CS 194-10, Fall 2011
Assignment 7 Solutions

1. *Markov blanket*

   (a) There are several ways to prove this. Probably the simplest is to work directly from the global semantics. First, we rewrite the required probability in terms of the full joint:

$$\begin{aligned} P(x_i|x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n) &= \frac{P(x_1,\ldots,x_n)}{P(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)} \\ &= \frac{P(x_1,\ldots,x_n)}{\sum_{x_i} P(x_1,\ldots,x_n)} \\ &= \frac{\prod_{j=1}^{n} P(x_j|parentsX_j)}{\sum_{x_i}\prod_{j=1}^{n} P(x_j|parentsX_j)} \end{aligned}$$

   Now, all terms in the product in the denominator that do not contain x_i can be moved outside the summation, and then cancel with the corresponding terms in the numerator. This just leaves us with the terms that do mention x_i, i.e., those in which X_i is a child or a parent. Hence, $P(x_i|x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)$ is equal to

$$\frac{P(x_i|parentsX_i)\prod_{Y_j\in Children(X_i)} P(y_j|parents(Y_j))}{\sum_{x_i} P(x_i|parentsX_i)\prod_{Y_j\in Children(X_i)} P(y_j|parents(Y_j))}$$

   Now, by reversing the argument in part (b), we obtain the desired result.

   (b) This is a relatively straightforward application of Bayes' rule. Let $\mathbf{Y} = Y_1,\ldots,y_\ell$ be the children of X_i and let $\mathbf{Z}_j$ be the parents of Y_j other than X_i. Then we have

$$\begin{aligned} &\mathbf{P}(X_i|MB(X_i)) \\ &= \mathbf{P}(X_i|Parents(X_i),\mathbf{Y},\mathbf{Z}_1,\ldots,\mathbf{Z}_\ell) \\ &= \alpha\mathbf{P}(X_i|Parents(X_i),\mathbf{Z}_1,\ldots,\mathbf{Z}_\ell)\mathbf{P}(\mathbf{Y}|Parents(X_i),X_i,\mathbf{Z}_1,\ldots,\mathbf{Z}_\ell) \\ &= \alpha\mathbf{P}(X_i|Parents(X_i))\mathbf{P}(\mathbf{Y}|X_i,\mathbf{Z}_1,\ldots,\mathbf{Z}_\ell) \\ &= \alpha\mathbf{P}(X_i|Parents(X_i)) \prod_{Y_j\in Children(X_i)} P(Y_j|Parents(Y_j)) \end{aligned}$$

   where the derivation of the third line from the second relies on the fact that a node is independent of its nondescendants given its children.

2. *Exponential Family* (15)
   This question deals with exponential family models, which have the form

$$p(x) = h(x)e^{\theta^T T(x) - A(\theta)} \ .$$

   (a) Fitting various models into the family:

      i. For Normal$(\mu, 1)$, we have

$$p(x) = h(x)e^{\theta^T T(x) - A(\theta)} = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} e^{\mu x - \frac{\mu^2}{2}}$$

      which fits the exponential family with parameter vector $\theta = (\mu)$, $h(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, $T(x) = (x)$, and $A(\theta) = \theta^T\theta/2$.

ii. For Bernoulli(p), we restrict x to the range $\{0,1\}$ and use the standard trick to express the Bernoulli as a product of exponentials:

$$p(x) = h(x)e^{\theta^T T(x) - A(\theta)} = p^x(1-p)^{(1-x)} = e^{x\log p + (1-x)\log(1-p)} = e^{x\log\frac{p}{1-p} + \log(1-p)}$$

which fits the exponential family with parameter vector $\theta = \left(\log\frac{p}{1-p}\right)$, $h(x) = 1$, $T(x) = (x)$, and $A(\theta) = \log(1+e^\theta)$.

