Varjatud Markovi ahelad jadade joondamisel

Õppematerjal ja lõputöö juhend

Enne töö alustamist tutvub üliõpilane Markovi ahela ja varjatud Markovi ahelaga seotud põhimõistetega: seisundid, üleminutõenäosused, emissioonitõenäosused, edasi-tagasi algoritmid. Samuti tutvub ta jadade joondamise tõhimõistetega. Selleks sobib hästi raamat (Koski, 01). Vaata ka K. Vimmi bakalaureusetööd (Vimm, 22) või raamatut (Durbin et al., 1998).

Lõputöö peamine fookus on erinevate dünaamilise planeerimise (*dynamic programming*) algoritmide väljatöötamine.

1 Juhuslik tee

Joondusi modelleeriv esialgne Markovi ahel. Seisundid:

$$\mathcal{Z} := \{M, I, D\} \quad \text{(match, insertion, deletion)}.$$

Üleminekumaatriks ja algtõenäosused:

$$P = \begin{pmatrix} p_{MM} & p_{MD} & p_{MI} \\ p_{DM} & p_{DD} & p_{DI} \\ p_{IM} & p_{ID} & p_{II} \end{pmatrix} \quad \pi = \begin{pmatrix} \pi_M \\ \pi_D \\ \pi_I \end{pmatrix}. \tag{1}$$

Joondus ja selle tõenäosus. Hulga $\{M, I, D\}^*$ (kõik lõplikud kolmest tähest moodustatud vektorid) elemente nimetame *joondusteks*. Igal joondusel on oma tõenäosus. Selle määravad üleminekumaatriks ja algtõenäosused:

$$p(z_{1:k}) :=: \pi_{z_1} \prod_{t=1}^{k-1} p_{z_t z_{t+1}}, \quad z_{1:k} := (z_1, \ldots, z_k).$$

Defineerime

$$\mathcal{A}(i,j) := \Big\{ z_{1:k} \in \{M,I,D\}^* : \sum_{t=1}^{k} I_{M,D}(z_t) = i, \quad \sum_{t=1}^{k} I_{M,I}(z_t) = j \Big\}, \quad i,j \in \mathbb{N}$$

$$\mathcal{A}(i,0) := \{\overbrace{(D,\ldots,D)}^{i}\}, \quad \mathcal{A}(0,j) := \{\overbrace{(I,\ldots,I)}^{j}\}.$$

Seega $\mathcal{A}(i,j)$ moodustavad sellised vektorid $z_{1:k}$ (pikkus k pole fikseeritud), mis rahuldavad tingimusi:

- vektoris on M ja D tähti kokku i;
- vektoris on M ja I tähti kokku j.

Näide: (üliõpilane lõpetab näite):

$$\mathcal{A}(4,2) = \{\overbrace{(M,M,D,D),\ldots,(D,D,M,M)}^{\text{permutatsioonid}},\overbrace{(M,I,D,D,D),\ldots,(D,D,D,I,M)}^{\text{permutatsioonid}}$$
$$(I,I,D,D,D,D),\ldots(D,D,D,D,I,I)\}$$

Ülesanne: Leia hulga $\mathcal{A}(i,j)$ võimsus $|\mathcal{A}(i,j)|$. Näita (või lükka ümber): $|\mathcal{A}(i,j)| = |\mathcal{A}(i-1,j-1)| + |\mathcal{A}(i-1,j)| + |\mathcal{A}(i,j-1)|$, kui $i,j > 0$.

Defineerime:

$$P(i,j) := \sum_{z_{1:k}\in\mathcal{A}(i,j)} p(z_{1:k}).$$

Seega $P(i,j)$ on hulga $\mathcal{A}(i,j)$ tõnäosus ja see ei ole üldiselt 1.

Näide: Leia $P(i,j)$, kui kõik üleminekutõenäosused ja algtõenäosused on $1/3$.

Joondus kahedimensionaalsena. Kuulugu $z = z_{1:k} \in \mathcal{A}(n,m)$. Defineerime $a(z) = a_{0:k}$, kus

$$a_0 = (0,0),\ a_t = \big(\sum_{l=1}^{t} I_{M,D}(z_l), \sum_{l=1}^{t} I_{M,I}(z_l)\big), \quad t = 1,\ldots,k.$$

Seega $a_t \in \{0,1,\ldots,n\}\times\{0,1,\ldots,n\}$ iga $t = 1,\ldots,k$ korral.

Näide:

$z_{1:8} = (D,M,M,I,D,M,I,I)$, $a(z) = \big((0,0),(1,0),(2,1),(3,2),(3,3),(4,3),(5,4),(5,5),(5,6)\big)$.

On selge, et kui $z_{1:k} \neq z'_{1:k}$, siis $a(z_{1:k}) \neq a(z'_{1:k})$ ehk $a(z_{1:k})$ on lihtsalt joonduse kahedimensionaalne esitus. Sellisele esitusele vastab joon/tee graafil, mille sõlmed on paarid (i,j) ja servad on naabrite vahel: (i,j) naabrid on $(i+1,j)$, $(i+1,j+1)$, $(i,j+1)$ (kui $i < n, j < m$). Sisuliselt võre, kus serv on veel tipu ja tema alumise parempoolse (SE) naabri vahel. Joondusele vastav tee algab punktist $(0,0)$ ja lõpeb alati punktis (n,m). Tähistame ka selliste teede hulka $\mathcal{A}(n,m)$. Igal sellisel joondusel on tõenäosus $p(a_{0:t}) = p(z_{1:k})$, kus $a_{0:k} = a(z_{1:k})$. Nende tõenäosuste summa pole 1.

