

---

## CS688: Graphical Models - Spring 2017

### Assignment 1

Assigned: Tuesday, Feb 7. Due: Tuesday, Feb 21 at 1:00

---

**General Instructions:** Please turn in two items: (1) A report with your answers, due at the start of lecture, and (2) a zip file with your code, uploaded to Moodle.

Submit a report with the answers to each question at the start of class on the date the assignment is due. You are encouraged to typeset your solutions. To help you get started, the full  $\text{\LaTeX}$  source of the assignment is included with the assignment materials. For your assignment to be considered “on time”, you must upload a zip file containing all of your code to Moodle by the due date. Make sure the code is sufficiently well documented that it’s easy to tell what it’s doing. You may use any programming language you like. For this assignment, you may not use existing code libraries for Bayesian network modeling, learning or inference. If you think you’ve found a bug with the data or an error in any of the assignment materials, please post a question to the Piazza discussion forum. Make sure to list in your report any outside references you consulted (books, articles, web pages, etc.) and any students you collaborated with.

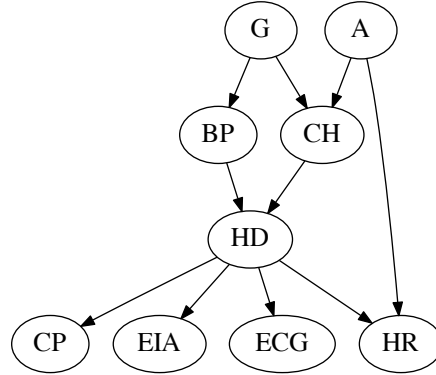
**Academic Honesty Statement:** You are encouraged to work with other students to understand the material, but not to share solutions. Copying solutions from external sources (books, web pages, etc.) or other students is considered cheating, and so is sharing your solutions. Please see the Course Policies listed in the syllabus.

**Introduction:** In this assignment, you will experiment with different aspects of modeling, learning, and applying a Bayesian network to answer probability queries. This assignment focuses on the heart disease diagnosis domain and uses part of a real clinical data set.

**Data Set:** The data set consists of 9 variables as described below. The number of each variable corresponds to its column number in the data set files. There are five sets of training and test data files in standard comma-separated-value (CSV) format. The files are named *data-train-i.txt* and *data-test-i.txt* for *i* from 1 to 5.

Number	Name	Description	Values
1	A	Age	1:< 45, 2: 45 – 55, 3:≥ 55
2	G	Gender	1:Female, 2:Male;
3	CP	Chest Pain	1:Typical, 2:Atypical, 3:Non-Anginal, 4:None
4	BP	Blood Pressure	1:Low, 2:High
5	CH	Cholesterol	1:Low, 2:High
6	ECG	Electrocardiograph	1:Normal, 2:Abnormal
7	HR	Exercise Heart Rate	1:Low, 2:High
8	EIA	Exercise Induced Angina	1:No, 2:Yes
9	HD	Heart Disease	1:No, 2:Yes

**Model:** We will consider applying a Bayesian network with the following structure to the data set.



**1. (10 points) Factorization:** Write down the factorization of the Bayesian network joint distribution implied by the structure shown above.

**2. (10 points) Likelihood Function:** Using the notation for the parameters of CPTs introduced in Lecture 4 (ie:  $P_{\theta}(HD = hd|CH = ch, BP = bp) = \theta_{hd|ch,bp}^{HD}$ ), write down the log likelihood of the Bayesian network model as a function of the parameters  $\theta$  given  $N$  data cases.

**3. (15 points) Maximum Likelihood Estimates:** Using the notation for the parameters of CPTs introduced in lecture, derive the maximum likelihood estimate for the parameter  $\theta_{L|1,Y}^{HR}$  starting from the log likelihood function. Be sure to account for the sum-to-one constraint  $\sum_{hr \in \{L,H\}} \theta_{hr|1,Y}^{HR} = 1$ . Show all of your work.

**4. (15 points) Learning:** Implement the maximum likelihood parameter estimates for all CPTs in the model. For this question, run your code on the data in the *first training data set only* to compute the maximum likelihood parameter estimates for each CPT in the model. Report the maximum likelihood values you computed for each of the following CPTs:

- (a)  $P_{\theta}(A)$
- (b)  $P_{\theta}(BP|G)$
- (c)  $P_{\theta}(HD|BP, CH)$
- (d)  $P_{\theta}(HR|A, HD)$

You may report the values of the above CPTs using the provided template *params.pdf*. Or, you may also write code to output the CPTs, which is strongly recommended. However, make sure to list the CPTs in your report in the order they appear in the supplied template. Also make sure to list the configurations within each CPT in the same order that they appear in the template. This is to facilitate grading.

**5. (15 points) Probability Queries:** For each of the two queries below, first show how the query can be expressed in terms of the factorized joint distribution for the Bayesian network. Simplify the expressions wherever possible using the conditional independence properties of the network structure. Finally, use the parameters obtained in the previous question (first training set file only) to compute the distribution over the

query variables. Display the result using a table or a bar chart. Note that there is an unobserved variable in the second query.

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

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

**6. (20 points) Classification:** In this question, we will assess the ability of the model to correctly predict the occurrence of heart disease given the values of all of the other variables in the network. Perform the following steps:

(a) Train the network on each of the five training data files, obtaining five sets of parameters. There is nothing to report for this step.

(b) Write down the probability distribution over the heart disease variable (HD) given the remaining variables. Simplify the result using the conditional independence properties of the network.

(c) We will follow a standard five-fold-cross validation protocol to assess the performance of the model. For each test file  $i$  and each test data case  $n$ , compute the most likely value of the heart disease variable  $\hat{h}_{d_{ni}}$  using the parameters learned with training file  $i$ . For each test file  $i$ , compute the prediction accuracy  $A_i$  as the number of cases correctly predicted divided by the total number of cases. Lastly, compute the mean prediction accuracy over the five test files (the average of  $A_1$  to  $A_5$ ) and the standard deviation of the prediction accuracy over the five test files (the standard deviation of  $A_1$  to  $A_5$ ). Report the mean and the standard deviation of the prediction accuracy.

**7. (15 points) Modeling:** Design your own network structure for the heart disease domain.

(a) Draw the graphical model for your network.

(b) Write down the factorization for your network.

(c) Briefly describe some of the choices that went into the design of your network structure.

(d) Use your network to repeat the heart disease classification experiment and report the mean and standard deviation of your network. Can you find a network with better accuracy than the given network?