CSC 7810: Data Mining: Algorithms and Applications

Homework 2

Total: 100 points

Problem 1. Hidden Markov Models (30pts). Assume that we have the Hidden Markov Model (HMM) depicted in the figure below. The transition, emission and prior probabilities are given in the table below.

State	$P(S_1)$
A	0.85
B	0.15

(a) Initial probs.

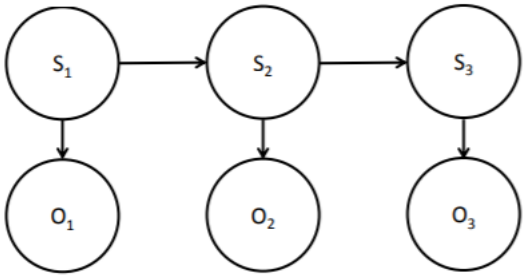
S_1	S_2	$P(S_2\|S_1)$
A	A	0.99
A	B	0.01
B	A	0.01
B	B	0.99

(b) Transition probs.

S	O	$P(O\|S)$
A	0	0.8
A	1	0.2
B	0	0.1
B	1	0.9

(c) Emission probs.

a. If each of the states can take on k different values and a total of m different observations are possible (across all states), how many parameters are required to fully define this HMM? Justify your answer (5pts).



b. Using the forward algorithm, compute the probability that we observe the sequence $O_1 = 1, O_2 = 1, and\ O_3 = 0$. Show your work (i.e., show each of your alphas) (10pts).

c. Use the Viterbi algorithm to compute (and report) the most likely sequence of states. Show your work (i.e., show each of your V_s) (10pts).

d. Is the most likely sequence of states the same as the sequence comprised of the most likely setting for each individual state? Does this make sense? Provide a 1-2 sentence justification for your answer (5 pts).

Problem 2. Modeling Data (20pts). Download the data file called geyser.txt from the course web site. This is a sequence of 295 consecutive measurements of two variables from Old Faithful geyser in Yellowstone National Park: the duration of the current eruption in minutes (to nearest 0.1 minute), and the waiting time until the next eruption in minutes (to nearest minute).

a. Examine the data by plotting the variables within and between consecutive time steps. E.g. plot(geyser(1:end-1,1),geyser(2:end,1),'o'); (10pts).

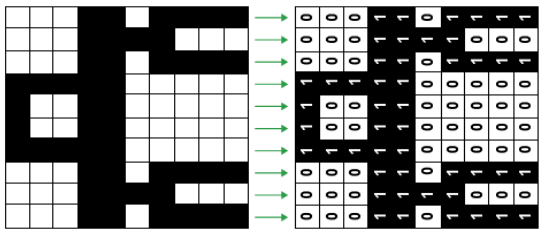
b. Discuss and justify based on your observations what kind of model might be most appropriate for this data set: e.g. a mixture of Gaussians, a hidden Markov model, a linear dynamical system, etc (10pts).

Problem 3. Modeling Data (25pts). Consider a data set consisting of the following string of 160 symbols from the alphabet {A, B, C}:

AABBBACABBBACAAAAAAAAAABBBACAAAAABACAAAAAABBBBACAAAAAAAA
AAAABACABACAABBACAAABBBBACAAABACAAAABACAABACAAABBACAAAA
BBBBACABBACAAAAAABACABACAAABACAABBBACAAAABACABBACA

Look carefully at the above string. Having analyzed the string, describe an HMM model for it. Your description should include the number of states in the HMM, the transition matrix including the values of the elements of the matrix, the emission matrix including the values of its elements, and the initial state probabilities. You need to provide some description/justification for how you arrived at these numbers. I am not expecting you to code the HMM algorithm—you should be able to answer this question just by examining the sequence carefully.

Problem 4. Expectation Maximization (25pts). Consider a data set of binary (black and white) images. Each image is arranged into a vector of pixels by concatenating the columns of pixels in the image.



The data set has N images $\{x^{(1)}, \ldots, x^{(N)}\}$ and each image has D pixels, where D is (number of rows X number of columns) in the image. For example, image $x^{(n)}$ is a vector $(x_1^{(n)}, \ldots, x_D^{(n)})$ where $x_d^{(n)} \in \{0, 1\}$ for all $n \in \{1, \ldots, N\}$ and $d \in \{1, \ldots, D\}$.

Write down the likelihood for a model consisting of a mixture of K multivariate Bernoulli distributions. Use the parameters $\pi_1, \ldots, \pi_K$ to denote the mixing proportions $(0 \leq \pi_k \leq 1; \sum_k \pi_k = 1)$ and arrange the K Bernoulli parameter vectors into a matrix P with elements p_{kd} denoting the probability that pixel d takes value 1 under mixture component k.