Edaspidi tähistame $a_t = (a_t^x, a_t^y)$ ehk kui $a_t = (i,j)$, siis $a_t^x = i, a_t^y = j$. Samuti tähistame $z_t = z(a_{t-1}, a_t)$ paarile (servale) (a_{t-1}, a_t) vastavat seisundit, täpsemalt

$$z(a_{t-1},a_t) := \begin{cases} M, & \text{kui } a_t^x - a_{t-1}^x = 1, a_t^y - a_{t-1}^y = 1; \\ D, & \text{kui } a_t^x - a_{t-1}^x = 1, a_t^y - a_{t-1}^y = 0; \\ I, & \text{kui } a_t^x - a_{t-1}^x = 0, a_t^y - a_{t-1}^y = 1. \end{cases}$$

Kui $a_{0:k} \in \mathcal{A}(n,m)$, siis $z_{1:k}$, kus $z_t = z(a_{t-1},a_t)$ on vastav seisundite jada.

Tinglik tõenäosus. Olgu n, m fikseeritud ja defineerime tõenäosusmõõdu

$$q(z_{1:k}) := \frac{p(z_{1:k})}{P(n,m)} \quad \forall z_{1:k} \in \mathcal{A}(n,m).$$

Seega kui $Z_1, Z_2, \ldots$ on ülaldefineeritud Markovi ahel, st $P(Z_{1:k} = z_{1:k}) = p(z_{1:k})$, siis iga $z_{1:k} \in \mathcal{A}(n,m)$ korral

$$q(z_{1:k}) = P\Big(Z_{1:k} = z_{1:k} \Big| \sum_{t=1}^{k} I_{M,D}(Z_t) = n, \quad \sum_{t=1}^{k} I_{M,I}(Z_t) = m\Big), \quad \forall z_{1:k} \in \mathcal{A}(n,m).$$

Ülesanne: Näita, et q ei rahulda Markovi omadust:

$$q(z_{1:k}) \neq q(z_1)q(z_2|z_1)\cdots q(z_k|q_{k-1}).$$

Analogiliselt defineerime tõenäosusmõõdu kahedimensionaalsete jaotuste hulgal

$$q(a_{0:k}) := q(z_{1:k}), \quad a_{0:k} = a(z_{1:k}) \quad \forall a_{0:k} \in \mathcal{A}(n,m).$$

Seega igal teel punktist (0,0) punkti (n,m) on oma tõenäosus, nende teede tõenäosuste summa on nüüd 1. Iga teelõigu tõenäosus on kõikide seda teelõiku sisaldavate teede tõenäosuste summa. Seega kui $a_{0,k} \in \mathcal{A}(n,m)$ ja $z_{1:k}$ on vastav seisundite jada, siis

$$\begin{aligned} q(a_{0:t}) &= \sum_{a'_{0:k} \in \mathcal{A}(n,m): a'_{0:t} = a_{0:t}} q(a'_{0:k}) = \frac{1}{P(n,m)} \sum_{a'_{0:k} \in \mathcal{A}(n,m): a'_{0:t} = a_{0:t}} p(z_{1:k}) \\ &= \frac{1}{P(n,m)} p(z_{1:t}) \cdot \Big[\sum_{z_{t+1,n} \in \mathcal{A}(n-i,m-j)} p_{z_t}(z_{t+1;n}) \Big] = \frac{1}{P(n,m)} p(z_{1:t}) P_{z_t}(n-i, m-j), \end{aligned}$$

kus $a_t = (i,j)$ ja

$$p_z(z_{t+1;n}) := p_{zz_{t+1}} \cdot p_{z_{t+1}z_{t+2}} \cdots p_{z_{n-1}z_n}, \quad P_z(i,j) := \sum_{z_{1:k} \in \mathcal{A}(i,j)} p_z(z_{1:k}),\ P_z(0,0) := 1 \quad z \in \mathcal{Z}.$$

Pane tähele, et

$$P_z(i,j) = p_{zD} P_D(i-1,j) + p_{zM} P_M(i-1,j-1) + p_{zI} P_I(i,j-1). \tag{2}$$

Seega

$$q(a_{t+1}|a_{0:t}) = \frac{q(a_{0:t+1})}{q(a_{0,t})} = \frac{p_{z_t z_{t+1}} P_{z_{t+1}}(n - a^x_{t+1}, m - a^y_{t+1})}{P_{z_{t+1}}(n - a^x_t, m - a^y_t)} =$$
$$\frac{p_{z_t z_{t+1}} P_{z_{t+1}}(n - a^x_{t+1}, m - a^y_{t+1})}{p_{z_t M} P_M(n - a^x_t - 1, m - a^y_t - 1) + p_{z_t D} P_D(n - a^x_t - 1, m - a^y_t) + p_{z_t I} P_I(n - a^x_t, m - a^y_t - 1)}.$$

Et z_t on (a_{t-1}, a_t) funktsioon $z_t = z(a_{t-1}, a_t)$, saame teist järku Markovi omaduse:

$$q(a_{t+1}|a_{0:t}) = q(a_{t+1}|a_{t-1:t}).$$

Ühe tee $a_{0:k}$ tõenäosus on nüüd :

$$q(a_{0:k}) = q(a_1|a_0)q(a_2|a_1,a_0)q(a_3|a_2,a_1)\cdots q(a_k|a_{k-1},a_{k-2}),$$

kusjuures

$$q(a_1|a_0) \propto \begin{cases} \pi_D P_D(n-1,m), & \text{kui } a_1=(1,0); \\ \pi_M P_M(n-1,m-1), & \text{kui } a_1=(1,1); \\ \pi_I P_I(n,m-1) & \text{kui } a_1=(0,1). \end{cases}$$

Et

$$\pi_D P_D(n-1,m) + \pi_M P_M(n-1,m-1) + \pi_I P_I(n,m-1) = P(n,m),$$

siis normaliseeriv konstant on $P(n,m)$.

