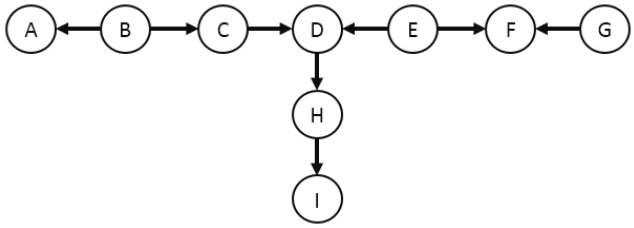
EE 583 Probabilistic Graphical Models

Homework 2

Due on Friday, March 25, 2016

1. **[9 pts]** According to the Bayesian Network shown below, list the cases, in which variables A and G become independent.



Solution:

A and G are independent if

a) B is observed
b) C is observed
c) E is observed
d) F and at least one of (D,H,I) are not simultaneously observed

2. **[20 pts]** Let G be a BN structure and P be a distribution over random variables $\boldsymbol{X} = \{X_1, \dots, X_n\}$. Show that if P factorizes according to G, then $\left(X_i \perp \boldsymbol{X}_{\mathrm{NonDes}(i)} \middle| \boldsymbol{X}_{\mathrm{Pa}(i)}\right) \in \mathrm{I}(P)$ for all X_i.
Hint: Write $P(\boldsymbol{X}) = P\left(X_i, \boldsymbol{X}_{\mathrm{Pa}(i)}, \boldsymbol{X}_{\mathrm{NonDes}(i)}, \boldsymbol{X}_{\mathrm{Des}(i)}\right)$ and using factorization of the RHS, show $P\left(X_i \middle| \boldsymbol{X}_{\mathrm{Pa}(i)}, \boldsymbol{X}_{\mathrm{NonDes}(i)}\right) = P\left(X_i \middle| \boldsymbol{X}_{\mathrm{Pa}(i)}\right)$.

Solution:

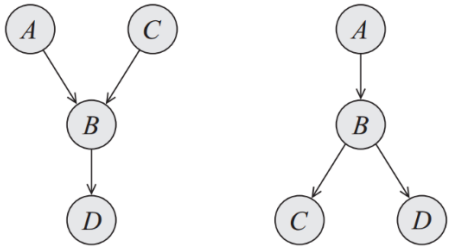
$$P\left(x_i \middle| \boldsymbol{x}_{\mathrm{Pa}(i)}, \boldsymbol{x}_{\mathrm{NonDes}(i)}\right) = \frac{\sum_{\boldsymbol{x}_{\mathrm{Des}(i)}} P\left(x_i, \boldsymbol{x}_{\mathrm{Pa}(i)}, \boldsymbol{x}_{\mathrm{NonDes}(i)}, \boldsymbol{x}_{\mathrm{Des}(i)}\right)}{\sum_{x_i}\sum_{\boldsymbol{x}_{\mathrm{Des}(i)}} P\left(x_i, \boldsymbol{x}_{\mathrm{Pa}(i)}, \boldsymbol{x}_{\mathrm{NonDes}(i)}, \boldsymbol{x}_{\mathrm{Des}(i)}\right)}$$

$$= \frac{\sum_{\boldsymbol{x}_{\mathrm{Des}(i)}} P\left(x_i \middle| \boldsymbol{x}_{\mathrm{Pa}(i)}\right) \prod_{j\in\mathrm{NonDes}(i)} P\left(x_j \middle| \boldsymbol{x}_{\mathrm{Pa}(j)}\right) \prod_{j\in\mathrm{Des}(i)} P\left(x_j \middle| \boldsymbol{x}_{\mathrm{Pa}(j)}\right)}{\sum_{x_i}\sum_{\boldsymbol{x}_{\mathrm{Des}(i)}} P\left(x_i \middle| \boldsymbol{x}_{\mathrm{Pa}(i)}\right) \prod_{j\in\mathrm{NonDes}(i)} P\left(x_j \middle| \boldsymbol{x}_{\mathrm{Pa}(j)}\right) \prod_{j\in\mathrm{Des}(i)} P\left(x_j \middle| \boldsymbol{x}_{\mathrm{Pa}(j)}\right)}$$

