Problem Set 4

MAS 622J/1.126J: Pattern Recognition and Analysis

Due: 5:00 p.m. on October 20

[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. 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.

   Solution:

$$\begin{aligned} p(\mathbf{x}|\underline{\theta}) = p(x_1, x_2|\underline{\theta}) &= p_{x1}(x_1|\underline{\theta})p_{x_2}(x_2|\underline{\theta}) \\ &= \begin{cases} \left(\frac{1}{\theta_1} e^{-x_1/\theta_1}\right)\left(\frac{1}{\theta_2}\right) & \text{if } x_1 \geq 0 \text{ and } 0 \leq x_2 \leq \theta_2 \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{1}{\theta_1\theta_2} e^{-x_1/\theta_1} & \text{if } x_1 \geq 0 \text{ and } 0 \leq x_2 \leq \theta_2 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

   With our initial parameters, our distribution becomes:

$$p(\mathbf{x}|\underline{\theta}^0) = p(x_1, x_2|\underline{\theta}^0)$$
$$= \begin{cases} \frac{1}{15}e^{-x_1/3} & \text{if } x_1 \geq 0 \text{ and } 0 \leq x_2 \leq 6 \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned} Q(\underline{\theta};\underline{\theta}^0) &= E_{x_{32}}[\ln p(\mathbf{x}_g, \mathbf{x}_b;)|\underline{\theta}^0; D_g] \\ &= \int_{-\infty}^{\infty} \left[\sum_{k=1}^{2} \ln p(\mathbf{x}_k|\underline{\theta}) + \ln p(\mathbf{x}_3|\underline{\theta})\right] p(x_{32}|x_{31}=2;\underline{\theta}^0)dx_{32} \\ &= \sum_{k=1}^{2} \ln p(\mathbf{x}_k|\underline{\theta}) + \int_{-\infty}^{\infty} \ln p\left(\begin{bmatrix} 2 \\ x_{32} \end{bmatrix}|\underline{\theta}\right) \frac{p\left(\begin{bmatrix} 2 \\ x_{32} \end{bmatrix}|\underline{\theta}^0\right)}{\int_{-\infty}^{\infty} p\left(\begin{bmatrix} 2 \\ x'_{32} \end{bmatrix}|\underline{\theta}^0\right)dx'_{32}} dx_{32} \end{aligned}$$

Note that x_1 and x_2 are independent with each other. Thus,

$$p\left(\begin{bmatrix} 2 \\ x_{32} \end{bmatrix}|\underline{\theta}^0\right) = p_{x_1}\left(x_1 = 2|\underline{\theta}^0\right) p_{x_2}(x_{32}|\underline{\theta}^0),$$

$$\int_{-\infty}^{\infty} p\left(\begin{bmatrix} 2 \\ x'_{32} \end{bmatrix}|\underline{\theta}^0\right)dx'_{32} = p_{x_1}\left(x_1 = 2|\underline{\theta}^0\right)\int_{-\infty}^{\infty} p_{x_2}(x'_{32}|\underline{\theta}^0)dx'_{32} = p_{x_1}\left(x_1 = 2|\underline{\theta}^0\right)$$

$$Q(\underline{\theta};\underline{\theta}^0) = \sum_{k=1}^{2} \ln p(\mathbf{x}_k|\underline{\theta}) + \int_{-\infty}^{\infty} \ln p\left(\begin{bmatrix} 2 \\ x_{32} \end{bmatrix}|\underline{\theta}\right) p_{x_2}(x_{32}|\underline{\theta}^0)dx_{32}$$

Note that θ_2 is necessarily greater or equal to 4 because $0 < x_{12} = 2 < x_{22} = 4 \leq \theta_2$.
Also, $p\left(\begin{bmatrix} 2 \\ x_{32} \end{bmatrix}|\underline{\theta}\right)$ is different than zero if $0 \leq x_{32} \leq \theta_2$, and $p_{x_2}(x_{32}|\underline{\theta}^0)$ is different than zero if $0 \leq x_{32} \leq \theta_2^0 = 6$.
Therefore, $p\left(\begin{bmatrix} 2 \\ x_{32} \end{bmatrix}|\underline{\theta}\right)p_{x_2}(x_{32}|\underline{\theta}^0)$ is different than zero between $[0, \min(\theta_2\ , 6)]$

$$\sum_{k=1}^{2} \ln p(\mathbf{x}_k|\underline{\theta}) = \ln\frac{1}{\theta_1\theta_2}e^{-1/\theta_1} + \ln\frac{1}{\theta_1\theta_2}e^{-4/\theta_1} = 2\ln\frac{1}{\theta_1\theta_2} - \frac{6}{\theta_1}$$