Näide: Olgu $a_t=(i,j)$, $a_{t+1}=(i+1,j)$. Seega $z_{t+1}=D$. Nüüd

$$q((i+1,j)|(i,j),a_{t-1}) = \frac{p_{z_t,D}P_D(n-i-1,m-j)}{p_{z_t,D}P_D(n-i-1,m-j)+p_{z_t,M}P_M(n-i-1,m-j-1)+p_{z_tI}P_I(n-i,m-j-1)}.$$

Veendu, et

$$q\Big((n,m-1)|((n-1),(m-1)),((n-2),(m-2))\Big) = \frac{p_{MD}p_{DI}}{p_{MD}p_{DI}+p_{MM}+p_{MI}p_{ID}},$$

sest $P_D(0,1)=p_{DI}$, $P_M(0,0)=1$, $P_I(1,0)=p_{ID}$. Näeme, et üleminekutõenäosused sõltuvad positsioonist (i,j) ehk teist järku Markovi ahel pole homogeenne. Selleks, et leida üleminekutõenäosused on vaja iga paari (i,j) ja iga seusundi z korral arvutada $P_z(i,j)$, $i=1,\ldots,n$, $j=1,\ldots,m$.

Ülesanne: Leia rekursioon üleminekutõenäosuste $q(b|a,a')$ arvutamiseks. Üleminekutõenäosused veelkord:

$$q(b|a,a') = \frac{p_{z',z_b}P_{z_b}(n-b^x,m-b^y)}{\sum_z p_{z',z}P_z(n-(a^x+z^x),m-(b^y+z^y))} =: q(b|a,z') :=: q(z_b|a,z') \qquad (3)$$

kus $z'=z(a',a)$, $z_b=z(a,b)$ ja

$$z^x = I_{\{M,D\}}(z), \quad z^y = I_{\{M,I\}}(z).$$

Üks võimalus on leida rekursioon $P_z(i,j)$ jaoks. Veendu:

$$P_z(i,0) = p_{z,D}p_{DD}^{i-1}, \quad P_z(0,j) = p_z I p_{II}^{j-1}.$$

Kasuta rekursiooni (2) ja leia $P_z(i,j)$ iga i,i ja z korral. Lõpuks leia $P(n,m)$. Kas tekib väikeste arvude probleem? Kui see probleem tekib, st tõenäosused $P_z(i,j)$ on liiga väikesed, siis tuleb leida skaleeritud rekursioon üleminekutõenäosuste (3) leidmiseks.

Näide: Leia üleminekutõenäosused (3), kui üleminekumaatriks on

$$\mathbf{A}: \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix} \qquad \mathbf{B}: \begin{pmatrix} 1-2\delta & \delta & \delta \\ \delta & 1-2\delta & \delta \\ \delta & \delta & 1-2\delta \end{pmatrix} \quad \pi = \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix}, \quad \delta \le 0.25. \tag{4}$$

Algtõenäosuste vektor mõlemal juhul π.

Serva läbimise tõenäosus. Olgu (a,b) naabrid (serv), st kui $b=(i,j)$, siis

$$a \in \{(i-1,j),(i-1,j-1),(i,j-1)\}.$$

Leiame tõenäosuse, et juhuslik tee (punktist $(0,0)$ punkti (n,m)) läbib serva (a,b). Kui tõenäosused (3) on teada, siis võime kasutada rekursiooni:

$$q(a,b) = \sum_{a' \in N(a)} q(a',a)q(b|a,a'), \quad N(a^x,a^y) := \{(a^x-1,a^y),(a^x-1,a^y-1),(a^x,a^y-1)\}, \quad a^x>0, a>0.$$

Rekursiooni algus: Kui $a=(0,0)=a_0$, siis $q(a,b)=q(b|a_0)$. Kui $a=(0,i)$, $b=(0,i+1)$, siis $q(a,b)=q(a',a)q(b|a,a')$, kus $a'=(0,i-1)$ jne.

Serva läbimise tõenäosus esialgse Markovi ahela (1) kaudu. Selleks defineerime

$$P(i,j,z) := \sum_{z_{1:k} \in \mathcal{A}(i,j): z_k = z} p_z(z_{1:k}),$$
$$P(i,0,I) = P(i,0,M) := 0, \quad (0,j,D) = P(0,j,M) := 0, \quad P(0,0,z) := 0.$$

On selge, et $P(i,j) = P(i,j,D) + P(i,j,M) + P(i,j,I)$.

Ülesanne: Leia rekursioon $P(i,j,z)$ arvutamiseks. Leia (skaleeritud) rekursioon $p(z|i,j) = P(i,j,z)/P(i,j)$ arvutamiseks.

