CS688: Graphical Models - Spring 2023

Assignment 1

Assigned: Tuesday, Feb 16. Due: Thursday, Mar 02 at 23:59.

General Instructions: Submit a **report** with the answers to each question and your **code** before the date the assignment is due. You may complete the assignment using any programming language you like. The data files for this assignment are in the `data` directory. 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 Piazza. Make sure to list in your report any outside references you consulted (books, articles, web pages, etc.) and any students you collaborated with.

Deliverables: This assignment has two types of deliverables: a report and code files.

- **Report:** The solution report will give your answers to the homework questions. Items that you should include in your report are marked with **(report)**. You can use any software to create your report, but your report must be submitted in PDF format. You will upload the PDF of your report to Gradescope under *HW01-Report* for grading. It is strongly recommended that you typeset your report. To assist with this if you wish to use Latex, the Latex source of the handout is also included in the homework archive. When you submit to Gradescope, please mark page numbers for the different questions.
- **Code:** The second deliverable is your code. Items that you should include in your code are marked with (**code**). You will upload a zip file (not rar, bz2 or other compressed format) containing all of your code to Gradescope under *HW01-Programming*.

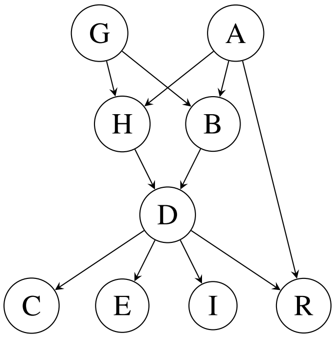
Academic Honesty Statement: Copying solutions from external sources (books, web pages, etc.) or other students is considered cheating. Sharing your solutions with other students is also considered cheating. Collaboration indistinguishable from copying is a violation of the course's collaboration policy and will be treated as cheating. Any detected cheating will result in a grade of 0 on the assignment for all students involved, and potentially a grade of F in the course.

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-m.txt` and `data-test-m.txt` for m from 1 to 5.

Number	RV	Description	Values
1	A	Age	1:<45, 2: $45-55$, 3:≥ 55
2	G	Gender	1:Female, 2:Male;
3	C	Chest Pain	1:Typical, 2:Atypical, 3:Non-Anginal, 4:None
4	B	Blood Pressure	1:Low, 2:High
5	H	Cholesterol	1:Low, 2:High
6	E	Electrocardiograph	1:Normal, 2:Abnormal
7	R	Exercise Heart Rate	1:Low, 2:High
8	I	Exercise Induced Angina	1:No, 2:Yes
9	D	Heart Disease	1:No, 2:Yes

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



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

P(A,G,C,B,H,E,R,I,D) = P(A)P(G)P(C|D)P(B|A,G)P(H|A,G)P(E|D)P(R|A,D)P(I|D)P(D|B,H)

2 **Likelihood Function:** Suppose that we have N training points $x^{(1)}, \ldots, x^{(N)}$ where $x^{(n)} = [a^{(n)}, g^{(n)}, c^{(n)}, b^{(n)}, h^{(n)}, e^{(n)}, r^{(n)}, i^{(n)}, d^{(n)}]$ for $1 \leq n \leq N$. Using the notation for the parameters of CPTs introduced in the lecture (ie: $P_\theta(D = d|H = h, B = b) = \theta^D_{d|h,b}$), write down the **average** log likelihood of the Bayesian network model as a function of the parameters θ given N data cases **(report)**:

$$\frac{1}{N}\sum_{n=1}^{N} \log p_\theta(x^{(n)}). \tag{1}$$

You may also use notation like $\#(G = g, A = a)$ for "counts" to refer to the number of times a set of particular value $(G = g, A = a)$ occur in the dataset, e.g.

$$\#(G = g, A = a) = \sum_{n=1}^{N} \mathbb{I}[g^{(n)} = g, a^{(n)} = a]. \tag{2}$$

1) $\log p_\theta(x^{(n)}) = \sum_a \mathbb{I}[a^{(n)} = a] \log \theta^A_a + \sum_g \mathbb{I}[g^{(n)} = g] \log \theta^G_g +$
$\sum_a \sum_g \sum_h \mathbb{I}[a^{(n)} = a, g^{(n)} = g, h^{(n)} = h] \log \theta^H_{h|a,g} + \sum_a \sum_g \sum_b \mathbb{I}[a^{(n)} = a, g^{(n)} = g, b^{(n)} = b] \log \theta^B_{b|a,g}$

