CSCI-580 Fall 2014 Project 2 Due: October 29-31

In this project you will practice implementation of learning *hidden Markov model* algorithm (EM for HMM).

Given: A sequence of observations at time 1, 2, …, *n*; and the number of states in an HMM.

Goal:

1. Implement EM for HMM and the Viterbi's algorithms;
2. Using EM for HMM and the Viterbi's algorithm, find:
   a) A most probable sequence of states that generated the given sequence of observations;
   b) A transition and sensory models of the HMM.
3. Given three different sets of starting transition and sensory probabilities, and given an original sequence of states ***X*** that generated the given sequence of observations, find ***accuracy*** of your classifier for each set of probabilities using the following formula:

$$\frac{The\ number\ of\ states\ identified\ by\ the\ classifier\ that\ match\ the\ states\ in\ X\ at\ time\ i}{Total\ number\ of\ states\ (=n)}$$

   where *i* varies from 1 to *n*, the length of the sequence of observations.

Consider an example, in which a casino flips a coin *n* times. A coin can be balanced (state *B*), loaded with higher probabilities for Heads (state *L*), or loaded with higher probabilities for Tails (state *M*). Assume that the casino selects at random a coin out of B, L, and M coins to start. This random selection has equal probability, i.e. $P(X_0 = x) = 1/3$ for $x = B, L, M$.

Input for your program is four files:

- the first file *observations.txt* has a sequence of observations HTTHTTTHHHHTTHTTTTT of length *n*;
- the second file *transition.txt* has the starting transition probabilities, $P(X_i|X_{i-1})$;
- the third file *sensory.txt* has the starting sensory probabilities, $P(e_i|X_i)$;
- the fourth file *original.txt* has an original sequence of states that generated the given observations

Format of files:

observations.txt

HTTHTTTHHHHTTHTTTTT

original.txt

BBMMMMMLLLLLBBBMMMMM

transition.txt

$P(B_i\|B_{i-1})\ P(B_i\|L_{i-1})\ P(B_i\|M_{i-1})$
$P(L_i\|B_{i-1})\ P(L_i\|L_{i-1})\ P(L_i\|M_{i-1})$
$P(M_i\|B_{i-1})\ P(M_i\|L_{i-1})\ P(M_i\|M_{i-1})$

example

0.45 0.52 0.25
0.35 0.3 0.13
0.2 0.18 0.62

sensory.txt

$P(H_i\|B_i)$
$P(H_i\|L_i)$
$P(H_i\|M_i)$

example

0.5
0.85
0.1

Output of your program consists of two parts: to the screen and to file output.txt. Format to the screen:

```
Transition probabilities learned:
  0.17   0.5   0.29
  0.66   0.17  0.14
  0.17   0.33  0.57

Sensory probabilities learned:
  0.60
  0.80
  0.17

Accuracy:
  70%
```

Note that this is only an example of the output format, not the correct values for the given example. Please use <iomanip> to format the output, and show at least 6 digits after the decimal points.

Output file ***output.txt*** must contain the sequence of states found by the Viterbi's algorithm in the last iteration of EM.

Command line to run your executable program em:

./*em* observations.txt transition.txt sensory.txt original.txt *k*

where *k* is an integer, the number of iterations to run the EM.

Submission includes:

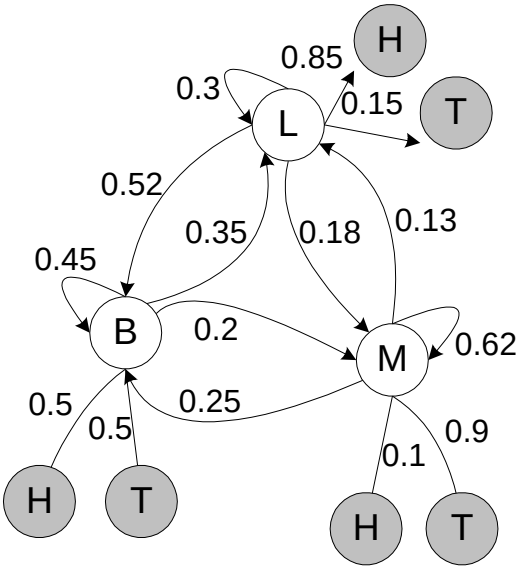
1. A personal demonstration of your running program to the instructor (10 minutes).
2. Zipped code submitted via Blackboard.

Grading rubric:
1. Submission of code is 10%.
2. Demo is up to 90%.

