Problem Set 1 Solutions

Calvin Walker

1.

(a) (i)

$$P(A|B,E)P(B|E) = \frac{P(A,B,E)}{P(B,E)}P(B|E) = \frac{P(A,B,E)}{P(E)P(B|E)}P(B|E)$$
$$= \frac{P(A,B,E)}{P(E)} = P(A,B|E)$$

So $P(A,B|E) = P(A|B,E)P(B|E)$ as needed.

(ii)

$$P(A|B,E) = \frac{P(A,B,E)}{P(B,E)} = \frac{P(B|A,E)P(A|E)P(E)}{P(E)P(B|E)}$$
$$= \frac{P(B|A,E)P(A|E)}{P(B|E)}$$

(b)

$$P(X,Y|Z,W) = \frac{P(X,Y,Z,W)}{P(Z,W)} = \frac{P(X,Y,W|Z)}{P(W|Z)}$$
$$= \frac{P(X|Z)P(Y,W|Z)}{P(W|Z)}$$
$$= P(X|Z)P(Y|Z,W)$$

So $(X \perp Y|Z,W)$

(c)

$$P(X,Y,W|Z) = P(X,W|Z,Y)P(Y|Z)$$
$$= P(X|Z,Y)P(W,Y|Z)P(Y|Z) = P(X|Z,Y)P(W,Y|Z)$$
$$= P(X|Z)P(W,Y|Z)$$

So $(X \perp Y,W|Z)$

(d) (i)

$$P(H|E_1,E_2) = \frac{P(H,E_1,E_2)}{P(E_1,E_2)} = \frac{P(H)P(E_1|H)P(E_2|E_1,H)}{P(E_1,E_2)}$$
$$= \frac{P(H)P(E_1,E_2|H)}{P(E_1,E_2)}$$

So option (b) is sufficient to compute $P(H|E_1,E_2)$

(ii) If $(E_1 \perp E_2|H)$, then:

$$\frac{P(H)P(E_1,E_2|H)}{P(E_1,E_2)} = \frac{P(H)P(E_1|H)P(E_2|H)}{\sum_i P(E_1|H_i)P(E_2|H_i)}$$

So option (c) is sufficient to compute $P(H|E_1,E_2)$

(e) (i)

$$E_p[-\log P(X)] \leq \log E_p\left[\frac{1}{P(X)}\right]$$
$$= \log|\mathrm{Val}(X)|$$

(ii)

$$
\begin{aligned}
-E_p[-\log P(X)] &= -\sum_X P(X)(-\log P(X)) \\
&= \sum_X P(X) \sum_X \log P(X) \\
&\leq \log \sum_X P(X) = 0
\end{aligned}
$$

So $E_p[-\log P(X)] \geq 0$ as needed

(iii)

$$
\begin{aligned}
-E_p[\log \frac{P(X)}{Q(X)}] &= -\sum_X P(X) \log \frac{P(X)}{Q(X)} \\
&= \sum_X P(X) \log \frac{Q(X)}{P(X)} \\
&\leq \log \sum_X Q(X) = 0
\end{aligned}
$$

So $E_p[\log \frac{P(X)}{Q(X)}] \geq 0$ as needed

(f) (i)

$$
\begin{aligned}
H_p(X|Y) - H_p(X) &= E_p[-\log P(X|Y)] - E_p[-\log P(X)] \\
&= E_p\left[\log \frac{P(X)}{P(X|Y)}\right] \\
&= \sum_{X,Y} P(X,Y) \log \frac{P(X)}{P(X|Y)} = \sum_{X,Y} P(X,Y) \log \frac{P(X)}{P(X,Y)/P(Y)} \\
&\leq \log \sum_{X,Y} P(X)P(Y) = 0
\end{aligned}
$$

