

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, write down the log likelihood of the Bayesian network model as a function of the parameters θ 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 θ HR L|1,Y starting from the log likelihood function. Be sure to account for the sum-to-one constraint P hr∈{L,H} θ HR hr|1,Y = 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) Pθ(A)

(b) Pθ(BP|G)

(c) Pθ(HD|BP, CH)

(d) Pθ(HR|A, HD)

1. (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 = T ypical, CH = H, ECG = Normal, HR = H, EIA = Y es, HD = No)

1. (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.

**Answer:**

**P(HD) = P(HD|BP,CH)\*P(CP|HD)\*P(EIA|HD)\*P(ECG|HD)\*P(HR|A,HD)**

(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 hdˆ ni using the parameters learned with training file i. For each test file i, compute the prediction accuracy Ai 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 A1 to A5) and the standard deviation of the prediction accuracy over the five test files (the standard deviation of A1 to A5). Report the mean and the standard deviation of the prediction accuracy.

**Accuracy of data-train-4.txt-- > 0.8**

**Accuracy of data-train-1.txt-- > 0.733333333333**

**Accuracy of data-train-3.txt-- > 0.666666666667**

**Accuracy of data-train-5.txt-- > 0.783333333333**

**Accuracy of data-train-2.txt-- > 0.8**

**Mean accuracy --> 0.756666666667,**

**Standard deviation --> 0.0512076383191**

1. (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?