$$\begin{aligned} \int_{-\infty}^{\infty} \ln p\left(\begin{bmatrix} 2 \\ x_{32} \end{bmatrix}|\underline{\theta}\right) p_{x_{32}}\left(x_{32}|\underline{\theta}^0\right)dx_{32} &= \int_0^{\min(\theta_2,6)} \left(\ln\frac{1}{\theta_1\theta_2}e^{-2/\theta_1}\right)\left(\frac{1}{6}\right)dx_{32} \\ &= \begin{cases} \frac{1}{6}\theta_2\left(\ln\frac{1}{\theta_1\theta_2} - \frac{2}{\theta_1}\right) & \text{if } \min(\theta_2, 6) = \theta_2, \quad \text{where } 4 \leq \theta_2 < 6 \\ \ln\frac{1}{\theta_1\theta_2} - \frac{2}{\theta_1} & \text{if } \min(\theta_2, 6) = 6, \quad \text{where } \theta_2 \geq 6 \end{cases} \end{aligned}$$

$$Q(\underline{\theta};\underline{\theta}^0) = \sum_{k=1}^{2} \ln p(\mathbf{x}_k|\underline{\theta}) + \int_{-\infty}^{\infty} \ln p\left(\left[\begin{array}{c} 2 \\ x_{32} \end{array}\right] |\underline{\theta}\right) p_{x_{32}}\left(x_{32}|\underline{\theta}^0\right) dx_{32}$$
$$= \begin{cases} 2\ln\frac{1}{\theta_1\theta_2} - \frac{6}{\theta_1} + \frac{1}{6}\theta_2\left(\ln\frac{1}{\theta_1\theta_2} - \frac{2}{\theta_1}\right) & \text{if } \min(\theta_2, 6) = \theta_2, \quad \text{where } 4 \le \theta_2 < 6 \\ 3\ln\frac{1}{\theta_1\theta_2} - \frac{7}{\theta_1} & \text{if } \min(\theta_2, 6) = 6, \quad \text{where } \theta_2 \ge 6 \end{cases}$$

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

Solution:

First,

$$\frac{\partial Q(\underline{\theta};\underline{\theta}^0)}{\partial \theta_2} = \begin{cases} -\frac{2}{\theta_2} - \frac{1}{5}\left(\frac{2}{\theta_1} - \ln\frac{1}{\theta_1\theta_2}\right) - \frac{1}{5} & \text{where } 4 \le \theta_2 < 5 \\ -\frac{3}{\theta_2} & \text{where } \theta_2 \ge 5 \end{cases}$$

Thus, $Q(\underline{\theta};\underline{\theta}^0) < 0$ where $\theta_2 \ge 4$. (Note $\frac{2}{\theta_1} - \ln\frac{1}{\theta_1\theta_2} = \frac{1}{\theta_1} + (\frac{1}{\theta_1} - \ln\frac{1}{\theta_1}) + \ln\theta_2 > 0$ because $\frac{1}{\theta_1} > 0$, $(\frac{1}{\theta_1} - \ln\frac{1}{\theta_1}) \ge 1$ and $\ln\theta_2 > 0$ where $\theta_1 > 0$ and $\theta_2 \ge 4$.)

Second, $Q(\underline{\theta};\underline{\theta}^0)$ is continuous at $\theta_2 = 5$, because the two equations corresponding to the two intervals ($4 \le \theta_2 < 5$, $\theta_2 \ge 5$) have the same value at $\theta_2 = 5$, which is $3\ln\frac{1}{5\theta_1} - \frac{7}{\theta_1}$.

From these two facts, we can conclude that $Q(\underline{\theta};\underline{\theta}^0)$ is a monotonically decreasing continuous function with respect to θ_2. Thus, the maximum occurs when $\theta_2 = 4$.

To find the value of θ_1 that maximizes Q when $\theta_2 = 4$, $\frac{\partial Q(\underline{\theta};\underline{\theta}^0)}{\partial \theta_1} = -\frac{2}{\theta_1} + \frac{5}{\theta_1^2} - \frac{4}{5\theta_1} + \frac{8}{5\theta_1^2} = 0$. Thus, we find that $\arg\max_{\theta_1} Q(\underline{\theta};\underline{\theta}^0) = \frac{33}{14}$ (i.e., $\theta_1 = \frac{33}{14}$ and $\theta_2 = 4$).

Thus, $\underline{\theta} = (\frac{33}{14}, 4)$ and $\max Q(\underline{\theta};\underline{\theta}^0) \cong -9.0825$.

