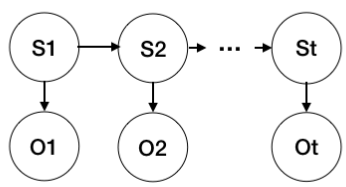
Hidden Markov Model (HMM) Project

Week 8

1 HMM Exercise

Infection is a common condition among patients in ICU settings and can have various roots, which makes it challenging to be determined. Assume that we want to model infection using an HMM, while *infection* is the hidden state and the only available observation is *blood pressure* (0 for normal and 1 for abnormal). When patients entering ICU, the probability of being infected is 0.75. At any given time, infected patients have 40% chance of improving to be uninfected and uninfected patients have 20% chance of becoming infected. There is 80% chance of observing an abnormal blood pressure for infected patients while only 10% chance of observing abnormal observation for uninfected patients.



(a) (3 points) Create initial, transition, and emission probability tables based on the problem statement given above.

Solution:

S_i	$S_{i+!}$	$P(S_{i+1}\|S_i)$
T	F	0.4
T	T	0.6
F	T	0.2
F	F	0.8

S_i	O_i	$P(O_i\|S_i)$
T	0	0.2
T	1	0.8
F	0	0.9
F	1	0.1

S_1	$P(S_1)$
T	0.75
F	0.25

(b) (6 points) Using the described HMM and the generated probability tables, apply the forward algorithm to compute the probability that we observe the sequence $\{0, 1, 1\}$ blood pressure. Show your work (i.e., show each of your αs).

Solution: The values of different alphas and the probability of the sequences are as follows:

$$\alpha_1^T = 0.75 \times 0.2 = 0.15$$
$$\alpha_1^F = 0.25 \times 0.9 = 0.225$$
$$\alpha_2^T = 0.8 \times (0.15 \times 0.6 + 0.225 \times 0.2) = 0.108$$
$$\alpha_2^F = 0.1 \times (0.15 \times 0.4 + 0.225 \times 0.8) = 0.024$$
$$\alpha_3^T = 0.8 \times (0.108 \times 0.6 + 0.024 \times 0.2) = 0.05568$$
$$\alpha_3^F = 0.1 \times (0.108 \times 0.4 + 0.024 \times 0.8) = 0.00624$$
$$P(\{O_t\}_{t=1}^T) = 0.06192$$

(c) (6 points) Using the backward algorithm, compute the probability that we observe the aforementioned sequence ($\{0, 1, 1\}$). Again, show your work (i.e., show each of your βs).

Solution: The values of different alphas and the probability of the sequences are as follows:

$$\beta_3^T = 1$$
$$\beta_3^F = 1$$
$$\beta_2^T = 0.6 \times 0.8 \times 1 + 0.4 \times 0.1 \times 1 = 0.52$$
$$\beta_2^F = 0.2 \times 0.8 \times 1 + 0.8 \times 0.1 \times 1 = 0.24$$
$$\beta_1^T = 0.6 \times 0.8 \times 0.52 + 0.4 \times 0.1 \times 0.24 = 0.2592$$
$$\beta_1^F = 0.2 \times 0.8 \times 0.52 + 0.8 \times 0.1 \times 0.24 = 0.1024$$
$$P(\{O_t\}_{t=1}^T) = 0.15 \times 0.2592 + 0.225 \times 0.1024 = 0.06192$$

(d) (4 points) Using the forward-backward algorithm, compute the most likely setting for each state. Show your work.

Solution: We already have the alphas and betas from the previous two computations. Note that the most likely state at time t is state T if $\alpha_t^T \beta_1^T > \alpha_1^F \beta_1^F$ and state F if the inequality is reversed. We predict that $\{T, T, T\}$ is the most likely setting of states. The relevant values for the computation are:

$$\alpha_1^T \beta_1^T = 0.03888$$
$$\alpha_1^F \beta_1^F = 0.02304$$
$$\alpha_2^T \beta_2^T = 0.05616$$
$$\alpha_2^F \beta_2^F = 0.00576$$
$$\alpha_3^T \beta_3^T = 0.055$$
$$\alpha_3^F \beta_3^F = 0.00624$$

(e) (6 points) Use the Viterbi algorithm to compute the most likely sequence of states. Show your work.
Solution: The Viterbi algorithm predicts that the most likely sequence of states is $\{T, T, T\}$, based on the following computations:

$$V_1^T = 0.75 \times 0.2 = 0.15$$
$$V_1^F = 0.25 \times 0.9 = 0.225$$
$$V_2^T = 0.8 \times \max\{0.225 \times 0.2, 0.15 \times 0.6\} = 0.072$$
$$V_2^F = 0.1 \times \max\{0.225 \times 0.8, 0.15 \times 0.4\} = 0.018$$
$$V_3^T = 0.8 \times \max\{0.072 \times 0.6, 0.018 \times 0.2\} = 0.03456$$
$$V_3^F = 0.1 \times \max\{0.072 \times 0.4, 0.018 \times 0.8\} = 0.00288$$

2 HMM Programming

Similar to the workshop programming exercise, in this project you will use the developed 2-state Hidden Markov Model (HMM) for student learning prediction. Specifically, using the sequence of student assignment data, we want to predict two outcomes: the student's quantile learning gain (QLG) (binary) and the student's post-test score (numerical).

Data The dataset used for this project is the same student assignment data from workshop exercise. Divide the students' records into train and test portions using 5-fold cross validation. Further, you are given a student outcome data in CSV format that includes pre-test, post-test, and learning gain for each student defined as improved test result of the post-test score as compared to the pre-test score. The quantile learning gain target data is binary as being learned (1) or unlearned (0). Post-test score and learning gain are the target data that your HMM algorithm should predict.

Model Train a 2-state HMM model for each KC on training data, using the initial parameters specified in the workshop exercise. Then, use the prediction of each HMM on the test set as features to learn a Logistic Regression model to predict the student learning gain and a Linear Regression to predict the student post-test score.

Finally, compare the performance of HMM model for both tasks against a deep recurrent model, Long Short Term Memory (LSTM).

Report In your report, include the following:

- Report the performance of your developed 2-state HMM model on the training and test sets for predicting the student learning gain and post-test score.
- Report the performance of the LSTM model on the same training and test sets for predicting the student learning gain and post-test score. Mention and discuss your choice of the hyper-parameters.

- compare the performance of the above two models. Mention the pros and cons of utilizing each model for analysis.

Solution:

The result for learning gain prediction is:

Model	Accuracy	Recall	F-measure	AUC
HMM	0.5820	0.6756	0.6410	0.5711
LSTM	0.6567	0.8513	0.7325	0.6340

The results for post test score prediction is:

Model	Training MSE	Test MSE
HMM	1.2853	1.4280
LSTM	7.2014	7.2014

The code to generate solutions are included in `https://github.ncsu.edu/mchi/AIA-MachineLearning/tree/master/HMM-BKT`