Time-Series: Computer Lab

David Barber
Department of Computer Science, University College London

1 Introduction

To get the MATLAB code for this tutorial, please go to

`http://web4.cs.ucl.ac.uk/staff/D.Barber/pmwiki/pmwiki.php?n=Main.Textbook`

and download the latest version.

You'll then need to unzip the file in a local directory. Then open matlab and change to the directory where you unzipped the file. Then, within matlab, type `setup`. This will set the paths required to run the demos.

2 Rock Paper Scissors

2.1 Markov Model

1. run the demo `demoRockPaperScissorsMarkovHuman.m` by typing `demoRockPaperScissorsMarkovHuman(false)`. This runs the demo without showing you what it thinks you will play the next timestep.

   The computer predict what you will do based on using a first order Markov Chain of your previous play. You will play 20 games against the computer.

2. Run `demoRockPaperScissorsMarkovHuman(true)` to see what the computer thinks you will play next.

3. Modify the code to make a new routine `demoRockPaperScissorsMarkovHumanComputer.m` that predicts your move based on both your previous move and the computer's previous move.

2.2 Hidden Markov Model

1. run the demo `demoHMMRockPaperScissors(false)` which predicts what you will play using a HMM based on the 4 strategies described in the lecture.

2. run the demo `demoHMMRockPaperScissors(true)` which shows the filtered distribution and next timestep prediction distribution. Do you think this makes sense?

3. Make a new demo which includes new strategies, namely 'forward cyclic' playing 'rock, paper, scissors, rock, paper, scissors...' and 'reverse cyclic', 'scissors, paper, rock, scissors,

paper, rock...', 'doing something different to what you did last time' and 'doing something different to what the computer did last time'. Now play this new game and try to beat the computer!

4. You may note from the code that we assume that we stay in the same strategy with a certain fixed probability. Try to adjust this probability and see if this makes much difference.

3 Clustering Genetic Sequences

3.1 Mixture of Markov Models

1. run the demo

   `demoMixMarkov.m`

   which corresponds to the Gene Clustering example in the notes on Mixture of Markov models.

   Try to understand whether or not there are local minima in the clustering problem by running the routine more than once.

   If you get different results after doing several runs, which one should you prefer?

   Examine the sequences to see if you can understand why the sequences are clustered the way they are.

2. Use the correspondence $A = 1, C = 2, G = 3, T = 4$:

   Define a 4×4 transition matrix p that produces sequences of the form

   $$A, C, G, T, A, C, G, T, A, C, G, T, A, C, G, T, \ldots$$

   Now define a new transition matrix

   ```
   pnew = 0.9*p + 0.1*ones(4)/4
   ```
   (1)

   Define a 4×4 transition matrix q that produces sequences of the form

   $$T, G, C, A, T, G, C, A, T, G, C, A, T, G, C, A, \ldots$$

   Now define a new transition matrix

   ```
   qnew = 0.9*q + 0.1*ones(4)/4
   ```
   (2)

   Assuming that probability of being in the initial state of the Markov Chain $p(h_1)$ is constant for all four states A, C, G, T.

   What is the probability that the Markov Chain `pnew` generated the sequence S given by

   $$S \equiv A, A, G, T, A, C, T, T, A, C, C, T, A, C, G, C \tag{3}$$

   Similarly what is the probability that S was generated by `qnew`?

3. Using the function `randgen.m` generate 100 sequences from the Markov Chain defined by `pnew`. Similarly, generate 100 sequences each of length 16 from the Markov Chain defined by `qnew`.

   Concatenate all these sequences into a cell array `v` so that `v{1}` contains the first sequence and `v{200}` the last sequence.

   Use `MixMarkov.m` to find learn the optimum Maximum Likelihood parameters that generated these sequences. Assume that there are $H = 2$ kinds of Markov Chain. The result returned in `phgvn` indicates the posterior probability that sequence `n` belongs to the two models. Do you agree with the solution found?

3.2 Hidden Markov Models

Examine `demoHMMinferenceSimple.m`.

Take the sequence S as defined in Equation (3). Define an emission distribution that has 4 output states such that

$$p(v = i|h = j) = \begin{cases} 0.7 & i = j \\ 0.1 & i \neq j \end{cases}$$

Using now `pnew` defined in Equation (1) adapt `demoHMMinferenceSimple.m` suitably to find the most likely hidden sequence that generated the observed sequence S.

Repeat the above computation for `qnew`. Which hidden sequence is to be preferred? Justify your answer.

Can you understand how this method could be used to 'clean up' corrupted gene sequences?

4 Noisy Pattern Search

Download `http://web4.cs.ucl.ac.uk/staff/D.Barber/code.zip` and unzip.
Modify `demoFirstnameSurname.m` to search for patterns of the form:

`firstname*middlename*surname` or `firstname*surname`.
and demonstrate your method.