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

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.

Figure 1: HI

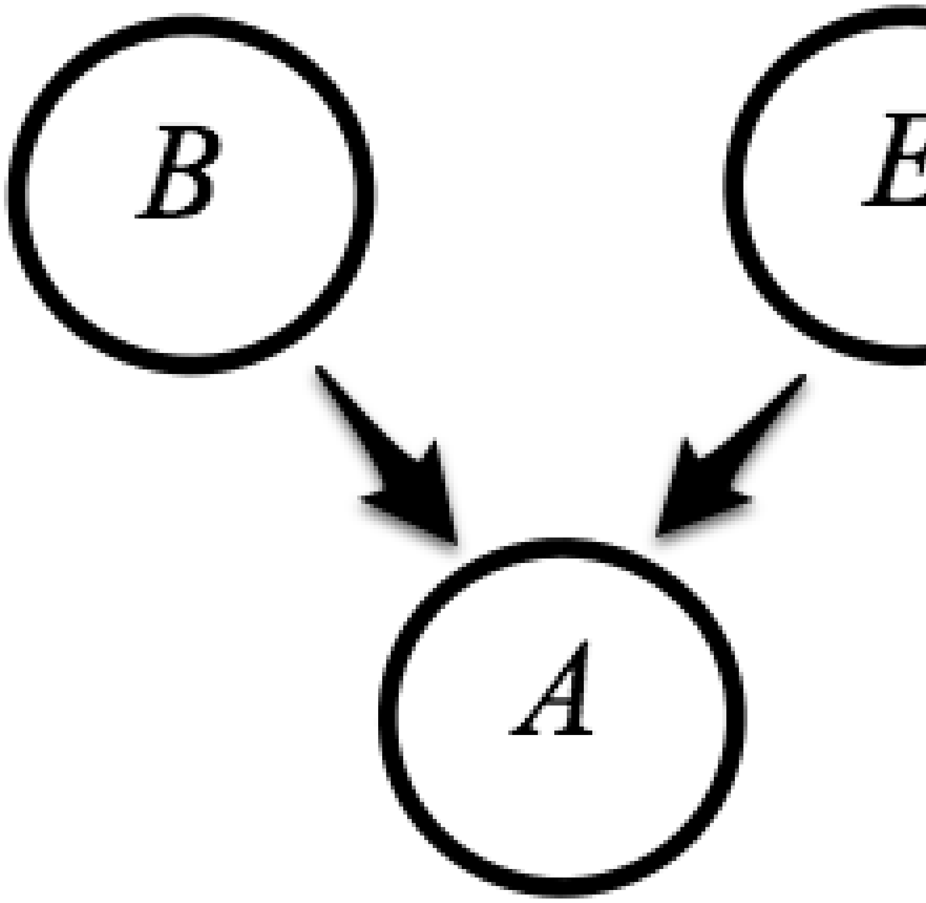
b. Implement the Viterbi algorithm to
probability of each test sequence usi

c. Compute the recognition accurary o

Include a complete listing of your source

Problem 3: Bayes Netw

Consider the following Bayesian Network
(Alarm). Assume that the three nodes ar



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 $P(E|B,A) \neq P(E|A)$.

Solution: From $P(B, E|A) \neq P(B|A)P(E|A)$, we get the relation $\frac{P(B,E|A)}{P(E|A)} \neq P(B|A)$

Also, we know that $P(B|E, A) = \frac{P(B,E|A)}{P(E|A)}$ by Bayes' rule.
Therefore, $P(B|E, A) \neq P(B|A)$

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 $P(A\ |B, E)$.

Solution:

$B,\ E$	$P(A = f\ \|B, E)$	$P(A = t\ \|B, E)$
f, f	1	0
t, f	0	1
f, t	0	1
t, t	0	1

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?

Solution: $P(B = t|A = t) = P(A = t, B = t)/P(A = t)$ by Bayes' rule.

