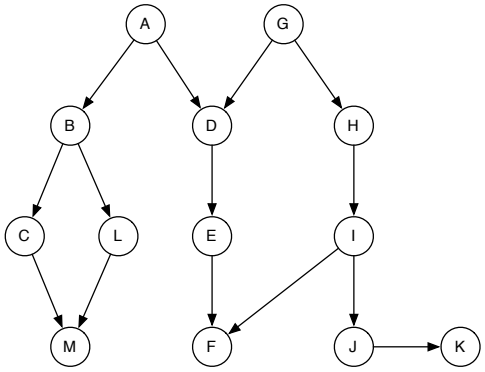
10-701/15-781 Machine Learning: Assignment 4

Released: Nov 29. Revised: Dec 6

- The assignment is due **December 8, 2005** at the beginning of class.
- Write your name in the top right-hand corner of each page submitted. No paperclips, folders, etc.
- If you have any questions, email `questions-10701@autonlab.org`.
- This assignment consists of five questions totalling 100 points.
- Each student must hand in an writeup. See the web page for the collaboration policy.

Q1 Independence [15 pts]

1. Which of the following statements are true with respect to the following graphical model, regardless of the conditional probability distributions ?



(a) $P(A, G|F) = P(A|F)P(G|F)$
(b) $P(B, F|E) = P(B|E)P(F|E)$
(c) $P(B, M|C, L) = P(B|C, L)P(M|C, L)$
(d) $P(G, K|F, I) = P(G|I)P(K|I)$
(e) $P(D, I|G) = P(D|G)P(I|G)$
(f) $P(D, I|G, F) = P(D|G, F)P(I|G, F)$
(g) $P(B, D, H|A, E) = P(B, D|A, E)P(H|A, E)$

Solution. *If X and Y are d-separated by Z then $P(X, Y|Z) = P(X|Z)P(Y|Z)$.*

$P(A, G|F) = P(A|F)P(G|F)$ – *FALSE (active trail A-D-G)*

$P(B, F|E) = P(B|E)P(F|E)$ – *FALSE (active trail B-A-D-G-H-I-F)*

$P(B, M|C, L) = P(B|C, L)P(M|C, L)$ – *TRUE*

$P(G, K|F, I) = P(G|I)P(K|I)$ – *FALSE. We know that G and K are d-separated given F,I so* $P(G, K|F, I) = P(G|F, I)P(K|F, I)$. *Additionally, K and F are d-separated given I so* $P(K|F, I) = P(K|I)$. *But G and F are not d-separated given I (there is an active trail G-D-E-F). Therefore, in general,* $P(G|F, I) \neq P(G|I)$.

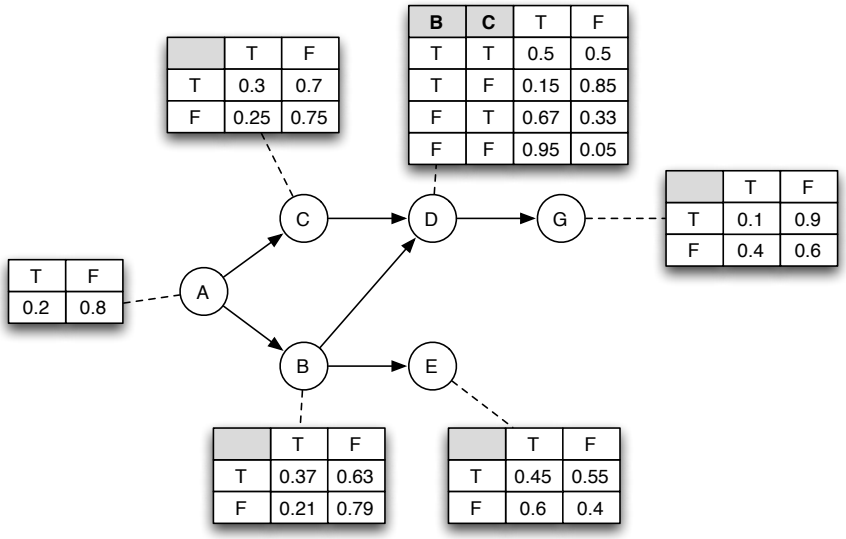
$P(D, I|G) = P(D|G)P(I|G)$ – *TRUE*

$P(D, I|G, F) = P(D|G, F)P(I|G, F)$ - *FALSE (active trail D-E-F-I)*

$P(B, D, H|A, E) = P(B, D|A, E)P(H|A, E)$ - *FALSE (active trail D-G-H)*

Q2 Inference [15 pts]

Compute the distribution $P(B|D = T)$ on the following Bayes network. Show your work.



