עיבוד שפה טבעית - תרגיל 2 - חלק תיאורטי

מגישים: בר רוזן 203765698, נועה רפפורט 318418837

א) תא יכול להיות באחד משני מצבים: $S \in \{H, L\}$

תא יכול לצור נוקלאוטיד אחד בכל מצב S: $nu \in \{A, C, T, G\}$

נתון: עבור מצב S_i בזמן i:

$$P_r(S_{i+1}=L \mid S_i=H)=0.5 \qquad P_r(S_{i+1}=H \mid S_i=L)=0.4$$
$$P_r(S_{i+1}=H \mid S_i=H)=0.5 \qquad P_r(S_{i+1}=L \mid S_i=L)=0.6$$

$$P_r(nu_i=A \mid S_i=L)=0.3 \qquad P_r(nu_i=A \mid S_i=H)=0.2$$
$$P_r(nu_i=C \mid S_i=L)=0.2 \qquad P_r(nu_i=C \mid S_i=H)=0.3$$
$$P_r(nu_i=G \mid S_i=L)=0.2 \qquad P_r(nu_i=G \mid S_i=H)=0.3$$
$$P_r(nu_i=T \mid S_i=L)=0.3 \qquad P_r(nu_i=T \mid S_i=H)=0.2$$

עבור סדרת הנוקלאוטידים: $NU = A,C,C,G,T,G,C,A$

נמצא את סדרת המצבים הסבירה ביותר $S_1,\dots,S_8$

בהינתן ש $S_0=H$ ונחשב את ההסתברות למעברי מצבים כך.

כלומר שנרצה לפתור את הבעיה:

$$\underset{S_1,\dots,S_8}{\arg\max}\ P_r(NU_1=A, NU_2=C, NU_3=C, NU_4=G, NU_5=T, NU_6=G, NU_7=C, NU_8=A, S_0=H, S_1, S_2, S_3, S_4, S_5, S_6, S_7, S_8)$$

$$= \underset{S_1,\dots,S_8}{\arg\max}\ P_r(S_1 \mid S_0=H)\cdot \prod_{i=2}^{8} P_r(S_i \mid S_{i-1})\cdot P_r(NU_i \mid S_i)$$

נשתמש באלגוריתם Viterbi כדי לפתור את הבעיה:

1:
$$\begin{cases} \pi(1, S_0=H, S_1=L) = P_0(L|H)\, P_0(A|L) = 0.5 \cdot 0.3 = 0.15 \\ \pi(1, S_0=H, S_1=H) = P_0(H|H) \cdot P_0(A|H) = 0.5 \cdot 0.2 = 0.1 \end{cases}$$

2:
$$\begin{cases} \pi(2, L, L) = \pi(1, H, L) \cdot P_0(L|L) \cdot P_0(C|L) = 0.15 \cdot 0.6 \cdot 0.2 = 0.018 \\ \pi(2, L, H) = \pi(1, H, L) \cdot P_0(H|L) \cdot P_0(C|H) = 0.15 \cdot 0.4 \cdot 0.3 = 0.018 \\ \pi(2, H, L) = \pi(1, H, H) \cdot P_0(L|H) \cdot P(C|L) = 0.1 \cdot 0.5 \cdot 0.2 = 0.01 \\ \pi(2, H, H) = \pi(1, H, H) \cdot P_0(H|H) \cdot P_0(C|H) = 0.1 \cdot 0.5 \cdot 0.3 = 0.015 \end{cases}$$

3:
$$\begin{cases} \pi(3, L, L) = \max\{\pi(2, L, L), \pi(2, H, L)\} \cdot P_0(L|L) \cdot P_0(C|L) = 2.16 \cdot 10^{-3} \\ \pi(3, L, H) = \max\{\pi(2, L, L), \pi(2, H, L)\} \cdot P_0(H|L) \cdot P_0(C|H) = 2.16 \cdot 10^{-3} \\ \pi(3, H, L) = \max\{\pi(2, L, H), \pi(2, H, H)\} \cdot P_0(L|H) \cdot P_0(C|L) = 1.8 \cdot 10^{-3} \\ \pi(3, H, H) = \max\{\pi(2, L, H), \pi(2, H, H)\} \cdot P_0(H|H) \cdot P_0(C|H) = 2.7 \cdot 10^{-3} \end{cases}$$