iii. For Categorical($p_1, \ldots, p_K$), we can extend the product-of-exponentials trick if we define the K-element vector random variable $\mathbf{x}$ such that each x_k is 0 or 1 and exactly one of the x_ks is 1. Then we have

$$p(\mathbf{x}) = \prod_{k=1}^{K} p_k^{x_k}$$

and we define the parameter vector $\theta = (\log p_1, \ldots, \log p_K)^T$. Then

$$p(\mathbf{x}) = \prod_{k=1}^{K} e^{x_k \log p_k} = e^{\sum_{k=1}^{K} x_k \log p_k} = e^{\theta^T \mathbf{x}}$$

which is trivially in the exponential family with $h(\mathbf{x}) = 1$, $T(\mathbf{x}) = \mathbf{x}$, and $A(\theta) = 0$.

(b) The derivative of the log likelihood is obtained as follows:

$$\begin{aligned} P(x_1, \ldots, x_N) &= \prod_{i=1}^{N} h(x_i) e^{\theta^T T(x_i) - A(\theta)} \\ L(x_1, \ldots, x_N) &= \sum_{i=1}^{N} \log h(x_i) + \theta^T T(x_i) - A(\theta) \\ \nabla L &= \left(\sum_{i=1}^{N} T(x_i)\right) - \nabla A(\theta) \ . \end{aligned}$$

Now define $g(\theta) \equiv \nabla A(\theta)$; at the maximum likelihood value of θ, the derivatives are zero, so $g(\theta) = \sum_{i=1}^{N} T(x_i)$ or $\theta = g^{-1}\left(\sum_{i=1}^{N} T(x_i)\right)$, which is a function of the data only through the sufficient statistic T.

(c) From part (a), for the Normal(μ, 1) distribution we have $T(x) = (x)$ and $A(\theta) = \theta^T\theta/2$. hence

$$\frac{\partial}{\partial\theta} A(\theta) = \theta = (\mu)$$

and

$$E(T(x)) = E((x)) = (\mu)$$

so the property is satisfied.

3. *EM with discrete variables* (10)
Define the paramaters to be

$$\begin{aligned} &\theta_X = P(X=1) \\ &\theta_{Y0} = P(Y=1|X=0) \qquad\qquad \theta_{Y1} = P(Y=1|X=1) \\ &\theta_{Z0} = P(Z=1|Y=0) \qquad\qquad \theta_{Z1} = P(Z=1|Y=1) \end{aligned}$$

(a) The "observable counts" are those for which both parent and child are observed. For example, if X and Y are observed, that gives a count for θ_{Y0} or θ_{Y1}, even if Z is unobserved. So the ML

values of the parameters are

$$\theta_X^{(0)} = 0.5$$
$$\theta_{Y0}^{(0)} = \frac{1}{3} \qquad \theta_{Y1}^{(0)} = 0.5$$
$$\theta_{Z0}^{(0)} = 0.5 \qquad \theta_{Z1}^{(0)} = 0.5$$

(b) To apply EM in this case we recall that the general EM formulation is

$$\theta^{(i+1)} = \operatorname*{argmax}_{\theta} \sum_{z} P(Z = z|\mathbf{x}, \theta^{(i)}) L(\mathbf{x}, Z = z|\theta)$$

where in this formula Z are the hidden variables and L is the log-likelihood. We should therefore proceed by computing:

$$P(z_2 = 0, y_5 = 0|\mathbf{x}, \theta^{(i)}) = (1 - \theta_{Z0}^{(i)}) \cdot \frac{(1 - \theta_{Y1}^{(i)}) \cdot \theta_{Z0}^{(i)}}{\theta_{Y1}^{(i)} \cdot \theta_{Z1}^{(i)} + (1 - \theta_{Y1}^{(i)}) \cdot \theta_{Z0}^{(i)}}$$