Saame

$$q(a,b) = \frac{1}{P(n,m)} \sum_z P(a^x,a^y,z) p_{zz'} P_{z'}(n-b^x, m-b^y) =: \frac{p(a,b)}{P(n,m)}, \quad z' = z(a,b).$$

Ülesanne: Veendu lihtsa näite korral (maatriksid (4)), et mõlemad rekursioonid/valemid $q(a,b)$ arvutamiseks annavad ühe ja sama tulemuse.

Punkti läbimise tõenäosus. Punkti $a=(i,j)$ läbimise tõenäosus: $q(a) = \sum_b q(a,b)$. Punkti läbimise tõenäosus esialgse Markovi ahela (1) kaudu:

$$q(a) = \frac{\sum_z P(a^x,a^y,z) P_z(n-i,m-j)}{P(n,m)} =: \frac{p(a)}{P(n,m)}.$$

Suurima tõenäosusega (Viterbi) tee. Otsime

$$v_{0:k} := \arg \max_{a_{0:k} \in \mathcal{A}(n,m)} q(a_{0:k}).$$

Viterbi algoritm: defineeri

$$\delta(i,j,z) := \max\{\ln q(z_{1:k}) : z_{1:k} \in \mathcal{A}(i,j) : z_k = z\}.$$

Rekursioon:

$$\begin{aligned}
\delta(i,j,D) &= \max_{z'}\{\delta(i-1,j,z') + \ln q(D|(i-1,j),z')\},\\
\psi(i,j,D) &:= \arg\max_{z'}\{\delta(i-1,j,z') + \ln q(D|(i-1,j),z')\}\\
\delta(i,j,I) &= \cdots\\
\psi(i,j,I) &:= \cdots\\
\delta(i,j,M) &= \cdots\\
\psi(i,j,M) &:= \cdots .
\end{aligned}$$

Otsitava joonduse $z_{1:k}$ viimane täht on $\arg\max_z \delta(n,m,z)$. Kui see on näiteks I, siis eelviimane täht on $\psi(n,m,I)$, järgmine täht (tagantpoolt) on $\psi(n,m-1,\psi(n,m,I))$ jne.

Ülesanne: Lõpeta algoritm (rekuriooni algus jne) ning et algoritm töötab. Näita, et kui ülaltoodud algoritmis q-tõenäosused asendada p-tõenäosustega (st $q(D|(i-1,j),z')$ asemel $p_{z'D}$ jne), siis tulemus ei muutu.Leia Viterbi tee mõnel lihtsal näitel (näiteks maatriksid (4)).

Suurima servade tõenäosuste summaga/korrutisega tee. Olgu meil antud kõikide servade tõenäosused $q(a,b)$, kus $a \in N(b)$. Otsime teed, mille servade tõenäosuste summa oleks maksimaalne:

$$w_{0:k} = \arg \max_{a_{0:k} \in \mathcal{A}(n,m)} \big(q(a_0,a_1) + \cdots + q(a_{k-1},a_k)\big).$$

Asendades ülaltoodud avaldises $q(a_{t-1},a_t)$ logaritmiga $\ln q(a_{t-1},a_t)$, saame suurima servade tõenäosuste korrutisega tee. See garanteerib, et iga serva tõenäosus lahendis on positiivne.

Kas servatõenäosuste summa (korrutis) soosib pikimaid jadu? Siit ülesanne (loomulikult võib q asemel olla ka $\ln q$).

$$w_{0:k} = \arg \max_{a_{0:k} \in \mathcal{A}(n,m)} \big(q(a_0,a_1) + \cdots + q(a_{k-1},a_k) - Ck\big),$$

kus $C \geq 0$ on regulariseeriv konstant.

Ülesanne: Leia algoritm(id) ülaltoodud ülesannete lahendamiseks (tegelikult ju vaid üks algoritm).

Pane tähele: kui tõenäosuste summas q-tõenäosused asendada p-tõenäosustega, siis tulemus sama, st:

$$\arg \max_{a_{0:k}\in\mathcal{A}(n,m)} \big(q(a_0,a_1)+\cdots+q(a_{k-1},a_k)\big) = \arg \max_{a_{0:k}\in\mathcal{A}(n,m)} \big(p(a_0,a_1)+\cdots+p(a_{k-1},a_k)\big).$$

Kas ülaltoodud samasus kehtib ka siis, kui $C > 0$?
Kui aga korrutises (logaritmide summas) asendada q-tõenäosused asendada p-tõenäosustega, siis tulemus ei pruugi sama olla, sest

$$\ln q(a_0,a_1)+\cdots+\ln q(a_{k-1},a_k) = \ln p(a_0,a_1)+\cdots+\ln p(a_{k-1},a_k) - k\ln P(n,m).$$

Hübriidjoondus. Optimeerimisülesanne

$$\arg \max_{a_{0:k}\in\mathcal{A}(n,m)} \Big(B\sum_{t=0}^{k-1} \ln q(a_t,a_{t+1}) + C\ln q(a_{0:k})\Big), \tag{5}$$

kuus $C, B \geq 0$ on regulariseerivad konstandid. Arusaadavalt piisab tegelikult ühest konstandist (miks?).

