Grad AI. 15-780 Fall, 2006

Homework 3 Solutions

1. **Bayesian Networks [33 pts]**. This problem involves a theoretical analysis of *admissible* Bayesian networks. Recall from lecture that an admissible Bayesian network must be a Directed Acyclic Graph (DAG).

(a)&(b) **[8 pts]** The networks and equivalence classes are shown in Figure 1 inside dashed-boxes.

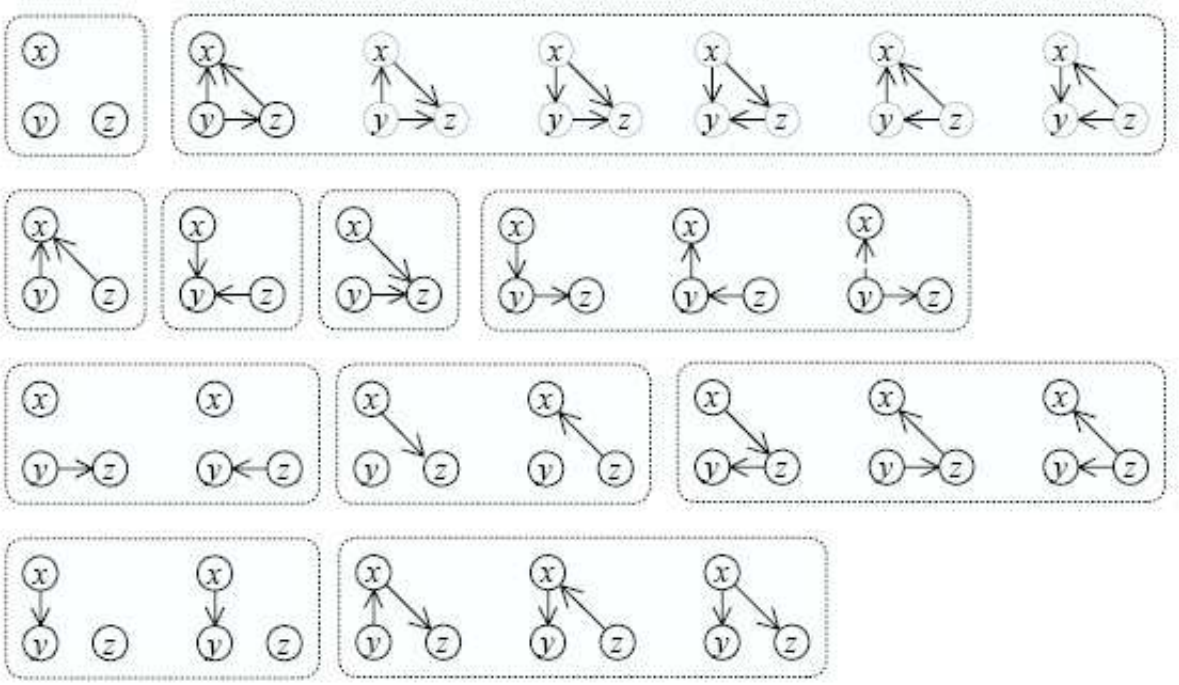

Figure 1: Equivalence Classes of Bayesian Networks

(c) **[10 pts]** A suitable upper-bound is given by considering all pairs of nodes. In a Bayesian network with n nodes there are $\binom{n}{2}$ pairs. For each pair of nodes i and j there can be an edge from i to j, an edge for j to i or no edge at all between them. This provides an upper bound of $3^{\binom{n}{2}}$. Note this is an upper-bound because this counts some networks with cycles (e.g. $i \rightarrow j$, $j \rightarrow k$, $k \rightarrow i$).

(d) **[15 pts]** For our lower bound we will impose an arbitrary ordering on the nodes from $1, \ldots, n$. We will count all networks with nodes i and j such that $i < j$ and there is either an edge from $i \rightarrow j$ or no edge between the two nodes. Counting only networks with edges leaving nodes earlier in the ordering and entering nodes later in the ordering ensures there are no cycles. Again there are $\binom{n}{2}$ such pairs making our bound $2^{\binom{n}{2}}$. The ratio between our bounds is,

$$r = \frac{3^{\binom{n}{2}}}{2^{\binom{n}{2}}} = \left(\frac{3}{2}\right)^{\binom{n}{2}}$$

2. **Maximum Likelihood Estimation [23 pts]**. In class we derived the Maximum Likelihood Estimator (MLE) for the single parameter of a Binomial distribution (e.g. the probability that a coin lands heads after observing the outcome of n independent flips of the coin). In this problem we will derive the MLE for the parameters of a multinomial distribution where the variable of interest, X, can take on k values rather than 2.

