (X_n : n \geq 0)$, a O i Ψ sljedeće:

- O skup emisijskih simbola tj. vrijednosti niza slučajnih varijabli Y,
- Ψ= $(\psi_{jk} : j \in S, k \in O)$ matrica emisijskih vjerojatnosti, uz

$$\psi_{jk} := \mathbf{P}(Y_n = k \,|\, X_n = j).$$

2.3. Pokazni primjer

Pretpostavimo da gledamo kronološki poredane fotografije na izložbi i zanima nas kakvo je bilo vrijeme u trenutku kada su nastale. Budući da nam je tadašnja vremenska prognoza nepoznata tretiramo ju kao skrivena stanja modela i razlikujemo dva stanja: toplo i hladno vrijeme. Iako nam je stanje nepoznato, na fotografijama su ljudi po čijoj odjeći možemo zaključiti nešto o vremenu.



Slika 2.2: Shematski prikaz modela iz primjera

Detaljnije, niz vremenskih prilika promatramo kao skriveni Markovljev lanac $X = (X_n : n \geq 0)$ koji emitira nošenu odjeću tj. opažanja $Y = (Y_n : n \geq 0)$. Uređenu petorku modela dakle čine:

- $S = \{t, h\}$ skup stanja lanca koji se sastoji od toplog i hladnog vremena,
- $O = \{\text{kaput}, \text{sandale}\}$ skup emisijskih simbola,
- P matrica prijelaza

$$\begin{pmatrix} 9/10 & 1/10 \\ 2/10 & 8/10 \end{pmatrix},$$

  gdje je $p_{11} = \mathbf{P}(X_{n+1} = t \,|\, X_n = t)$ vjerojatnost da je nakon toplog vremena opet bilo toplo vrijeme, $p_{21} = \mathbf{P}(X_{n+1} = t \,|\, X_n = h)$ vjerojatnost da je nakon hladnog vremena došlo toplo i analogno dalje,

- Ψ matrica emisijskih vjerojatnosti

$$\begin{pmatrix} 2/10 & 8/10 \\ 9/10 & 1/10 \end{pmatrix}$$

  gdje je npr. $\psi_{11} = \mathbf{P}(Y_n = \text{kaput} \,|\, X_n = t)$ vjerojatnost da čovjek nosi kaput ako je vrijeme toplo.

2.4. Osnovni problemi modela

Pri korištenju ovog modela, prirodno se nameću neka pitanja. Sukladno gore navedenom primjeru, ona bi glasila:

1. Koliko dobro naš pretpostavljeni model opisuje situaciju?
2. Promatrajući odjeću, koji niz vremenskih prilika je najvjerojatniji?
3. Kako odrediti model koji maksimizira vjerojatnost danog niza odjeće?

Ova tri pitanja predstavljaju osnovne probleme modela.

2.4.1. Problem evaluacije

Kod evaluacijskog problema želimo saznati kolika je vjerojatnost da je model emitirao dani niz opažanja $k_1, k_2, ..., k_N$ tj.

$$\mathbf{P}(Y_1 = k_1, Y_2 = k_2, ..., Y_N = k_N \,;\, \text{model}). \tag{2.2}$$

Rješenje ovog problema omogućava uspoređivanje različitih modela. Od više ponuđenih modela (i naših pretpostavki o broju skrivenih stanja), može se odabrati onaj koji najbolje odgovara opažanjima. Dovoljno je usporediti vjerojatnosti (2.2) te odabrati model za koji se dobije najveća. Ovaj problem efikasno se riješava pomoću *forward* algoritma (1) i forward varijabli. Uz dana opažanja $k_1, k_2, ..., k_N \in O$, definira se:

$$\alpha_n(j) := \mathbb{P}(Y_1 = k_1, Y_2 = k_2, ..., Y_t = k_n, X_n = j), \text{ za svaki } n \leq N,\ j \in S.$$

Ovako definirani parametri obično se nazivaju *forward* varijablama. $\alpha_n(j)$ predstavlja vjerojatnost da se model u trenutku n nalazi u stanju j i da je emitirao prvih n elemenata niza opažanja. *Forward* algoritam radi tako da rekurzivno procjenjuje vjerojatnosti svih mogućih putanja kroz model te kumulativno kombinira te vjerojatnosti kako bi dobili ukupnu vjerojatnost promatranog niza. Algoritam se može opisati sljedećim koracima:

1. Inicijalizacija:
$$\alpha_1(j) = \pi_j \psi_{jk_1},\ j \in S$$
2. Rekurzija:
$$\alpha_n(j) = \sum_{i \in S} \alpha_{n-1}(i) p_{ij} \psi_{jk_n},\ j \in S, 1 \leq n \leq N$$
3. Kraj:
$$\mathbb{P}(Y_1 = k_1, Y_2 = k_2, ..., Y_N = k_N) = \sum_{i=1}^{N} \alpha_N(i).$$

2.4.2. Problem dekodiranja

Pretpostavimo da imamo niz opažanja te da su nam poznati parametri modela (početna distribucija skrivenih stanja, prijelazna i emisijska matrica). Problem dekodiranja odnosi se na pronalaženje navjerojatnijeg niza skrivenih stanja s obzirom na opažanja, tj.

$$X = \text{argmax}_x \, \mathbf{P}(X \mid Y).$$

Rješenje ovog problema dobije se pomoću Viterbijevog algoritma (1). Radi tako što rekurzivno procjenjuje vjerojatnosti svih mogućih skrivenih stanja na temelju opažanja i prethodno izračunatih vjerojatnosti te odabire najvjerojatniji niz stanja putem povratnog praćenja. Kroz ovu "Viterbijevu putanju", algoritam pruža optimalno dekodiranje uz pretpostavku da je skriveni Markovljev model poznat.

Za objašnjenje algoritma potrebno je uvesti sljedeću oznaku uz dana opažanja $k_1, k_2, ..., k_n \in O$:

$$\delta_n(i) := \max_{j_1, j_2, ..., j_{n-1} \in S} \mathbb{P}(X_1 = j_1, X_2 = j_2, ..., X_n = i, Y_1 = k_1, Y_2 = k_2, ..., Y_n = k_n).$$

Dakle, $\delta_n(i)$ je maksimalna vjerojatnost koju niz stanja duljine n može postići tako da završi u stanju i i emitira prvih n zadanih simbola $k_1, k_2, ..., k_n$. *Viterbijev* algoritam može se opisati sljedećim koracima:

1. Inicijalizacija:
$$\delta_1(i) = \pi_i \, \psi_{ik_1}, \;\; i \in S$$

2. Rekurzija:
$$\delta_n(j) = \max_{i \in S} \delta_{n-1}(i) \, p_{ij} \, \psi_{jk_n}, \;\; j \in S, \; 2 \leq n \leq N$$

3. Kraj:
$$P* = \max_{i \in S} \delta_N(i),$$

Parametar $\delta_n(i)$, $\max_{i \in S} \delta_N(i)$ daje najveću moguću traženu vjerojatnost. Cilj dekodiranja je pronaći niz stanja na kojem se ona postiže pa uz rekurzivno računanje vjerojatnosti *Viterbijevim* algoritmom dodatno treba pratiti skrivena stanja dobivenog puta.

2.4.3. Problem učenja

Problemi evaluacije i dekodiranja pretpostavljaju da su parametri modela poznati. Za razliku od njih, problem učenja se upravo fokusira na pronalazak najboljih parametara modela. Preciznije, potrebno je pronaći parametre početne distribucije, emisijske matrice te matrice prijelaza koji maksimiziraju opaženi niz Y:

$$(\Pi, P, \Psi) = \text{argmax}_{(\pi, p, \psi)} \, \mathbf{P}(Y \mid \text{model}).$$

Algoritam koji je razvijen za tu svrhu naziva se Baum-Welch algoritam (1), a ime je dobio po svojim kreatorima Leonardu E. Baumu i Lloydu R. Welchu. Baum-Welch postupak spada u algoritme maksimiziranja očekivanja (expectation-maximization (EM) algorithm) te je iterativan.