Solution.

$$
\begin{aligned}
P(B=k, D=T) &= \sum_A \sum_C \sum_E \sum_G P(A, B=k, C, D=T, E, G) \\
&= \sum_A \sum_C \sum_E \sum_G P(A)P(C|A)P(B=k|A)P(D=T|B=k, C)P(E|B=k)P(G|D=T) \\
&= \sum_A P(A)P(B=k|A) \sum_C P(C|A)P(D=T|B=k, C) \sum_E P(E|B=k) \sum_G P(G|D=T) \\
&= \sum_A P(A)P(B=k|A) \sum_C P(C|A)P(D=T|B=k, C) \sum_E P(E|B=k) \\
&= \sum_A P(A)P(B=k|A) \sum_C P(C|A)P(D=T|B=k, C)
\end{aligned}
$$

$P(B=T, D=T) = 0.0588$

$P(B=F, D=T) = 0.6653$

Thus $P(B=T|D=T) = 0.0812, P(B=F|D=T) = 0.9188$.

Q3 Structure Learning [5 pts]

Explain why structure scoring metrics are typically decomposable, *i.e.*

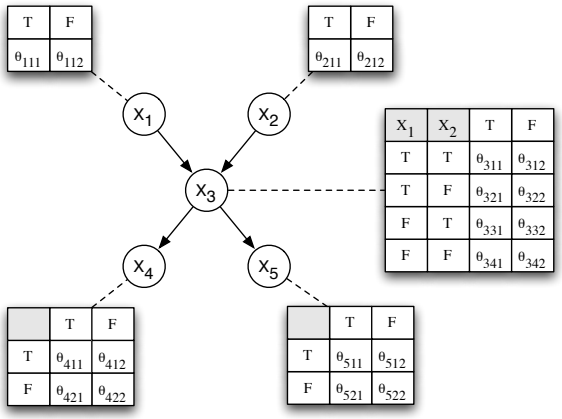
$$\text{Score}(S) = \sum_{i=1}^{n} \text{NodeScore}(X_i|\text{Parents}(X_i))$$

Your explanation should not be more than 4-5 sentences long.

Solution. *Structure learning typically involves greedy search over the space of structures, where operations include adding, deleting, and reversing a single arc. If the scoring metric is decomposable, then these changes involve recomputing at most two NodeScore terms (which require computing a few sufficient statistics). If the scoring metric was not decomposable, recomputing the score would require computing many more sufficient statistics.*

Q4 Parameter Estimation with Missing Values [65 pts]

Consider a Bayesian network with the following structure:



In class we covered how to learn the parameters of the network from complete data, where the values of all the attributes are specified for each record. Here, we want to learn the parameters of the network from incomplete data, where some of the records have missing values. Unlike learning with latent variables, the value of a variable may be known for some records but not for others.

A Bayesian network on discrete variables represents a multinomial distribution on $X_1, \ldots, X_n$. $\text{pa}(X_i)$ represents the variables that are parents of X_i in the directed acyclic graph. X_i takes on r_i values and $\text{pa}(X_i)$ takes on q_i values. The parameters of the network $p(X_i = k|\text{pa}(X_i) = j)$ are denoted θ_{ijk}. The entire set of network parameters are denoted θ.

A data set containing R iid records is denoted $\mathbf{x} = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^R\}$. $\mathbf{x}_i^m$ denotes the value of X_i in the m^{th} record. N_{ijk} is a count of how many records have $X_i = k$ and $\text{pa}(X_i) = j$. If we use $\delta(\cdot)$ to represent the indicator function

$$N_{ijk} = \sum_{m=1}^{R} \delta(\mathbf{x}_i^m = k, \mathbf{x}_{\text{pa}(X_i)}^m = j)$$
$$N_{ij} \equiv \sum_k N_{ijk}$$

If the records are complete then the log-likelihood of the data (ignoring the normalization constant) is

$$\ell(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ijk} \log \theta_{ijk} \quad \text{where} \quad \sum_{k=1}^{r_i} \theta_{ijk} = 1$$

If the records are not complete then let $\mathbf{x}_{\text{obs}}^m$ and $\mathbf{x}_{\text{hid}}^m$ denote the observed and unobserved (hidden) part of the m^{th} record.