So $H_p(X|Y) - H_p(X) \leq 0$ and $H_p(X|Y) \leq H_p(X)$ as needed

(ii)

$$
\begin{aligned}
-I(X;Y) &= -E_p\left[\log \frac{P(X|Y)}{P(X)}\right] \\
&\leq -\log \sum_{X,Y} P(X,Y) \frac{P(X|Y)}{P(X)} = -\log \sum_{X,Y} P(X,Y) \frac{P(X,Y)/P(Y)}{P(X)} \\
&= \log \sum_{X,Y} P(X)P(Y) = 0
\end{aligned}
$$

So $I(X;Y) \geq 0$ as needed

(g) (i)

$$
\begin{aligned}
I_p(X;Y|Z) &= E_p\left[\log \frac{P(X|Y,Z)}{P(X|Z)}\right] \\
&= E_p[\log P(X|Y,Z) - \log P(X|Z)] \\
&= E[-\log P(X|Z)] - E_p[-\log P(X|Y,Z)] \\
&= H_p(X|Z) - H_p(Z|Y,Z)
\end{aligned}
$$

(ii)

$$
\begin{aligned}
I_p(X;Y,Z) &= E_p\left[\log \frac{P(X|Y,Z)}{P(X)}\right] \\
&= E_p[\log P(X|Y,Z) - \log P(X)] \pm E_p[-\log P(X|Y)] \\
&= E_p\left[\log \frac{P(X|Y)}{P(X)}\right] - E_p\left[\log \frac{P(X|Y,Z)}{P(X|Y)}\right] \\
&= I_p(X;Y) + I_p(X;Z|Y)
\end{aligned}
$$

2.1

(a) No, there is an active trail $W \leftrightharpoons F \leftrightharpoons M$

(b) Yes, d-sep$(W, M|F)$

(c) No, there is an active trail $W \leftrightharpoons D \leftrightharpoons H$

(d) Yes, d-sep$(W, H|F, D)$

(e) No, there is an active trail $W \leftrightharpoons F \leftrightharpoons H \leftrightharpoons L \leftrightharpoons N$

(f) Yes, d-sep$(W, N|D, H)$

(g) No, F and D have a common cause W

(h) No, there is an active trail $F \leftrightharpoons H \leftrightharpoons D$

(i) Yes, d-sep$(F, D|W)$

(j) No, conditioning on N activates the v-structure so there is an active trail $F \leftrightharpoons H \leftrightharpoons L \leftrightharpoons N \leftrightharpoons D$

(k) No, M and N share the common cause F

(l) No, there is an active trail $M \leftrightharpoons F \leftrightharpoons H \leftrightharpoons D \leftrightharpoons N$

2.2 : $P(W, F, D, M, H, L, N) = P(W)P(F|W)P(D|W)P(M|F)P(H|F, D)P(L|H)P(N|L, D)$

2.3

(a) $P(F) = P(F|W)P(W) + (F|\neg W)P(\neg W) = 0.25$

(b) $P(F|W, H) = \frac{P(F,W,H)}{P(W,H)} = \frac{\sum_D P(W,F,D,H)}{\sum_D \sum_F P(W,F,D,H)} = \frac{0.162}{0.267} = 0.61$

(c) $P(F|W, D, H) = \frac{P(F,W,D,H)}{P(W,D,H)} = 0.43$

3.

(a)

$$\begin{aligned}
P(H|V) &= \frac{P(V,H)}{\sum_h (V,h)} \\
&= \frac{\exp(h^T W v + a^T v + b^T h)/Z}{\sum_h \exp(h^T W v + a^T v + b^T h)/Z} \\
&= \frac{\prod_j \exp(\sum_i h_j w_{ij} v_i + b_j h_j)}{\sum_h \prod_j \exp(\sum_i h_j w_{ij} v_i + b_j h_j)} \\
&= \frac{\prod_j \exp(\sum_i h_j w_{ij} v_i + b_j h_j)}{\prod_j \sum_{h_j \in \{0,1\}} \exp(\sum_i h_j w_{ij} v_i + b_j h_j)} \\
&= \frac{\prod_j \exp(\sum_i h_j w_{ij} v_i + b_j h_j)}{\prod_j (1 + \exp(b_j + \sum_i w_{ij} v_i))} \\
&= \prod_j P(H_j|V)
\end{aligned}$$

$$\begin{aligned}
P(H_j = 1|V) &= \frac{\exp(b_j + \sum_i h_i w_{ij} v_i)}{(1 + \exp(b_j + \sum_i w_{ij}))} \\
&= \sigma(b_j + \sum_i w_{ij} v_i)
\end{aligned}$$

