Semester 2 2014
COMP3702/7702 ARTIFICIAL INTELLIGENCE
ASSIGNMENT 3: Learning Bayesian Network

Note:

- You can do this assignment individually or in a group of 2-3 students.
- For those who choose to work in a group:
  - All students in the group must be enrolled in the same course code, i.e., all COMP3702 students or all COMP7702 students.
  - Please register your group members at http://goo.gl/w96zDE before **11.59pm on Saturday, November, 1st**. If you have not registered your group by then, you will need to work on the assignment individually.
- There will be **NO** demo for this assignment.
- You need to submit:
  - Your codes.
  - The results of your codes.
  - The answers to the questions in this assignment (in a .pdf file).

  via the submission website before **11.59pm Thursday, November, 13th**. We will announce the submission website in Piazza.
- **IMPORTANT:**
  - The source code should be zipped, and named:
    src-studentID-name.zip if you work individually
    src-studentID#1-studentID#2-name#1-name#2.zip if you work in a group of two.
    src-studentID#1-studentID#2-studentID#3-name#1-name#2-name#3.zip if you work in a group of three.
  - Answers to non-programming questions should be written as one .pdf file, and the file should be named:
    doc-studentID-name.pdf if you work individually
    doc-studentID#1-studentID#2-name#1-name#2.pdf if you work in a group of two.
    doc-studentID#1-studentID#2-studentID#3-name#1-name#2-name#3.pdf if you work in a group of three.
  - Please put all your files in one folder and submit a .zip file of the folder. Both the .zip and folder should be named:
    studentID-name-a3 if you work individually.
    studentID#1-studentID#2-name#1-name#2-a3 if you work in a group of two.
    studentID#1-studentID#2-studentID#3-name#1-name#2-name#3-a3 if you work in a group of three.

Your task in this assignment is to implement and explore methods to generate Bayesian Network from training data. In this assignment, we assume all variables are binary variables. The values are either 0 or 1.

Part I [50 points]: Learning Conditional Probability Tables.

The training data sets for this part are the .txt files whose names start with CPTNoMissingData. The format of each data file is:

- The first line contains two numbers separated by a white space. The first number is

the number of nodes in the Bayesian Network. The second number is the number of data in the file. Let's denote the number of nodes as K and the number of data as N.

- Each line at line 2 to K+1 represents a node of the Bayesian Network and its parents. Each line contains one or more words, separated by a white space. The first word represents the name of the node, while the rest represents the names of the node's parents. For example, A B C means Node B and C are the parents of node A.
- Each line at line K+2 to K+N+1 represents the data, in the same order the nodes are written. For instance, in CPTNoMissingData-d1.txt, each line of data represents the value of A B C.

Task-1 [25 points]. Given a set of training data together with the structure of the Bayesian Network, please write a program that computes the CPT for each node of the Bayesian Network using Maximum Likelihood Estimate as discussed in class. Please output the CPTs to a file named cpt-[NameOfTrainingData].txt in the following format:

- The output file contains 2K lines, representing K blocks of CPTs (K is the number of nodes). Each CPT block consists of two lines, where:
  - The first line contains one or more words, separated by a white space. The first word represents the name of the node, while the rest represents the names of the node's parents., i.e.,
    Node Parent-1 Parent-2 … Parent-L
    where L is the number of parents of node "Name".
  - The second line represent the CPT of the node. It contains 2^L numbers. Each number represents the conditional probability of the Node being True given the value of the parents, sorted in ascending order. For example, suppose the parents of Node A are B and C, and P(A | B, C) = 0.3, P(A | B, ~C) = 0.5, P(A | ~B, C) = 0.4, P(A | ~B, ~C) = 0.7. Then the output format for the CPT of A will be: 0.7 0.4 0.5 0.3
- Line 2K+1 is the log-likelihood of the data given the Bayesian Network model.

Task-2 [10 points]. Please write a program that computes the likelihood and the log likelihood of the training data set given the CPTs.

Task-3 [15 points]. Comparison of the likelihood and log-likelihood results.

- Please run the program you've written for Task-1 and Task-2 on each training data set. Please write the likelihood and log-likelihood of the CPT for each training data set.
- Please explain how the likelihood and log-likelihood measure of the Bayesian Network differs as the number of training data set increases.
- Please explain how the likelihood and log-likelihood measure of the Bayesian Network differs as the number of variables (nodes) increases.
- Please write a short discussion on how the likelihood and log-likelihood measure will differ when the possible values of each variable increases.

Part II [50 points]: Learning Structure and Conditional Probability Tables.

The training data sets for this part are the .txt files whose names start with noMissingData. The format of each data file is the same as the input format for Part I, but without the parents information, i.e.:

- The first line contains two numbers separated by a white space. The first number is the number of nodes in the Bayesian Network. The second number is the number of

data in the file. Let's denote the number of nodes as K and the number of data as N.
- The second line represents the names of the nodes, separated by a white space.
- Each line at line 3 to N+2 represents the data.

