PREDICTING HEART-DISEASE RISK WITH A BAYESIAN NETWORK

RE: 601

Cole M Helmer

1675690 | 11/18/2024

Abstract:

The objective of this project is to develop an AI-based system within Python. Using a Bayesian Network model, the class was tasked with building a module capable of predicting the likelihood of heart disease within patients. The system processes a data, stored in a local data file, which contains records of various medical attributes, such as age, sex, chest pain type, and fasting blood sugar levels, to come to a prediction of whether or not a patient may suffer from heart disease. The data is analyzed by the Bayesian Network, which characterizes the relationships between the various attributes and their conditional dependencies. The system will then output a percentage indicating the likelihood of heart disease. Which is then categorized into one of five labels, a value of 0 for an absence, and values 1-4 to represent varying levels of severity for when heart disease is detected.

Theory & Principle:

As mentioned above, the goal behind this project is to develop a predictive AI algorithm capable of analyzing and interpreting raw a data set, in order to create a system that can understand and process a set of variable attributes related to a patient's heart health. The data utilized in the project was sourced from patient records concerning heart disease diagnoses at the V.A. Medical Center in Long Beach and the Cleveland Clinic Foundation. The dataset is built from the data of 303 patients, with 76 attributes representing various symptoms, such as age, chest pain type, sex, resting heart rate, and more. For analysis by the system, a subset of 14 key attributes commonly used by experts when diagnosing heart disease, was selected for comparison to assess the risk of heart disease. This analysis of attributes was carried out using a Bayesian Network Model (BNM), with more detailed information about the model provided in the 'Methods' section below. By utilizing the BNM, the systems task was simplified as it just needed to evaluate the provided data and output a prediction indicating the presence or absence of heart disease with its according value in the range one through five. We were then given the values associated with four individual patient's attributes and asked to compare two patients using one BNM and the other two patients using a separate BNM.

Methods:

First the data set provided must be processed and sorted by the system before it can begin utilizing the information. In the case of heart disease, researchers typically rely on a subset of 14 specific attributes in order to help professionals conclude for the diagnosis of a patient. This subset constitutes the core of our model. To ensure the system can correctly interpret the data, we developed methods for understanding each attribute. The first attribute is age, which is stored as a numerical value. Next, sex is encoded as a binary value, with 1 representing male and 0 representing female. For chest pain (denoted in the code as 'cp'), there are four categories, 1 for typical angina, 2 for atypical angina, 3 for non-anginal pain, and 4 for asymptomatic pain. The model also considers resting blood pressure ('trestbps') and cholesterol levels ('chol') in the patient. Another key attribute is whether a patient's fasting blood sugar exceeds 120 mg/dl, if this rings true, the value for 'fbs' is set to 1, otherwise, it is set to 0. The model also processes the results of the patient's resting electrocardiogram ('restecg'). A value of 0 indicates normal results, 1 indicates an ST-T wave abnormality, and 2 indicates definite left ventricular hypertrophy according to Estes' criteria. The system then also considers the maximum heart rate achieved ('thalach'), exercise-induced angina ('exang'), and oldpeak, which represents ST depression induced by exercise relative to rest, all stored as numerical values. Additionally, the slope of the peak exercise ST segment ('slope') and the number of major vessels ('ca') are included, also as numerical values. For the thalassemia attribute ('thal'), values are encoded as 3 for normal, 6 for a fixed defect, and 7 for a reversible defect. Finally, the attribute num represents the model's diagnosis/prediction, which is the output indicating whether the patient has heart disease. The prediction is expressed, as a percentage also known as the posterior probability, for the 'at-risk level' of the patient for that value of heart disease. A

value of 0 indicates no heart disease, while values from 1 to 4 indicate varying levels of heart disease. The model makes this determination by comparing the posterior probability values associated with each level, which represent the likelihood of heart disease presence.

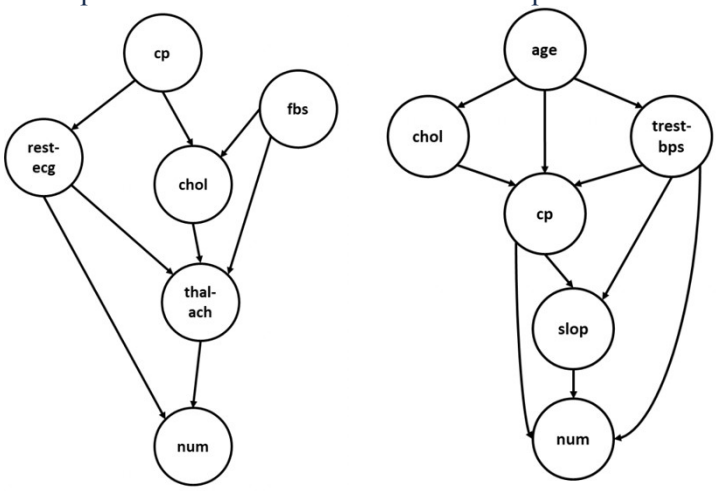

Figure 1. The acyclic graphs used in Predicting Heart-Disease Risk with a Bayesian Network to represent the Bayesian Network Model.