$$= \frac{\cancel{\prod_{j\in\mathrm{NonDes}(i)} P\left(x_j \middle| \boldsymbol{x}_{\mathrm{Pa}(j)}\right)} P\left(x_i \middle| \boldsymbol{x}_{\mathrm{Pa}(i)}\right) \boxed{\sum_{\boldsymbol{x}_{\mathrm{Des}(i)}} \prod_{j\in\mathrm{Des}(i)} P\left(x_j \middle| \boldsymbol{x}_{\mathrm{Pa}(j)}\right)}}{\cancel{\prod_{j\in\mathrm{NonDes}(i)} P\left(x_j \middle| \boldsymbol{x}_{\mathrm{Pa}(j)}\right)} \sum_{x_i} P\left(x_i \middle| \boldsymbol{x}_{\mathrm{Pa}(i)}\right) \boxed{\sum_{\boldsymbol{x}_{\mathrm{Des}(i)}} \prod_{j\in\mathrm{Des}(i)} P\left(x_j \middle| \boldsymbol{x}_{\mathrm{Pa}(j)}\right)}} \to 1$$

$$= \frac{P\left(x_i \middle| \boldsymbol{x}_{\mathrm{Pa}(i)}\right)}{\sum_{x_i} P\left(x_i \middle| \boldsymbol{x}_{\mathrm{Pa}(i)}\right)} = P\left(x_i \middle| \boldsymbol{x}_{\mathrm{Pa}(i)}\right)$$

3. **[8 pts]** Consider the two networks.



(a) (b)

For each of them, determine whether there can be any other Bayesian network that is I-equivalent to it.

Solution:

a) No, since assigning different directions to any edge in the skeleton will either make $A \rightarrow B \leftarrow C$ no longer an immorality, or introduce new ones $A \rightarrow B \leftarrow D$, and $C \rightarrow B \leftarrow D$.

b) Yes, since there is no immorality. Designating any node as the root in the skeleton, and directing edges outward will result in an I-equivalent Bayesian network.

4. **[20 pts]** Show that global Markov property of Bayesian networks imply the local Markov property. In other words, show that in a Bayesian network, each variable is d-separated from its non-descendants by its parents.

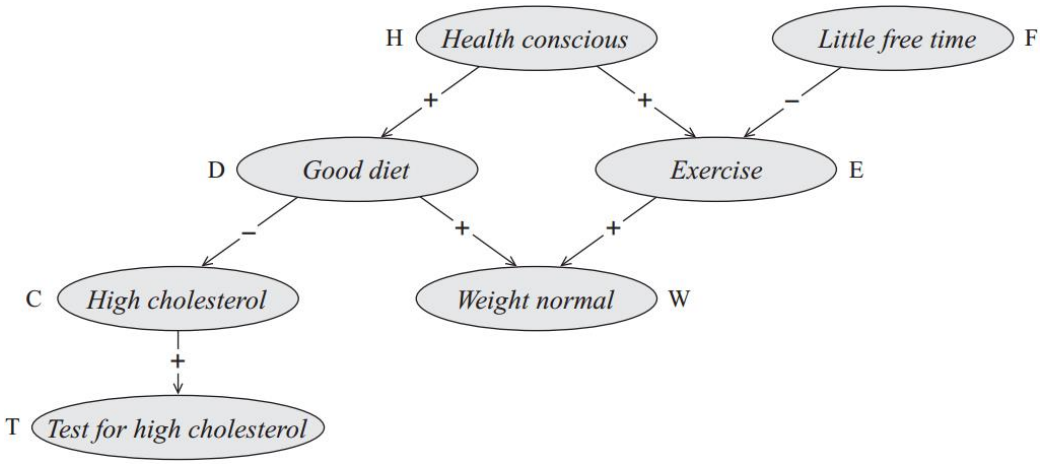
Solution:

Suppose Y is a non-descendant of X and $Y \notin \mathrm{Pa}(X)$. Let T be a trail between Y and X. We analyze two cases:

a) ***T* contains a parent of** X: Call that parent Z. Since $Z \rightarrow X$, Z cannot be a v-junction along T. Observing $\mathrm{Pa}(X)$ makes Z observed, and therefore $\mathrm{Pa}(X)$ blocks T.

