Supervision Assignments in MLRW - Set 2

Dionysis Manousakas

February 25, 2017

1 Dependence of p-values on irrelevant information -or why we should become Bayesians :-) (from McKay's book)

1.1 In an expensive laboratory, Dr. Bloggs tosses a coin labelled a and b twelve times and the outcome is the string

$$aaabaaaabaab,$$

which contains three bs and nine as. What evidence do these data give that the coin is biased in favour of a? Is it significant at the level of 5% ?

1.2 Dr. Bloggs pays careful attention to the calculation of 1.1, and responds *'no, no, the random variable in the experiment was not the number of bs: I decided before running the experiment that I would keep tossing the coin until I saw three bs; the random variable is thus the total number of tosses, n'*. A different calculation is required in order to assess the 'significance' of the result $n = 12$. Now, the probability distribution of n given $\mathcal{H}_0$ is the probability that the first $n-1$ tosses contain exactly $r-1$ bs and then the nth toss is a b. What evidence do these data give that the coin is biased in favour of a in this case? Is the evidence significant at the level of 5% ?

2 HMM (from CMU 10-701, Fall 2015 Homework)

2.1 Let's use the notation $I < X, Y, Z >$: "X is conditionally independent of Z given Y". Assuming the notation of HMM in which x_t is the hidden state at time t and o_t is the observation at time t, which of the following are true of all HMMs?

1. $I < x_{t+1}, x_t, x_{t-1} >$
2. $I < x_{t+2}, x_t, x_{t-1} >$
3. $I < x_{t+1}, x_t, x_{t-2} >$
4. $I < o_{t+1}, o_t, o_{t-1} >$
5. $I < o_{t+2}, o_t, o_{t-1} >$
6. $I < o_{t+1}, o_t, o_{t-2} >$

	0	S_1	S_2	S_3	S_4	S_5	S_6	f	A	C	G	T
0	0	1	0	0	0	0	0	0				
S_1	0	0	1	0	0	0	0	0	0.2	0.3	0	0.5
S_2	0	0	0	0.3	0	0.7	0	0	0.6	0.1	0.2	0.1
S_3	0	0	0	0	1	0	0	0	0.7	0	0.1	0.2
S_4	0	0	0	0	0	0	0	1	0.2	0.3	0.4	0.1
S_5	0	0	0	0	0	0	1	0	0.3	0.3	0.3	0.1
S_6	0	0	0	0	0	0	0	1	0.5	0.3	0	0.2

Table 1: The transition and emission probabilities.

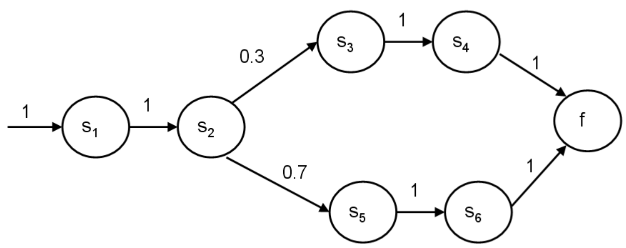

Figure 1: The state diagram of the HMM.

2.2 Consider a HMM with 6 states (plus a start and end states) and an alphabet $\{A, C, G, T\}$. Table 1 lists the transition and emission probabilities, and Figure 1 shows the state diagram.

Let z denote latent variables and x denote observed variables. Place $<$, $>$, or $=$ between the two components of each of the following pairs. Justify your answer.

1. $P(x_1 = T, x_2 = C, x_3 = A, x_4 = T, z_3 = S_3, z_4 = S_4)$
   $P(x_1 = T, x_2 = C, x_3 = A, x_4 = T | z_3 = S_3, z_4 = S_4)$
2. $P(x_1 = T, x_2 = C, x_3 = A, x_4 = T)$
   $P(x_1 = T, x_2 = C, x_3 = A, x_4 = T | z_3 = S_3, z_4 = S_4)$
3. $P(x_1 = T, x_2 = C, x_3 = A, x_4 = T, z_1 = S_1, z_2 = S_2)$
   $P(x_1 = T, x_2 = C, x_3 = A, x_4 = T | z_1 = S_1, z_2 = S_2)$
4. $P(x_1 = T, x_2 = C, x_3 = A, x_4 = T)$
   $P(x_1 = T, x_2 = A, x_3 = A, x_4 = G)$

2.3 Prove that $p(x_1, \ldots, x_i, z_i) = p(x_i | z_i) \sum_{z_{i-1}} p(x_1, \ldots, x_{i-1}, z_{i-1}) p(z_i | z_{i-1})$.

3 Weary dealer (from Pevzner's book)

To avoid suspicion, the dealer in the Fair Bet Casino keeps every coin for at least ten tosses, regardless of whether it is fair or biased. Describe the corresponding HMM.

4 HMM Complexity (from CMU 10-701 2003 Final Exams)

(True/False) In general when are trying to learn an HMM with a small number of states from a large number of observations, we can almost always increase the training data likelihood by permitting more hidden states.