The key theory or driving concept behind this project is that of a '*Bayesian Network Model'*. Also known as a *'Bayes Network'* or a '*Belief Network'*, is a graphical model which represents sets of variables and their conditional dependencies using a directed acyclic graph, examples of these acyclic graphs can be found to the left in figure 1. A BNM is built fundamentally off of *Bayes Rule*, which allows for evidence to be viewed as an effect of some unknown cause. BNM's are often used to display and model uncertainty within relationships between certain variables. For example, in a case like ours where we are analyzing for heart disease, it is commonly known that as you age, your risk for heart disease increases, however our model looks into a multitude of variables all known to have some sort of 'correlation' to heart disease in order to determine the patient's 'at-risk' level.

Results:

The AI-based heart disease prediction system was a success. To the right can be found our 'cleaned-up' data pool. This is how our system sees and understands the information it is receiving from this data, without this processing and organizing of data, this information is a bunch of random letters and numbers to our model. The hardest part of developing this system was simply functionality with certain operating systems and the technicalities that come along with working on projects within the realm of AI. In order to utilize a BNM, the *Python* libraries *'Pandas' & 'PGMPY'*, need to be installed. Normally, a very simple and easy process turned out to be a troubleshooting nightmare. Eventually, the system was developed and able to run proficiently, outputting predictions for the four subjects.

```
Python 3.13.0 (tags/v3.13.0:60403a5, Oct  7 2024, 09:38:07) [MSC v.1941 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.

========== RESTART: C:/Users/gma4r/OneDrive/Apps/Scripts/CHPJ2.py ==========
      age  sex   cp  trestbps   chol  fbs  ...  exang  oldpeak  slope   ca  thal num
0    63.0  1.0  1.0     145.0  233.0  1.0  ...    0.0      2.3    3.0  0.0   6.0   0
1    67.0  1.0  4.0     160.0  286.0  0.0  ...    1.0      1.5    2.0  3.0   3.0   2
2    67.0  1.0  4.0     120.0  229.0  0.0  ...    1.0      2.6    2.0  2.0   7.0   1
3    37.0  1.0  3.0     130.0  250.0  0.0  ...    0.0      3.5    3.0  0.0   3.0   0
4    41.0  0.0  2.0     130.0  204.0  0.0  ...    0.0      1.4    1.0  0.0   3.0   0
..    ...  ...  ...       ...    ...  ...  ...    ...      ...    ...  ...   ...  ..
298  45.0  1.0  1.0     110.0  264.0  0.0  ...    0.0      1.2    2.0  0.0   7.0   1
299  68.0  1.0  4.0     144.0  193.0  1.0  ...    0.0      3.4    2.0  2.0   7.0   2
300  57.0  1.0  4.0     130.0  131.0  0.0  ...    1.0      1.2    2.0  1.0   7.0   3
301  57.0  0.0  2.0     130.0  236.0  0.0  ...    0.0      0.0    2.0  1.0   3.0   1
302  38.0  1.0  3.0     138.0  175.0  0.0  ...    0.0      0.0    1.0    ?   3.0   0
```

Figure 2. The output of the data set after being processed and categorized by the system.

Discussion:

```
+--------+------------+--------+------------+
| num    |   phi(num) | num    |   phi(num) |
+========+============+========+============+
| num(0) |     0.3750 | num(0) |     0.0000 |
+--------+------------+--------+------------+
| num(1) |     0.1250 | num(1) |     0.5000 |
+--------+------------+--------+------------+
| num(2) |     0.2500 | num(2) |     0.0000 |
+--------+------------+--------+------------+
| num(3) |     0.1250 | num(3) |     0.5000 |
+--------+------------+--------+------------+
| num(4) |     0.1250 | num(4) |     0.0000 |
+--------+------------+--------+------------+
```

Figure 3. Above are the posterior probability charts for patient 1(left) & patient 2(right).

