

Ispit iz predmeta Istraživanje podataka u bioinformatiči

Trajanje: 1 h

Maksimalan broj poena: 60

Dozvoljeno je koristiti programski jezik Python i standardne biblioteke, kao što su math, csv i collections.

Nije dozvoljena upotreba biblioteka ili funkcija koje direktno implementiraju HMM modele, Viterbijev algoritam ili dekodiranje skrivene putanje.

Rešenje treba da sadrži programski kod, ispis rezultata i kratko objašnjenje dobijenih rezultata.

Zadatak - Viterbijev algoritam i najverovatnija putanja skrivenih stanja

60 poena

Dat je skriveni Markovljev model za razlikovanje regiona DNK sekvence koji pripadaju CpG ostrvima i regiona koji ne pripadaju CpG ostrvima.

Model ima dva skrivena stanja:

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

C	CpG region
N	region koji nije CpG

Početne verovatnoće su:

$$\pi(C) = 0.5, \quad \pi(N) = 0.5$$

Matrica prelaza je:

	sledeće C	sledeće N
iz C	$P(C C) = 0.8$	$P(N C) = 0.2$
iz N	$P(C N) = 0.3$	$P(N N) = 0.7$

Emisione verovatnoće su:

Stanje	A	C	G	T
C	0.15	0.35	0.35	0.15
N	0.30	0.20	0.20	0.30

Data je DNK sekvenca:

$$O = \text{AAGCGT}$$

Opis Viterbijevog algoritma

Neka je data sekvenca posmatranja

$$O = (o_1, o_2, \dots, o_T)$$

i neka je $Q = (q_1, q_2, \dots, q_T)$, gde je $q_t \in S$, putanja skrivenih stanja.

Za $j \in S$, Viterbijeva promenljiva definiše se kao

$$\delta_t(j) = \max_{(q_1, \dots, q_{t-1}) \in S^{t-1}} P(q_1, \dots, q_{t-1}, q_t = j, o_1, \dots, o_t \mid \lambda),$$

gde je $\lambda = (\pi, A, B)$ dati HMM model. Vrednost $\delta_t(j)$ predstavlja najveću zajedničku verovatnoću posmatranja $o_1, \dots, o_t$ i putanje koja se u trenutku t završava u stanju j .

Za $t \geq 2$, promenljiva $\psi_t(j)$ pamti stanje iz trenutka $t - 1$ iz kojeg se dobija ta najveća vrednost.

1. Inicijalizacija

Za prvo posmatranje i svako $i \in S$ važi:

$$\delta_1(i) = \pi(i)b_i(o_1), \quad \psi_1(i) = 0.$$

Vrednost 0 je samo oznaka da na prvoj poziciji prethodnik ne postoji; u ispisu se može prikazati znakom -.

Za dati model:

$$\delta_1(C) = \pi(C)b_C(o_1), \quad \delta_1(N) = \pi(N)b_N(o_1).$$

2. Rekurzija

Za $t = 2, 3, \dots, T$ i svako $j \in S$ važi:

$$\delta_t(j) = b_j(o_t) \max_{i \in S} \{ \delta_{t-1}(i) a_{ij} \}, \quad \psi_t(j) = \arg \max_{i \in S} \{ \delta_{t-1}(i) a_{ij} \}.$$

Ovde je $a_{ij} = P(q_t = j \mid q_{t-1} = i)$ verovatnoća prelaska iz stanja i u stanje j , dok je $b_j(o_t) = P(o_t \mid q_t = j)$ emisiona verovatnoća.

Pošto model ima dva stanja, formule za Viterbijeve vrednosti su:

$$\begin{aligned} \delta_t(C) &= b_C(o_t) \max \{ \delta_{t-1}(C)P(C \mid C), \delta_{t-1}(N)P(C \mid N) \}, \\ \delta_t(N) &= b_N(o_t) \max \{ \delta_{t-1}(C)P(N \mid C), \delta_{t-1}(N)P(N \mid N) \}. \end{aligned}$$

Vrednost $\psi_t(j)$ jednaka je stanju iz kojeg potiče maksimum. Ne zaokruživati tokom računanja; pri izjednačenju izabrati stanje C .

