Machine Learning- Assignment 5

Due Thursday December 4, 2009 No penalty until December 9, 2009

1. [20 points] Consider the data in file "hw5data.txt". There are three lines of numbers. Each line contains 200 numbers, each from the set $\{1, 2, 3, 4, 5\}$.

   For this problem, imagine that the numbers $\{1, 2, 3, 4, 5\}$ may come from one of two different sources, A and B. In the first line of data, all the numbers are generated from source A. Each number is generated independently and randomly, based on a discrete distribution over $\{1, 2, 3, 4, 5\}$ that is specific to source A. Likewise, the second line of numbers is generated from source B, but according to a different discrete distribution.

   (a) Compute the maximum likelihood estimates for A's distribution and for B's distribution, based on the data on lines 1 and 2 respectively. Report the distributions and show your code.

   The third line of numbers was generated from a mixture of A's and B's distributions in the following way. The first number was chosen to come from A's or B's distribution with equal probability. Subsequent numbers were generated in the following way. For each $i = 2, 3, 4, \ldots, 200$, number i was generated from the same distribution (A's or B's) as number $i - 1$ with probability 0.9. Otherwise (with probability 0.1) the other distribution was used.

   One way of modeling the process of generated the third line of numbers is as an HMM. Specifically, the HMM has two states, $S = \{a, b\}$ and observation set $O = \{1, 2, 3, 4, 5\}$. In state a, the HMM generates an observation o according to the distribution of source A. Similarly, in state b, the observation is generated according to the distribution of source B. The start state probabilities are $p_a = p_b = 0.5$. The transition probabilities are $p_{aa} = p_{bb} = 0.9$, and $p_{ab} = p_{ba} = 0.1$.

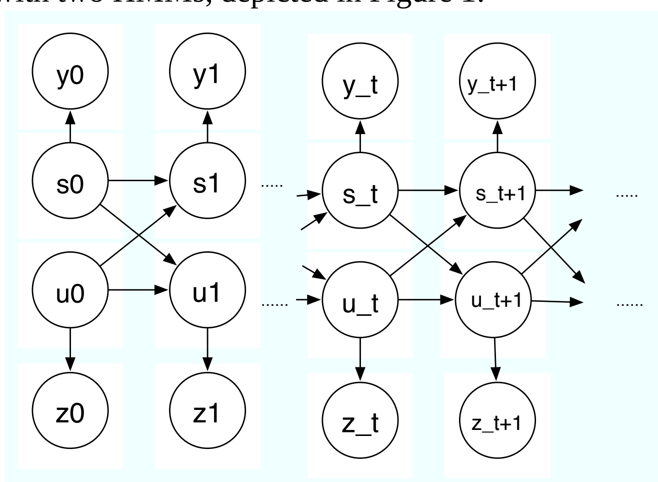
   (b) Implement the forward-backward algorithm and use it, along with the 3rd line of numbers and the HMM model described above, to compute the probabilities of each possible state at each time: $P(S_t = s|o_1, \ldots, o_{200})$, for all $t = 1, 2, 3, \ldots, 200$ and all $s = a, b$. Graph these probabilities as a function of time, t, and turn in your code.

   (c) Finally, imagine that we know the third line of numbers comes from a mixture of A's and B's distributions, but we don't know the start probabilities or transition probabilities for the HMM state. (We will assume, however, the we know the observation probabilities, simply taking them to be as computed in part (A).) Suppose we initially take $p_a = p_b = 0.5$ and $p_{aa} = p_{ab} = p_{ba} = p_{bb} = 0.5$. Implement the Baum-Welch reestimation algorithm (except do not update the observation probabilities), and use it to fit the start state distribution and state transitions of the HMM based on the third line of numbers. Report the optimized probabililities ($p_a, p_b, p_{aa}, p_{ab}, p_{ba}, p_{bb}$). Using these optimized probabilities, recompute $P(S_t = s|o_1, \ldots, o_{200})$, as in part (B). How do the new $P(S_t = s|o_1, \ldots, o_{200})$ compare with the old ones? Finally, turn in your code.

2. [50 points] **Coupled Hidden Markov Models**

We discussed in class several models for reasoning with sequences of data (trajectories). The HMM is the simplest such example, in which states are hidden, and we see observations that depend on the state. The Coupled Hidden Markov model (CHMM) is a similar kind of graphical model: we have several hidden Markov models running in parallel, and their states interact. This model is quite useful, for example, when you try to parse video, and you consider the observations as being sound and visual data, respectively.

Consider a system with two HMMs, depicted in Figure 1:



Here, s_i and u_i are the states of the two coupled HMMs, y_i and z_i are the observations coming from the two chains, and the two chains interact in the way depicted in the picture.

(a) [10 points] Specify what are the parameters of this model.

(b) [10 points] Derive an algorithm for computing the joint probability of a sequence of observations $(y_0, z_0), (y_1, z_1) \dots (y_T, z_T)$.

(c) [10 points] Derive a forward algorithm that computes the most likely sequence of hidden states given a sequence of observations. You recall that in order to do this, in the case of a simple HMM, you maintain a "belief state", which gives the probability of each hidden state based on the observations seen so far. You can use a similar idea here. Alternatively, you may consider how you can apply the junction tree algorithm to this situation.

(d) [10 points] Suppose that instead of the chains being coupled at every time step, the coupling only happens every k time steps (on time step 0, k $2k$ etc). For $k = 1$, you get the same model as above. If k is fairly large compared to the length of sequences, the chains are called **loosely coupled**. Describe how your model and the inference algorithms change in this case.

(e) [10 points] Suppose that you observe several sequences of two time series and you know that they come from a loosely coupled HMM; you know the number of possible states for each individual chain, but you do not know k. Describe a learning algorithm for this problem.

3. [30 points] **Gibbs sampling for partially observed Markov chains**

Consider a simple Markov chain where each state is 0 or 1. The initial value s_0 is drawn uniformly. The transition matrix is such that $p(s_{t+1} = s_t) = 0.9$ and $p(s_{t+1} \neq s_t) = 0.1$. Suppose that we observe $s_4 = 1$ and we want to compute $p(s_0|s_4 = 1)$.

(a) [10 points] Show how you would use Gibbs sampling in order to compute this conditional probability.

(b) [10 points] Extend your algorithm for the case in which we observe $s_t = 1$, and no other data is observed.

(c) [10 points] Describe what happens with the Gibbs sampling approach as t increases. If t was very large and you had to do approximate inference for such a problem, would you use Gibbs sampling or likelihood weighting? Justify your answer.