$\sum_h \sum_b \sum_d \mathbb{I}[h^{(n)} = h, b^{(n)} = b, d^{(n)} = d] \log \theta^D_{d|h,b} + \sum_d \sum_c \mathbb{I}[d^{(n)} = d, c^{(n)} = c] \log \theta^C_{c|d}$
$\sum_d \sum_e \mathbb{I}[d^{(n)} = d, e^{(n)} = e] \log \theta^E_{e|d} + \sum_d \sum_i \mathbb{I}[d^{(n)} = d, i^{(n)} = i] \log \theta^I_{i|d}$
$+ \sum_a \sum_d \sum_r \mathbb{I}[a^{(n)} = a, d^{(n)} = d, r^{(n)} = r] \log \theta^R_{r|a,d}$

2) $\frac{1}{N} \sum_{n=1}^N \log p_\theta(x^{(n)}) = \sum_a \frac{\#[A=a]}{N} \log \theta^A_a + \sum_g \frac{\#[G=g]}{N} \log \theta^G_g + \sum_a \sum_g \sum_h \frac{\#[A=a,G=g,H=h]}{N} \log \theta^H_{h|a,g}$
$+ \sum_a \sum_g \sum_b \frac{\#[A=a,G=g,B=b]}{N} \log \theta^B_{b|a,g} + \sum_h \sum_b \sum_d \frac{\#[H=h,B=b,D=d]}{N} \log \theta^D_{d|h,b} + \sum_d \sum_c \frac{\#[D=d,C=c]}{N} \log \theta^C_{c|d}$
$+ \sum_d \sum_e \frac{\#[D=d,E=e]}{N} \log \theta^E_{e|d} + \sum_d \sum_i \frac{\#[D=d,I=i]}{N} \log \theta^I_{i|d} + \sum_a \sum_d \sum_r \frac{\#[A=a,D=d,R=r]}{N} \log \theta^R_{r|a,d}$

3 **Maximum Likelihood Estimates:** Using the notation for the parameters of CPTs introduced in the leture, derive the maximum likelihood estimate (MLE) for the parameter

$$\theta^R_{r|d,a} = P_\theta(R = r|D = d, A = a) \tag{3}$$

starting from the log likelihood function. What will be the MLE of $\theta^R_{1|1,3}$? Show your work. **(report)**
1) we know that $\theta^R_{1|1,3} + \theta^R_{2|1,3} = 1$
2) $\frac{\delta}{\delta\theta^R_{1|1,3}} \mathcal{L}(\theta|x^{1:N}) = 0$, $\frac{\delta}{\delta\theta^R_{2|1,3}} \mathcal{L}(\theta|x^{1:N}) = 0$
3) $\frac{\delta}{\delta\theta^R_{r|d,a}} \mathcal{L}(\theta|x^{1:N}) = \frac{\#(R=r,D=d,A=a)}{N} \frac{1}{\theta^R_{r|d,a}} - \frac{\#(D=d,A=a)-\#(R=r,D=d,A=a)}{N} \frac{1}{1-\theta^R_{r|d,a}}$
4) $\frac{\#(R=r,D=d,A=a)}{N} \frac{1}{\theta^R_{r|d,a}} = \frac{\#(D=d,A=a)-\#(R=r,D=d,A=a)}{N} \frac{1}{1-\theta^R_{r|d,a}}$
5) remove N $\frac{\#(R=r,D=d,A=a)}{\theta^R_{r|d,a}} = \frac{\#(D=d,A=a)-\#(R=r,D=d,A=a)}{1-\theta^R_{r|d,a}}$
6) from 5) $\#(R = r, D = d, A = a)(1 - \theta^R_{r|d,a}) = (\#(D = d, A = a) - \#(R = r, D = d, A = a))(\theta^R_{r|d,a})$
7) from 6) $\#(R = r, D = d, A = a) = \#(D = d, A = a)\theta^R_{r|d,a}$

8) $\theta^R_{r|d,a} = \frac{\#(R=r,D=d,A=a)}{\#(D=d,A=a)}$

9) $\theta^R_{1|1,3} = \frac{\#(R=1,D=1,A=3)}{\#(D=1,A=3)}$

4 **Uniform Probabilities:** Suppose we choose a distribution with all probabilities set to be uniform. In this case, $P_\theta(A = a) = \frac{1}{3}$, $P_\theta(B = b|G = g, A = a) = \frac{1}{2}$, etc. What will be the log-likelihood for one datapoint $\log p(x)$? Give closed-form solution. Show your work. **(report)**