Now for $P(A = t, B = t)$,

$$\begin{aligned}
&P(A = t, B = t)\\
&= \sum_{E=t,f} P(A = t, B = t, E) \quad \text{(by marginalization)}\\
&= \sum_{E=t,f} P(B = t)P(E)P(A = t|B = t, E) \quad \text{(from the graphical model)}\\
&= P(B = t)P(E = t)P(A = t|B = t, E = t) + P(B = t)P(E = f)P(A = t|B = t, E = f)\\
&= P(B = t)P(E = t) + P(B = t)P(E = f)\\
&= P(B = t)\,(P(E = t) + P(E = f))\\
&= P(B = t)
\end{aligned}$$

Therefore, $P(B = t|A = t) = P(B = t)/P(A = t)$

Since $P(A = t) \leq 1$, $P(B = t|A = t) \geq P(B = t)$

Note that $P(B = t)$ means your belief on $B = t$ (Burglary happened) without being given any information about E and A. However, $P(B = t|A = t)$ means your belief on $B = t$ (Burglary happened) when you know $A = t$ (Alarm happened). Thus, $P(B = t|A = t) \geq P(B = t)$ means that your belief of $B = t$ (Burglary happend) increases when you know $A = t$ (Alarm happened).

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 $P(B = t)$?

Solution:

$$
\begin{aligned}
&P(A=t) \\
&= \sum_{B=t,f} \sum_{E=t,f} P(A=t,B,E) \\
&= \sum_{B=t,f} \sum_{E=t,f} P(B)P(E)P(A=t|B,E) \\
&= P(B=t)P(E=t)P(A=t|B=t,E=t) + P(B=t)P(E=f)P(A=t|B=t,E=f) \\
&+P(B=f)P(E=t)P(A=t|B=f,E=t) + P(B=f)P(E=f)P(A=t|B=f,E=f) \\
&= P(B=t)P(E=t) + P(B=t)P(E=f) + P(B=f)P(E=t) \\
&= P(B=t)\left(P(E=t)+P(E=f)\right) + \left(1-P(B=t)\right)P(E=t) \\
&= P(B=t) + P(E=t) - P(B=t)P(E=t)
\end{aligned}
$$

Thus, as $P(B=t)$ and $P(E=t)$ gets smaller, $P(A=t)$ also gets smaller. Since $P(B=t|A=t) = P(B=t)/P(A=t)$ (from (d)), $P(B=t|A=t)$ becomes much greater than $P(B=t)$.

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."

Solution:

$$
\begin{aligned}
&P(B=t|E=t,A=t) \\
&= \tfrac{P(B=t,E=t,A=t)}{P(E=t,A=t)} \\
&= \tfrac{P(B=t)P(E=t)P(A=t|B=t,E=t)}{P(E=t,A=t)} && \text{(from the graphical model)} \\
&= \tfrac{P(B=t)P(E=t)P(A=t|B=t,E=t)}{P(E=t)} && \text{(since } P(E=t,A=t)=P(E=t)\text{)} \\
&= P(B=t)
\end{aligned}
$$

Note that in (c), we proved $P(B=t,A=t) = P(B=t)$. In like manner, we can prove $P(E=t,A=t) = P(E=t)$.

Since $P(B=t|E=t,A=t) = P(B=t)$ and $P(B=t|A=t) \geq P(B=t)$ (from (c)), we get

$P(B=t|A=t) \geq P(B=t) = P(B=t|E=t,A=t)$

Note that $P(B=t|E=t,A=t)$ denotes your belief on $B=t$ (Burglary happened) when you know both Earthquake and Alarm happened $(E=t,A=t)$. Here the observation of Earthquake $(E=t)$ alone is enough to explain the cause of Alarm $(A=t)$. Therefore, since the fact that Earthquake happened can "explain away" the fact that Alarm happened, there is no increase in your belief on $B=t$ (Burglary happened). Thus, $P(B=t|E=t,A=t) = P(B=t)$.

f. We assume that $P(B=t) = 0.1$ and $P(E=t) = 0.01$. 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.

Solution:

$P(A=t) = P(B=t) + P(E=t) - P(B=t)P(E=t) = 0.1 + 0.01 - 0.1 \times 0.01 = 0.109$
$P(B=t|A=t) = P(B=t)/P(A=t) = 0.1/0.109 = 0.9174$
$P(B=t|E=t,A=t) = P(B=t) = 0.1$

$P(E=t|A=t) = P(E=t)/P(A=t) = 0.09174$
$P(E=t|B=t, A=t) = P(E=t) = 0.01$

g. Now install Kevin Murphy's Bayes Net Toolbox for Matlab. You can download the toolbox and find the install and usage instructions here:

`http://bnt.sourceforge.net/`

We compute the above $P(A=t), P(B=t|A=t), P(B=t|E=t, A=t), P(E=t|A=t), P(E=t|B=t, A=t)$ using this toolbox. To help you use the toolbox, we provide you with the basic Matlab code for this problem.

Download the Matlab code **"conditional_dep.m"** from the course website. Fill in the core part in the Matlab code and compute $P(A=t), P(B=t|A=t), P(B=t|E=t, A=t), P(E=t|A=t), P(E=t|B=t, A=t)$. Are the computed values the same to what you've obtained by hand?

