Assignment3

Part1 - Task3

The Bayesian network is a compact representation of conditional dependency relation between random variables, consisting of a graph and conditional probability tables (CPT). The result of likelihood and log-likelihood can be used to examine the performance of the Bayesian network.

Likelihood formula:

$$\text{Likelihood: } P\left(Data \middle| Model\right) = \prod_i P\left(Data_i \middle| Model\right)$$
$$= \prod_i \prod_j P\left(N_i^j \middle| Parent\left(N^j\right), Model\right)$$

Log-likelihood formula:

$$\text{Log likelihood:}$$
$$P\left(Data \middle| Model\right) = \log\left(\prod_i \prod_j P\left(N_i^j \middle| Parent\left(N^j\right), Model\right)\right)$$
$$= \sum_i \sum_j \log\left(P\left(N_i^j \middle| Parent\left(N^j\right), Model\right)\right)$$

Please explain how the likelihood and log-likelihood measure of the Bayesian Network differs as the number of training data set increases.

If more training data is added to the model to compute the likelihood and log-likelihood, then a more accurate result will be displayed.

For example, if the number of training data set increases by 1, the likelihood of the data set will decrease, which is the product of probability of each row: P (Data-1)*P (Data-2)*…*P (Data-n)*P (Data-n+1). The log-likelihood is defined as the summation of the log of probability of each row: log (P (Data-1))+ log (P (Data-2))+… log (P (Data-n))+ log (P(Data-n+1)), the result of log-likelihood of will increase.

Please explain how the likelihood and log-likelihood measure of the Bayesian Network differs as the number of variables (nodes) increases.

With increasing number of nodes, the structure of Bayesian Network will be changed, so the result of likelihood and log-likelihood which may lead to completely different results of likelihood and log-likelihood. The results of likelihood and log-likelihood depend on the structure of Bayesian Network. If the structure of Bayesian Network changes, the results will change as well.

For example, if there are two nodes A and B in a Bayesian Network (A -> B, data: 0 0), and then add a new node C (A -> B -> C, data: 0 0 1). So for the formula of likelihood will change from P(~A)∗P(~B|~A) to P(~A)∗P(~B|~A)∗P(C|~B), the corresponding result will decrease (CPT of node is from 0 to 1). For the log-likelihood, the formula will change from log(P(~A))+log(P(~B|~A)) to log(P(~A))+log(P(~B|~A))+log(P(C|~B)), and result will increase.

Please write a short discussion on how the likelihood and log-likelihood measure will differ when the possible values of each variable increases.

If the possible values of each variable (node) increases, the structure of Bayesian network will not change, but the CPT of each node will change. As the results of likelihood and log-likelihood denpend on CPT, if CPT of node is changed, the results of likelihood and log-likelihood will also change.
In this case, the possible values of each node will increase. For example, the data set of A is 0 1 1 0, so CPT of node A is P (A) = 2/4 = 0.5, and then the value of node A will change from 0 to 2 (2 1 1 0). Thus, the CPT of node A now is P (A) = 1+1+2/2+1+1+1 = 4/5 = 0.8. In addition, (P (A) is necessary for computing, so the results of likelihood and log-likelihood will change (the equation of likelihood and log-likelihood will be same).

In conclusion, likelihood usually is a decimal number (such as 0.1 and 0.2) while log-likelihood is an integer (such as -1 and -2), so log-likelihood is easily to read and understand for the user.

Part2

Task 5
Please explain how the final Bayesian Network changes as the parameter increases/decreases.
In order to solve this question, I have two comparing strategies.

1. **C is constant.** It will take longer search time (more search times) if the threshold is bigger. It also requires a better Bayes Network structure
2. **Threshold is constant**. It requires a simpler Bayes Networks structure (Less CPT relations between nodes) if the C is bigger.

Some issues may occur if the threshold is not big enough, the search operation will stop before the best structure is found.

In conclusion, the smaller threshold leads to the less accuracy of result for choosing the Bayesian network model. In other word, C affects the complexity of Bayes Network structure and threshold affects the accuracy of the Bayesian network structure.