b) ***T* contains no parent of** X: Then T must contain at least one v-junction, because otherwise Y would be a descendant of X. To complete the proof, we have to show that the statement "all v-junctions along T contain a parent of X among their descendants" connot be true, because if it were true observing $\mathrm{Pa}(X)$ would activate T. Suppose W is the first v-junction we encounter when tracing edge directions on T starting from X. Clearly, we must have a path $X \rightarrow \cdots \rightarrow W$, because reversing the first edge from X would introduce a parent of X to T, and reversing other edge directions would introduce another v-junction before W. Then, if there is a parent $Z \in \mathrm{Pa}(X)$ among W's descendants, then we must have a cycle $X \rightarrow \cdots \rightarrow W \rightarrow \cdots \rightarrow Z \rightarrow X$ violating the acyclicity of the Bayesian network, hence completing the proof.

5. **[18 pts]** Consider the Bayesian network shown below. Assume all variables are binary valued. Suppose the signs at edges indicate how each random variable affects its child qualitatively, that is, $X \overset{+}{\rightarrow} Y$ indicates $P(Y = 1|X = 1, \boldsymbol{U} = \boldsymbol{u}) > P(Y = 1|X = 0, \boldsymbol{U} = \boldsymbol{u})$ for all values of $\boldsymbol{u}$ of Y's other parents (accordingly, the inequality reverses when $X \overset{-}{\rightarrow} Y$).



For each pair of the probabilities listed in the rows of the table below, identify which one is larger than the other, or if they are equal, or if they are incomparable with the given edge-sign information alone.

(a)	$P(T = 1\|D = 1)$	$P(T = 1)$
(b)	$P(D = 1\|T = 0)$	$P(D = 1)$
(c)	$P(H = 1\|E = 1, F = 1)$	$P(H = 1\|E = 1)$
(d)	$P(C = 1\|F = 0)$	$P(C = 1)$
(e)	$P(C = 1\|H = 0)$	$P(C = 1)$
(f)	$P(C = 1\|H = 0, F = 0)$	$P(C = 1\|H = 0)$
(g)	$P(D = 1\|H = 1, E = 0)$	$P(D = 1\|H = 1)$
(h)	$P(D = 1\|E = 1, F = 0, W = 1)$	$P(D = 1\|E = 1, F = 0)$
(i)	$P(T = 1\|W = 1, F = 0)$	$P(T = 1\|W = 1)$

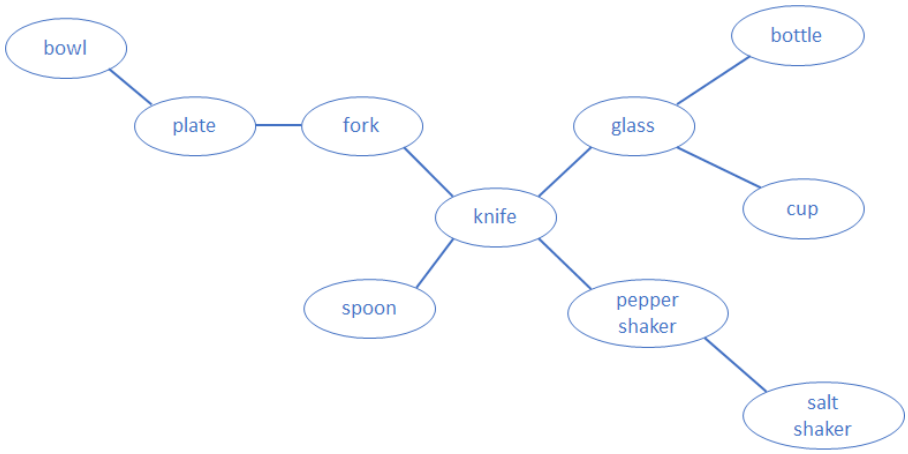
Solution:

