CM20220: Fundamentals of Machine Learning

Lab Sheet 3: Hidden Markov Models

Deadline: 30th April 2021, 8pm.

In this lab you will implement the Viterbi algorithm with backtracking to determine the most likely set of states that would produce a given set of observations.

Here is the HMM you will use:

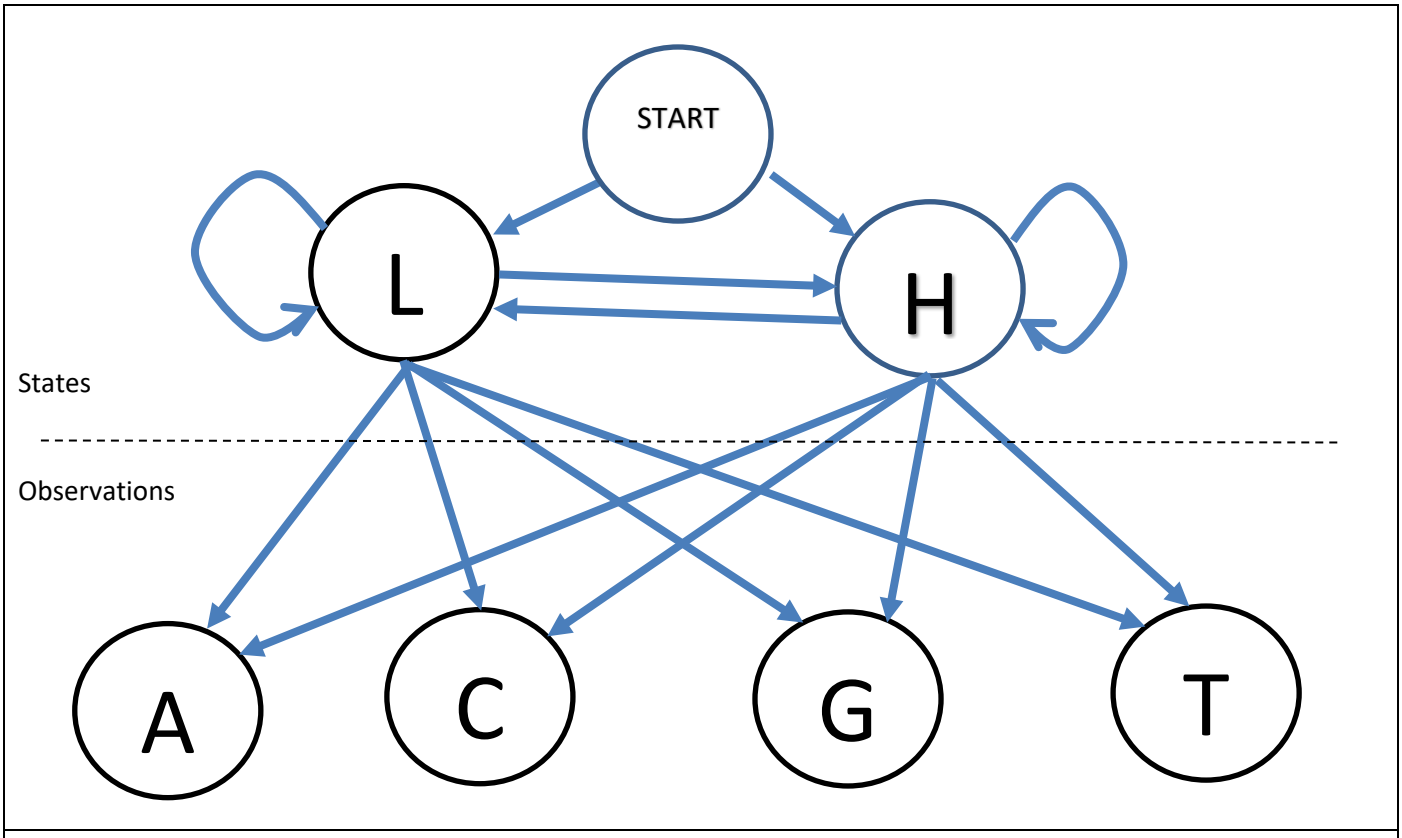

Figure 1

States: L, H Observations: A, C, T, G

Start Probability:

L	H
-1	-1

Transitions:

	To L	To H
From L	-0.737	-1.322
From H	-1	-1

Observation Probabilities:

	A	C	G	T
L	-1.737	-2.322	-2.322	-1.737
H	-2.322	-1.737	-1.737	-2.322

All probabilities given as $\log_2()$ values.

Viterbi Algorithm (5 Marks)

$$p_{\cdot}(i,x) = e_x(i) \max_k(p_k(j,x-1).p_{kl})$$

Equation 1

Given the sequence of observations **GGCACTGAA**.

Implement the dynamic Viterbi algorithm to compute the missing entries in the table below and then backtrack to identify a most probable path and therefore the most likely sequence of states that would produce the observations.

	G	G	C	A	C	T	G	A	A
H	-2.737								-25.658
L	-3.322								-24.488

Do not confuse the Viterbi algorithm with the forward-backward algorithm cover in the lectures. In this task, you will write two loops. The first will calculate the missing values making use of Equation 1. At each step you must also record which decision was taken in the max() calculation. Your second loop will then backtrack using this decision information not the values calculated to determine the most likely path.

Your code should output the values calculated, the decisions made and the most likely sequence (in the right order.)

Task 1 (3 marks)

Calculating the values and decisions.

Task 2 (2 marks)

Calculating the most probable path.

Lab Support

Tutors will be available during the LOIL sessions to help you to complete the labs. This will be done via Microsoft Teams. A series of rooms have been created to allow for small groups to cover specific issues and you can also ask for 1 to 1 help in order to share your screen without other students seeing your work.

Marking Guidance

The deadline for all four tasks of this lab sheet is **Friday 30th April 2021, 8pm**.

You must upload your Jupyter Notebook containing all the tasks attempted to Moodle for this unit by the deadline for this assignment or by any agreed extension deadline. Failing to do so will mean you do not receive the marks for the work. Marks will be given for each task successfully completed. This labsheet gives you an indicate of the last result you should be expecting for task 1. Tasks that are incomplete or produce the wrong answer will receive no marks. An allowance will be made for rounding errors in calculations. You must upload a version of your notebook that includes the output of running the code. We only expect to run your notebooks in exceptional cases.