4:
$$\begin{cases} \pi(4, L, L) = \max\{\pi(3, L, L), \pi(3, H, L)\} \cdot P_0(L|L) \cdot P_0(G|L) = 2.592 \cdot 10^{-4} \\ \pi(4, L, H) = \max\{\pi(3, L, L), \pi(3, H, L)\} \cdot P_0(H|L) \cdot P_0(G|H) = 2.592 \cdot 10^{-4} \\ \pi(4, H, L) = \max\{\pi(3, L, H), \pi(3, H, H)\} \cdot P_0(L|H) \cdot P_0(G|L) = 2.7 \cdot 10^{-4} \\ \pi(4, H, H) = \max\{\pi(3, L, H), \pi(3, H, H)\} \cdot P_0(H|H) \cdot P_0(G|H) = 4.05 \cdot 10^{-4} \end{cases}$$

5:
$$\begin{cases} \pi(5, L, L) = \max\{\pi(4, L, L), \pi(4, H, L)\} \cdot P_0(L|L) \cdot P_0(T|L) = 4.86 \cdot 10^{-5} \\ \pi(5, L, H) = \max\{\pi(4, L, L), \pi(4, H, L)\} \cdot P_0(H|L) \cdot P_0(T|H) = 2.16 \cdot 10^{-5} \\ \pi(5, H, L) = \max\{\pi(4, L, H), \pi(4, H, H)\} \cdot P_0(L|H) \cdot P_0(T|L) = 6.075 \cdot 10^{-5} \\ \pi(5, H, H) = \max\{\pi(4, L, H), \pi(4, H, H)\} \cdot P_0(H|H) \cdot P_0(T|H) = 4.05 \cdot 10^{-5} \end{cases}$$

6:
$$\begin{cases} \pi(6, L, L) = \max\{\pi(5, L, L), \pi(5, H, L)\} \cdot P_0(L|L) \cdot P_0(G|L) = 7.29 \cdot 10^{-6} \\ \pi(6, L, H) = \max\{\pi(5, L, L), \pi(5, H, L)\} \cdot P_0(H|L) \cdot P_0(G|H) = 7.29 \cdot 10^{-6} \\ \pi(6, H, L) = \max\{\pi(5, L, H), \pi(5, H, H)\} \cdot P_0(L|H) \cdot P_0(G|L) = 4.05 \cdot 10^{-6} \\ \pi(6, H, H) = \max\{\pi(5, L, H), \pi(5, H, H)\} \cdot P_0(H|H) \cdot P_0(G|H) = 6.075 \cdot 10^{-6} \end{cases}$$

4:
$$\begin{cases} \pi(7,L,L) = \max\{\pi(6,L,L), \pi(6,H,L)\} \cdot P_8(L|L) \cdot P(C|L) = 8.748 \cdot 10^{-7} \\ \pi(7,L,H) = \max\{\pi(6,L,L), \pi(6,H,L)\} \cdot P_8(H|L) \cdot P_8(C|H) = 8.748 \cdot 10^{-7} \\ \pi(7,H,L) = \max\{\pi(6,L,H), \pi(6,H,H)\} \cdot P_8(L|H) \cdot P_8(C|L) = 7.29 \cdot 10^{-7} \\ \pi(7,H,H) = \max\{\pi(6,L,H), \pi(6,H,H)\} \cdot P_8(H|H) \, P_8(C|H) = 1.0935 \cdot 10^{-6} \end{cases}$$

3:
$$\begin{cases} \pi(8,L,L) = \max\{\pi(7,L,L), \pi(7,H,L)\} \cdot P_8(L|L) \cdot P_8(A|L) = 1.57464 \cdot 10^{-7} \\ \pi(8,L,H) = \max\{\pi(7,L,L), \pi(7,H,L)\} \cdot P_8(H|L) \cdot P_8(A|H) = 6.9984 \cdot 10^{-8} \\ \pi(8,H,L) = \max\{\pi(7,L,H), \pi(7,H,H)\} \cdot P_8(L|H) \cdot P_8(A|L) = 1.64025 \cdot 10^{-7} \\ \pi(8,H,H) = \max\{\pi(7,L,H), \pi(7,H,H)\} \cdot P_8(H|H) \cdot P_8(A|H) = 1.0935 \cdot 10^{-7} \end{cases}$$

$$\max\{\pi(8,L,L), \pi(8,L,H), \pi(8,H,L), \pi(8,H,H)\} = \pi(8,H,L) =$$
$$= 1.64025 \cdot 10^{-7}$$

נקבל אחורה ונקבל:

$$\pi(8,H,L) \Rightarrow S_8 = L, \; S_7 = H$$

$$\max\{\pi(7,L,H), \pi(7,H,H)\} = \pi(7,H,H) \Rightarrow S_6 = H$$
$$\max\{\pi(6,L,H), \pi(6,H,H)\} = \pi(6,L,H) \Rightarrow S_5 = L$$
$$\max\{\pi(5,L,L), \pi(5,H,L)\} = \pi(5,H,L) \Rightarrow S_4 = H$$
$$\max\{\pi(4,L,H), \pi(4,H,H)\} = \pi(4,H,H) \Rightarrow S_3 = H$$
$$\max\{\pi(3,L,H), \pi(3,H,H)\} = \pi(3,H,H) \Rightarrow S_2 = H$$
$$\max\{\pi(2,L,H), \pi(2,H,H)\} = \pi(2,L,H) \Rightarrow S_1 = L$$