Uz sljedeće pomoćne oznake i jednakosti

- $\beta_n(i) := \mathbb{P}(Y_{n+1} = k_{n+1}, Y_{n+2} = k_{n+2}, ..., Y_N = k_N \mid X_n = i)$,
- $\gamma_n(j) := \mathbb{P}(X_n = j \mid Y_1 = k_1, Y_2 = k_2, ..., Y_N = k_N) = \frac{\alpha_n(j)\,\beta_n(j)}{\sum_{j \in S} \alpha_n(j)\,\beta_n(j)}$,
- $\xi_n(i,j) := \mathbb{P}(X_n = i, X_{n+1} = j \mid Y_1 = k_1, ..., Y_N = k_N) = \frac{\alpha_n(i)\,p_{ij}\,\psi_{jk_{n+1}}\,\beta_{n+1}(j)}{\sum_{i \in S} \alpha_n(i)\,\beta_n(i)}$,

glavni koraci algoritma su:

1. Inicijalizacija: parametri modela P, Ψ, Π se zadaju najčešće nasumično.
2. Računanje $\alpha_n(i)$ i $\beta_n(i)$.
3. Update korak:

$$\hat{\pi}_i = \gamma_1(i), \quad i \in S,$$

$$\hat{p}_{ij} = \frac{\sum_{n=1}^{N-1} \xi_n(i,j)}{\sum_{n=1}^{N-1} \gamma_n(i)}, \quad i,\, j \in S$$

$$\hat{\psi}_{jk} = \frac{\sum_{\substack{n=1 \\ Y_n = k}}^{N} \gamma_n(j)}{\sum_{n=1}^{N} \gamma_n(j)}, \quad j \in S,\, k \in O.$$

Dakle, u prvom koraku parametri se nasumično odabiru. Računaju se α i β pomoću kojih se računaju γ i ξ. U svakoj iteraciji dobiju se nove procjene parametara modela. Postupak se ponavlja do željene razine konvergencije i za svaku iteraciju k vrijedi

$$\mathbb{P}(Y \mid P^{(k)}, \Psi^{(k)}, \Pi^{(k)}) \geq \mathbb{P}(Y \mid P^{(k-1)}, \Psi^{(k-1)}, \Pi^{(k-1)}).$$

Na taj način vrijednost konvergira u (lokalni) maksimum funkcije vjerodostojnosti.

3. Primjena

3.1. Uvod u problem

U svijetu financijskih ulaganja izrazi "bull" i "bear" često se koriste za označavanje tržišnih uvjeta. Ovi izrazi opisuju kako tržišta financijskih instrumenata rade općenito - to jest, raste li ili pada vrijednost financijske imovine. Budući da smjer tržišta ima ogroman utjecaj na imovinu, bitno ga je razumjeti i pokušati detektirati. Nažalost, ne postoji fiksna definicija kada je tržište u kojem režimu, što čini problem zahtjevnim. "Bull" tržište, ili režim, je tržište koje je u usponu i gdje su uvjeti gospodarstva općenito povoljni. "Bear" režim vlada u gospodarstvu koje se povlači i gdje većini dionica pada vrijednost. Budući da su financijska tržišta pod velikim utjecajem stavova ulagača, ovi izrazi također označavaju kako ulagači misle o tržištu i gospodarskim trendovima koji iz njega proizlaze (6).

3.2. Podaci i postavke modela

Za primjenu modela korišteni su mjesečni povrati dionica i obveznica od 1871. do 2022. godine, dostupni online na web stranici Roberta Shillera. Dugački vremenski period je prikladan za detekciju režima, budući da se kroz dugi niz godina desilo puno promjena te je tržište prošlo kroz razne faze.

Budući da su povrati dionica i obveznica kontinuirani podaci, pretpostavit ćemo da dolaze iz Gaussove distribucije. Gaussovi skriveni Markovljevi modeli (2), vrsta su HMM-a u kojima skrivena stanja generiraju vrijednosti koje slijede Gaussovu distribuciju. Svi algoritmi za modele s kontinuiranim opažanjima slijede analogno kao i za modele s diskretnim opažanjima (4). Precizno, $Y|X \sim \mathcal{N}(\mu_{x_n}, \Sigma_{x_n})$, gdje x_n skriveno stanje. Dakle, u nastavku problema postojat će dva skupa parametara normalne distribucije ovisno o režimu: jedan za "bear", a drugi za "bull" režim.

Niz režima promatramo kao skriveni Markovljev lanac $X = (X_n : n \geq 0)$ koji emitira povrate dionica i obveznica tj. dvodimenzionalna opažanja $Y = (Y_n : n \geq 0)$. Preciznije:

- $S = \{0, 1\}$ skup stanja lanca koji se sastoji od 0-"bull" režima i 1-"bear" režima,
- O kontinuirani dvodimenzionalni skup povrata,

– $P = (p_{ij})_{i,j}$ matrica prijelaza gdje je npr. $p_{11} = \mathbf{P}(X_{n+1} = 0 \,|\, X_n = 0)$ vjerojatnost da je nakon "bull" tržišta opet krenulo "bull" tržište.