The first two subjects' posterior probability output charts can be found to the left in figure 3. These two were analyzed using the graph on the right from Figure 1 above. Subject 1 had recorded an age of 65, 'chol' at 230, 'cp' level 4, a 'trestbps' of 130 and a 'slope' valued 2. With that record inputted into the system, it is assumed that Subject 1, has a 37.50% chance of an absence or no heart disease at all. However, the system also predicts a 12.50% chance at Subject 1 having type 1, 3, or 4 heart disease and a 25% chance at type 2 heart disease. Analyzing Subject 2 now, who had reported an age of 51, 'chol' of 260, 'cp' level 2, 'trestbps' of 120, and a slope of 2. As shown in Figure 3, our system has

determined that heart disease is present in our subject, however it is a 50% chance of being type 1 and a 50% chance of being type 3.

num	phi(num)	num	phi(num)
num(0)	0.6000	num(0)	0.0000
num(1)	0.2000	num(1)	0.0000
num(2)	0.0000	num(2)	0.0000
num(3)	0.2000	num(3)	1.0000
num(4)	0.0000	num(4)	0.0000

Figure 4. Posterior Probability output charts for Subject 3(left) & Subject 4(right)

Moving on to Subjects 3 & 4 now, which were analyzed using the graph on the left in Figure 1. Subject 3, had reported a 'cp' type 3, 'chol' at 260, 'restecg' at 2, 'fbs' at 0 and a 'thalach' of 160. Shown to the left, Subject 3 is expected to have an absence of heart disease at 60% while there is still a 20% chance for both type 1 & 3 heart disease. The fourth and final subject reported a 'cp' type 2, 'chol' of 260, 'restecg' at 2, 'fbs' of 1 and a 'thalach' of 1. Our system has come to an absolute decision for Subject 4. That prediction being there is a 100% chance that Subject 4 has type 3 heart disease.

Conclusion:

AI algorithms, much like the one built in this project, can prove to be incredibly valuable as society grows to embrace technology/AI in healthcare. While many still hesitate to entrust their absolute health and common well-being to machines, yet this algorithm is not designed to replace doctors. Instead, it is a useful predictive tool that serves alongside that of healthcare professionals to help in the diagnosis process. The system built, can identify the potential for heart disease by analyzing just 14 key symptoms or attributes. This approach allows for early detection in patients, as patterns in underlying symptoms become more apparent over time with the growth of the datasets. For example, imagine a future where a predictive algorithm similar to the one discussed above, is developed yet it's for breast cancer. If the system detects such cancer early, imagine if it detected such a disease within a family member early on, it could allow for the proper assistance to be given in order for recovery, potentially saving their life. Without such a model, doctors might not identify the condition until it is too late.

Appendix:

```
import pandas as pd
from pgmpy.models import BayesianNetwork
from pgmpy.estimators import MaximumLikelihoodEstimator, BayesianEstimator
from pgmpy.inference import VariableElimination

fnames= "age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,oldpeak,slope,ca,thal,num"
fnames=fnames.split(",")

#classify & organize data
dat=pd.read_csv("processed.cleveland.data",names=fnames)
print(dat)

#First Bayesian Model#

#Subject 1
model=BayesianNetwork([("age","chol"),("age","cp"),("age","trestbps"),("chol","cp"),("trestbps","cp"),
("cp","slope"),("cp","num"),("trestbps","slope"),("trestbps","num"),("slope","num")])
model.fit(dat,estimator=MaximumLikelihoodEstimator)
infer=VariableElimination(model)
p11=infer.query(['num'],evidence={'age':65,'cp':4,'trestbps':130,'chol':230,'slope':2})
print(p11)

#Subject2
model=BayesianNetwork([("age","chol"),("age","cp"),("age","trestbps"),("chol","cp"),("trestbps","cp"),
("cp","slope"),("cp","num"),("trestbps","slope"),("trestbps","num"),("slope","num")])
model.fit(dat,estimator=MaximumLikelihoodEstimator)
infer=VariableElimination(model)
p12=infer.query(['num'],evidence={'age':51,'cp':2,'trestbps':120,'chol':260,'slope':2})
print(p12)

#Second Bayesian Model#

#Subject 1
model=BayesianNetwork([("cp","chol"),("cp","restecg"),("fbs","chol"),("fbs","thalach"),
("restecg","thalach"),("restecg","num"),("chol","thalach"),("thalach","num")])
model.fit(dat,estimator=MaximumLikelihoodEstimator)
infer=VariableElimination(model)
p21=infer.query(['num'],evidence={'cp':3,'chol':260,'restecg':2,'thalach':160,'fbs':0})
print(p21)

#Subject 2
model=BayesianNetwork([("cp","chol"),("cp","restecg"),("fbs","chol"),("fbs","thalach"),
("restecg","thalach"),("restecg","num"),("chol","thalach"),("thalach","num")])
model.fit(dat,estimator=MaximumLikelihoodEstimator)
infer=VariableElimination(model)
p22=infer.query(['num'],evidence={'cp':2,'chol':260,'restecg':2,'thalach':120,'fbs':1})
print(p22)
```