(b)

$$\begin{aligned}
P(V|H) &= \frac{P(V,H)}{\sum_v (v,H)} \\
&= \frac{\exp(h^T W v + a^T v + b^T h)/Z}{\sum_v \exp(h^T W v + a^T v + b^T h)/Z} \\
&= \frac{\prod_i \exp(\sum_j h_j w_{ij} v_i + a_i v_i)}{\sum_h \prod_i \exp(\sum_j h_j w_{ij} v_i + a_i v_i)} \\
&= \frac{\prod_i \exp(\sum_j h_j w_{ij} v_i + a_i v_i)}{\prod_i \sum_{v_i \in \{0,1\}} \exp(\sum_j h_j w_{ij} v_i + a_i v_i)} \\
&= \frac{\prod_i \exp(\sum_j h_j w_{ij} v_i + a_i v_i)}{\prod_i (1 + \exp(a_i + \sum_j h_j w_{ij}))} \\
&= \prod_i P(V_i|H)
\end{aligned}$$

$$\begin{aligned}
P(V_i = 1|H) &= \frac{\exp(a_i + \sum_j h_i w_{ij} v_i)}{(1 + \exp(a_i + \sum_j h_j w_{ij}))} \\
&= \sigma(a_i + \sum_j h_j w_{ij})
\end{aligned}$$

(c) Yes, as shown in parts (a) and (b), the hidden units are conditionally independent given the visible units and vice versa. So the Markov network is an I-map for the distribution in Equation 2.

(d)

$$\begin{aligned}
\frac{\partial \log P(V=v)}{\partial w_{ij}} &= \frac{\partial}{\partial w_{ij}} \log \frac{1}{Z} \sum_h \exp(-E(v,h)) \\
&= \frac{\partial}{\partial w_{ij}} \log \sum_h \exp(-E(v,h)) - \frac{\partial}{\partial w_{ij}} \log Z \\
&= \frac{\frac{\partial}{\partial w_{ij}} \sum_h \exp(-E(v,h))}{\sum_h \exp(-E(v,h))} - \frac{\frac{\partial}{\partial w_{ij}} Z}{Z} \\
&= \frac{\sum_h \exp(-E(v,h)) \frac{\partial}{\partial w_{ij}} (-E(v,h))}{\sum_h \exp(-E(v,h))} - \frac{\sum_{v,h} \exp(-E(v,h)) \frac{\partial}{\partial w_{ij}} (-E(v,h))}{Z}
\end{aligned}$$

Where $\frac{\partial}{\partial w_{ij}}(-E(v,h)) = v_i h_j$. So

$$\begin{aligned}
&\frac{\sum_h \exp(-E(v,h)) \frac{\partial}{\partial w_{ij}} (-E(v,h))}{\sum_h \exp(-E(v,h))} - \frac{\sum_{v,h} \exp(-E(v,h)) \frac{\partial}{\partial w_{ij}} (-E(v,h))}{Z} \\
&= \sum_h P(H=h|V=v) v_i h_j - \sum_{v,h} P(V=v, H=h) v_i h_j \\
&= \mathbb{E}[V_i H_j | V=v] - \mathbb{E}[V_i H_j]
\end{aligned}$$

(e) w_{00}, w_{10}, w_{20} and w_{30} would be positive as all the films are in the action drama except "The Notebook". Similarly, w_{11} and w_{41} would be positive since "Casablanca" and "The Notebook" are in the romance genre.

4.

(a)

(b) For all $(X \perp Y \mid MB_H(X))$, $Y \notin MB_H(X)$ and $X \notin MB_H(Y)$, so $X - Y \notin H$ and thus $(X \perp Y \mid \mathbf{X} - \{X, Y\})$. So if $P \models I_l(H)$, then $P \models I_p(H)$.

