תרגיל 3 NLP

הדר באון - 207034182

ענבר בחנוף - 207009721

1. (10 pts) Consider this (toy) biological setup:
A cell can be in one of two states - H, for high GC-content, and L for low GC. On each time step the cell produces one nucleotide, A,C,T or G, and might also change its state. The probability of changing from state H to L is 0.5, and from state L to H is 0.4.
In state H the probabilities for producing nucleotides are 0.2 for A, 0.3 for C, 0.3 for G and 0.2 for T. In L the probabilities are 0.3 for A, 0.2 for C, 0.2 for G and 0.3 for T.
Consider the nucleotide sequence $S = ACCGTGCA$. Use the Viterbi algorithm to find the best state-sequence and calculate the probability of S given this state-sequence. Assume the previous state before S was H.

$$\text{tansitions} = \begin{cases} P(H|H) = 0.5 & P(H|L) = 0.4 \\ P(L|H) = 0.5 & P(L|L) = 0.6 \end{cases}$$

$$\text{emmitions} = \begin{cases} P(A|H) = 0.2 & P(C|H) = 0.3 & P(G|H) = 0.3 & P(T|H) = 0.2 \\ P(A|L) = 0.3 & P(C|L) = 0.2 & P(G|L) = 0.2 & P(T|L) = 0.3 \end{cases}$$

$S = \overset{1}{A}\overset{2}{C}\overset{3}{C}\overset{4}{G}\overset{5}{T}\overset{6}{G}\overset{7}{C}\overset{8}{A}$, $P(\text{start with } H) = 1$

$V_{1H} = P(H|H) \cdot P(A|H) = 0.2 \cdot 0.5 = 0.1$ – כיון - המצב ההתחלתי הוא H

$V_{1L} = P(L|H) \cdot P(A|L) = 0.3 \cdot 0.5 = 0.15$

$$V_{2H} = \max\left(\overset{0.1}{V_{1H}} \cdot \overset{0.5}{P(H|H)}, \overset{0.15}{V_{1L}} \cdot \overset{0.4}{P(H|L)}\right) \cdot \overset{0.3}{P(C|H)} = 0.018$$

$$V_{2L} = \max\left(\overset{0.1}{V_{1H}} \cdot \overset{0.5}{P(L|H)}, \overset{0.15}{V_{1L}} \cdot \overset{0.6}{P(L|L)}\right) \cdot \overset{0.2}{P(C|L)} = 0.09 \cdot 0.2 = 0.018$$

$$V_{3H} = \max\left(\overset{0.018}{V_{2H}} \cdot \overset{0.5}{P(H|H)}, \overset{0.018}{V_{2L}} \cdot \overset{0.4}{P(H|L)}\right) \cdot \overset{0.3}{P(C|H)} = 2.7 \cdot 10^{-3}$$

$$V_{3L} = \max\left(\overset{0.018}{V_{2H}} \cdot \overset{0.5}{P(L|H)}, \overset{0.018}{V_{2L}} \cdot \overset{0.6}{P(L|L)}\right) \cdot \overset{0.2}{P(C|L)} = 2.16 \cdot 10^{-3}$$

$$V_{4H} = \max\left(\overset{2.7\cdot10^{-3}}{V_{3H}} \cdot P(\overset{0.5}{H}|H), \overset{2.16\cdot10^{-3}}{V_{3L}} \cdot P(\overset{0.4}{H}|L)\right) \cdot P(\overset{0.3}{G}|H) = 4.05 \cdot 10^{-4}$$

$$V_{4L} = \max\left(V_{3H} \cdot P(\overset{0.5}{L}|H), V_{3L} \cdot P(\overset{0.6}{L}|L)\right) \cdot P(\overset{0.2}{G}|L) = 2.7 \cdot 10^{-4}$$

$$V_{5H} = \max\left(\overset{4.05\cdot10^{-4}}{V_{4H}} \cdot P(\overset{0.5}{H}|H), \overset{2.7\cdot10^{-4}}{V_{4L}} \cdot P(\overset{0.4}{H}|L)\right) \cdot P(\overset{0.2}{T}|H) = 4.05 \cdot 10^{-5}$$