Task-4 [25 points]. Please write a program that will generate the structure and CPT of a Bayesian network given a set of training data. You can use a greedy search method and the scoring function as discussed in class. Please use at most 3 minutes search time. Please output the results to a file named bn-[NameOfTrainingData].txt in the following format:

- Each line at line 1 to K represents a node of the Bayesian Network and its parents. Each line contains one or more words, separated by a white space. The first word represents the name of the node, while the rest represents the names of the node's parents. For example, A B C means Node B and C are the parents of node A.
- Line K+1 to N+K represents blocks of CPTs (K is the number of nodes). Each CPT block consists of two lines, where:
  - The first line contains one or more words, separated by a white space. The first word represents the name of the node, while the rest represents the names of the node's parents., i.e.,
    Node Parent-1 Parent-2 … Parent-L
    where L is the number of parents of Node.
  - The second line represents the CPT of the node. It contains 2^L numbers. Each number represents the conditional probability that the Node is True given the value of the parents, sorted in ascending order. For example, suppose the parents of Node A are B and C, and P(A | B, C) = 0.3, P(A | B, ~C) = 0.5, P(A | ~B, C) = 0.4, P(A | ~B, ~C) = 0.7. Then the output format for the CPT of A will be: 0.7 0.4 0.5 0.3
- Line N+K+1 consists of two number separated by a white space. The first number is the log-likelihood of the data given the Bayesian Network model. The second number is the score of the Bayesian Network.

Task-5 [7 points]. Please experiment with the scoring function by changing the constant parameter. For each parameter, please run the program you've written for Task-4 on each data set. For each data set and each parameter, please write the score function of the final Bayesian Network. Please explain how the final Bayesian Network change as the parameter increases/decreases.

Task-6 [8 points]. Please implement "no edge" and "random chain" to initialize the structure. Please run the Bayesian Network generation program (Task-4) with these two initialization methods on each data set and compare the final Bayesian Network (in terms of the scoring function and structural complexity) after 3 minutes searching time.

Task-7 [10 points]. Please implement the "best tree network" to initialize the structure, and compare the final Bayesian Networks results with the Bayesian Networks generated with "no edge" and "random chain" initialization method (Task-6), in a similar manner as in Task-6.

Part III – Bonus [15 points]: Learning Structure and Conditional Probability Tables.

The training data sets for this part are files that start with someMissingData. The format of each data file is the same as the input format for Part II, but each data input may have a value of 0, 1, or H1/H2/…/HM, where Hi means missing data-i and M is the number of missing data.

Task-8 [15 points]. Please write a program to generate the structure and CPT of a Bayesian network given a set of training data where some of these data are missing. Please use the greedy search method with "best tree network" initialization to generate the structure. Please use the "fill in with distribution" strategy (as discussed in class) to learn the CPTs.
Please output the results to a file named bn-[NameOfTrainingData].txt. The format of the file should start with the same output format as in Part II, but appended with the probabilities of each missing data having the value `1'. To be more precise, the format is:

- Each line at line 1 to K represents a node of the Bayesian Network and its parents. Each line contains one or more words, separated by a white space. The first word represents the name of the node, while the rest represents the names of the node's parents. For example, A B C means Node B and C are the parents of node A.
- Line K+1 to N+K represents blocks of CPTs (K is the number of nodes). Each CPT block consists of two lines, where:
  - The first line contains one or more words, separated by a white space. The first word represents the name of the node, while the rest represents the names of the node's parents., i.e.,
    Node Parent-1 Parent-2 … Parent-L
    where L is the number of parents of Node.
  - The second line represents the CPT of the node. It contains 2^L numbers. Each number represents the conditional probability that the Node is True given the value of the parents, sorted in ascending order. For example, suppose the parents of Node A are B and C, and P(A | B, C) = 0.3, P(A | B, ~C) = 0.5, P(A | ~B, C) = 0.4, P(A | ~B, ~C) = 0.7. Then the output format for the CPT of A will be: 0.7 0.4 0.5 0.3
- Line N+K+1 consists of two number separated by a white space. The first number is the log-likelihood of the data given the Bayesian Network model. The second number is the score of the Bayesian Network.
- Line N+K+2 to line M+N+K+2 consists of two words separated by a white space, where M is the number of missing data. The first word is the name of the missing data (e.g., H1, H2, etc.). The second word is the probability that the missing data has value `1'. For example, if P(H1 | Data, Model) = 0.6, then the output will be: "H1 0.6" (of course without the quotation mark).

Please explain your implementation of the "fill in with distribution" strategy.
Please discuss how your program will perform as the number of missing data increases.