3. Završni korak

Poslednje stanje optimalne putanje i Viterbijev skor određuju se kao:

$$q_T^* = \arg \max_{i \in S} \delta_T(i), \quad P^* = \delta_T(q_T^*) = \max_{i \in S} \delta_T(i).$$

Pošto je O fiksirano, Q^* maksimizuje i $P(Q \mid O, \lambda)$. Ipak, $P^* = P(Q^*, O \mid \lambda)$ je zajednička, a ne ukupna verovatnoća sekvence $P(O \mid \lambda)$.

4. Povratno praćenje

Preostala stanja optimalne putanje dobijaju se unazad:

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \quad t = T - 1, T - 2, \dots, 1.$$

Na taj način dobija se najverovatnija putanja skrivenih stanja

$$Q^* = (q_1^*, q_2^*, \dots, q_T^*).$$

Zahtevi zadatka

1. Predstaviti početne, prelazne i emisione verovatnoće odgovarajućim strukturama podataka. **(8 poena)**
2. Implementirati inicijalizaciju i rekurziju Viterbijevog algoritma, uključujući čuvanje prethodnika u tabeli ψ , bez upotrebe biblioteka ili funkcija koje već implementiraju HMM ili Viterbijev algoritam. **(16 poena)**
3. Ispisati kompletnu Viterbijevu tabelu sa vrednostima $\delta_t(C)$ i $\delta_t(N)$ za svaku poziciju sekvence. Vrednosti ispisati kao obične verovatnoće; zaokruživanje je dozvoljeno samo prilikom ispisa. **(8 poena)**
4. Ispisati kompletnu tabelu prethodnika $\psi_t(C)$ i $\psi_t(N)$ za svaku poziciju, pri čemu za prvu poziciju treba ispisati da prethodnik ne postoji. **(8 poena)**
5. Izračunati i ispisati Viterbijev skor $P^* = \max_{Q \in S^T} P(Q, O \mid \lambda)$ i poslednje stanje optimalne putanje q_T^* . **(8 poena)**
6. Primenom povratnog praćenja rekonstruisati i ispisati najverovatniju putanju skrivenih stanja poravnatu sa sekvencom AAGCGT. **(8 poena)**
7. Kratko objasniti šta predstavlja vrednost P^* , zašto ona nije isto što i $P(O \mid \lambda)$ i zašto se optimalna putanja ne dobija nezavisnim izborom stanja sa većom vrednošću δ_t na svakoj poziciji. **(4 poena)**

Očekivani format ispisa

```

Pozicija 1, simbol A: delta(C) = ..., psi(C) = -; delta(N) = ..., psi(N) = -
Pozicija 2, simbol A: delta(C) = ..., psi(C) = ...; delta(N) = ..., psi(N) = ...
Pozicija 3, simbol G: delta(C) = ..., psi(C) = ...; delta(N) = ..., psi(N) = ...
Pozicija 4, simbol C: delta(C) = ..., psi(C) = ...; delta(N) = ..., psi(N) = ...
Pozicija 5, simbol G: delta(C) = ..., psi(C) = ...; delta(N) = ..., psi(N) = ...
Pozicija 6, simbol T: delta(C) = ..., psi(C) = ...; delta(N) = ..., psi(N) = ...

Viterbijev skor P*: ...
Poslednje stanje optimalne putanje: ...
Simboli: A A G C G T
Stanja: ... ..

```

Bodovanje

Stavka	Poeni
1. Predstavljanje početnih, prelaznih i emisionih verovatnoća	8
2. Implementacija inicijalizacije i rekurzije Viterbijevog algoritma	16
3. Kompletan i čitljiv ispis tabele δ	8
4. Kompletan i čitljiv ispis tabele ψ	8
5. Izračunavanje Viterbijevog skora P^* i poslednjeg stanja	8
6. Povratno praćenje i ispis najverovatnije sekvence stanja	8
7. Objašnjenje značenja P^* i uloge povratnog praćenja	4
Ukupno	60

Ukupan broj poena: 60