Demo grading rubric:

1. Using log space for calculations: 10%
2. Overall flow of the EM algorithm: 20%
3. Calculation of transition and sensory probabilities from the Viterbi's state sequence: 20%
4. The Viterbi's algorithm: filling the table with values of the maximum probability $\mathrm{P}(\mathrm{e}_{1:\mathrm{i}}, \pi_0, \pi_1, \pi_2, \ldots, \pi_\mathrm{i})$ for $1 \le i \le n$: 10%
5. The Viterbi's algorithm: using backtracking, find the most likely state sequence that generated the given sequence of observations: 20%
6. Output format: 5%
7. Correctness of the state sequence found: 5%
8. Correctness of the transition and sensory probabilities and accuracy: 10%

Example 1 Solutions **(You can use this example to test your program).** Given the starting transition and sensory probabilities for an example of flipping a coin when three kinds of coins are used, namely, balanced B, loaded with higher probabilities of Heads L and loaded with higher probabilities of Tails M; find the most likely sequence of states that had generated the given sequence of observations. Assume that $P(X_0 = x) = 1/3$ for $x = B, L, M$. $-\log_2(P(X_0 = x)) = 1.58496$.



Original sequence of states: `BBBLLMMMMB`

Given sequence of observations: `HHHTHTTTTH`

Negative logs of transition probabilities:

	X_{i-1}=B	X_{i-1}=L	X_{i-1}=M
-log(P(B_i|X_{i-1}))	1.15200	0.94342	2.00000
-log(P(L_i|X_{i-1}))	1.51457	1.73697	2.94342
-log(P(M_i|X_{i-1}))	2.32193	2.47393	0.68966

Negative logs of sensory probabilities:

	-log(P(H_i|X_i))	-log(P(T_i|X_i))
X_i =B	1	1
X_i =L	0.23447	2.73697
X_i =M	3.32193	0.15200

Step 1. Calculate the most likely sequence of states that had generated the given sequence of observations `HHHTHTTTTH`, using the Viterbi's algorithm and the starting transition and sensory probabilities.

	H_1	H_2	H_3	T_4	H_5	T_6	T_7	T_8	T_9	H_{10}
1.5849	3.52838	5.27742	7.22083	8.96987	11.12188	12.66233	14.81433	16.96633	18.0281	18.8698
1.5849	3.33400	5.27742	7.02646	11.47237	10.71891	15.19284	16.91387	19.06587	20.7085	19.0477
1.5849	5.59655	9.12986	10.92127	9.65239	13.66398	13.34485	14.18651	15.02817	15.8698	19.8814

The most likely state sequence is: B_0,L_1,B_2,L_3,B_4,L_5,M_6,M_7,M_8,M_9,B_{10}

Iteration 1:

Expectation Step: Calculate transition and sensory probabilities, given the state sequence and using the Laplacian smoothing with k = 1.

Transition Model:

	X_{i-1}=B	X_{i-1}=L	X_{i-1}=M
P(B_i|X_{i-1})	1/6	3/6	2/7
P(L_i|X_{i-1})	4/6	1/6	1/7
P(M_i|X_{i-1})	1/6	2/6	4/7

	X_{i-1}=B	X_{i-1}=L	X_{i-1}=M
-log(P(B_i|X_{i-1}))	2.58496	1.00000	1.80735
-log(P(L_i|X_{i-1}))	0.58496	2.58496	2.80735
-log(P(M_i|X_{i-1}))	2.58496	1.58496	0.80735

Sensory Model:

	$P(H_i \mid X_i)$	$P(T_i \mid X_i)$
X_i =B	3/5	2/5
X_i =L	4/5	1/5
X_i =M	1/6	5/6

	$-\log(P(H_i \mid X_i))$	$-\log(P(T_i \mid X_i))$
X_i =B	0.73697	1.32193
X_i =L	0.32193	2.32193
X_i =M	2.58496	0.26303

Maximization Step: Calculate the most likely sequence of states that had generated the given sequence of observations `HHHTHTTTTH`, using the Viterbi's algorithm and the learned transition and sensory probabilities.

	H_1	H_2	H_3	T_4	H_5	T_6	T_7	T_8	T_9	H_{10}
1.58496	3.32193	4.22882	5.96578	7.45764	9.52803	10.68646	13.34181	14.4122	15.4825	15.9680
1.58496	2.49185	4.22882	5.13571	8.87267	8.36453	12.43492	13.59335	16.2487	17.3190	16.3894
1.58496	4.97728	6.66178	8.39874	6.98371	10.37602	10.21252	11.28291	12.3533	13.4236	16.8160

The most likely state sequence is: $B_0,L_1,B_2,L_3,B_4,L_5,M_6,M_7,M_8,M_9,B_{10}$

Iteration 2: Since the state sequence has not changed, the transition and sensory probabilities will be the same, so no need to continue.

Find accuracy of our classifier:

Original sequence: $B_0, B_1, B_2, B_3, L_4, L_5, M_6, M_7, M_8, M_9, B_{10}$

Resulting sequence: $L_1, B_2, L_3, B_4, L_5, M_6, M_7, M_8, M_9, B_{10}$

Accuracy = $\frac{7}{10} * 100\% = 0.7 * 100\% = 70\%$