Ülesanne: Olgu $a_{0:k} \in \mathcal{A}(n,m)$. Defineerime: $r(s:t) := \ln q(a_{s:t})$

$$R(l,a_{0:k}) := r(0:1)+\cdots+r(0:l)+r(1:l+1)+r(2:l+2)+\cdots r(k-l:k)+\cdots+r(k-1:k).$$

Seega

$$R(1,a_{0:k}) = \sum_{t=0}^{k-1} \ln q(a_t,a_{t+1}).$$

Tõesta, et

$$R(l,a_{0:k}) = R(l-1,a_{0:k}) + \ln q(a_{0:k}).$$

Järelda sellest, et

$$R(l,a_{0:k}) = \sum_{t=0}^{k-1} \ln q(a_t,a_{t+1}) + (l-1)\ln q(a_{0:k}).$$

Ülesanne: Leia dünaamilise planeerimise algoritm (5) lahendamiseks. Millise ülesande lahendame siis, kui algoritmis asendame q-tõenäosused p-tõenäosustega?

Rekursiooni idee (kontrolli!):

$$\delta(i,j,D) = \max_z \big(\delta(i-1,j,z) + C\ln q((i,j)|(i-1,j),z)\big) + B\ln q((i-1,j),(i,j)).$$

Joondus kui Markovi ahel graafil. Võtame ülaltoodu kokku: Nägime, et esialgne (homogeenne) Markovi ahel seisundite hulgal $\mathcal{Z}$ (1) defineerib (mittehomogeense) Markovi ahela graafil (läbi tingumustamise). Kas see seos on üksühene: kas iga Markovi ahel graafil vastab mingisugusele (võib olla ka mittehomogeensele) ahelale seisundite hulgal? Millisele esialgsele seisundite ahelale vastab järgmine ahel graafil:

$$q(b|a,a') = \begin{cases} \frac{1}{3}, & \text{kui } a^x < n, a^y < m; \\ 1, & \text{kui } a^x = n \text{ või } a^y = m; \end{cases} \quad ?$$

See on juhuslik ekslemine graafil kuni servani ja siis serva mööda lõpuni.

Nii või teisiti, edaspidi tasubki vaadelda juhuslikke joondusi kui Markovi ahelaid graafil, see on paljuski lihtsam (pole vaja arvutada üleminekutõenäosusi (3)) ja võibolla ka üldisem (kui üksühest vastavust pole). Teisisõnu, juhuslike joonduste tõenäosused on määratud üleminekutõenäosustega

$$\{q(z|(i,j),z'), \quad i = 0,\dots,m, \quad j = 0,\dots,n\}. \tag{6}$$

2 Pair HMM

Mudel. Seisund M emiteerib tähepaarid $(x,y) \in \mathcal{X}^2$, seisund D emiteerib paari $(x,-)$, $x \in \mathcal{X}$ (täht X jadas ja indel Y jadas), seisund D emiteerib paari $(-,y)$, $y \in \mathcal{X}$ (täht Y jadas ja indel X jadas).

Emissioonitõenäosused: $p(x,y)$ (seisund M), $p(x,-)$ (seisund D) ja $p(-,y)$ (seisund I). Kui lisada tähestikule indel, saame nn laiendatud tähestiku $\mathcal{X}^+ = \{\mathcal{X},-\}$, võime formaalselt kõik emissioonitõenäosused defineerida hulgal $\mathcal{X}^+ \times \mathcal{X}^+$: $p(x,y|M), p(x,y|D), p(x,y|I)$, kuid nüüd on kitsendused (siin $x,y \in \mathcal{X}$)

$$p(x,-|M) = p(-,y|M) = P(-,-|M) = 0,$$
$$p(x,y|D) = p(-,-|D) = p(-,y|D) = 0, \quad p(x,y|I) = p(-,-|I) = p(x,-|I) = 0.$$

Mudel: Markovi ahel (n,m) graafil liigub punktist $(0,0)$ punkti (n,m) vastavalt tõenäosustele (6). Igale servale/naabripaarile (a,b) vastab seisund $z(a,b) \in \{I,D,M\}$. Kui ahel läbib serva (a,b), genereeritakse paar $(x,y) \in \mathcal{X}^+ \times \mathcal{X}^+$ jaotusest $p(x,y|z(a,b))$. Seega kokku genereeritakse n X-tähte ja m Y-tähte.

Olgu $a_{0:k}$ Markovi ahela realisatsioon – joondus – ning olgu $x^* = (x'_1,\dots,x'_k), y^* = y'_1,\dots,y'_k$ genereeritud jadad. Edaspidi nimetame neid *laiendatud jadadeks*. Pane tähele, et paar (x^*,y^*) määrab üheselt teda genereerinud joonduse $a(x^*,y^*)$. Seega selle laiendatud paari saamise tõenäosus meie mudelis:

$$p(x^*,y^*) = p(x'_1,\dots,x'_k,y'_1,\dots,y'_k) = q(a_{0:k}) \prod_{t=1}^{k} p(x'_t,y'_t|z_t),$$

kus $a_{0:k} = a(x^*, y^*)$ ja $z_{1:k}$ vastab joondusele $a_{0:k}$. Laiendatud jadas x^* on täpselt n tähte (ja $k-n$ indelit), olgu vastav alamjada $x_{1:n} \in \mathcal{X}^n$; analoogiliselt olgu $y_{1:m}$ laiendatud jada y^* tähtedest moodustatud alamjada. Edaspidi nimetame neid jadu *vaatlusteks*. Pane tähele, et vaatlused $(x_{1:n}, y_{1:m})$ ja joondus $a_{0:k}$ määravad üheselt laiendatud jadad (x^*, y^*). Sellest lähtudes defineerime

$$p(x_{1:n}, y_{1:m}|a_{0:k}) := \prod_{t=1}^{k} p(x'_t, y'_t|z_t),$$

millest saame vaatluste ja joonduse *ühistõenäosusele* kuju

$$p(x_{1:n}, y_{1:m}, a_{0:k}) := p(x^*, y^*) = q(a_{0:k})p(x_{1:n}, y_{1:m}|a_{0:k}).$$