1. Why is N_{ijk} a latent variable when we deal with incomplete data ? Your explanation must not exceed 2 sentences.

   Solution. *When the data is incomplete we do not know the exact value of N_{ijk}, and so it is latent*[1]

2. Prove that $E[N_{ijk}|\mathbf{x}, \theta] = \sum_{m=1}^{R} P(\mathbf{x}_i^m = k, \mathbf{x}_{\text{pa}(X_i)}^m = j|\mathbf{x}_{\text{obs}}^m, \theta)$.

   Solution.

$$\begin{aligned} E[N_{ijk}|\mathbf{x}, \theta] &= E[\sum_{m=1}^{R} \delta(\mathbf{x}_i^m = k, \mathbf{x}_{pa(X_i)}^m = j)|\mathbf{x}^1, \ldots, \mathbf{x}^R, \theta] \\ &= \sum_{m=1}^{R} E[\delta(\mathbf{x}_i^m = k, \mathbf{x}_{pa(X_i)}^m = j)|\mathbf{x}_{obs}^m, \theta] \\ &= \sum_{m=1}^{R} P(\mathbf{x}_i^m = k, \mathbf{x}_{pa(X_i)}^m = j|\mathbf{x}_{obs}^m, \theta) \end{aligned}$$

3. (E-Step) Prove that the expected complete log-likelihood

$$Q(\theta|\theta^{(t)}) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} E[N_{ijk}|\mathbf{x}, \theta^{(t)}] \log \theta_{ijk}$$

   Solution. *Since we know the complete log-likelihood (to within a normalization constant) is*

$$\ell(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ijk} \log \theta_{ijk}$$

   it follows immediately by noting that $Q(\theta|\theta^{(t)}) = E_{N_{ijk}|\mathbf{x}, \theta^{(t)}}[\ell(\theta)]$.

[1] For a given data set, it is possible that some N_{ijk} are known precisely.

4. (M-Step) Prove that the expected complete log-likelihood is maximized when

$$\theta_{ijk}^{(t+1)} = \frac{E[N_{ijk}|\mathbf{x}, \theta^{(t)}]}{\sum_k E[N_{ijk}|\mathbf{x}, \theta^{(t)}]}$$

Solution. *Let $E[N_ijk]$ be shorthand for $E[N_ijk|\mathbf{x}, \theta^{(t)}]$. Using Lagrange multipliers*

$$\begin{aligned}
\Lambda(\theta_{ij1}, \ldots, \theta_{ijk}, \lambda) &= \sum_{ijk} E[N_{ijk}] \log \theta_{ijk} - \lambda \left(\sum_k \theta_{ijk} - 1 \right) \\
0 = \frac{\partial \Lambda}{\partial \theta_{ijk}} &= \frac{E[N_{ijk}]}{\theta_{ijk}} - \lambda \\
E[N_{ijk}] &= \lambda \theta_{ijk} \quad \forall k \quad \textit{(sum the equations)} \\
\sum_k E[N_{ijk}] &= \lambda \sum_k \theta_{ijk} \implies \lambda = \sum_k E[N_{ijk}]
\end{aligned}$$

5. If $\mathbf{x}^m$ has no missing values, write down pseudocode for an algorithm that returns $P(\mathbf{x}_i^m = k, \mathbf{x}_{\text{pa}(X_i)}^m = j|\mathbf{x}_{\text{obs}}^m, \theta)$ in time polynomial in n and the number of parameters $|\theta|$.

