CS688 Assignment 01

Sourabh Kulkarni | 29572060

Problem 1

The factorization of the graph is as follows:

$$
\begin{aligned}
&P(A, G, CH, BP, HD, HR, CP, EIA, ECG) \\
&= P(A)P(G)P(CH|A, G)P(BP|G)P(HD|BP, CH)P(HR|A, BP, HD)P(CP|HD)P(EIA|HD)P(ECG|HD)
\end{aligned}
$$

Problem 2

The log likelihood of the BN model as a function of its parameters θ given data N is as follows:

$$
\begin{aligned}
\mathcal{L}(\theta|\boldsymbol{x}_{1:N}) &= \frac{1}{N}\sum_{n=1}^{N} \log P(X = x_n) \\
&= \frac{1}{N}\sum_{n=1}^{N} \log P(A = a_n, G = g_n, CH = ch_n, BP = bp_n, HD = hd_n, HR = hr_n, CP = cp_n, \\
&\quad EIA = eia_n, ECG = ecg_n) \\
&= \frac{1}{N}\sum_{n=1}^{N} (\log P(A = a_n) + \log P(G = g_n) + \log P(CH = ch_n|A = a_n, G = g_n) \\
&\quad + \log P(BP = bp_n|\, G = g_n) + \log P(HD = hd_n|CH = ch_n, BP = bp_n) \\
&\quad + \log P(HR = hr_n|\, A = a_n, BP = bp_n, HD = hd_n) + \log P(CP = cp_n|HD = hd_n) \\
&\quad + \log P(EIA = eia_n|HD = hd_n) + \log P(ECG = ecg_n|HD = hd_n)) \\
&= \frac{1}{N}\sum_{n=1}^{N} \Bigg(\sum_{a} [a_n = a] \log \theta_a^A + \sum_{g} [g_n = g] \log \theta_g^G \\
&\quad + \sum_{ch,a,g} [ch_n = ch, a_n = a, g_n = g\,] \log \theta_{ch|a,g}^{CH} + \sum_{bp,g} [bp_n = bp, g_n = g] \log \theta_{bp|g}^{BP} \\
&\quad + \sum_{hd,ch,bp} [hd_n = hd, ch_n = ch, bp_n = bp] \log \theta_{hd|ch,bp}^{HD} \\
&\quad + \sum_{hr,a,bp,hd} [hr_n = hr, a_n = a, bp_n = bp, hd_n = hd\,] \log \theta_{hr|a,bp,hd}^{HR} \\
&\quad + \sum_{cp,hd} [cp_n = cp, hd_n = hd] \log \theta_{cp|hd}^{CP} + \sum_{eia,hd} [eia_n = eia, hd_n = hd] \log \theta_{eia|hd}^{EIA} \\
&\quad + \sum_{ecg,hd} [ecg_n = ecg, hd_n = hd] \log \theta_{ecg|hd}^{ECG} \Bigg)
\end{aligned}
$$

Problem 3

Forming the Lagrangian of the log likelihood function we get:

