Ciências ULisboa | Informática

Aprendizagem Automática Avançada
(2º Ciclo de Informática)

Hidden Markov Models Exercises

For the following problems, we will be using the implementations of Hidden Markov Models provided in the hmmlearn package (`https://hmmlearn.readthedocs.io/en/latest/`).

Problem 1

Run the experiments in hmmlearn's tutorial (`https://hmmlearn.readthedocs.io/en/latest/tutorial.html`) and note the number of states, transition matrix / parameters, most likely state sequence. Visualize the samples.

Problem 2 *

Recall the dishonest casino problem presented during the lecture (slide 23):

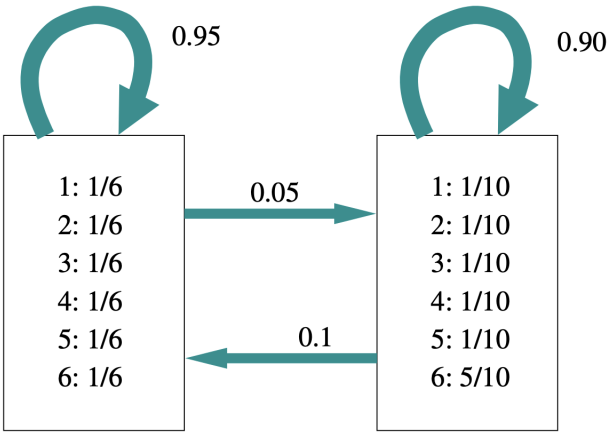

Figure 1: Dishonest Casino Hidden Model. The state on the left represents a fair dice and on the right a loaded dice. Each state contains the probabilities for the dice to land on each number face.

Consider the following sample:

- Rolls: `664153216162115234653214356634261655234232315142464156663246`.
- Dice: `LLLLLLLLLLLLLLFFFFFFLLLLLLLLLLLLLLFFFFFFFFFFFFFFFFFFLLLLLLL`.

For the next two question (a and b), use the score function (`https://hmmlearn.readthedocs.io/en/latest/api.html#hmmlearn.hmm.CategoricalHMM.score`) to assess how well the HMM

models the problem, in terms of log-likelihood over the rolls, using the last 10 data points of the sample.

a) construct the model assuming all parameters known.

b) train the model assuming the transition probability matrix is unknown. Compare the estimated transition matrix with the original one. Visualize the generated dice series against the given one.

c) Train HMM's with half of the data points and observe the log-likelihood with the remaining half. Compare the estimated parameters (transition and emission matrices) with the original model from question a). Visualize the generated dice series of the last half of data against the given one. (Note: Use the dataset provided in files `rolls.txt` and `dice.txt` provided on Moodle):

   (i) Assume both transition and emission matrices are unknown.

   (ii) Assume all the parameters unknown, including the number of states (suggestion: try with the number of states from 1 to 4 and compare the resulting log-likehoods).

d) Generate 30000 samples from the model of question a). Repeat question c) using this new dataset. (Note: you don't need to visualize the dice states.)