כלומר הרצף ה"סודי" הוא:

$$\arg\max_{S_1,\ldots,S_8} P_1(A,C,C,G,T,G,C,A,S_0=H,S_1,S_2,S_3,\ldots,S_8) =$$
$$= S_1=L, S_2=H, S_3=H, S_4=H, S_5=L, S_6=H, S_7=H, S_8=L$$

והסתברות הרצף הזה היא $1.64025 \cdot 10^{-7}$.

Input: Integer n,

Parameters: $q(w|x,x,x)$, $q(w|x,x,v)$, $q(w|x,u,v)$, $q(w|t,u,v)$, $q(STOP|t,u,v)$ $\forall\ w,t,u,v \in K$

$e(x|s) \quad \forall\ x \in V,\ s \in K$

Definitions: K = set of all possible tags

$K_{-2} = K_{-1} = K_0 = \{*\}$, $K_k = K\ \forall\ 1 \le k \le n$

$K_{n+1} = \{STOP\}$

V = set of all possible words

$\forall\ 1 \le k \le n$ $\pi(k,u,v)$ = maximum probability of a tag sequence ending in tags u,v at position k

Initialization: $\pi(-1,*,*) = \pi(0,*,*) = 1$

$bp(0,*,*) = *$

Algorithm: For $k = 1, \dots, n$:

For $u \in K_{k-1}$, $v \in K_k$:

$$\pi(k,u,v) = \max_{\substack{y \in K_{k-2} \\ x \in V}} \{\pi(k-1,y,u) \cdot q(v|bp(k-1,x,u)[0],y,u) \cdot e(x|v)\}$$

$$bp(k,u,v) = [y_{k-1}, x_k] = \arg\max_{\substack{y \in K_{k-1} \\ x \in V}} \{\pi(k-1,y,u) \cdot q(v|bp(k-1,w,u)[0],w,u) \cdot e(x|v)\}$$

$$\text{Set } (y_n, y_{n-1}) = \arg\max_{\substack{u \in K_{n-1} \\ v \in K_n}} \{\pi(n,u,v) \cdot q(STOP|bp(n,u,v)[0],u,v)\}$$

FOR $k = n-2, \ldots, 1$

$y_k = bp(k+2, y_{k+1}, y_{k+2})[0]$

$x_k = bp(k+2, y_{k+1}, y_{k+2})[1]$

return $y_1, \ldots, y_n, x_1, \ldots, x_n$

NLP - Exercise 2 - Practical part

Bar Rousso 203765698, Noa Rapoport 318418837

Results for question 3 (b ii):

MLE Error rate for known words: 0.07
MLE Error rate for un-known words: 0.743
MLE General error rate: 0.147

Results for question 3 (c iii):

HMM-Bigram error rate for known words: 0.213
HMM-Bigram error rate for un-known words: 0.784
HMM-Bigram general error rate: 0.279

Results for question 3 (d ii):

HMM-Bigram-Laplace error rate for known words: 0.144
HMM-Bigram-Laplace error rate for un-known words: 0.743
HMM-Bigram-Laplace general error rate: 0.212

Results for question 3(e ii):

HMM-Bigram-Pseudo error rate for known words: 0.202
HMM-Bigram-Pseudo error rate for un-known words: 0.586
HMM-Bigram-Pseudo general error rate: 0.246

Results for question 3(e iii):

HMM-Bigram-Pseudo-Laplace error rate for known words: 0.141
HMM-Bigram-Pseudo-Laplace error rate for un-known words: 0.558
HMM-Bigram-Pseudo-Laplace general error rate: 0.188

10 most frequent errors:

1. True tag: NNS, Predicted tag: NN, Count: 333
2. True tag: NP, Predicted tag: NN, Count: 294
3. True tag: JJ, Predicted tag: NN, Count: 224
4. True tag: VB, Predicted tag: NN, Count: 107
5. True tag: VBN, Predicted tag: NN, Count: 91
6. True tag: VBG, Predicted tag: NN, Count: 88
7. True tag: VBD, Predicted tag: NN, Count: 78
8. True tag: RB, Predicted tag: NN, Count: 73

9. True tag: CD, Predicted tag: NN, Count: 70
10. True tag: TO, Predicted tag: IN, Count: 44