Za potrebe modeliranja i analize rezultata koristit će se programski jezik Python, zajedno s programskim paketima pandas, numpy i hmmlearn.

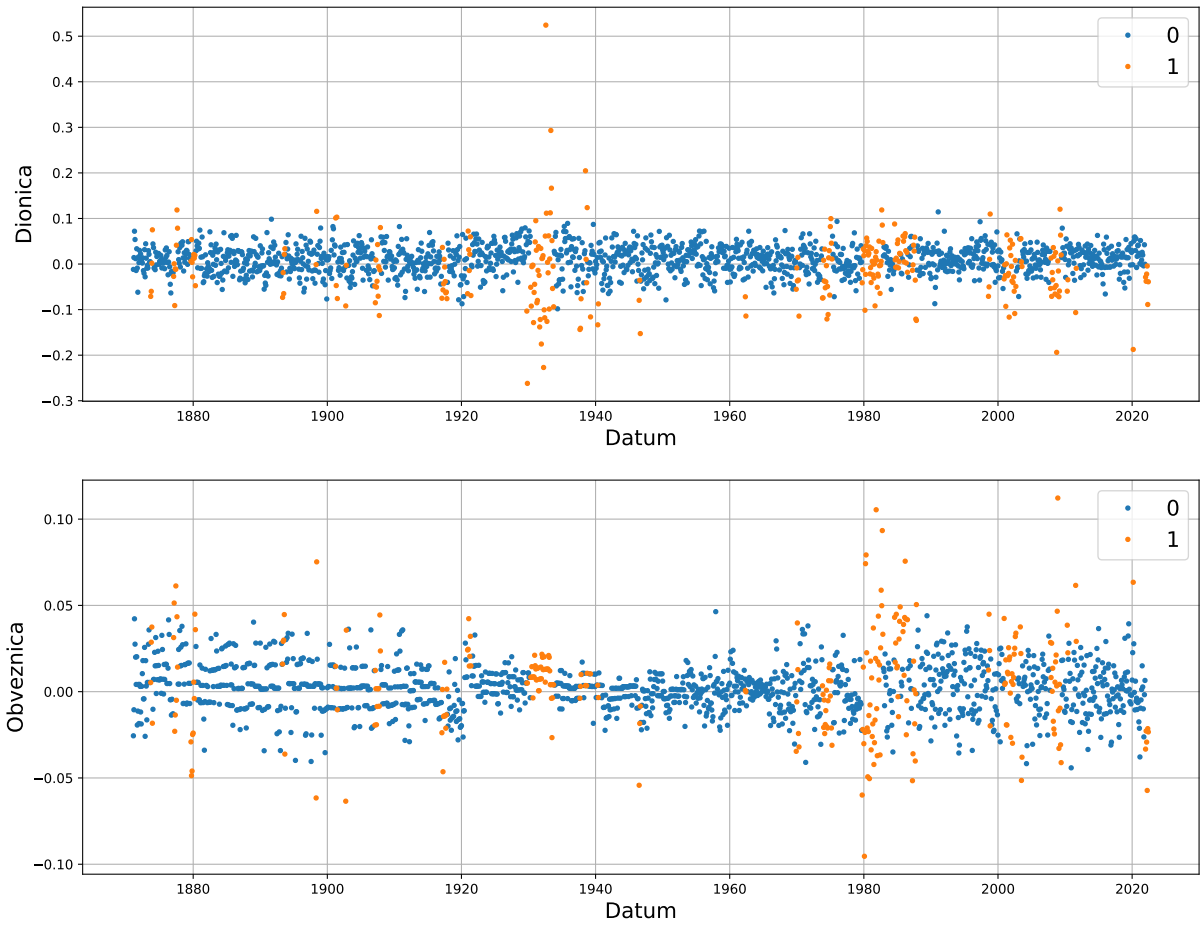
3.3. Rezultati

Pomoću 100000 iteracija Baum-Welch algoritma pronađeni su parametri modela. Na slici 3.1 su prikazani povijesni povrati dionica i obveznica te režimi detektirani pomoću Viterbijevog algoritma. Dobivena početna distribucija skrivenih stanja je:

$$\Pi = [0.8503026967528894, 0.1496973032471106].$$

Dobivena matrica prijelaza je:

$$P = \begin{bmatrix} 0.96439555 & 0.03560445 \\ 0.18235763 & 0.81764237 \end{bmatrix}$$