and similarly

$$P(z_2 = 0, y_5 = 1|\mathbf{x}, \theta^{(i)}) = (1 - \theta_{Z0}^{(i)}) \cdot \frac{\theta_{Y1}^{(i)} \cdot \theta_{Z1}^{(i)}}{\theta_{Y1}^{(i)} \cdot \theta_{Z1}^{(i)} + (1 - \theta_{Y1}^{(i)}) \cdot \theta_{Z0}^{(i)}}$$
$$P(z_2 = 1, y_5 = 0|\mathbf{x}, \theta^{(i)}) = \theta_{Z0}^{(i)} \cdot \frac{(1 - \theta_{Y1}^{(i)}) \cdot \theta_{Z0}^{(i)}}{\theta_{Y1}^{(i)} \cdot \theta_{Z1}^{(i)} + (1 - \theta_{Y1}^{(i)}) \cdot \theta_{Z0}^{(i)}}$$
$$P(z_2 = 1, y_5 = 1|\mathbf{x}, \theta^{(i)}) = \theta_{Z0}^{(i)} \cdot \frac{\theta_{Y1}^{(i)} \cdot \theta_{Z1}^{(i)}}{\theta_{Y1}^{(i)} \cdot \theta_{Z1}^{(i)} + (1 - \theta_{Y1}^{(i)}) \cdot \theta_{Z0}^{(i)}} \ .$$

In particular,

$$P(z_2 = 0, y_5 = 0|\mathbf{x}, \theta^{(0)}) = 0.5 \cdot \frac{0.5 \cdot 0.5}{0.5 \cdot 0.5 + 0.5 \cdot 0.5} = 0.25$$

and similarly

$$P(z_2 = 0, y_5 = 1|\mathbf{x}, \theta^{(0)}) = 0.25$$
$$P(z_2 = 1, y_5 = 0|\mathbf{x}, \theta^{(0)}) = 0.25$$
$$P(z_2 = 1, y_5 = 1|\mathbf{x}, \theta^{(0)}) = 0.25$$

Next, we compute

$$\begin{aligned} L(\mathbf{x}, z_2 = 0, y_5 = 0|\theta) = & 3\log(\theta_X) + 3\log(1 - \theta_X) + \\ & \log(\theta_{Y0}) + 2\log(1 - \theta_{Y0}) + \log(\theta_{Y1}) + 2\log(1 - \theta_{Y1}) + \\ & 2\log(\theta_{Z0}) + 2\log(1 - \theta_{Z0}) + \log(\theta_{Z1}) + \log(1 - \theta_{Z1}) \end{aligned}$$

and similarly

$$\begin{aligned}
L(\mathbf{x}, z_2 = 0, y_5 = 1|\theta) =& 3\log(\theta_X) + 3\log(1-\theta_X) + \\
& \log(\theta_{Y0}) + 2\log(1-\theta_{Y0}) + 2\log(\theta_{Y1}) + \log(1-\theta_{Y1}) + \\
& \log(\theta_{Z0}) + 2\log(1-\theta_{Z0}) + 2\log(\theta_{Z1}) + \log(1-\theta_{Z1}) \\
L(\mathbf{x}, z_2 = 1, y_5 = 0|\theta) =& 3\log(\theta_X) + 3\log(1-\theta_X) + \\
& \log(\theta_{Y0}) + 2\log(1-\theta_{Y0}) + \log(\theta_{Y1}) + 2\log(1-\theta_{Y1}) + \\
& 3\log(\theta_{Z0}) + \log(1-\theta_{Z0}) + \log(\theta_{Z1}) + \log(1-\theta_{Z1}) \\
L(\mathbf{x}, z_2 = 1, y_5 = 1|\theta) =& 3\log(\theta_X) + 3\log(1-\theta_X) + \\
& \log(\theta_{Y0}) + 2\log(1-\theta_{Y0}) + 2\log(\theta_{Y1}) + \log(1-\theta_{Y1}) + \\
& 2\log(\theta_{Z0}) + \log(1-\theta_{Z0}) + 2\log(\theta_{Z1}) + \log(1-\theta_{Z1})
\end{aligned}$$

