COMPUTER SCIENCE TRIPOS Part IA – 2025 – Paper 3

9 Machine Learning and Real-world Data (fm611)

Hidden Markov Models (HMMs) can be used for modelling the consonant-vowel sequences in words, using the letters of the words as the hidden states. An HMM of this type has two output options: $q_1 = V$ (for a vowel) and $q_2 = C$ (for a consonant). Each letter in the training data is labelled with either V or C. The training data is as follows (each column below is an example):

t a b l e	t r a c k	c a r	b e t	t a l l	l a b	b e l l
C V C C V	C C V C C	C V C	C V C	C V C C	C V C	C V C C

(a) Provide the full transition matrix A for this HMM based on the training data shown (ignoring the end states). [4 marks]

(b) Give the general formula for calculating emission probabilities from training data, and calculate the emission probabilities for a and ℓ. [2 marks]

(c) Suppose the system is in hidden state a, what is the most likely V, or C, for the next letter, and what are their probabilities. [2 marks]

(d) Suppose the HMM is in hidden state e, what is the most likely V, or C, for the letter 2 letters ahead, and what are their probabilities. [4 marks]

(e) What would be the probability of the following sequence of hidden states?

```
b e a t
```

[2 marks]

(f) How could you improve the model so that unseen combinations of letters do not make a previously unseen word totally impossible? Write the equations and new necessary matrices. What would the probability of

```
b e a t
```

become with this change? [6 marks]