Meie mudelis on seega vaatluste saamise tõenäosus *marginaaltõenäosus*:

$$p(x_{1:n}, y_{1:m}) = \sum_{a_{0:k} \in \mathcal{A}(n,m)} p(x_{1:n}, y_{1:m}|a_{0:k})q(a_{0:k}). \tag{7}$$

Näide: $\mathcal{X} = \{0, 1\}$, $x_{1:n} = (1, 0, 1, 1, 0, 1, 1)$, $y_{1:m} = (0, 0, 0, 1, 0, 1, 0, 0)$. Tavaliselt esitatkse joondused tabelina.

x^*	1	0	-	-	1	1	0	1	-	1
y^*	-	0	0	0	1	0	-	1	0	0

Vastav joondus ja selle tõenäosus

$$p(z_{1:k}) = p(D, M, I, I, M, M, D, M, I, M) = \pi_D \cdot p_{DM} \cdot p_{MI} \cdot p_{II} \cdot p_{IM} \cdot p_{MM} p_{MD} \cdot p_{DM} \cdot p_{MI} \cdot p_{IM}$$

$$q(z_{1:k}) = \frac{p(z_{1:k})}{P(7, 8)}.$$

Vaatluste tinglik tõenäosus:

$$p(x_{1:n}, y_{1:m}|z_{1:k}) = p(1, -)p(0, 0)p(-, 0)p(-, 0)p(1, 1)p(1, 0)p(0, -)p(1, 1)p(-, 0)p(1, 0).$$

Joonduse tinglik jaotus. Vastavalt tingliku tõenäosuse definitsioonile:

$$p(a_{0:k}|x_{1:n}, y_{1:m}) = \frac{q(a_{0:k})p(x_{1:n}, y_{1:m}|a_{0:k})}{p(x_{1:n}, y_{1:m})}.$$

Ülesanne: Näita, et säilib (teist järku) Markovi omadus:

$$p(a_{t+1}|a_{0:t}, x_{1:n}, y_{1:m}) = p(a_{t+1}|a_{t-1:t}, x_{1:n}, y_{1:m}).$$

Selleks näita, et kui $a_t = (i, j)$, siis

$$p(x_{1:n}, y_{1:m}, a_{0:t}) = \sum_{a'_{0:k} \in \mathcal{A}(n,m): a'_{0:t} = a_{0:t}} p(x_{1:n}, y_{1:m}, a'_{0:k}) = q(a_{0:t})p(x_{1:i}, y_{1:j}|a_{0:t})p(x_{i+1:n}, y_{j+1:m}|a_t, z_t),$$

$$p(x_{i+1:n}, y_{j+1:m}|a_t, z_t) := \sum_{a_{t+1:k} \in \mathcal{A}(n-i, m-j)} q(a_{t+1:k}|a_t, z_t)p(x_{i+1:n}, y_{j+1:m}|a_{t+1:k}).$$

Seega kui $a_{t+1} = (i+1, j)$ (st $z_t = D$)

$$p(a_{t+1}|a_{0:t}, x_{1:n}, y_{1:m}) = \frac{p(x_{1:n}, y_{1:m}, a_{0:t+1})}{p(x_{1:n}, y_{1:m}, a_{0:t})} = q(a_{t+1}|a_{t-1:t})p(x_{i+1}, -)\frac{p(x_{i+2:n}, y_{j+1:m}|z_t)}{p(x_{i+1:n}, y_{j+1:m}|z_{t+1})}.$$

Näeme, et parem pool ei sõltu $a_{0:t-2}$ -st.

Asjaolu, et Markovi omadus säilib, on väga oluline. Nüüd on meil jällegi üleminekutõenäosused kujul (6). Üleminekutõenäosused sõltuvad nüüd vaatlustest $x_{1:n}$ ja $y_{1:m}$, aga need on fikseeritud. Kõik ülaltoodud algoritmid rakenduvad. Küsimus on selles, kuidas üleminekutõenäosusi efektiivselt arvutada. See taandub küsimusele: kuidas arvutada tõenäosusi $p(x_{i+2:n}, y_{j+1:m}|z_t)$ (või seda suhet seal valemis).

β-rekursioon. Antud tõenäosused (6) ning emissioonitõenäosused. Rekursioon:

$$\begin{aligned}
p(y_m|(n, m-1), z) &= p(-, y_m), \\
p(x_n|(n-1, m), z) &= p(x_n, -) \\
p(x_{i+1:n}, y_{j+1:m}|(i, j), z) &= q(D|(i, j), z)p(x_{i+1}, -)p(x_{i+2:n}, y_{j+1:m}|(i+1, j), D) \\
&\quad + q(I|(i, j), z)p(-, y_{j+1})p(x_{i+1:n}, y_{j+2:m}|(i, j+1), I) \\
&\quad + q(M|(i, j), z)p(x_{i+1}, y_{j+1})p(x_{i+2:n}, y_{j+2:m}|(i+1, j+1), M)
\end{aligned}$$

Seega

$$q(D|(i, j), z, x_{1:n}, y_{1:m}) = \frac{q(D|(i, j), z)p(x_{i+1}, -)p(x_{i+2:n}, y_{j+1:m}|z_t)}{p(x_{i+1:n}, y_{j+1:m}|z)},$$

kus $p(x_{i+1:n}, y_{j+1:m}|z)$ avaldub summana. Võrdle saadud valemit valemiga (3). Tekib väikeste arvude ja skaleerimise probleem. Ülatoodud tõenäosused sõltuvad $x_{i+1,n}$ ja $y_{j+1,m}$, mistõttu võime kirjutada $q(D|(i, j), z, x_{i+1:n}, y_{j+1:m})$.