$$\mathcal{L}(\theta|\lambda) = \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{a}[a_n = a]\log\theta_a^A\right) - \lambda^A\left(\sum_a \theta_a^A - 1\right)$$
$$+\frac{1}{N}\sum_{n=1}^{N}\left(\sum_{g}[g_n = g]\log\theta_g^G\right) - \lambda^G\left(\sum_g \theta_g^G - 1\right)$$
$$+\frac{1}{N}\sum_{n=1}^{N}\left(\sum_{ch,a,g}[ch_n = ch, a_n = a, g_n = g\,]\log\theta_{ch|a,g}^{CH}\right) - \sum_{a,g}\lambda_{a,g}^{CH}\left(\sum_{ch}\theta_{ch|a,g}^{CH} - 1\right)$$
$$+\frac{1}{N}\sum_{n=1}^{N}\left(\sum_{bp,g}[bp_n = bp, g_n = g]\log\theta_{bp|g}^{BP}\right) - \sum_{g}\lambda_{g}^{BP}\left(\sum_{bp}\theta_{bp|g}^{BP} - 1\right)$$
$$+\frac{1}{N}\sum_{n=1}^{N}\left(\sum_{hd,ch,bp}[hd_n = hd, ch_n = ch, bp_n = bp]\log\theta_{hd|ch,bp}^{HD}\right)$$
$$-\sum_{ch,bp}\lambda_{ch,bp}^{HD}\left(\sum_{ch}\theta_{hd|ch,bp}^{HD} - 1\right)$$
$$+\frac{1}{N}\sum_{n=1}^{N}\left(\sum_{hr,a,bp,hd}[hr_n = hr, a_n = a, bp_n = bp, hd_n = hd\,]\log\theta_{hr|a,bp,hd}^{HR}\right)$$
$$-\sum_{a,bp,hd}\lambda_{a,bp,hd}^{HR}\left(\sum_{ch}\theta_{hr|a,bp,hd}^{HR} - 1\right)$$
$$+\frac{1}{N}\sum_{n=1}^{N}\left(\sum_{cp,hd}[cp_n = cp, hd_n = hd]\log\theta_{cp|hd}^{CP}\right) - \sum_{hd}\lambda_{hd}^{CP}\left(\sum_{cp}\theta_{cp|hd}^{CP} - 1\right)$$
$$+\frac{1}{N}\sum_{n=1}^{N}\left(\sum_{eia,hd}[eia_n = eia, hd_n = hd]\log\theta_{eia|hd}^{EIA}\right) - \sum_{hd}\lambda_{hd}^{EIA}\left(\sum_{eia}\theta_{eia|hd}^{EIA} - 1\right)$$
$$+\frac{1}{N}\sum_{n=1}^{N}\left(\sum_{ecg,hd}[ecg_n = ecg, hd_n = hd]\log\theta_{ecg|hd}^{ECG}\right) - \sum_{hd}\lambda_{hd}^{ECG}\left(\sum_{ecg}\theta_{ecg|hd}^{ECG} - 1\right)$$

To obtain MLE for $\theta_{L|1,H,Y}^{HR}$, we take partial derivatives of the above expression w.r.t $\theta_{L|1,H,Y}^{HR}$ and $\theta_{H|1,H,Y}^{HR}$. We then equate those to 0:

$$\frac{\partial\mathcal{L}(\theta|\lambda)}{\partial\theta_{L|1,H,Y}^{HR}} = \frac{1}{N}\sum_{n=1}^{N}\frac{[hr_n = L, a_n = 1, bp_n = H, hd_n = Y\,]}{\theta_{L|1,H,Y}^{HR}} - \lambda_{1,H,Y}^{HR} = 0 \ldots (1)$$

$$\frac{\partial \mathcal{L}(\theta|\lambda)}{\partial \theta^{HR}_{H|1,H,Y}} = \frac{1}{N}\sum_{n=1}^{N}\frac{[hr_n = H, a_n = 1, bp_n = H, hd_n = Y\,]}{\theta^{HR}_{H|1,H,Y}} - \lambda^{HR}_{1,H,Y} = 0 \ldots (2)$$

But we know that $\theta^{HR}_{L|1,H,Y}$ + $\theta^{HR}_{H|1,H,Y} = 1$:

$$\frac{1}{N}\sum_{n=1}^{N}\frac{[hr_n = L, a_n = 1, bp_n = H, hd_n = Y\,]}{\lambda^{HR}_{1,H,Y}} + \frac{1}{N}\sum_{n=1}^{N}\frac{[hr_n = H, a_n = 1, bp_n = H, hd_n = Y\,]}{\lambda^{HR}_{1,H,Y}} = 1$$

$$\frac{1}{N}\sum_{n=1}^{N}\frac{[\,a_n = 1, bp_n = H, hd_n = Y\,]}{\lambda^{HR}_{1,H,Y}} = 1 \ldots (3)$$

Substituting (3) in (1), we get:

$$\frac{1}{N}\sum_{n=1}^{N}\frac{[hr_n = L, a_n = 1, bp_n = H, hd_n = Y\,]}{\theta^{HR}_{L|1,H,Y}} - \frac{1}{N}\sum_{n=1}^{N}[\,a_n = 1, bp_n = H, hd_n = Y\,] = 0$$

$$\theta^{HR}_{L|1,H,Y} = \frac{\sum_{n=1}^{N}[hr_n = L, a_n = 1, bp_n = H, hd_n = Y\,]}{\sum_{n=1}^{N}[\,a_n = 1, bp_n = H, hd_n = Y\,]}$$

Which is the desired expression.