$logp(1, 1, 4, 1, 1, 1, 1, 1, 1, 1) = \log \theta^A_1 + \log \theta^G_1 + \log \theta^H_{1|1,1} + \log \theta^B_{1|1,1} + \log \theta^D_{1|1,1} + \log \theta^C_{4|1} + \log \theta^E_{1|1} + \log \theta^I_{1|1} + \log \theta^R_{1|1,1} = \log(1/3) + \log(1/2) + \log(1/2) + \log(1/2) + \log(1/2) + \log(1/4) + \log(1/2) + \log(1/2) + \log(1/2) = -7.3369$
5 **Learning:** Implement maximum likelihood learning for all factors in the directed model. **(code)** For this question, use the data in `data-train-1.txt` only. What is the **average** log-likelihood over all the datapoints? Give a specific number rounding to 4 decimal places. How does it compare to the log-likelihood in Question 4? Please use **natural logarithms** to compute your log-likelihood for this and the following questions. **(report)**

For the result I got -6.1952

Compare to uniform distribution, probability of average log likelihood is depends on the training data

6 **Conditioning:** For each of the following probabilities, show how it can be expressed in terms of the factorized joint distribution for the directed model. Simplify the expressions as far as possible. Give the final expression as an equation.

$$\begin{aligned}
&p(e|a, g, c, b, h, r, i, d)\\
&p(b|a, g, c, h, e, r, i, d)\\
&p(a, h|g, c, b, e, r, i, d)\\
&p(d|a, g, c, b, h, e, i)
\end{aligned}$$

Note that r is not present in the last equation! **(report)**
Hint: in each case your answer should be simplified to a single conditional distribution e.g., $p(g|a, c, b, h, e, r, i, d)$ can be simplified to $p(g|a, b, h)$. The most direct solution is to use the model's factorization and derive the answer algebraically; it may also be possible to use conditional independence properties to simplify expressions.

$p(e|a, g, c, b, h, r, i, d) = \frac{P(A)P(G)P(H|A,G)P(B|A,G)P(D|H,B)P(C|D)P(E|D)P(I|D)P(R|D,A)}{\sum_e P(A)P(G)P(H|A,G)P(B|A,G)P(D|H,B)P(C|D)P(E=e|D)P(I|D)P(R|D,A)} = P(e|d)$

$p(b|a, g, c, h, e, r, i, d) = \frac{P(A)P(G)P(H|A,G)P(B|A,G)P(D|H,B)P(C|D)P(E|D)P(I|D)P(R|D,A)}{\sum_b P(A)P(G)P(H|A,G)P(B=b|A,G)P(D|H,B=b)P(C|D)P(E|D)P(I|D)P(R|D,A)} = P(b|a, g, h, d)$

$p(a, h|g, c, b, e, r, i, d) = \frac{P(A)P(G)P(H|A,G)P(B|A,G)P(D|H,B)P(C|D)P(E|D)P(I|D)P(R|D,A)}{\sum_{a,h} P(A=a)P(G)P(H=h|A=a,G)P(B|A=a,G)P(D|H=h,B)P(C|D)P(E|D)P(I|D)P(R|D,A=a)} = P(a, h|g, b, d, r)$

$(d|a, g, c, b, h, e, i) = \frac{P(A)P(G)P(H|A,G)P(B|A,G)P(D|H,B)P(C|D)P(E|D)P(I|D)}{\sum_d P(A)P(G)P(H|A,G)P(B|A,G)P(D=d|H,B)P(C|D=d)P(E|D=d)P(I|D=d)} = P(d|h, b, c, e, i)$

7 **Probability Queries:** Compute each of the following four probabilities, using your maximum-likelihood parameters estimated on `data-train-1.txt.` **(code)**

$$\begin{aligned}
&P(E = 1|A = 2, G = 1, C = 1, B = 1, H = 2, R = 2, I = 1, D = 1)\\
&P(B = 1|A = 3, G = 1, C = 3, H = 1, E = 2, R = 1, I = 1, D = 2)\\
&P(A = 1, H = 1|G = 2, C = 3, B = 2, E = 1, R = 2, I = 1, D = 1)\\
&P(D = 2|A = 1, G = 2, C = 1, B = 1, H = 1, E = 2, I = 2)
\end{aligned}$$

You should round to 4 decimal places.