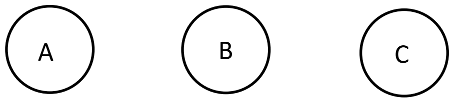
Task 6

The result file ends with "-n" is no edge model.

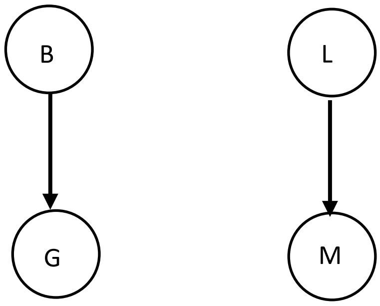
The result file ends with "-r" is random chain model.

The following Bayes Network structures are based on the result file.

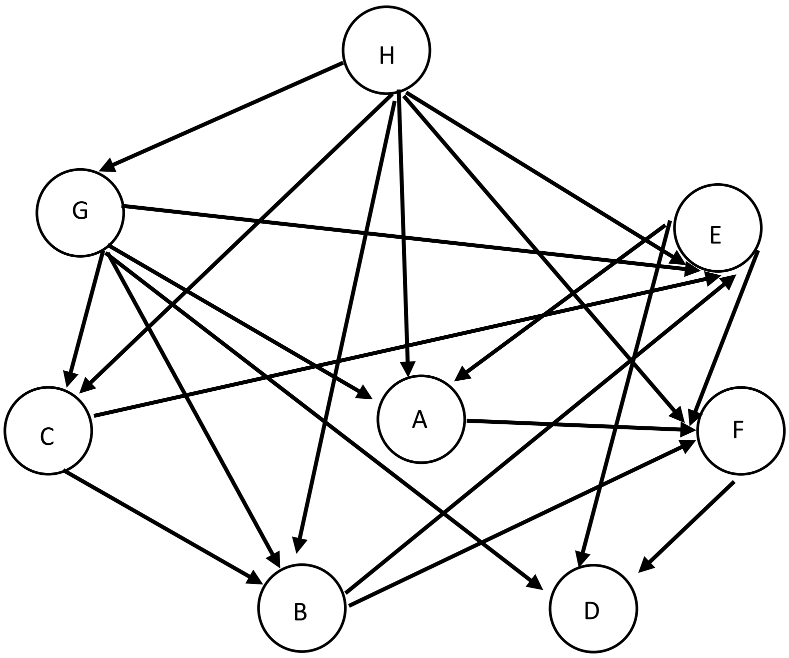
Task6 d1-n



Task6 d2-n

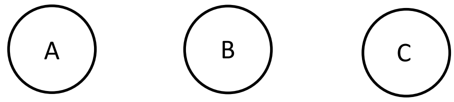


Task6 d3-n

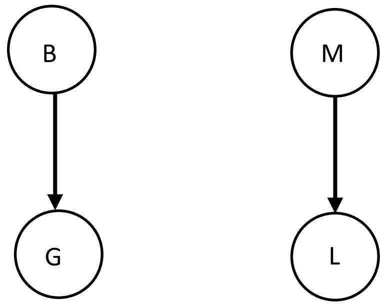


20 edges

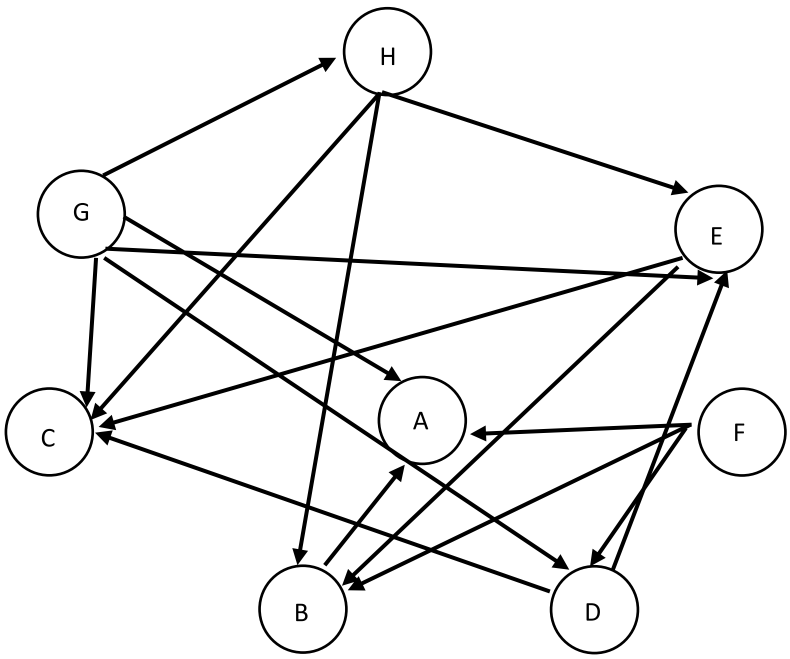
Task6 d1-r



Task6 d2-r



Task6 d3-r



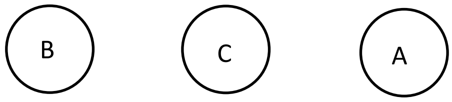
15 edges

According to the result file, the random chain model has better score than the no edge model based on complex CPT relations. But they almost have same score base on independent Bayes network structure.