Solution:

```
% MIT MAS622j/1.126J, Problem Set 3
% Conditional Dependence, Explaining Away
% (Burglary-->Alarm<---Earthquake example)
%
% (for Problem 2(g))

clear all
N = 3;
dag = zeros(N,N);
B = 1; E = 2; A = 3;
dag(B,A) = 1;
dag(E,A) = 1;

false= 1; true= 2;
ns = 2*ones(1,N); % binary nodes
names = {'B', 'E', 'A'};

% Set up the topology of the Bayes Net
bnet = mk_bnet(dag, ns, 'names', names, 'discrete', 1:N);

% Display the graph: download graph_draw from the link on the class webpage
graph_draw(bnet.dag, 'node_labels', names);

% Set up conditional probability tables for each node
% ### Your code comes here! ###
% Hint: Fill in the row vectors with CPD values
bnet.CPD{B} = tabular_CPD(bnet, B, [0.9  0.1]);
bnet.CPD{E} = tabular_CPD(bnet, E, [0.99 0.01]);
```

```
bnet.CPD{A} = tabular_CPD(bnet, A, [1 0 0 0 0 1 1 1]);

% Choose an inference engine
engine = jtree_inf_engine(bnet);

% First, compute P(A)
evidence = cell(1,N);
[engine, ll] = enter_evidence(engine, evidence);

m = marginal_nodes(engine, A);
disp(sprintf('p(A) = %g', m.T(true)));

% Second, compute P(B|A)
% ### Your code comes here! ###
% Hint: Enter the observation that Alarm = true
evidence = cell(1,N);
evidence{A} = true;
[engine, ll] = enter_evidence(engine, evidence);

m = marginal_nodes(engine, B);
disp(sprintf('p(B|A) = %g', m.T(true)));

% Third, compute P(B|E,A)
% ### Your code comes here! ###
% Hint: Enter the observation that Earthquake = true and Alarm = true
evidence = cell(1,N);
evidence{E} = true;
evidence{A} = true;
[engine, ll] = enter_evidence(engine, evidence);

m = marginal_nodes(engine, B);
disp(sprintf('p(B|E,A) = %g', m.T(true)));

% Fourth, compute P(E|A)
% ### Your code comes here! ###
% Hint: Enter the observation that Earthquake = true
evidence = cell(1,N);
evidence{A} = true;
[engine, ll] = enter_evidence(engine, evidence);

m = marginal_nodes(engine, E);
disp(sprintf('p(E|A) = %g', m.T(true)));

% Fifth, compute P(E|B,A)
% ### Your code comes here! ###
```

```
% Hint: Enter the observation that Burglary = true and Alarm = true
evidence = cell(1,N);
evidence{B} = true;
evidence{A} = true;
[engine, ll] = enter_evidence(engine, evidence);

m = marginal_nodes(engine, E);
disp(sprintf('p(E|B,A) = %g', m.T(true)));
```

The answers by toolbox:
$p(A = t) = 0.109$
$p(B = t|A = t) = 0.917431$
$p(B = t|E = t, A = t) = 0.1$
$p(E = t|A = t) = 0.0917431$
$p(E = t|B = t, A = t) = 0.01$

h. Second, we consider a more complex case. Assume that $P(B = t) = 0.1$, $P(E = t) = 0.01$ and $P(A\ |B, E)$ is described by the following CPT.

B, E	$P(A = f\ \|B, E)$,	$P(A = t\ \|B, E)$
f, f	0.99	0.01
t, f	0.05	0.95
f, t	0.01	0.99
t, t	0.001	0.999

Compute $P(A = t), P(B = t|A = t), P(B = t|E = t, A = t), P(E = t|A = t), P(E = t|B = t, A = t)$ using the toolbox.

Solution: The CPDs for this problem:

```
bnet.CPD{B} = tabular_CPD(bnet, B, [0.9  0.1]);
bnet.CPD{E} = tabular_CPD(bnet, E, [0.99 0.01]);
bnet.CPD{A} = tabular_CPD(bnet, A, [0.99 0.05 0.01 0.001 0.01 0.95 0.99 0.999]);
```

The answers by toolbox:
$p(A = t) = 0.112869$
$p(B = t|A = t) = 0.842118$
$p(B = t|E = t, A = t) = 0.100817$
$p(E = t|A = t) = 0.087792$
$p(E = t|B = t, A = t) = 0.0105104$
(Note that $P(B = t|A = t) > P(B = t|E = t, A = t)$ here.)