   (a) **[5 pts]** Given data describing n independent identically distributed observations of X, $\mathbf{d} = \{d_1, \ldots, d_n\}$, each of which can be one of k values, express the likelihood of the data given $k-1$ parameters for the distribution over X. Let n_i represent the number of times X takes on value i in the data.

$$\prod_i^{k-1} \theta_i^{n_i} \left(1 - \sum_i^{k-1} \theta_i\right)^{n_k}$$

   (b) **[6 pts]** Find the MLE for one of the $k-1$ parameters, θ_j in terms of the other parameters.

$$\begin{aligned}
l(x \mid \theta) &= \log\left(\prod_i^{k-1} \theta_i^{n_i} \left(1 - \sum_i^{k-1} \theta_i\right)^{n_k}\right) \\
&= \sum_i^{k-1} \log\left(\theta_i^{n_i}\right) + \log\left(1 - \sum_i^{k-1} \theta_i\right)^{n_k} \\
\frac{\partial}{\partial \theta_j} l(\theta \mid x) &= \frac{\partial}{\partial \theta_j} n_j \log\left(\theta_j\right) + \frac{\partial}{\partial \theta_j} n_k \log\left(1 - \sum_i^{k-1} \theta_i\right) \\
&= \frac{n_j}{\theta_j} - \frac{n_k}{1 - \sum_i^{k-1} \theta_i}
\end{aligned}$$

   Setting this equal to 0 we have the following,

$$\begin{aligned}
0 &= \frac{n_j}{\hat{\theta}_j} - \frac{n_k}{1 - \sum_i^{k-1} \theta_i} \\
\hat{\theta}_j &= \frac{n_j \left(1 - \sum_i^{k-1} \theta_i\right)}{n_k}
\end{aligned}$$

   (c) **[12 pts]** At this point you should have $k-1$ equations describing MLE's of different parameters. Show how those equations imply that the MLE for a parameter

θ_j representing the probability that X takes on value j is equal to $\frac{n_j}{n}$. You may find the following hint useful for this: In order to remove the k'th parameter from the likelihood in part (a) you had to represent it with an equation, $\theta_k = f()$. At this point you may find it helpful to replace all occurrences of $f()$ with θ_k. After replacing $f()$ with θ_k you can substitute all occurrences of each other parameter in $f()$ with its MLE from part (b). This should allow you to solve for the MLE of θ_k, which can then be used to simplify all of the other equations.

$$\begin{aligned}
\hat{\theta_k} &= 1 - \sum_i^{k-1} \hat{\theta_i} \\
\hat{\theta_j} &= \frac{n_j \hat{\theta_k}}{n_k} \\
\hat{\theta_k} &= 1 - \sum_i^{k-1} \frac{n_i \hat{\theta_k}}{n_k} \\
&= 1 - \frac{\hat{\theta_k}}{n_k}(n - n_k) \\
&= \frac{n_k}{n}
\end{aligned}$$

Now we can substitute the value of $\hat{\theta_k}$ back into the original equation.

$$\begin{aligned}
\hat{\theta_j} &= \frac{n_j n_k}{n_k n} \\
&= \frac{n_j}{n}
\end{aligned}$$

3. **Hidden Markov Models [44 pts]**. Many of you may be familiar with the T9 input paradigm commonly found on cell phones for interpreting a series of key presses on the 9 button keypad as text. Typically, each of the keys 2-9 represents about 3 different letters (Figure 2 provides the exact mapping). When the user inputs a series of key presses, such as 3-6-4, the T9 system provides a list of words from its dictionary that potentially match the sequence (e.g. "dog" and "fog" both match the sequence above). In this problem you will build a next generation text messaging cell phone input system, SmarT9, by training an HMM on an English text corpus. For simplicity, we will ask you to build your system only for 5 letter words and we will provide a small corpus on the course website.

   (a) **[5 pts]** The hidden states in our HMM will be letters of the word intended by the user. The observable outputs are the digits 2-9 entered on the key pad.

   (b) **[3 pts]** The probability that a state emits a particular output is 1 if its one of the letters on the key corresponding to that digit, and 0 otherwise. Let $b_i(o_j)$ be 1 if letter i emits observation o_j.

Figure 2: A typical phone keypad.

(c) **[10 pts]** Programming problem. You can verify your table by checking that the probability u follows q is close to 1. Let a_{ij} be the probability that state i is followed by state j.

(d) **[6 pts]** Programming problem. You can verify your table by checking that the probability the initial state starts with 'S' is largest and ≈ 0.1. Let π_i be the probability that state 1 has value i.

(e) **[10 pts]** Programming problem. The definition of α from lecture can be extended to account for missing observations in the following way,

$$\begin{aligned} \alpha_1^*(i) &= \pi_i \\ (\forall t \le k) \quad \alpha_t^*(i) &= b_i(o_t) \sum_j \alpha_{t-1}^*(j) a_{ij} \\ (\forall t > k) \quad \alpha_t^*(i) &= \sum_j \alpha_{t-1}^*(j) a_{ij} \\ P(S_t = i \mid O_1, \ldots, O_k) &= \frac{\alpha_t^*(i)}{\sum_i \alpha_t^*(i)} \end{aligned}$$

(f) **[10 pts]** Programming problem. The definition of δ from lecture can be extended to account for missing observations in the following way,

$$\begin{aligned} \delta_1^*(i) &= \pi_i \\ (\forall t \le k) \quad \delta_t^*(i) &= \max_j b_i(o_t) \delta_{t-1}^*(j) a_{ij} \\ (\forall t > k) \quad \delta_t^*(i) &= \max_j \delta_{t-1}^*(j) a_{ij} \\ S_t^* &= \arg\max_j \delta_t^*(j) \end{aligned}$$