Bakatöö teema: Viterbi-tüüpi algoritmid ja murdepunktid

Tähistus $a_{1:n} := a_1, \ldots, a_n$ jne

Kadu ja risk. Olgu $Y_1, Y_2, \ldots, Y_n$ juhuslikud suurused väärtuste hulgaga $\{1, \ldots, K\}$ (ei pruugi olla sõltumatud ega sama jaotusega ega Markovi ahel). Juhusliku vektori $Y_{1:n}$ realisatsiooni üritame hinnata/prognoosida/esitada mingi (teatavas mõttes parima) konstantse (st mittejuhusliku) vektoriga $\hat{a}_{1:n} \in \{1, \ldots, K\}^n$. Olgu $L : K^n \times K^n \to \mathbb{R}^+$ kaofunktsioon, kus $L(y_{1:n}, a_{1:n})$ on kadu, mis tekib kui $y_{1:n}$ on tegelik $Y_{1:n}$ realisatsioon ja $a_{1:n}$ on prognoos. Keskmine kadu on risk

$$R(a_{1:n}) := EL(Y_{1:n}, a_{1:n}) = \sum_{y_{1:n}} L(y_{1:n}, a_{1:n}) P(Y_{1:n} = y_{1:n}).$$

Et üldiselt n on suur, siis nimetame vektoreid jadadeks või ka teeks. Otsime jada, mis minimeerib riski:

$$\hat{a}_{1:n} = \arg\min_{a_{1:n}} R(a_{1:n})$$

Näita, et kui

$$L(y_{1:n}, a_{1:n}) = \begin{cases} 0, & \text{kui } y_{1:n} = a_{1:n}; \\ 1, & \text{mujal.} \end{cases},$$

siis $\hat{a}_{1:n}$ on suurima tõenäosusega:

$$\hat{a}_{1:n} := \arg\max_{y_{1:n}} P(Y_{1:n} = y_{1:n}).$$

Näita, et kui kaofunktsioon on Hammingu kaugus ehk vigade arv

$$L(y_{1:n}, a_{1:n}) = \sum_{t=1}^{n} l(a_t, y_t), \quad l(a_t, y_t) = \begin{cases} 0, & \text{kui } a_t = y_t; \\ 1, & \text{mujal.} \end{cases}$$

siis risk $R(a_{1:n})$ on keskmine vigade arv ning $\hat{a}_{1:n}$ on järgmine:

$$\hat{a}_t = \arg\max_{k=1,\ldots,K} P(Y_t = k), \quad t = 1, \ldots, n.$$

Murdepunktid (change points). Olgu $y_{1:n}$ vektor/tee. Nimetame indeksit (ajahetke) $t > 1$ murdepunktiks kui $y_{t-1} \neq y_t$. Olgu $u_{2:n} \in \{0, 1\}^{n-1}$. Defineerime kaofunktsiooni

$$L : K^n \times 2^{n-1} \to \mathbb{R}^+$$

järgmiselt

$$L(y_{1:n}, u_{2:n}) = \sum_{t=2}^{n} l((y_{t-1}, y_t), u_t), \quad l((y_{t-1}, y_t), u_t) = \begin{cases} 0, & \text{kui } y_{t-1} = y_t \text{ ja } u_t = 0; \\ 0, & \text{kui } y_{t-1} \neq y_t \text{ ja } u_t = 1; \\ A, & \text{kui } y_{t-1} = y_t \text{ ja } u_t = 1; \\ B, & \text{kui } y_{t-1} \neq y_t \text{ ja } u_t = 0. \end{cases}$$

Seega $u_t = 1$ tähendab murdepunkti; $A > 0$ on murdepunkti valesti prognoosimise hind (st tegelikult murdepunkti pole) ja $B > 0$ on murdepunkto mahamagamise hind. Näiteks kui $y_{1:8} = 22333411$ ja $u_{2:8} = 0100110$ siis $L(22333411, 0100110) = 0$ aga $L(22333411, 1100000) = A + 2B$. Risk

$$R(u_{2;n}) = EL(Y_{1:n}; u_{2:n}).$$

Ülesanne 1: Leida parim murdepunktide prognoos

$$\hat{u}_{2:n} = \arg \min_{u_{2:n} \in \{0,1\}^{n-1}} R(u_{2:n}).$$

Olgu $\hat{u}_{2:n}$ on murdepunktide prognoos. Tähistame $\tau_1 < \tau_2 < \cdots < \tau_k$ murdepunktide indekseid/ajahetki (antud prognoosi põhjal). Näiteks kui $\hat{u}_{2:20} = 10000100000110000000$, siis $k = 4$ ja $\tau_1 = 2, \tau_2 = 7, \tau_3 = 13, \tau_4 = 14$.

