COMPUTER SCIENCE TRIPOS Part IA – 2022 – Paper 3

8 Machine Learning and Real-world Data (av308)

A local government wants to use the COVID-19 test positivity rate (i.e. the proportion of tests with a positive result) to estimate the number of COVID-19 infections in the community. In the data collected, the positivity rate is labelled with one of three levels of positivity (`+`, `++` and `+++`) and the number of people infected are categorised as high (H), medium (M) or low (L).

timestep	1	2	3	4	5	6	7
infections	L	M	H	H	H	M	M
positivity	+	++	++	++	+++	++	++

They decided to use a first-order hidden Markov model (HMM), modelling the infections as the hidden states and the positivity as the observations.

(*a*) Define and estimate the components of an appropriate HMM for this application, without smoothing. [4 marks]

(*b*) What assumptions are implicit with the use of an HMM? Are they appropriate in the context of this application? [4 marks]

(*c*) You are in timestep 7 and you now observe the following positivity rates, but you do not know the number of infections:

timestep	8	9	10
positivity	+++	++	++

Predict the number of infections for each timestep using the Viterbi algorithm, showing the equations and calculations you make. [8 marks]

(*d*) Briefly describe two shortcomings of the HMM developed for predicting the number of infections. [4 marks]

[*Note:* This version fixes a typesetting mistake that had appeared in the exam.]