EECS545 - Homework 6

March 27th, 2018

Instructions

- This homework is Due Wednesday, April the 11th at 5pm. No late submissions will be accepted.
- As always, submit a write-up and your code for this homework.
- You are expected to use Python for the programming questions in this homework. Use of Python 3 is recommended.
- **Submit your write-up to Gradescope under the assignment titled** `Homework 6`. Make sure to tag the pages with the corresponding problems numbers.
- Submit all your python code to Canvas in a compressed zip file named `uniqname_hw6_code.zip`. Your zip file should contain `prob4.py`.

1. Suppose we have binary states (labeled A and B) and binary observations (labeled 0 and 1) and the initial, transition, and emission probabilities as in the given table. Please answer

State	$P(S_1)$
A	0.80
B	0.20

(a) Initial Probs.

S_1	S_2	$P(S_2\|S_1)$
A	A	0.80
A	B	0.20
B	A	0.30
B	B	0.70

(b) Transition Probs.

S	O	$P(O\|S)$
A	0	0.80
A	1	0.20
B	0	0.10
B	1	0.90

(c) Emission Probs.

   the questions below. You need do the calculation by hand instead of using any programs or packages.

   (a) Using the forward algorithm, compute and report the probability that we observe the sequence $O_1O_2O_3 = 010$ and $O_1O_2O_3 = 101$. Briefly justify/explain your answer.

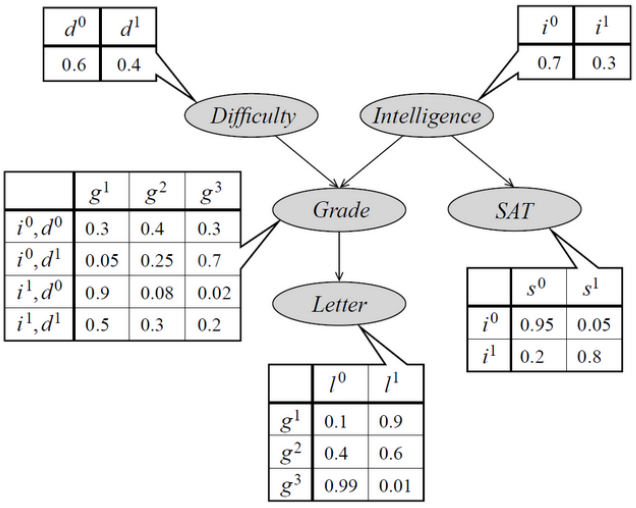
   (b) Using the backward algorithm, compute and report the probability that we observe the sequence $O_1O_2O_3 = 010$ and $O_1O_2O_3 = 101$. Briefly justify/explain your answer.

   (c) Use the Viterbi algorithm to compute and report the most likely sequence of states for $O_1O_2O_3 = 010$. Show your derivation.

2. In this problem, you need to implement in code the forward-backward and Viterbi algorithms, and learn a HMM model with K hidden states from several simple DNA sequences. A DNA sequence is regarded as a series of components from A, C, G, T. Assume the initial state is uniformly-randomly picked from all hidden states. We got two DNA sequences $X_1 = CCTACACGCA$ and $X_2 = CTACGCAAT$, calculate and report:

- the transition probability and emission probability using the forward-backward algorithm for $K = 2$ and $K = 4$. We provide the initial transition matrices T and emission matrices E in starter code, where $T(i, j)$ stands for the transition probability from S_i to S_j, and $E(i, j)$ stands for the emission probability for S_i to emit O_j.($O_1 = A, O_2 = C, O_3 = G, O_4 = T$) Please use 30 iterations in this algorithm.
- the most likely sequence of states of length 4 when the first observation is A, for the two models you have learned.

3. Consider the following Bayesian student network. The network models the relationship between the following random variables. The grade received by a student (G) is influenced by the difficulty of the course (D) and the intelligence of the student (I). Our student asks the professor for a recommendation letter. Due to the large class size the professor does not remember the names of students, and hence writes the letter based on the student's course grade. L represents the quality of the recommendation letter. The network also shows how a student's SAT score is influenced by his/her intelligence.



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

(a) Write an expression for the joint distribution $P(D, I, G, L, S)$ exploiting the conditional independence relationships in the network.

(b) What is the probability that a student gets a strong letter ($L = l^1$) ?

(c) Given that the student is intelligent ($I = i^1$), how likely is it that he/she will get a strong letter ?

(d) If the student did well on his/her SAT ($S = s^1$), what is the probability that they will do well in the course ($G = g^1$) ?

(e) If the student was intelligent, did well on his/her SAT and got a great letter, how likely is it that the course was difficult ($D = d^1$) ?

4. In this problem we will perform **principal component analysis (PCA)** to reduce the dimensionality of data in the Boston Housing dataset from 13 to 2, and visualize the dataset in this lower-dimensional space.

   Normalize the features (such that each feature has mean 0 and standard deviation 1 over the data), perform PCA, extract the first two principal components for each datapoint, and include the following in your report:

   - Report the values of the first two principal directions of the data. This should be two vectors of dimension 13.
   - Create a scatter plot of the first two principal components of each datapoint in the dataset. Color each point based on its target value (see hint #2 for how to do this).

   Please submit your code as **prob4.py**.

   Hint #1: You can include the boston housing dataset in your code with

   ```
   from sklearn.datasets import load_boston
   data, target = load_boston().data, load_boston().target
   ```

   Hint #2: You can create a scatterplot that assigns each point to a color based on it's target value with

   ```
   plt.scatter(first_principal_components, second_principal_components,
               c=target / max(target))
   ```