Olgu meil antud nüüd murdepunktide arv k ja nende indeksid τ_i, $i = 1, \ldots, k$, teisõnu olgu meil murdepunktide prognoos $\hat{u}$. Neid teid, mille murdepunktid on täpselt τ_i on rohkem kui üks, kui $K = 2$, siis on selliseid teid täpselt kaks, kui $K > 2$, siis rohkem (kui palju?). Nende seast tuleks valida mingis mõttes parim. Näiteks suurima tõenäosusega tee

$$\hat{a} = \arg \max_{y_{1:n}: L(y_{1:n}, \hat{u}_{2:n}) = 0} p(y_{1:n}), \quad p(y_{1:n}) := P(Y_{1:n} = y_{1:n}). \tag{0.1}$$

Ülesande (0.1) lahend on suurima tõenäosusega tee nende teede seast mille murdepunktid on täpselt $\tau_1, \ldots, \tau_k$. Kui me avaldises (0.1) olevas kaofunktsioonis L võtame $A = 0$ (ja $B > 0$), siis (0.1) on suurima tõenäosusega tee nende teede seast, mille murdepunktid on hulgas $\{\tau_1, \ldots, \tau_k\}$ (aga mitte iga τ_i ei pruugi olla murdepunkt). Neid teid on rohkem. Kui (0.1) olevas kaofunktsioonis L võtame $B = 0$ (ja $A > 0$), siis (0.1) on suurima tõenäosusega tee nende teede seast, mille murdepunktid sisaldavad indekseid $\{\tau_1, \ldots, \tau_k\}$ aga neid murdepunkte võib olla ka mujal. Veendu selles.

Markovi ahel ja L-parim Viterbi algoritm. Olgu nüüd (ja edaspidi) $Y_1, Y_2, \ldots, Y_n$ mittehomogeene Markovi ahel seisundite hulgaga $\{1, \ldots, K\}$. Mittehomogeensus tähendab, et üleminekutõenäosused $P(Y_{t+1} = i | Y_t = j) =: p_t(i|j)$, $i, j \in \{1, \ldots, K\}$ sõltuvad ajast $t = 1, \ldots, n-1$. *Viterbi algoritm* leiab ahela suurima tõenäosusega realisatsiooni – nn *Viterbi või MAP* tee, olgu see $v^1_{1:n}$:

$$v^1_{1:n} := \arg \max_{y_{1:n}} p(y_{1:n}).$$

Algoritmi kompleksus $O(Kn)$. Pane tähele, et $v^1_{1:n}$ ei pruugi olla ühene, kuid kui $v^1_{1:n}$ ja $\tilde{v}^1_{1:n}$ on kaks erinevat Viterbi teed, siis $p(v^1_{1:n}) = p(\tilde{v}^1_{1:n})$, Olgu $p_1 := p(v^1_{1:n})$ kõikide Viterbi teede ühine tõenäosus.

Ülesanne 2: Üldista Viterbi algoritmi (leia vastav rekursioon) nii, et ta leiaks paremuselt teise, kolmanda, neljanda jne tee. Paremuselt teine tee

$$v^2_{1:n} := \arg \max_{y_{1:n}: p(y_{1:n}) < p_1} p(y_{1:n}).$$

Jällegi ei pruugi $v^2_{1:n}$ olla ühene, kuid vastav tõenäosus $p_2 := p(v^2_{1:n})$ kindlasti on. Paremuselt kolmas tee

$$v^3_{1:n} := \arg \max_{y_{1:n}: p(y_{1:n}) < p_2} p(y_{1:n}),$$

jne. Näita, et paremuselt L-nda teed saab leida kompleksusega $O(KLn)$.