Kui joonduste ahel on defineeritud esialgsete tõenäosuste (1) kaudu, siis võib ülaltoodud rekursioonis kasutada tõenäosuste $q(z|(i, j), z')$ asemel esialgseid tõenäosusi $p_{z'z}$ (ja algtõenäosusi π), β-rekursiooni tulemus küll muutub, kuid $q(D|(i, j), z, x_{1:n}, y_{1:m})$ jäävad samaks. Veendu selles!

α-rekursioon. Antud tõenäosused (6) ning emissioonitõenäosused. Eesmärk arvutada:

$$\alpha((i, j), z) := \sum_{z_{1:k} \in \mathcal{A}(i,j): z_k = z} p(x_{1:i}, y_{1:j}, z_{1:k}).$$

Ülesanne: Leia rekursioon.

Idee:

$$\alpha((i, j), D) = \big(\sum_z \alpha((i-1, j), z)q(D|(i-1, j), z)\big)p(x_i, -).$$

Kas rekursiooni tulemus muutub, kui q-tõenäosuste asemel kasutada p-tõenäosusi? Skaleerimine.

Serva (a,b) läbimise tõenäosus. Olgu $z = z(a,b)$.

$$p\big((a,b)|x_{1:n},y_{1:m}\big) = \frac{\alpha((b^x,b^y),z)p(x_{i+1:n},y_{j+1:m}|(i,j),z)}{p(x_{1:n},y_{1:m})} = \frac{\alpha((b^x,b^y),z)p(x_{i+1:n},y_{j+1:m}|(i,j),z)}{\sum_{z'}\alpha((b^x,b^y),z')p(x_{i+1:n},y_{j+1:m}|(i,j),z')}.$$

Seega tuleb läbi viia nii α kui β-rekursioon. Veendu, et ülaltoodud valem kehtib. Näita, et $p(a,b|x_{1:n},y_{1:m})$ ei muutu, kui q-tõenäosuste asemel kasutada p tõenäosusi ja seega võib ülaltoodud α ja β rekursioonis asendada q-tõenäosused p-tõenäosustega.

Alternatiiv: arvutada tinglikud tõenäosused $q(z'|(i,j),z,x_{1:n},y_{1:m})$ ja siis arvutada kõikide paaride tinglikud tõenäosused (0,0)-st alates rekursiivselt.

Hübriidjoondus. Optimeerimisülesanne

$$\arg\max_{a_{0:k}\in\mathcal{A}(n,m)}\Big(B\sum_{t=0}^{k-1}\ln p((a_t,a_{t+1})|x_{1:n},y_{1:m}) + C\ln p(a_{0:k}|x_{1:n},y_{1:m})\Big) =$$
$$\arg\max_{a_{0:k}\in\mathcal{A}(n,m)}\Big(B\sum_{t=0}^{k-1}\ln p((a_t,a_{t+1})|x_{1:n},y_{1:m}) + C\ln p(x_{1:n},y_{1:m},a_{0:k})\Big)$$

kuus $C,B\geq 0$ on regulariseerivad konstandid. Veendu, et üleande lahend ei muutu, kui tinglike tõenäosuste arvutamisel q-tõenäosused asendada p-tõenäosustega. Lahendamiseks tuleb arvutada kõikide servade läbimise tõenäosused $p\big((a,b)|x_{1:n},y_{1:m}\big)$ ja siis on mitu varianti:

- Leida tõenäosused $\{q(z'|(i,j),z,x_{1:n},y_{1:m})\}$ ja siis kasutada sama algoritmi, mis ülesande (5) lahendamisel
- Kasutada tõenäosusi (6). Rekursiooni idee (kontrolli!):

$$\begin{aligned}\delta(i,j,D) = \max_z\big(\delta(i-1,j,z) + C\ln q((i,j)|(i-1,j),z)\big) + \ln p(x_i.-)\\ + B\ln q((i-1,j),(i,j)|x_{1:n},y_{1:m}).\end{aligned}$$

Kui tõenäosused on antud esialgsel viisil (1), siis ülaltoodud rekirsioonis võib üleminekutõenäosused $q((i,j)|(i-1,j),z)$ asendada p_{zD}-ga.

Ülesanne: Tööta välja rekursioonid hübriidjoonduste leidmiseks.

Suurima tõepäraga ehk Viterbi joondus. Kui hübriidjoonduses võtta $B=0$, saame Viterbi joonduse.