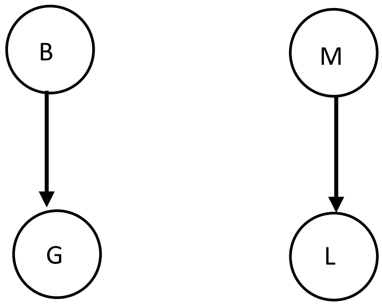
In conclusion, the no Edge Model is the better model for independent Bayes network structure

while Random Chain Model is better for complex CPT relations.

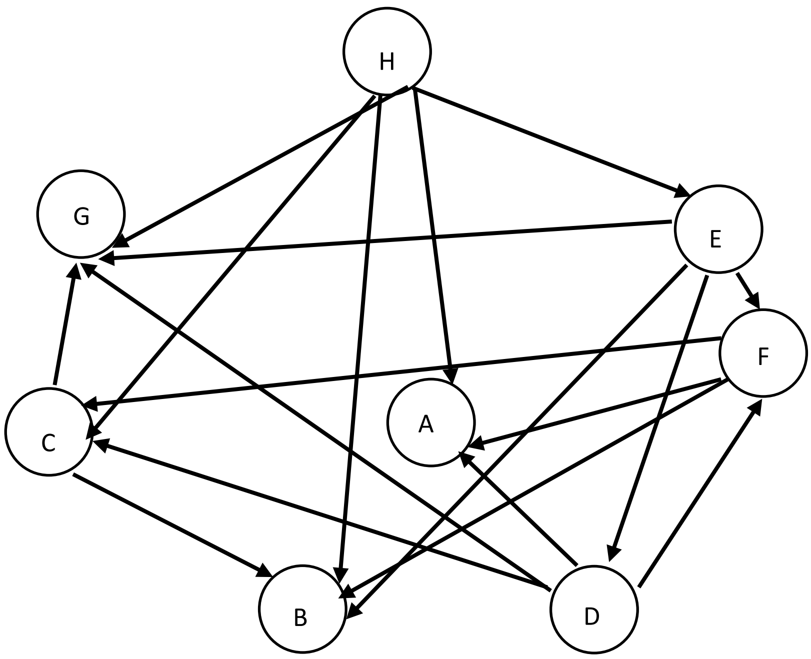
Task7 d1-b



Task7 d2-b



Task7 d3-b



17 edges

According to the result file, the score of Best Tree Model is better than No edge but worse than the Random Chain based on complex CPT relations.

In conclusion, if it is hard to make choice whether No edge or Random chain is better for training data set. The Best tree model is an alternative choice.