STAT 535 Final Exam

Wednesday December 12, 2012, 4:30-6:20

Student name: ..........................................................

Problem 1 (Independence, parametrization, junction tree, VE) 11
Problem 2 (ML estimation) 5
Problem 3 (K-means) 3
Problem 4 (Inference) 3

Total 22 points

• notes and books are allowed
• electronic devices are not allowed
• OK to write on backs of pages

• Do Well!
Problem 1 – Independencies, parametrization, junction tree, VE

No proofs are required for this problem.

1.1 List 3 V-structures in $G$; at least one of these should involve $Y$.

1.2 answer the independence questions below. When an independence is false, give an open path.

<table>
<thead>
<tr>
<th>Independence</th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A \perp Y \mid CE$</td>
<td>True</td>
<td>False</td>
</tr>
<tr>
<td>$A \perp U$</td>
<td>True</td>
<td>False</td>
</tr>
<tr>
<td>$G \perp Y$</td>
<td>True</td>
<td>False</td>
</tr>
<tr>
<td>$G \perp Y \mid T$</td>
<td>True</td>
<td>False</td>
</tr>
<tr>
<td>$S \perp U \mid TY$</td>
<td>True</td>
<td>False</td>
</tr>
</tbody>
</table>
1.3 Moralize graph $\mathcal{G}$ above.

1.4 Is the moral graph you obtained chordal? Prove or disprove by Tarjan elimination. Show the cliques appearing and whether they are maximal or not.

1.5 Construct a junction tree from the moral graph above; make a clearly labeled drawing of the junction tree. Denote the junction tree you obtain by $\mathcal{J}$. 
1.6 Assume $P_V, V = \{A, B, C, \ldots, Y, Z\}$ is a joint distribution that factors according to graph $\mathcal{G}$. Denote by $\{P_{X|\text{pa}(X)}, X \in V\}$ the Bayes net parametrization of $P_V$. Write the junction tree representation of $P_V$ in terms of clique and separator marginals for the junction tree you obtained in 1.5. Denote by $\phi_{\text{clique}}$, respectively $\phi_{\text{sep}}$, the clique and separator tables in the representation of $\mathcal{J}$.

JT parametrization of $P_V$ according to $\mathcal{J} 

1.7 Initialize the $\phi_{\text{clique}}, \phi_{\text{sep}}$ potentials using the given $P_{X|\text{pa}(X)}$ potentials, i.e. write the clique potentials of $\mathcal{J}$ as a product of factors from the Bayes Net representation. It is sufficient to do this for

- a clique that contains $U$
- a clique that contains $S$
- a clique that contains $G$ and $F$

If there are multiple solutions, select any of them you want.
Starting From the Bayes Net parametrization of $P_V$, describe the steps of an efficient Variable Elimination procedure to compute $P_{Y|U=u}$. For each step, give the variable eliminated, and the potential that is created. If the new potential is 1, just state so; if the potential is not identical to 1, then show the expression of this potential table in terms of existing potentials. For full credit you must eliminate in the most efficient possible way.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Optional [Neighbors]</th>
<th>Optional [Potentials deleted]</th>
<th>New potential and how obtained</th>
<th>Optional [Graph]</th>
</tr>
</thead>
</table>

3 points
Problem 2 – ML Estimation

Consider the Bayes Net defined by the graph below, where variables $A, B, C, D$ take values in $\{0, 1\}$. This Bayes Net is parametrized as follows:

$$P_A(1) = p \quad P_A(0) = q$$

$$P_{B|A}(x|a) = P_{C|A}(x|a) = P_{D|A}(x|a) = \theta(x|a) \text{ for } x \in \{0, 1\} \text{ and } a \in \{0, 1\}$$

For full credit, you must bring results to their simplest form. For questions 2.2, 2.3, no need to show your work.

2.1 Write the expression of $l = \ln P_{ABCD}(D|p, q, \theta)$ the log-likelihood of the parameters, given a data set $D$ of size $N$. 

more space on next page
2.2 What are the sufficient statistics for this parametrization?
2.3 For the sample $D$ below, calculate the sufficient statistics and Maximum Likelihood estimates of the parameters $p, q, \theta$ (numerical answers only).

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>$C$</th>
<th>$D$</th>
<th>times observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

$N = 20$

\[
\hat{\theta} =
\begin{array}{c|cc}
  x = 0 & x = 1 \\
  A = 0 & & \\
  A = 1 & & \\
\end{array}
\]

$\hat{p} = \quad \hat{q} =$
Problem 3 – A special case of the K-means algorithm

The following variant of the K-means algorithm is called SEQUENTIAL K-MEANS (SKM). We shall denote as usual by \( C_k, n_k \) and \( k(i) \) the \( k \)-th cluster, its size, and the assignment of point \( i \) respectively.

**Input** Data \( \mathcal{D} = \{x_1, \ldots, x_n\} \) with \( x_i \in \mathbb{R}^d \) for \( i = 1 : n \), number of clusters \( K \)

**Initialize** with centers \( \mu_{1:K} \)

Repeat

for \( i = 1, 2, \ldots, n \)

1. re-assign \( x_i \): set \( k(i) \leftarrow \arg \min_{k=1:K} ||\mu_k - x_i|| \)

2. re-calculate centers \( \mu_k \leftarrow \left( \sum_{x_j \in C_k} x_j \right) / n_k \)

until no change in \( k(i) \)

Denote by \( \Delta \) the K-means cost, i.e. \( \Delta(\mu_{1:K}, k(1 : n)) = \sum_{i=1}^{n} ||x_i - \mu_{k(i)}||^2 \).

1 point

**3.1** Prove that step 1 of SKM always decreases \( \Delta \) or leaves it the same.

1 point

**3.2** Prove that step 2 of SKM always decreases \( \Delta \) or leaves it the same.
3.3 Show how the centers $\mu_{1:K}$ can be updated in step 2 of SKM with a constant number of operations per data point.

A note about this algorithm and a puzzle: From question 3.3 it follows that SKM is almost as efficient as standard K-means. SKM also has provably no more local minima than standard K-means, which means that in theory it should be less sensitive to initialization. I have run extensive comparison experiments between the two algorithms, and did not find one case in which the two algorithms gave different answers.
Consider the following decomposable graphical model. All variables are binary and all potentials are strictly positive.

Graph $G$

\[ P_V \propto \phi_V = \frac{\phi_{ABC}\phi_{BCE}\phi_{BDE}\phi_{DEG}\phi_{CEF}}{\phi_{BC}\phi_{BE}\phi_{DE}\phi_{CE}} \]

where $V = \{A, B, C, D, E, F, G\}$

(This is the graph in your homeworks 4 and 5.)

Below is a partial “answer sheet” obtained by running the Junction Tree algorithm on $P_V$. All runs of the Junction Tree algorithm start from the same uncalibrated initial state of the clique and separator potentials. The literals $x, x_0, \ldots, p$ below stand for known numerical values, and $Z, Z_0, Z_1$ are the normalization constants obtained in the Normalize step of the Junction Tree algorithm.

\[
\begin{array}{c|c|c}
P_{ABC}(0, 0, 0) &= x & Z = z \\
P_{ABC|G=0}(0, 0, 0) &= x_0 & Z_0 = z_0 \\
P_{ABC|G=1}(0, 0, 0) &= ? & Z_1 = ? \\
P_G(1) &= ? \\
\end{array}
\]

Show how to complete the missing values in this table.
That’s all, folks!