Problem Set 4

MAS 622J/1.126J: Pattern Recognition and Analysis

Due: 5:00 p.m. on October 22

[Note: All instructions to plot data or write a program should be carried out using Matlab. In order to maintain a reasonable level of consistency and simplicity we ask that you do not use other software tools.]

If you collaborated with other members of the class, please write their names at the end of the assignment. Moreover, you will need to write and sign the following statement: "In preparing my solutions, I did not look at any old homeworks, copy anybody's answers or let them copy mine."

Problem 1: Expectation Maximization and Missing Data [20 points]

Consider data, $D = \left\{ \begin{pmatrix} 1 \\ 3 \end{pmatrix}, \begin{pmatrix} 4 \\ 5 \end{pmatrix}, \begin{pmatrix} 2 \\ * \end{pmatrix} \right\}$, sampled from a two-dimensional (separable) distribution, $p(x_1, x_2) = p_{x_1}(x_1)p_{x_2}(x_2)$, with

$$p_{x_1}(x_1) = \begin{cases} \frac{1}{\theta_1} e^{-x_1/\theta_1} & \text{if } x_1 \geq 0 \\ 0 & \text{otherwise} \end{cases} \qquad p_{x_2}(x_2) = \begin{cases} \frac{1}{\theta_2} & \text{if } 0 \leq x_2 \leq \theta_2 \\ 0 & \text{otherwise} \end{cases}$$

and a missing feature value, *.

a. What can you infer from θ_2 by looking at D?

b. Start with an initial estimate $\underline{\theta}^0 = \begin{pmatrix} 3 \\ 6 \end{pmatrix}$ and analytically calculate $Q(\underline{\theta}; \underline{\theta}^0)$. This is the *estimate* step in the EM algorithm.

c. Find the $\underline{\theta}$ that maximizes your $Q(\underline{\theta}; \underline{\theta}^0)$ – the *maximization* step of the EM algorithm.

Problem 2: Baum-Welch Algorithm and Discrete HMMs [40 points]

Download the datasets from the course webpage. The datasets consist of training and testing sequences belonging to two classes. We assume the two HMMs for the two classes have the same configuration, i.e. the same number of states, zero transition probabilities and the number of output states.

Implement the Baum-Welch algorithm for training a discrete HMM. Train HMMs with one, three, and five states with transition probabilities in a strictly left-to-right configuration (see the figure below for a two-state HMM in left-to-right configuration). The visible output has four possible states 0, 1, 2 or 3. and

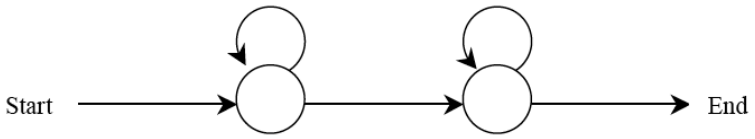


Figure 1: HMM in a left-to-right configuration

Repeat the following steps for each of the three HMM configurations with one, three, and five states:

a. Train two HMMs, one for each class of data. State very clearly the threshold you are using and the maximum number of iterations. List the output probabilities and state transition probabilities of each HMM.

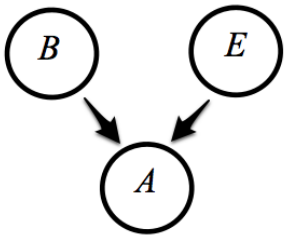
b. Implement the Viterbi algorithm to decode each test sequence using both HMMs. Show the log probability of each test sequence using each HMM.

c. Show the confusion matrix of the test set.

Include a complete listing of your source code.

Problem 3: Bayes Networks [20 points]

Consider the following Bayesian Network where the three events are B (Burglary), E (Earthquake) and A (Alarm). Assume that the three nodes are binary nodes that can take on the value f (false) or t (true).



Here, B and E are marginally independent. That is, $P(B, E) = P(B)P(E)$ (Note that this relation implies $P(B|E) = P(B)$ or $P(E|B) = P(E)$.) Now we want to show B and E are conditionally dependent given A. In other words, $P(B, E|A) \neq P(B|A)P(E|A)$.

a. Prove that the relation $P(B, E|A) \neq P(B|A)P(E|A)$ implies $P(B|E, A) \neq P(B|A)$. Note that by the symmetry between B and E, this also means $P(E|B, A) \neq P(E|A)$.

b. First, we consider a very simple case. Suppose $A = B + E$ where the $+$ sign means "Logical OR." This means that B and E are independent deterministic causes of A. Construct the CPT (conditional probability table) for $P(A\ |B, E)$.

c. For the obtained CPT above, prove that $P(B = t|A = t) = P(B = t)/P(A = t)$. Which is greater between $P(B = t|A = t)$ and $P(B = t)$? What is the meaning of this?

d. For the obtained CPT above, prove that $P(A = t) = P(B = t)+P(E = t) - P(B = t)P(E = t)$. If $P(B = t)$ and $P(E = t)$ are small, what happens to $P(B = t|A = t)$, compared with $P(B = t)$?

e. For the obtained CPT above, prove that $P(B = t|E = t, A = t) = P(B = t)$. Which is greater between $P(B = t|E = t, A = t)$ and $P(B = t|A = t)$? What is the meaning of this? *Hint:* Note that the observation of Earthquake ($E = t$) alone is enough to explain the cause of Alarm ($A = t$). This occurrence is called "Explaining Away."

f. We assume that $P(B = t) = 0.15$ and $P(E = t) = 0.005$. Calculate $P(A = t), P(B = t|A = t), P(B = t|E = t, A = t), P(E = t|A = t), P(E = t|B = t, A = t)$ by hand.