Fall 2020 CIS 262

Automata, Computability and Complexity
Jean Gallier

Homework 6

February 25 2020; Due March 5, 2020, beginning of class

"B problems" must be turned in.

Problem B1 (100 pts). (1) Implement the Viterbi algorithm, as described in Section 4.2 of the notes. **Do not implement any other version of the Viterbi algorithm found on the web or anywhere else**.

The input should consist of the matrices A (an $n \times n$ matrix), B (an $n \times m$ matrix), the vector π (of dimension n), and a sequence $(\omega_1, \ldots, \omega_T)$ of length T consisting of the indices associated with the observation sequence $(O_1, \ldots, O_T)$ (given by the bijection $\omega \colon \mathbb{O} \to \{1, \ldots, m\}$).

The output should be

(a) The sequence $(i_1, \ldots, i_T)$ of indices associated with the state sequence $(q_1, \ldots, q_T)$ (given by the bijection $\sigma \colon Q \to \{1, \ldots, m\}$) that yields the highest probability of producing the observation sequence $(O_1, \ldots, O_T)$,

(b) The highest probability $maxscore = \max_{1 \leq j \leq n} score(j, T)$ found at time T.

In Example 4.1 of the notes, we have $Q = \{\text{Cold}, \text{Hot}\}$, the bijection σ is given by $\sigma(\text{Cold}) = 1$ and $\sigma(\text{Hot}) = 2$, the output alphabet is $\mathbb{O} = \{\text{N}, \text{D}\}$, and the bijection ω is given by $\omega(\text{N}) = 1$, and $\omega(\text{D}) = 2$.

The output sequence NNND corresponds to the sequence $(1, 1, 1, 2)$, and the state sequence $(\text{Hot}, \text{Cold}, \text{Cold}, \text{Hot})$ corresponds to $(2, 1, 1, 2)$.

The matrices A, B and the vector π are given in the notes.

Test your program on the HMM of Example 4.1 of the notes for the following observation sequences:

1. NNND (**2 points**)
2. NNNDN (**2 points**)
3. NNNDNN (**2 points**)

4. NNNDNDDN (**2 points**)

   In all four cases, print the most likely sequence of states.

5. The sequence of length 1200 consisting of the following four blocks: (**3 points**)

$$\underbrace{\mathrm{N}\cdots\mathrm{N}}_{300}\underbrace{\mathrm{D}\cdots\mathrm{D}}_{300}\underbrace{\mathrm{N}\cdots\mathrm{N}}_{300}\underbrace{\mathrm{D}\cdots\mathrm{D}}_{300}$$

   Print states q_1-q_5, q$_{300}$-q_{304}, q_{600}-q_{604}, q_{900}-q_{904}, and q_{1196}-q_{1200}.

6. The sequence of length 2000 consisting of the following four blocks: (**3 points**)

$$\underbrace{\mathrm{N}\cdots\mathrm{N}}_{500}\underbrace{\mathrm{D}\cdots\mathrm{D}}_{500}\underbrace{\mathrm{N}\cdots\mathrm{N}}_{500}\underbrace{\mathrm{D}\cdots\mathrm{D}}_{500}$$

   Print states q_1-q_5, q$_{500}$-q_{504}, q_{1000}-q_{1004}, q_{1500}-q_{1504}, and q_{1996}-q_{2000}.

7. The sequence of length 2004 consisting of the following four blocks, followed by NNND: (**3 points**)

$$\underbrace{\mathrm{N}\cdots\mathrm{N}}_{500}\underbrace{\mathrm{D}\cdots\mathrm{D}}_{500}\underbrace{\mathrm{N}\cdots\mathrm{N}}_{500}\underbrace{\mathrm{D}\cdots\mathrm{D}}_{500}\mathrm{NNND}$$

   Print states q_1-q_5, q$_{500}$-q_{504}, q_{1000}-q_{1004}, q_{1500}-q_{1504}, and q_{1999}-q_{2004}.

For the output sequences in (6) and (7) you will find *maxscore* $= 0$, which means that the numbers are smaller than machine precision; you run into *underflow*.

(2) To overcome underflow, modify your program by using logarithms as suggested in Section 4.2 of the notes.

Run your new version of Viterbi on the sequences (5), (6), (7) of part (1). (**9 points**)

Hint. In (6), you should expect that *maxscore* $= -1.2275e{+}03$, and in (7), that *maxscore* $= -1.2318e + 03$.

(3) Consider the example of an HMM given online as Example-Viterbi-DNA. The set of states is $Q = \{\mathrm{L}, \mathrm{H}\}$, and the set of outputs is $\mathbb{O} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$. Assume we use the bijections $\sigma\colon Q \to \{1, 2\}$ and $\omega\colon \mathbb{O} \to \{1, 2, 3, 4\}$ given by $\sigma(\mathrm{H}) = 1$, $\sigma(\mathrm{L}) = 2$, and by $\omega(\mathrm{A}) = 1$, $\omega(\mathrm{C}) = 2$, $\omega(\mathrm{G}) = 3$, and $\omega(\mathrm{T}) = 4$. Then the probability matrices are

$$A = \begin{pmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \end{pmatrix}, \quad B = \begin{pmatrix} 0.2 & 0.3 & 0.3 & 0.2 \\ 0.3 & 0.2 & 0.2 & 0.3 \end{pmatrix}, \quad \pi = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}.$$

Verify that the sequence GGCACTGAA corresponds to the state sequence HHHLLLLLL, and that the corresponding probability is $4.2515e - 08$. (**2 points**)

Which state sequence corresponds to the DNA sequence GAGATATACATAGAATTACG, and what is the corresponding highest probability? (**2 points**)

Run both versions of your Viterbi and compare the highest probabilities. By taking the exponential of the value given by the second version you should get the probability given by the first version (not using logs). (**5 points**)

TOTAL: 100 points