Ülesande (0.1) lahendamise algoritm Markovi ahela korral. Ülesanne 3: Üldista Viterbi algoritmi (leia vastav rekursioon) nii, et ta lahendaks ülesande (0.1) juhul kui

1. Tingimuses on $A > 0$ ja $B > 0$, st kõik murdepunktid on ette antud;
2. Tingimuses on $A > 0$ ja $B = 0$, st kõik murdepunktid sisalduvad etteantud hulgas;
3. Tingimuses on $A = 0$ ja $B > 0$, st kõik murdepunktid sisaldavad etteantud hulka.

Teha sama juhul kui ülesande (0.1) asemel on järgmine ülesanne (väiksema keskmise vigade arvuga tee leidmine etteantud hulgast)

$$\hat{a}_{1:n} = \arg \max_{y_{1:n}: L(y_{1:n}, \hat{u}_{2:n}) = 0} \sum_{t=1}^{n} p_t(y_t), \quad p_t(y_t) := P(Y_t = y_t). \tag{0.2}$$

Etteantud murdepunktidega arvuga parima tee leidmine. Praktikas on teinekord teada murdepunktide arv või selle ülemine/alumine piir. Ehk otsitakse (mingis mõttes) parimat teed nende teede seast kui on k murdepunkti (või ülimalt/vähemalt) k murdepunkti. Olgu $\mu(y_{1:n})$ tee $y_{1:n}$ murdepunktide arv. Kui headuse kriteerium on tõenäosus, siis vaja leida

$$\hat{a}_{1:n} = \arg \max_{y_{1:n}: \mu(y_{1:n}) = k} p(y_{1:n}), \tag{0.3}$$

kusjuures märk $=$ tingimuses võib olla nii $\geq$ või ka $\leq$.

Ülesanne 4: Üldista Viterbi algoritmi (leia vastav rekursioon) nii, et ta lahendaks ülesande (0.3) juhul kui murdepunktide arv on 1) täpselt k, 2) vähemalt k, 3) maksimaalselt k. Lahenda ülesanne ka siis kui (0.3) asemel on järgmine ülesanne

$$\hat{a}_{1:n} = \arg \max_{y_{1:n}: \mu(y_{1:n}) = k} \sum_{t=1}^{n} p_t(y_t), \tag{0.4}$$

Paarikaupa Markovi mudel (PMM). 2D-protsess (X, Y) on Markovi ahel. Varjatud Markovi ahel (HMM) on erijuht. Antud $X_{1:n}$ realisatsioon $x_{1:n}$. Tinglikult on $Y_{1:n}$ mittehomogeene Markovi ahel, st

$$P(Y_{t+1} = j | Y_{1:t-1} = i_{1:t-1}, Y_t = i, X_{1:n} = x_{1:n}) = P(Y_{t+1} = j | Y_t = i, X_{1:n} = x_{1:n}).$$

Teglikult kehtib ka

$$P(Y_{t+1} = j | Y_t = i, X_{1:n} = x_{1:n}) = P(Y_{t+1} = j | Y_t = i, X_{1:n} = x_{t:n}),$$

näita seda. Seega kõik ülaltoodud algoritmid ja rekursioonid kehtivad, formaalselt on vaja asendada tõenäosused $p(y_{1:n})$ tinglike tõenäosustega

$$p(y_{1:n}|x_{1:n}) := P(Y_{1:n} = y_{1:n}|X_{1:n} = x_{1:n}).$$

Edasi-tagasi algoritmid võimaldavad efektiivselt arvutada tõenäosusi $P(Y_{t:t+k} = a_{t:t+k})$ iga $k = 0, 1, 2, \cdots$ korral ja iga $t = 1, \ldots, n - k$ korral.

Ülesanne 5: Esita kõik ülaltoodud algoritmid/rekursioonid PMM jaoks. Kuidas näevad need algoritmid välja HMM korral? Kui vaja, arvesta skaleerimist.

Arvutisimulatsioonid.

References

[1] J. Lember, Tehisõpe I, loengukonspekt 2017 (ptk 7), https://courses.ms.ut.ee/MTMS.02.046/2017

[2] K. Kuljus, J. Lember, Hybrid classifiers of pairwise Markov models, https://arxiv.org/pdf/2203.10574.pdf

[3] L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE* , Vol. 77(2), pp. 257 - 286, 1989.

[4] J. Sõnajalg, Segmenteerimine peidetud Markovi mudelite segude korral, Magistritöö, 2016

[5] T. Koski, Hidden Markov models in Bioinformatics, Kluwer, 2001