(a)	$P(T = 1\|D = 1)$	<	$P(T = 1)$
(b)	$P(D = 1\|T = 0)$	>	$P(D = 1)$
(c)	$P(H = 1\|E = 1, F = 1)$	>	$P(H = 1\|E = 1)$
(d)	$P(C = 1\|F = 0)$	=	$P(C = 1)$
(e)	$P(C = 1\|H = 0)$	>	$P(C = 1)$
(f)	$P(C = 1\|H = 0, F = 0)$	=	$P(C = 1\|H = 0)$
(g)	$P(D = 1\|H = 1, E = 0)$	=	$P(D = 1\|H = 1)$
(h)	$P(D = 1\|E = 1, F = 0, W = 1)$	>	$P(D = 1\|E = 1, F = 0)$
(i)	$P(T = 1\|W = 1, F = 0)$	>	$P(T = 1\|W = 1)$

6. **[25 pts]** Download the `diningData.mat` and `categoryNames.mat` files from Moodle. Each column of the `diningData` matrix corresponds to one real dining scene image from JHU table setting dataset (http://cis.jhu.edu/entropy-pursuit/images.php), and each row corresponds to a particular object category from a refined set of $n = 10$ categories commonly seen on table settings. The (i, j)th entry of `diningData` is the binary presence indicator of category i in image j. Similarly, the ith entry of `categoryNames` array contains the name for the ith category.
   a) Let $X_i \in \{0,1\}$ denote the indicator random variable for $i = 1, \dots, n$. Using data, estimate and provide the mutual information $\hat{I}(X_i; X_j)$ for each pair $i \neq j$.
   b) Setting $\hat{I}(X_i; X_j)$ as edge weights between each possible variable pair, write a program to compute the maximally spanning tree T^* over $\boldsymbol{X} = (X_i, i = 1, \dots, n)$, and plot (or draw) T^* providing category names as node labels.
   c) Let $P(\boldsymbol{X})$ be the true joint distribution over $\boldsymbol{X}$, and let $P_T(\boldsymbol{X})$ denote some approximation to P that uses tree structure T. Show that $T^* = \arg\min_T D_{\mathrm{KL}}(P||P_T)$.

Solution:

b)



c) $\arg\min_T D_{\mathrm{KL}}(P||P_T) = \arg\min_T \left(\sum_{\boldsymbol{x}} P(\boldsymbol{x}) \log \frac{P(\boldsymbol{x})}{P_T(\boldsymbol{x})}\right)$

$= \arg\min_T \left(-H(P) - \sum_{\boldsymbol{x}} P(\boldsymbol{x}) \sum_i \log P\left(x_i \middle| x_{\mathrm{Pa}_T(i)}\right)\right)$

$= \arg\max_T \left(\sum_i \sum_{\boldsymbol{x}} P(\boldsymbol{x}) \log \frac{P(x_i, x_{\mathrm{Pa}_T(i)})}{P(x_{\mathrm{Pa}_T(i)})}\right) = \arg\max_T \left(\sum_i \sum_{x_i, x_{\mathrm{Pa}_T(i)}} P(x_i, x_{\mathrm{Pa}_T(i)}) \log \frac{P(x_i, x_{\mathrm{Pa}_T(i)})}{P(x_{\mathrm{Pa}_T(i)})}\right)$

$= \arg\max_T \left(\sum_i \sum_{x_i, x_{\mathrm{Pa}_T(i)}} P(x_i, x_{\mathrm{Pa}_T(i)}) \log \frac{P(x_i) P(x_i, x_{\mathrm{Pa}_T(i)})}{P(x_i) P(x_{\mathrm{Pa}_T(i)})}\right)$

$= \arg\max_T \left(\sum_i \sum_{x_i, x_{\mathrm{Pa}_T(i)}} P(x_i, x_{\mathrm{Pa}_T(i)}) \log \frac{P(x_i, x_{\mathrm{Pa}_T(i)})}{P(x_i) P(x_{\mathrm{Pa}_T(i)})} + \sum_i \sum_{x_i} P(x_i) \log P(x_i)\right)$

$= \arg\max_T \left(\sum_i \sum_{x_i, x_{\mathrm{Pa}_T(i)}} P(x_i, x_{\mathrm{Pa}_T(i)}) \log \frac{P(x_i, x_{\mathrm{Pa}_T(i)})}{P(x_i) P(x_{\mathrm{Pa}_T(i)})}\right) = \arg\max_T \left(\sum_i I(X_i; X_{\mathrm{Pa}_T(i)})\right) = T^*$