(c) (i) Let U_1 and U_2 be sets in $\mathcal{U}$. Then,

$$P \models (X \perp U_2 - U^*, \mathbf{X} - \{X\} - U^* \mid U_1 - U^*, U^*) \;\&\; (X \perp U_1 - U^*, \mathbf{X} - \{X\} - U^* \mid U_2 - U^*, U^*)$$

So by intersection and decomposition

$$P \models (X \perp U_1 - U^*, U_2 - U^* \mid U^*)$$

Combining the results

$$P \models (X \perp U_1 - U^*, U_2 - U^*, \mathbf{X} - \{X\} - \{U_1, U_2\} \mid U^*)$$

So $P \models (X \perp \mathbf{X} - \{X\} - U^* \mid U^*)$, and $U^* \in \mathcal{U}$. By definition, $MB_P(X)$ is the minimal set $U \in \mathcal{U}$, which is U^*.

(ii)

$$P \models (X \perp Y | \mathbf{X} - \{X, Y\})$$
$$P \models (X \perp \mathbf{X} - \{X\} - (\mathbf{X} - \{X, Y\}) | \mathbf{X} - \{X, Y\})$$

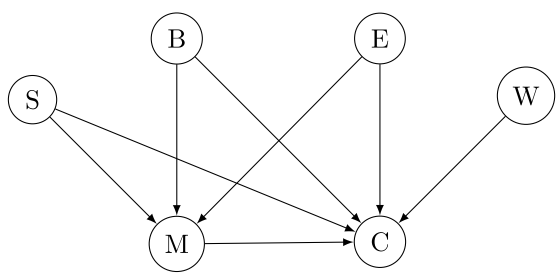
So $P \models (X \perp \mathbf{X} - \{X\} - U | \, U)$. But, $Y \notin U$ so $Y \notin U^* = MB_P(X)$

(iii) If $Y \notin MB_P(X)$ then $Y \in \mathbf{X} - \{X\} - MB_P(X)$ and $(X \perp Y, \mathbf{X} - \{X, Y\} - MB_P(X) \mid MB_P(X))$. By weak union, $(X \perp Y \, | \mathbf{X} - \{X, Y\} - MB_P(X), MB_P(X))$. Thus, $(X \perp Y \mid \mathbf{X} - \{X, Y\})$

(iv) Parts (ii) and (iii) demonstrate that $MB_P(X)$ is exactly the set of neighbors of X in H, and thus a minimal I-map for P.

5.

(a)



The above Bayesian Network is a minimal I-map for the marginal distribution over the remaining variables. When removing A from the original network, we preserve the active trails from B and E to M and C that previously passed through A. Furthermore, in the original network, if conditioning on M, due to the v-structure, there was a dependency between S and C (the symmetric dependency between W and M when conditioning on C is also preserved with this edge). The dependency between M and C (common cause A) in the orginal network is preserved without loss of generality by a new edge from M to C.

(b) We can generalize the above process as follows. When removing X_i, for each child of X_i in BN, we add edges from the parents of X_i, an edge from the other decendents of X_i, and an edge from the parents of the other children of X_i. Formally, let Pa'_{X_j} denote the parents of X_j in BN'. For each child $X_j \in \mathrm{Children}_{X_i}$

$$\mathrm{Pa}'_{X_j} = \mathrm{Pa}_{X_j} \cup \mathrm{Pa}_{X_i} \cup \{X_k, \mathrm{Pa}_{X_k} | X_k \in \mathrm{Children}_{X_i}, X_j \notin \mathrm{Pa}'_{X_k}\}$$

Where the last condition prevents adding redundant dependencies as the algorithm progresses, for instance, over a topological ordering of BN. For non-children of X_i in BN, $\mathrm{Pa}'_{X_j} = \mathrm{Pa}_{X_j}$.

6.

(a) I did not collaborate

(b)