$$P(E = 1|A = 2, G = 1, C = 1, B = 1, H = 2, R = 2, I = 1, D = 1) = 0.6136$$
$$P(B = 1|A = 3, G = 1, C = 3, H = 1, E = 2, R = 1, I = 1, D = 2) = 0.2920$$
$$P(A = 1, H = 1|G = 2, C = 3, B = 2, E = 1, R = 2, I = 1, D = 1) = 0.0783$$
$$P(D = 2|A = 1, G = 2, C = 1, B = 1, H = 1, E = 2, I = 2) = 0.6735$$

(report)

8 **Test Log-likelihood:** We will follow a standard five-fold cross-validation protocol to assess the performance of our model. For each training file m, use your Bayesian network implementation to learn the parameters of the model, then compute:

- **Average** training log-likelihood for the training points in training file m;
- **Average** test log-likelihood for the test datapoints in test file m.

Give all the computed average (training and test) log-likelihoods as a 5x2 table. Also for each of the two metrics, compute the mean and standard deviation of the five average log-likelihoods. You can assume that every combination of values for a variable and its parents that appear in the test data also appear in the corresponding training data, so you do not have to worry about zeros in the estimated conditional probabilities. **(code, report)**

```
train                  test
-6.207972760799755 -6.500465340866037
-6.237086133492926 -6.412707981858421
-6.299980630507212 -6.129661001734047
-6.217116199969979 -6.460096316363466
-6.223901979821554 -6.405579759515236
mean and std for training
-6.237211540918286
0.03666008640906743
mean and std for testing
-6.381702080067441
0.1460364535733252
```

9 **Classification Accuracy:** For each training file m, use your Bayesian network implementation to learn the parameters of the model, then make predictions for the heart disease variable for each data case n in test file m. **(code)** For each element $x^{(n)}$ of the corresponding test data, compute the most likely value of the heart disease variable given all other variables, i.e.

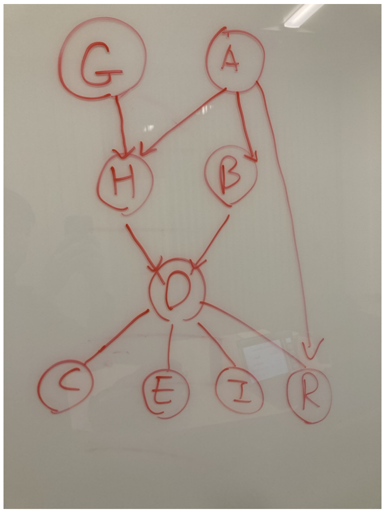
$$\arg\max_{d} p_\theta(d|a^{(n)}, g^{(n)}, c^{(n)}, b^{(n)}, h^{(n)}, e^{(n)}, r^{(n)}, i^{(n)}). \tag{4}$$

Compute the mean prediction accuracy for each of the 5 datasets. Also compute the mean and standard deviation of this metric over the five experiments.

```
0.6166666666666667, test1
0.8166666666666667, test2
0.7333333333333333, test3
0.7333333333333333, test4
0.7833333333333333, test5

mean :0.7366666666666667,
std: 0.07582875444051548
```

(report)



10 **Modeling:** Use your own intuition about heart disease to design your own network structure for the heart disease domain.

- Draw a graphical model for your network (hand-drawn is OK); **(report)**
  i missed arrow d-> c, d->e d->i d-> r

- Write down the factorization for your network; **(report)**

  P(A,G,C,B,H,E,R,I,D) = P(A)P(G)P(C|D)P(B|A)P(H|A,G)P(E|D)P(R|A,D)P(I|D)P(D|B,H)

- Describe the choices that led to your network; **(report)**

  according to the research blood pressure is more likely depends on the size of the human body so I remove realtionship between gender and blood pressure

- Repeat the experiment from the previous question to compute the average training and test log-likelihoods for each of the 5 datasets. **(code, report)**

```
    train                  test
-6.195981991842887 -6.129372167037158
-6.216834666366294 -5.970213587807827
-6.288217749256831 -5.721886955495516
-6.210829428128073 -6.145478392262523
-6.217377814738477 -6.162546296117825
mean and std for training
-6.225848330066512
0.0359183302357072
mean and std for testing
```

```
-6.02589947974417
0.18654932271422509
```

- Repeat the experiment from the previous questions to compute the mean prediction accuracy for each of the 5 datasets. **(code, report)**

```
([0.6166666666666667, data 1
0.8166666666666667,   data 2
0.7333333333333333,  data 3
0.7333333333333333,   data 4
0.7833333333333333],  data 5
0.7366666666666667,   mean
0.07582875444051548)  std
```

You do not have to be better than the example model to get full points.