Slika 3.1: Grafički prikaz povrata i režima

Kako ne postoji fiksna definicija kada je tržište bilo u kojem režimu te pojmovi "bear" i "bull" više služe kao opisni pojmovi za uvjete na tržištu, ne može se definitno reći je li model prikladan ili nije. Međutim, "bear" režim je uglavnom vezan za krizna razdoblja. Ono što se jasno vidi na slici 3.1 je da su dobro detektirani krizni periodi: 1930e, 1980e, kriza 2008. te Covid kriza kada je tržište uistinu bilo u padu.

Za detaljniju analizu treba pogledati parametre normalnih distribucija vezanih uz svaki režim.

U režimu rasta, povrati dolaze iz $\mathcal{N}(\mu_0, \Sigma_0)$, tj. $Y_{|X=0} \sim \mathcal{N}(\mu_0, \Sigma_0)$, gdje su:

$$\mu_0 = [0.00989553, 0.00164454]$$

,

$$\Sigma_0 = \begin{bmatrix} 8.62184843 \times 10^{-4} & 8.93105088 \times 10^{-5} \\ 8.93105088 \times 10^{-5} & 1.77307706 \times 10^{-4} \end{bmatrix}$$

. U ovom režimu očita je pozitivna srednja vrijednost povrata kao što je i očekivano.

U režimu pada, distribucija povrata je sljedeća $Y_{|X=1} \sim \mathcal{N}(\mu_1, \Sigma_1)$, gdje su parametri distribucije:

$$\mu_1 = [-0.010467, 0.00490739],$$

$$\Sigma_1 = \begin{bmatrix} 5.30741231 \times 10^{-3} & 2.01451425 \times 10^{-4} \\ 2.01451425 \times 10^{-4} & 8.14336145 \times 10^{-4} \end{bmatrix}$$

Za "bear" režim dobiven je negativan prosječni povrat za dionicu što je također očekivano budući je krizni režim povezan s velikim padom vrijednosti dionica. Također, uspoređujući volatilnost s obzirom na režim, vidi se da je volatilnost veća i kod dionice i kod obveznice u kriznom režimu što prikazuje veću nesigurnost ulaganja u tim periodima.

Kako bi se procijenilo koliko dobro model paše podacima (problem evaluacije - *forward* algoritam) može se izračunati log-izglednost cijelog skupa. U slučaju modela s dva režima, log-izglednost iznosi 8476.03. Za usporedbu, provedeno je i treniranje modela sa samo jednim režimom. U tom slučaju log-izglednost iznosi 8197.48 što je manje. Pomoću tog rezultata može se naslutiti da je model s 2 režima uistinu vjerniji prikaz podataka.

4. Zaključak

U ovom seminaru obrađeni su osnovni koncepti i teorija skrivenih Markovljevih modela koji predstavljaju snažan alat u analizi vremenskih serija i sekvencijalnih podataka. HMM-ovi se temelje na ideji o skrivenim stanjima koja evoluiraju prema Markovljevom lancu, a njihovi izlazi (opažanja) se generiraju iz odgovarajućih diskretnih ili neprekidnih distribucija. Uvedeni su osnovni pojmovi HMM-a, uključujući prijelazne i emisijske vjerojatnosti te algoritmi za rješavanje osnovnih problema modela.

Ovi modeli se primjenjuju u različitim područjima, uključujući obradu prirodnog jezika, biomedicinsku analizu, financije, i mnoga druga područja. U ovom seminaru prikazan je primjer primjene HMM-a na analizi financijskih podataka, gdje je HMM korišten za detekciju režima. HMM je generativan model te kao takav omogućuje stvaranje sintetičkih podataka što je korisno i u brzorastućem području strojnog i dubokog učenja.

LITERATURA

[1] L.R. Rabiner, *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proceedings of the IEEE (vol. 77, no.2), 1989.

[2] Wikipedia, *Hidden Markov model*

[3] Z. Vondraček, *Markovljevi lanci, predavanja*, Prirodoslovno-matematički fakultet, Zagreb, 2008.

[4] L. Nguyen, *Continuous Observation Hidden Markov Model*, 2016.

[5] C. Kohlschein, *An introduction to Hidden Markov Models*, 2006.

[6] Investopedia, *An Overview of Bull and Bear Markets*