Problem 4

4.1 The CPT implementation was done in 3 steps:

1. The network itself was encoded as a dictionary, where every node was a key and the parents of that node were its values

2. Another dictionary was initialized with the nodes of the BN as keys. Each node has for its value, another dictionary which is the CPT for that node

3. The dictionary which represents the CPT for each node has a tuple for its key which is a combination of all the values the node and its parents can take. This was achieved by iterating over all possible combinations using the itertools library from python.

Problem 5

5.1 To perform MLE, I initialized counters corresponding to each CPT in a similar fashion as setting up the CPTs themselves. The fit method then goes through the input array, incrementing each corresponding count. The MLE estimate is then done by taking the ratio of counts of a certain value of node to sum of all counts for all values of node for same combination of parent values.

Problem 6

6.1 Probability queries:

(a) $P(CH = L|A = 2, G = M, CP = None, ECG = Normal, HR = L, EIA = No, HD = No)$

$$= \frac{P(CH = L, A = 2, G = M, CP = None, ECG = Normal, HR = L, EIA = No, HD = No)}{P(A = 2, G = M, CP = None, ECG = Normal, HR = L, EIA = No, HD = No)}$$

$$= \frac{P(CH = L, A = 2, G = M, CP = None, ECG = Normal, HR = L, EIA = No, HD = No)}{\sum_{ch} P(CH = ch, A = 2, G = M, CP = None, ECG = Normal, HR = L, EIA = No, HD = No)}$$

$$= \frac{P(CH = L, A = 2, G = M)P(HD = No|CH = L, BP = L)}{\sum_{ch} P(CH = ch| A = 2, G = M)P(HD = No|CH = ch, BP = L)}$$

$$P\big((CH = H|A = 2, G = M, CP = None, ECG = Normal, HR = L, EIA = No, HD = No)\big)$$
$$= 1 - P((CH = L|A = 2, G = M, CP = None, ECG = Normal, HR = L, EIA = No, HD = No))$$

(b) $P(BP = L|A = 2, CP = Typical, CH = H, ECG = Normal, HR = H, EIA = Yes, HD = No)$

$$= \frac{P(BP = L, A = 2, CP = Typical, CH = H, ECG = Normal, HR = EIA = Yes, HD = No)}{\sum_{bp} P(BP = bp, A = 2, CP = Typical, CH = H, ECG = Normal, HR = EIA = Yes, HD = No)}$$

$$= \frac{\sum_g P(BP = L, A = 2, G = g, CP = Typical, CH = H, ECG = Normal, HR = EIA = Yes, HD = No)}{\sum_g \sum_{bp} P(BP = bp, A = 2, CP = Typical, CH = H, ECG = Normal, HR = EIA = Yes, HD = No)}$$

$$= \frac{\sum_g P(G = g)P(CH = H|G = g, A = 2)P\big(BP = L|G = g)\big)P(HR = H|A = 2, BP = L, HD = No)P(HD = No|BP = L, CH = H)}{\sum_g \sum_{bp} P(G = g)P(CH = H|G = g, A = 2)P\big(BP = bp|G = g)\big)P(HR = H|A = 2, BP = bp, HD = No)P(HD = No|BP = L, C}$$

$$P(BP = h, A = 2, CP = Typical, CH = H, ECG = Normal, HR = EIA = Yes, HD = No)$$
$$= 1 - \ P(BP = L, A = 2, CP = Typical, CH = H, ECG = Normal, HR = EIA = Yes, HD = No)$$

Problem 7

7.1

$$P(HD = N|A = a, G = g, CH = ch, BP = bp, HR = hr, CP = cp, EIA = eia, ECG = ecg)$$

$$= \frac{P(A = a)P(G = g)P(CH = CH|A = a, G = g)P(BP = bp|G = g)P(HD = N|CH = ch, BP = bp)P(HR = hr|A = a, BP = bp, HD = N)P(CP = CP|HD = N)P(EIA = eia|HD = N)P(ECG = ecg|HD = N)}{\sum_{hd} P(A = a)P(G = g)P(CH = CH|A = a, G = g)P(BP = bp|G = g)P(HD = hd|CH = ch, BP = bp)P(HR = hr|A = a, BP = bp, HD = hd)P(CP = CP|HD = hd)P(EIA = eia|HD = hd)P(ECG = ecg|HD = hd)}$$