$$V_{5L} = \max\left(V_{4H} \cdot P(\overset{0.5}{L}|H), V_{4L} \cdot P(\overset{0.6}{L}|L)\right) \cdot P(\overset{0.3}{T}|L) = 6.075 \cdot 10^{-5}$$

$$V_{6H} = \max\left(\overset{4.05\cdot10^{-5}}{V_{5H}} \cdot P(\overset{0.5}{H}|H), \overset{6.075\cdot10^{-5}}{V_{5L}} \cdot P(\overset{0.4}{H}|L)\right) \cdot P(\overset{0.3}{G}|H) = 7.29 \cdot 10^{-6}$$

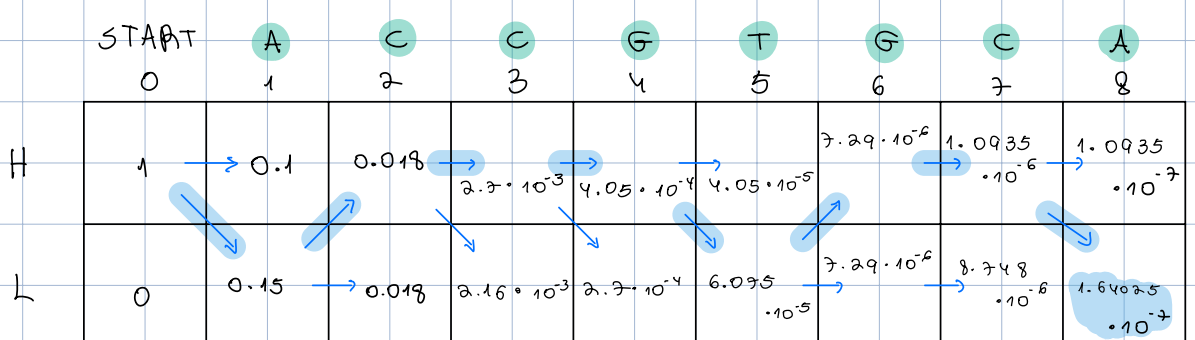
$$V_{6L} = \max\left(V_{5H} \cdot P(\overset{0.5}{L}|H), V_{5L} \cdot P(\overset{0.6}{L}|L)\right) \cdot P(\overset{0.2}{G}|L) = 7.29 \cdot 10^{-6}$$

$$V_{7H} = \max\left(\overset{7.29\cdot10^{-6}}{V_{6H}} \cdot P(\overset{0.5}{H}|H), \overset{7.29\cdot10^{-6}}{V_{6L}} \cdot P(\overset{0.4}{H}|L)\right) \cdot P(\overset{0.3}{C}|H) = 1.0935 \cdot 10^{-6}$$

$$V_{7L} = \max\left(V_{6H} \cdot P(\overset{0.5}{L}|H), V_{6L} \cdot P(\overset{0.6}{L}|L)\right) \cdot P(\overset{0.2}{C}|L) = 8.748 \cdot 10^{-6}$$

$$V_{8H} = \max\left(\overset{1.0935\cdot10^{-7}}{V_{7H}} \cdot P(\overset{0.5}{H}|H), \overset{8.748\cdot10^{-6}}{V_{7L}} \cdot P(\overset{0.4}{H}|L)\right) \cdot P(\overset{0.2}{A}|H) = 1.0935 \cdot 10^{-7}$$

$$V_{8L} = \max\left(V_{7H} \cdot P(\overset{0.5}{L}|H), V_{7L} \cdot P(\overset{0.6}{L}|L)\right) \cdot P(\overset{0.3}{A}|L) = 1.64025 \cdot 10^{-7}$$