Solution. *If we condition on a record with no missing values, then we know with certainty whether an event (in this case the values of a node and its parents) occurred. Another way of saying this is that all of the variables are observed, so inference reduces to checking the value of the observations.*

LIKELIHOOD-NO-MISSING($\mathbf{x}^m$)
1 ***if*** $(\mathbf{x}_i^m = k)$ *and* $(\mathbf{x}_{pa(X_i)}^m = j)$
2 ***then return*** *1*
3 ***else return*** *0*

6. For the 5-node Bayesian network given above, write down the formulae for $P(x_i^m = k, \mathbf{x}_{\text{pa}(X_i)}^m = j|\mathbf{x}_{\text{obs}}^m, \theta)$ when exactly one value is missing from $\mathbf{x}^m$.

Solution. *We show only the formulae that are affected by the missing value. The rest are computed as in part 5.*

$$\begin{aligned}
P(X_1 = k|x_2^m, x_3^m, x_4^m, x_5^m) &= \frac{P(X_1 = k, x_2^m, x_3^m, x_4^m, x_5^m)}{\sum_{k=1}^2 P(X_1 = k, x_2^m, x_3^m, x_4^m, x_5^m)} \\
P(x_3^m, X_1 = k, x_2^m|x_2^m, x_3^m, x_4^m, x_5^m) &= \delta(X_3 = x_3^m, X_2 = x_2^m) P(X_1 = k|x_2^m, x_3^m, x_4^m, x_5^m) \\
P(X_2 = k|x_1^m, x_3^m, x_4^m, x_5^m) &= \frac{P(X_2 = k, x_1^m, x_3^m, x_4^m, x_5^m)}{\sum_{k=1}^2 P(X_2 = k, x_1^m, x_3^m, x_4^m, x_5^m)} \\
P(x_3^m, x_1^m, X_2 = k|x_1^m, x_3^m, x_4^m, x_5^m) &= \delta(X_3 = x_3^m, X_1 = x_1^m) P(X_2 = k|x_1^m, x_3^m, x_4^m, x_5^m) \\
P(X_3 = k|x_1^m, x_2^m, x_4^m, x_5^m) &= \frac{P(X_3 = k, x_1^m, x_2^m, x_4^m, x_5^m)}{\sum_{k=1}^2 P(X_3 = k, x_1^m, x_2^m, x_4^m, x_5^m)} \\
P(X_3 = k, x_1^m, x_2^m|x_1^m, x_2^m, x_4^m, x_5^m) &= \delta(X_1 = x_1^m, X_2 = x_2^m) P(X_3 = k|x_1^m, x_2^m, x_4^m, x_5^m) \\
P(x_4^m, X_3 = k|x_1^m, x_2^m, x_4^m, x_5^m) &= \delta(X_4 = x_4^m) P(X_3 = k|x_1^m, x_2^m, x_4^m, x_5^m) \\
P(x_5^m, X_3 = k|x_1^m, x_2^m, x_4^m, x_5^m) &= \delta(X_5 = x_5^m) P(X_3 = k|x_1^m, x_2^m, x_4^m, x_5^m) \\
P(X_4 = k|x_1^m, x_2^m, x_3^m, x_5^m) &= \frac{P(X_4 = k, x_1^m, x_2^m, x_3^m, x_5^m)}{\sum_{k=1}^2 P(X_4 = k, x_1^m, x_2^m, x_3^m, x_5^m)} \\
P(X_4 = k, x_3^m|x_1^m, x_2^m, x_3^m, x_5^m) &= \delta(X_3 = x_3^m) P(X_4 = k|x_1^m, x_2^m, x_3^m, x_5^m)
\end{aligned}$$

$$P(X_5 = k|x_1^m, x_2^m, x_3^m, x_4^m) = \frac{P(X_5 = k, x_1^m, x_2^m, x_3^m, x_4^m)}{\sum_{k=1}^{2} P(X_5 = k, x_1^m, x_2^m, x_3^m, x_4^m)}$$
$$P(X_5 = k, x_3^m|x_1^m, x_2^m, x_3^m, x_4^m) = \delta(X_3 = x_3^m)P(X_5 = k|x_1^m, x_2^m, x_3^m, x_4^m)$$

