9 Machine Learning and Real-world Data (SHT)

Hidden Markov Models (HMM) can be used to find names in text. In the following HMM created for this purpose, the two emitting states are q_1 = `i` (for "inside a name") and q_2 = `o` (for "outside a name"). Each word in the training data is labelled with either `i` or `o`. There are two sequences in the training data, as follows:

```
today may bakes a nice cake
 o      i   o   o  o    o

peter bakes and mary bakes may like sue
  i    i     o    i    i    o   o    i
```

(*a*) Give the general formula for estimating transition probabilities from training data. Provide the full transition matrix A for this HMM based on the training data shown. [6 marks]

(*b*) Give the general formula for calculating emission probabilities from training data, and calculate the emission probabilities $P(may|o)$, $P(may|i)$, $P(bakes|o)$, $P(bakes|i)$. [3 marks]

(*c*) An HMM trained with the above training observations is exposed to the following test observation:

```
may bakes
```

Which probabilities does the HMM assign to the following two interpretations?

(*i*) `may` is a name, and `bakes` is not [2 marks]

(*ii*) `bakes` is a name, and `may` is not [2 marks]

(*d*) The first training observation is now replaced with:

```
today peter bakes a nice cake
```

How does this change your answers to part (*c*)? [2 marks]

(*e*) A comparable situation to part (*d*) can arise even with substantial amounts of training data. Describe why this is a problem and indicate a solution to it. [5 marks]