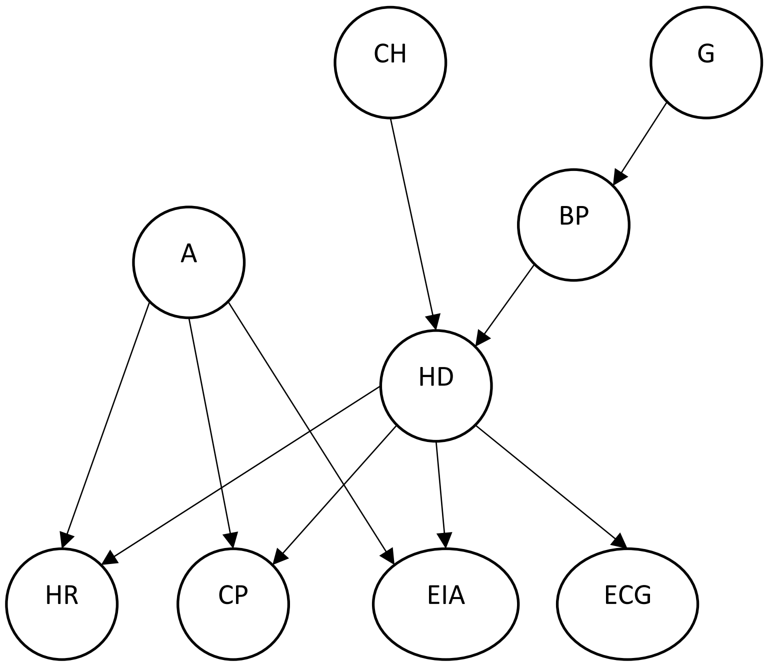
$$= \frac{P(HD = N|CH = ch, BP = bp)P(HR = hr|A = a, BP = bp, HD = N)P(CP = CP|HD = N)P(EIA = eia|HD = N)P(ECG = ecg|HD = N)}{\sum_{hd} P(HD = hd|CH = ch, BP = bp)P(HR = hr|A = a, BP = bp, HD = hd)P(CP = CP|HD = hd)P(EIA = eia|HD = hd)P(ECG = ecg|HD = hd)}$$

7.3 Model crossvalidation:

- Individual accuracies: [0.8 0.76666667 0.73333333 0.71666667 0.75]
- Mean:75.33333333333334%
- Standard deviation:2.8674417556808782%
- Execution time:0.0639998912811

Problem 8

8.1 Proposed model BN:



8.2 Factorization of the BN:

$$P(A, G, CH, BP, HD, HR, CP, EIA, ECG)$$
$$= P(A)P(G)P(CH)P(BP|G)P(HD|BP, CH)P(HR|A, HD)P(CP|A, HD)P(EIA|A, HD)P(ECG|HD)$$

8.3 Here are the changes made to the original network and a brief reasoning for those changes:

1. Removing age and gender as parents of cholesterol levels: Causal relations of age and gender toward cholesterol levels were removed, signifying that there is a genetic predisposition to higher cholesterol levels. A possible parent node here might be the presence of the genetic markers for high cholesterol in the patient's genome.

2. Removed blood pressure as a parent of heart rate: The increase in heart rate due to increased blood pressure usually signifies a heart disease, hence the additional connection is redundant.

3. Adding age as a parent to chest pain: Age has a correlation with chest pain, as there can be various other reasons which cause chest pain in an aged person, and can ‘explain away’ the chest pain in certain cases.

4. Adding age as a parent to EIA: Age is an important factor in whether exercise induces angina. If an aged patient has EIA, the age might partially 'explain away' the presence of EIA in certain cases.

8.5 Modified Network Model crossvalidation:

- Individual accuracies: [0.8 0.75 0.75 0.75 0.78333333]
- Mean:76.66666666666666%
- Standard deviation:2.108185106778921%
- Execution time:0.0460000038147

This model has a better mean accuracy (76.6% vs. 75.3% for original), and has a lower standard deviation (2.11% vs. 2.86% for original). The reduced complexity of the network also resulted in a shorter execution time (0.046s vs. 0.064s for original).