Plugging in we have

$$\begin{aligned}
&\sum_z P(Z = z|\mathbf{x}, \theta^{(i)}) L(\mathbf{x}, Z = z|\theta) = \\
&\frac{3+3+3+3}{4}\log(\theta_X) + \frac{3+3+3+3}{4}\log(1-\theta_X) + \\
&\frac{1+1+1+1}{4}\log(\theta_{Y0}) + \frac{2+2+2+2}{4}\log(1-\theta_{Y0}) + \frac{1+2+1+2}{4}\log(\theta_{Y1}) + \frac{2+1+2+1}{4}\log(1-\theta_{Y1}) + \\
&\frac{2+1+3+2}{4}\log(\theta_{Z0}) + \frac{2+2+1+1}{4}\log(1-\theta_{Z0}) + \frac{1+2+1+2}{4}\log(\theta_{Z1}) + \frac{1+1+1+1}{4}\log(1-\theta_{Z1})
\end{aligned}$$

which is maximized by

$$\begin{aligned}
\theta_X^{(0)} &= 0.5 \\
\theta_{Y0}^{(0)} &= \frac{1}{3} & \theta_{Y1}^{(0)} &= 0.5 \\
\theta_{Z0}^{(0)} &= \frac{4}{7} & \theta_{Z1}^{(0)} &= 0.6
\end{aligned}$$

We iteratively compute the probabilities and maximizing parameters and after about 10 iterations it converges (up to the 4th digit), to

$$\begin{aligned}
\theta_X^{(0)} &= 0.5 \\
\theta_{Y0}^{(0)} &= \frac{1}{3} & \theta_{Y1}^{(0)} &= 0.5053 \\
\theta_{Z0}^{(0)} &= 0.5983 & \theta_{Z1}^{(0)} &= 0.6151
\end{aligned}$$

(c) In this case it takes infinitely many iterations to converge exactly. There are cases where convergence is immediate and exact—for example, if the missing data is only on Z.

4. *Learning with continuous variables* (15)

(a) The log-likelihood is

$$\begin{aligned}
L &= \sum_{i=1}^{N} \log P(x_i) + \log P(\mathbf{y}_i \mid x_i) + \log P(z_i \mid \mathbf{y}_i) \\
&= \sum_{i=1}^{N} \log 0.5 + \sum_{j=1}^{D} -\log\sqrt{2\pi} - \log\sigma - (y_{ij} - \mu_{jx_i})^2/2\sigma^2 - \log(1 + exp(-z_i \sum_j w_j y_{ij}))
\end{aligned}$$

and the partial derivatives are

$$\begin{aligned}
\frac{\partial L}{\partial \mu_{jk}} &= \sum_{\{i:x_i=k\}} (y_{ij} - \mu_{jk})/\sigma^2 \\
\frac{\partial L}{\partial \sigma} &= \sum_{i=1}^{N} \sum_{j=1}^{D} -1/\sigma + (y_{ij} - \mu_{jx_i})^2/\sigma^3 \\
\frac{\partial L}{\partial w_j} &= \sum_{i=1}^{N} \frac{z_i y_{ij} exp(-z_i \sum_j w_j y_{ij})}{1 + exp(-z_i \sum_j w_j y_{ij})}
\end{aligned}$$

(b) EM is doing something familiar in each case:

i. If X and $\mathbf{Y}$ are observed and Z is not, then Z is irrelevant and EM is solving an observable naive Bayes model with a Boolean class and D continuous attributes with Gaussian class-conditional distributions.

ii. If Z and $\mathbf{Y}$ are observed and X is not, then EM is solving two problems: the first is fitting a mixture of two spherical Gaussians with unknown mean and variance, while the second is an observable logistic regression problem.

iii. If $\mathbf{Y}$ is observed and X and Z are not, then Z is irrelevant and EM is fitting a mixture of two spherical Gaussians with unknown mean and variance.