2. (10 pts) In class we saw the trigram HMM model and the corresponding Viterbi algorithm. We will now make two main changes. First, we will consider a four-gram tagger, where p takes the form:

$$p(x_1 \cdots x_n, y_1 \cdots y_{n+1}) = \prod_{i=1}^{n+1} q(y_i | y_{i-3}, y_{i-2}, y_{i-1}) \prod_{i=1}^{n} e(x_i | y_i) \quad (1)$$

We assume in this definition that $y_0 = y_{-1} = y_{-2} = *$, where $*$ is the START symbol, $y_{n+1} = STOP$, and $y_i \in \mathcal{K}$ for $i = 1 \cdots n$, where $\mathcal{K}$ is the set of possible tags in the HMM.
Second, we consider a version of the Viterbi algorithm that takes as input **an integer** n (and not a sentence $x_1 \cdots x_n$ as we saw in class) and finds

$$\max_{y_1 \cdots y_{n+1}, x_1 \cdots x_n} p(x_1 \cdots x_n, y_1 \cdots y_{n+1})$$

for a four-gram tagger, as defined in Equation 1. $x_1 \cdots x_n$ may range over the values of some fixed vocabulary $\mathcal{V}$. Complete the following pseudo-code of this version of the Viterbi algorithm for this model . The pseudo-code must be efficient.

Input: An integer n, parameters $q(w|t,u,v)$ and $e(x|s)$.
Definitions: Define $\mathcal{K}$ to be the set of possible tags. Define $\mathcal{K}_{-2} = \mathcal{K}_{-1} = \mathcal{K}_0 = \{*\}$, and $\mathcal{K}_k = \mathcal{K}$ for $k = 1 \cdots n$. Define $\mathcal{V}$ to be the set of possible words.

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

Algorithm:

For $i = 1, \ldots, n$

For $u \in \mathcal{K}_{i-2}$, $v \in \mathcal{K}_{i-1}$, $w \in \mathcal{K}_i$

$$\pi(i, u, v, w) = \max_{s \in \mathcal{K}_{i-3}, x \in \mathcal{V}} \{\pi(i-1, s, u, v) \times q(w | s, u, v) \times e(x_i | w)\}$$

$$dp(i, u, v, w) = \arg\max_{s \in \mathcal{K}_{i-3}, x \in \mathcal{V}} \{\pi(i-1, s, u, v) \times q(w | s, u, v) \times e(x_i | w)\}$$

Set $(y_{n-2}, y_{n-1}, y_n) = \arg\max \{\pi(n, u, v, w) \cdot p(STOP | u, v, w)\}$

denote $dp(k, x, y, z)[0] \in \mathcal{K}$, $dp(k, x, y, z)[1] \in \mathcal{V}$

For $k = 1, \ldots, n-3$

$y_k = dp(k+3, y_{k+1}, y_{k+2}, y_{k+3})[0]$

For $k = 1, \ldots, n$

$x_k = dp(k, y_{k-2}, y_{k-1}, y_k)[1]$

Return: $(x_1, \ldots, x_n, y_1, \ldots y_n)$

שאלה 3

```
Q3b (ii)
Known words error rate: 0.0814
Unknown words error rate: 0.7897
Overall error rate: 0.1623
```

```
Q3c (iii) viterbi algorithm
Bigram HMM Error Rates (Viterbi on Test Set):
Known words error rate: 0.2755
Unknown words error rate: 0.8316
Overall error rate: 0.3390
```

```
Q3d (ii) viterbi algorithm with add-1 smoothing
Bigram HMM Error Rates (Viterbi on Test Set):
Known words error rate: 0.7876
Unknown words error rate: 1.0000
Overall error rate: 0.8118
```

```
Q3e (ii) viterbi algorithm with pseudo words
Bigram HMM Error Rates (Viterbi on Test Set):
Known words error rate: 0.2774
Unknown words error rate: 0.2705
Overall error rate: 0.2766
```

```
Q3e (iii)  viterbi algorithm with pseudo words and add-1 smoothing
Bigram HMM Error Rates (Viterbi on Test Set):
Known words error rate: 0.1329
Unknown words error rate: 0.6780
Overall error rate: 0.1952
```