COL 776: Assignment 2 (Part B)

Due: Friday October 16, 12 noon

Updated: 12:40 pm, Tuesday Oct 20

Max Points: 24 (+ 10 extra credit)

- You should submit all your code as well as any graphs that you might plot (see below).
- Include a **single write-up (pdf) file** which includes a brief description for each question explaining what you did. Include any observations and/or plots required by the question in this single write-up file.
- You can use any programming language from the set C++, Java, Python, Matlab. If you would like to use any other language, please check with us before you start.
- Your code should have appropriate documentation for readability.
- You will be graded based on what you submit as well as your ability to explain your code.
- Refer to the course website for assignment submission instructions.
- This assignment is supposed to be done individually. You should carry out all the implementation by yourself.
- We plan to run Moss on the submissions. Any cheating will result in a penalty of **-10** points on your total course score (in addition to a 0 on the assignment). Stricter penalties (**including a fail grade**) may follow.

CpG Islands prediction using Hidden Markov Model: In this problem, we will play around with some biological data. As is common knowledge, genomic sequence consists of the base pairs A,G,C,T. Among these the "CG" combination is known to occur quite infrequently in a genome due to some chemical properties of the associated molecules. It is known that in the areas which codify genes, the chemical processes are altered leading to frequent occurrences of "CG" pairs. Such sequences with high concentration of "CG" pairs are known as CpG islands and are of importance since they can help narrow down the search for genes in the whole sequence.

You are given a long sequence of bases and a corresponding annotation which tells for each base whether it belongs a CpG island or not. Go through the readme file to understand the format of these files. You will model the given sequence as an HMM where the hidden states are a (base,sign) pair, where the base denotes one of A,G,C and T. Sign is '+' if the base came from a CpG island, '-' otherwise. Observation is simply the associated base i.e. A,G,C,T. Figure **??** shows the associated transition diagram between the 8 possible hidden states i.e. $\{A^+,G^+,C^+,T^+,A^-,G^-,C^-,T^-\}$. (for each base, there are 2 states for each sign). Note that observation set is simply {A,C,T,G}. You can learn the transition as well as emission probabilities (in the HMM model) from the given training sequence by simply counting the fraction of desired transitions among all possible transitions. For example, to calculate $P(Y^{t+1} = A^+|Y^t = C^-)$, ~~we will simply count the number of times this pair occurs in the sequence divided by the total number of pairs i.e. $T-1$ where T is the length of the sequence.~~ we will simply count the number of times this pair occurs in the sequence divided by the total number of pairs where the first element of the pair is C^-. Since both the hidden states A^+ and A^- can only result in the observation A, the associated emission probabilities are equal to 1 (the probability of emitting any other base is 0). i.e. $P(X^t = A|Y^t = A^+) = P(X^t = 1|Y^t = A^-) = 1$. And similarly for other bases. In order to calculate the probability of the start state, assume that the first base of the sequence is randomly chosen with the probability being proportional to the frequency of its occurrence in the non-CpG islands parts of the training sequecne (note that in this formulation, the first base can never belong a CpG island).

- (**20 points**) You are given another test sequence and your task in this problem is to find all the CpG islands in the sequence. To do this, you will perform the MPE inference using the HMM model learned

above. Specifically, you need to find the joint most likely state of all the hidden variables in the sequence, and correspondingly output all the CpG islands in the most likely state (given the observed sequence of bases).

Doing this naïvely would result in an exponential algorithm as we will need to try every possible state. Read about the Viterbi algorithm. See the links posted on website (you are also encouraged to search online for additional resources if required). Implement it to solve the above problem. Try it on the provided test data and produce an output file containing one CpG island per line (in the sequence in which they occur). Do not use any separators (e.g., ' ' ',' etc.) among the bases in your output file (this is to ensure a standard output format for easy grading). Also, output the log-probability of the most likely joint state of the hidden variables.

- (**4 points**) What is the connection of Viterbi with the max-product VE algorithm as discussed in class? Justify your answer.

Extra Fun and Learning (no formal credits): Suppose you are given a randomly drawn small test sequence which either completely lies inside a CpG island or lies completely outside a CpG island. How would you use the model above to calculate the probability that the given test sequence was in fact generated by a CpG island? Remember to take into account the probability of the start state. Assume that the start state was chosen randomly based on the frequency of occurrence in the training set (note that in this problem, the starting state can now come from a CpG island or not as the case may be). You can try your approach on the set of test sequences given along with the problem dataset. We will provide you the model answers at the time of evaluation.
Note: This problem requires very little extra coding effort.