Project 5: HMMs

Objectives

The objective is to solve the three problems associated with an HMM with a discrete observation probability distribution for a sequence of observations O that you will generate yourself!

As usual, you can use any statistical package, such as MatLab, R, SAS, or use Python with all the available statistical functions, or functions from different packages.

Task 1: Data Set Generation

Generate a set of observations by simulating an HMM consisting of 4 states and 3 objects, i.e., $S = \{1,2,3,4\}$, $v = \{1,2,3\}$. For this, set the initial vector $\pi = \{1,0,0,0\}$, and generate your personalized one-step transition probability matrix P and event matrix B:

$$P = \begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{23} & p_{33} & p_{34} \\ p_{41} & p_{24} & p_{43} & p_{44} \end{pmatrix}, B = \begin{pmatrix} b_1(1) & b_1(2) & b_1(3) \\ b_2(1) & b_2(2) & b_2(3) \\ b_3(1) & b_3(2) & b_3(3) \\ b_4(1) & b_4(2) & b_4(3) \end{pmatrix},$$

As described below.

We first describe how to generate the one-step transition probabilities p_{ij} and the probabilities $b_i(k)$, and then we describe how to simulate the resulting HMM in order to generate a sequence of observations O that you will use in the subsequent tasks.

Use your student ID number as the seed to the pseudo-random number generator that you will use in the assignment.

Generation of the p_{ij} probabilities

- Draw a pseudo-random number r, and set $p_{11} = r$. Repeat the same for the remaining probabilities on the same row, i.e., $p_{12}, p_{13},$ and p_{14}. Normalize the resulting probabilities, by dividing each by the sum $p_{11} + p_{12} + p_{13} + p_{14}$, so that they all add up to 1. That is, set $p_{11} = p_{11}/(p_{11} + p_{12} + p_{13} + p_{14})$, $p_{12} = p_{12}/(p_{11} + p_{12} + p_{13} + p_{14}), p_{13} = p_{13}/(p_{11} + p_{12} + p_{13} + p_{14})$, and $p_{14} = p_{14}/(p_{11} + p_{12} + p_{13} + p_{14})$.

- Repeat the same for the remaining rows of P.

Generation of the $b_i(k)$ probabilities

- Draw a pseudo-random number r, and set $b_1(1) = r$. Repeat the same for the remaining probabilities on the same row $b_1(2), b_1(3)$. Normalize the resulting probabilities, by dividing each by the sum $b_1(1) + b_1(2) + b_1(3)$. That is, set $b_1(1) = b_1(1)/(b_1(1) + b_1(2) + b_1(3)), b_1(2) = b_1(2)/(b_1(1) + b_1(2) + b_1(3)), b_1(3) = b_1(3)/(b_1(1) + b_1(2) + b_1(3))$.

- Repeat the same for the remaining rows of B.

Generation of the sequence of observations O.

Let us assume that you generated the following values for P and B:

$$P = \begin{pmatrix} 0.4 & 0.4 & 0.1 & 0.1 \\ 0.2 & 0.3 & 0.3 & 0.2 \\ 0.1 & 0.1 & 0.4 & 0.4 \\ 0.4 & 0.3 & 0.2 & 0.1 \end{pmatrix}, \text{ and } B = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.3 & 0.4 & 0.3 \\ 0.1 & 0.2 & 0.7 \\ 0.2 & 0.5 & 0.3 \end{pmatrix}.$$

Use the following steps to generate O.

$t = 1$:

- The HMM is in state $q_1 = 1$, since we assumed that $\pi = \{1,0,0,0\}$.
- Draw a pseudo-random number r and use the probabilities $b_1(k), k = 1,2,3$ to determine observation O_1. That is:
  - if $r \leq 0.8$ then $O_1 = 1$
  - If $0.8 < r \leq 0.9$ then $O_1 = 2$.
  - Otherwise, $O_1 = 3$.

$t = 2$:

- The HMM shifts to state $j, j = 1,2,3,4$. For this, draw a pseudo-random number r and use the branching probabilities corresponding to state 1 to determine the next state q_2. That is:
  - If $r \leq 0.4$ then $q_2 = 1$.
  - If $0.4 < r \leq 0.8$ then $q_2 = 2$.
  - If $0.8 < r \leq 0.9$ then $q_2 = 3$.
  - Otherwise, $q_2 = 4$.
- Having determined the state, use matrix B to determine O_2 as described above for $t = 1$.

$t = 3$:

- Assuming that you determined that the HMM shifts to state j at time $t = 2$, draw a pseudo-random number r and use the branching probabilities corresponding to state j to determine the next state q_3 as described above for $t = 2$.
- Having determined the next state, use matrix B to determine O_3 as described above for $t = 1$.

$t \geq 4$

- Repeat above steps for $t = 3$ until you have generated 1000 observations, that is, $t = 1000$. Remember to also record the generated sequence of states.

Task 2: Estimate $p(O|\lambda)$

- Given that you know the parameters of the HMM, calculate the probability $p(O|\lambda)$ that the sequence of observations $O = 123312331233$ came from the HMM.

- Discuss your results.

Task 3: Estimate the Most Probable Sequence Q

- Given that you know the parameters of the HMM, calculate the most probable sequence of

states Q that gave rise to the sequence of observations $O = 12331233123$

- Discuss your results.

Task 4: Train the HMM

- Now forget that you know the parameters of the HMM, and use an HMM solver to estimate its parameters $\lambda = (P, \pi, B)$ using the 1000 generated observations, assuming that the HMM has four states, i.e., $S = \{1,2,3,4\}$, and that the number of objects is 3, i.e., $v = \{1,2,3\}$.

- Give the estimated parameters in one or more tables along with their corresponding p-values. Compare your results against the actual parameters $\lambda = (P, \pi, B)$ of the HMM used to generate the data.

- Discuss your results.

What to submit

You will submit your report along with the code that you wrote to obtain the results. It is very important that you provide enough results to support your conclusions. Conclusions without insufficient results will make you lose grades. Remember: no code sharing!

How to submit your work

Submit your report and code in a single .zip file (not .tar or .7z, or any other format). Your report should should be a single pdf file that contains all your graphs, tables and conclusions.

Grading

The TA will first verify that your code works and produces the results you submitted. The break down of the grades will be as follows:

Task 1: 30 points
Task 2: 20 points
Task 3: 20 points
Task4: 30 points

Extra credit

Assume now that you do not know the number of states of the HMM, but you know that the number of objects is 3, i.e., $v = \{1,2,3\}$. Train different HMMs each with a different number of states, starting with an HMM with two states, and for each HMM calculate the likelihood, AIC, and BIC. Plot these three quantities as a function of the number of states. Keep increasing the number of states until you begin to discern a pattern in each of the three plots. Select the best HMM. Discuss your results.

Note that the number of parameters increases as the number of states increases. A rule of thumb is that for each parameter, you need at least 10 observations. As you increase the number of states,

you may require more than 1000 observations. In this case, simply generate additional observations as described above.

How to submit your extra credit

Submit your report and any additional code to the one you used above in tasks 1 to 4, in a single .zip file (not .tar or .7z, or any other format). Your report should should be a single pdf file.

Submit your .zip file separately from project 5, where it says: "Submit your extra credit work".

Grading of extra credit

This task will be graded separately from the above assignment from 0 to 100. The grade will then be transformed to the range [0,3], and it will be added to your final grade for the class. For instance, if you get a 3 for the extra credit and you class grade is 83, then your final grade will be 86. This extra credit typically will push your final grade up one +/- bracket, but not always. See also syllabus for the +/- grading scale.

Remember that you will be graded mostly on your ability to interpret the results