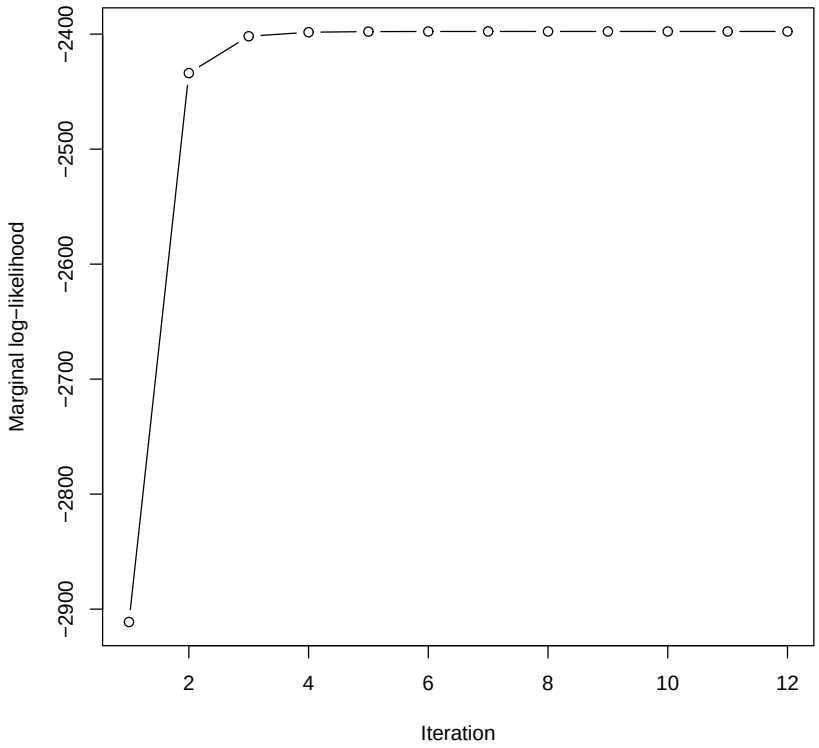
The nice thing about knowing there is only one value missing is that the inference on each record reduces to computing two likelihoods, which can be done in polynomial time.

7. The data set `missing.csv` contains data on $X_1, \ldots, X_5$ where each record has at most one missing value. Implement the EM algorithm for estimating θ. Use a uniform starting configuration for θ, *i.e.*, $\theta_{ijk}^{(0)} = 1/r_i$. Run until $\max_{ijk} |\theta_{ijk}^{(t+1)} - \theta_{ijk}^{(t)}| < 10^{-4}$. Plot the marginal log-likelihood $\ell(\theta) = \sum_{m=1}^{R} \log p(\mathbf{x}_{obs}^m|\theta)$ vs. the number of iterations. What is the final estimate for θ ?

Note: `missing.csv` contains 1000 records where '0' corresponds to false, '1' corresponds to true, and 'NA' corresponds to a missing value.

Solution. *The maximum likelihood estimate is* $(\theta_{111} = 0.1549933, \theta_{112} = 0.8450067, \theta_{211} = 0.5929645, \theta_{212} = 0.4070355, \theta_{311} = 0.2689665, \theta_{312} = 0.7310335, \theta_{321} = 0.6117814, \theta_{322} = 0.3882186, \theta_{331} = 0.8218331, \theta_{332} = 0.1781689, \theta_{341} = 0.2922545, \theta_{342} = 0.7077455, \theta_{411} = 0.3141978, \theta_{412} = 0.6858022, \theta_{421} = 0.09749617, \theta_{422} = 0.9025038, \theta_{511} = 0.5828414, \theta_{512} = 0.4171586, \theta_{521} = 0.5508529, \theta_{522} = 0.4491471)$.

The network the data was generated from had parameters $(\theta_{111} = 0.15, \theta_{112} = 0.85, \theta_{211} = 0.6, \theta_{212} = 0.4, \theta_{311} = 0.25, \theta_{312} = 0.75, \theta_{321} = 0.5, \theta_{322} = 0.5, \theta_{331} = 0.8, \theta_{332} = 0.2, \theta_{341} = 0.3, \theta_{342} = 0.7, \theta_{411} = 0.3, \theta_{412} = 0.7, \theta_{421} = 0.1, \theta_{422} = 0.9, \theta_{511} = 0.6, \theta_{512} = 0.4, \theta_{521} = 0.55, \theta_{522} = 0.45)$.



EM converges in 12 iterations with a final marginal likelihood of -2397.669.