Näide 1: Olgu emissiooni- ja üleminekutõenäosused

$$p(x,y) = \begin{cases}0, & \text{kui } x\neq y;\\ 1/4, & \text{kui } x=y.\end{cases}, \quad p(x,-)=p(-,y)=\frac{1}{4}, \quad \forall x,y \quad P = \begin{pmatrix}\frac{1}{3}&\frac{1}{3}&\frac{1}{3}\\ \frac{1}{3}&\frac{1}{3}&\frac{1}{3}\\ \frac{1}{3}&\frac{1}{3}&\frac{1}{3}\end{pmatrix} \quad \pi = \begin{pmatrix}\frac{1}{3}\\ \frac{1}{3}\\ \frac{1}{3}\end{pmatrix}.$$

Nüüd iga joonduse korral $p(z_{1:k}) = 3^{-k}$,

$$p(x_{1:n}, y_{1:m}|z_{1:k}) = \begin{cases} 0, & \text{kui leidub } t \text{ nii, et } x'_t \neq y'_t,\ x'_t, y'_t \in \mathcal{X} \text{ (mismatch);} \\ 4^{-k}, & \text{mujal.} \end{cases}$$

Üheski positiivse ühistõenäosusega joonduses pole mismatchi, ja selliste joonduste seast on suurima (ühis)tõenäosusega lühimad joondused. Näita, et need vastavad pikimale ühisjadale – M-ide arv Viterbi joonduses on pikima ühisjada pikkus.

Näide 2: Olgu emissioonitõenäosused nagu enne, üleminekumaatriks olgu $\delta < 1/4$ ja

$$\begin{pmatrix} 1-2\delta & \delta & \delta \\ \delta & 1-2\delta & \delta \\ \delta & \delta & 1-2\delta \end{pmatrix}.$$

Selline üleminekumaatriks soosib seda, et sarnased seisundid on rohkem blokikaupa. Näiteks bioloogias eelistatakse, et indeld on blokiti.
Olgu $x_{1:4} = (1, 0, 0, 0)$ ja $y_{1:4} = (1, 1, 0, 1)$. Võrdleme joondusi:

1	-	0	0	0	-	-	1	0	0	0	-	-	1	0	0	0	-	-	1	0	0	0	-
1	1	-	-	0	1	1	1	0	-	-	1	1	1	-	0	-	1	1	1	-	-	0	1
M	I	D	D	M	I	I	M	M	D	D	I	I	M	D	M	D	I	I	M	D	D	M	I

Kõikide nende joonduste korral

$$p(x_{1:4}, y_{1:4}|z_{1:6}) = \frac{1}{4^6},$$

kuid joonduste tõenäosused on erinevad:

$$\ln P(M, I, D, D, M, I) = \ln \frac{1}{3} + \ln(\delta) + \ln(\delta) + \ln(1-\delta) + \ln(\delta) + \ln(\delta)$$
$$\ln P(I, M, M, D, D, I) = \ln \frac{1}{3} + \ln(\delta) + \ln(1-\delta) + \ln(\delta) + \ln(1-\delta) + \ln(\delta)$$
$$\ln P(I, M, D, M, D, I) = \ln \frac{1}{3} + \ln(\delta) + \ln(\delta) + \ln(\delta) + \ln(\delta) + \ln(\delta)$$
$$\ln P(I, M, D, D, M, I) = \ln \frac{1}{3} + \ln(\delta) + \ln(\delta) + \ln(1-\delta) + \ln(\delta) + \ln(\delta)$$

Näeme, et teise joonduse (I, M, M, D, D, I) tõenäosus in suurim (neli blokki: "I", "MM", "DD", "I"; teistes joondustes rohkem blokke). Kõik ülaltoodud joondused vastavad pikimale ühisjadale, kas aga antud juhul pikim ühisjada on Viterbi joondus? Vaatleme veel ühte joondust:

-	-	1	0	0	0	-	-
1	1	-	-	-	-	0	1
I	I	D	D	D	D	I	I

Selle joonduse korral $p(x_{1:4}, y_{1:4}|z_{1:8}) = \frac{1}{4^8}$, ning

$$p(I, I, D, D, D, D, I, I) = \frac{1}{3} \cdot (1-\delta)^5 \cdot \delta^2.$$

Seega ühistõenäosuste suhe ($z_{1:6} = (I, M, M, D, D, I)$, $z_{1:8} = (I, I, D, D, D, D, I, I)$)

$$\frac{p(x_{1:4}, y_{1:4}, z_{1:8})}{p(x_{1:4}, y_{1:4}, z_{1:6})} = \frac{(1-\delta)^3}{4\delta} > 1.$$

kui δ on piisavalt väike. Seega indelid on kõik kenasti blokikaupa kuid Viterbi joondus ei vasta pikimale ühisjadale. Leia antud näites Viterbi joondus.

Tee näide läbi $\delta = 1/4$ korral. Kas siis viterbi joondus vastab pikimale ühisjadale (tõestus).

Probleem: Leia piisavad (ja tarvilikud) tingimused (üleminekumaatriksi ja emissioonitõenäosuste (1) kaudu), mis garanteerivad, et Viterbi joondus vastaks pikimale ühisjadale ja kõikide pikimale ühisjadale vastavate joonduste seast valiks Viterbi joondus sellise kus indelid oleks võimalikult blokiti. Kas neid tungimusi on võimalik leida üleminekutõenäosuste (6) kaudu?

Kirjandus

Vimm, K (2022). *Suboptimaalsed meetodid jadade sarnasuse võrdlemiseks kahetähelise tähestiku puhul.* Bakalureusetöö. TÜ

Koski. T. (2001). *Hidden Markov Models for Bioinformatics.* Springer

Durbin, R, Eddy, S., Krogh, A., Mitchinson, G (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* J. Wiley sons.