

MS2a, Week 5

Rune Lyngsø

November 10, 2011

A Hidden Markov Model Use

- a. Consider a hidden Markov model emitting sequences over the alphabet $\{A, C, G, T\}$. The model has two states, 0 and 1, that are equiprobable start states. Transition probabilities are 0.75 for remaining in a state and 0.25 for switching to the other state. In state 0 A and G are emitted with probability 0.40 while C and T are emitted with probability 0.10. In state 1 A and G are emitted with probability 0.10 while C and T are emitted with probability 0.40. What is the probability of observing the sequence ACTG? Observe that we do not have an end state providing explicit termination, so the model will not model a sequence length distribution. Rather, for every sequence length it models a distribution over sequence content.
- b. What is the most likely sequence of hidden states, and how probable is it?
- c. What is the most likely hidden state at position 2, summing over all possible paths, and how probable is it?

B Hidden Markov Model Design

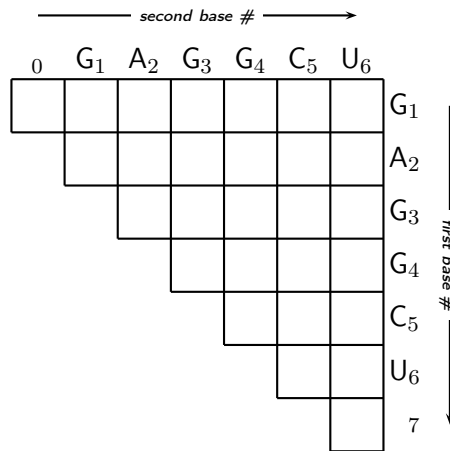
- d. The occasionally dishonest casino is a standard HMM example. Given the season and the problems ludomania causes in modern society, we will consider a rephrased version of this example. **Infinity Road** is an endless row of houses with a child living in most. Given the length of the street, Santa and his elves cannot constantly check the behaviour of all the children in the road, but checks just one deed per child each year to see whether it is Naughty or Nice. However, even a Nice child may transgress and perform a Naughty act (with probability 10% i.i.d. for all Nice children), and a Naughty child may inadvertently find himself doing something that can be classified as Nice (with probability 5%, again i.i.d. for all Naughty children). Santa would like to do better than basing his judgement on just

a single observation, and it just so happens that **Infinity Road** segments into Good neighbourhoods and Bad neighbourhoods, both of lengths that are geometrically distributed and with an expected length of a Good neighbourhood of 20 houses and an expected length of a Bad neighbourhood of 40 houses. All children living in a Good neighbourhood are Nice, and all children living in a Bad neighbourhood are Naughty (it's all about peer pressure). It is equally likely that **Infinity Road** starts with a Good neighbourhood as with a Bad neighbourhood. Houses with no children living in them are distributed uniformly at random, with on average one out of every 500 houses not having a child living in it. Unfortunately the records of which houses are childless has been lost, all Santa has to go by is the sequence of Nice and Naughty for the deed checked for each child as you go down **Infinity Road**. Neighbourhoods either side of one or more childless houses are uncorrelated, such that the neighbourhood starting after a childless house has equal chance of being Good and Bad. Design a hidden Markov model that can help Santa use the observations for all the children to annotate each child as either Naughty or Nice – don't worry that it would normally take infinitely long time to annotate an infinitely long sequence.

- e. Construct a HMM that generates the sequence A^i , i.e. the sequence of i As, with probability 2^{-i} for $i \geq 1$, if possible. Otherwise argue it is not possible.
- f. Construct a HMM that generates sequences over the alphabet $\{A, G\}$ with probability $(k-1)2^{-k}$ for generating a sequence of length $k \geq 2$, and for which all sequences that are generated of length k are on the form $A^i G^{k-i}$ with $1 \leq i < k$ and equiprobable.
- g. Construct a HMM that generates the sequence $A^i G^i$ with probability 2^{-i} for $i \geq 1$, if possible. Otherwise argue that it is not possible.

C RNA Secondary Structure Prediction

- h. Use Algorithm 1 and Algorithm 2 of the lecture notes on RNA secondary structure prediction to find the maximum number of base pairs for the sequence GAGGCU, and a structure with this number of base pairs. Two bases can form a valid base pair if *i*) they are separated by at least three bases in the sequence, i.e. their indices differ by at least 4, and *ii*) they form one of the three types of base pairs shown in Figure 2 in the lecture notes. For added convenience, the table you need to fill out and backtrack (cf. Figure 5 in the lecture notes) is:



- i. Forgetting about Algorithms 1 and 2, can you find a structure with more valid base pairs than the one you found above? If so, why does Algorithm 1 fail to find this number of base pairs?
- j. How many different structures with no crossing base pairs can you find for this sequence?
- k. Describe a recursion for determining the number of different possible structures possible for an RNA sequence s . Equation (1) in the RNA notes is a perfect source for inspiration, but remember that we need to count all structures rather than finding the score of the best one. Would it be as easy to modify equations (2) and (3) to obtain a recursion for the number of structures when all base pairs are required to be stacking, i.e. have a neighbouring base pair (why/why not)?